Parametrisations for Inference on the Dependency Structures in Multivariate Markov Chains

Submitted Oct 2025, in partial fulfillment of
the conditions for the award of the degree **PhD Mathematics.**

Jiale Tao
16522227

Supervised by Theodore Kypraios, Karthik Bharath

School of Mathematical Science
University of Nottingham

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature _____

Date _____ / _____ / _____

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation via the University of Nottingham's e-dissertation archive.

Public access to this dissertation is restricted until: DD/MM/YYYY

Abstract

Multivariate Markov Chains (MMCs) provide a powerful framework for capturing dependencies across multiple interrelated processes that evolve over time. In literature, the natural way to represent MMC dynamics is through a joint transition matrix on the extended state space formed by the Cartesian product of all marginal state spaces. However, such a joint transition matrix by the Cartesian product does not explicitly reveal the underlying local dependency structures—for example, whether one group of chains depends on another, or whether interactive effects exist between groups.

To address this, this thesis develops two novel reparametrisations, the *Univariate Marginal Model* and the *Hierarchical Marginal Model*, which re-express the joint transition matrix in terms of marginal probabilities. These probabilities are further decomposed via multinomial logistic regression, yielding regression coefficients that not only provide clear interpretations of dependency structures but also allow explicit encoding and detection of these dependency structures such as conditional independence, contemporaneous independence, and Granger non-causality.

In many practical applications, the underlying MMCs are unobserved, giving rise to Hidden Markov Models (HMMs) where the hidden dynamics have to be inferred from the emitted observations. Classical methods such as the Forward–Backward algorithm become computationally intractable as the number of chains grows. To solve this, we develop the *Generalised Individual Forward–Backward* (GIFB) algorithm, an exact (in Monte Carlo sense) Gibbs sampler that updates hidden states group-wise, reducing computational cost from exponential to quadratic in the number of chains. Since GIFB operates on marginal transition probabilities, it naturally integrates with our earlier marginal models, enabling efficient full inference in Multivariate Hidden Markov Models (MHMMs). Finally, we applied this framework to real Twitter/X data in the context of UK railways to detect dependency structures and disruptions, demonstrating its effectiveness in practice.

Acknowledgements

I would first like to express my sincere gratitude to my supervisors, Theodore Kypraios and Karthik Bharath. Theo has always been patient, supportive, and encouraging, and has provided invaluable help whenever I encountered difficulties in my research. Karthik, while strict and precise, consistently pushed me to think more carefully and rigorously, which greatly improved the quality of my work. I feel extremely fortunate to have spent more than four years working under their guidance, and I am deeply grateful for everything they have taught me. I would also like to thank my examiners, Richard Wilkinson and Panayiota Touloupou, for taking the time to carefully read my thesis and for providing insightful comments and valuable suggestions.

I am grateful to the School of Mathematics and to UK Research and Innovation (UKRI) for providing the funding and financial support that made my studies in Nottingham possible.

I would like to thank my friends in the UK—Huachen Su, Jiaxiang Li, Weiyi Peng, and especially Haodong Zhu, who has been my close friend for more than 15 years since middle school. Their companionship made my life in the UK far more enjoyable, and I am grateful for the many memorable moments we shared in Nottingham.

I would also like to express my deepest gratitude to my parents and grandparents for their unwavering financial and emotional support throughout my studies in the UK. In particular, I wish to thank my grandfather, who enlightened my interest in mathematics during my primary school years-an interest that ultimately led me to pursue my studies through to the PhD level.

Lastly, I would like to thank my wonderful girlfriend, Ruoxi Pan, for standing by my side throughout the past years of my PhD. Without her support, my PhD journey would have been far more difficult. I am truly fortunate to have her, with whom I can share both my happiness and my struggles.

Contents

List of Tables

List of Figures

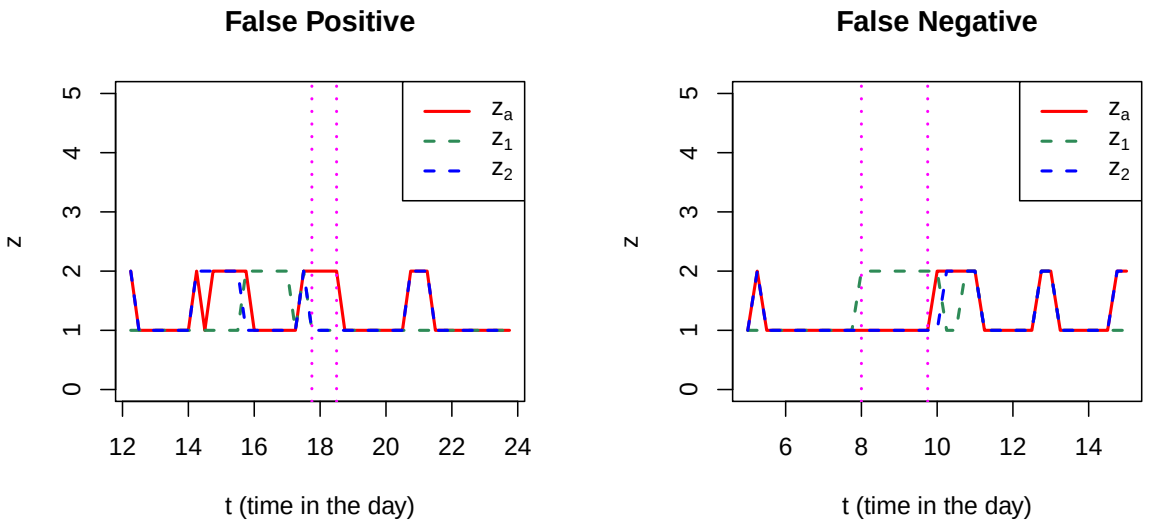
Chapter 1

Introduction

1.1 Motivation and Objectives

Consider the following problems: predict if the value of a particular stock within a portfolio will go above a threshold; model the probability of failure of a system of connected electrical components, before a fixed time; predict if a bank within a financial network will experience a catastrophic loss; if an entire railway network experiences delay because of delays on one or two of its routes. A first-order multivariate Markov chain (MMC) model, comprising several univariate Markov chains, in which the joint state at each time depends only on the immediately preceding joint state, is a sensible statistical model to address the problems. In fact, the last of the problems on delays in a UK train network will be considered in Section 5.5.

The problems seek answers at both the individual (e.g. fate of a single bank) and collective levels (e.g. delay in an entire rail network). Summarizing K univariate chains $Z_1, \ldots, Z_K$ via an aggregate (e.g. $\sum_k Z_k$) will be inappropriate. Figure 1.1 below illustrates how this can lead to incorrect conclusions on a toy example, wherein the aggregate's behaviour is markedly different from that of the individual chains. Modelling and quantifying dependency between $Z_1, \ldots, Z_K$ is a crucial requirement in the above problems; it is important to be able to directly access how the behaviour of a single or a subset of chains influence the rest or a subset thereof.



(a) False Positive (b) False Negative

Figure 1.1: Motivating binary example: z_a is inferred from the aggregated process of z_1 and z_2

Much of the existing literature on MMC models is based on the joint probability transition matrix, which encodes the conditional probability $\mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)\big)$ of the state of the K-dimensional chain $\boldsymbol{Z}$ at time t given its state at time $t-1$. Accessing information on how the subset $\boldsymbol{Z}_j$ influences another subset $\boldsymbol{Z}_{j'}$ with $g_j, g_{j'} \subset \{1, \ldots, K\}, g_j \cap g_{j'} = \emptyset$ via the joint transition probability is difficult since these are derived quantities from $\mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)\big)$, and subject to constraints in a composition of addition followed by multiplication. As such, the dependence structure between the individual chains within $\boldsymbol{Z}$ is not transparent. This issue will be detailed in Section 2.8.2.

Inferring the dependency structure of $\boldsymbol{Z}$ is exacerbated when data is unavailable on $\boldsymbol{Z}$ but instead on a variable $\boldsymbol{X}$; for example, in the problem of predicting delay in the train network to be considered in Section 5.5, $\boldsymbol{X}$ is the number of tweets about delays on individual trains within the network. This represents a case wherein the MMC is a hidden (i.e. not directly observed) Markov chain. Computational challenges are numerous in this setting, mainly owing to large dimension of the parameter space of $\boldsymbol{Z}$.

Dependence between component chains $Z_1, \ldots, Z_K$ in a first-order MMC manifest in two ways (i) given the past, Z_k is independent of each other or $\boldsymbol{Z}_j$ is independent of $\boldsymbol{Z}_{j'}$ for two disjoint subsets $g_j, g_{j'} \subset \{1, \ldots, K\}$, known as *conditional* (see Section 2.4) and *contemporaneous* independence (see Section 2.5), respectively; (ii) current state of $\boldsymbol{Z}_j$ does not influence the future state of $\boldsymbol{Z}_{j'}$, given the current state of the rest $\{1, \ldots, K\} \setminus g_j$, known as the *Granger non-causality* (see Section 2.6). As such, any dependence structure may be represented by a simple graph on $2K$ nodes, representing the states of the K components at times $t-1$ and t, with directed edges. For instance, the following Figure 1.2 illustrates the case when $K = 3$.

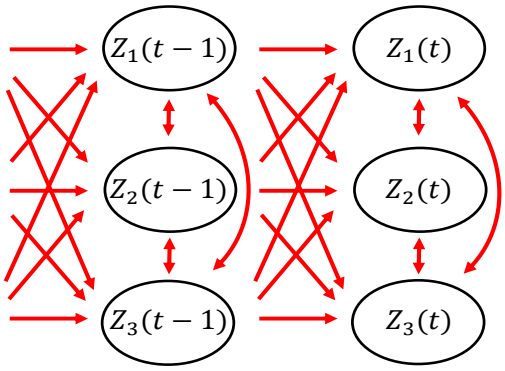


Figure 1.2: The dependency graph for MMC with $K = 3$ chains under full dependencies

Inferring such a dependency graph from data on $\boldsymbol{Z}$ or on another variable $\boldsymbol{X}$ when $\boldsymbol{Z}$ is hidden is key to accurate modelling and prediction using MMCs. Such a graph, however, is not directly available from inference on a joint transition matrix, but instead requires inferring a constrained transition matrix.

Consequently, for a first-order MMC $\boldsymbol{Z} = (Z_1, \ldots, Z_K)$ in discrete time with a common state space $\mathcal{S}$, the objectives of the work in this thesis are as follows:

(a) develop novel parameterizations that directly encode the dependence structure amongst the components of $\boldsymbol{Z}$ in a manner that allows for the parameters to be interpreted as quantifying correlations between the individual chains;

(b) utilize the parameterizations to develop efficient computational algorithms for inference on $\boldsymbol{Z}$ under the high-dimensional parameter space that when it is not directly observable.

1.2 Literature Review

1.2.1 Multivariate Markov Chains (MMC)

First introduced by Markov [1906], Markov chains provide a fundamental framework for modelling stochastic systems that evolve over time under the memoryless property, where the current state depends only on the immediately preceding state. A natural extension of this framework is the multivariate Markov chain (MMC), which combines several univariate chains into a joint system. In this setting, the state evolves in multiple dimensions simultaneously, enabling the model to capture dependencies across interacting stochastic processes [Howard, 2012, Norris, 1997].

As a result, MMCs have been widely applied across diverse fields to model the dependencies, for instance finance Bernardi et al. [2013], cyber security Abraham and Nair [2014] and reliability engineering Yera et al. [2021]. However, none of these works explicitly specifies the dependency structures within the MMC. Specifically, Bernardi et al. [2013] developed a multivariate Markov-switching model to study tail risk interdependence across asset returns. However, all the dependencies are embedded in the covariance matrix of the regime-specific multivariate Student t distribution, rather than through an explicitly defined dependency structure within the Markov chain itself. In cybersecurity, Abraham and Nair [2014] applied absorbing Markov chains to quantify security risks. Here, dependencies among vulnerabilities and exploits were represented implicitly through the associated transition probabilities and did not provide a formal specification of the dependency structure. Yera et al. [2021] captured only the positive correlation between railway door failures, without specifying a broader or more general dependency structure for the MMCs. This highlights the importance of moving beyond implicit treatments to

frameworks that define dependency structures explicitly.

1.2.2 Dependency Structures within MMC

In the literature, three main notions of dependence have been developed for MMCs: *conditional independence*, *contemporaneous independence*, and *Granger non-causality*. The first two concepts characterize concurrent interactions among components at a given time step, whereas Granger non-causality, a concept borrowed from time series analysis, formalizes the predictive influence of one component's past on another within the Markov setting. As we will demonstrate in Section 2.7, the combination of these three notions provides a complete framework for formalizing all possible dependency structures within a first-order MMC.

The most widely adopted framework is *conditional independence*, under which each individual chain is assumed to evolve independently of the others when conditioned on the previous states [Saul and Jordan, 1999]. This independence relationship removes all the concurrent interactions among the chains while retaining the influence from the past states as in the Markov setting. Additionally, conditional independence has been central to the development of Coupled Hidden Markov Models (CHMMs), originally formalized by Brand et al. [1997] and subsequently applied in areas such as speech and activity recognition [Natarajan and Nevatia, 2007], signal processing [Nefian et al., 2002], and in epidemic modelling [Sherlock et al., 2013, Touloupou et al., 2020].

Nevertheless, conditional independence has often been criticized as overly restrictive, since it eliminates all concurrent interactions. To address this limitation, Richardson [2003] introduced the concept of separation in the context of the Markov property for mixed graphs, which allows for independence relations between groups of vertices. Building on this framework, Colombi and Giordano [2012] adapted the idea to the MMCs and formalized *contemporaneous independence*, where groups of chains are independent of another conditioned on the past. Contemporaneous independence generalises conditional inde-

pendence by allowing partially existing concurrent interaction: interactions are preserved within groups while removed across groups, thus providing a more flexible way to structure dependence among components.

While both conditional independence and contemporaneous independence focus on concurrent interactions among component chains, *Granger non-causality* specifies the influence from the past states, consistent with the Markovian framework. Originally developed in econometrics [Granger, 1969a], Granger non-causality formalizes the idea that the past of one process does not improve the prediction of another. Colombi and Giordano [2012] adapted this concept to multivariate Markov chains by embedding it within a mixed-graph framework, which allows one group of chains to be independent of another group's past, conditional on the remaining chains. This provides a principled way to represent settings where some dependency structures among chain groups are absent.

1.2.3 Parameterizations for MMC

Alongside the dependency structures discussed previously, several parameterizations had been developed for MMCs, among which the most widely used was the *Cartesian Product Model* [see, e.g., Brand et al., 1997, Ghosh et al., 2017, Pohle et al., 2021]. This approach constructed a joint transition matrix over an extended state space defined as the Cartesian product of the state spaces of the individual chains, to capture the dynamics of the MMC. Each entry within the joint transition matrix specified the probability of moving from one joint state to another, analogous to the transition probabilities in the univariate Markov setting. Its popularity in the literature stemmed not only from its ability to fully represent both within-chain and cross-chain dependencies in a first-order MMC, but also from the straightforward extension it offered to existing inference methods from the univariate case. However, as discussed in Section 2.8.2, these dependencies were only implicitly encoded in the joint transition probabilities, which were expressed through complex summations and multiplications. This implicit representation complicated the detection and testing of potential dependency structures. More details will be reviewed

in Section 2.9.2.

To explicitly represent dependency structures, an alternative was to use models based on linear decompositions. A well-known example was the *Inter-Chain Transition* (ICT) model proposed by Zhong and Ghosh [2001]. In this approach, the joint transition probabilities were factorized into marginals under conditional independence, and each marginal was then expressed as a linear combination of weighted univariate transitions. The coupling weights naturally reflected the marginal influence of one chain on another. More details will be reviewed in Section 2.9.1.

Another prominent linear decomposition method was the *Mixture Transition Distribution* (MTD) model introduced by Raftery [1985], which expressed an l^{th}-order univariate Markov chain as a convex mixture of single-lag transitions, providing a parsimonious alternative to full higher-order representations. Multivariate extensions [Ching and Ng, 2006, Ching et al., 2004] applied the same mixture idea across chains, where each chain's next-step distribution was a weighted sum of contributions from all chains, encoded in a structured block transition operator Q. To simplify inference under simplex constraints on mixture weights, Nicolau [2014] proposed the *MTD-Probit* model, which retained the linear mixture structure but replaced constrained weights with unconstrained probit coefficients while preserving normalization. More details will be reviewed in Section 2.9.2.

However, the linear decomposition–based models such as ICT and MTD suffered from several limitations. Firstly, the linear specification of marginal transition probabilities restricted their ability to capture non-linear or more complex dependency structures. Second, their reliance on the conditional independence prevented them from representing concurrent interactions. Finally, the introduction of the additional coupling parameters can lead to the potential non-identifiability issue.

Lastly, we reviewed a class of marginal models that were most closely related to our work.

Colombi et al. [2013] first introduced the marginal modelling scheme using log-odds contrasts, which could be naturally incorporated into the logistic regression framework of McCullagh and Nelder [1989], with the resulting regression coefficients providing meaningful interpretations. This setting was later reviewed by Rudas and Bergsma [2023]. However, as argued by Lupparelli et al. [2009], under the GM logistic framework the regression coefficients lay in an unconstrained parameter space only in the case of two chains. When extended to three or more chains, no closed-form inverse mapping from marginal contrasts to the joint distribution existed, rendering the model invalid. More recent work by Colombi et al. [2024] is still limited to the binary chain. Building on the work of Colombi et al. [2013], we model the marginal probabilities directly, ensuring that the inverse mapping to the joint distribution is a bijection (see Section 4.2).

1.2.4 Hidden Markov Models and Their Inference Frameworks

In many practical applications, however, the Markov chains are unobserved with only an associated process is observed instead. This naturally leads to the framework of *Hidden Markov Model* (HMM) [Baum et al., 1970, 1972], where only a process emitted by the hidden Markov process is available. Since then, HMMs have been widely applied across diverse domains. Unlike MMCs, which could be inferred directly using both Frequentist (Section A.1) and Bayesian (Section A.2) approaches, HMMs required more advanced techniques to recover the hidden chains from observed data, which then can be utilized to investigate the potential dependency structures.

The canonical approach in the literature was the *Forward–Backward* (FB) algorithm [Baum et al., 1970, 1972], which factorized the marginal probability of each hidden state into the product of the forward and backward probabilities, both of which could then be computed recursively. Building on this, Baum et al. [1970, 1972] introduced the Expectation–Maximization–based *Baum–Welch* (BW) algorithm, in which the FB procedure served as the E-step to estimate the hidden chains. While the FB algorithm was efficient in univariate cases, it quickly became computationally prohibitive in multivariate settings.

As noted by Carter and Kohn [1994], extending the FB algorithm to the multivariate case required applying it to an equivalent univariate HMM defined on the Cartesian product of the individual state spaces. In this setting, the forward and backward recursions operated over an exponentially growing joint state space, causing the computational cost to scale accordingly and rendering the computation infeasible.

Alternatively, by adopting the data-augmentation strategy, Fearnhead and Sherlock [2006], Sherlock et al. [2013] proposed a Bayesian Gibbs sampler that iteratively samples from the full conditionals of the hidden chain and the emission parameters. However, updating the hidden chain still relied on the FB algorithm, whose cost scaled exponentially with the number of chains. To address this, several approximate alternatives to FB have been developed, including *approximate Bayesian computation* [Martin et al., 2014, Visani et al., 2021] and *variational inference* [Foti et al., 2014, Zhai et al., 2024]. While these approaches substantially reduced computational cost, they achieved this through approximation, which might obscure the true dependency structures, thereby limiting their suitability for our purposes.

As a result, more scalable yet exact samplers are required. Touloupou [2016], Touloupou et al. [2020] proposed the *individual Forward Filtering Backward Sampling* (iFFBS) algorithm, which operates as an exact Gibbs sampler by updating each chain iteratively rather than updating all chains jointly as in the standard FB algorithm. This modification substantially reduces the computational burden, replacing the exponential dependence of FB with a quadratic dependence on the number of chains. However, since iFFBS was developed within the CHMM framework, it inherently relies on conditional independence, without which its recursion is not valid. Building on this Gibbs-like formulation, our work introduces an alternative way of expanding the likelihood that allows the method to extend naturally to general multivariate hidden Markov chains without assuming conditional independence.

1.3 Contributions and Structure of the Thesis

The thesis consists the following contributions to the literature of identifying the potential dependency structures within the multivariate Markov chains and the hidden multivariate Markov chains:

1. We present a novel geometric interpretation for each of the dependency structures, including *Condition Independence* in Section 2.4.2, *Contemporaneous Independence* in Section 2.5.2 and *Granger non-Causality* in Section 2.6.2.

2. We develop two novel reparameterizations, namely the *Univariate Marginal Model* (Chapter 3.2) and the *Hierarchical Marginal Model* (Chapter 4.2) for the joint transition matrix. We further incorporate both models with multinomial logistic regressions to obtain the corresponding regression coefficients that effectively characterize the dependency structures embedded within MMCs.

3. We develop a novel *Generalised Individual Forward Backward* algorithm upon the iFFBS by Touloupou [2016], Touloupou et al. [2020] in Section 5.3.4 and incorporate it with the marginal transition probabilities modelled within the *Hierarchical Marginal Model*. This enables efficient sampling of the hidden states from the exact full conditionals under a general multivariate hidden Markov model, as detailed in Section 5.3.6.

This chapter aims to motivate our research on identifying dependency structures in MMCs under both observed and unobserved settings. It also provides a detailed review of the existing literature, emphasizing the need for more effective approaches to detect dependency structures in multivariate contexts.

Chapter 2 begins by reviewing dependency structures in MMCs, including *conditional independence*, *contemporaneous independence*, and *Granger non-causality*, accompanied by a novel geometric interpretation of these relationships. Proposition 2.18 is then established, stating that all dependency structures within a first-order MMC can be expressed

as combinations of contemporaneous independence and Granger non-causality. With these foundations in place, the chapter reviews the limitations of existing modelling approaches in some detail, most notably the Cartesian Product Model and the linear decomposition based models in representing such dependencies.

To address these limitations, Chapter 3 introduces the *Univariate Marginal Model*, which reparameterizes the joint transition probabilities into marginal transition probabilities and auxiliary probabilities conditioned additionally on the concurrent states. By incorporating multinomial logistic regression, the model yields regression coefficients that explicitly encode dependency structures in an interpretable form. Beyond inference, the Univariate Marginal Model is also formulated as a generative framework, enabling the simulation of MMCs with specified dependency structures directly through the regression coefficients. The effectiveness of this approach is subsequently evaluated through a comprehensive simulation study.

In parallel, Chapter 4 introduces the *Hierarchical Marginal Model*, which constructs the marginal transition probabilities hierarchically, from the univariate up to higher-orders. Through the use of multinomial logistic regression, this model produces regression coefficients that provide a direct and interpretable representation of dependency structures. As with the Univariate Marginal Model, the framework serves both inferential and generative purposes, allowing the simulation of MMCs with specified dependencies via the regression coefficients. Its performance is then systematically assessed through a dedicated simulation study.

Chapter 5 shifts the focus to the setting where the Markov process is unobserved and only data from another related process are available. The chapter initiates by introducing the framework of the multivariate hidden Markov models, highlighting the challenge of recovering hidden paths. Building on the classical approaches such as the *Forward–Backward* (FB) algorithm and the *individual Forward Filtering Backward Sampling* (iFFBS) algo-

rithm, the chapter presents the novel *Generalised Individual Forward Backward* (GIFB) algorithm, which accommodates full dependency structures in general MHMMs while maintaining feasible computational cost. Operating on marginal transition probabilities, GIFB naturally integrates with the Hierarchical Marginal Model. Its performance is then assessed through simulation studies by comparing with the standard FB. It is further applied to real Twitter data on UK railways to infer dependency structures among railway groups and identify disruptive events.

Chapter 2

The Dependency Structure of the Multivariate Markov Chain

2.1 Introduction

2.1.1 Motivation and Aim of the Chapter

Consider a simplified ecological system consisting of grass, rabbits, and wolves, where the population level of each species evolves over time and takes discrete values (e.g. low, medium, high). Let $G(t), R(t)$, and $W(t)$ denote the population levels of grass, rabbits, and wolves at time t, respectively. We model the joint dynamics of this system using a first-order multivariate Markov chain, whose dependency structure is illustrated in Figure 2.1.

The red arrows denote the self-dependence, reflecting the tendency of each population to persist across consecutive time points. The black arrows capture cross-species causal effects from the previous time step: for example, the rabbit population at time t depends on both grass availability (food supply) and wolf population (predation pressure), while the wolf population depends primarily on the rabbit population and not directly on grass. Finally, the blue arrows denote contemporaneous interactions within the same time step, reflecting immediate interactions between grass and rabbits, and between rabbits and

wolves.

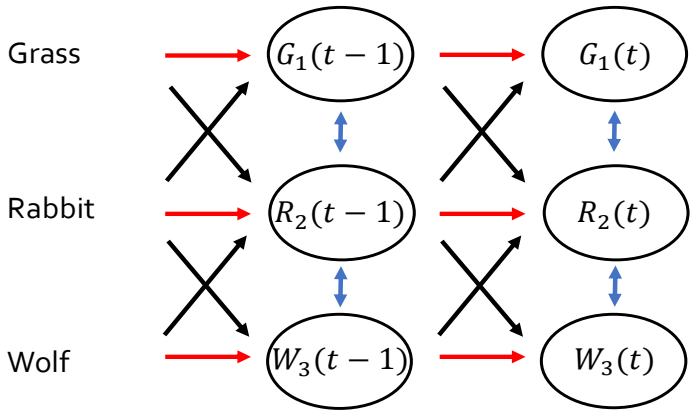


Figure 2.1: The dependency graph for the motivating example on grass-rabbit-wolf system.

While such dependency structures are explicit and interpretable in ecological systems, they are typically unknown in more complex settings, such as interacting stock prices or interconnected machine components, thus highlighting the importance of flexible models capable of learning dependency structures in multivariate Markov systems. In literature, MMCs have long been modelled using transition matrices, as they fully encode the dynamics of all possible state transitions [e.g. Brand et al., 1997, Ghosh et al., 2017, Pohle et al., 2021]. Alongside this, various dependency structures have also been studied—such as *conditional independence* [Saul and Jordan, 1999], widely used in Coupled Hidden Markov Models [e.g. Brand et al., 1997, Natarajan and Nevatia, 2007, Touloupou et al., 2020]; *contemporaneous independence* [Colombi and Giordano, 2012, Lupparelli et al., 2009], which generalises conditional independence; and *Granger non-causality*, a concept borrowed from time series [Granger, 1969a] that has been adapted to MMCs by Colombi and Giordano [2012].

Although transition matrices capture the full joint dynamics of the MMCs, they do not explicitly reflect the underlying dependency structures, such as how the current states depend on the past and the concurrent states across the chains. This leads to a discon-

nection between transition matrix representations and the dependency structures that are extensively studied in the literature. Bridging this gap is crucial, as it not only enables the straightforward inference of such dependencies within MMCs but also facilitates the explicit modelling of MMCs under these structural assumptions.

Hence, the primary goal of this chapter is to establish a connection between these first order dependency structures and the transition matrix. More specifically, we aim to develop a convenient approach to interpret these dependency structures within the framework of transition probability matrices through the marginal transition probabilities. To make this abstract concept more tangible, we further investigate these relationships through a novel geometric perspective and provide visualizations in low-dimensional settings.

In parallel, we review several modelling frameworks from the literature that have been proposed to represent MMCs, highlighting the difficulty of these models in modelling dependency structures. These include the *Cartesian Product Model* [see Brand et al., 1997, Ghosh et al., 2017, Pohle et al., 2021, and others], which aggregates the joint state space using a Cartesian product of individual chains; the *Inter-Chain Transition (ICT) Model* [Zhong and Ghosh, 2001], which expresses marginal transition probabilities as linear combinations of the inter-chain transition probabilities; and the *Mixture Transition Distribution (MTD) model*, originally introduced by Raftery [1985] for high-order univariate Markov chains, which represents the transition probabilities as weighted averages of single-lag transitions. This MTD formulation significantly reduces the number of parameters compared to conventional high-order models. Ching and Ng [2006], Ching et al. [2004] extended MTD model to incorporate with MMC, enabling it to capture both temporal and cross-chain dependencies with a structured transition matrix that brings interpretability. More recently, Nicolau [2014] proposed the *MTD-Probit* model, which embeds the linear mixture of past influences within a Probit link function, freeing the weight constraints of the original MTD model and hence facilitating the parameter estimation.

2.1.2 Chapter Layout

Based on the definitions of the dependency structures, the main contribution of this chapter is twofold. First, we derive how each dependency structure induces explicit constraints on the parametrisation in terms of the joint transition probabilities. Second, we give a geometric interpretation of these constraints, illustrating the resulting parameter space and providing visualisations in low-dimensional settings.

Specifically, Section 2.2 specifies the general notation for MMCs, along with the definition of the joint transition matrix P encoding the MMC. Section 2.3 outlines the dependency structure of the MMCs, encompassing both the effects of the past states and the concurrent interactions among the chains. Based on this structural insight, we introduce the concept of marginal transition probabilities to formally capture these dependencies. In Section 2.4, we present the conditional independence assumption, and in Section 2.4.1 we explore the basic properties of the conditional independence assumption including the structure it imposes to the joint transition matrix as well as the bijection it induces between the joint and the marginal transition probabilities. A two-dimensional visualization of this is provided in Section 2.4.2. Building on this, Section 2.5 introduces the contemporaneous independence. Section 2.5.1 examines the constraints it imposes on the transition matrix via the marginal transition probabilities, while Section 2.5.2 shows how this structure can be reduced to conditional independence through specific projections. After that, Section 2.6 introduces the Granger non-causality. Section 2.6.1 describes the structures it imposed on the joint transition matrix, via the column-wise constraints on the marginal transition probabilities, and Section 2.6.2 provides its geometric interpretation. Section 2.7 summarizes the dependency assumptions discussed previously and formalizes Proposition 2.18 that connects them with all the dependency structure within a first-order MMC.

With these dependencies in place, the chapter then turns to a review of existing parametrisations for MMCs, highlighting how they model the underlying dependency structures. Section 2.8 introduces the Cartesian Product Model, a natural and widely used approach

to modelling MMCs. We discuss how this model encodes various dependency structures in Section 2.8.2, followed by a evaluation of its advantages and limitations in Section 2.8.3. In Section 2.9, we review models based on linear decomposition, including the Inter-Chain Transition model proposed by Zhong and Ghosh [2001] and the Mixture Transition Distribution model proposed by Ching and Ng [2006], Ching et al. [2004]. These models and their ability to capture dependencies are discussed in Section 2.9.3. Finally, Section 2.10 highlights the key findings of this chapter and offers general concluding remarks.

2.2 Notation for the Multivariate Markov Chain

Let $[K] = \{1, 2, \ldots, K\}$ be the set of indices labelling a collection of chains. Define

$$\boldsymbol{Z} = \{\, Z_k : k \in [K]\} = \{\, Z_k(t) : k \in [K],\, t = 1, 2, \ldots T\},$$

as a first-order time-homogeneous Multivariate Markov chain, evolving over the discrete time points $t = 1, 2, \ldots, T$, where $T \in \mathbb{N}^+$.

For each time points t, the vector $\boldsymbol{Z}(t) = \{Z_k(t) : k \in [K]\}$ is a discrete random vector, where each component $Z_k(t)$ takes values from its discrete state space $\mathcal{S}_k = \{1, 2, \ldots, s_k\}$ where $s_1, \ldots, s_K \in \mathbb{N}^+$. For simplicity, we assume that all chains share the same state discrete space $\mathcal{S} = \{1, 2, \ldots, s\}$, meaning that $s_1 = \cdots = s_K = s$ for some $s \in \mathbb{N}^+$. As a result, the joint state space $\boldsymbol{\mathcal{S}}$ for $\boldsymbol{Z}$ can be written as:

$$\boldsymbol{\mathcal{S}} = \mathcal{S}^{(1)} \times \mathcal{S}^{(2)} \times \cdots \times \mathcal{S}^{(K)} = \underbrace{\mathcal{S} \times \mathcal{S} \times \cdots \times \mathcal{S}}_{K \text{ times}}$$

It's worth noting that the model above can be extended in a straightforward manner to accommodate the general case where each chain has a distinct state space.

Rather than exploring each chain on its own, the multivariate Markov chain $\boldsymbol{Z}$ can be divided into subgroups and analysed in a group-wise manner. Suppose we have a partition of the index set $[K]$ into J groups $g_1, g_2, \ldots, g_J$ with $J \leq K$ and $J \in \mathbb{N}^+$. For any subgroup $g_j \subset [K]$, define

$$\boldsymbol{Z}_j = \{\boldsymbol{Z}_j(t) : t = 1, 2, \ldots, T\} = \{Z_k(t) : k \in g_j,\ t = 1, 2, \ldots, T\}$$

as the marginal process of the multivariate chain $\boldsymbol{Z}$ corresponding to the group g_j. The joint state space $\boldsymbol{\mathcal{S}}_j$ for $\boldsymbol{Z}_j$ is given as:

$$\boldsymbol{\mathcal{S}}_j = \prod_{k \in g_j} \mathcal{S}_k$$

Specifically, if $g_j = \{k\}$ is a singleton in $[K]$, then $\boldsymbol{Z}_j$ is a univariate marginal process. In addition, let

$$\boldsymbol{Z}_{-j} = \{Z_k(t) : k \in [K] \setminus g_j,\ t = 1, 2, \ldots, T\}$$

be the marginal process excluding the group g_j.

MMCs are typically represented using joint transition matrices. Given two state vectors $\boldsymbol{a} = (a_1, \ldots, a_K)^\top$ and $\boldsymbol{b} = (b_1, \ldots, b_K)^\top$ from the state space $\boldsymbol{\mathcal{S}}$, we define the transition matrix $P = \left\{p_{\boldsymbol{ab}}\right\}$, where each element represents the probability of transitioning from state $\boldsymbol{a}$ at time $t-1$ to state $\boldsymbol{b}$ at time t:

$$p_{\boldsymbol{ab}} = \mathbb{P}(\boldsymbol{Z}(t) = \boldsymbol{b} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}).$$

It follows that P is a squared transition matrix of dimension $s^K \times s^K$, with each row summing to 1 and is invariant of time t under the time-homogeneous setting.

2.3 Dependency Structures and the Marginal Transition Probabilities

Under full dependency, each state $Z_k(t)$ in a first-order MMC depends not only on the previous states $\boldsymbol{Z}(t-1)$ but also interacts with the concurrent states $\boldsymbol{Z}_{-k}(t)$, which is induced by the dynamics going from previous time $t-1$ to current time t. Figure 2.2 illustrates a multivariate Markov chain with three chains under a full dependency structure, where each current state is connected with all previous states across chains as well as the concurrent states in the other chains:

- **Influence from previous states (red and black arrows):** These arrows depict how the previous states $\boldsymbol{Z}(t-1)$ influence the current state $Z_k(t)$. Red arrows

indicate self-dependence, such as the influence of $Z_1(t-1)$ on $Z_1(t)$, while black arrows indicate cross-chain effects, such as $Z_2(t-1)$ and $Z_3(t-1)$ influencing $Z_1(t)$.

- **Concurrent state interaction (blue arrows):** These arrows reflect dependencies between components of $\boldsymbol{Z}(t)$, conditioned on the past. For example, a blue arrow between $Z_1(t)$ and $Z_2(t)$ represents a contemporaneous relationship between chains 1 and 2.

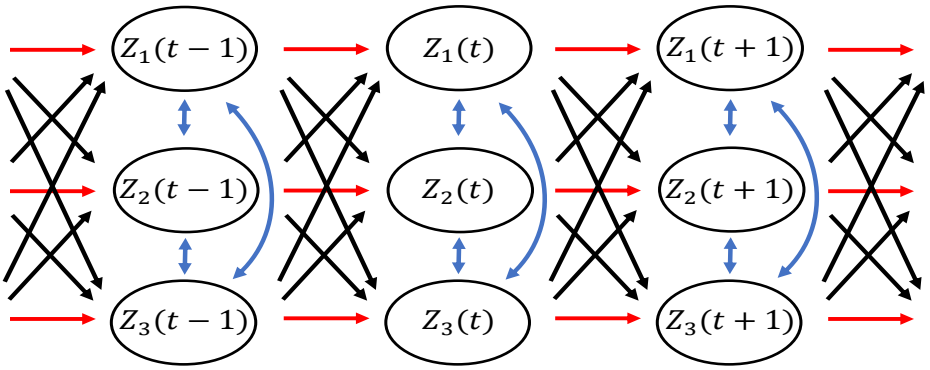


Figure 2.2: An illustration of the MMC with $K = 3$ chains under full dependency. The red arrows depict within-chain dependencies, capturing how the past state of a chain influences its own current state. The black arrows indicate cross-chain dependencies, that is how the previous states of one chain affect the current states of other chains. Lastly, the blue vertical arrows represent concurrent interactions, describing the dependencies among chains at the same time.

As illustrated in Figure 2.2, the dependency structures encoded by the arrows can be classified into three categories. *Conditional independence* corresponds to the absence of all blue arrows, indicating no concurrent interactions between chain groups. *Contemporaneous independence* arises when only a subset of the blue arrows is removed, allowing concurrent interactions within groups while excluding them across groups. Finally, *Granger non-causality* is represented by the absence of red and black arrows, indicating that past states of one group do not influence the future states of another.

Alternatively, an adjacency matrix can be constructed to represent the full dependency graph of a first-order Markov chain in Figure 2.2. Under such a case, the corresponding 6×3 adjacency matrix A that captures the full dependency structure is defined as:

$$
A = \begin{array}{c} \\ Z_1(t-1) \\ Z_2(t-1) \\ Z_3(t-1) \\ Z_1(t) \\ Z_2(t) \\ Z_3(t) \end{array}
\begin{array}{c}
\begin{array}{ccc} Z_1(t) & Z_2(t) & Z_3(t) \end{array} \\
\left(\begin{array}{ccc}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1
\end{array}\right)
\end{array}
\begin{array}{l}
\\
\\
\\
\left.\begin{array}{l}\\\\\\\end{array}\right\} \text{Effects from the previous time} \\
\left.\begin{array}{l}\\\\\\\end{array}\right\} \text{Effects from the current time}
\end{array}
\tag{2.1}
$$

In this adjacency matrix A, a 1 indicates dependency and a 0 indicates independence. Specifically, the 1 in the top-left corner of A signifies that the states $Z_1(t)$ of chain 1 at time t depend on its previous state $Z_1(t-1)$. This is depicted as the red arrow from $Z_1(t-1)$ to $Z_1(t)$ in Figure 2.2.

- The top 3 rows (highlighted in green in (2.1)) represent the effects from the previous states $Z(t-1)$. These correspond to the red (diagonal terms) and black (off-diagonal terms) arrows in the dependency graph shown in Figure 2.2. Under full dependency, this section is a matrix filled with ones.

- The bottom 3 rows (highlighted in red in (2.1)) account for the effects from the concurrent states $Z(t)$, corresponding to the blue vertical arrows in Figure 2.2. This matrix is assumed to be symmetric, capturing the symmetric dependency relationships among concurrent chains. The diagonal entries are set to 1, reflecting that each chain is dependent on itself. Under full dependency, all entries in this matrix are 1.

A complete adjacency matrix A would have dimensions of 6×6 under 3 chains to represent all possible dependencies. However, in the first-order Markov setting, we are not concerned with the left half of A, which depicts how the previous states $Z(t-1)$ depend on both the current states $Z(t)$ and the previous states themselves. The right half of

the matrix, shown in Equation (2.1), adequately captures all the dependency structures relevant to a first-order Markov chain by considering effects from both the previous and current times. Therefore, we focus only on this half of the adjacency matrix A.

Motivated by investigating such marginal relationships, such as how the state of chain 1 at time t is influenced by chain 2 at time $t-1$, represented by the black arrow from $Z_2(t-1)$ to $Z_1(t)$ in Figure 2.2, we further define the marginal transition probabilities as:

Definition 2.1. *Given a subgroup $g_j \subseteq [K]$, the marginal transition probability η_j is defined as the probability of the current state of chain group g_j conditioned on all the previous state, formally as:*

$$\eta_j = \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_{j'}(t-1)\big)$$

More precisely, given a state vector $\boldsymbol{a}_j \in \boldsymbol{\mathcal{S}}_j$for chain group g_j and $\boldsymbol{b} \in \boldsymbol{\mathcal{S}}$, we define the realization of η_j as:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) = \mathbb{P}\big(\boldsymbol{Z}_j(t) = \boldsymbol{b}_j \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big)$$

In the following Sections 2.4, 2.5, and 2.6, we will utilize these marginal transition probabilities η_j to formally characterize how each of the three dependency structures imposes constraints on the joint transition matrix.

2.4 Conditional Independence

Saul and Jordan [1999] first proposed the following conditional independence assumption 2.2 which corresponding to the blue vertical arrows within the Figure 2.2:

Assumption 2.2. *[Saul and Jordan, 1999] When conditioned on all the previous states $\boldsymbol{Z}(t-1)$, all the current states $Z_1(t), Z_2(t), \ldots, Z_K(t)$ are mutually independent.*

$$Z_1(t) \perp\!\!\!\perp Z_2(t) \perp\!\!\!\perp \cdots \perp\!\!\!\perp Z_K(t) \mid \boldsymbol{Z}(t-1) \tag{2.2}$$

Or equivalently:

$$\mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)\big) = \prod_{k=1}^{K} \mathbb{P}\big(Z_k(t) \mid \boldsymbol{Z}(t-1)\big) \tag{2.3}$$

Figure 2.3 illustrates how the structure of a fully dependent multivariate Markov chain with $K = 3$ chains is simplified under the conditional independence assumption. In this setting, each current state $Z_k(t)$ depends only on the full set of past states $\boldsymbol{Z}(t-1)$, and there are no contemporaneous (i.e., same-time) dependencies among the chains. As a result, there are no blue vertical arrows linking $Z_1(t)$, $Z_2(t)$, and $Z_3(t)$ in the diagram, in contrast to the full dependency case shown in Figure 2.2.

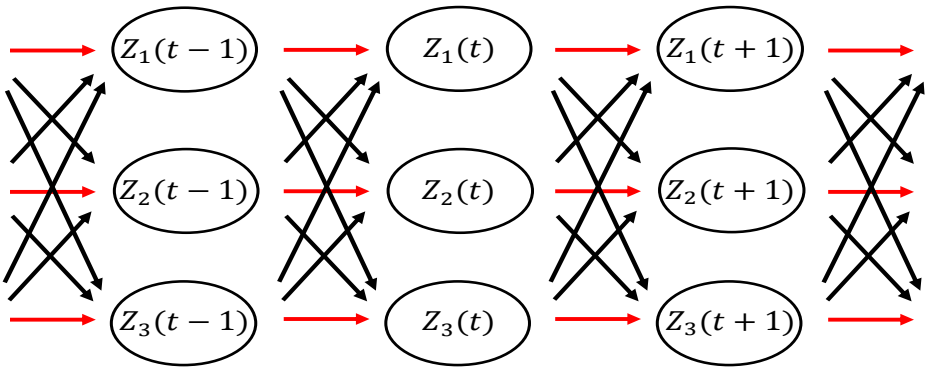


Figure 2.3: An illustration of the MMC with $K = 3$ chains under the conditional independence, given as $Z_1(t) \perp\!\!\!\perp Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$. The missing vertical blue arrows comparing to Figure 2.2 under full dependency refers to the conditional independence, where all the chains are mutually independent when conditioned on the past.

It is also worth noting that this conditional independence structure forms the basis of Coupled Hidden Markov Models (CHMMs), which have been extensively used in diverse domains such as computer vision [Brand et al., 1997], social network analysis [Raghavan et al., 2013], and epidemiology [Sherlock et al., 2013, Touloupou et al., 2020].

Building on the Definition 2.1, we now define the marginal transition probabilities η_k that appears on the right-hand side of Equation 2.3 by setting $g_j = k$ for each individual chain

$k \in [K]$, and letting $g_{j'} = [K]$, representing all chains, formally as:

$$\eta_k = \mathbb{P}\big(Z_k(t) \mid Z_1(t-1), Z_2(t-1), \ldots, Z_K(t-1)\big) \tag{2.4}$$

More precisely, given a state vector $\boldsymbol{a} = (a_1, \ldots, a_K) \in \boldsymbol{\mathcal{S}}$ and a possible state $b_k \in \mathcal{S}$ for chain k, we define the realization of η_k as:

$$\eta_k(b_k \mid \boldsymbol{a}) = \mathbb{P}\big(Z_k(t) = b_k \mid Z_1(t-1) = a_1, \ldots, Z_K(t-1) = a_K\big)$$

Under the time-homogeneous setting, η_k is invariant with time t. We can interpret $\eta_k(b_k \mid \boldsymbol{a})$ as the marginal probability of chain k being in state b_k given that the previous states are $\boldsymbol{a}$. Based on its definition, this marginal can be expressed in terms of the joint transition probabilities $p_{\boldsymbol{ab}}$ using the following marginalization formula:

$$\eta_k(b_k \mid \boldsymbol{a}) = \sum_{\boldsymbol{b'}: b'_k = b_k} p_{\boldsymbol{ab'}} \tag{2.5}$$

However, Equation (2.5) suggests that multiple sets of joint transition probabilities $\{p_{\boldsymbol{ab}}\}_{\boldsymbol{b} \in \boldsymbol{\mathcal{S}}}$ can marginalize to the same set of $\{\eta_k(b_k|\boldsymbol{a})\}_{b_k \in \mathcal{S}}$, indicating a non-identifiability issue. Incorporating conditional independence assumption 2.2 serves as a remedy to this issue. Since under conditional independence, we can rewrite transition probabilities (see Equation (2.3)) in terms of the marginal transition probabilities η as defined in Equation (2.4).

$$p_{\boldsymbol{ab}} = \prod_{k=1}^{K} \eta_k(b_k \mid \boldsymbol{a}) \tag{2.6}$$

for state vectors $\boldsymbol{a} = (a_1, \ldots, a_K), \boldsymbol{b} = (b_1, \ldots, b_K) \in \boldsymbol{\mathcal{S}}$

This implies that instead of the marginalization relationship between the marginal transition probabilities η and the entries p in transition matrix P, the conditional independence assumption establishes another relationship and connects the η with p through multiplication. This connection will later be shown to form a bijection in Proposition 2.6.

2.4.1 The Properties of Conditional Independence Assumption

At the same time of simplifying the conditional dependencies among different chains, the conditional independence assumption also imposes the row constraints in P (Proposition 2.3) and reduces the number of parameters needed to parametrise it (Proposition 2.5).

Proposition 2.3. *Consider an arbitrary row $\boldsymbol{a}$ of P. For a randomly picked $I \in \mathbb{N}^+$ and state vectors $\boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(I)} \in \boldsymbol{\mathcal{S}}$. For each chain $k \in [K]$ and state $c \in \mathcal{S}$, define the number of times chain k is in state c across these I vectors by*

$$\sum_{i=1}^{I} \mathbb{I}\left\{b_k^{(i)} = c\right\} = n_{k,c}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function.

*Now, take two sets of the state vectors $\forall \boldsymbol{b}^{(1)}, \ldots, \boldsymbol{b}^{(I)}$ and $\boldsymbol{b}'^{(1)}, \ldots, \boldsymbol{b}'^{(I)} \in \boldsymbol{\mathcal{S}}$ with their corresponding counts $n = \{n_{k,c}\}$ and $n' = \{n'_{k,c}\}$. The **Conditional Independence Assumption** introduces the row constraints such that if:*

$$n_{k,c} = n'_{k,c} \qquad \forall k \in [K], c \in \mathcal{S}$$

Then,

$$\prod_{i=1}^{I} p_{\boldsymbol{a}\boldsymbol{b}^{(i)}} = \prod_{i=1}^{I} p_{\boldsymbol{a}\boldsymbol{b}'^{(i)}} \tag{2.7}$$

Proof. By expanding $p_{\boldsymbol{a}\boldsymbol{b}^{(j)}}$ according to Assumption 2.2, we yield:

$$\begin{aligned}
\prod_{i=1}^{I} p_{\boldsymbol{a}\boldsymbol{b}^{(i)}} &= \prod_{i=1}^{I}\prod_{k=1}^{K} \eta_k(b_k^{(i)} \mid \boldsymbol{a}) \\
&= \prod_{c=1}^{s}\prod_{k=1}^{K} \left(\eta_k(c \mid \boldsymbol{a})\right)^{n_{k,c}} && \text{(By aggregating all the } \eta\text{)} \\
&= \prod_{i=1}^{I}\prod_{k=1}^{K} \eta_k(b_k'^{(i)} \mid \boldsymbol{a}) = \prod_{i=1}^{I} p_{\boldsymbol{a}\boldsymbol{b}'^{(i)}}
\end{aligned}$$

□

Example 2.4. Consider a MMC with $K = 3$ chains sharing a common state space $\mathcal{S} = \{1, 2\}$. Take $I = 2$ and define the target states as $\boldsymbol{b}^{(1)} = (1, 1, 2)$, $\boldsymbol{b}^{(2)} = (2, 2, 2)$, and an alternative pair $\boldsymbol{b}'^{(1)} = (1, 2, 2)$, $\boldsymbol{b}'^{(2)} = (2, 1, 2)$. Under the conditional independence assumption, we have:

$$p_{\boldsymbol{a}\boldsymbol{b}^{(1)}} p_{\boldsymbol{a}\boldsymbol{b}^{(2)}} = p_{\boldsymbol{a}\boldsymbol{b}'^{(1)}} p_{\boldsymbol{a}\boldsymbol{b}'^{(2)}} \qquad \forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$$

since

$$n_{1,2} = n'_{1,2} = 1 \qquad n_{2,2} = n'_{2,2} = 1 \qquad n_{3,2} = n'_{3,2} = 2$$

Introducing additional constraints inevitably leads to a reduction in the number of independent parameters, as demonstrated in the following Proposition 2.5.

Proposition 2.5. *The imposing of Assumption 2.2 reduces the number of independent parameters in each row of the transition matrix P from $s^K - 1$ to $(s-1)K$ for a K-variate Markov chain with each chain having s possible states.*

Proof. Fix the previous state to be $\boldsymbol{Z}(t-1) = \boldsymbol{a}$. Under conditional independence, the probability of transitioning to $\boldsymbol{Z}(t) = \boldsymbol{b}$ for $\forall \boldsymbol{b} \in \boldsymbol{\mathcal{S}}$ can be expressed as:

$$\mathbb{P}(\boldsymbol{Z}(t) = \boldsymbol{b} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}) = \prod_{k=1}^{K} \mathbb{P}(Z_k(t) = b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}) = \prod_{i=1}^{I} \eta_k(b_k \mid \boldsymbol{a}) \tag{2.8}$$

For each element, $\eta_k(b_k \mid \boldsymbol{a})$, in the product on the right-hand side, we have the constraint:

$$\sum_{b_k=1}^{s} \eta_k(b_k \mid \boldsymbol{a}) = \sum_{b_k=1}^{s} \mathbb{P}(Z_k(t) = b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}) = 1 \tag{2.9}$$

This constraint arises from the fact that the future state of chain k must lie within the state space $\mathcal{S}$. Hence, each $\eta_k(\cdot \mid \boldsymbol{a})$ has $s - 1$ independent parameters. Consequently, to recover all the elements in that row, which is the left-hand side of Equation (2.8), we require $(s-1)K$ parameters for K multiplicand representing K chains. □

At the same time, the conditional independence assumption constructs a bijection between the marginal transition probabilities η and the entries p of the joint transition matrix P.

Proposition 2.6. *If we define the process of multiplying marginal probability η to obtain the joint transition probabilities p, depicted in Equation (2.6) as the map $\mathcal{H} \to \hat{\mathcal{P}}$, where $\hat{\mathcal{P}}$ represents the restricted parameter space of $P = \left\{p_{\boldsymbol{ab}}\right\}_{\boldsymbol{a},\boldsymbol{b}\in\boldsymbol{\mathcal{S}}}$ satisfying conditional independence, which is formally defined as:*

$$\hat{\mathcal{P}} = \left\{P = \{p_{\boldsymbol{ab}}\} : \quad \textit{satisfies Equation 2.7}\right\},$$

and $\mathcal{H}$ is the full parameter space for $H = \left\{\eta_k(b_k \mid \boldsymbol{a})\right\}_{k\in[K],b_k\in\mathcal{S},\boldsymbol{a}\in\boldsymbol{\mathcal{S}}}$. Then this map $\mathcal{H} \to \hat{\mathcal{P}}$ is a bijection.

Proof. Consider the map $\mathcal{H} \to \hat{\mathcal{P}}$

- **Surjective**: If we pick an arbitrary $P \in \hat{\mathcal{P}}$, due to Proposition 2.3, we have:

$$\prod_{i=1}^{I} p_{\boldsymbol{ab}^{(i)}} = \prod_{c=1}^{s}\prod_{k=1}^{K} \left(\eta_k(c \mid \boldsymbol{a})\right)^{n_{k,c}}$$

  We can accordingly pick a large I such that $n_{k,c} > 0$ for $\forall k \in [K], c \in \mathcal{S}$. Then, the corresponding $H = \left\{\eta_k(b_k \mid \boldsymbol{a})\right\}$ on the right-hand-side will map to the given $P = \{p_{\boldsymbol{ab}}\}$.

- **Injective**: Suppose we pick $\hat{H} = \left\{\hat{\eta}_k(b_k \mid \boldsymbol{a})\right\}, \tilde{H} = \left\{\tilde{\eta}_k(b_k \mid \boldsymbol{a})\right\} \in \mathcal{H}$ with

$$\mathcal{F}_1(\hat{H}) = \hat{P} \qquad \text{and} \qquad \mathcal{F}_1(\tilde{H}) = \tilde{P}$$

  Equivalently, $\forall \boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{S}}$, we have :

$$\prod_{i=1}^{I} \hat{\eta}_k(b_k \mid \boldsymbol{a}) = \hat{p}_{\boldsymbol{ab}} \qquad \text{and} \qquad \prod_{i=1}^{I} \tilde{\eta}_k(b_k \mid \boldsymbol{a}) = \tilde{p}_{\boldsymbol{ab}}$$

  Note that these are the sets of equations for different $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{S}}$. We can solve for $\hat{\eta}, \tilde{\eta}$ in terms of $\hat{p}_{\boldsymbol{ab}}$ and $\tilde{p}_{\boldsymbol{ab}}$ respectively. If we sum over all the $\hat{p}_{\boldsymbol{ab}}$ with $b_k = c \in \mathcal{S}$, we obtain:

$$\sum_{\boldsymbol{b}:b_k=c} \hat{p}_{\boldsymbol{ab}} = \sum_{\boldsymbol{b}:b_k=c} \left(\prod_{k'=1}^{K} \hat{\eta}_{k'}(b_{k'} \mid \boldsymbol{a}) \right) \tag{2.10}$$

$$= \hat{\eta}_k(c \mid \boldsymbol{a}) \left[\sum_{b_{k'}=1}^{s} \left(\prod_{k'=1,k'\neq k}^{K} \hat{\eta}_{k'}(b_{k'} \mid \boldsymbol{a}) \right) \right] \tag{2.11}$$

$$= \hat{\eta}_k(c \mid \boldsymbol{a}) \left[\prod_{k'=1,k'\neq k}^{K} \left(\sum_{b_{k'}=1}^{s} \hat{\eta}_{k'}(b_{k'} \mid \boldsymbol{a}) \right) \right] \tag{2.12}$$

$$= \hat{\eta}_k(c \mid \boldsymbol{a}) \tag{2.13}$$

Since all the elements in the summation on the right-hand side of Equation (2.10) involve $\hat{\eta}_k(c \mid \boldsymbol{a})$, we can factor it out and obtain (2.11). We can then interchange the order of summation and product, as we are summing over all possible combinations of $\prod_{k'=1,k'\neq k}^{K} \hat{\eta}_{k'}(b_{k'} \mid \boldsymbol{a})$, resulting in (2.12). Finally, in accordance with Equation (2.9), all terms in the product are 1, yielding $\hat{\eta}_k(c \mid \boldsymbol{a})$.

These steps establish the marginalization relationship between η and p which is identical to Equation (2.5). Repeating steps (2.10) to (2.13) for $\tilde{\eta}_k(b_k \mid \boldsymbol{a})$ gives:

$$\hat{\eta}_k(c \mid \boldsymbol{a}) = \sum_{\boldsymbol{b}:b_k=c} \hat{p}_{\boldsymbol{ab}} \qquad \text{and} \qquad \tilde{\eta}_k(c \mid \boldsymbol{a}) = \sum_{\boldsymbol{b}:b_k=c} \tilde{p}_{\boldsymbol{ab}}$$

Thus, $\hat{p}_{\boldsymbol{ab}} = \tilde{p}_{\boldsymbol{ab}}$ for all $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{S}}$ implies $\hat{\eta}_k(c \mid \boldsymbol{a}) = \tilde{\eta}_k(c \mid \boldsymbol{a})$, establishing injectivity.

□

2.4.2 The Geometric Interpretation of Conditional Independence

The conditional independence can be interpreted geometrically. Consider a $s^K \times s^K$ joint transition matrix P for an MMC with K chains and a shared state space $\mathcal{S} = \{1, 2, \ldots, s\}$, then each row of P, denoted as $P_{\boldsymbol{a}\cdot}$, lies within an $s^K - 1$ dimensional simplex, $\boldsymbol{\Delta}^{s^K-1}$ subject to the rows sum up to 1. The entire matrix P can be seen as a product of these $s^K - 1$ dimensional simplices, as each row of P is independent of each other.

Consider the equivalence relation $\sim$ on $\mathbf{\Delta}^{s^K-1}$ defined as $P_{\boldsymbol{a}\cdot} \sim P'_{\boldsymbol{a}\cdot}$ if they marginalize to the same $H = \left\{\eta_k(b_k \mid \boldsymbol{a})\right\}_{k\in[K],b_i\in\mathcal{S},\boldsymbol{a}\in\boldsymbol{\mathcal{S}}}$ through Equation (2.5). We denote the equivalence class as:

$$[P_{\boldsymbol{a}\cdot}] = \left\{P'_{\boldsymbol{a}\cdot} \in \mathbf{\Delta}^{s^K-1} : P'_{\boldsymbol{a}\cdot} \sim P_{\boldsymbol{a}\cdot}\right\}$$

The corresponding quotient space $\mathbf{\Delta}^{s^K-1}/\sim$, which identifies joint transition matrices P that induce the same marginal transition matrix H, is the set

$$\{[P_{\boldsymbol{a}\cdot}] : P_{\boldsymbol{a}\cdot} \in \mathbf{\Delta}^{s^K-1}\}$$

of all the equivalence classes.

Proposition 2.7. *The section $S \subset \mathbf{\Delta}^{s^K-1}$ identified by conditional independence assumption 2.2 is a* ***cross section*** *such that*

$$S \cap [P_{\boldsymbol{a}\cdot}] = \hat{P}_{a\cdot} \qquad \forall\, [P_{\boldsymbol{a}\cdot}] \in \mathbf{\Delta}^{I^S-1}/\sim$$

This result follows directly from the fact that the map $\mathcal{H} \to \hat{\mathcal{P}}$ is a bijection, as established in Proposition 2.6. Consequently, imposing conditional independence restricts the parameter space $\mathcal{P}$ to $\hat{\mathcal{P}}$ so that the identifiability are preserved, since each element of the reduced transition space $\hat{\mathcal{P}}$ corresponds uniquely to an element of $\mathcal{H}$.

2.4.3 Visualize the Map from Transition Probabilities to Marginal Transition Probabilities under Conditional Independence

In this section, we aim to visualize the transformation from the joint transition probabilities p to the marginal transition probabilities η under the conditional independence assumption. To make the visualization tractable, we focus on a simplified setting with $K = s = 2$. That is, we consider a bivariate Markov chain where each chain takes values in the state space $\mathcal{S} = \{1, 2\}$.

2.4.3.1 The Parameter Spaces

We start with presenting the parameter space of p and η. Under the two chains two states setting, we will have a 4×4 transition matrix P as the $|\boldsymbol{\mathcal{S}}| = |\{1,2\} \times \{1,2\}| = 4$. Explicitly, we write it as:

$$P = \begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,2) & (2,1) & (2,2) \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} \end{array}$$

Without loss of generality, we assign 1 in subscription of p refers to the bivariate chain being in state $(1,1)$, 2 for $(1,2)$, 3 for $(2,1)$ and 4 for $(2,2)$. For a proper transition matrix, each row should sum up to 1, that is:

$$p_{a1} + p_{a2} + p_{a3} + p_{a4} = 1 \qquad \text{for} \quad a = 1,2,3,4$$

If we define $P_{a\cdot} = (p_{a1}, p_{a2}, p_{a3}, p_{a4})$ as the a-th row of the transition matrix P, then the row constraint implies that $P_{a\cdot}$ lies within a standard simplex in 3-D, say $\boldsymbol{\Delta}^3$, under the constraints:

$$0 < p_{a1} < 1, \qquad 0 < p_{a2} < 1, \qquad 0 < p_{a3} < 1$$
$$0 < p_{a4} = 1 - p_{a1} - p_{a2} - p_{a3} < 1$$

As we do not impose any column structure on the P, four rows of P are independent. Consequently, P lies in the space $\mathcal{P} = \boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^3$. By choosing (p_{a1}, p_{a2}, p_{a3}) to be the coordinates, we can visualize $P_{a\cdot}$, the a-th row of P, as the gray simplex $\boldsymbol{\Delta}^3_{ABCD}$ formed by $A = (0,1,0)$, $B = (0,0,0)$, $C = (1,0,0)$ and $D = (0,0,1)$ in Figure 2.4a.

Then, we consider the parameter space for marginal transition matrix H. Under this special case, we have a 4×4 matrix H, which can be written with the abbreviated notation:

$$H = \begin{array}{c} \\ (1,1) \\ (1,2) \\ (2,1) \\ (2,2) \end{array} \begin{array}{c} \begin{array}{cccc} (1,\cdot) & (2,\cdot) & (\cdot,1) & (\cdot,2) \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{pmatrix} \end{array}$$

We retain the same row indexing as in the transition matrix P, which corresponds to the previous states. The columns, however, are restructured: columns 1 and 2 represent the marginal transition probabilities of chain 1 being in states 1 and 2, respectively; columns 3 and 4 correspond to the marginal transition probabilities of chain 2 being in states 1 and 2, respectively. For instance, $\eta_{22} \equiv \eta_1\big(2 \mid (1,2)\big)$.

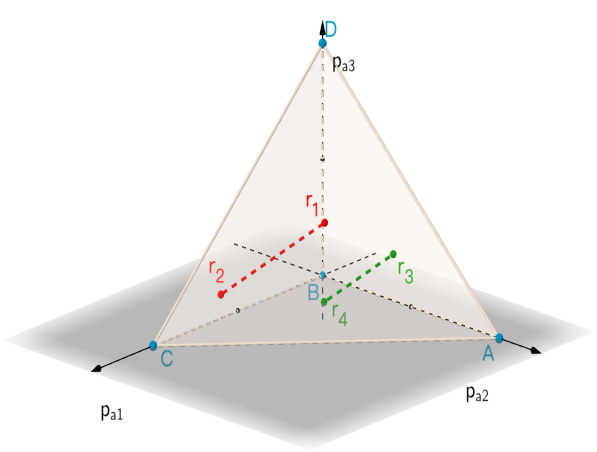
According to the row constraints given for marginal transition probability η in Equation (2.9), we know that

$$\eta_{a1} + \eta_{a2} = \eta_{a3} + \eta_{a4} = 1 \qquad \text{for} \quad a = 1,2,3,4$$

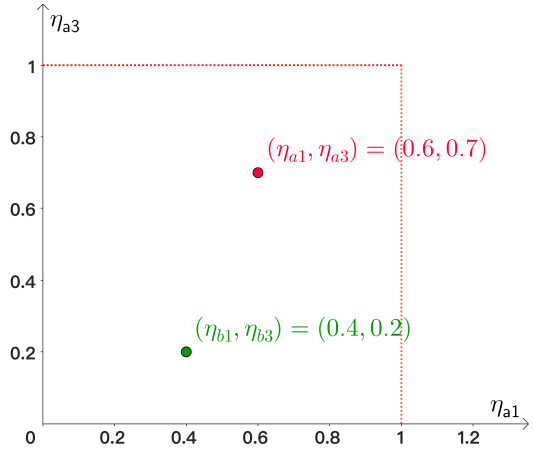
Let $H_{a\cdot} = (\eta_{a1}, \eta_{a2}, \eta_{a3}, \eta_{a4})$ denote the a-th row of the marginal transition matrix H. The row constraint implies that $H_{a\cdot}$ lies within the square $[0,1]^2$. As with P, we do not impose any column structures on H, which means that four rows of H are mutually independent. Consequently, H lies in the hypercube $\mathcal{H} = [0,1]^8$. We try to visualize the parameter space of $H_{a\cdot}$, the a-th row of H, with a choice of coordinates to be η_{a1} and η_{a3}, as the square outlined by the x-axis, the y-axis, and two red dashed lines in Figure 2.4b.

2.4.3.2 The Map between p and η

Having identified the parameter spaces for $P_{a\cdot}$ and $H_{a\cdot}$, we now turn to the map between them. The transformation from $P_{a\cdot}$ to $H_{a\cdot}$ is direct via marginalization. We define this



(a) The parameter space of $P_{a\cdot}$ when we choose (p_{a1}, p_{a2}, p_{a3}) as the coordinates, given as a standard simplex $\mathbf{\Delta}^3$ in 3D



(b) The parameter space of $H_{a\cdot}$ when we choose (η_{a1}, η_{a3}), given as a unit square given as $[0,1]^2$

Figure 2.4: An illustration of the transformation from $P_{a\cdot}$ to $H_{a\cdot}$. The red and green doted lines in the left Figure 2.4a represent two different equivalence classes $[P_{a\cdot}]$ and $[P_{b\cdot}]$ constituted by the points $P_{a\cdot}$ and $P_{b\cdot}$ on that line. All these points within their equivalence classes marginalize to the same point (η_{a1}, η_{a3}) in red and (η_{b1}, η_{b3}) in green in the parameter space of $H_{a\cdot}$ respectively, as shown on the right Figure 2.4b.

map as $\mathcal{F}_2 : \mathcal{P}_{a\cdot} \to \mathcal{H}_{a\cdot}$ given by:

$$\mathcal{F}_2(P_{a\cdot}) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_{a1} \\ p_{a2} \\ p_{a3} \\ p_{a4} \end{pmatrix} = \begin{pmatrix} \eta_{a1} \\ \eta_{a2} \\ \eta_{a3} \\ \eta_{a4} \end{pmatrix} = H_{a\cdot}$$

where $\mathcal{P}_{a\cdot} = \mathbf{\Delta}^3$ and $\mathcal{H}_{a\cdot} = [0,1]^2$ are the parameter spaces for $P_{a\cdot}$ and $H_{a\cdot}$ respectively. Alternatively, we may write it in equation form as:

$$p_{a1} + p_{a2} = \eta_{a1} \qquad \text{and} \qquad p_{a1} + p_{a3} = \eta_{a3} \tag{2.14}$$

Clearly, this map is surjective but not injective:

- There can be infinitely many $P_{a\cdot}$ map to the same $H_{a\cdot}$. This can also be verified by the plots. A randomly chosen $H_{a\cdot} = (\alpha, 1-\alpha, \beta, 1-\beta)$ with $\eta_{a1} = \alpha$ and $\eta_{a3} = \beta$

$(\alpha, \beta \in [0,1])$ identifies two planes within the parameter space $\mathbf{\Delta}^3$ as:

$$p_{a1} + p_{a2} = \alpha \qquad \text{and} \qquad p_{a1} + p_{a3} = \beta$$

These two planes are not parallel and intersect at the line ℓ:

$$\begin{pmatrix} p_{a1} \\ p_{a2} \\ p_{a3} \end{pmatrix} = \begin{pmatrix} \alpha \\ 0 \\ \beta - \alpha \end{pmatrix} + t \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \tag{2.15}$$

This implies that all the $P_{a\cdot} \in \ell$ locates on this line will map to the same $H_{a\cdot} = (\alpha, 1-\alpha, \beta, 1-\beta)$

- The surjectivity refers to check if ℓ cuts $\mathbf{\Delta}^3$ for $\forall \alpha, \beta \in [0,1]$:

  1. For $\beta - \alpha \geq 0$, let $t = 0$, the point $(\alpha, 0, \beta - \alpha)$ is where ℓ cuts $\mathbf{\Delta}^3$
  2. For $\alpha - \beta < 0$, let $t = \beta - \alpha$, the point $(\beta, \alpha - \beta, 0)$ is where ℓ cuts $\mathbf{\Delta}^3$.

  This implies that the surjectivity of $\mathcal{F}_2$.

Figure 2.4 illustrates the map $\mathcal{F}_2$: all the $P_{a\cdot}$ locates on the red doted line $\ell_a : \overline{\mathrm{r_1 r_2}}$ maps to the same $H_{a\cdot} = (\eta_{a1}, \eta_{a2}, \eta_{a3}, \eta_{a4})$ while all the $P_{b\cdot}$ locates on the green doted line $\ell_b : \overline{\mathrm{r_3 r_4}}$ maps to the same $H_{b\cdot} = (\eta_{b1}, \eta_{b2}, \eta_{b3}, \eta_{b4})$.

2.4.3.3 Conditional Independence and the Cross-Section Identified

Consider the equivalence relation $\sim$ on $\mathbf{\Delta}^3$ defined as $P_{a\cdot} \sim P_{b\cdot}$ if they map to the same $H_{a\cdot}$ through $\mathcal{F}_2$. We denote the equivalence class as:

$$[P_{a\cdot}] = \{P_{b\cdot} \in \mathbf{\Delta}^3 : P_{b\cdot} \sim P_{a\cdot}\}$$

The corresponding quotient space $\mathbf{\Delta}^3 / \sim$ is the set

$$\{[P_{a\cdot}] : P_{a\cdot} \in \mathbf{\Delta}^3\}$$

of all the equivalence classes.

We verify that all the equivalence classes $[P_{a\cdot}]$ are the straight line ℓ given in Equation (2.15). We further notice that all these lines ℓ identified by different $H_{a\cdot}$ are parallel to each other as the direction vector $(1, -1, -1)$ is constant and independent of $H_{a\cdot}$, which implies:

$$[P_{a\cdot}] \cap [P_{b\cdot}] = \emptyset \qquad \text{for} \quad a \neq b$$

Geometrically, the quotient space $\mathbf{\Delta}^3/\sim$ is hence a union of the straight parallel line ℓ forming $\mathbf{\Delta}$.

Conditional independence assumption 2.2 says that:

$$p_{a1} = \eta_{a1}\eta_{a3} \qquad p_{a2} = \eta_{a1}\eta_{a4}$$
$$p_{a3} = \eta_{a2}\eta_{a3} \qquad p_{a4} = \eta_{a2}\eta_{a4}$$

Together with Equation (2.14), we are able to cancel η and yield a manifold $\mathcal{M} \subset \mathbf{\Delta}^3$ given by:

$$\mathcal{M}: \quad p_{a3} = \frac{p_{a1}}{p_{a1} + p_{a2}}(1 - p_{a1} - p_{a2}) \tag{2.16}$$

Proposition 2.8. *The manifold $\mathcal{M} \subset \mathbf{\Delta}^3$ identified by Assumption 2.2 is a cross section such that the map $\mathcal{M} \to \mathbf{\Delta}^3/\sim$ is bijective. Or alternatively,*

$$\mathcal{M} \cap [P_{a\cdot}] = \hat{P}_{a\cdot} \qquad \forall\, [P_{a\cdot}] \in \mathbf{\Delta}^3/\sim$$

Proof. We arbitrarily pick two different points $\hat{P}_{a\cdot}, \hat{P}_{b\cdot} \in \mathbf{\Delta}^3$ given by

$$\hat{P}_{a\cdot} = (\alpha\beta, \alpha(1-\beta), (1-\alpha)\beta, (1-\alpha)(1-\beta))$$
$$\hat{P}_{b\cdot} = (\alpha'\beta', \alpha'(1-\beta'), (1-\alpha')\beta', (1-\alpha')(1-\beta'))$$
$$\text{with} \qquad \alpha \neq \alpha' \quad \text{and} \quad \beta \neq \beta'$$

They belong to two lines ℓ_a and ℓ_b given by:

$$\ell_a : \begin{pmatrix} p_{a1} \\ p_{a2} \\ p_{a3} \end{pmatrix} = \begin{pmatrix} \alpha\beta \\ \alpha(1-\beta) \\ (1-\alpha)\beta \end{pmatrix} + t \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \qquad \ell_b : \begin{pmatrix} p_{a1} \\ p_{a2} \\ p_{a3} \end{pmatrix} = \begin{pmatrix} \alpha'\beta' \\ \alpha'(1-\beta') \\ (1-\alpha')\beta' \end{pmatrix} + t \begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix}$$

Equating two lines gives us $\alpha = \alpha'$ and $\beta = \beta'$, which implies that different $\hat{P}_{a\cdot}$ and $\hat{P}_{b\cdot}$ cannot lies on the same line. Alternatively, two arbitrarily different points on the section $\mathcal{M}$ must belong to two different equivalence classes $[P_{a\cdot}]$. Consequently, $\mathcal{M}$ is a cross-section. □

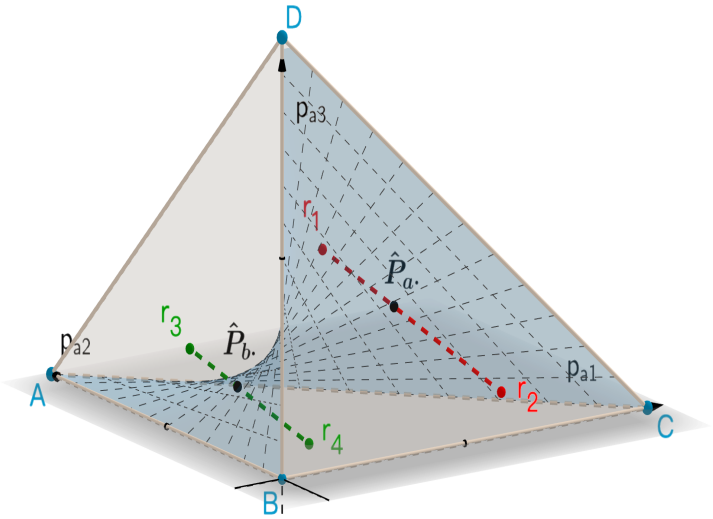


Figure 2.5: An illustration of the cross-section $\mathcal{M}$ defined by the conditional independence, which is the manifold in light blue grid. Two different equivalence classes, $[P_{a\cdot}]$ and $[P_{b\cdot}]$ in red and green dotted line, are both cut by the cross-section $\mathcal{M}$ by a unique point, $\hat{P}_{a\cdot}$ and $\hat{P}_{b\cdot}$ respectively. The intersected points within $\mathcal{M}$ form a bijection with the $H_{a\cdot}$ in $[0,1]^2$

Figure 2.5 gives an illustration of this. We label the cross-section $\mathcal{M}$ defined by Assumption 2.2 in light blue grid. The red and green dotted lines refer to two different equivalence classes $[P_{a\cdot}]$ and $[P_{b\cdot}]$. The intersection between each of the equivalence classes and the cross section $\mathcal{M}$ gives a unique point, $\hat{P}_{a\cdot}$ and $\hat{P}_{b\cdot}$ respectively as the definition of the cross

section.

2.4.3.4 Further Discussion on the Necessary Condition of the Cross Section

One necessary condition for the surface S to be a cross section $\mathcal{M}$ in the bivariate case is that it must be bounded by following four segments:

$$\begin{aligned} \ell_{AB} &: p_{a1} = p_{a3} = 0 \qquad \forall 0 \leq p_{a2} \leq 1 \\ \ell_{BD} &: p_{a1} = p_{a2} = 0 \qquad \forall 0 \leq p_{a3} \leq 1 \\ \ell_{CD} &: p_{a1} + p_{a3} = 1 \qquad \text{for} \quad p_{a2} = 0 \\ \ell_{AC} &: p_{a1} + p_{a2} = 1 \qquad \text{for} \quad p_{a3} = 0 \end{aligned}$$

Where ℓ_{AB} stands for the segment starting from point A and ending with point B.

These four segments are identified by assigning the pair (η_{a1}, η_{a3}) the extreme values located at the edges of the unit square, outlined by the red dotted line in Figure 2.4b. If we let $(\eta_{a1}, \eta_{a3}) = (\alpha, \beta)$ and line ℓ to be the one in Equation (2.15) representing the equivalence class $[P_{a\cdot}]$, then:

- When fixing $\alpha = 0$, ℓ intersects $\mathbf{\Delta}^3$ only at point $(0, 0, \beta)$. Varying β gives ℓ_{BD}
- When fixing $\beta = 0$, ℓ intersects $\mathbf{\Delta}^3$ only at point $(0, \alpha, 0)$. Varying α gives ℓ_{AB}
- When fixing $\alpha = 1$, ℓ intersects $\mathbf{\Delta}^3$ only at point $(\beta, 1 - \beta, 0)$. Varying α gives ℓ_{AC}
- When fixing $\beta = 1$, ℓ intersects $\mathbf{\Delta}^3$ only at point $(\alpha, 0, 1 - \alpha)$. Varying α gives ℓ_{CD}

Since a cross section $\mathcal{M}$ must have exactly one intersection with each equivalence class $[P_{a\cdot}]$. For the extreme cases where $\alpha = 0, 1$ or $\beta = 0, 1$, there will only be one element in each equivalence class. Therefore, the cross section S must include those points and, consequently, the segments formed by those points. It is evident both graphically and algebraically that the cross section S identified by Assumption 2.2 includes all the four segments.

Conditional independence assumption provides an approach to identify a cross section $\mathcal{M} \subset \boldsymbol{\Delta}^3$ and constructs a bijective map between all the points $\hat{P}_{a\cdot}$ on $\mathcal{M}$ and $H_{a\cdot} \in [0,1]$, which makes the problem identifiable under marginal transition probability η parametrisation. It is also worth noting that other model assumptions could also apply as long as it identifies a cross section $\mathcal{M}' \subset \boldsymbol{\Delta}^3$ and the conditional independence assumption is just one of the possible model choices.

2.4.4 The General Cases

Consider the general case of a K-variate MMC with a shared state space $\mathcal{S}$. As discussed in Section 2.4.2, without any restrictions, each row $P_{\boldsymbol{a}\cdot}$ independently occupies the parameter space $\boldsymbol{\Delta}^{S^K-1}$. When introducing the conditional independence assumption, as demonstrated in Proposition 2.6, there exists a bijection between the row $P_{\boldsymbol{a}\cdot}$ and the corresponding $H_{\boldsymbol{a}\cdot}$. Therefore, under the conditional independence assumption, $P_{\boldsymbol{a}\cdot}$ are restricted into the same subspace as $H_{\boldsymbol{a}\cdot}$, which is given by:

$$\underbrace{\boldsymbol{\Delta}^{S-1} \times \boldsymbol{\Delta}^{S-1} \times \cdots \times \boldsymbol{\Delta}^{S-1}}_{K \text{ times}} \tag{2.17}$$

Each $\boldsymbol{\Delta}^{S-1}$ corresponds to the parameter space of $\eta_k(\cdot \mid \boldsymbol{a})$ for $k \in [K]$, resulting in the product space of K copies of $\boldsymbol{\Delta}^{S-1}$. Note that in the previous sample with $S = K = 2$ in Section 2.4.2, $P_{\boldsymbol{a}\cdot}$ locates in the subspace $\boldsymbol{\Delta}^1 \times \boldsymbol{\Delta}^1$, which is a point in $\boldsymbol{\Delta}^3$ corresponding to $\hat{P}_{\boldsymbol{a}\cdot}$ in Figure 2.5.

2.5 Contemporaneous Independence

2.5.1 Properties

The conditional independence assumption can sometimes be overly restrictive, as it ignores all inter-chain effects and only considers the influence from previous states. In a more realistic scenario, only a portion of the inter-chain effects may be negligible, while the others may significantly interact with each other. Therefore, more flexible assumptions

are required. Richardson [2003] introduced first the concept of separation in the context of the Markov property for mixed graphs. For any three disjoint sets of vertices $A, B, C \subset V$, the separation condition is expressed as:

$$A \perp\!\!\!\perp B \mid V \setminus (A \cup B \cup C)$$

which means that the vertices in A are separated from those in B by the vertices in C. Colombi and Giordano [2012] later adapted this framework to the context of MMC and introduced the contemporaneous independence assumption, defined as follows:

Assumption 2.9. *[Colombi and Giordano, 2012] Given two disjoint groups of chains $g_j, g_{j'} \subset [K]$, the marginal processes $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_{j'}$ of $\boldsymbol{Z}$ are said to be* ***contemporaneously independent*** *with respect to $\boldsymbol{Z}$ if and only if for every discrete time $t = 2, \ldots, T$*

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}(t) \quad | \quad \boldsymbol{Z}\big(1 : (t-1)\big) \tag{2.18}$$

This assumption states that the two marginal processes, $\boldsymbol{Z}_j(t)$ and $\boldsymbol{Z}_{j'}(t)$, are independent at all time points, given the past information. In the context of the first-order MMC, Equation (2.18) can be further simplified by disregarding the history prior to time t.

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{2.19}$$

It is straightforward to see that the conditional independence assumption 2.2 is a special case of the contemporaneous independence assumption, by choosing:

$$g_j = \{k\}, \quad g_{j'} = [K] \setminus g_j \quad \text{for each } k \in [K]$$

where g_j is a singleton that iterates over all chains, and $g_{j'}$ denotes its complement. This construction implies that all individual chains are mutually independent at time t, given the previous time step $t-1$, which is by definition the conditional independence assumption in 2.2.

Figure 2.6 gives an illustration of the contemporaneous independence. Consider the case where $K = 3$, $g_1 = \{1, 2\}$ and $g_2 = \{3\}$ form a partition of $[K]$.

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{2.20}$$

The contemporaneous independence states that chain 1 and chain 2, within group g_1, are correlated with each other, which is represented by the blue bidirectional arrow between $Z_1(t)$ and $Z_2(t)$. However, when conditioned on the previous states $\boldsymbol{Z}(t-1)$, the joint of $Z_1(t)$ and $Z_2(t)$ are independent of $Z_3(t)$ in group g_2, as indicated by the absence of the blue vertical arrows between them.

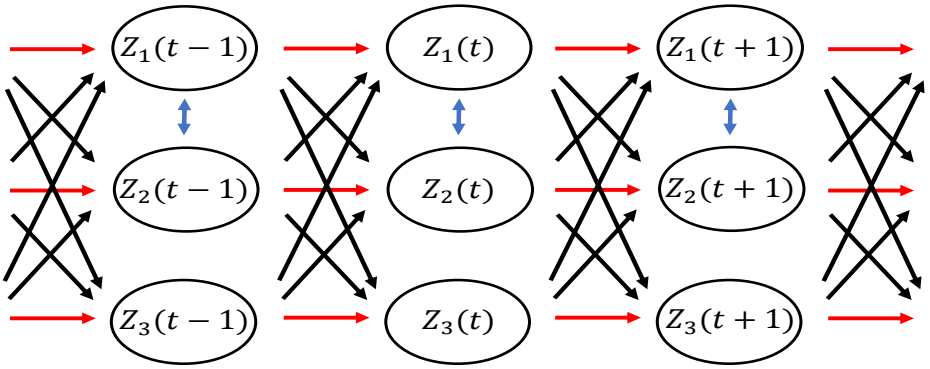


Figure 2.6: An illustration of the MMC with $K = 3$ chains under the contemporaneous independence, by picking $g_1 = \{1, 2\}$ and $g_2 = \{3\}$, given as $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$. We denote it as **Case I** since chain groups g_1 and g_2 form a partition of all the chains. The missing blue vertical arrows between $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2\}$ and $\boldsymbol{Z}_3$ comparing to Figure 2.2 identify the contemporaneous independence between chain group g_1 and g_2.

Figure 2.7 presents a more general case for $K = 3$, $g_1 = \{1\}$ and $g_2 = \{3\}$ do not form a partition of the chains $[K]$.

$$Z_1(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{2.21}$$

In this scenario, the absence of the blue vertical arrow between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_3$ indicates that these two marginal processes are independent, although both may still interact with $\boldsymbol{Z}_2$.

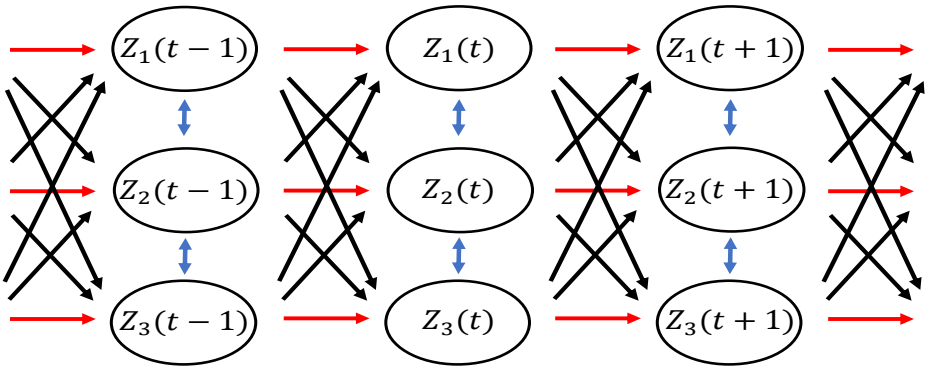


Figure 2.7: An illustration of the MMC with $K = 3$ chains under the contemporaneous independence, by picking $g_1 = \{1\}$ and $g_2 = \{3\}$, given as $Z_1(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$. We denote it as **Case II** since chain groups g_1 and g_2 do not form a partition of all the chains. The missing blue vertical arrows between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_3$ comparing to Figure 2.2 identify the contemporaneous independence between chain group g_1 and g_2.

Figure 2.6 and 2.7 also provide intuition that, although both the conditional independence assumption and the contemporaneous independence assumption focus on concurrent interactions among the chains, the contemporaneous independence assumption 2.9 generalises the former by allowing for partially missing concurrent interactions among the chains.

Catering to the contemporaneous independence assumption, we extend the definition of the univariate marginal η in Equation (2.4) to the group-wise as:

$$\eta_j = \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1)\big) = \sum_{\boldsymbol{Z}_{-j}(t)} \mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)\big) \tag{2.22}$$

where $\boldsymbol{Z}_j(t)$ is the marginal process of MMC $\boldsymbol{Z}$ with respect to the group g_j. Similarly, given a state vector $\boldsymbol{a} = (a_1, \ldots, a_K) \in \boldsymbol{\mathcal{S}}$ for $\boldsymbol{Z}$ and $\boldsymbol{b}_j \in \boldsymbol{\mathcal{S}}_j$ for the marginal process $\boldsymbol{Z}_j$, we define the realization of η_j as:

$$\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}) = \mathbb{P}\big(\boldsymbol{Z}_j(t) = \boldsymbol{b}_j \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) \tag{2.23}$$

$\eta_k(b_k \mid \boldsymbol{a})$ can be interpreted as the marginal probability of marginal process $\boldsymbol{Z}_j$ being in

state $\boldsymbol{b}_j$ at current time given that the previous states are $\boldsymbol{a}$. This allows us to rewrite the contemporaneous independence assumption in Equation (2.18) in terms of the newly defined η for chain group $g_{j''} = g_j \cup g_{j'}$ as:

$$\begin{aligned}\eta_{j''} =& \mathbb{P}\big(\boldsymbol{Z}_{j''}(t) \mid \boldsymbol{Z}(t-1)\big) \\ =& \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1)\big)\mathbb{P}\big(\boldsymbol{Z}_{j'}(t) \mid \boldsymbol{Z}(t-1)\big) = \eta_j \eta_{j'}\end{aligned} \tag{2.24}$$

In Case I, as illustrated in Figure 2.6, the chain subgroups $g_1, g_2, \ldots, g_J$ form a partition of all chains in $[K]$. In this setup, the group-wise marginal transition probabilities η_j for $j \in [J]$ can still be directly linked to the joint transition matrix using a product structure similar to Equation 2.6 under conditional independence:

$$p_{\boldsymbol{ab}} = \prod_{j=1}^{J} \eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}).$$

This formulation allows one to trivially extend the properties of conditional independence, such as the bijection between the marginal transitions η and the joint transition probabilities p, to the contemporaneous independence setting by aggregating each chain group into a single univariate chain over an extended state space.

Example 2.10. Consider the contemporaneous independence assumption in Equation (2.20). For simplicity, we assume that all three chains share the same state space $\mathcal{S} = \{1, 2\}$. The idea is to aggregate chain 1 and 2 in group g_1 into an univariate chain $Z_1'(t)$ over an extended state space according to the following:

$$\begin{aligned} Z_1'(t) = 1 &\iff \boldsymbol{Z}_1(t) = (1,1) & \qquad Z_1'(t) = 2 &\iff \boldsymbol{Z}_1(t) = (1,2) \\ Z_1'(t) = 3 &\iff \boldsymbol{Z}_1(t) = (2,1) & \qquad Z_1'(t) = 4 &\iff \boldsymbol{Z}_1(t) = (2,2) \end{aligned}$$

We retain chain 3 as a univariate chain $Z_3(t)$ in group g_2. As a result, each original 3-dimensional state vector $\boldsymbol{Z}(t) = \big(Z_1(t), Z_2(t), Z_3(t)\big)$ is transformed into a 2-dimensional state vector $\boldsymbol{Z}'(t) = \big(Z_1'(t), Z_3(t)\big)$, where $Z_1'(t)$ and $Z_3(t)$ take values in state spaces of

sizes 4 and 2, respectively. The transformed $\boldsymbol{Z}'(t)$ aligns with the structure of conditional independence for two chains, enabling us to apply the properties associated with the conditional independence, as discussed in Section 2.4.1, directly to the contemporaneous independence setting.

In the more general Case II, we do not obtain straightforward properties like the bijective relationship seen in the previous case. This is because the conditional formulation in Equation (2.21) only outlines the relationship among different η_j, instead of directly linking them to the joint transition probabilities p. Nonetheless, we can work out the number of the constraints introduced to the transition matrix P given as the following Proposition 2.11:

Proposition 2.11. *Imposing the contemporaneous independence assumption 2.19 introduces an additional* $(s^{|g_j|}-1)(s^{|g_{j'}|}-1)$ *constraints for each of the row in the transition matrix* P*, where* $|g_j|$ *and* $|g_{j'}|$ *denote the number of chains in the chain groups* g_j *and* $g_{j'}$*, respectively, and* s *is the number of possible states per chain.*

Proof. We express the contemporaneous independence assumption in terms of marginal transition probabilities η as in Equation (2.23). Consider the realization of this relationship as:

$$\eta_{j''}(\boldsymbol{b}_{j''} \mid \boldsymbol{a}) = \eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})\eta_{j'}(\boldsymbol{b}_{j'} \mid \boldsymbol{a})$$

which holds for all combinations of $\boldsymbol{b}_{j''} = (\boldsymbol{b}_j, \boldsymbol{b}_{j'})$, with $\boldsymbol{b}_j \in \boldsymbol{\mathcal{S}}_j$ and $\boldsymbol{b}_{j'} \in \boldsymbol{\mathcal{S}}_{j'}$, under the contemporaneous independence assumption. This yields a total of $s^{|g_j|} \times s^{|g_{j'}|}$ equations, one for each combination of states.

However, since $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$ and $\eta_{j'}(\boldsymbol{b}_{j'} \mid \boldsymbol{a})$ each lie in a simplex as described in Equation (2.9), only $(s^{|g_j|}-1)(s^{|g_{j'}|}-1)$ of these equations are linearly independent. This result justifies Proposition 2.11. □

Furthermore, after a suitable transformation, the general contemporaneous independence

structure observes a similar geometric interpretation as conditional independence in Section 2.4.2, which we will explore in the next section.

2.5.2 The Geometric Interpretation of Contemporaneous Independence

We now turn to a geometric interpretation of the contemporaneous independence assumption, using Case I (2.20) and Case II (2.21) as illustrative examples. For simplicity, we consider the case where all three chains share a common state space $\mathcal{S} = \{1, 2\}$.

2.5.2.1 Case I

We start with Case I, as illustrated in Figure 2.6, where $K = 3$, $g_1 = 1, 2$, and $g_2 = 3$. Under full dependency, the corresponding transition matrix P will be of dimension 8×8, since the size of the state space is $|\boldsymbol{\mathcal{S}}| = |\{1, 2\}^3| = 8$. Each row of P, say $P_{\boldsymbol{a}\cdot} = \{p_{\boldsymbol{ab}}\}_{\boldsymbol{b}\in\boldsymbol{\mathcal{S}}}$, belongs to the standard 7-dimensional simplex, denoted as $\boldsymbol{\Delta}^7$.

More precisely, we define $P_{\boldsymbol{a}\cdot}$ and $H_{\boldsymbol{a}\cdot}$ with the abbreviated notation as:

$$P_{\boldsymbol{a}\cdot} = \begin{matrix}\boldsymbol{a}\end{matrix}\begin{pmatrix} \overset{(1,1,1)}{p_{\boldsymbol{a}1}} & \overset{(1,1,2)}{p_{\boldsymbol{a}2}} & \overset{(1,2,1)}{p_{\boldsymbol{a}3}} & \overset{(1,2,2)}{p_{\boldsymbol{a}4}} & \overset{(2,1,1)}{p_{\boldsymbol{a}5}} & \overset{(2,1,2)}{p_{\boldsymbol{a}6}} & \overset{(2,2,1)}{p_{\boldsymbol{a}7}} & \overset{(2,2,2)}{p_{\boldsymbol{a}8}} \end{pmatrix} \tag{2.25}$$

$$H_{\boldsymbol{a}\cdot} = \begin{matrix}\boldsymbol{a}\end{matrix}\begin{pmatrix} \overset{(1,1,\cdot)}{\eta_{\boldsymbol{a}1}} & \overset{(1,2,\cdot)}{\eta_{\boldsymbol{a}2}} & \overset{(2,1,\cdot)}{\eta_{\boldsymbol{a}3}} & \overset{(2,2,\cdot)}{\eta_{\boldsymbol{a}4}} & \overset{(\cdot,\cdot,1)}{\eta_{\boldsymbol{a}5}} & \overset{(\cdot,\cdot,2)}{\eta_{\boldsymbol{a}6}} \end{pmatrix}$$

Rewrite the contemporaneous independence assumption 2.20 in terms of η and p as:

$$\begin{aligned} &\eta_{\boldsymbol{a}1}\eta_{\boldsymbol{a}5} = p_{\boldsymbol{a}\mathbf{1}} \quad \eta_{\boldsymbol{a}1}\eta_{\boldsymbol{a}6} = p_{\boldsymbol{a}\mathbf{2}} \quad \eta_{\boldsymbol{a}2}\eta_{\boldsymbol{a}5} = p_{\boldsymbol{a}\mathbf{3}} \quad \eta_{\boldsymbol{a}2}\eta_{\boldsymbol{a}6} = p_{\boldsymbol{a}\mathbf{4}} \\ &\eta_{\boldsymbol{a}3}\eta_{\boldsymbol{a}5} = p_{\boldsymbol{a}\mathbf{5}} \quad \eta_{\boldsymbol{a}3}\eta_{\boldsymbol{a}6} = p_{\boldsymbol{a}\mathbf{6}} \quad \eta_{\boldsymbol{a}4}\eta_{\boldsymbol{a}5} = p_{\boldsymbol{a}\mathbf{7}} \quad \eta_{\boldsymbol{a}4}\eta_{\boldsymbol{a}6} = p_{\boldsymbol{a}\mathbf{8}} \end{aligned} \tag{2.26}$$

Additionally, the marginalization relationships give:

$$p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2} = \eta_{\boldsymbol{a}1} \qquad p_{\boldsymbol{a}3} + p_{\boldsymbol{a}4} = \eta_{\boldsymbol{a}2} \qquad p_{\boldsymbol{a}5} + p_{\boldsymbol{a}6} = \eta_{\boldsymbol{a}3} \qquad p_{\boldsymbol{a}7} + p_{\boldsymbol{a}8} = \eta_{\boldsymbol{a}4}$$
$$p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3} + p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7} = \eta_{\boldsymbol{a}5} \qquad p_{\boldsymbol{a}2} + p_{\boldsymbol{a}4} + p_{\boldsymbol{a}6} + p_{\boldsymbol{a}8} = \eta_{\boldsymbol{a}5} \tag{2.27}$$

Combining Equations (2.26) and (2.27) and eliminating η, we obtain the section $\mathcal{M} \subset \boldsymbol{\Delta}^7$ identified by the contemporaneous independence assumption:

$$\frac{p_{\boldsymbol{a}1}}{p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2}} = \frac{p_{\boldsymbol{a}3}}{p_{\boldsymbol{a}3} + p_{\boldsymbol{a}4}} = \frac{p_{\boldsymbol{a}5}}{p_{\boldsymbol{a}5} + p_{\boldsymbol{a}6}} = \frac{p_{\boldsymbol{a}7}}{p_{\boldsymbol{a}7} + p_{\boldsymbol{a}8}} = p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3} + p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7} \tag{2.28}$$

Note that only 3 of these set of four equations are independent as each of the equality can be induced by the rest. Hence, there remains 3 constraints in the set of Equation 2.28.

Consider the equivalence relation $\sim$ on $\boldsymbol{\Delta}^7$ defined as $P_{\boldsymbol{a}\cdot} \sim P'_{\boldsymbol{a}\cdot}$ if they map to the same $H_{\boldsymbol{a}\cdot}$ through Equation (2.27). We denote the equivalence class as:

$$[P_{a\cdot}] = \{P'_{a\cdot} \in \boldsymbol{\Delta}^7 : P'_{a\cdot} \sim P_{a\cdot}\}$$

By Proposition 2.7, the section $\mathcal{M}$ in Equation 2.28 identified by contemporaneous independence assumption is actually a cross section. By letting $q_1 = p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2}$, $q_2 = p_{\boldsymbol{a}3} + p_{\boldsymbol{a}4}$, $q_3 = p_{\boldsymbol{a}5} + p_{\boldsymbol{a}6}$, $q_4 = p_{\boldsymbol{a}7} + p_{\boldsymbol{a}8}$, it is straight forward to see that

$$\big(q_1, q_2, q_3, q_4\big) \in \boldsymbol{\Delta}^3$$

Furthermore, we know that for each pair, take $p_{\boldsymbol{a}1}$ and $p_{\boldsymbol{a}2}$ for instance:

$$\big(p_{\boldsymbol{a}1}, p_{\boldsymbol{a}2}\big) \in \boldsymbol{\Delta}^1(q_1) = \Big\{p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2} = q_1, \quad \frac{p_{\boldsymbol{a}1}}{p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2}} = k \quad p_{\boldsymbol{a}1}, p_{\boldsymbol{a}2} \geq 0\Big\}$$

lie within $\boldsymbol{\Delta}^1$ subject to the same proportion $k = p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3} + p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7}$ indicated by Equation (2.28). This implies that $P_{\boldsymbol{a}\cdot}$ locates in the product space $\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1$, which is of

the same dimension as $H_{\boldsymbol{a}\cdot}$. The dimension also matches as (2.28) introduces 3 additional constraints, which restricts $P_{\boldsymbol{a}\cdot}$ from $\boldsymbol{\Delta}^7$ (7 dimensional) into $\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1$ (4 dimensional).

It is also worth noting that from Equation (2.17), we know that the transformed two chain system $\boldsymbol{Z}'(t)$ under the conditional independence assumption locates in the subspace:

$$\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1$$

which is consistent with what we obtained for the restricted parameter space of $\hat{P}_{\boldsymbol{a}\cdot}$ under contemporaneous independence assumption (2.20).

Finally, since no structure is imposed on the columns of P and H. Hence, the 8 rows of P and H are independent. As a result, under contemporaneous independence assumption in Equation (2.20)

$$P \in \hat{\mathcal{P}} = \underbrace{\left(\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1\right) \times \left(\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1\right) \times \cdots \times \left(\boldsymbol{\Delta}^3 \times \boldsymbol{\Delta}^1\right)}_{8 \text{ times}}$$

for the restricted parameter space $\hat{\mathcal{P}}$.

2.5.2.2 Case II

For the general Case II, we continue using the abbreviated notation for $P_{\boldsymbol{a}\cdot}$ introduced in Equation (2.25). Rewriting the contemporaneous independence condition (2.21) in terms of marginal transition probabilities η gives:

$$\eta_1 \eta_3 = \eta_{\{1,3\}}$$

Expressing $\eta_1, \eta_3, \eta_{\{1,3\}}$ in terms of p by marginalizing relationship (2.5) yields

$$(p_{\boldsymbol{a}1} + p_{\boldsymbol{a}2} + p_{\boldsymbol{a}3} + p_{\boldsymbol{a}4})(p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3} + p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7}) = p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3}$$

Letting $p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3} = p'_{\boldsymbol{a}1}$, $p_{\boldsymbol{a}2} + p_{\boldsymbol{a}4} = p'_{\boldsymbol{a}2}$, $p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7} = p'_{\boldsymbol{a}3}$, we obtain:

$$(p'_{\boldsymbol{a}1} + p'_{\boldsymbol{a}2})(p'_{\boldsymbol{a}1} + p'_{\boldsymbol{a}3}) = p'_{\boldsymbol{a}1} \tag{2.29}$$

Rearranging the equation leads to the expression:

$$p'_{\boldsymbol{a}3} = \frac{p'_{\boldsymbol{a}1}}{p'_{\boldsymbol{a}1} + p'_{\boldsymbol{a}2}} \cdot (1 - p'_{\boldsymbol{a}1} - p'_{\boldsymbol{a}2}), \tag{2.30}$$

which share the same form of the cross-section $\mathcal{M}$ previously introduced in Equation (2.16) under the conditional independence setting.

By taking $\left(p'_{\boldsymbol{a}1}, p'_{\boldsymbol{a}2}, p'_{\boldsymbol{a}3}\right)$ as coordinates, this substitution of variables can be interpreted geometrically as a projection from the 7-dimensional simplex $\boldsymbol{\Delta}^7$, which represents the parameter space of $P_{\boldsymbol{a}\cdot}$, into a lower-dimensional space with the following map:

$$\begin{aligned} x &= p'_{\boldsymbol{a}1} = p_{\boldsymbol{a}1} + p_{\boldsymbol{a}3}, \\ y &= p'_{\boldsymbol{a}2} = p_{\boldsymbol{a}2} + p_{\boldsymbol{a}4}, \\ z &= p'_{\boldsymbol{a}3} = p_{\boldsymbol{a}5} + p_{\boldsymbol{a}7}. \end{aligned}$$

From the perspective of dimensionally, this corresponds to projecting from $\boldsymbol{\Delta}^7$ onto 4 coordinates. The resulting 3 degrees of freedom span the 3-dimensional simplex $\boldsymbol{\Delta}^3$ depicted in Figure 2.5. Note that the projection towards the forth coordinates $p'_{\boldsymbol{a}4} = p_{\boldsymbol{a}6} + p_{\boldsymbol{a}8}$ is not illustrated in Figure 2.5 as $p'_{\boldsymbol{a}4}$ is constrained subject to the simplex constraint:

$$p'_{\boldsymbol{a}4} = 1 - p'_{\boldsymbol{a}1} - p'_{\boldsymbol{a}2} - p'_{\boldsymbol{a}3}$$

This leads to the insight that the geometric representation of the restricted parameter space $\hat{P}_{\boldsymbol{a}\cdot}$ under the contemporaneous independence assumption (2.21), when projected onto the four coordinates $p'_{\boldsymbol{a}1}, \ldots, p'_{\boldsymbol{a}4}$, forms a 2-dimensional manifold $\mathcal{M}$ as defined in Equation (2.30). This implies that the original restricted space $\hat{P}$ lies on a 6-dimensional manifold within the 7-simplex.

This projection with respect to $p'_{\boldsymbol{a}1}, \ldots, p'_{\boldsymbol{a}4}$ can also be interpreted as marginalizing out the influence of $\boldsymbol{Z}_2$, which does not appear explicitly in the contemporaneous independence assumption (2.21). Notably, each of the projected coordinates sums over different realizations of $Z_2(t)$ while holding $Z_1(t)$ and $Z_3(t)$ fixed. Once the influence of $\boldsymbol{Z}_2$ is marginalized out, the resulting structure between $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_3$ matches that of conditional independence.

2.5.3 The General Case

Consider the general case where the system of the MMC consists K chains with a shared state space $\mathcal{S} = \{1, 2, \ldots, s\}$:

2.5.3.1 Case I

Let the K Markov Chains partitioned into subgroups $g_1, g_2, \ldots, g_J$ where

$$g_1 \cup g_2 \cup \cdots \cup g_J = [K] \qquad \text{and} \qquad g_j \cap g_{j'} = \emptyset \qquad \forall j, j' \in [J]$$

Note that $J = K$ refers to the case of conditional independence as all the subsets $g_1, g_2, \ldots, g_j$ being a singleton.

The contemporaneous independence relationship among $g_1, g_2, \ldots, g_J$ can then be expressed as:

$$\boldsymbol{Z}_1(t) \perp\!\!\!\perp \boldsymbol{Z}_2(t) \perp\!\!\!\perp \ldots \perp\!\!\!\perp \boldsymbol{Z}_J(t) \mid \boldsymbol{Z}(t-1) \tag{2.31}$$

By aggregating each chain group into a single univariate chain over an extended state space:

$$Z'_1(t) \perp\!\!\!\perp Z'_2(t) \perp\!\!\!\perp \ldots \perp\!\!\!\perp Z'_j(t) \mid \boldsymbol{Z}(t-1)$$

where $Z'_j(t)$ is the aggregated univariate chain for group g_j with an extended state space of $s^{|g_j|}$ states. Here, $|g_j|$ represents the cardinality of the set g_j, or in other words the number of chains involved in the subgroup g_j.

The geometric interpretation becomes clear after the transformation. Consider a row of the transition matrix $P_{\boldsymbol{a}\cdot}$. Initially, this row lies within the simplex $\boldsymbol{\Delta}^{s^K-1}$ subject to the row constraints for a proper transition matrix. The contemporaneous independence assumption then restricts this row to a subspace formed by the product of K simplexes:

$$\boldsymbol{\Delta}^{s^{|g_1|}-1} \times \boldsymbol{\Delta}^{s^{|g_2|}-1} \times \cdots \times \boldsymbol{\Delta}^{s^{|g_J|}-1}$$

defined by $Z'_1(t), Z'_2(t) \ldots, Z'_J(t)$, respectively.

2.5.3.2 Case II

Consider a contemporaneous independence relationship between two chain groups g_j and $g_{j'}$, which do not form a partition of all the chains ($g_j \cup g_{j'} \neq [K]$):

$$\boldsymbol{Z}_1(t) \perp\!\!\!\perp \boldsymbol{Z}_2(t) \mid \boldsymbol{Z}(t-1)$$

Then, according to Proposition 2.11, the parameter space for each row $P_{\boldsymbol{a}\cdot}$ is now constrained to a lower-dimensional manifold with dimension:

$$\underbrace{s^K - 1}_{\text{Original Dimension}} - \underbrace{(s^{|g_j|}-1)(s^{|g_{j'}|}-1)}_{\substack{\text{Additional Constraints by} \\ \text{Contemporaneous Independence}}} = s^K + s^{|g_j|} + s^{|g_{j'}|} - s^{|g_j|+|g_{j'}|}$$

This dimensionality represents the space of feasible transition probabilities initially inside the simplex $\boldsymbol{\Delta}^{s^K}$ but now restricted by the contemporaneous independence assumption.

Moreover, consider the manifold defined by the contemporaneous independence condition involving chain groups g_j and $g_{j'}$, where $\boldsymbol{b}_j \in \boldsymbol{\mathcal{S}}_j$, $\boldsymbol{b}_{j'} \in \boldsymbol{\mathcal{S}}_{j'}$, and $\boldsymbol{b}_{j''} = (\boldsymbol{b}_j, \boldsymbol{b}_{j'})$. Then the

assumption implies:

$$\eta_{j''}(\boldsymbol{b}_{j''} \mid \boldsymbol{a}) = \eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})\eta_{j'}(\boldsymbol{b}_{j'} \mid \boldsymbol{a})$$

Rewriting each η in terms of the joint transition probabilities $p_{\boldsymbol{ab}}$ leading to the following relationship:

$$\sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j,\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} = \Big(\sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j} p_{\boldsymbol{ab'}}\Big)\Big(\sum_{\boldsymbol{b'}:\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}}\Big) \tag{2.32}$$

Letting the coordinates (x, y, z) as:

$$\begin{aligned} x &= \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j,\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} \\ y &= \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j} p_{\boldsymbol{ab'}} - \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j,\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} = \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j,\boldsymbol{b}'_{j'}\neq\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} \\ z &= \sum_{\boldsymbol{b'}:\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} - \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j=\boldsymbol{b}_j,\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} = \sum_{\boldsymbol{b'}:\boldsymbol{b}'_j\neq\boldsymbol{b}_j,\boldsymbol{b}'_{j'}=\boldsymbol{b}_{j'}} p_{\boldsymbol{ab'}} \end{aligned} \tag{2.33}$$

Substituting into Equation (2.32), we obtain:

$$x = (x + y)(x + z)$$

which matches the form of the 2-dimensional manifold previously defined in Equation (2.29). This implies that the manifold induced by the contemporaneous independence constraint (2.32) together with the full parameter space $\boldsymbol{\Delta}^{s^K-1}$, when projected onto the coordinate system (x, y, z) as defined in Equation (2.33), lie within a 3-dimensional space as the 2-dimensional cross-section $\mathcal{M}$ and the 3-dimensional simplex $\boldsymbol{\Delta}^3$ respectively, as visualized in Figure 2.5.

Furthermore, this projection can be naturally interpreted as a marginalization. When we marginalize over the chains not included in groups g_1 and g_2 in the contemporaneous independence constraint (2.17), the resulting geometric structure simplifies to the conditional independence case.

2.6 Granger non-Causality

2.6.1 Definition of Granger non-Causality and its Properties

In the previous section on marginal models, we considered the scenario each chain or group of chains depends on all prior states. More generally, it may be the case that only certain subsets of states influence the future states of a given chain, leading to the concept of Granger causality. Granger causality, originally introduced by Granger [1969b], refers to whether one time series contains information useful for forecasting another. In the context of multivariate Markov chains, Colombi and Giordano [2012] provides a definition of Granger non-causality as follows:

Assumption 2.12. *[Colombi and Giordano, 2012] Given two disjoint groups of chains* $g_j, g_{j'} \subset [K]$ *and the corresponding marginal processes* $\boldsymbol{Z}_j$ *and* $\boldsymbol{Z}_{j'}$ *of* $\boldsymbol{Z}$*,* $\boldsymbol{Z}_j$ *is* ***not Granger caused*** *by* $\boldsymbol{Z}_{j'}$ *with respect to* $\boldsymbol{Z}$ *if and only if for every discrete time* $t = 1, 2, \ldots, T$*:*

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}\big(1:(t-1)\big) \mid \boldsymbol{Z}_{-j'}\big(1:(t-1)\big)$$

Under the first-order Markov model, the Granger non-causality assumption can be further simplified as:

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}(t-1) \mid \boldsymbol{Z}_{-j'}(t-1) \tag{2.34}$$

which implies that the current states of chain group $g_{j'}$ do not provide any additional information about the future states of chain group g_j, given the current states of the rest chains $[K] \setminus g_{j'}$

It follows that we can represent the Granger non-causality in Equation (2.34) in terms of the conditional probabilities:

$$\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1)\big) = \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_{-j'}(t-1)\big) \tag{2.35}$$

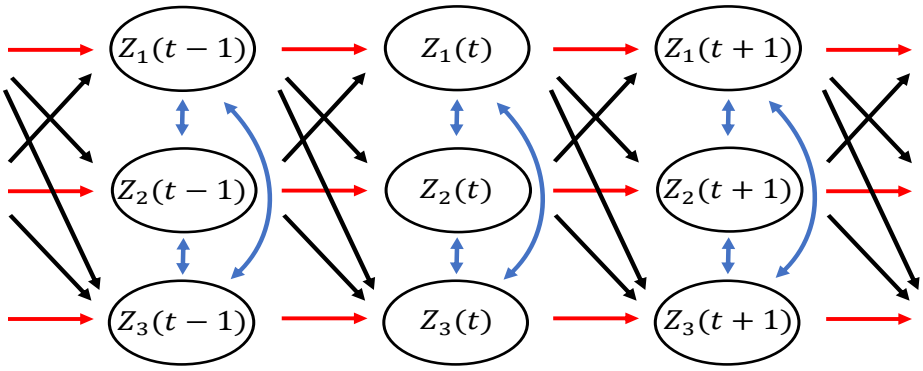


Figure 2.8: An illustration of the MMC with $K = 3$ chains under the Granger non-causality, by picking $g_1 = \{1, 2\}$ and $g_2 = \{3\}$, given as $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid \boldsymbol{Z}_{-3}(t-1)$. The black arrows pointing $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ at current time from $\boldsymbol{Z}_3$ at previous time comparing to Figure 2.2, indicates that $Z_1(t)$ and $Z_2(t)$ do not depend on the previous state $Z_3(t)$, or in other words, $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are not Granger caused by $\boldsymbol{Z}_3$.

Figure 2.8 gives an illustration of the Granger non-causality. Consider the case where $K = 3$, $g_1 = \{1, 2\}$ and $g_2 = \{3\}$:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1) \tag{2.36}$$

The Granger non-causality in this case states that the current state of chain 3 within group g_2 does not provide additional information of future state of chain 1 and chain 2 within group g_1, as indicated by the absence of black arrows directing $Z_1(t)$ and $Z_2(t)$ from $Z_3(t-1)$ in Figure 2.8, comparing to those under full dependency in Figure 2.2.

We begin by introducing the definition of the row group $\mathcal{A}$, which will serve as a key concept for analyzing the structural constraints that Granger non-causality imposes on the joint transition matrix.

Definition 2.13. *Suppose the chain groups g_j, $g_{j'}$, and $g_{j''}$ form a partition of all the chains, that is, $[K] = g_j \cup g_{j'} \cup g_{j''}$. Then, we define the set of rows $\mathcal{A}(\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''})$ as:*

$$\mathcal{A}(\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''}) = \{\boldsymbol{a} = (\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''}) : \boldsymbol{Z}_j = \boldsymbol{a}_j, \boldsymbol{Z}_{j'} = \boldsymbol{a}_{j'}, \boldsymbol{Z}_{j''} = \boldsymbol{a}_{j''}\}$$

This includes the rows with the states being $\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''}$ *for chain groups* $g_j, g_{j'}$ *and* $g_{j''}$ *respectively. This defines a singleton set since there is only one row* $P_{\boldsymbol{a}\cdot}$ *associated with the specific state* $\boldsymbol{a} = (\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''})$.

We further define $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ *to represent the set:*

$$\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''}) = \{\boldsymbol{a} = (\boldsymbol{a}_j, \boldsymbol{a}_{j'}, \boldsymbol{a}_{j''}) : \boldsymbol{Z}_j = \boldsymbol{a}_j, \boldsymbol{Z}_{j'} \in \boldsymbol{\mathcal{S}}_{j'}, \boldsymbol{Z}_{j''} = \boldsymbol{a}_{j''}\}$$

This refers to the collection of all state vectors where the first and third subgroups are fixed to $\boldsymbol{a}_j$ *and* $\boldsymbol{a}_{j''}$ *respectively, while the second subgroup is allowed to vary over its state space* $\boldsymbol{\mathcal{S}}_{j'}$. *Trivially, the set for all the rows can be denoted as* $\mathcal{A}(\cdot, \cdot, \cdot)$.

Based on the definition of the row set $\mathcal{A}$, we can now formalize Proposition 2.14, which characterizes the structure of the joint transition matrix P as imposed by the Granger non-causality.

Theorem 2.14. *Under the* ***Granger non-causality*** ***2.12****, the joint transition matrix* P *must satisfy the following constraint where for all* $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)} \in \mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ *and for all* $\boldsymbol{b} = (\boldsymbol{b}_j, \boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}) \in \boldsymbol{\mathcal{S}}$*:*

$$\sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}^{(1)}\boldsymbol{b}} = \sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}^{(2)}\boldsymbol{b}} \tag{2.37}$$

Proof. The definition of the Granger non-causality in Equation 2.35 shares a similar structure with that of the group-wise marginal transition probabilities η_j defined under the contemporaneous independence assumption in Equation (2.22). This similarity enables us to re-express it using the marginal transition probabilities η as follows:

$$\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}^{(1)}) = \eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}^{(2)}) \tag{2.38}$$

where for $\forall \boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)} \in \mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ and $\boldsymbol{b}_j \in \boldsymbol{\mathcal{S}}_j$. This Equation (2.38) conveys the same interpretation as Equation (2.35): the current states in subgroup $\boldsymbol{Z}_{j'}$ do not provide information about the future states of subgroup $\boldsymbol{Z}_j$. By expanding the group marginal

probabilities η in Equation (2.38) in terms of the transition probabilities p:

$$\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}) = \sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{ab}} \tag{2.39}$$

we obtain Equation (2.37). □

Equation (2.37) shows that the Granger non-causality imposes constraints on the *columns* of the transition matrix P, in contrast to the *row* constraints induced by conditional independence and contemporaneous independence in Equation (2.7). This is because the summation in Equation (2.37) over different rows of P yields equal values, highlighting that the influence of past states on the current state $\boldsymbol{Z}_j(t)$ remains the same across various combinations of previous states $\boldsymbol{Z}_{j'}(t-1)$ of group $g_{j'}$

Example 2.15. Consider the Granger non-causality for $K = 3$ chains demonstrated in Equation (2.36) with $g_1 = \{1, 2\}$ and $g_2 = \{3\}$. According to Proposition 2.14, this assumption imposes the following constraint on the entries p of the transition matrix P:

$$\sum_{b_3} p_{(a_1,a_2,1)(b_1,b_2,b_3)} = \sum_{b_3} p_{(a_1,a_2,2)(b_1,b_2,b_3)} \qquad \forall a_1, a_2, b_1, b_2 \in \{1, 2\}$$

Expanding the summation explicitly gives:

$$p_{(a_1,a_2,1)(b_1,b_2,1)} + p_{(a_1,a_2,1)(b_1,b_2,2)} = p_{(a_1,a_2,2)(b_1,b_2,1)} + p_{(a_1,a_2,2)(b_1,b_2,2)} \tag{2.40}$$

This condition effectively imposes a combination of row and column constraints on the joint transition probabilities p, as Equation 2.40 equates entries of p across different rows and columns. These constraints will be elaborated later in Section 2.6.2.

Proposition 2.16. *The imposition of the Granger non-causality Assumption 2.12 reduces the number of independent parameters for each group of rows $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ by:*

$$(s^{|g_j|} - 1)(s^{|g_{j'}|} - 1) = (n_j - 1)(n_{j'} - 1)$$

Here, s represents the number of states shared by all chains. $|g_j|$ and $|g_{j'}|$ denote the

number of chains within subgroups g_j and $g_{j'}$, respectively. To simplify the notation, we set $n_j = s^{|g_j|}$ and $n_{j'} = s^{|g_{j'}|}$ representing the number of possible states within the chain groups g_j and $g_{j'}$, respectively.

Proof. The proof can be divided into two parts:

1. Consider a group of rows $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$. There are $n_{j'}$ such distinct values for $\boldsymbol{a}$ within this group, resulting in $n_{j'} - 1$ equations of the form:

$$\sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}^{(1)}\boldsymbol{b}} = \sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}^{(2)}\boldsymbol{b}} = \cdots = \sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}} \tag{2.41}$$

   for distinct $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(n_{j'})} \in \mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$. These equations impose $n_{j'} - 1$ constraints on p.

2. There are n_j such sets of Equation (2.41), depending on the value of $\boldsymbol{b}_j$. However, the last one, say $\boldsymbol{b}^{(n_j)}$, depends entirely on the previous $n_j - 1$ sets of equations as follows:

$$\sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}\boldsymbol{b}^{(n_j)}} = 1 - \sum_{n'=1}^{n_j - 1} \left(\sum_{\boldsymbol{b}_{j'}, \boldsymbol{b}_{j''}} p_{\boldsymbol{a}\boldsymbol{b}^{(n')}} \right)$$

   Due to the summation of each row of the transition matrix P should be 1.

Combining the above, there are $(n_j - 1)(n_{j'} - 1)$ constraints within each row group $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$. Let $n_{j''}$ denotes the number of possible states within the chain group $g_{j''} = [K] \setminus (g_j \cup g_{j'})$. Specifically, $n_{j''} = 1$ when $g_{j''} = \emptyset$. As there are $n_j n_{j''}$ distinct row groups depending on the values of $\boldsymbol{a}_j$ and $\boldsymbol{a}_{j''}$, the total reduction in the number of independent parameters under Granger non-causality Assumption 2.12 is given by:

$$n_j n_{j''} (n_j - 1)(n_{j'} - 1) \tag{2.42}$$

□

Example 2.17. Continuing with Example 2.15, we observe that Equation (2.40) yields $2^4 = 16$ different equations when considering all possible combinations of $a_1, a_2, b_1, b_2 \in$

$\{1, 2\}$. Among these, only 12 equations are linearly independent. This is because, for each fixed pair (a_1, a_2), the equation corresponding to $(b_1, b_2) = (2, 2)$ can be derived from the equations for $(b_1, b_2) = (1, 1)$, $(1, 2)$, and $(2, 1)$ due to the row sum constraint. These 12 independent equations represent the additional constraints imposed by the Granger non-causality assumption, as specified in the above Equation (2.42) in Proposition 2.16.

$$n_j n_{j''}(n_j - 1)(n_{j'} - 1) = 4 \times 1 \times (4 - 1) \times (2 - 1) = 12$$

where $n_j = 2^2 = 4$, $n_{j'} = 2$ and $n_{j''} = 1$ for empty $g_{j''}$.

2.6.2 The Geometric Interpretation of Granger Non-Causality

In Sections 2.4.2 and 2.5.2, we explored the geometric structure of conditional independence and contemporaneous independence. These two assumptions only impose constraints within each row of the transition matrix P, meaning each row can be analysed independently. Consequently, our geometric analysis focused on a row-by-row basis. In contrast, Granger non-causality imposes constraints across columns of P, thereby introducing structured dependencies between different rows. This necessitates a different approach to geometric interpretation that accounts for these column-wise interactions.

Consequently, we will now investigate the geometric structure of P group by group, using the same grouping $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ as defined in Proposition 2.14, which we used to explore the properties of Granger non-causality. This group-based analysis is adopted because, under Granger non-causality, the transition probabilities p associated with different row groups $\mathcal{A}$ are mutually independent.

2.6.2.1 Simple Case with $K = 3$ Chains

Consider the simple case of an MMC with a shared state space $\mathcal{S} = \{1, 2\}$ and $K = 3$ chains as illustrated in Figure 2.8 with the following Granger non-causality relationship:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1) \tag{2.43}$$

with $g_1 = \{1, 2\}$, $g_2 = \{3\}$ and $g_3 = [K] \setminus (g_1 \cup g_2) = \emptyset$.

In this case, we omit the last term in each row group because $\mathcal{A}(\boldsymbol{a}_1, \cdot, \boldsymbol{a}_3)$ reduces to $\mathcal{A}(\boldsymbol{a}_1, \cdot)$ as $g_3 = \emptyset$. Each group $\mathcal{A}(\boldsymbol{a}_1, \cdot)$ contains two rows corresponding to $\boldsymbol{a}_2 \in \{1, 2\}$. In the absence of any constraints, each of these rows belongs to the standard simplex $\boldsymbol{\Delta}^7$, due to the requirement that each row sums to 1. Therefore, the pair of rows indexed by $\boldsymbol{a} \in \mathcal{A}(\boldsymbol{a}_1, \cdot)$ lies in the product space of two 7-dimensional simplixes:

$$\boldsymbol{\Delta}^7 \times \boldsymbol{\Delta}^7$$

Under the Granger non-causality described in Equation (2.38), column constraints are directly imposed on the marginal transition matrix H. Equation (2.44) provides an illustration, where rows $(1, 1, 1)$ and $(1, 1, 2)$ both belong to the same row group $\mathcal{A}\big((1, 1), \cdot\big)$.

$$H = \begin{array}{c} \\ (1,1,1) \\ (1,1,2) \\ (1,2,1) \\ (1,2,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccc} (1,1,\cdot) & (1,2,\cdot) & (2,1,\cdot) & (2,2,\cdot) & (\cdot,\cdot,1) & (\cdot,\cdot,1) \end{array} \\ \left(\begin{array}{cccc:cc} \eta_{11} & \eta_{12} & \eta_{13} & \eta_{14} & \eta_{15} & \eta_{16} \\ \eta_{21} & \eta_{22} & \eta_{23} & \eta_{24} & \eta_{25} & \eta_{26} \\ \hdashline \eta_{31} & \eta_{32} & \eta_{33} & \eta_{34} & \eta_{35} & \eta_{36} \\ \eta_{41} & \eta_{42} & \eta_{43} & \eta_{44} & \eta_{45} & \eta_{46} \\ \hdashline \vdots & & \vdots & & \vdots & \end{array}\right) \end{array} \tag{2.44}$$

where η_{11} in the matrix H is the abbreviated version of $\eta_1\big((1, 1, \cdot) \mid (1, 1, 1)\big)$.

According to the Granger non-causality in Equation (2.43), the marginal transition probabilities η with respect to g_1 in these rows must be identical. This is visualized in Equation (2.44), where the η within each block, outlined by dotted lines, must match those of the

same colour within the same block. Specifically, in the top-left block,

$$\eta_{11} = \eta_{21}, \qquad \eta_{12} = \eta_{22}, \qquad \eta_{13} = \eta_{23}, \qquad \eta_{14} = \eta_{24} \tag{2.45}$$

While the Granger non-causality in Equation (2.43) does not impose any constraints on the η values shown in black in the right blocks of the marginal transition matrix H, as these η stands for the marginal of chain group $g_2 = \{3\}$ that does not involved in Granger causality. As a result, each pair of these η resides independently in a 2 dimensional simplex $\boldsymbol{\Delta}^2$.

If we expand the η terms in the left blocks of Equation (2.44) using the marginalization relationship defined in Equation (2.22), we obtain the structure that Granger non-causality, as expressed in Equation (2.43), imposes on the joint transition matrix P, illustrated in Equation (2.46).

$$P = \begin{array}{c} \\ (1,1,1) \\ (1,1,2) \\ (1,2,1) \\ (1,2,2) \\ \vdots \end{array} \begin{array}{c} \begin{array}{cccccccc} (1,1,1) & (1,1,2) & (1,2,1) & (1,2,2) & (2,1,1) & (2,1,2) & (2,2,1) & (2,2,2) \end{array} \\ \left(\begin{array}{cccccccc} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & p_{18} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} & p_{28} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} & p_{37} & p_{38} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} & p_{47} & p_{48} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right) \end{array} \tag{2.46}$$

where the entries p_{11} is the abbreviated version of $p_{(1,1,1)(1,1,1)}$.

The entries p highlighted in the same colour within each row of P sum up to the η values of the same colour in the same row, via marginalization. For example:

$$p_{11} + p_{12} = \eta_{11} \qquad p_{13} + p_{14} = \eta_{12}$$

The equality among different η given in Equation (2.45) now transformed to p. Take first row group $\mathcal{A}\big((1,1),\cdot\big)$ in the first and second row of P in Equation (2.46) as an example:

$$p_{11}+p_{12}=\eta_{11}=\eta_{21}=p_{21}+p_{22}$$
$$p_{13}+p_{14}=\eta_{12}=\eta_{22}=p_{23}+p_{24}$$
$$p_{15}+p_{16}=\eta_{13}=\eta_{23}=p_{25}+p_{26}$$
$$p_{17}+p_{18}=\eta_{14}=\eta_{24}=p_{27}+p_{28}$$

This formulation makes the geometric representation of P more straightforward. Consider the first group of rows, specifically $(1,1,1)$ and $(1,1,2)$. Under the Granger non-causality, the first row $P_{1\cdot}$ remains free to reside anywhere in the simplex $\boldsymbol{\Delta}^7$. However, the second row $P_{2\cdot}$, conditioned on the first, must lie in a lower-dimensional subspace given as:

$$P_{2\cdot} \mid P_{1\cdot} \in \boldsymbol{\Delta}^2(p_{11}+p_{12}) \times \boldsymbol{\Delta}^2(p_{13}+p_{14}) \times \boldsymbol{\Delta}^2(p_{15}+p_{16}) \times \boldsymbol{\Delta}^2(p_{17}+p_{18}) \subset \boldsymbol{\Delta}^7$$

where $\boldsymbol{\Delta}^2(p_{11}+p_{12}) = \big\{(x_1,x_2) : x_1+x_2 = p_{11}+p_{12}, x_1, x_2 \geq 0\big\}$

This implies that, when fixing the first row $P_{1\cdot}$, the second row $P_{2\cdot}$ must lie in a subspace that is defined by the product of four 2-dimensional simplexes, $\boldsymbol{\Delta}^2$. Since each $\boldsymbol{\Delta}^2$ corresponds to a line segment, the overall space can be written as:

$$P_{2\cdot} \mid P_{1\cdot} \in [0, p_{11}+p_{12}] \times [0, p_{13}+p_{14}] \times [0, p_{15}+p_{16}] \times [0, p_{17}+p_{18}] \qquad \text{with} \quad \sum_{i=1}^{8} p_{1i} = 1$$

Hence, the joint of these two rows locates in the product space as:

$$\big(P_{1\cdot}, P_{2\cdot}\big) \in \boldsymbol{\Delta}^7 \times \boldsymbol{\Delta}^2(p_{11}+p_{12}) \times \boldsymbol{\Delta}^2(p_{13}+p_{14}) \times \boldsymbol{\Delta}^2(p_{15}+p_{16}) \times \boldsymbol{\Delta}^2(p_{17}+p_{18}) \quad (2.47)$$

Since under Granger non-causality these 4 distinct row groups $\mathcal{A}(\boldsymbol{a}_1,\cdot)$ corresponding to

the different values of $\boldsymbol{a}_1 \in \{1,2\}^2$ are mutually independent, that is

$$\left(P_{1\cdot}, P_{2\cdot}\right) \perp\!\!\!\perp \left(P_{3\cdot}, P_{4\cdot}\right) \perp\!\!\!\perp \left(P_{5\cdot}, P_{6\cdot}\right) \perp\!\!\!\perp \left(P_{7\cdot}, P_{8\cdot}\right)$$

It then follows that the entire transition matrix P lies in the product space formed by four such subspaces, as defined in Equation (2.47), where each is constructed by substituting $P_{1\cdot}$ with $P_{3\cdot}, P_{5\cdot}$, and $P_{7\cdot}$, respectively.

It is also easy to verify that, in this case, Granger non-causality imposes 3 additional constraints on the second row compared to the fully unconstrained case. Since there are 4 distinct row groups, this results in a total of 12 additional constraints imposed on the entire transition matrix P, which is consistent with the results in Example 2.17.

2.6.2.2 General Case

We now turn to the general case of a multivariate Markov chain (MMC) with K chains, where each chain shares a common state space $\mathcal{S}$ of size s. Furthermore, we assume that the chain group $g_{j''}$ is non-empty. As in the previous analysis, we examine the structure of the constraints imposed by the Granger non-causality assumption on the transition matrix P, by evaluating each group of rows separately.

Consider a group of rows from the set $\mathcal{A}(\boldsymbol{a}_j, \cdot, \boldsymbol{a}_{j''})$ defined in Proposition 2.14. Under full dependency, there is no additional constraints imposed on these rows. Therefore, each row independently lies within a standard simplex $\boldsymbol{\Delta}^{s^K-1}$, subjecting to the constraint that each row sums to 1. Consequently, the group of rows $\boldsymbol{a}^{(1)}, \boldsymbol{a}^{(2)}, \ldots, \boldsymbol{a}^{(n_{j'})}$ resides jointly in the product space of $n_{j'}$ simplexes

$$\underbrace{\boldsymbol{\Delta}^{s^K-1} \times \boldsymbol{\Delta}^{s^K-1} \times \cdots \times \boldsymbol{\Delta}^{s^K-1}}_{n_{j'} \text{ times}}$$

Similarly, we start with the marginal transition matrix H, since column constraints imposed by Granger non-causality, as shown in Equation (2.38), are directly on the marginal

transition matrix H. This is further illustrated in Equation (2.48), where the parameters η labelled with the same colour are forced to take the same value, rather than being independent as they would be without Granger non-causality.

$$
\begin{array}{c|cccc:c:c}
 & \boldsymbol{b}_u & & & & \boldsymbol{b}_v & \boldsymbol{b}_w \\
\boldsymbol{a}^{(1)} & \eta_{j'}(\boldsymbol{b}_{j'}^{(1)} \mid \boldsymbol{a}^{(1)}) & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(1)}) & \cdots & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(1)}) & & \\
\boldsymbol{a}^{(2)} & \eta_{j'}(\boldsymbol{b}_{j'}^{(1)} \mid \boldsymbol{a}^{(2)}) & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(2)}) & \cdots & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(2)}) & \text{Unrestricted} & \text{Unrestricted} \\
\vdots & \vdots & \vdots & & \vdots & & \\
\boldsymbol{a}^{(n_{j'})} & \eta_{j'}(\boldsymbol{b}_{j'}^{(1)} \mid \boldsymbol{a}^{(1)}) & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(1)}) & \cdots & \eta_{j'}(\boldsymbol{b}_{j'}^{(2)} \mid \boldsymbol{a}^{(1)}) & & \\
\hdashline
\vdots & \multicolumn{4}{c:}{\vdots} & \vdots & \vdots
\end{array}
\tag{2.48}
$$

Consequently, within the row groups $\mathcal{A}$, as depicted as the top three blocks of H, there are only $(n_j - 1)$ independent parameters remaining, instead of the $(n_j - 1)n_{j'}$ parameters present without constraints, as specified by Equation (2.9). Note that the reduction in the number of independent parameters by $(n_j - 1)(n_{j'} - 1)$ aligns with the result provided in Proposition 2.16.

Similar as previous, we then expand the marginal transition probabilities η within H according to the marginalization relationship defined in Equation (2.39), to obtain the transition matrix P, as illustrated in Equation (2.49).

According to the marginalization process, there are $n_1 = n_{j'} \times n_{j''}$ entries of the joint transition probabilities p that sum up to each marginal transition probability η. Without loss of generality, we assign the first n_1 entries, namely $p_{\boldsymbol{a}^{(i)}\boldsymbol{b}^{(1)}}, \ldots, p_{\boldsymbol{a}^{(i)}\boldsymbol{b}^{(n_1)}}$ in red, which marginalizes to the red $\eta_{j'}(\boldsymbol{b}_{j'}^{(1)} \mid \boldsymbol{a}^{(i)})$, one for each of the $i = 1, \ldots, n_{j'}$ rows. Likewise, the green and blue elements in each block of p marginalize to their respective η values in H, as indicated by matching colours.

$$
\begin{array}{c}
\\
\boldsymbol{a}^{(1)} \\
\boldsymbol{a}^{(2)} \\
\vdots \\
\boldsymbol{a}^{(n_{j'})} \\
\vdots
\end{array}
\begin{pmatrix}
p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(1)}} & \cdots & p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(n_1)}} & p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(n_1+1)}} & \cdots & p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(2n_1)}} & \cdots & p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(n_2)}} & \cdots & p_{\boldsymbol{a}^{(1)}\boldsymbol{b}^{(s^K)}} \\
p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(1)}} & \cdots & p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(n_1)}} & p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(n_1+1)}} & \cdots & p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(2n_1)}} & \cdots & p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(n_2)}} & \cdots & p_{\boldsymbol{a}^{(2)}\boldsymbol{b}^{(s^K)}} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(1)}} & \cdots & p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(n_1)}} & p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(n_1+1)}} & \cdots & p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(2n_1)}} & \cdots & p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(n_2)}} & \cdots & p_{\boldsymbol{a}^{(n_{j'})}\boldsymbol{b}^{(s^K)}} \\
\vdots & \vdots & & \vdots & & & \vdots & & \vdots &
\end{pmatrix}
\tag{2.49}
$$

where we denote $n_1 = n_{j'} \times n_{j''}$, $n_2 = s^K + 1 - n_1$ to simplify the notation.

This formulation makes the geometric representation of P more straightforward. For each row $\boldsymbol{a} \in \mathcal{A}$, there are n_j different blocks, each labelled with a different colour in Equation (2.49). Each of these blocks lies within a simplex defined as:

$$
\boldsymbol{\Delta}^{n_1-1}(\delta_i) = \left\{ \left(p_{\boldsymbol{a}\boldsymbol{b}^{(1)}}, \ldots, p_{\boldsymbol{a}\boldsymbol{b}^{(n_1)}}\right) : \quad p_{\boldsymbol{a}\boldsymbol{b}^{(1)}}, \ldots, p_{\boldsymbol{a}\boldsymbol{b}^{(n_1)}} \geq 0 \quad \text{and} \quad \sum_{n'=1}^{n_1} p_{\boldsymbol{a}\boldsymbol{b}^{(n')}} = \delta_i \right\}
$$

where $n_1 = n_{j'} n_{j''}$ and δ_i is a fixed value shared by all blocks within a column labelled with the same colour, as introduced by the structure imposed by Granger non-causality. This simplex $\boldsymbol{\Delta}^{n_1-1}(\delta_i)$ is of dimension $n_1 - 1$ because it contains n_1 elements of p.

This leads to the geometric picture of the joint of rows $\boldsymbol{a}^{(1)}, \ldots, \boldsymbol{a}^{(2)}, \ldots, \boldsymbol{a}^{(n_{j'})} \in \mathcal{A}$ as:

$$
\boldsymbol{\Delta}^{s^K-1} \times \underbrace{\prod_{n=1}^{n_j} \boldsymbol{\Delta}^{n_1-1}(\delta_1) \times \prod_{n=1}^{n_j} \boldsymbol{\Delta}^{n_1-1}(\delta_2) \times \cdots \times \prod_{n=1}^{n_j} \boldsymbol{\Delta}^{n_1-1}(\delta_{n_{j'}})}_{n_{j'}-1 \text{ times}} \tag{2.50}
$$

The first term, $\boldsymbol{\Delta}^{s^K-1}$, represents having a completely independent row as that without any constraints, to determine the constants δ_i, which are shared by all the blocks within a

column. For the remaining $n_{j'}-1$ rows, there are n_j different blocks within each, and each block lies within its corresponding $\boldsymbol{\Delta}^{n_1-1}(\delta_i)$, defined by the independent row through δ_i. Note that, compared to unconstrained model where we have $n_{j'} \times (s^K - 1)$ independent parameters, here we have:

$$\underbrace{s^K - 1 + (n_{j'} - 1)n_j(n_{j'}n_{j''} - 1)}_{\text{From Equation (2.50)}} = \underbrace{n_{j'} \times (s^K - 1)}_{\text{Unconstrained model}} - \underbrace{(n_j - 1)(n_{j'} - 1)}_{\substack{\text{Number of Independent} \\ \text{Parameters Reduced}}}$$

This is consistent with the reduction in the number of independent parameters demonstrated in Proposition 2.16.

2.7 Dependency Structures within MMC

Having explored conditional independence, contemporaneous independence, and Granger non-causality, we are now able to combine these assumptions to fully characterize any desired dependency structure in a multivariate Markov chain.

In a first-order MMC under full dependency, each group of states $\boldsymbol{Z}_j(t)$ depends exactly on two components: interactions with concurrent states $\boldsymbol{Z}_{-j}(t)$ and influence from the previous time step $\boldsymbol{Z}(t-1)$. Graphically, this means that each groups of states $\boldsymbol{Z}_j(t)$ receives arrows from all the other state groups $\boldsymbol{Z}_{j'}(t-1)$ explaining the dependency from the previous and is bidirectionally connected to all other current state groups $\boldsymbol{Z}_{-j}(t)$ capturing mutual interactions. This is the structure when no additional dependency assumptions are imposed to the MMC.

- Any absence of arrows from a previous group of states $\boldsymbol{Z}_{j'}(t-1)$ to the current group of states $\boldsymbol{Z}_j(t)$ can be expressed by Granger non-causality as:

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}(t-1) \mid \boldsymbol{Z}_{-j'}(t-1)$$

indicating that the chains in group g_j are not Granger caused by those in group $g_{j'}$.

- Any absence of double-directed arrows between concurrent groups of states, for instance between $\boldsymbol{Z}_j(t)$ and $\boldsymbol{Z}_{j'}(t)$, can be written in terms of the contemporaneous independence assumption:

$$\boldsymbol{Z}_j(t) \perp\!\!\!\perp \boldsymbol{Z}_{j'}(t) \mid \boldsymbol{Z}(t-1)$$

indicating that the chains in group g_j are contemporaneously independent with the chains in group $g_{j'}$.

Thus, combining Granger non-causality with contemporaneous independence fully characterizes all possible dependency structures in a first-order MMC. This specification is formalized in the following Theorem 2.18.

Theorem 2.18. *Any dependency structures of a first-order MMC can be represented as a combination of* ***Contemporaneous Independence*** *and* ***Granger non-Causality****.*

Accordingly, we can also modify the adjacency matrix A constructed in Equation (2.1) to reflect the dependencies according to the Granger non-causality and contemporaneous independence assumptions.

- Granger non-Causality: Imposing that chain group $g_{j'}$ does not cause chain group g_j for $j, j' \in [J]$ is equivalent to setting the (j', j)-th element in the top part of A to zero.

- Contemporaneous Independence: Imposing contemporaneous independence between chain groups g_j and $g_{j'}$, for $j, j' \in [J]$, is equivalent to setting the $(j' + J, j)$ and $(j + J, j')$-th element in the bottom part of A to zero.

Example 2.19. Figure 2.9 illustrates the Theorem 2.18 with an MMC with 3 chains, where the different types of dependencies are represented using different coloured arrows. All these dependencies fall into two categories:

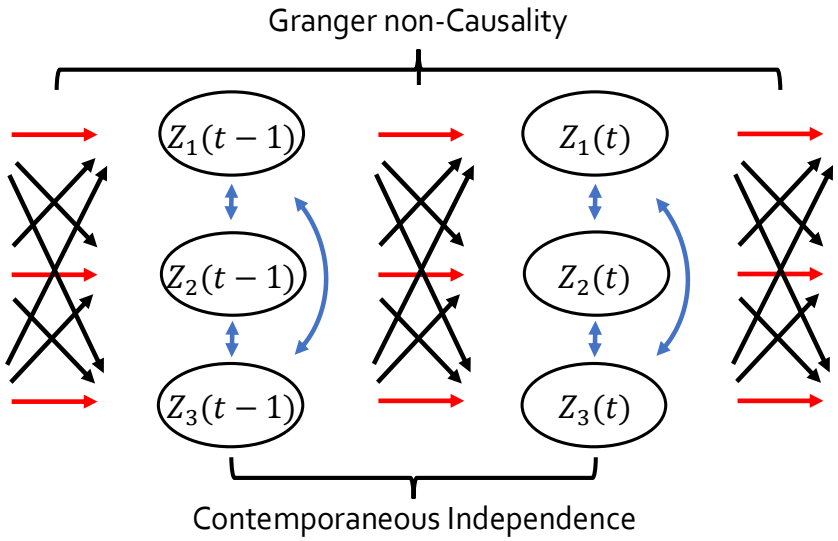


Figure 2.9: An illustration of the Theorem 2.18 using an MMC with 3 chains. All of the dependencies can be categorized into two types: (I) blue vertical arrows between concurrent states, characterized by the contemporaneous independence, and (II) black and red arrows from previous states to the current, governed by the Granger non-causality.

- **Dependency from the past (red and black arrows):** These arrows represent how the states at time $t-1$ affect the current states at time t. Red arrows indicate self-dependency within a chain, while black arrows denote inter-chain dependencies. These relationships are governed by **Granger non-causality**, which determines whether a chain is conditionally independent of other past states. For example, the absence of an arrow from $Z_3(t-1)$ to $Z_2(t)$ implies $Z_2(t)$ is not Granger caused by $Z_3(t-1)$, explicitly as $Z_2(t) \perp Z_3(t-1) \mid \boldsymbol{Z}_{-3}(t-1)$.

- **Concurrent dependencies (blue bidirectional arrows):** These arrows capture the mutual influence between chains at the same time point t, conditional on the past $\boldsymbol{Z}(t-1)$. The **Contemporaneous Independence** assumption formalizes this, stating which subsets of chains are conditionally independent of each other. For instance, the missing blue vertical arrow between $Z_2(t)$ and $Z_3(t)$ implies they are contemporaneously independent given $\boldsymbol{Z}(t-1)$, explicitly as $Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$.

Building on these dependencies, this chapter proceeds to review several widely used parametrisations of MMCs, including the *Cartesian Product* Model (see Section 2.8) and

linear decomposition based approaches including the *Inter-Chain Transition* (ICT) Model (see Section 2.9.1) and the *Mixture Transition Distribution* (MTD) Model (see Section 2.9.2), with particular attention to the challenges they confront in representing dependency structures.

2.8 Cartesian Product Model

A common method found in literature to represent the transition of a first-order MMC is by consolidating the state space into the joint K-dimensional state vector, which is subsequently modelled using a joint transition matrix P [see Brand et al., 1997, Ghosh et al., 2017, Pohle et al., 2021, and others]. As it is formulating the joint state space using the Cartesian product method, it's often referred to as the Cartesian Product model.

2.8.1 Formulation of the Cartesian Product Model

Consider a multivariate Markov chain (MMC) with K chains, where each chain k takes values in a discrete state space $\mathcal{S}^{(k)}$. The overall state space $\boldsymbol{\mathcal{S}}$ is then defined as the Cartesian product of the individual state spaces:

$$\boldsymbol{\mathcal{S}} = \mathcal{S}^{(1)} \times \mathcal{S}^{(2)} \times \cdots \times \mathcal{S}^{(K)}.$$

Assuming that all chains share a common discrete state space of size s, the joint state space has cardinality $|\boldsymbol{\mathcal{S}}| = s^K$. The joint transition matrix P is defined over this extended space, where each entry $p_{\boldsymbol{ab}}$ denotes the probability of transitioning from state $\boldsymbol{a}$ at time $t-1$ to state $\boldsymbol{b}$ at time t:

$$P = \{p_{\boldsymbol{ab}}\}_{\boldsymbol{a},\boldsymbol{b}\in\boldsymbol{\mathcal{S}}}, \qquad p_{\boldsymbol{ab}} = \mathbb{P}(\boldsymbol{Z}(t) = \boldsymbol{b} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a})$$

Here, $\boldsymbol{a} = (a_1, \ldots, a_K)$ and $\boldsymbol{b} = (b_1, \ldots, b_K)$ represent state vectors. The matrix P is of dimension $s^K \times s^K$, each row sums to 1, and under the time-homogeneous setting, P remains constant over time.

Example 2.20. Consider a MMC with $K = 3$ chains with each chain having a state space $\{1, 2\}$. Under the Cartesian Product model, this MMC can be fully encoded by a joint transition matrix P of dimension 8×8 given as:

$$P = \begin{array}{c} \\ (1,1,1) \\ (1,1,2) \\ (1,2,1) \\ (1,2,2) \\ (2,1,1) \\ (2,1,2) \\ (2,2,1) \\ (2,2,2) \end{array} \begin{array}{c} \begin{array}{cccccccc} (1,1,1) & (1,1,2) & (1,2,1) & (1,2,2) & (2,1,1) & (2,1,2) & (2,2,1) & (2,2,2) \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & p_{13} & p_{14} & p_{15} & p_{16} & p_{17} & p_{18} \\ p_{21} & p_{22} & p_{23} & p_{24} & p_{25} & p_{26} & p_{27} & p_{28} \\ p_{31} & p_{32} & p_{33} & p_{34} & p_{35} & p_{36} & p_{37} & p_{38} \\ p_{41} & p_{42} & p_{43} & p_{44} & p_{45} & p_{46} & p_{47} & p_{48} \\ p_{51} & p_{52} & p_{53} & p_{54} & p_{55} & p_{56} & p_{57} & p_{58} \\ p_{61} & p_{62} & p_{63} & p_{64} & p_{65} & p_{66} & p_{67} & p_{68} \\ p_{71} & p_{72} & p_{73} & p_{74} & p_{75} & p_{76} & p_{77} & p_{78} \\ p_{81} & p_{82} & p_{83} & p_{84} & p_{85} & p_{86} & p_{87} & p_{88} \end{pmatrix} \end{array}$$

where, for instance p_{11} specifies the transition probabilities of remaining in the joint state $(1, 1, 1)$ and p_{12} specifies the transition probabilities from the joint state $(1, 1, 1)$ to $(1, 1, 2)$

2.8.2 Encode dependency structures under Cartesian Product Model

While all the dependency structures are embedded entirely within the joint transition matrix P under the Cartesian Product model, it is usually implicit and challenging to derive. Take the simple case where an MMC with 3 chains and a state space $\mathcal{S} = \{1, 2\}$ is considered. Recall the contemporaneous independence assumption specified in Equation (2.21) where chain 1 and chain 3 are independent, such an independent relationship can be expressed via the entries p as:

$$\sum_{b_2 \in \mathcal{S}} p_{\boldsymbol{a}(b_1,b_2,b_3)} = \left(\sum_{b_2,b_3 \in \mathcal{S}} p_{\boldsymbol{a}(b_1,b_2,b_3)} \cdot \sum_{b_1,b_2 \in \mathcal{S}} p_{\boldsymbol{a}(b_1,b_2,b_3)} \right) \tag{2.51}$$

Similarly, one simple example of Granger non-causality in Equation (3.14) where chain 1 and chain 2 is not Granger caused by chain 3, can also be described in terms of the p. Specifically, for arbitrary $\boldsymbol{b} = (b_1, b_2, b_3) \in \boldsymbol{\mathcal{S}}$, we have:

$$\sum_{b_3 \in \mathcal{S}} p_{(a_1,a_2,1)(b_1,b_2,b_3)} = \sum_{b_3 \in \mathcal{S}} p_{(a_1,a_2,2)(b_1,b_2,b_3)} \qquad \forall a_1, a_2 \in \mathcal{S} \tag{2.52}$$

Expressing these dependency structure via the joint transition matrix P under the Cartesian Product model inevitably involves a lot of summation to marginalize out the chains that does not appear in the dependency relationship and some multiplications. In the more general case with K chains and general state space $\mathcal{S}$, expressing the implicit constraints on P becomes more tricky.

2.8.3 Discussion on the Cartesian Product Model

The Cartesian Product Model's distinct advantage lies in its single joint transition matrix, allowing us to treat it as a univariate chain with an expanded state space. Consequently, we can apply all the inference methods, for both Frequentist and Bayesian, developed for univariate chains to this model.

However, although the joint transition matrix P under the Cartesian Product Model, as demonstrated in Section 2.8.2, captures all within-chain and inter-chain dependencies, these dependencies are not modelled explicitly. As shown in Equations (2.51) and (2.52), even simple structures such as contemporaneous independence and Granger non-causality in a 3-chain, 2-state setting translate into complex constraints on the transition probabilities p, involving tedious combinations of summations and multiplications. These complex constraints make the inference challenging. In the Frequentist framework, maximizing the constrained likelihood becomes computationally intensive, and in the Bayesian setting, it is difficult to specify compatible priors or efficiently sample from the constrained posterior distribution straightforwardly.

2.9 Models with Linear Decomposition

In this section, we review two widely used models in the literature: the Inter-Chain Transition (ICT) model and the Mixture Transition Distribution (MTD) model. Although these models are constructed differently, they share a common strategy of decomposing joint transition probabilities into linear combinations of lower-dimensional conditional probabilities, which significantly reduces the parameter space compared to the Cartesian Product Model. However, both the ICT and MTD models inherently assume conditional independence, which prevents them from capturing concurrent interactions such as contemporaneous independence. Moreover, their reliance on linear decomposition oversimplifies the underlying dependency structures, especially in representing group-wise influences and non-linear interactions. As a result, these models may overlook intricate relationships across the groups of chains.

2.9.1 Inter-Chain Transition (ICT) Model

Rather than modelling the joint transition matrix P defined by the Cartesian Product Model, an alternative approach proposed by Zhong and Ghosh [2001] involves modelling inter-chain transition matrices. Their method is based on the Coupled Markov Chain framework, which explicitly assumes conditional independence 2.2 among chains.

2.9.1.1 The construction of ICT model

Let $M^{(kk')}$ represent the inter-chain transition matrix capturing transitions from chain k to chain k':

$$M^{(kk')} = \left\{ m_{ab}^{(kk')} \right\}_{a,b \in \mathcal{S}}$$

where each element $m_{ab}^{(kk')}$ denotes an inter-chain transition probability, defined as:

$$m_{ab}^{(kk')} = \mathbb{P}\big(Z_{k'}(t) = b \mid Z_k(t-1) = a\big)$$

In other words, $m_{ab}^{(kk')}$ is the probability of chain k' being in state b at time t, conditioned

on chain k being in state a at the previous time point $t-1$. Under this setting, there will be k^2 such matrix M.

Building upon the conditional independence (Assumption 2.2), Zhong and Ghosh [2001] introduced an additional Assumption 2.21 to decompose the marginal transition probabilities explicitly as linear combinations of inter-chain transition probabilities and coupling coefficients, formalized as follows:

Assumption 2.21. *[Zhong and Ghosh, 2001] The marginal transition probability $\eta_{k'}$ of each chain k' can be split as a linear combination of the effect from each previous state multiplied by some coupling weight $\theta_{kk'}$:*

$$\mathbb{P}\big(Z_{k'}(t)|\boldsymbol{Z}(t-1)\big) = \sum_{k=1}^{K} \theta_{kk'} \mathbb{P}\big(Z_{k'}(t)|Z_k(t-1)\big) \qquad k, k' \in [K]$$

To simplify the notation, we denote the marginal transition probabilities as η and the inter-chain transition probabilities as m, yielding to the concise form:

$$\eta_{k'} = \sum_{k=1}^{K} \theta_{kk'} m^{(kk')} \qquad k, k' \in [K]$$

The weights are normalized such that $\sum_{k=1}^{K} \theta_{kk'} = 1$.

Combining Assumption 2.21 with the conditional independence assumption 2.2, we can explicitly represent the entries of the joint transition matrix p in terms of the coupling weights θ and the inter-chain transition probabilities m as follows:

$$\begin{aligned}
p_{\boldsymbol{ab}} =& \mathbb{P}(\boldsymbol{Z}(t) = \boldsymbol{b} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}) \\
=& \prod_{k'=1}^{K} \eta^{(k')}(b_{k'} \mid \boldsymbol{a}) && \text{(By Assumption 2.2)} \\
=& \prod_{k'=1}^{K} \left(\sum_{k=1}^{K} \theta_{kk'} m^{(kk')}_{a_k b_{k'}} \right) && \text{(By Assumption 2.21)}
\end{aligned}$$

Example 2.22. Consider an MMC with 3 chains, each having 2 possible states. The

ICT Model suggests that this MMC can be parametrised using 9 2×2 marginal transition matrices $M^{(kk')}$ for $k, k' = 1, 2, 3$:

$$M^{(kk')} = \begin{pmatrix} m_{11}^{(kk')} & m_{12}^{(kk')} \\ m_{21}^{(kk')} & m_{22}^{(kk')} \end{pmatrix}$$

and a corresponding 3×3 coupling coefficient matrix Θ:

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \end{pmatrix}$$

subject to the column constraints: $\theta_{1k'} + \theta_{2k'} + \theta_{3k'} = 1$ for $k' = 1, 2, 3$.

For a general MMC consisting of K chains, each with s possible states, the number of independent parameters required to represent the joint transition matrix P, under the ICT Model using coupling coefficients θ and inter-chain transition probabilities m, is:

$$\underbrace{s(s-1)K^2}_{\substack{K^2 \text{ inter-chain} \\ \text{transition matrices}}} + \underbrace{K(K-1)}_{\text{coupling coefficients}} \tag{2.53}$$

- There are K^2 inter-chain transition matrices M, each of dimension $s \times s$, requiring $s(s-1)$ independent parameters per matrix due to row-sum constraints.

- A $K \times K$ coupling coefficient matrix Θ, requiring $K(K-1)$ independent parameters, with each column summing to 1 as specified in Assumption 2.21.

Example 2.23. Continue with Example 2.22, considering an MMC with $K = 3$ chains and state space size $s = 2$, the total number of independent parameters is computed as:

$$2(2-1) \times 3^2 + 3(3-1) = 24$$

including 6 independent coupling coefficients θ and 18 independent inter-chain transition probabilities $m^{(kk')}$, according to Equation (2.53), in comparison with $2^3 \times (2^3 - 1) = 56$ independent parameters in the Cartesian Product Model, showing a significant reduction in the number of independent parameters.

2.9.2 Mixture Transition Distribution (MTD) model

Originally introduced by Raftery [1985], the MTD model is designed for univariate l-order Markov chains. It models the conditional probability of $Z(t)$ given its l previous states as a weighted sum of the individual contributions from each lag:

$$\mathbb{P}\big(Z(t) = a_0 \mid Z(t-1) = a_1, \ldots, Z(t-l) = a_l\big) = \sum_{i=1}^{l} \lambda_i q_{a_0 a_i},$$

where λ_i are non-negative weights satisfying $\sum_{i=1}^{l} \lambda_i = 1$, and $Q = \{q_{ab}\}_{a,b \in \mathcal{S}}$ is a non-negative matrix with each column summing to 1, ensuring:

$$0 \leq \sum_{i=1}^{l} \lambda_i q_{a_0 a_i} \leq 1.$$

Ching and Ng [2006], Ching et al. [2004] then extended this model to incorporate with MMC. In the first-order setting, the MTD model can be constructed as follows:

- Let $Y_k(t)$ denote the state vector of the k-th sequence at time t. If sequence k is in state a, then:

  $$Y_k(t) = (0, \ldots, 0, \underbrace{1}_{a\text{-th entry}}, 0, \ldots, 0)^\top.$$

  Note that $Y_k(t)$ denotes the same hidden-chain state previously written as $Z_k(t)$. We switch to the notation $Y_k(t)$ here because it uses the alternative one-hot representation.

- Each $Y_k(t)$ is modelled as a convex combination of the individual contribution of

the influences from all chains at time t:

$$Y_k(t+1) = \sum_{k'=1}^{K} \lambda_{k'k}\, P^{(k'k)}\, Y_{k'}(t) \qquad k, k' \in [K] \tag{2.54}$$

subject to constraints:

$$\lambda_{k'k} \geq 0 \quad \text{and} \quad \sum_{k'=1}^{K} \lambda_{k'k} = 1 \qquad \forall k', k \in [K] \tag{2.55}$$

where $P^{(k'k)}$ is the $s \times s$ transition probability matrix from the state of chain k' to chain k, and $\lambda_{k'k}$ is the associated weight representing the influence of chain k' on chain k.

Explicitly, Equation (2.54) can be expanded as the following Equation (2.56):

$$\boldsymbol{Y}(t) \equiv \begin{pmatrix} Y_1(t) \\ Y_2(t) \\ \vdots \\ Y_K(t) \end{pmatrix} = \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}P^{(12)} & \cdots & \lambda_{1K}P^{(1K)} \\ \lambda_{21}P^{(21)} & \lambda_{22}P^{(22)} & \cdots & \lambda_{2K}P^{(2K)} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{K1}P^{(K1)} & \lambda_{K2}P^{(K2)} & \cdots & \lambda_{KK}P^{(KK)} \end{pmatrix} \begin{pmatrix} Y_1(t-1) \\ Y_2(t-1) \\ \vdots \\ Y_K(t-1) \end{pmatrix} \equiv Q\boldsymbol{Y}(t-1) \tag{2.56}$$

Equation (2.54) shows that the MTD model resembles the ICT model from Section 2.9.1, as both express transition dynamics as a weighted sum of contributions from previous states. In both cases, the influence from chain k' to k is captured by the linear combination of an inter-chain transition matrix $P^{(k'k)}$ and a corresponding weight $\lambda_{k'k}$. However, rather than modelling joint states in the Cartesian product space, the ICT model operates on state vectors, enabling a more compact matrix formulation.

From Equation (2.56), it is evident that the structured transition matrix Q in the MTD model requires the same number of parameters as the ICT model, both involving inter-chain transition matrices $P^{(k'k)}$ and weights $\lambda_{k'k}$. However, despite using the same number

of parameters, the structured transition matrix Q in the MTD model has a significantly lower dimension, $sK \times sK$, compared to the joint transition matrix P in the ICT model, offering a more compact and convenient representation.

It is also worth noting that, although not explicitly stated in Ching and Ng [2006], Ching et al. [2004], the conditional independence is inherently embedded in the MTD model. This is evident in the matrix form (2.56), where each current state vector $Y_k(t)$, when conditioned on the entire previous state vector $\boldsymbol{Y}(t-1)$, are mutually independent, as they are determined solely by the k-th row of the structured transition matrix Q.

Example 2.24. Consider an MMC with 3 chains, each taking values in a binary state space $\mathcal{S} = \{1, 2\}$. The overall state vector $\boldsymbol{Y}(t)$ has length $3 \times 2 = 6$. For example, if the system is in state $(1, 1, 1)$ at time t, then:

$$\boldsymbol{Y}(t) \equiv \big(Y_1(t), Y_2(t), Y_3(t)\big)^\top = \big(\underbrace{1\ ,\ 0}_{\text{chain 1}}\ ,\ \underbrace{1\ ,\ 0}_{\text{chain 2}}\ ,\ \underbrace{1\ ,\ 0}_{\text{chain 3}}\big)^\top$$

Under the MTD model, the transition of the full system can be expressed using a structured transition matrix Q of dimension 6×6 as:

$$\begin{pmatrix} Y_1(t) \\ Y_2(t) \\ Y_3(t) \end{pmatrix} = \begin{pmatrix} \lambda_{11}P^{(11)} & \lambda_{12}P^{(12)} & \lambda_{13}P^{(13)} \\ \lambda_{21}P^{(21)} & \lambda_{22}P^{(22)} & \lambda_{23}P^{(23)} \\ \lambda_{31}P^{(31)} & \lambda_{32}P^{(32)} & \lambda_{33}P^{(33)} \end{pmatrix} \begin{pmatrix} Y_1(t-1) \\ Y_2(t-1) \\ Y_3(t-1) \end{pmatrix}$$

where each $P^{(k'k)}$ is a 2×2 transition matrix and the weights $\lambda_{k'k}$ satisfy the column normalization condition: $\lambda_{1k} + \lambda_{2k} + \lambda_{3k} = 1$ for each $k = 1, 2, 3$.

However, the constraints on the coefficients $\lambda_{kk'}$ in the MTD model, as specified in Equation (2.55), complicate inference. To address this, Nicolau [2014] introduced the MTD-Probit model, which embeds the linear combination of inter-chain transition matrices $P^{(k'k)}$ and coefficients $\lambda_{k'k}$ into the cumulative distribution function of a standard normal

distribution, and is formalized as:

$$\mathbb{P}\big(Z_k(t) = b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) = \frac{\Phi\left(\lambda^\star_{0k} + \sum_{k'=1}^{K} \lambda^\star_{k'k}\, p^{(k'k)}_{a_{k'}b_k}\right)}{\sum_{b'_k=1}^{s} \Phi\left(\lambda^\star_{0k} + \sum_{k'=1}^{K} \lambda^\star_{k'k}\, p^{(k'k)}_{a_{k'}b'_k}\right)} \tag{2.57}$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal.

Equation (2.57) shows that the linear structure of the original MTD model is preserved within a Probit link, with an additional intercept term $\lambda^\star_{0k}$. This formulation allows the coefficients $\lambda^\star$ to be unconstrained in $\mathbb{R}$, as the normalization condition $\sum_{b_k} \mathbb{P}(Z_k(t) = b_k \mid \boldsymbol{Z}(t-1)) = 1$ is automatically satisfied. Additionally, it involves one additional intercept parameter $\lambda^\star_{0k}$ in comparison with the MTD case, to improve the fit.

2.9.3 Discussion on the Models with Linear Decomposition

Despite introducing additional parameters through coupling coefficients θ and λ in ICT and MTD models respectively, these models effectively reduces complexity and enhances interpretability via Assumption 2.21:

- Parametrisation with coupling coefficients and inter-chain transition probabilities m reduces the parameter space significantly from $O(s^{2K})$ to $O(s^2K^2)$, addressing the exponential growth in parameters.

- Each coefficient $\theta_{kk'}$, $\lambda_{kk'}$ provides an intuitive interpretation as the marginal influence that chain k exerts on chain k'. Typically, the matrix for these coefficients is set to be positive definite, implying that each chain's current state depends predominantly on its own previous state rather than states from other chains.

However, both model's reliance on the linear decomposition provided by Assumption 2.21 and conditional independence (Assumption 2.2) significantly limits its flexibility and ability to capture more complex interactions. Specifically:

- The linear decomposition of marginal transition probabilities restricts the model from capturing nonlinear or comprehensive dependency structures.
- Conditional independence inherently constrains the ICT and the MTD model, preventing them from modelling contemporaneous independence.
- For Granger non-causality, both models can only handle scenarios involving single chains rather than general subgroups of chains.
- The introduction of additional parameters, particularly these coefficients, can lead to non-identifiability, as multiple combinations of coupling coefficients and inter-chain transition probabilities may correspond to the same joint transition matrix under row and column restrictions.

2.10 Discussion

In this chapter, we aim to bridge the gap between the dependency structures of MMCs and their commonly used representation via transition matrices. While the transition matrices efficiently encode the joint dynamics of the MMC, they do not explicitly reveal the underlying dependency relationships, such as how the current states relate to the past and the concurrent states across chains. Specifically, we examine how different types of dependency structures impose constraints on the transition matrix.

To systematically explore these relationships, we first categorize the dependency structures in MMCs into two types: concurrent interactions and influence from the past. We then examined three key assumptions: *Conditional Independence*, *Contemporaneous Independence*, and *Granger non-Causality*. Conditional independence represents a special, highly restrictive case in which all concurrent interactions are absent. Contemporaneous independence generalises this by allowing the partial absence of the correlations among concurrent states, thereby capturing the full spectrum of concurrent interactions. Granger non-causality, on the other hand, characterizes the influence from past states, asserting that the future states are conditionally independent of certain subsets of past states

given the rest. The combination of the contemporaneous independence assumption and the Granger non-causality collectively represent and govern all the dependency structure within a first-order MMCs, as demonstrated in Theorem 2.18. The following Figure 2.10 also illustrates this relationship.

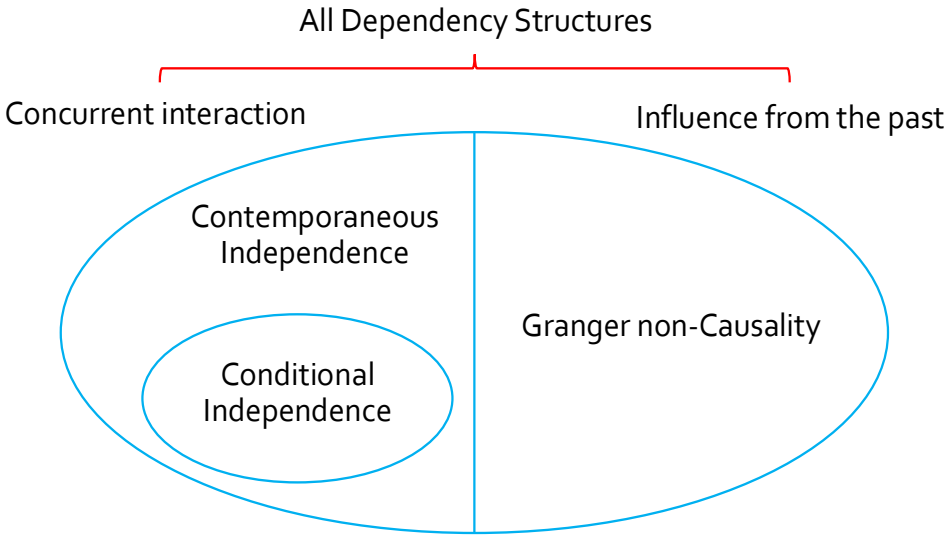


Figure 2.10: The Venn diagram of all the dependency structures in a first-order MMC, partitioned into concurrent interaction and influence from the past. Concurrent interactions are fully characterized by the *contemporaneous independence* assumption, with *conditional independence* being a special case. Meanwhile, influences from past states are fully described by the *Granger non-causality* assumption. Together, contemporaneous independence and Granger non-causality comprehensively capture the entire dependency structure of an MMC, as stated in Theorem 2.18.

It is also intuitive to interpret how each type of independence assumption influences the structure of the joint transition matrix P. Both conditional independence and contemporaneous independence refer to the relationships among states at the same time, and thus impose constraints within each row of P, since rows correspond to different concurrent realizations. In contrast, Granger non-causality involves dependencies from the past to the current states, and therefore imposes constraints across both rows and columns of P.

Then, to make these abstract structural assumptions more concrete, we explore their geometric interpretations in terms of how they constrain the joint transition matrix. For conditional independence, we illustrate the structure using the special case of two chains

with two possible states. As shown in Figure 2.5, the constrained parameter space under conditional independence forms a two-dimensional manifold $\mathcal{M}$ embedded within the full parameter space, which is the three-dimensional simplex $\boldsymbol{\Delta}^3$. For contemporaneous independence, we demonstrate that after applying a projection, based on coordinate definitions derived from the marginalization process in Equation 2.33, the resulting geometric representation matches that of conditional independence. Lastly, for Granger non-causality, the geometric structure is more intricate. For each group of rows $\mathcal{A}$, the constrained space can be viewed as a product of several lower-dimensional simplexes, hierarchically organized as described in Equation 2.50, and embedded within the full parameter space, which is a Cartesian product of simplexes representing the unconstrained rows.

After that, motivated by the need to capture interactions among subsets of chains, we introduce the concept of *marginal transition probabilities* η, which represent the probability that a subgroup of chains g_j transitions to a specific state, conditioned on the full system state $\boldsymbol{Z}$ at the previous time step. Importantly, key dependency structures, including conditional independence, contemporaneous independence, and Granger non-causality, can all be reformulated succinctly using these marginals, as shown in Equations (2.6), (2.24), and (2.38), respectively. This marginal-based formulation offers a more intuitive and tractable way to express dependency assumptions, in contrast to working directly with the joint transition probabilities p, which often requires tedious operation of summations and multiplications. As a result, modelling dependency structures using η can be significantly more efficient and interpretable. In the next chapter, we build upon this insight and propose two novel reparametrisations of the joint transition probabilities founded on these marginal transitions to more effectively represent and encode dependencies in MMCs.

With these dependency structures in place, we reviewed two classes of popular models in the literature: the Cartesian Product Model and models with linear decompositions, such as the ICT and MTD models. The Cartesian Product Model parametrised the MMC

with a single joint transition matrix, which enabled it to capture all possible dependency structures, including contemporaneous independence and Granger non-causality. However, these dependencies were not modelled explicitly. Even simple structures translated into complicated constraints on transition probabilities, making inference challenging in both Frequentist and Bayesian frameworks. In contrast, the ICT and MTD models relied on linear decompositions of marginal transition probabilities, which limited their ability to capture non-linear or richer dependencies. Their assumption of conditional independence further prevented them from representing contemporaneous independence. Moreover, the additional parameters, particularly coupling coefficients, could lead to non-identifiability, since multiple parameter settings yielded the same joint transition matrix under row and column restrictions.

These drawbacks of the Cartesian Product Model and the linear decomposition based models motivated us to seek more effective parametrisations of MMCs that could fully capture all first-order dependencies. As noted previously, the marginal transition probabilities not only served as the bridge linking joint transition probabilities to dependency structures, but also appeared frequently in geometric representations of these structures. This naturally suggested modelling dependency structures directly through the marginal transition probabilities. Building on this idea, we introduced the Univariate Marginal Model and the Hierarchical Marginal Model, which will be developed in detail in Chapters 3 and 4, respectively.

Chapter 3

Univariate Marginal Model

3.1 Introduction

3.1.1 Motivation and Aim of the Chapter

Recall that in previous Chapter 2, we explored a variety of dependency structures in multivariate Markov chains (MMCs), including concurrent interactions, (e.g. *Conditional Independence* and *Contemporaneous Independence*), as well as temporal influences from the past (i.e. *Granger non-Causality*). A key result, as presented in Theorem 2.18, is that any dependency structure in a first-order MMC can be expressed as a combination of contemporaneous independence and Granger non-causality. This foundational result motivates modelling MMCs through frameworks that explicitly capture these two core types of dependency, contemporaneous independence and Granger non-causality, thereby bringing enhanced interpretability and supporting the structure-driven analysis.

One widely studied approach to modelling multivariate dependence is provided by copula models, which separate marginal behaviour from the dependence structure of a joint distribution. Following the seminal result of Sklar [1959], copulas have been extensively developed and applied in statistics, econometrics, and finance to flexibly capture non-linear and tail dependence while allowing arbitrary marginal distributions [Joe, 2014]. Extensions such as vine copulas further enable high-dimensional constructions through pairwise

dependence decompositions [Aas et al., 2009, Bedford and Cooke, 2002]. However, in discrete multivariate settings, copulas are non-unique and make dependency structures difficult to identify [Genest and Nešlehová, 2007], resulting in limited interpretability of both contemporaneous independence and Granger non-causality when compared with transition-probability–based representations.

As reviewed in the previous section, existing transition-probability–based parametrisations of MMCs in the literature face important limitations. *Cartesian Product Model* [see Brand et al., 1997, Ghosh et al., 2017, Pohle et al., 2021, and others] represents dependency structures only implicitly through tedious summations and productions, while linear decomposition–based approaches, including *Inter-Chain Transition (ICT) Model* [Zhong and Ghosh, 2001] and the *Mixture Transition Distribution (MTD) Model*, suffer from non-identifiability due to the introduction of coupling coefficients, and the reliance on the conditional independence that inherently prevents these models from capturing concurrent interactions.

Therefore, there is a need to develop efficient models for MMCs that can explicitly represent the dependency structures discussed in Chapter 2. In this Chapter 3, we introduce a novel class of model based on the marginal transition probabilities, namely the *Univariate Marginal Model*, which factorizes the joint transition probabilities p into a product of the marginal transition probabilities η and auxiliary probabilities ζ that conditioned on both past and contemporaneous states, according to a predefined permutation σ. In this case, the joint transition matrix P is fully characterized, ensuring that no information related to the transition is lost and enabling more interpretable tracking of the underlying dependency structures. The *Univariate Marginal Model* is then paired with a multinomial logistic regression that further decompose these marginal and auxiliary probabilities into unrestricted and interpretable coefficients, offering a clear representation of the underlying dependency structures (see Figure 3.1).

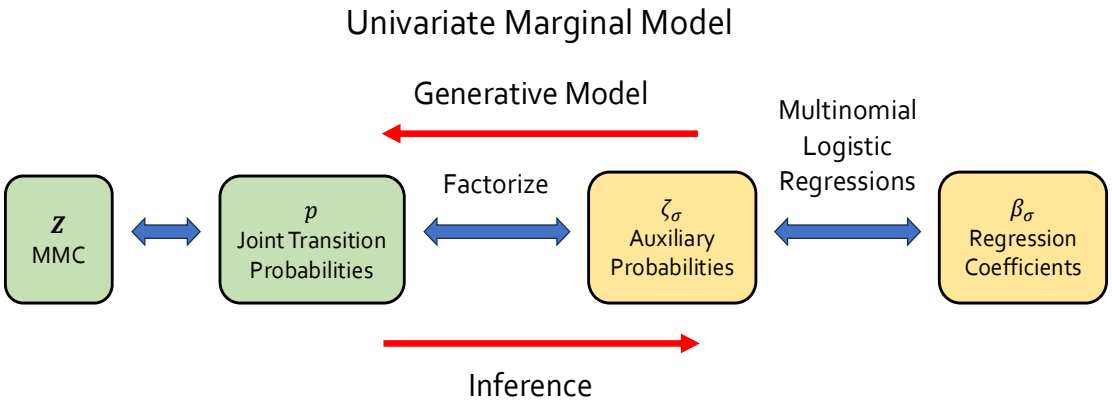


Figure 3.1: An illustration of the reparametrisation procedure of the *Univariate Marginal Model* is shown. It starts with factorizing the joint transition probabilities p into auxiliary probabilities ζ_σ according to the predefined permutation, followed by further decompose these ζ_σ into the regression coefficients $\boldsymbol{\beta}_\sigma$ through a multinomial logistic regression. In the generative direction (the top arrow pointing left), the process is initiated with the regression coefficients $\boldsymbol{\beta}_\sigma$ and generates an MMC that exhibits the desired dependency structures. In contrast, the inference direction (the bottom arrow pointing right) starts with observed data from an MMC and infers the corresponding regression coefficients that encode the underlying dependency structures within the MMC.

This chapter goes one step further by exploring the utilizations of the *Univariate Marginal Model* from both generative and inference perspectives:

- **Generative Model**: The generative model simulates an MMC with specified dependency structures, starting from the regression coefficients $\boldsymbol{\beta}_\sigma$ and proceeding, as shown in Figure 3.1, from coefficients on the right to MMC data on the left.
- **Inference**: The inference starts from the observed MMC data $\boldsymbol{Z}$ and proceeds in the reverse direction, as shown in Figure 3.1, aiming to estimate the regression coefficients $\boldsymbol{\beta}_\sigma$ under permutation σ within the Univariate Marginal Models.

3.1.2 Chapter Layout

Noting the shortcomings of existing models in effectively encoding dependency structures, we propose a novel *Univariate Marginal Model* in Section 3.2. This approach factorizes the joint transition probabilities into auxiliary probabilities, in a similar fashion as Bayesian networks, according to a predefined permutation. We further elaborate on the geometric

relationship between these auxiliary probabilities and the original joint transition probabilities concerning different permutations in Section 3.2.1. The encoding capability of auxiliary probabilities for dependency structures is thoroughly discussed in Section 3.2.2.

To achieve an unconstrained parameter space and enhance interpretability, we utilize a multinomial logistic regression reviewed in Section A.4 to further decompose these auxiliary probabilities in Section 3.3.1, followed by demonstrating how the regression coefficients encode various dependency structures in Section 3.3.2. Building on this encoding, Section 3.3.3 explains how to select an appropriate permutation to target specific dependence structures. Several illustrative examples are provided in Section 3.3.4.

In utilizing the Univariate Marginal Model, Section 3.4 considers a generative perspective, showing how an MMC can be simulated under the desired dependency structures from the regression coefficients, with the procedure summarized in Algorithm 1. As reviewed in Section A, the majority of inference approaches estimate the regression coefficients $\boldsymbol{\beta}_\sigma$ either indirectly by first inferring the joint transition probabilities p through Frequentist maximum likelihood estimation (Section A.1) or Bayesian conjugate Dirichlet priors (Section A.2), or directly by applying Pólya–Gamma data augmentation (Section A.3).

With these methods in place, Section 3.5 presents a simulation study, with details of the simulation design in Section 3.5.1. Simulation results are then presented in Section 3.5.2 under four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures. The aim is to evaluate the ability of both models to estimate regression coefficients accurately and to recover the embedded dependency structures.

Finally, Section 3.6 evaluates the strengths and limitations of the proposed Univariate Marginal Models.

3.2 Formulation of the Univariate Marginal Model

Consider an MMC $\boldsymbol{Z}$ consisting of K chains, partitioned into J non-empty subgroups $g_1, \ldots, g_J \subseteq [K]$. Let $[J] = \{1, 2, \ldots, J\}$ denote the index set of these chain groups. We define a permutation σ over the set $[J]$ such that:

$$\sigma : [J] \mapsto [J]$$

where $\sigma(j)$ denotes the j-th term of the set $[J]$ under permutation σ.

Example 3.1. Consider an MMC with three chain groups indexed by $[J] = \{1, 2, 3\}$. One possible permutation σ on $[J]$ is $\{3, 1, 2\}$. Under this permutation:

$$\sigma(1) = 3 \quad \sigma(2) = 1 \quad \sigma(3) = 2$$

$\sigma(1)$ refers to the first element of the set $[J]$ under permutation σ, which is 3. $\sigma(2)$ and $\sigma(3)$ refers to the second and the third being 1 and 2 respectively.

The following Definition 3.2 introduces the auxiliary probabilities ζ_σ, defined under a permutation σ of the chain groups. These probabilities serve as a bridge connecting the joint transition probabilities p with the marginal transition probabilities η.

Definition 3.2. *Given a permutation σ of $[J]$, the auxiliary probabilities $\zeta_{\sigma(j)}$ for $j = 2, \ldots, J$ are of the form:*

$$\begin{aligned} \zeta_{\sigma(j)} &= \mathbb{P}\left(\boldsymbol{Z}_{\sigma(j)}(t) \mid \boldsymbol{Z}_{\sigma(1)}(t), \ldots, \boldsymbol{Z}_{\sigma(j-1)}(t),\ \boldsymbol{Z}_{\sigma(1)}(t-1), \ldots, \boldsymbol{Z}_{\sigma(J)}(t-1)\right) \\ &= \mathbb{P}\left(\boldsymbol{Z}_{\sigma(j)}(t) \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t),\ \boldsymbol{Z}(t-1)\right) \end{aligned} \tag{3.1}$$

where $\boldsymbol{Z}_{\sigma(1:(j-1))}(t)$ stands for $\left(\boldsymbol{Z}_{\sigma(1)}(t), \ldots, \boldsymbol{Z}_{\sigma(j-1)}(t)\right)$.

Specially for $j = 1$:

$$\zeta_{\sigma(1)} = \mathbb{P}\left(\boldsymbol{Z}_{\sigma(1)}(t) \mid \boldsymbol{Z}(t-1)\right) = \eta_{\sigma(1)} \tag{3.2}$$

Note that the auxiliary probabilities $\zeta_{\sigma(j)}$ are different from the marginal probability $\eta_{\sigma(j)}$ since it also involves the contemporaneous terms $\boldsymbol{Z}_{\sigma(1:(j-1))}(t)$ in the conditioning part. It follows immediately from Equation (3.2) that $\eta_{\sigma(1)} \equiv \zeta_{\sigma(1)}$ as there will be no additional $\boldsymbol{Z}(t)$ terms in the conditioning part when we set $j = 1$.

Example 3.3. Continue with Example 3.1 with three chain groups $[J] = \{1, 2, 3\}$ and the same permutation $\sigma = \{3, 1, 2\}$, we define $\zeta_{\sigma(1)}$, $\zeta_{\sigma(2)}$, and $\zeta_{\sigma(3)}$ as follows:

$$\begin{aligned}
\zeta_{\sigma(1)} &= \mathbb{P}(\boldsymbol{Z}_{\sigma(1)}(t) \mid \boldsymbol{Z}(t-1)) = \mathbb{P}(\boldsymbol{Z}_3(t) \mid \boldsymbol{Z}(t-1)) \\
\zeta_{\sigma(2)} &= \mathbb{P}\left(\boldsymbol{Z}_{\sigma(2)}(t) \mid \boldsymbol{Z}_{\sigma(1)}(t), \boldsymbol{Z}(t-1)\right) = \mathbb{P}\left(\boldsymbol{Z}_1(t) \mid \boldsymbol{Z}_3(t), \boldsymbol{Z}(t-1)\right) \\
\zeta_{\sigma(3)} &= \mathbb{P}\left(\boldsymbol{Z}_{\sigma(3)}(t) \mid \boldsymbol{Z}_{\sigma(1)}(t), \boldsymbol{Z}_{\sigma(2)}(t), \boldsymbol{Z}(t-1)\right) = \mathbb{P}\left(\boldsymbol{Z}_2(t) \mid \boldsymbol{Z}_3(t), \boldsymbol{Z}_1(t), \boldsymbol{Z}(t-1)\right)
\end{aligned}$$

More precisely, the auxiliary probabilities corresponding to the specific realization of states $\boldsymbol{a}$, $\boldsymbol{b}$, and $\boldsymbol{c}$ can be defined as follows:

$$\zeta_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b},\ \boldsymbol{a}) = \mathbb{P}\left(\boldsymbol{Z}_{\sigma(j)}(t) = \boldsymbol{c} \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t) = \boldsymbol{b},\ \boldsymbol{Z}(t-1) = \boldsymbol{a}\right) \tag{3.3}$$

where $\boldsymbol{c} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{c}} = \boldsymbol{\mathcal{S}}_{\sigma(j)}$ denotes the current-time realization of the states $\boldsymbol{Z}_{\sigma(j)}$ in chain group g_j and $\boldsymbol{b} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{b}} = \boldsymbol{\mathcal{S}}_{\sigma(1:(j-1))} = \boldsymbol{\mathcal{S}}_{\sigma(1)} \times \cdots \times \boldsymbol{\mathcal{S}}_{\sigma(j-1)}$ denotes the current-time realization of the states $\boldsymbol{Z}_{\sigma(1:(j-1))}$ for the joint of chain groups $g_1, g_2, \ldots g_{j-1}$. Finally, $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}} = \boldsymbol{\mathcal{S}}$ represents the realized states of all the chain groups at previous time $t-1$.

The following Definition 3.4 provides the explicit vector and matrix forms of the auxiliary probabilities $\zeta_{\sigma(j)}$, which will later be used to express their transformation with the joint transition probabilities p in matrix form.

Definition 3.4. *The vector $\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot, \boldsymbol{a}^{(i)})$ for $i = 1, 2, \ldots, s^K$ and $\boldsymbol{a}^{(i)} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$ are defined as a column vector of the collection of the auxiliary probabilities in the form:*

$$\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot, \boldsymbol{a}) = \begin{pmatrix} \mathbb{P}(\boldsymbol{Z}_{\sigma(j)}(t) = \boldsymbol{c}^{(1)} \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t) = \boldsymbol{b}^{(1)},\ \boldsymbol{Z}(t-1) = \boldsymbol{a}^{(i)}) \\ \mathbb{P}(\underbrace{\boldsymbol{Z}_{\sigma(j)}(t) = \boldsymbol{c}^{(2)} \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t) = \boldsymbol{b}^{(1)}}_{\text{iterate over } \boldsymbol{b}\in\boldsymbol{\mathcal{S}}_{\boldsymbol{b}},\ \boldsymbol{c}\in\boldsymbol{\mathcal{S}}_{\boldsymbol{c}}},\ \underbrace{\boldsymbol{Z}(t-1) = \boldsymbol{a}^{(i)}}_{\text{fixed}}) \\ \vdots \end{pmatrix}$$

$\forall \boldsymbol{b} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{b}}$ *and* $\forall \boldsymbol{c} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{c}}$.

The auxiliary probability matrix $\boldsymbol{\zeta}_{\sigma(j)}$ *collects every auxiliary probability* $\zeta_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b},\ \boldsymbol{a})$ *for all possible states* $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$. *It is constructed by stacking the column vectors* $\boldsymbol{\zeta}_{\sigma(j')}(\cdot \mid \cdot, \boldsymbol{a}^{(i)})$ *for all* $\boldsymbol{a}^{(i)} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$, *formalized as:*

$$\boldsymbol{\zeta}_{\sigma(j)} = \begin{pmatrix} \longleftarrow & \left(\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot,\ \boldsymbol{a}^{(1)})\right)^{\top} & \longrightarrow \\ \longleftarrow & \left(\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot,\ \boldsymbol{a}^{(2)})\right)^{\top} & \longrightarrow \\ & \vdots & \\ \longleftarrow & \left(\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot,\ \boldsymbol{a}^{(s^K)})\right)^{\top} & \longrightarrow \end{pmatrix} \tag{3.4}$$

The number of the probabilities involved in this column vector $\boldsymbol{\zeta}_{\sigma(j)}(\cdot \mid \cdot, \boldsymbol{a}^{(i)})$ is equal to the cardinality of $\boldsymbol{\mathcal{S}}_{\boldsymbol{b}} \times \boldsymbol{\mathcal{S}}_{\boldsymbol{c}} = \boldsymbol{\mathcal{S}}_{\sigma(1:j)}$, which is equal to $s^{\sum |g_{\sigma(\cdot)}|}$ with $|g_{\sigma(j)}|$ refers to the number of chains within subgroup $g_{\sigma(j)}$ and the summation goes from 1 to j. It follows that the matrix $\boldsymbol{\zeta}_{\sigma(j)}$ is of dimension $s^K \times s^{\sum |g_{\sigma(\cdot)}|}$ since it has $|\boldsymbol{\mathcal{S}}| = s^K$ rows for all the different $\boldsymbol{a}^{(i)} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$.

Example 3.5. Continuing with the three chain groups $[J] = \{1, 2, 3\}$ and the permutation $\sigma = \{3, 1, 2\}$ from Example 3.1, assume that each chain group contains exactly one chain and these three chains share the same state space $\mathcal{S} = \{1, 2\}$. As an example, we construct the vector $\boldsymbol{\zeta}_{\sigma(2)}(\cdot \mid \cdot, (1, 1, 1))$ as follows:

$$\boldsymbol{\zeta}_{\sigma(2)}\left(\cdot \mid \cdot, (1,1,1)\right) = \begin{pmatrix} \mathbb{P}\big(\boldsymbol{Z}_1(t) = 1 \mid \boldsymbol{Z}_3(t) = 1, \boldsymbol{Z}(t-1) = (1,1,1)\big) \\ \mathbb{P}\big(\boldsymbol{Z}_1(t) = 2 \mid \boldsymbol{Z}_3(t) = 1, \boldsymbol{Z}(t-1) = (1,1,1)\big) \\ \mathbb{P}\big(\boldsymbol{Z}_1(t) = 1 \mid \boldsymbol{Z}_3(t) = 2, \boldsymbol{Z}(t-1) = (1,1,1)\big) \\ \mathbb{P}\big(\boldsymbol{Z}_1(t) = 2 \mid \boldsymbol{Z}_3(t) = 2, \boldsymbol{Z}(t-1) = (1,1,1)\big) \end{pmatrix}$$

and the matrix $\boldsymbol{\zeta}_{\sigma(2)}$ by stacking the vectors:

$$\boldsymbol{\zeta}_{\sigma(2)} = \begin{pmatrix} \longleftarrow & \big(\boldsymbol{\zeta}_{\sigma(2)}(\cdot \mid \cdot, (1,1,1))\big)^\top & \longrightarrow \\ \longleftarrow & \big(\boldsymbol{\zeta}_{\sigma(2)}(\cdot \mid \cdot, (1,1,2))\big)^\top & \longrightarrow \\ & \vdots & \\ \longleftarrow & \big(\boldsymbol{\zeta}_{\sigma(2)}(\cdot \mid \cdot, (2,2,2))\big)^\top & \longrightarrow \end{pmatrix}$$

It is straightforward to see that matrix $\boldsymbol{\zeta}_{\sigma(2)}$ is of dimension $s^k \times s^{\sum |g_{\sigma(\cdot)}|} = 8 \times 4$

Finally, the Univariate Marginal Model suggests that instead of modelling the dynamics of an MMC using the joint transition matrix P, one can alternatively use the set of auxiliary probability matrices $\boldsymbol{\zeta}_\sigma = \big\{\boldsymbol{\zeta}_{\sigma(1)}, \boldsymbol{\zeta}_{\sigma(2)}, \ldots, \boldsymbol{\zeta}_{\sigma(J)}\big\}$.

The following Proposition 3.6 establishes the equivalence between modelling an MMC using the joint transition probabilities and using the auxiliary probabilities proposed in the Univariate Marginal Model.

Proposition 3.6. *Any first-order time-homogeneous Multivariate Markov chain to be modelled with a joint transition matrix P can be equivalently modelled with a set of auxiliary probability matrices $\boldsymbol{\zeta}_\sigma = \big(\boldsymbol{\zeta}_{\sigma(1)}, \ldots, \boldsymbol{\zeta}_{\sigma(J)}\big)$ for an arbitrary permutation σ of $[J]$.*

Proof.

$$\begin{aligned} P =& \mathbb{P}\left(\boldsymbol{Z}_1(t), \boldsymbol{Z}_2(t), \ldots, \boldsymbol{Z}_J(t) \mid \boldsymbol{Z}_1(t-1), \boldsymbol{Z}_2(t-1), \ldots, \boldsymbol{Z}_J(t-1)\right) \\ =& \mathbb{P}\left(\boldsymbol{Z}_{\sigma(1)}(t) \mid \boldsymbol{Z}(t-1)\right) \prod_{j=2}^{J} \mathbb{P}\left(\boldsymbol{Z}_{\sigma(j)}(t) \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t), \boldsymbol{Z}(t-1)\right) \\ =& \eta_{\sigma(1)} \prod_{j=2}^{J} \zeta_{\sigma(j)} \end{aligned} \tag{3.5}$$

□

Alternatively, we can write Equation (3.5) in the matrix form as:

$$P = \bigodot_{j=1}^{J} \left[\boldsymbol{\zeta}_{\sigma(j)} \bigotimes \mathbf{1}_{\sigma(j)}^{\top}\right] \tag{3.6}$$

P is the $s^K \times s^K$ joint transition matrix. $\bigodot$ and $\bigotimes$ stand for Hadamard and Kronecker products, respectively. $\mathbf{1}_{\sigma(j)}$ is a column vector of 1 with length:

$$\dim(\mathbf{1}_{\sigma(j)}) = s^{K-\sum |g_{\sigma(j)}|}$$

where $|g_{\sigma(j)}|$ refers to the number of chains within subgroup $g_{\sigma(j)}$ and the summation runs from $\sigma(1)$ to $\sigma(j)$.

Equation (3.6) is constructed through the following two steps:

1. For each $j \in 1, \ldots, J$, we apply the Kronecker product to the matrices $\boldsymbol{\zeta}_{\sigma(j)}$ defined in Equation (3.4), to expand them into a matrix of dimension $s^K \times s^K$, ensuring the dimension matches across all terms.

2. We then take the Hadamard product of these expanded matrices across $j = 1, \ldots, J$, which element-wise multiplies each of the auxiliary probabilities ζ_σ under the permutation σ, thereby reconstructing the full joint transition matrix P.

Remark 3.7 (Hadamard and Kronecker Products)**.** Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two matrices of the same dimensions. Their Hadamard product, denoted by $\boldsymbol{A} \bigodot \boldsymbol{B}$, is the element-wise

product defined as:

$$\left(\boldsymbol{A} \bigodot \boldsymbol{B}\right)_{ij} = a_{ij} b_{ij}$$

The Kronecker product refers to the block-wise product. For matrices $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{p \times q}$, the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B}$ is defined as a block matrix of size $(mp \times nq)$ where each entry a_{ij} is replaced by the block $a_{ij}\boldsymbol{B}$. Concretely,

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\boldsymbol{B} & \cdots & a_{mn}\boldsymbol{B} \end{pmatrix}$$

Lemma 3.8. *There are $J!$ possible ways to factorize the joint transition probability p in terms of Equation (3.5) by alternating the order of conditioning according to the permutation σ. But all of them end up with the same joint transition probability p.*

Proof. Consider an alternative permutation $\tilde{\sigma}$ and factorize the joint transition probability p into the product of marginal transition probability $\eta_{\tilde{\sigma}(1)}$ and auxiliary probabilities $\zeta_{\tilde{\sigma}(j)}$ for $j = 2, \ldots, J$:

$$\prod_{j=1}^{J} \zeta_{\tilde{\sigma}(j)} = \mathbb{P}(\boldsymbol{Z}_{\tilde{\sigma}(1)}(t) \mid \boldsymbol{Z}(t-1)) \prod_{j=2}^{J} \mathbb{P}(\boldsymbol{Z}_{\tilde{\sigma}(j)}(t) \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t),\ \boldsymbol{Z}(t-1)) \tag{3.7}$$

By the formula of the conditional probability, the right hand side of Equation (3.7) is exactly the joint transition probability p. Combining Equation (3.5) and (3.7) gives

$$\prod_{j=1}^{J} \zeta_{\sigma(j)} = p = \prod_{j=1}^{J} \zeta_{\tilde{\sigma}(j)}$$

which demonstrates that two different sets of η and ζ under different permutations σ and $\tilde{\sigma}$ gives the same joint transition probability p. □

Proposition 3.9. *Let $\mathcal{H}_\sigma$ and $\mathcal{P}$ denote the parameter spaces for the auxiliary probability matrices $\boldsymbol{\zeta}_\sigma = \left(\boldsymbol{\zeta}_{\sigma(1)}, \boldsymbol{\zeta}_{\sigma(2)}, \ldots, \boldsymbol{\zeta}_{\sigma(J)}\right)$ and the joint transition matrix P, respectively.*

Then, under a given permutation σ, the map $\mathcal{H}_\sigma \to \mathcal{P}$ from the auxiliary probability matrices $\boldsymbol{\zeta}_\sigma$ to the joint transition matrix P, as well as its inverse map $\mathcal{P} \to \mathcal{H}_\sigma$, is a diffeomorphism.

Proof. Without loss of generality, consider the given permutation σ to be the identity permutation $\sigma^\star$ of $[J]$ with $\sigma(j) = j$ for each $j \in [J]$. For notational simplicity, we omit the subscript $\sigma^\star$ for the collection of auxiliary probabilities under the identity permutation, that is $\boldsymbol{\zeta} \equiv \boldsymbol{\zeta}_{\sigma^\star} = (\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_J)$.

- Injectivity means that given two identical joint transition probabilities $p = \tilde{p}$, then two sets of marginal probabilities and auxiliary probabilities that maps to p and $\tilde{p}$ through Equation (3.5) respectively must be identical.

  Consider the contrapositive, if $(\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots, \boldsymbol{\zeta}_J)$ and $\left(\tilde{\boldsymbol{\zeta}}_1, \tilde{\boldsymbol{\zeta}}_2, \ldots, \tilde{\boldsymbol{\zeta}}_J\right)$ are not identical, then there exists at least a $j \in [J]$ such that $\zeta_j \neq \tilde{\zeta}_j$, plugging it back to Equation (3.5) gives:

$$p = \eta_1 \prod_{j=2}^{J} \zeta_j \neq \tilde{\eta}_1 \prod_{j=2}^{J} \tilde{\zeta}_j = \tilde{p}$$

- Surjectivity refers to for an arbitrary p, there exists a set of marginal probability and auxiliary probabilities $\boldsymbol{\zeta} = (\eta_1, \zeta_2, \ldots, \zeta_J)$ that maps to it. For η_1, we have:

$$\begin{aligned} \eta_1 =& \mathbb{P}\left(\boldsymbol{Z}_1(t) \mid \boldsymbol{Z}(t-1)\right) \\ =& \sum_{\boldsymbol{Z}_2(t)} \cdots \sum_{\boldsymbol{Z}_j(t)} \mathbb{P}\left(\boldsymbol{Z}_1(t), \boldsymbol{Z}_2(t), \ldots, \boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1)\right) \end{aligned} \tag{3.8}$$

  For ζ_j with $j = 2, \ldots, J$, we have:

$$\begin{aligned} \zeta_j =& \mathbb{P}\left(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_{1:(j-1)}(t), \boldsymbol{Z}(t-1)\right) \\ =& \frac{\mathbb{P}\left(\boldsymbol{Z}_{1:j}(t) \mid \boldsymbol{Z}(t-1)\right)}{\mathbb{P}\left(\boldsymbol{Z}_{1:(j-1)}(t) \mid \boldsymbol{Z}(t-1)\right)} \end{aligned} \tag{3.9}$$

  where

$$\mathbb{P}\left(\boldsymbol{Z}_{1:j}(t) \mid \boldsymbol{Z}(t-1)\right) = \sum_{\boldsymbol{Z}_{j+1}(t)} \cdots \sum_{\boldsymbol{Z}_J(t)} \mathbb{P}\left(\boldsymbol{Z}_1(t), \boldsymbol{Z}_2(t), \ldots, \boldsymbol{Z}_J(t) \mid \boldsymbol{Z}(t-1)\right)$$

That is, for an arbitrary joint transition matrix P, we can construct the corresponding auxiliary probability matrices $\boldsymbol{\zeta}_\sigma$ according to Equations (3.8), (3.9) through marginalizing, that maps back to P.

- As given in Equations (3.8), (3.9), the map $\mathcal{H}_\sigma \to \mathcal{P}$ involves only basic arithmetic operations, namely addition, multiplication, and division, hence both the map and its inverse are differentiable.

□

Under a given permutation σ, this diffeomorphism ensures that the dependency structures encoded in p are fully preserved in $\boldsymbol{\zeta}_\sigma$ and that the two parameter spaces share the same smooth geometric structure, as will be discussed in the following section.

3.2.1 The Geometric Interpretation

The relationship between joint transition probability p and the auxiliary probabilities ζ can also be interpreted geometrically. Let $\mathcal{H}$ denote the joint parameter space for auxiliary probability matrices $\boldsymbol{\zeta}_\sigma = \left\{\boldsymbol{\zeta}_{\sigma(1)}, \boldsymbol{\zeta}_{\sigma(2)}, \ldots, \boldsymbol{\zeta}_{\sigma(J)}\right\}$ with $\forall\sigma$ of $[J]$. Consider the equivalence relation $\sim$ on $\mathcal{H}$ defined as $\boldsymbol{\zeta}_\sigma \sim \boldsymbol{\zeta}_{\sigma'}$ if they are obtained from factorizing the same joint transition matrix P via different permutations σ and σ'. The corresponding equivalence class is denoted as:

$$[\boldsymbol{\zeta}_\sigma] = \{\boldsymbol{\zeta}_{\sigma'} \in \mathcal{H} : \boldsymbol{\zeta}_\sigma \sim \boldsymbol{\zeta}_{\sigma'}\}$$

The corresponding quotient space $\mathcal{H}/\sim$ is the set

$$\{[\boldsymbol{\zeta}_\sigma] : \boldsymbol{\zeta}_\sigma \in \mathcal{H}\}$$

of all the equivalence classes.

Due to Proposition 3.9, the section $S_\sigma \subset \mathcal{H}$ identified by the permutation σ of $[J]$ is a cross section such that $\forall\, [\boldsymbol{\zeta}_\sigma] \in \mathcal{H}/\sim$

$$S_\sigma \cap [\boldsymbol{\zeta}_\sigma] = \boldsymbol{\zeta}_\sigma$$

Figure 3.2 illustrates the geometric structure of the quotient space $\mathcal{H}/\sim$. The large ellipse represents the entire quotient space, which is partitioned into several blocks, each corresponding to an equivalence class $[\boldsymbol{\zeta}_\sigma]$ that maps to the same joint transition matrix P. For example, all auxiliary probabilities $\boldsymbol{\zeta}_\sigma$ in the top-middle block map to the same matrix P_1. According to Proposition 3.9, for a chosen permutation $\hat{\sigma}$, there exists one unique $\boldsymbol{\zeta}_{\hat{\sigma}}$ within each equivalence class, marked as red points. For instance, $\boldsymbol{\zeta}_{\hat{\sigma}}^{(1)}$ in the top-middle block is the unique representative under permutation $\hat{\sigma}$ corresponding to P_1. The cross-section $S_{\hat{\sigma}}$ is then defined as the collection of these red points across all equivalence classes.

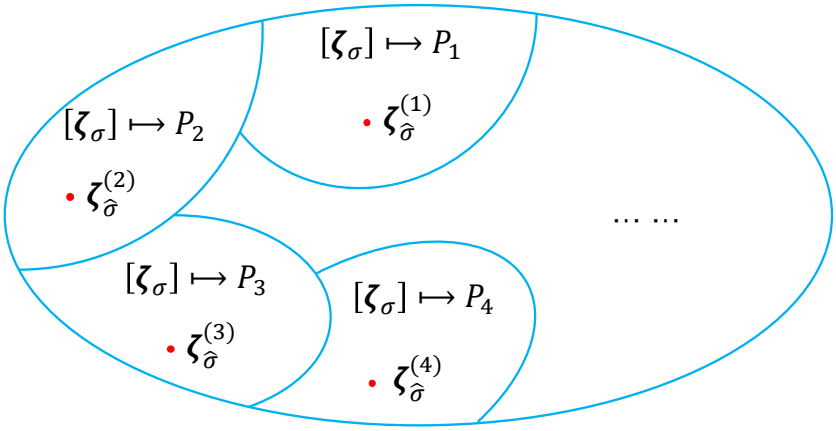


Figure 3.2: An illustration of the entire parameter space $\mathcal{H}$ for the auxiliary probability matrices $\boldsymbol{\zeta}_\sigma$ for all the permutation σ, is shown. The quotient space $\mathcal{H}/\sim$ are partitioned into multiple blocks, where each block represents an equivalence class of auxiliary probabilities $\boldsymbol{\zeta}_\sigma$ that map to the same joint transition matrix P. Selecting a specific permutation $\hat{\sigma}$ allows us to identify a unique representative $\boldsymbol{\zeta}_{\hat{\sigma}}$ within each equivalence class, highlighted as red points. The cross-section $S_{\hat{\sigma}}$ is then formed by these red points.

3.2.2 Encode Dependency Structures

Though the auxiliary probabilities do not have a clear interpretation as marginal transition probabilities do, but on the one hand it associates the marginal transition probability η_1 with the transition probability p with a bijection, ensuring that all the dependency structures embedded within the joint transition matrix P are preserved within the auxiliary probability matrices $\boldsymbol{\zeta}$ and on the other hand these auxiliary probabilities $\boldsymbol{\zeta}$ conveniently encode assumptions such as contemporaneous independence and Granger non-causality within their conditioning sets, as demonstrated in Proposition 3.10.

Proposition 3.10. *We can directly impose arbitrary contemporaneous independence and Granger non-causality through modifying the conditioning term of the auxiliary probabilities $\zeta_{\sigma(j)}$ for a specifically chosen permutation σ.*

Proof.

- For an arbitrary contemporaneous independence assumption with disjoint chain groups $g_{j_1}, g_{j_2} \in [K]$:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}(t-1) \tag{3.10}$$

Consider the permutation $\hat{\sigma}$ such that $\hat{\sigma}(1) = j_1$ and $\hat{\sigma}(2) = j_2$. We can impose the contemporaneous independence assumption (3.10) by removing $\boldsymbol{Z}_{j_1}(t)$ in the conditioning part of $\zeta_{\hat{\sigma}(2)}$ to show this independence. The modified $\zeta_{\hat{\sigma}(2)}$ is given as:

$$\zeta_{\hat{\sigma}(2)} = \mathbb{P}\big(\boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}_{j_1}(t),\ \boldsymbol{Z}(t-1)\big) = \mathbb{P}\big(\boldsymbol{Z}_{j_2}(t) \mid \underbrace{\boldsymbol{Z}(t-1)}_{\text{remove } \boldsymbol{Z}_{j_1}(t)}\big)$$

- For an arbitrary Granger non-causality assumption with with disjoint chain groups $g_{j_1}, g_{j_2} \in [K]$:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t-1) \mid \boldsymbol{Z}_{-j_2}(t-1) \tag{3.11}$$

Consider the permutation $\hat{\sigma}$ such that $\hat{\sigma}(1) = j_1$. We can impose this Granger non-causality (3.11) by removing $\boldsymbol{Z}_{j_2}(t-1)$ in the conditioning part of $\zeta_{\hat{\sigma}(1)}$ to show this

independence. The modified ζ_{j_1} is given as:

$$\zeta_{\hat{\sigma}(1)} = \mathbb{P}(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}_1(t-1), \ldots, \underbrace{\boldsymbol{Z}_{j_2-1}(t-1), \boldsymbol{Z}_{j_2+1}(t-1)}_{\text{remove } \boldsymbol{Z}_{j_2}(t-1)}, \ldots, \boldsymbol{Z}_j(t-1))$$

□

To sum up, the contemporaneous independence and the Granger non-causality uniquely determine the structure of the conditioning part in the auxiliary probabilities ζ. Imposing contemporaneous independence assumption corresponds to removing $\boldsymbol{Z}_{j_2}(t)$ at time t in the conditioning part of ζ_{j_1} while imposing Granger non-causality assumption corresponds to removing $\boldsymbol{Z}_{j_2}(t-1)$ at time $t-1$ in the conditioning part of ζ_{j_1}. This demonstrates that any combination of these two assumptions can be applied to ζ_{j_1} because it accommodates all the combination of $\boldsymbol{Z}_{j_2}(t)$ and $\boldsymbol{Z}_{j_2}(t-1)$ that accounts for these two assumptions in the conditioning part.

Example 3.11. Continuing with Example 3.5, we consider three chains sharing the state space $\mathcal{S} = \{1, 2\}$.

- Assume that we have the contemporaneous independence assumption that:

$$Z_2(t) \perp\!\!\!\perp Z_1(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{3.12}$$

Choose the identity permutation $\sigma^\star = \{1, 2, 3\}$. It modifies the structure of ζ_2 by excluding $Z_1(t)$ from the conditioning part.

$$\zeta_2 = \mathbb{P}\left(Z_2(t) \mid Z_1(t),\ \boldsymbol{Z}(t-1)\right) = \mathbb{P}\left(Z_2(t) \mid \boldsymbol{Z}(t-1)\right))$$

Then, for each of the vector $\boldsymbol{\zeta}_2(\cdot \mid \cdot, \boldsymbol{a})$ defined in Example 3.5 with $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, imposing (3.12) translates into

$$\boldsymbol{\zeta}_2(1 \mid 1, \boldsymbol{a}) = \boldsymbol{\zeta}_2(1 \mid 2, \boldsymbol{a}) \qquad \text{and} \qquad \boldsymbol{\zeta}_2(2 \mid 1, \boldsymbol{a}) = \boldsymbol{\zeta}_2(2 \mid 2, \boldsymbol{a}) \tag{3.13}$$

This can be interpreted as chain 1's current state does not influence chain 2's current

state. The constraints specified by Equation (3.13) reduce the number of constraints on each $\boldsymbol{\zeta}_2(\cdot \mid \cdot, \boldsymbol{a})$ from 2 to 1 subject to the original row constraints to make sure the total probability being 1.

- Assume that we have the Granger non-causality that:

$$Z_2(t) \perp\!\!\!\perp Z_1(t-1) \quad | \quad \boldsymbol{Z}_{-1}(t-1) \tag{3.14}$$

Choose the permutation $\sigma'(1) = 2$. It modifies the structure of $\zeta_1 \equiv \eta_1$ by excluding $Z_1(t-1)$ from the conditioning part.

$$\zeta_1 = \mathbb{P}\left(Z_2(t) \mid \boldsymbol{Z}(t-1)\right) = \mathbb{P}\left(Z_2(t) \mid Z_2(t-1), Z_3(t-1)\right)$$

Then, for the vector $\boldsymbol{\zeta}_1(\cdot \mid \boldsymbol{a})$, imposing (3.14) translates into

$$\boldsymbol{\zeta}_1(c_2 \mid (1, a_2, a_3)) = \boldsymbol{\zeta}_1(c_2 \mid (2, a_2, a_3)) \qquad \forall c_2, a_2, a_3 \in \mathcal{S} \tag{3.15}$$

This can be interpreted as the state of chain 1 at previous will not influence the current state of chain 2. In terms of the transition matrix $\boldsymbol{\eta}_1$, Equation 3.15 specifies that row i and row $i+4$ of matrix should take the same value for $i = 1, 2, 3, 4$, which reduces the number of independent rows of $\boldsymbol{\zeta}_2$ in this toy example from 8 to 4.

It is also worth noting that the auxiliary probabilities ζ are closely related to the adjacency matrix A defined in Equation (2.1). Consider its group-wise extension A_σ under a permutation σ, where rows and columns correspond to chain groups are permuted according to σ, as shown in the following Equation (3.16).

Since the bottom $J \times J$ sub-matrix of A_σ is symmetric, it suffices to consider only the upper triangular portion of that. In this case, each column of A_σ serves as an indicator vector that reflects the dependency structure for the corresponding auxiliary probabilities ζ_σ. Specifically, the j-th column of A_σ indicates that, under permutation σ, $\boldsymbol{Z}_j(t)$ depends on $\boldsymbol{Z}_1(t-1), \ldots, \boldsymbol{Z}_J(t-1)$ as well as $\boldsymbol{Z}_1(t), \ldots, \boldsymbol{Z}_{j-1}(t)$, aligning with the structure of

$\zeta_{\sigma(j)}$ defined in Equation (3.1) under full dependency.

$$
A_\sigma = \begin{array}{c|cccc}
 & \boldsymbol{Z}_{\sigma(1)}(t) & \boldsymbol{Z}_{\sigma(2)}(t) & & \boldsymbol{Z}_{\sigma(J)}(t) \\
\boldsymbol{Z}_{\sigma(1)}(t-1) & 1 & 1 & \cdots & 1 \\
\boldsymbol{Z}_{\sigma(2)}(t-1) & 1 & 1 & \cdots & 1 \\
 & \vdots & \vdots & & \vdots \\
\boldsymbol{Z}_{\sigma(J)}(t-1) & 1 & 1 & \cdots & 1 \\
\boldsymbol{Z}_{\sigma(1)}(t) & 1 & 1 & \cdots & 1 \\
\boldsymbol{Z}_{\sigma(2)}(t) & & 1 & \ddots & 1 \\
 & & & \ddots & \vdots \\
\boldsymbol{Z}_{\sigma(J)}(t) & & & & 1
\end{array}
$$

Effects from the previous time

Effects from the current time

(3.16)

3.3 Multinomial Logistic Regressions

Section 3.2 demonstrate that modelling the joint transition probability p using auxiliary probabilities ζ_σ with a predefined permutation σ offers several advantages, such as establishing bijection $\boldsymbol{\zeta}_\sigma \mapsto P$ under a fixed permutation $\sigma \in [J]$ and hence ensuring both approaches preserve all dependency structures embedded in the joint transition matrix P, as well as providing explicit encoding of the dependency structures like Granger non-causality and contemporaneous independence in first-order MMCs.

However, both models operate within constrained parameter spaces, as ζ_σ are bounded within $[0, 1]$. Moreover, we would like to retain the interpretability of the inter-chain effects, similar to that provided by the Inter-Chain Transition Models introduced in Section 2.9.1, where the joint transition probabilities p are expressed as linear combinations of inter-chain transition probabilities M and coupling coefficients θ. To achieve both of them, we aim to further decompose ζ_σ via the multinomial logistic regression.

3.3.1 Formulations of the Multinomial Logistic Regressions

The Univariate Marginal Model, structured in the same fashion as Bayesian network that factorizes the joint transition probabilities according to a prescribed permutation σ, requires K regressions, with one for each group of the chains. However, the design matrix in this model varies across regressions due to the increasing number of conditioning terms for the chain groups $g_{\sigma(1)}, \ldots, g_{\sigma(J)}$, as we will detail below.

Consider a first-order MMC $\boldsymbol{Z}$ composed of K chains, which are grouped into J non-empty subsets $g_1, g_2, \ldots, g_J$. Each individual chain Z_k for $k \in [K]$ shares the same finite state space $\mathcal{S} = \{1, 2, \ldots, s\}$. For simplicity, the identity permutation $\sigma^\star$ is applied to demonstrate the construction of the multinomial regressions. We represent the auxiliary probabilities ζ_j for each $j \in [J]$ using multinomial logistic regression. The link function employed in this modelling framework is built on each column $\boldsymbol{\zeta_j}(\cdot \mid \boldsymbol{b}, \boldsymbol{a})$ of the matrix $\boldsymbol{\zeta_j}$ constructed in Equation (3.4). It is defined as follows:

$$\zeta_j(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a}) = \frac{e^{\varphi_j(\boldsymbol{c}|\boldsymbol{b},\boldsymbol{a})}}{\sum e^{\varphi_j(\,\cdot\,|\boldsymbol{b},\boldsymbol{a})}} \tag{3.17}$$

where $\zeta_j(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})$ is defined the same as that in Equation (3.3) by applying identity permutation:

$$\zeta_j(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a}) = \mathbb{P}\left(\boldsymbol{Z}_j(t) = \boldsymbol{c} \mid \boldsymbol{Z}_{1:(j-1)}(t) = \boldsymbol{b}, \boldsymbol{Z}(t-1) = \boldsymbol{a}\right)$$

The summation in the denominator in Equation (3.17) is taken over all possible states $\boldsymbol{c} \in \boldsymbol{\mathcal{S}_c}$ for the group g_j. The associated log odds, denoted by $\varphi_j(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})$ for $j \in [J]$, take on $s_{\boldsymbol{c}} = |\boldsymbol{\mathcal{S}_c}| = s^{|g_j|}$ possible values, one for each distinct state $\boldsymbol{c}$. Each of these log odds $\varphi_j(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})$ can be modelled as the product of a design matrix $\boldsymbol{\mathcal{X}}$ and a corresponding regression coefficient vector $\boldsymbol{\beta_c}$. Formally as:

$$\boldsymbol{\varphi}_j(\boldsymbol{c} \mid \cdot\,,\, \cdot) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta_c} \qquad \forall \boldsymbol{c} \in \boldsymbol{\mathcal{S}_c} \tag{3.18}$$

Or explicitly in the matrix form as:

$$\begin{pmatrix} \varphi_j(\boldsymbol{c} \mid \boldsymbol{b}^{(1)}, \boldsymbol{a}^{(1)}) \\ \varphi_j(\boldsymbol{c} \mid \boldsymbol{b}^{(2)}, \boldsymbol{a}^{(1)}) \\ \vdots \\ \varphi_j(\boldsymbol{c} \mid \boldsymbol{b}^{(s_{\boldsymbol{b}})}, \boldsymbol{a}^{(s_{\boldsymbol{a}})}) \end{pmatrix}_{s_{\boldsymbol{a}} s_{\boldsymbol{b}} \times 1} = \begin{pmatrix} \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(1)}, \boldsymbol{b}^{(1)}) & \longrightarrow \\ \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(1)}, \boldsymbol{b}^{(2)}) & \longrightarrow \\ & \vdots & \\ \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(s_{\boldsymbol{a}})}, \boldsymbol{b}^{(s_{\boldsymbol{b}})}) & \longrightarrow \end{pmatrix}_{s_{\boldsymbol{a}} s_{\boldsymbol{b}} \times s_{\boldsymbol{a}} s_{\boldsymbol{b}}} \begin{pmatrix} \beta_{\boldsymbol{c}}^{(1)} \\ \beta_{\boldsymbol{c}}^{(2)} \\ \vdots \\ \beta_{\boldsymbol{c}}^{(s_{\boldsymbol{a}} s_{\boldsymbol{b}})} \end{pmatrix}_{s_{\boldsymbol{a}} s_{\boldsymbol{b}} \times 1} \tag{3.19}$$

- The vector $\boldsymbol{\varphi}_j(\boldsymbol{c} \mid \cdot\,, \cdot)$ is a column vector that consists all log odds for a given $\boldsymbol{c}$ and every $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$ and $\boldsymbol{b} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{b}}$. Consequently, it is of length $s_{\boldsymbol{a}} s_{\boldsymbol{b}}$, where $s_{\boldsymbol{a}}$ and $s_{\boldsymbol{b}}$ are the cardinality of the state space $\boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$ and $\boldsymbol{\mathcal{S}}_{\boldsymbol{b}}$ respectively. Precisely:

$$\begin{aligned} s_{\boldsymbol{a}} &= |\boldsymbol{\mathcal{S}}_{\boldsymbol{c}}| = |\boldsymbol{\mathcal{S}}_1 \times \cdots \times \boldsymbol{\mathcal{S}}_J| = s^K \\ s_{\boldsymbol{b}} &= |\boldsymbol{\mathcal{S}}_{\boldsymbol{b}}| = |\boldsymbol{\mathcal{S}}_1 \times \cdots \times \boldsymbol{\mathcal{S}}_{j-1}| = s^{\sum |g_i|} \end{aligned}$$

  where $|g_i|$ refers to the number of chains within the group g_i and the summation runs from 1 to $j-1$

- The design matrix $\boldsymbol{\mathcal{X}}$ is constructed by stacking together the transpose of $s_{\boldsymbol{a}} s_{\boldsymbol{b}}$ individual components $\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$, one for each $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$ and $\boldsymbol{b} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{b}}$. The explicit construction of each $\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$ through Kronecker product $\bigotimes$ is defined as:

$$\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b}) = \bigotimes_{k=1}^{K} v(a_k) \bigotimes_{k' \in g_{1:(j-1)}} v(b_{k'}) \tag{3.20}$$

$$\text{with} \quad v(s') = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^\top}_{\substack{\text{First element and the } s'\text{-th} \\ \text{element are set to be 1}}} \qquad s' \in \mathcal{S}$$

  where $g_{1:(j-1)} = g_1 \cup g_2 \cup \cdots \cup g_{j-1}$. The vector $v(a_k)$ and $v(b_{k'})$ is a one-hot encoding of the state a_k and $b_{k'}$ respectively with their first component fixed to 1. It follows that they are both of length s.

- The column vector $\boldsymbol{\beta}_{\boldsymbol{c}}$ represents the coefficients for the multinomial logistic re-

gression model and is of length $s_{\boldsymbol{a}}s_{\boldsymbol{b}}$. These coefficients provide interpretability of inter-chain effects. To ensure model identifiability, the coefficient corresponding to the baseline category, $\boldsymbol{\beta}_{\boldsymbol{c}_1}$, is set to zero.

It is straightforward to see that the entire multinomial logistic regression depends on the permutation σ. In what follows, we will comment on how this choice of permutation σ influences the multinomial logistic regression through the design matrix $\boldsymbol{\mathcal{X}}$. Under an alternative permutation σ of the set $[J]$, the row vector of the design matrix $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ can be rewritten according to Equation (3.20) as:

$$\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b}) = \bigotimes_{k=1}^{K} v(a_k) \bigotimes_{k' \in g_{\sigma(1:(j-1))}} v(b_{k'})$$

We attach the subscript σ in $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ to indicate that the design vector is constructed under the permutation σ. In general, $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ differs from $\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$ defined under the identity permutation, primarily because the Kronecker product of $v(b_{k'})$ in the former is taken over $k' \in g_{\sigma(1:(j-1))}$ instead of $k' \in g_{1:(j-1)}$. Consequently, the lengths of $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ and $\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$ generally differ due to the varying cardinalities of $g_{\sigma(1:(j-1))}$ and $g_{1:(j-1)}$. The only case where $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$ is when these cardinalities match for every $j = 2, \ldots, J$, under our assumption that all chains share the same state space $\mathcal{S}$. For instance, if we set each chain group g_j to be a singleton, then $|g_{\sigma(1:(j-1))}| = |g_{1:(j-1)}| = j - 1$ for each j, resulted in the agreement between $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ and $\boldsymbol{\mathcal{X}}(\boldsymbol{a}, \boldsymbol{b})$.

However, when the permutation σ is predefined based on the dependency structure of interest, the diffeomorphism is established between the regression coefficients $\boldsymbol{\beta}_\sigma$ and the auxiliary probabilities $\boldsymbol{\zeta}_\sigma$. This relationship is formalized in Proposition 3.12.

Proposition 3.12. *Under a given permutation σ, the transformation $\mathcal{F} : \mathcal{B}_\sigma \to \mathcal{H}_\sigma$ from the full space $\mathcal{B}_\sigma$ of $\boldsymbol{\beta}_\sigma$ to the full space $\mathcal{H}_\sigma$ of $\boldsymbol{\zeta}_\sigma$ is a diffeomorphism for $\boldsymbol{\zeta}_\sigma \neq 0$:*

- *The map is a bijection*
- *The map is differentiable as well as its inverse*

Proof. To show the diffeomorphism, we decompose the map into two: $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\varphi}_\sigma$ and $\boldsymbol{\varphi}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$ such that the composition of these two sub-maps gives the original map.

1. $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\varphi}_\sigma$: Matrix multiplication with the design matrix $\boldsymbol{\mathcal{X}}_\sigma$:

   This map is linear. Since each multiplicand v within the Kronecker products takes value from the modified one-hot encoding, which is linearly independent, the resulting Kronecker products $\boldsymbol{\mathcal{X}}_\sigma(\boldsymbol{a}, \boldsymbol{b})$ are also linearly independent. Consequently, the design matrix is of full rank and is therefore invertible. As a result, the linear map $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\varphi}_\sigma$ is bijective. Furthermore, being a linear map ensures it is differentiable, and its inverse is also differentiable. Thus, $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\varphi}_\sigma$ is a diffeomorphism.

2. $\boldsymbol{\varphi}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$: Link function for multinomial logistic regression given in Equation (3.17): This is also known as the standard softmax function: $\mathbb{R}^{s_c - 1} \to \boldsymbol{\Delta}^{s_c}$ when we restrict the coefficient of baseline level to be 0 for identifiability. Firstly, the softmax function is smooth since it involves only exponentials, sums, and divisions. For fixed $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{a}}$ and $\boldsymbol{b} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{b}}$, given $\zeta_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})$ for $\forall \boldsymbol{c} \in \boldsymbol{\mathcal{S}}_{\boldsymbol{c}}$, we can derive the inverse to recover $\varphi_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})$ as:

   $$\varphi_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a}) = \log \left(\frac{\zeta_{\sigma(j)}(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a})}{\zeta_{\sigma(j)}(\boldsymbol{c}_1 \mid \boldsymbol{b}, \boldsymbol{a})} \right)$$

   Where $\zeta_{\sigma(j)}(\boldsymbol{a}_1 \mid \boldsymbol{b}, \boldsymbol{c})$ refers to the auxiliary probability for baseline level $\boldsymbol{a}_1$. Thus, the softmax is a bijection with a smooth inverse, making the map $\boldsymbol{\varphi}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$ a diffeomorphism.

Therefore, the map $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$, composited by $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\varphi}_\sigma$ and $\boldsymbol{\varphi}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$, is a diffeomorphism.

□

By combining Proposition 3.9 and 3.12, we yield the following Lemma 3.13 that demonstrates the diffeomorphism between the joint transition matrices $P \in \mathcal{P}$ and the regression coefficients in the induced parameter space $\mathcal{B}_\sigma$ given the permutation σ

Lemma 3.13. *Given a permutation σ, the transformation $g_2 : P \mapsto \boldsymbol{\beta}_\sigma$ from the joint transition matrices P within the full parameter space $\mathcal{P}$ to the regression coefficients $\boldsymbol{\beta}_\sigma$*

under permutation σ, is a diffeomorphism.

This diffeomorphism not only prevents the issue of non-identifiability arising in Linear Decomposition Models discussed in Section 2.9, but also provides a smooth, invertible reparametrisation. In the Frequentist setting, it permits maximum likelihood estimation to be carried out directly on the joint transition matrix P and guarantees Jacobian-based change-of-variables to obtain the regression coefficients $\boldsymbol{\beta}_\sigma$, with a invertible Jacobian matrix (see Appendix A.1). In the Bayesian setting, the diffeomorphism allows priors to be specified on the auxiliary probabilities $\boldsymbol{\zeta}_\sigma$ through the regression coefficients $\boldsymbol{\beta}_\sigma$, which lie in a fully unconstrained parameter space. Furthermore, the diffeomorphism guarantees that the dependency structures captured by $\boldsymbol{\zeta}_\sigma$ are fully preserved in $\boldsymbol{\beta}_\sigma$. It turns out that $\boldsymbol{\beta}_\sigma$ provides an efficient and interpretable representation of these structures, as formalized in the next section.

3.3.2 Encode Dependency Structures

In this section, we formalize how to encode dependency structures via the regression coefficients $\boldsymbol{\beta}_\sigma$ in Equations (3.18) in Univariate Marginal Model.

Proposition 3.14. *We can encode any desired Granger non-causality or contemporaneous independence by setting specific coefficients in $\boldsymbol{\beta}_\sigma$, which would initially be unconstrained under full independency, to zero, under a carefully chosen permutation σ.*

Proof. Since the map $\boldsymbol{\beta}_\sigma \mapsto \boldsymbol{\zeta}_\sigma$ is a diffeomorphism, the number of free parameters remains unchanged across the two representations. When Granger non-causality and contemporaneous independence are imposed, they introduce constraints on the auxiliary probabilities $\boldsymbol{\zeta}_\sigma$.

We consider the contemporaneous independence and Granger non-causality for each pair of disjoint chain groups g_{j_1} and g_{j_2}, (i.e. $g_{j_1}, g_{j_2} \subset [k]$ and $g_{j_1} \cap g_{j_2} = \emptyset$):

- For an arbitrary contemporaneous independence relationship with chain groups g_{j_1}

and g_{j_2}:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}(t-1)$$

Consider the permutation σ on $[J]$ where $\sigma(1) = j_1$ and $\sigma(2) = j_2$. As discussed in Proposition 3.10, this contemporaneous independence can be transferred into removing $\boldsymbol{Z}_{j_1}(t)$ in the conditioning part of $\zeta_{\sigma(2)}$ as:

$$\begin{aligned}\zeta_{\sigma(2)} &= \mathbb{P}\big(\boldsymbol{Z}_{\sigma(2)}(t) \mid \boldsymbol{Z}_{\sigma(1)}(t), \boldsymbol{Z}(t-1)\big) \\ &= \mathbb{P}\big(\boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}_{j_1}(t), \boldsymbol{Z}(t-1)\big) \\ &= \mathbb{P}\big(\boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}(t-1)\big)\end{aligned}$$

which means that the corresponding log odds $\varphi_{\sigma(2)}$ must satisfy:

$$\varphi_{\sigma(2)}\big(\boldsymbol{c} \mid \boldsymbol{b}, \boldsymbol{a}\big) = \varphi_{\sigma(2)}\big(\boldsymbol{c} \mid \boldsymbol{a}\big) \equiv \varphi_{j_2}\big(\boldsymbol{c} \mid \boldsymbol{a}\big)$$

Hence, the corresponding regression coefficients $\boldsymbol{\beta}_{\boldsymbol{c}}$ must have:

$$\beta_{\boldsymbol{c}}(\boldsymbol{b}_j) = 0 \qquad \forall \boldsymbol{c} \in \boldsymbol{Z}_{\sigma(j)} \quad \text{and} \quad g_j \cap g_{j_1} \neq \emptyset$$

$\boldsymbol{\beta}_{\boldsymbol{c}}(\boldsymbol{b}_j)$ denotes the subset of regression coefficients within $\boldsymbol{\beta}_{\boldsymbol{c}}$ associated with $\zeta_{j_2}(\boldsymbol{c} \mid \cdot, \cdot)$ that capture the influence of past states in the chain group g_j. The group g_j contains at least one element from g_{j_1} and therefore represents the effect arising from the concurrent chains $\boldsymbol{Z}_{j_1}$. Under contemporaneous independence, these coefficients are constrained to zero, indicating that all concurrent effects between g_{j_1} and g_{j_2} are removed.

- For an arbitrary Granger non-causality that chain group g_{j_2} does not cause g_{j_1}:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t-1) \mid \boldsymbol{Z}_{-j_2}(t-1)$$

Consider the permutation σ on $[J]$ where $\sigma(1) = j_1$ and $\sigma(2) = j_2$. As discussed in Proposition 3.10, this Granger non-causality can be transferred into removing

$\boldsymbol{Z}_{j_2}(t-1)$ in the conditioning part of $\zeta_{\sigma(1)}$ as:

$$\begin{aligned}\zeta_{\sigma(1)} &= \mathbb{P}\big(\boldsymbol{Z}_{\sigma(1)}(t) \mid \boldsymbol{Z}(t-1)\big) = \mathbb{P}\big(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}_{j_2}(t-1), \boldsymbol{Z}_{-j_2}(t-1)\big) \\ &= \mathbb{P}\big(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}_{-j_2}(t-1)\big)\end{aligned}$$

which means that the corresponding log odds $\varphi_{\sigma(1)}$ must satisfy:

$$\varphi_{\sigma(1)}\big(\boldsymbol{c} \mid \boldsymbol{a}\big) = \varphi_{\sigma(1)}\big(\boldsymbol{c} \mid \boldsymbol{a}_{-\sigma(2)}\big) \equiv \varphi_{j_1}\big(\boldsymbol{c} \mid \boldsymbol{a}_{-j_2}\big)$$

Hence, the corresponding regression coefficients $\boldsymbol{\beta}_{\boldsymbol{c}}$ must have:

$$\beta_{\boldsymbol{c}}(\boldsymbol{a}_j) = 0 \qquad \forall \boldsymbol{c} \in \boldsymbol{Z}_{\sigma(j)} \quad \text{and} \quad g_j \cap g_{j_2} \neq \emptyset$$

$\boldsymbol{\beta}_{\boldsymbol{c}}(\boldsymbol{a}_j)$ denotes the subset of regression coefficients within $\boldsymbol{\beta}_{\boldsymbol{c}}$ associated with $\zeta_{j_1}(\boldsymbol{c} \mid \cdot, \cdot) \equiv \zeta_{j_1}(\boldsymbol{c} \mid \cdot)$ that capture the effect of past states in the chain group g_j. The group g_j contains at least one element from g_{j_2} and thus represents the influence of the past states of g_{j_2} on the current state of g_{j_1}. Under the Granger non-causality, these coefficients are constrained to zero, indicating that the past states of g_{j_2} have no effect on the current state of g_{j_1}.

□

Followed by the Proposition 3.14, we can alternatively define the contemporaneous independence and Granger non-causality in terms of the regression coefficients in the Univariate Marginal Model, instead of the conditional probabilities. This formulation is formalized as follows:

Definition 3.15 (Define Dependency Structures via the Regression Coefficients)**.** *Consider the MMC $\boldsymbol{Z}$ with chains $[K] = 1, 2, \ldots, K$. Given two disjoint groups of chains $g_{j_1}, g_{j_2} \subset [K]$ and the corresponding marginal process $\boldsymbol{Z}_{j_1}$ and $\boldsymbol{Z}_{j_2}$ of $\boldsymbol{Z}$ and the permutation σ on $[J]$ where $\sigma(1) = j_1$ and $\sigma(2) = j_2$:*

- *$\boldsymbol{Z}_{j_1}$ and $\boldsymbol{Z}_{j_2}$ are said to be contemporaneously independent if and only if the corre-*

sponding regression coefficients satisfy:

$$\boldsymbol{\beta}_{\boldsymbol{c}}(\boldsymbol{b}_j) = 0 \qquad \forall \boldsymbol{c} \in \boldsymbol{Z}_{\sigma(j)} \quad \textit{and} \quad g_j \cap g_{j_1} \neq \emptyset$$

where $\boldsymbol{\beta}_{\boldsymbol{c}}(\boldsymbol{b}_j)$ denotes the subset of regression coefficients within $\boldsymbol{\beta}_{\boldsymbol{c}}$ associated with $\zeta_{j_2}(\boldsymbol{c} \mid \cdot, \cdot)$ that capture the influence of past states in the chain group g_j.

- *$\boldsymbol{Z}_j$ is not Granger caused by $\boldsymbol{Z}_{j'}$ if and only if the corresponding regression coefficients satisfy:*

$$\boldsymbol{\beta}_{\boldsymbol{c}}(\boldsymbol{a}_j) = 0 \qquad \forall \boldsymbol{c} \in \boldsymbol{Z}_{\sigma(j)} \quad \textit{and} \quad g_j \cap g_{j_2} \neq \emptyset$$

where $\beta_{\boldsymbol{c}}(\boldsymbol{a}_j)$ denotes the part of the regression coefficients $\boldsymbol{\beta}_{\boldsymbol{c}}$ that associated with $\zeta_{j_1}(\boldsymbol{c} \mid \cdot, \cdot) \equiv \zeta_{j_1}(\boldsymbol{c} \mid \cdot)$ that capture the effect of past states in the chain group g_j.

The way the Univariate Marginal Model encodes dependency structures can also be expressed using an indicator vector $\mathbb{I}_\sigma$ defined under a permutation σ. Specifically, for each regression coefficient vector $\boldsymbol{\beta}_{\boldsymbol{c}}$ associated with the chain group $g_{\sigma(j)}$, we construct the corresponding indicator vector $\mathbb{I}_{\sigma(j)}(\boldsymbol{\beta}_{\boldsymbol{c}})$ using a Kronecker product.

$$\mathbb{I}_{\sigma(j)}(\boldsymbol{\beta}_{\boldsymbol{c}}) = \bigotimes_{i=1}^{j-1} \mathbb{I}_{\sigma(j)|\sigma(i)}(t) \bigotimes_{i'=1}^{J} \mathbb{I}_{\sigma(j)|\sigma(i')}(t-1)$$

where

$$\mathbb{I}_{\sigma(j)|\sigma(i)}(t) = \begin{cases} \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix}^\top}_{\text{a vector of 1}} & \text{if } \boldsymbol{Z}_{\sigma(j)}(t) \text{ correlated with } \boldsymbol{Z}_{\sigma(i)}(t) \\ \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \end{pmatrix}^\top}_{\substack{\text{first element being 1} \\ \text{the rest being 0}}} & \text{if } \boldsymbol{Z}_{\sigma(j)}(t) \text{ does not depend on } \boldsymbol{Z}_{\sigma(i)}(t) \end{cases} \tag{3.21}$$

The marginal indicator vector $\mathbb{I}_{\sigma(j)|\sigma(i)}(t)$ is a binary vector of length $s^{|g_{\sigma(i)}|}$ that indicates whether the states $\boldsymbol{Z}_{\sigma(j)}(t)$ is related with the concurrent states $\boldsymbol{Z}_{\sigma(i)}(t)$ of chain group $g_{\sigma(i)}$. Specifically, if there is dependence, all elements of $\mathbb{I}_{\sigma(j)|\sigma(i)}(t)$ are set to 1; otherwise,

only the first element is 1 and the rest are 0. Similarly, $\mathbb{I}_{\sigma(j)|\sigma(i')}(t-1)$ is defined in the same way but for the dependency on the previous states $\boldsymbol{Z}_{\sigma(i')}(t-1)$ of group $g_{\sigma(i')}$, and its length is $s^{|g_{\sigma(i')}|}$.

By definition, the indicator vector $\mathbb{I}_j(\boldsymbol{\beta_c})$, whose length matches that of $\boldsymbol{\beta_c}$, is a binary vector consisting only zeros and ones. An entry of one (1) permits the corresponding regression coefficient to be non-zero, and a 0 enforces it to be zero. Under full dependency, $\mathbb{I}_j(\boldsymbol{\beta_c})$ is a vector of all ones, imposing no restrictions. Independence assumptions modify this structure:

- **Contemporaneous Independence Assumption**: If chain group $g_{\sigma(j_1)}$ is contemporaneously independent of chain group $g_{\sigma(j_2)}$, then the marginal indicator $\mathbb{I}_{\sigma(j_2)|\sigma(j_1)}(t)$ is set to be $\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^\top$.
- **Granger non-Causality**: If chain group $g_{\sigma(j_1)}$ is not Granger caused by $g_{\sigma(j_2)}$, then the marginal indicator $\mathbb{I}_{\sigma(j_1)|\sigma(j_2)}(t-1)$ is set to be $\begin{pmatrix}1 & 0 & \cdots & 0\end{pmatrix}^\top$.

In both cases, modifying these marginal indicators introduce zeros into the indicator $\mathbb{I}_j(\boldsymbol{\beta_c})$, hence enforcing the corresponding coefficients in $\boldsymbol{\beta_c}$ to be zero. Generally, the construction of the indicator vector $\mathbb{I}_{\sigma(j)}(\boldsymbol{\beta_c})$ under permutation σ is also directly related to the adjacency matrix A_σ defined under the same permutation σ, as shown in Equation (3.16). Each element of A_σ is either 0 or 1, indicating independence or dependency, respectively. This binary nature corresponds to the two possible cases of $\mathbb{I}_{\sigma(j)|\sigma(i)}(t)$ and $\mathbb{I}_{\sigma(j)|\sigma(i')}(t-1)$ as defined in Equation (3.21) respectively. Consequently, each column $\boldsymbol{Z}_{\sigma(j)}(t)$ in the adjacency matrix A_σ, which outlines the dependency structure for the chain group $g_{\sigma(j)}$, uniquely determines the pattern of zeros and ones in $\mathbb{I}_{\sigma(j)}(\boldsymbol{\beta_c})$. This pattern, in turn, dictates which elements in $\boldsymbol{\beta_c}$ are zero or non-zero.

3.3.3 Choice of Permutation

Based on Definition 3.15, prior knowledge about the dependency structures of interest can be incorporated into an appropriate choice of permutation σ. Specifically:

- To investigate *contemporaneous independence* between two chain groups g_{j_1} and g_{j_2}, we choose $\sigma(1) = j_1$ and $\sigma(2) = j_2$. The corresponding auxiliary probability

$$\zeta_{\sigma(2)} = \mathbb{P}\big(\boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}_{j_1}(t), \boldsymbol{Z}(t-1)\big)$$

  includes regression coefficients associated with $\boldsymbol{Z}_{j_1}(t)$, which represent the contemporaneous interactions between g_{j_1} on g_{j_2}.

- To investigate *Granger non-causality* for a particular chain group, say g_{j_1}, we set $\sigma(1) = j_1$. The first auxiliary probability then becomes

$$\zeta_{\sigma(1)} = \mathbb{P}\big(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}(t-1)\big),$$

  which allows us to directly model the influence of all past states $\boldsymbol{Z}(t-1)$ on the current state of g_{j_1}.

When no prior structural information is available, a natural starting point is the identity permutation σ^*. However, it is important to note that certain combinations of dependency structures may not be representable under a single permutation, and alternative permutations may need to be considered. For example, consider the contemporaneous independence between chain groups $g_{\sigma(2)}$ and $g_{\sigma(3)}$ under a given permutation σ:

$$\boldsymbol{Z}_{\sigma(2)}(t) \perp\!\!\!\perp \boldsymbol{Z}_{\sigma(3)}(t) \mid \boldsymbol{Z}(t-1) \tag{3.22}$$

This independence is not encoded in the current permutation σ because the Univariate Marginal Model defines the conditional distribution

$$\mathbb{P}\big(\boldsymbol{Z}_{\sigma(3)}(t) \mid \boldsymbol{Z}_{\sigma(2)}(t), \boldsymbol{Z}_{\sigma(1)}(t), \boldsymbol{Z}(t-1)\big)$$

which conditions on $\boldsymbol{Z}_{\sigma(1)}(t)$ and thereby prevents us from isolating the independence between $g_{\sigma(2)}$ and $g_{\sigma(3)}$. To properly encode the structure in Equation (3.22), a new permutation σ' must be considered such that $\sigma'(1) = \sigma(2), \sigma'(2) = \sigma(3)$ and by considering

the term:

$$\mathbb{P}\big(\boldsymbol{Z}_{\sigma'(2)}(t) \mid \boldsymbol{Z}_{\sigma'(1)}(t), \boldsymbol{Z}(t-1)\big)$$

This illustrates the importance of choosing permutations carefully, depending on the prior information of the dependency structure under investigation. In general, a single set of regressions under a given permutation may not fully capture all possible dependency structures. For dependencies not explicitly reflected under a specific permutation, one must first recover the transition probability matrix P based on that permutation, and then factorize P under a new permutation σ that aligns with the target structure. While this process is deterministic, it introduces additional computational steps. An alternative approach is to fit models under multiple permutations, allowing for a more comprehensive identification of dependency structures. We will later introduce Hierarchical Marginal Model in Chapter 4 to avoid the effect of the permutations.

3.3.4 Examples on Encoding Dependency Structures

In this section, we present examples on how to encode dependency structures via the Univariate Marginal Model. Sticking on the same setting in Example 3.11, we consider an MMC with 3 chains, each taking values in $\mathcal{S} = \{1, 2\}$. Under the Univariate Marginal Model, with the predefined σ on $\{1, 2, 3\}$, the quantities that need to be specified, as shown in Example 3.3, are:

$$\begin{aligned}
\zeta_{\sigma(1)} &= \mathbb{P}\big(Z_{\sigma(1)}(t) \mid \boldsymbol{Z}(t-1)\big) \\
\zeta_{\sigma(2)} &= \mathbb{P}\big(Z_{\sigma(2)}(t) \mid Z_{\sigma(1)}(t), \boldsymbol{Z}(t-1)\big) \\
\zeta_{\sigma(3)} &= \mathbb{P}\big(Z_{\sigma(3)}(t) \mid Z_{\sigma(1)}(t), Z_{\sigma(2)}(t), \boldsymbol{Z}(t-1)\big)
\end{aligned}$$

Moreover, with a bivariate state space $\mathcal{S} = \{1, 2\}$, the multinomial regression simplifies to a standard logistic regression. We construct 3 logistic regressions, with each corresponding to one of the auxiliary probabilities introduced above, to explicitly encode conditional independence, contemporaneous independence, and Granger non-causality with a specifically chosen permutation σ.

(I) Full Dependency

Recall the full dependency scenario illustrated in Figure 2.2, meaning there are no constraints on the regression coefficients $\boldsymbol{\beta}$. Thus, each coefficient β can vary freely in $\mathbb{R}$. The corresponding logistic regression can be written as:

$$\operatorname{logit}\big(\boldsymbol{\zeta}_j(\boldsymbol{c} \mid \cdot\,,\, \cdot\,)\big) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{\boldsymbol{c}} \qquad \forall g_j \subseteq \{1,2,3\}$$

Such a logistic regression can be written in one equation via the indicator function $\mathbb{I}(\cdot)$. Take the logistic regression on $\zeta_2 = \mathbb{P}\big(Z_2(t) \mid Z_1(t), \boldsymbol{Z}(t-1)\big)$ under identity permutation $\sigma^\star$ as an example:

$$\begin{aligned}
&\operatorname{logit}\Big(\zeta_2\big(c \mid b, \boldsymbol{a}\big)\Big) \\
&= \underbrace{\beta_{\emptyset}}_{\text{intercept}} + \underbrace{\mathbb{I}\big(a_1=2\big)\beta_{(\{1\},\emptyset)} + \mathbb{I}\big(a_2=2\big)\beta_{(\{2\},\emptyset)} + \mathbb{I}\big(a_3=2\big)\beta_{(\{3\},\emptyset)} + \mathbb{I}\big(b=2\big)\beta_{(\emptyset,\{1\})}}_{\text{univariate effects}} \\
&+ \underbrace{\mathbb{I}\big(a_1=2,a_2=2\big)\beta_{(\{1,2\},\emptyset)} + \mathbb{I}\big(a_1=2,a_3=2\big)\beta_{(\{1,3\},\emptyset)} + \mathbb{I}\big(a_2=2,a_3=2\big)\beta_{(\{2,3\},\emptyset)}}_{\text{bivariate effects}} \\
&+ \underbrace{\mathbb{I}\big(a_1=2,b=2\big)\beta_{(\{1\},\{1\})} + \mathbb{I}\big(a_2=2,b=2\big)\beta_{(\{2\},\{1\})} + \mathbb{I}\big(a_3=2,b=2\big)\beta_{(\{3\},\{1\})}}_{\text{bivariate effects}} \\
&+ \underbrace{\mathbb{I}\big(a_1=2,a_2=2,a_3=2\big)\beta_{(\{1,2,3\},\emptyset)} + \mathbb{I}\big(a_1=2,a_2=2,b=2\big)\beta_{(\{1,2\},\{1\})}}_{\text{trivariate effects}} \\
&+ \underbrace{\mathbb{I}\big(a_1=2,a_3=2,b=2\big)\beta_{(\{1,3\},\{1\})} + \mathbb{I}\big(a_2=2,a_3=2,b=2\big)\beta_{(\{2,3\},\{1\})}}_{\text{trivariate effects}} \\
&+ \underbrace{\mathbb{I}\big(a_1=2,a_2=2,a_3=2,b=2\big)\beta_{(\{1,2,3\},\{1\})}}_{\text{4-variate effects}}
\end{aligned} \tag{3.23}$$

- $\boldsymbol{\zeta}_j(\boldsymbol{b}_j \mid \cdot)$ is a 16×1 column vector that stacks the values $\zeta_2(c \mid b, \boldsymbol{a})$ for all combinations of $b \in \mathcal{S}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$.

- $\boldsymbol{\mathcal{X}}$ is the 16×16 design matrix associated with ζ_2, constructed according to Equation 3.20. For other auxiliary probabilities, the dimension of $\boldsymbol{\mathcal{X}}$ is 8×8 and 32×32 for ζ_1 and ζ_3 respectively.

- $\boldsymbol{\beta}$ is a 16×1 vector of logistic regression coefficients associated with the marginal

transition of chain $Z_2(t)$ under the identity permutation. For notational simplicity, we write:

$$\beta_{(j',j)} = \beta_{(j',j)\,2}$$

where the subscript 2 indicates that the target is chain $Z_2(t)$ is omitted. Each coefficient $\beta_{(j',j)}$ reflects the joint log-odds effect of the previous state of chain group $g_{j'}$ and the current state of chain group g_j being in an alternative state on the current state of chain $Z_2(t)$ being in an alternative state, say $s = 2$. The term $\beta_\emptyset$ denotes the intercept, capturing the baseline level.

In the case of full dependency, none of the regression coefficients $\boldsymbol{\beta}_2$ are restricted to 0, allowing all coefficients to vary freely.

(II) Conditional Independence

Recall the Conditional Independence on 3 chains as illustrated in Figure 2.3, where all the blue vertical arrows are missing, indicating all concurrent interactions among the chains are eliminated after conditioning on past states:

$$Z_1(t) \perp\!\!\!\perp Z_2(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1)$$

According to Proposition 3.14, this conditional independence can be incorporated into the logistic coefficients $\boldsymbol{\beta}_{\sigma(2)}$ and $\boldsymbol{\beta}_{\sigma(3)}$ for $\zeta_{\sigma(2)}$ and $\zeta_{\sigma(3)}$ as:

$$\begin{aligned}
&\boldsymbol{\beta}_{(j,\{\sigma(1)\})\,\sigma(2)} = 0 \qquad \forall g_j \in \{1,2,3\} \\
&\boldsymbol{\beta}_{(j,j')\,\sigma(3)} = 0 \qquad \forall g_j \in \{1,2,3\} \quad \text{and} \quad \forall g_{j'} \in \{\sigma(1), \sigma(2)\}
\end{aligned}$$

In this case, all three logistic regressions for $\zeta_{\sigma(1)}$, $\zeta_{\sigma(2)}$, and $\zeta_{\sigma(3)}$ exclude the concurrent effects in the conditioning terms and thus reduce to:

$$\operatorname{logit}\big(\boldsymbol{\zeta}_{\sigma(k)}(2 \mid \boldsymbol{b}, \boldsymbol{a})\big) = \operatorname{logit}\big(\boldsymbol{\zeta}_{\sigma(k)}(2 \mid \boldsymbol{a})\big) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_k \qquad k = 1,2,3$$

Notably, under the conditional independence assumption, this formulation coincides with the logistic regression structure used in the Hierarchical Marginal Model that will be discussed in the Chapter 4, where the design matrix $\boldsymbol{\mathcal{X}}$ retains the same form as in Equation (4.33). As a result, the Univariate Marginal Model shares the same regression structure as Equation (4.34) and becomes independent of the permutation, since the only permutation related term, $\boldsymbol{b}$, is eliminated from the conditioning part.

(III) Contemporaneous Independence

Recall the Contemporaneous Independence on 3 chains as illustrated in Figure 2.6, where the missing blue vertical arrows between $\big(Z_1(t), Z_2(t)\big)$ and $Z_3(t)$ indicating the partially missing concurrent interaction between them:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{3.24}$$

According to Proposition 3.14, if we choose the identity permutation $\sigma^\star = \{1, 2, 3\}$ and consider $\zeta_3 = \mathbb{P}\big(Z_3(t) \mid Z_1(t), Z_2(t), \boldsymbol{Z}(t-1)\big)$, then, contemporaneous independence is incorporated into the logistic regression for ζ_3 by setting specific elements of the regression coefficient vector $\boldsymbol{\beta}_3$ to zero. Specifically, we set

$$\beta_{(j,j')} = 0 \qquad \text{for} \quad g_{j'} \neq \emptyset$$

to eliminate the influence of the concurrent states on chain 3. This enforces the contemporaneous independence condition described in Equation (3.24). As a result, the logistic regression for ζ_3 simplifies to:

$$\begin{aligned}
&\operatorname{logit}\Big(\zeta_3\big(2 \mid \boldsymbol{b}, \boldsymbol{a}\big)\Big) \;=\; \operatorname{logit}\Big(\zeta_3\big(2 \mid \boldsymbol{a}\big)\Big) \\
&= \beta_\emptyset \;+\; \mathbb{I}\big(a_1 = 2\big)\beta_{(\{1\},\emptyset)} \;+\; \mathbb{I}\big(a_2 = 2\big)\beta_{(\{2\},\emptyset)} \;+\; \mathbb{I}\big(a_3 = 2\big)\beta_{(\{3\},\emptyset)} \\
&+ \mathbb{I}\big(a_1 = 2, a_2 = 2\big)\beta_{(\{1,2\},\emptyset)} \;+\; \mathbb{I}\big(a_1 = 2, a_3 = 2\big)\beta_{(\{1,3\},\emptyset)} \;+\; \mathbb{I}\big(a_2 = 2, a_3 = 2\big)\beta_{(\{2,3\},\emptyset)} \\
&+ \mathbb{I}\big(a_1 = 2, a_2 = 2, a_3 = 2\big)\beta_{(\{1,2,3\},\emptyset)}
\end{aligned}$$

This example also illustrates that contemporaneous independence generalises conditional independence: unlike conditional independence, it allows to retain concurrent interactions partially. Specifically, it places no restrictions on the concurrent terms in $\boldsymbol{\beta}_2$ for ζ_2, enabling the model to capture concurrent interactions between $Z_1(t)$ and $Z_2(t)$.

(IV) Granger non-Causality

Recall the Granger non-Causality on 3 chains as illustrated in Figure 2.8, where the missing black arrows from $Z_3(t-1)$ to $Z_1(t)$ and $Z_2(t)$ represent that chain 3 at previous time does not affect the states of chain 1 or chain 2 at current time. Formally, this is expressed as

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \quad | \quad Z_1(t-1), Z_2(t-1)$$

According to Proposition 4.13, by defining the chain group $g_1 = \{1,2\}$, $g_2 = \{3\}$ and choosing identity permutation $\sigma^\star = \{1,2\}$, the Granger non-causality can be encoded within the multinomial logistic regression on the regression coefficients $\boldsymbol{\beta}_1$ on ζ_1:

$$\beta_{(j,\emptyset)} = 0 \qquad \forall\, g_j \supseteq \{3\}$$

meaning that, the regression on ζ_1 simplifies to

$$\begin{aligned}
&\operatorname{logit}\Big(\zeta_1\big(\boldsymbol{b} \mid \boldsymbol{a}\big)\Big) \;=\; \operatorname{logit}\Big(\zeta_1\big(\boldsymbol{b} \mid (a_1,a_2)\big)\Big) \\
&= \beta_\emptyset \;+\; \mathbb{I}\big((a_1,a_2)=(2,1)\big)\beta^{(1)}_{(\{1\},\emptyset)} \;+\; \mathbb{I}\big((a_1,a_2)=(1,2)\big)\beta^{(2)}_{(\{1\},\emptyset)} \;+\; \mathbb{I}\big((a_1,a_2)=(2,2)\big)\beta^{(3)}_{(\{1\},\emptyset)}
\end{aligned}$$

Note that logit here refers to a multinomial logistic regression, as the chain group $g_1 = \{1,2\}$ has four possible levels. This equation indicates that all the terms corresponding to the effects from chain group g_2, namely $\beta_{(\{2\},\emptyset)}$ and $\beta_{(\{1,2\},\emptyset)}$, are set to be 0 and consequently eliminates the effects from chain 3 at previous time to the joint of chain 1 and chain 2.

3.4 Generative Model

In this section, we examine the utilization of the Univariate Marginal Model as generative model, that is to generate an MMC $\boldsymbol{Z}$ that exhibits desired dependency structures, based on the regression coefficients $\boldsymbol{\beta}_\sigma$ under a carefully chosen permutation σ.

For the Univariate Marginal Model, there are no constraints on the regression coefficients $\boldsymbol{\beta}_\sigma$, meaning that one can jointly generate a full set of $\boldsymbol{\beta}_\sigma$ under a given permutation σ, and the resulting coefficients will automatically correspond to a valid joint transition probability matrix P. To impose specific dependency structures, it is sufficient to set the relevant regression coefficients to zero while drawing the remaining coefficients from a chosen distribution, such as a normal distribution. The entire procedure of the generating process via the Univariate Marginal Model can be summarized in the Algorithm 1:

Algorithm 1 Generating Data from the Univariate Marginal Model

Input: the predefined permutation σ as well as the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ for the prior distribution of regression coefficients $\boldsymbol{\beta}_\sigma$ under σ

Output: an MMC $\boldsymbol{Z}$ generated by the Univariate Marginal Model

if impose dependency structures **then**

  Set the corresponding $\boldsymbol{\beta}_\sigma = 0$ according to Proposition 3.14

end if

Generate the rest of the regression coefficients $\boldsymbol{\beta}_\sigma \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Compute the joint transition matrix P by $\boldsymbol{\beta}_\sigma$ via $g_2(\cdot)$ defined in Lemma 3.13

Generate MMC $\boldsymbol{Z}$ based on the the joint transition matrix P

Note that since the regression coefficients $\boldsymbol{\beta}_\sigma$ lie in an unconstrained parameter space, they can be generated from any distribution supported on $\mathbb{R}$. Here, the normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is just one particular choice among all the possible distributions.

Nevertheless, as discussed in Section 3.6.1, a key limitation of the Univariate Marginal Model lies in the selection of the permutation σ. When attempting to impose multiple

dependency structures simultaneously, it can be the case that no single permutation σ is able to accommodate all of them. This can prevent the model from generating an MMC that reflects a specific combination of the dependency structures, limiting its flexibility in certain applications.

3.5 Simulation Studies

In this section, we examine the performance of the Univariate Marginal Model, in particular the ability in recovering the true underlying regression coefficients, $\boldsymbol{\beta}_\sigma$, as well as detecting the potential dependency structures embedded within the MMC based on the simulated data.

3.5.1 Simulation Methods

For simplicity, we consider an MMC with 3 chains and 2 states in our simulations for the Univariate Marginal Model. This setup can be readily extended to more chains and states. The simulation proceeds as follows:

1. Generate regression coefficients $\boldsymbol{\beta}_\sigma$ representing:

   - The full model (all chains fully connected)
   - Contemporaneous independence

   $$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1) \tag{3.25}$$

   - Granger non-causality

   $$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1) \tag{3.26}$$

   - A mixture of the contemporaneous independence and Granger non-causality

   Given the contemporaneous independence (3.25) and Granger non-causality (3.26), it is sufficient to consider the identity permutation $\sigma^\star = \{1, 2, 3\}$, that is to model:

$$\eta_1 = \mathbb{P}\big(Z_1(t) \mid \boldsymbol{Z}(t-1)\big)$$
$$\zeta_2 = \mathbb{P}\big(Z_2(t) \mid Z_1(t), \boldsymbol{Z}(t-1)\big)$$
$$\zeta_3 = \mathbb{P}\big(Z_3(t) \mid Z_1(t), Z_2(t), \boldsymbol{Z}(t-1)\big)$$

since this permutation inherently accommodates both dependency structures, as will be demonstrated below. Note that we drop the subscript $\sigma^\star$, indicating that the regression coefficients $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_{\sigma^\star}$ are specified under the identity permutation.

Once the permutation $\sigma^\star$ is specified, the generating procedure is straightforwardly provided in Algorithm 1. We choose to generate each coefficient independently from a normal distribution with mean 0 and variance 2:

$$\beta \sim \mathcal{N}(0, 2)$$

2. Compute the corresponding transition matrix P based on $\boldsymbol{\beta}$ under the Univariate Marginal Model respectively.

3. Conduct 500 repetitive experiments on that transition matrix P, that is to simulate 500 different MMC sequences of length 2000 according to this P.

4. For each simulation, a full model is fitted via:

   - Frequentist inference by maximizing the likelihood with respect to p followed by transforming it into $\boldsymbol{\beta}$ (see Section A.1)
   - Bayesian inference using a Gibbs sampler with Pólya Gamma latent variables with 5000 MCMC iterations to directly obtain $\boldsymbol{\beta}$ (see Section A.3)

   Note that, in the Bayesian setting, we could alternatively use the conjugate Dirichlet distribution (see Section A.2.1). However, because it is defined on the joint transition probabilities p, additional manipulation is required to map it to the target regression coefficients $\boldsymbol{\beta}$. We therefore prefer the Pólya-Gamma approach, which acts directly on $\boldsymbol{\beta}$.

In each repetitive experiments, we compute the MLE and the posterior median after discarding the first 500 samples as burn-in.

5. To evaluate the recovery of the true regression coefficients $\boldsymbol{\beta}$, we present box plots and histograms of the MLE (Frequentist) and posterior median (Bayesian).

6. To evaluate detection of dependency structures, we:

   - Apply Hotelling's T^2 test and the likelihood ratio test in the Frequentist setting
   - Compute contour probabilities of the posterior medians in the Bayesian setting

   and record the number of times the null hypothesis H_0 (i.e. existence of certain dependency structures) is retained across the 500 repetitive experiments.

The significance level α for the hypothesis testing is set to 5%. For Bayesian inference, we use the `pgdraw` package [Polson et al., 2013, Windle, 2013] to sample Pólya-Gamma auxiliary variables ω, and the `bayesSurv` package [Besag et al., 1995, Held, 2004] to compute the contour probabilities. The convergence and the stationarity of the MCMC chains was assessed visually by inspecting the trace plots of the sampled regression coefficients, as attached in Appendix B.1.1.

The simulation results for the Univariate Marginal Model under: (I) Full model, (II) Contemporaneous independence, (III) Granger non-causality, (IV) Mixed dependency structures are presented in the following section 3.5.2.

3.5.2 Simulation Results

(I) Full model

Recall the full model illustrated in Figure 3.3. There is no independent relationship among the chains and all the chains are fully connected, indicating no additional restrictions are required to place on the regression coefficients $\boldsymbol{\beta}$.

As an illustration, consider the regression coefficients $\boldsymbol{\beta}_1$ associated with η_1. Figure 3.4 displays box plots of the estimates for all 8 regression coefficients obtained across repetitions. The Bayesian estimates of the posterior median are coloured in light blue, the Frequentist

estimates are shown in yellow, and the true values are indicated by the red points. It can be seen that the regression coefficients $\boldsymbol{\beta}$ are well estimated by both Bayesian and Frequentist approaches as the true value lies close to the median of each estimates.

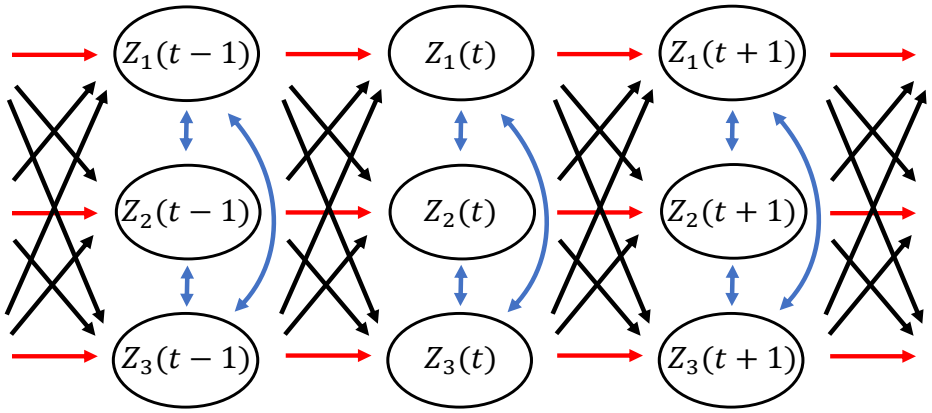


Figure 3.3: An illustration of an MMC with 3 chains under full dependencies, with all the chains fully connected with each other.

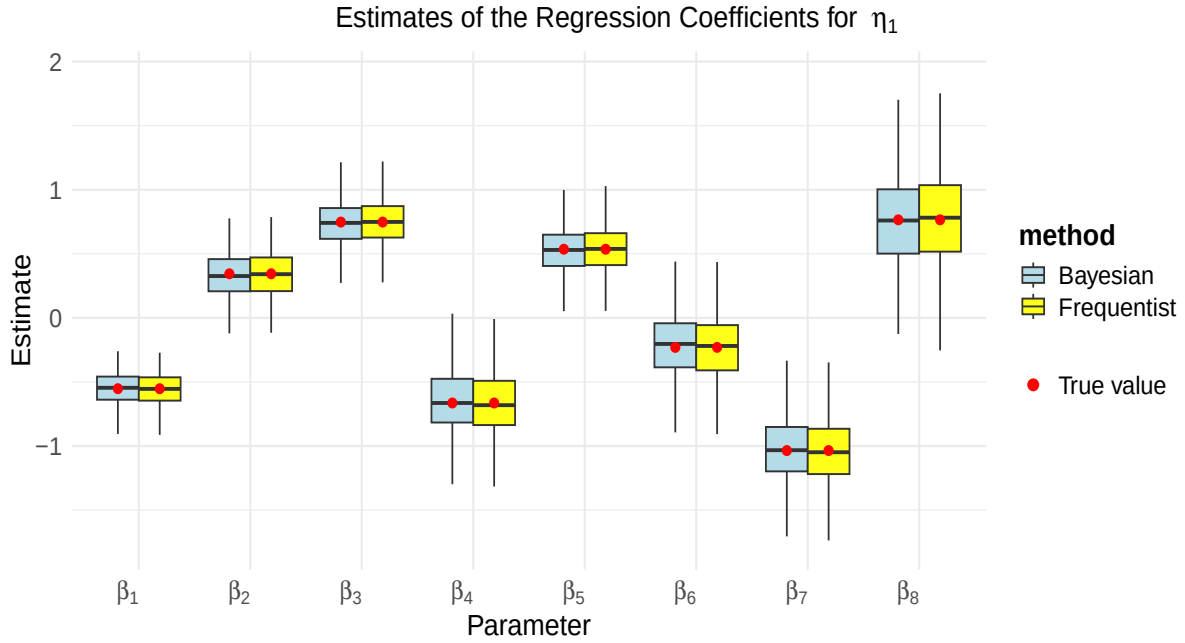


Figure 3.4: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_1$ for ζ_1, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points in the Univariate Marginal Model. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

(II) Contemporaneous Independence

Recall the contemporaneous independence setting illustrated in Figure 3.5, where chains 1 and 2 are contemporaneously independent from chain 3:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1) \tag{3.27}$$

This independence is visually represented by the absence of the blue vertical arrows connecting $Z_1(t)$ and $Z_2(t)$ with $Z_3(t)$, in contrast to the full model shown in Figure 3.3.

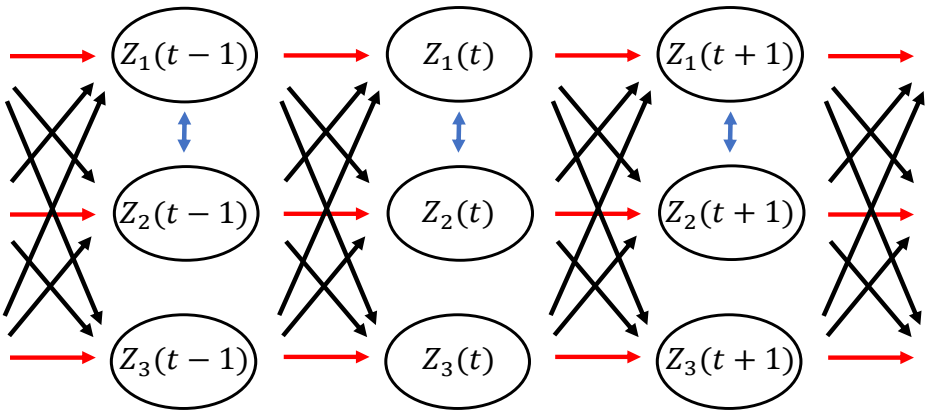


Figure 3.5: An illustration of an MMC with 3 chains under the contemporaneous independence where chain 1 and chain 2 are contemporaneously independent from chain 3.

This contemporaneous independence can be encoded within the auxiliary probabilities $\boldsymbol{\zeta}$ under the identity permutation $\sigma^\star$ as

$$\begin{aligned}\zeta_3 &= \mathbb{P}\big(Z_3(t) \mid Z_1(t), Z_2(t), \boldsymbol{Z}(t-1)\big) \\ &= \mathbb{P}\big(Z_3(t) \mid \boldsymbol{Z}(t-1)\big) \qquad \text{(under Contemporaneous Independence)}\end{aligned} \tag{3.28}$$

This implies that, among the $2^5 = 32$ regression coefficients in $\boldsymbol{\beta}_3$ for chain 3, only 8 are non zero, reflecting the effects purely from $\boldsymbol{Z}(t-1)$, while no explicit constraints are imposed on ζ_1 and ζ_2.

For illustration, consider the first eight regression coefficients of $\boldsymbol{\beta}_3$ associated with ζ_3. Figure 3.6 presents the box plots of the estimates for these coefficients obtained across 500 repetitions, with the posterior mean in Bayesian and the MLE in Frequentist shown in light blue and yellow, respectively, and the true values indicated by red points. It is evident that among these eight coefficients, only β_1 (intercept) and β_5 (the effect of $Z_1(t-1)$ on $Z_3(t)$) are non zero, while the rest regression coefficients corresponding to the concurrent interaction with $Z_1(t)$ and $Z_2(t)$ are all zero indicating the contemporaneous independence. Furthermore, both the Bayesian and Frequentist approaches provide accurate estimates, as the red points representing the true values lie almost at the centre of the box plots.

To test the contemporaneous independence in Frequentist approach, we perform a likeli-

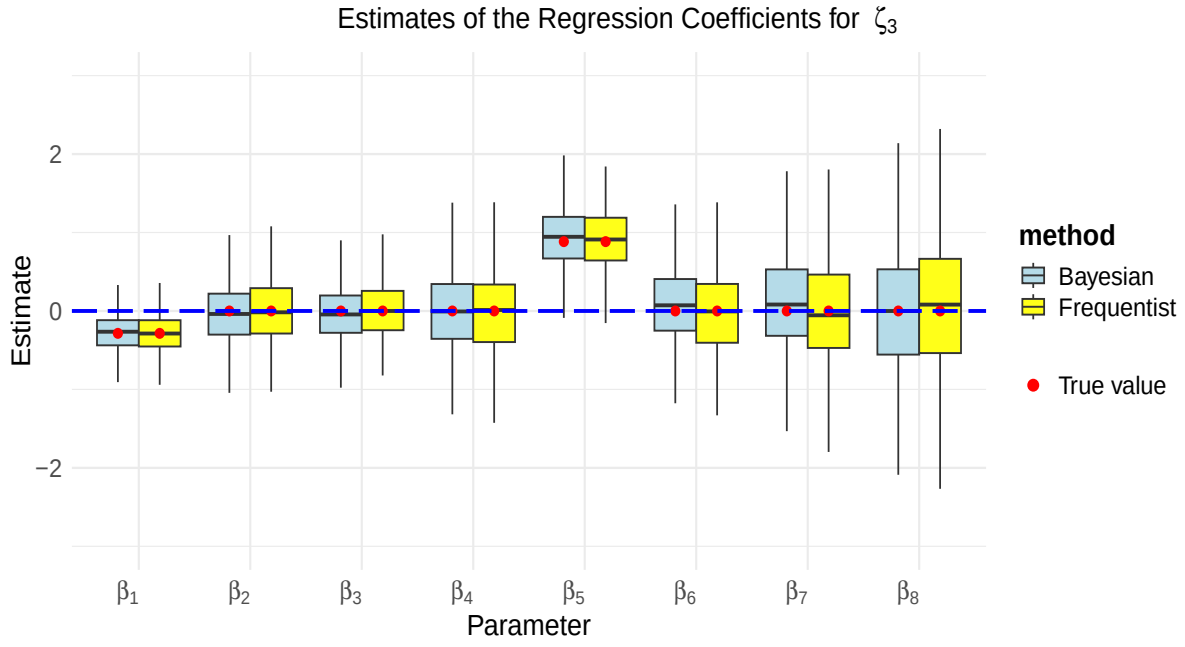


Figure 3.6: An illustration of the posterior medians of the first 8 regression coefficients $\boldsymbol{\beta}_3$ for ζ_3, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points in the Univariate Marginal Model. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

hood ratio test within each of the 500 repeated experiments for the nested model:

$$H_0: \quad Z_1 \text{ and } Z_2 \text{ are contemporaneously independent from } Z_3$$
$$H_1: \quad Z_1 \text{ and } Z_2 \text{ are correlated with } Z_3$$

with the test statistic

$$-2\log\Lambda = 2\left(\max_{\boldsymbol{\beta} \text{ under } H_1} \ell(\boldsymbol{\beta}) - \max_{\boldsymbol{\beta}_0 \text{ under } H_0} \ell(\boldsymbol{\beta}_0)\right) \sim \chi^2_\nu \tag{3.29}$$

where the degree of freedom is $\nu = 32 - 8 = 24$, since under the restricted model H_0 (contemporaneous independence), only 8 regression c.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 24 out of 500 experiments (4.8%), which is fully consistent with the random variation under H_0. Furthermore, we record the p-value from each experiment and present their distribution in the histogram as the following Figure 3.7:

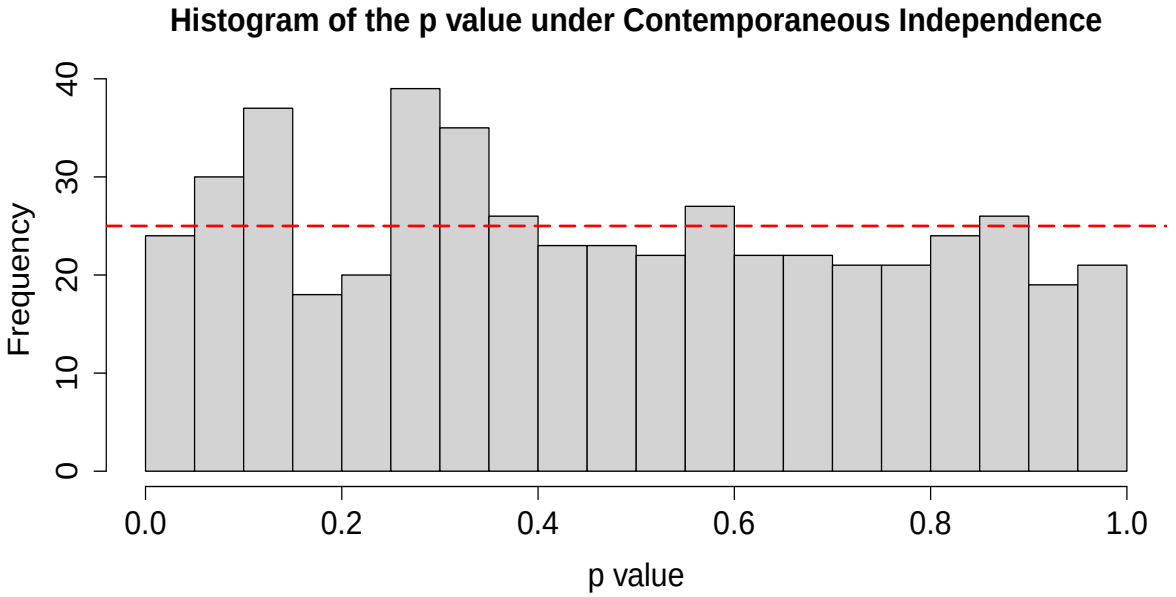


Figure 3.7: The histogram of the p-value from each of the 500 repetitive experiments under contemporaneous independence for the Univariate Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

As shown, the 500 p-values from the repetitive experiments are approximately uniformly distributed over $[0, 1]$, closely matching the Uniform distribution indicated by the red dashed line. This suggests that the test is well-calibrated and that the null hypothesis H_0 of contemporaneous independence is plausible.

In the Bayesian framework, we can evaluate the posterior contour probabilities of the corresponding 24 regression coefficients in $\boldsymbol{\beta}_3$ indicating the concurrent interaction between $Z_1(t), Z_2(t)$ and $Z_3(t)$, centred at $\mathbf{0}$. Using the `bayesSurv` package, we compute the posterior contour probabilities exceed 0.5, providing sufficient evidence that these coefficients are centred at $\mathbf{0}$ and thereby supporting contemporaneous independence.

(III) Granger non-Causality

Recall the Granger non-causality illustrated in Figure 3.8, where chains 1 and 2 are contemporaneously independent of chain 3 given the past:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1) \tag{3.30}$$

This independence is characterized by the missing black arrow pointing $Z_1(t)$ and $Z_2(t)$ from $Z_3(t-1)$, in comparison with the full model in Figure 3.3.

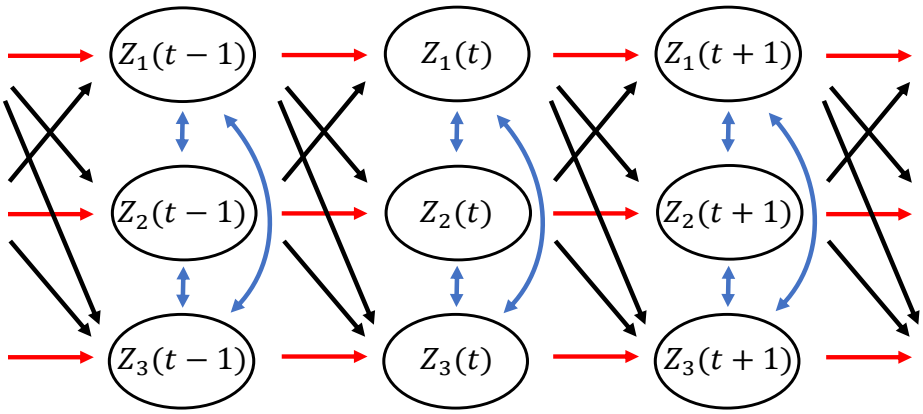


Figure 3.8: An illustration of an MMC with 3 chains under the Granger non-causality where chain 1 and chain 2 are not Granger caused by chain 3.

This Granger non-causality can be expressed within the auxiliary probabilities $\boldsymbol{\zeta}$ under the identity permutation $\sigma^\star$ as

$$\begin{aligned}
\eta_1 \cdot \zeta_2 &= \mathbb{P}\big(Z_1(t) \mid \boldsymbol{Z}(t-1)\big) \cdot \mathbb{P}\big(Z_2(t) \mid Z_1(t), \boldsymbol{Z}(t-1)\big) \\
&= \mathbb{P}\big(Z_1(t), Z_2(t) \mid \boldsymbol{Z}(t-1)\big) \\
&= \mathbb{P}\big(Z_1(t), Z_2(t) \mid \boldsymbol{Z}_{-3}(t-1)\big) \qquad \text{(under Granger non-causality)} \\
&= \mathbb{P}\big(Z_1(t) \mid \boldsymbol{Z}_{-3}(t-1)\big) \cdot \mathbb{P}\big(Z_2(t) \mid Z_1(t), \boldsymbol{Z}_{-3}(t-1)\big) \qquad (3.31)
\end{aligned}$$

This implies that, among the 8 regression coefficients in $\boldsymbol{\beta}_1$ for chain 1, 4 of them are zero, reflecting the absence of effects from $Z_3(t-1)$. Similarly, among the 16 regression coefficients in $\boldsymbol{\beta}_2$ for chain 2, 8 of them are zero, again reflecting the absence of effects from $Z_3(t-1)$.

For illustration, consider the regression coefficients of $\boldsymbol{\beta}_1$ associated with ζ_1. Figure 3.9

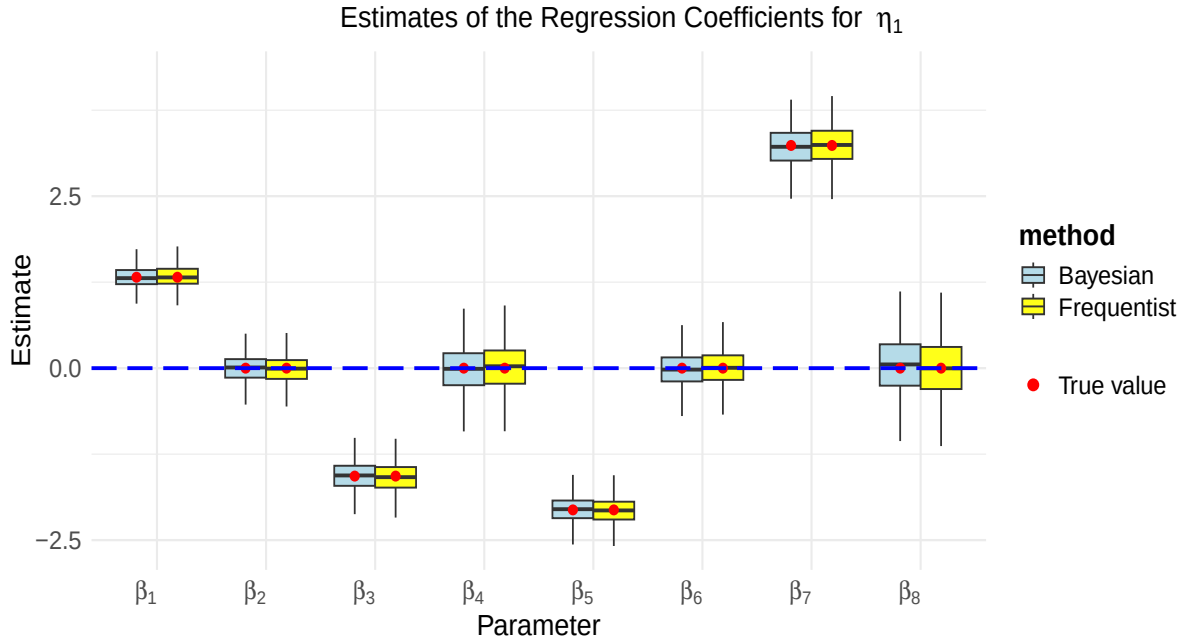


Figure 3.9: An illustration of the estimation of the regression coefficients $\boldsymbol{\beta}_1$ for η_1, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points in the Univariate Marginal Model. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

presents the box plots of the estimates for these coefficients obtained across 500 repetitions, with the posterior mean in Bayesian and the MLE in Frequentist shown in light blue and yellow, respectively, and the true values indicated by red points. It is evident that among these eight coefficients, $\beta_2, \beta_4, \beta_6, \beta_8$ accounting for the effects from chain 3 at previous time to chain 1 at current time are all zero, which is consistent with the Granger non-causality. Furthermore, both the Bayesian and Frequentist approaches provide accurate estimates, as the red points representing the true values lie almost at the centre of the box plots.

To test the Granger non-causality in Frequentist approach, we perform a likelihood ratio test within each of the 500 repeated experiments for the following nested model:

$$\begin{aligned} H_0: &\quad Z_1 \text{ and } Z_2 \text{ are not Granger caused by } Z_3 \\ H_1: &\quad Z_1 \text{ and } Z_2 \text{ are Granger caused by } Z_3 \end{aligned}$$

The test statistic $-2\log\Lambda$ is the same as is given in Equation (3.29) but with a degree of freedom of $\nu = 4 + 8 = 12$, since under the restricted model H_0 (Granger non-causality), 4 coefficients in $\boldsymbol{\beta}_1$ and 8 coefficients in $\boldsymbol{\beta}_2$, accounting for the effects from $Z_3(t-1)$, are constrained to be zero for both cases.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 33 out of 500 experiments (6.6%), which is fairly consistent with the random variation under H_0. Furthermore, we record the p-value from each experiment and present their distribution in the histogram as the following Figure 3.10:

As shown, the 500 observed p-values are approximately uniformly distributed over $[0, 1]$, closely matching the Uniform distribution indicated by the red dashed line. This suggests that the test is well-calibrated and that the null hypothesis H_0 of Granger non-causality is plausible.

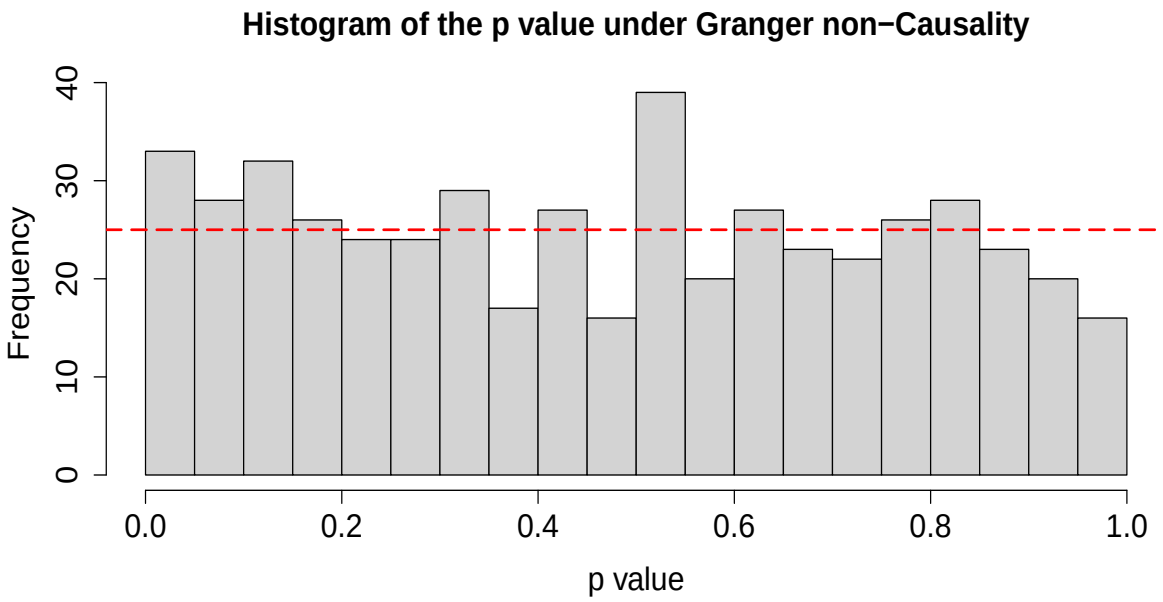


Figure 3.10: The histogram of the p-value from each of the 500 repetitive experiments under Granger non-causality for the Univariate Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

As in the previous case of contemporaneous independence, within the Bayesian framework, we can evaluate the posterior contour probabilities of the 12 regression coefficients in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, corresponding to the potential Granger causality from $Z_3(t-1)$ to $Z_1(t)$ and $Z_2(t)$, cente red at $\mathbf{0}$. Using the `bayesSurv` package, we find that the posterior contour probabilities exceed 0.5, providing sufficient evidence that these coefficients are centred at $\mathbf{0}$ and thereby supporting Granger non-causality.

(IV) Mixed Dependency

Finally, consider the a mixed dependency structures with both contemporaneous independence and the Granger non-causality, as illustrated in Figure 3.11, where we assume:

- **Contemporaneous Independence**: Chain 1 and chain 2 are contemporaneously independent from chain 3 (see Equation (3.27)), characterized by the missing black arrow pointing $Z_1(t)$ and $Z_2(t)$ from $Z_3(t-1)$, in comparison with the full model.
- **Granger non-Causality**: Chain 1 and chain 2 are not Granger caused by chain 3

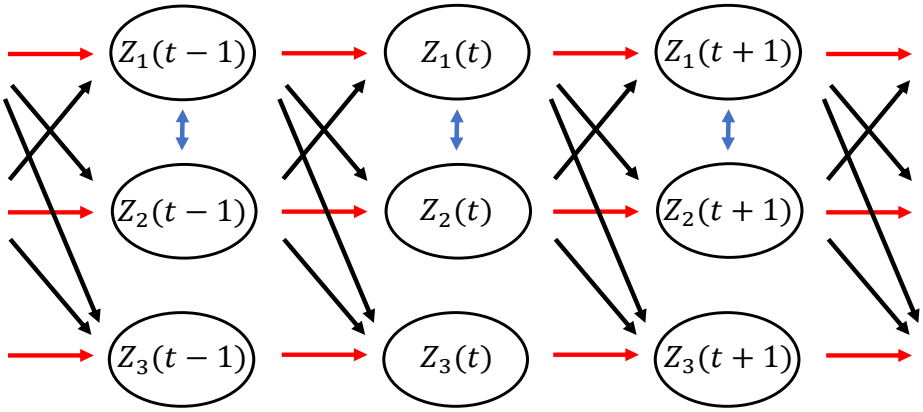


Figure 3.11: An illustration of an MMC with 3 chains under the mixed dependencies including both Contemporaneous Independence where chain 1 and chain 2 are contemporaneously independent from chain 3 and the Granger non-causality where chain 1 and chain 2 are not Granger caused by chain 3.

(see Equation (3.30)), characterized by the missing blue vertical arrow connecting $Z_1(t)$ and $Z_2(t)$ with $Z_3(t)$, in comparison with the full model in Figure 3.3.

This mixed dependency structure can be represented within the auxiliary probabilities $\boldsymbol{\zeta}$ under the identity permutation $\sigma^\star$ by combining Equation 3.28, which encodes contemporaneous independence in ζ_3, and Equation 3.31, which embeds Granger non-causality through η_1 and ζ_2. This implies that, among the $8+16+32=56$ regression coefficients in $\boldsymbol{\zeta}$, a total of 36 are constrained to zero: 24 in $\boldsymbol{\beta}_3$ due to contemporaneous independence, and 12 in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ due to Granger non-causality.

Hence, for illustration, we choose to visualize:

- The estimates of $\boldsymbol{\beta}_1$ for η_1, to examine whether $Z_1(t)$ is Granger caused by $Z_3(t-1)$ by inspecting the relevant coefficients.
- The first eight estimates of $\boldsymbol{\beta}_3$ for ζ_3, to check whether $Z_3(t)$ is contemporaneously independent of $Z_1(t)$ and $Z_2(t)$ by inspecting the relevant coefficients.

Figure 3.12 presents the box plots of the estimates of $\boldsymbol{\beta}_1$ for η_1 obtained across 500 repetitions, with the posterior mean in Bayesian and the MLE in Frequentist shown in light

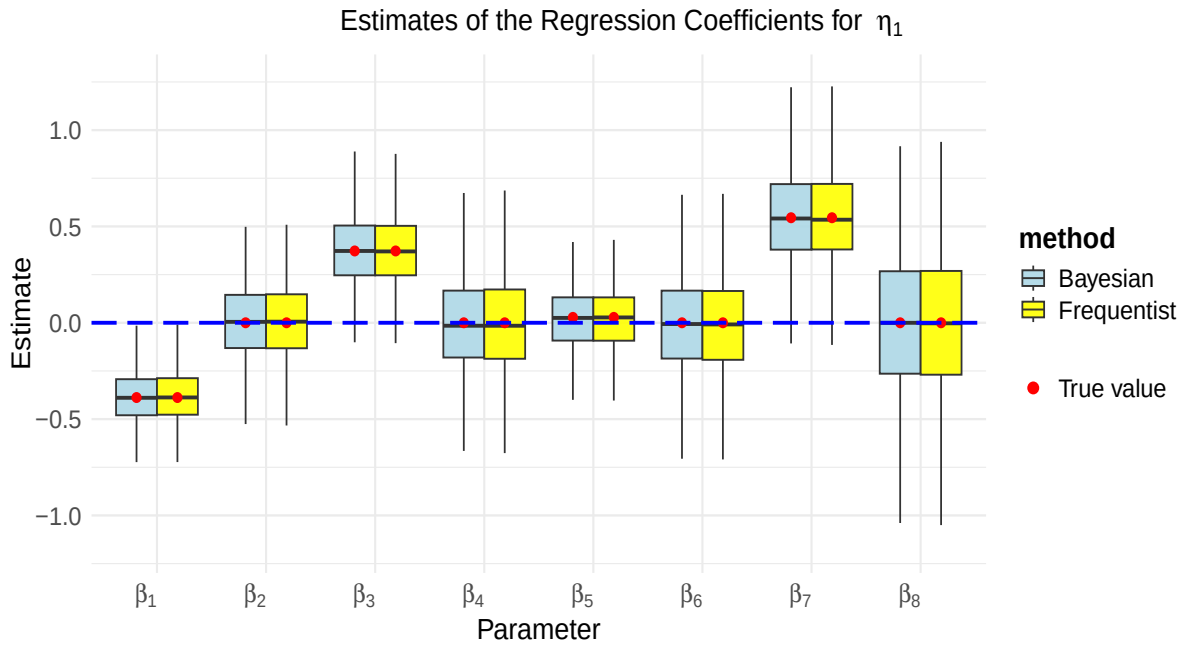


Figure 3.12: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_1$ for η_1, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points in the Univariate Marginal Model. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

blue and yellow, respectively, and the true values indicated by red points. It is evident that $\beta_2, \beta_4, \beta_6$ and β_8, which correspond to the effects of chain 3 at the previous time step on chain 1 at the current time, are all zero, consistent with Granger non-causality. Note that it is a coincidence that β_5, indicating the effect of chain 1 on its own, is set to a value close to zero.

Figure 3.13 presents the box plots of the estimates for these coefficients obtained across 500 repetitions, with the Bayesian and Frequentist estimates shown in light blue and yellow, respectively, and the true values indicated by red points. It is evident that among these eight coefficients, only β_1 (intercept) and β_5 (the effect of $Z_1(t-1)$ on $Z_3(t)$) are non zero, while the rest regression coefficients corresponding to the concurrent interaction with $Z_1(t)$ and $Z_2(t)$ are all zero, implying the contemporaneous independence.

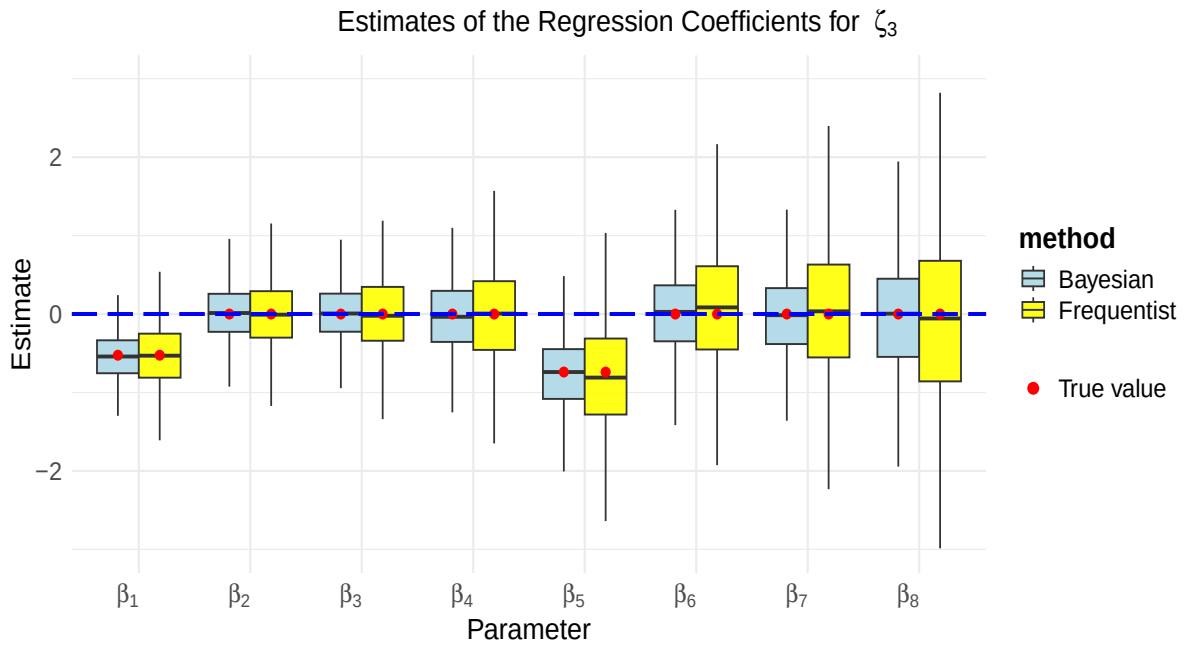


Figure 3.13: An illustration of the estimation of the first 8 regression coefficients $\boldsymbol{\beta}_3$ for ζ_3, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points in the Univariate Marginal Model. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

For both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_3$ illustrated in Figure 3.12 and 3.13 respectively, the Bayesian and Frequentist approaches provide accurate estimates for $\boldsymbol{\beta}_1$, as the red points representing the true values lie almost at the centre of the box plots in both cases.

To test the mixed dependency structures of contemporaneous independence and the Granger non-causality in Frequentist approach, we perform a likelihood ratio test within each of the 500 repeated experiments for the following nested model:

$$
\begin{aligned}
H_0: \quad & Z_1 \text{ and } Z_2 \text{ are not Granger caused by } Z_3 \quad \textbf{and} \\
& Z_1 \text{ and } Z_2 \text{ are contemporaneously independent from } Z_3 \\
H_1: \quad & \text{There is no such independence relationships}
\end{aligned}
$$

The test statistic $-2\log\Lambda$ is the same as is given in Equation (3.29) but with a degree of freedom of $\nu = 24 + 12 = 36$ since there are 24 constraints from contemporaneous

independence and 12 from the Granger non-causality.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 30 out of 500 experiments (6%), which is fairly consistent with the random variation under H_0. Furthermore, we record the p-value from each experiment and present their distribution in the histogram as the following Figure 3.14:

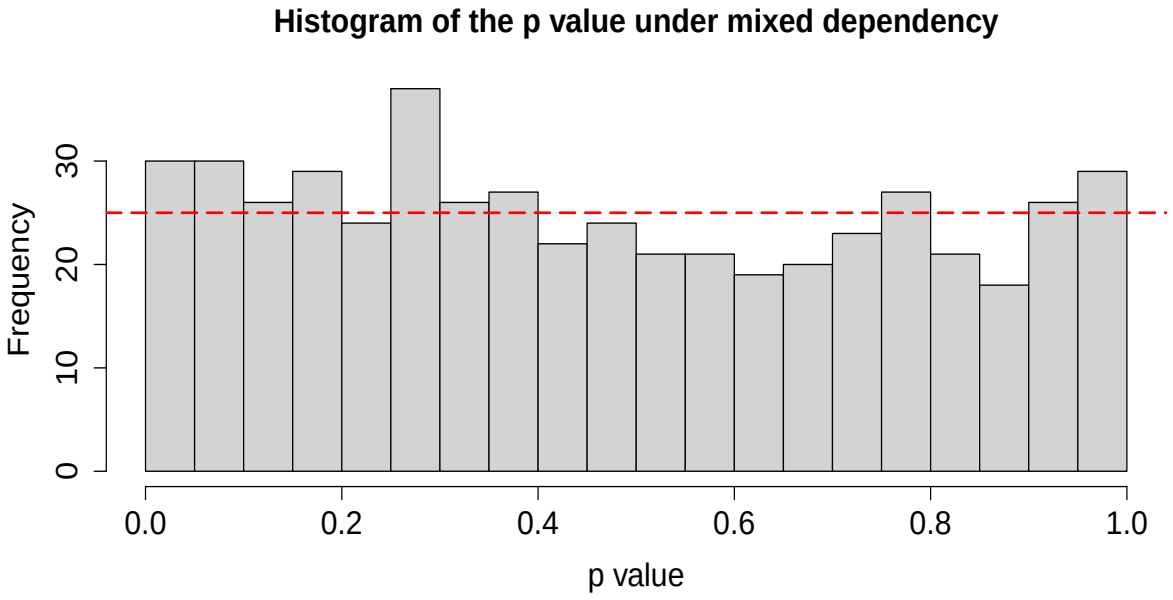


Figure 3.14: The histogram of the p-value from each of the 500 repetitive experiments under mixed dependencies for the Univariate Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

As shown, the 500 observed p-values are approximately uniformly distributed over $[0, 1]$ with a slight skewness towards 1, indicating a larger p is more likely to be observed in the experiments and the null hypothesis H_0 for the mixed independency is favoured.

Within the Bayesian framework, we evaluate the posterior contour probabilities of the 24 coefficients in $\boldsymbol{\beta}_3$ associated with contemporaneous independence and the 12 coefficients in $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ associated with Granger non-causality, yielding a total of 36 regression coefficients. These coefficients correspond to the potential contemporaneous independence

between $Z_3(t)$ and $Z_1(t), Z_2(t)$, as well as the potential Granger causality from $Z_3(t-1)$ to $Z_1(t)$ and $Z_2(t)$. The independence relationships hold if the joint vector of these 36 coefficients is centred at $\mathbf{0}$. Using the `bayesSurv` package, we find that the posterior contour probabilities exceed 0.5, providing sufficient evidence that these coefficients are centred at $\mathbf{0}$ and thereby supporting the mixed independence structures.

3.6 Discussion

In this chapter, we introduced a novel framework for modelling dependency structures among chain groups in an MMC: the *Univariate Marginal Model*. Figure 3.15 summarise the reparametrisation process within this model. It begins with the joint transition probabilities p used in the *Cartesian Product Model* (see Section 2.8), which naturally capture arbitrary dependency structures in the MMC. The Univariate Marginal Model factorizes these joint transition probabilities p into auxiliary probabilities ζ_σ, following a structure analogous to a Bayesian network under a specific permutation σ. Proposition 3.9 establishes that the map $\mathcal{P} \to \mathcal{H}_\sigma$ from the full parameter space $\mathcal{P}$ of joint transition matrices to the parameter space $\mathcal{H}_\sigma$ of auxiliary probabilities ζ_σ under the specified permutation σ is a diffeomorphism, thereby not only provides a smooth, invertible reparametrisation that is beneficial for both Frequentist and Bayesian inference but also ensures that the dependency structures encoded in the joint transition matrix P are equivalently represented within the auxiliary probabilities ζ_σ.

We subsequently incorporated multinomial logistic regression frameworks for the auxiliary probabilities ζ_σ under a specified permutation σ in the Univariate Marginal Model. These probabilities are thus further decomposed via a multinomial logistic link function into

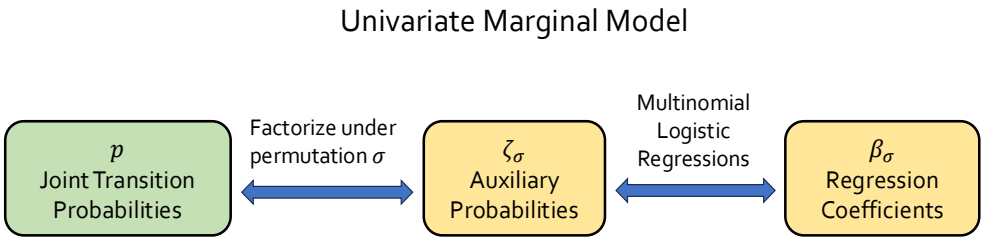


Figure 3.15: The summary of the parametrisations used in the *Univariate Marginal Model*. It begin with the joint transition probabilities p, followed by factorizes p into auxiliary probabilities ζ_σ based on a specified permutation σ. These ζ_σ are subsequently decomposed into the regression coefficients β_σ through a multinomial logistic regression.

linear combinations of marginal effects from all the marginal groups. This reparametrisation not only offers clear interpretability of these coefficients and enables straightforward encoding of dependency structures comparing to the Cartesian Product Model mentioned in 2.8.2, but also provides an alternative way to define the dependency structures through the corresponding regression coefficients (see Definition 3.15). It is worth noting that alternative link functions, such as the probit link, can also be employed, and the multinomial logistic link is merely one modelling choice among many.

This chapter then discusses the generative process under the Univariate Marginal Model, as described in Section 3.4. As given in Algorithm 1, this model admits a straightforward procedure for simulation, since the regression coefficients $\boldsymbol{\beta}_\sigma$ are fully unconstrained in this setting.

Finally, a comprehensive simulation study demonstrates that the regression coefficients in the Univariate Marginal Model can be accurately recovered under all four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures. The results show that hypothesis tests on the structures of the regression coefficients $\boldsymbol{\beta}$ from the Univariate Marginal Model provide a powerful tool for detecting and validating the potential dependency structures embedded within the MMC.

3.6.1 Strength and Limitations

The Univariate Marginal Model factorizes the joint transition probabilities p into a product of auxiliary probabilities ζ_σ, based on a predefined permutation σ. These auxiliary probabilities ζ_σ are then further decomposed into regression coefficients β_σ using multinomial logistic regression, thereby fully reparametrising the first-order MMC. This alternative parametrisation offers the following advantages:

- The Univariate Marginal Model defines a parametrisation $\boldsymbol{\beta}$ over an fully unconstrained space $\sigma \in \mathbb{R}^D$. Every $\boldsymbol{\beta}_\sigma \in \mathbb{R}^D$ yields a valid P, making this model

well-suited for generative purposes and allowing priors to be placed directly on $\boldsymbol{\zeta}_\sigma$ via the regression coefficients.

- The coefficients $\boldsymbol{\beta}_\sigma$ provide an explicit and direct mechanism for identifying and inferring any dependency structures. Both Granger non-causality and contemporaneous independence can be encoded by simply setting specific elements of $\boldsymbol{\beta}_\sigma$ to zero, thus avoiding the tedious summation and multiplication operations over the entries p required in the Cartesian representation.

- Additionally, the Univariate Marginal Model naturally corresponds to the adjacency matrices A, as shown in Equation 3.16, enabling us to visualize the dependencies between Markov chains in an explicit way.

Along with its advantages, the Univariate Marginal Model has a key limitation: its reliance on a predefined permutation σ of the chain groups (see Section 3.3.1). This dependence is manageable when prior knowledge of chain relationships is available—for example, when testing a specific structure, one can select an appropriate permutation, as illustrated in Section 3.3.2. However, each permutation captures only a subset of the dependency structures explicitly, while the remaining ones are represented only implicitly through further rearrangements. Hence, in some cases, no single σ can fully encode all the desired dependency structures explicitly. As a result, for generative purposes, the model may fail to produce MMCs that reflect particular combinations of dependencies, reducing flexibility. For inference multiple dependency structures, this issue might result in fitting multiple Univariate Marginal Models under different permutations, which increases computational burden.

Additionally, while the model uses only J regressions, which is a huge reduction compared to the Hierarchical Marginal Model, the number of covariates in the design matrix increases with each regression $\sigma(1), \sigma(2), \ldots, \sigma(J-1)$. This leads to the exponentially growing numbers of explanatory variables, resulting in increased computational cost, as will be elaborated in Section 4.6.2.

Chapter 4

Hierarchical Marginal Model

4.1 Introduction

As noted at the end of the previous chapter, the *Univariate Marginal Model* provides an unrestricted and efficient reparametrisation of the joint transition probabilities p through regression coefficients $\boldsymbol{\beta}_\sigma$, thereby offering a clear and interpretable way to detect dependency structures. Despite these advantages, the model relies heavily on a predefined permutation σ of the chain groups. Each permutation captures only a subset of the dependency structures explicitly, while the remainder are represented only implicitly through further rearrangements. This reliance reduces flexibility in generating MMCs that align with specific dependency structures and increases the computational burden when inferring them, since multiple models with different permutations are required to fit.

As a result, in parallel with the Univariate Marginal Model, we introduce in this chapter another novel marginal-based framework, the *Hierarchical Marginal Model*, which avoids the reliance on permutations and thereby overcomes these limitations. This model hierarchically constructs marginal transition probabilities, starting from univariate and extending up to K-variate levels. Similar to the Univariate Marginal Model, it employs multinomial logistic regression to decompose the marginal probabilities, ensuring an unconstrained parameter space and enhancing interpretability (see Section 4.3.1). Unlike the Univariate Marginal Model, where only a subset of dependency structures is explicitly

represented, the regression coefficients in the Hierarchical Marginal Model are capable of encoding the full range of dependency structures. The complete reparametrisation of the Hierarchical Marginal Model is illustrated in Figure 4.1.

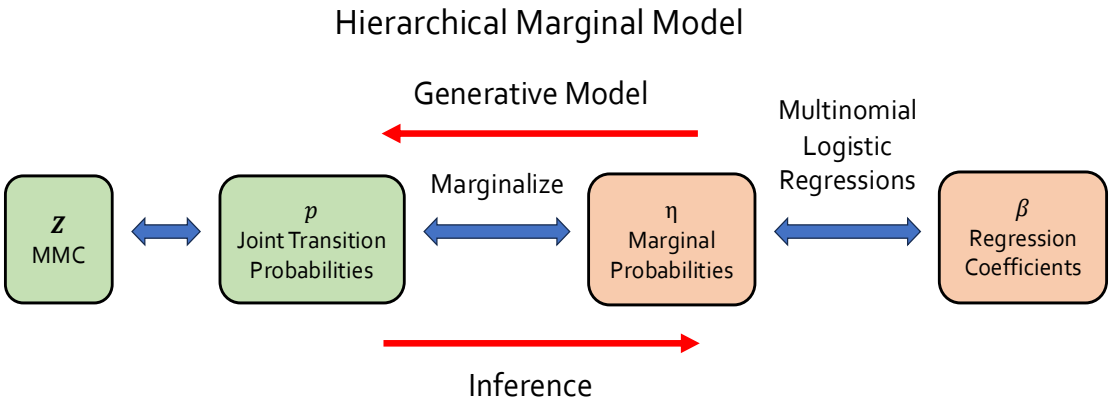


Figure 4.1: An illustration of the reparametrisation procedure of the *Hierarchical Marginal Model* is shown. It starts with hierarchically constructs marginal transition probabilities η from the univariate to the K-variates, followed by further decompose these η into the regression coefficients $\boldsymbol{\beta}_\sigma$ through a multinomial logistic regression. In the generative direction (the top arrow pointing left), the process is initiated with the regression coefficients $\boldsymbol{\beta}_\sigma$ and generates an MMC that exhibits the desired dependency structures. In contrast, the inference direction (the bottom arrow pointing right) starts with observed data from an MMC and infers the corresponding regression coefficients that encode the underlying dependency structures within the MMC.

In line with the previous chapter, we will also extend the discussion to the Hierarchical Marginal Model and examine its use from both generative and inference perspective, as illustrated in Figure 4.1.

4.1.1 Chapter Layout

This chapter develops in parallel with Chapter 3 and focuses on the *Hierarchical Marginal Model*. In Section 4.2, we introduce this framework and explain how it hierarchically constructs marginal transition probabilities η from the univariate to the K-variate level. The manner in which these marginal transition probabilities encode dependency structures is then detailed in Section 4.2.3.

Similar to the Univariate Marginal Model, we employ multinomial logistic regression, as reviewed in Section A.4, to further decompose the marginal transition probabilities in the Hierarchical Marginal Model into the corresponding regression coefficients, thereby achieving an unconstrained parameter space and enhancing interpretability. Section 4.3.1 introduces this decomposition, Section 4.3.2 explains how the resulting regression coefficients encode different dependency structures, and illustrative examples are provided in Section 4.3.3.

In applying the Hierarchical Marginal Model, Section 4.4 considers a generative model, introducing a hierarchical generating scheme in Algorithm 2, in accordance with the way Hierarchical Marginal Model constructs the marginal transition probabilities. Again, the corresponding inference methods have already been reviewed in Section A, including both Frequentist approaches (Section A.1) and Bayesian approaches (Sections A.2 and A.3).

Section 4.5 presents a simulation study, using the same setup as for the Univariate Marginal Model in Section 3.5.1. The corresponding simulation results are then presented in Section 4.5.1 under four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures. The objective is to assess the model's ability to estimate regression coefficients accurately and to recover the underlying dependency structures.

Finally, Section 4.6 evaluates the strengths and limitations of the proposed Hierarchical and Univariate Marginal Models, concluding with a comparative analysis among the different modelling approaches in Section 4.6.2.

4.2 Hierarchical Marginal Model

As discussed in Section 3.6.1, the *Univariate Marginal Model* relies heavily on a predefined permutation σ of the chain groups. Each permutation captures only a subset of the dependency structures explicitly, while the remaining ones are represented only implicitly through further rearrangements. Consequently, in some cases no single σ can directly encode all the desired dependency structures, creating difficulties for both generation and inference. To overcome this limitation, we introduce the *Hierarchical Marginal Model*, which represents MMCs through hierarchically constructed marginal transition probabilities η.

4.2.1 Formulation of the Hierarchical Marginal Model

Consider an MMC with k component chains, all sharing the same state space $\mathcal{S} = \{1, 2, \ldots, s\}$. Recall that for any subgroup of chains $g_j = \{k_1, \ldots, k_i\} \subseteq [K]$, we define the marginal transition probabilities of this subgroup as η_j according to Equation (2.39) as:

$$\eta_j = \mathbb{P}(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1)) = \mathbb{P}(Z_{k_1}(t), \ldots, Z_{k_i}(t) \mid \boldsymbol{Z}(t-1))$$

Concretely, given the state $\boldsymbol{b}_j = \{b_k : k \in g_j\} \in \mathcal{S}_j$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, the realization of the marginal transition probabilities $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$ is defined as:

$$\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}) = \mathbb{P}(\boldsymbol{Z}_j(t) = \boldsymbol{b}_j \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}) \tag{4.1}$$

By definition, this marginal transition probability η_j is obtained via marginalizing out all the chains k that is not in g_j:

$$\eta_j = \sum_{Z_k : k \notin g_j} \mathbb{P}(Z_1(t), \ldots, Z_K(t) \mid \boldsymbol{Z}(t-1))$$

As is usual in marginal models, we introduce the *strong compatibility* presented by Bergsma and Rudas [2002] that ensures the consistency of the marginal models:

Definition 4.1 (Strong Compatibility)**.** *The marginals $\eta_1, \eta_2, \ldots, \eta_J$ are said to be* ***strongly compatible*** *if there exists at least one joint transition probability matrix P that exactly matches all these given marginals simultaneously. Formally as:*

$$\exists \quad \mathbb{P}(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)) \in \mathcal{P} \qquad \textit{such that} \qquad \eta_j = \sum_{Z_k : k \notin g_j} \mathbb{P}(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1)) \quad \forall j \in [J]$$

where $\mathcal{P}$ denotes the parameter space for the joint transition probability.

In other words, *strong compatibility* ensures the existence of at least one joint transition distribution that aligns with all the specified marginal transition probabilities. This condition prevents scenarios where marginal distributions are only pairwise consistent but no single joint distribution satisfies all of them simultaneously. A toy example is given by considering an MMC with three chains, where all the bivariate marginals $\eta_{\{1,2\}}, \eta_{\{1,3\}}, \eta_{\{2,3\}}$ are specified. In this case, strong compatibility ensures that there exists a joint transition matrix P that exactly matches all three bivariate marginals.

Assuming that all the marginals are strongly compatible, we can then construct the Hierarchical Marginal Model as following:

1. We start with the univariate marginal, that is for all K different singleton $g_j = \{k_1\}$, and $\forall b_{k_1} \in \mathcal{S}, \boldsymbol{a} \in \boldsymbol{\mathcal{S}}, k_1 \in [K]$:

   $$\eta_{k_1}(b_{k_1} \mid \boldsymbol{a}) = \mathbb{P}(Z_{k_1}(t) = b_{k_1} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a})$$

   subject to the simplex constraint:

   $$\sum_{b_{k_1}=1}^{s} \eta_{k_1}(b_{k_1} \mid \boldsymbol{a}) = 1 \tag{4.2}$$

   The simplex constraint allows us only to model b_{k_1} that does not belongs to baseline level, $b_{k_1} \in \mathcal{S}^\star$, with $\mathcal{S}^\star = \mathcal{S} \setminus \{1\}$ denoting the state space excluding the baseline level.

2. We then model the bivariate marginal, where for $\binom{K}{2}$ different $g_{j_2} = \{k_1, k_2\}$ contains exactly two elements. Without loss of generality, we set $k_1 < k_2$ and choose to model $\forall b_{k_1}, b_{k_2} \in \mathcal{S}, \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$:

$$\eta_{j_2}(\boldsymbol{b}_{j_2} \mid \boldsymbol{a}) = \mathbb{P}(Z_{k_1}(t) = b_{k_1}, Z_{k_2}(t) = b_{k_2} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a})$$

When given all the univariate marginals in the previous step, the independent bivariate marginals η_{j_2} need to be modelled are only for $b_{k_1}, b_{k_2} \in \mathcal{S}^\star$. Since either $b_{k_1} = 1$ or $b_{k_2} = 1$ can be regarded as a baseline level and derived as:

$$\begin{aligned}
\eta_{j_2}\big((b_{k_1}, 1) \mid \boldsymbol{a}\big) &= \eta_{k_1}\big(b_{k_1} \mid \boldsymbol{a}\big) - \sum_{b_{k_2}=2}^{s} \eta_{j_2}\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big) \\
\eta_{j_2}\big((1, b_{k_2}) \mid \boldsymbol{a}\big) &= \eta_{k_2}\big(b_{k_2} \mid \boldsymbol{a}\big) - \sum_{b_{k_1}=2}^{s} \eta_{j_2}\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big)
\end{aligned} \tag{4.3}$$

3. Following that, we model trivariate marginals to K-variate marginals. For each i with $3 \leq i \leq K$, consider the $\binom{K}{i}$ different groups $g_{j_i} = \{k_1, \ldots, k_i\}$. For each such group and any $b_1, \ldots, b_i \in \mathcal{S}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$,

$$\eta_{j_i}(\boldsymbol{b}_{j_i} \mid \boldsymbol{a}) = \mathbb{P}(Z_{k_1}(t) = b_{k_1}, \ldots, Z_{k_i}(t) = b_{k_i} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a})$$

Again, only the states do not belong to the baseline level, $b_{k_1}, \ldots, b_{k_i} \in \mathcal{S}^\star$, require explicit modelling, because states in baseline level $a_{k_i} = 1$ can be derived from the lower-order marginals in a similar manner as Equation (4.3).

The strong compatibility by Bergsma and Rudas [2002] ensures that, at each step $i = 2, \ldots, K$, there exists a set of i-variate marginals that is consistent with the current $(i-1)$-variate marginals.

By construction, the number of independent parameters for marginal transition probabilities η_j considered by Hierarchical Marginal Model is given as:

$$
\begin{aligned}
&\underbrace{\binom{K}{1}(s-1)\times s^K}_{\text{univariate}} + \underbrace{\binom{K}{2}(s-1)^2\times s^K}_{\text{bivariate}} + \cdots + \underbrace{\binom{K}{K}(s-1)^K\times s^K}_{K\text{-variate}} \\
=&s^K\sum_{k=1}^{K}\binom{K}{k}(s-1)^K \\
=&s^K(s^K-1)
\end{aligned}
$$

which matches precisely the number of free parameters in a K-variate Markov chain with state space $\mathcal{S}$.

Example 4.2. Consider a multivariate Markov chain consisting three chains, each taking values in $\mathcal{S}=\{1,2\}$. Under the hierarchical marginal model, we directly parametrise the following non-baseline probabilities with $\forall \boldsymbol{a}\in\boldsymbol{\mathcal{S}}$:

Univariate Marginal: $\eta_k(2\mid \boldsymbol{a})=\mathbb{P}(Z_k(t)=2\mid \boldsymbol{Z}(t-1)=\boldsymbol{a}) \qquad \forall k\in\{1,2,3\}$

Bivariate Marginal: $\eta_{k,k'}((2,2)\mid \boldsymbol{a})=\mathbb{P}(Z_k(t)=2,Z_{k'}(t)=2\mid \boldsymbol{Z}(t-1)=\boldsymbol{a}) \qquad \forall k,k'\in\{1,2,3\}$

Trivariate marginal: $\eta_{1,2,3}((2,2,2)\mid \boldsymbol{a})=\mathbb{P}(Z_1(t)=2,Z_2(t)=2,Z_3(t)=2\mid \boldsymbol{Z}(t-1)=\boldsymbol{a})$

We have $\binom{3}{1}=3$ different univariate marginals, $\binom{3}{2}=3$ different bivariate marginals and $\binom{3}{3}=1$ trivariate marginal, which sums up to 7, timing all the $2^3=8$ different $\boldsymbol{b}\in\boldsymbol{\mathcal{S}}$, ending up with 56 targeting η to model, which is consistent with the number of independent parameters in the $2^3\times 2^3$ joint transition matrix P.

The map from the joint transition matrix P to the marginal transition probability matrix H:

Alternatively, we can build the matrix of η_j, denoted by H, through applying a matrix transformation to the joint transition matrix P. This involves a $\left(s^K-1\right)\times s^K$ marginal indicator matrix M, constructed by stacking rows of marginal indicators m. Specifically, for a k-variate Markov chain with state space $\mathcal{S}=\{1,2,\ldots,s\}$, the marginal indicator $m_j(\boldsymbol{b})$

aggregating the joint transition probabilities $p_{\boldsymbol{ab}}$ into $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$, where $g_j = \{k_1, \ldots, k_i\}$ is given as:

$$m_j(\boldsymbol{b}) = \bigotimes_{k=1}^{K} v(b_k)$$

$$\text{with} \quad v(b_k) = \begin{cases} \underbrace{\begin{pmatrix} 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^\top}_{\substack{\text{the } a_k\text{-th element} \\ \text{is set to be 1}}} & k \in g_j \\ \underbrace{\begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 \end{pmatrix}^\top}_{\text{a vector of 1}} & k \notin g_j \end{cases}$$

Here, $\bigotimes$ denotes the Kronecker product. For each chain k, the vector $v(b_k)$ has length s. If $k \in g_j$, $v(b_k)$ is the one-hot encoding of the state b_k; otherwise, it is a vector of all 1 which indicates the marginalization. The marginal indicator matrix M is then formed by stacking these $s^K - 1$ different row vectors $m_j(\boldsymbol{b})$ as:

$$M = \begin{pmatrix} \longleftarrow & m_1^\top(\boldsymbol{b}) & \longrightarrow \\ \longleftarrow & m_2^\top(\boldsymbol{b}) & \longrightarrow \\ & \vdots & \\ \longleftarrow & m_J^\top(\boldsymbol{b}) & \longrightarrow \end{pmatrix} \tag{4.4}$$

It follows that the marginal indicator matrix M is of dimension $s^{K-1} \times s^K$. Finally, the marginal transition matrix H for η_j is of $s^K \times s^{K-1}$ and can be written as a linear transformation of the joint transition matrix P with respect to M as:

$$H = PM^T \tag{4.5}$$

Example 4.3. Following the setup in Example 4.2, the corresponding marginal indicator matrix M of size $(2^3 - 1) \times 2^3$ is constructed by stacking the seven marginal indicator vectors $m_j(\boldsymbol{b})$, which represent all univariate marginals ($g_j = \{1\}, \{2\}, \{3\}$), bivariate

marginals ($g_j = \{1,2\},\{1,3\},\{2,3\}$), and the full trivariate marginal ($g_j = \{1,2,3\}$):

$$M = \begin{array}{r} \eta_{\{1\}}(2|\cdot) \\ \eta_{\{2\}}(2|\cdot) \\ \eta_{\{3\}}(2|\cdot) \\ \eta_{\{1,2\}}((2,2)|\cdot) \\ \eta_{\{1,3\}}((2,2)|\cdot) \\ \eta_{\{2,3\}}((2,2)|\cdot) \\ \eta_{\{1,2,3\}}((2,2,2)|\cdot) \end{array} \begin{pmatrix} (0 & 1) \otimes (1 & 1) \otimes (1 & 1) \\ (1 & 1) \otimes (0 & 1) \otimes (1 & 1) \\ (1 & 1) \otimes (1 & 1) \otimes (0 & 1) \\ (0 & 1) \otimes (0 & 1) \otimes (1 & 1) \\ (0 & 1) \otimes (1 & 1) \otimes (0 & 1) \\ (1 & 1) \otimes (0 & 1) \otimes (0 & 1) \\ (0 & 1) \otimes (0 & 1) \otimes (0 & 1) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

And the corresponding $2^3 \times (2^3 - 1)$ marginal transition matrix H follows the linear transformation in Equation (4.5) can be constructed as:

$$\begin{array}{r} \\ \boldsymbol{a}_1{=}(1,1,1) \\ \boldsymbol{a}_2{=}(1,1,2) \\ \boldsymbol{a}_3{=}(1,2,1) \\ \boldsymbol{a}_4{=}(1,2,2) \\ \boldsymbol{a}_5{=}(2,1,1) \\ \boldsymbol{a}_6{=}(2,1,2) \\ \boldsymbol{a}_7{=}(2,2,1) \\ \boldsymbol{a}_8{=}(2,2,2) \end{array} \begin{array}{c} \begin{array}{ccccccc} \eta_1(2|\cdot) & \eta_2(2|\cdot) & \eta_3(2|\cdot) & \eta_{1,2}((2,2)|\cdot) & \eta_{1,3}((2,2)|\cdot) & \eta_{2,3}((2,2)|\cdot) & \eta_{1,2,3}((2,2,2)|\cdot) \end{array} \\ \begin{pmatrix} \eta_1(2 \mid \boldsymbol{a}_1) & \eta_2(2 \mid \boldsymbol{a}_1) & \eta_3(2 \mid \boldsymbol{a}_1) & \cdots & & & \\ \eta_1(2 \mid \boldsymbol{a}_2) & \eta_2(2 \mid \boldsymbol{a}_2) & \eta_3(2 \mid \boldsymbol{a}_2) & \cdots & & & \\ \vdots & \vdots & \vdots & \ddots & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{pmatrix} \end{array}$$

Equation (4.5) defines the forward map from the joint transition probabilities p to the

marginal transition probabilities η. To recover the joint probabilities from the marginals, we consider the inverse map:

$$P = H(M^\top)^{-1} \tag{4.6}$$

Here, $(M^\top)^{-1}$ denotes the pseudo-inverse of the marginal indicator matrix $M^\top$, satisfying $(M^\top)^{-1}M^\top = I_{s^\kappa - 1}$ for non-squared matrix $M^\top$. This inversion is valid only under the assumption of *strong compatibility* given in Definition 4.1, which requires that the given marginals are mutually consistent and can be derived from some valid joint distribution. If this condition is violated, or in other words if these marginals transition probabilities η are incompatible, then there will be no joint transition matrix P that corresponds to them.

The *strong compatibility* will in turn impose some constraints on the marginal transition probabilities η. The following Example 4.4 illustrates a simple scenario in which a set of marginal transition probabilities η does not satisfy *strong compatibility* and therefore cannot be mapped back to a valid joint distribution.

Example 4.4. Consider the simplest scenario with two chains Z_1 and Z_2, each taking values from the binary state space $\mathcal{S} = \{1, 2\}$. Given a previous state $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$ and consider the corresponding row $\boldsymbol{a}$ of the transition probability matrix P, suppose we specify the following marginal transition probabilities with the abbreviated notation:

$$\eta_1 = \eta_1(2 \mid \boldsymbol{a}) = 0.6, \quad \eta_2 = \eta_2(2 \mid \boldsymbol{a}) = 0.3, \quad \eta_{1,2} = \eta_{1,2}((2,2) \mid \boldsymbol{a}) = 0.4$$

Attempting to reconstruct the joint transition probabilities p from these marginals, we obtain:

$$\begin{aligned}
\mathbb{P}\big((1,1) \mid \boldsymbol{a}\big) &= 1 - \eta_1 - \eta_2 + \eta_{1,2} = 0.7, \\
\mathbb{P}\big((1,2) \mid \boldsymbol{a}\big) &= \eta_2 - \eta_{1,2} = -0.1, \\
\mathbb{P}\big((2,1) \mid \boldsymbol{a}\big) &= \eta_1 - \eta_{1,2} = 0.2, \\
\mathbb{P}\big((2,2) \mid \boldsymbol{a}\big) &= \eta_{1,2} = 0.4.
\end{aligned}$$

However, this results in an invalid joint distribution since $\mathbb{P}((1,2) \mid \boldsymbol{a})$ is negative. This

indicates a violation of the *strong compatibility* condition, where the specified marginals do not correspond to any valid joint distribution.

4.2.2 Constraints on Marginal Transition Probabilities

In this section, we discuss the potential constraints on the marginal transition probabilities in detail. For two chains with binary states, the constraints on the marginals imposed by *strong compatibility* can be explicitly presented as:

$$\max(0, \eta_1 + \eta_2 - 1) \;\leq\; \eta_{1,2} \;\leq\; \min(\eta_1, \eta_2) \tag{4.7}$$

where η_k abbreviates $\eta_k(2 \mid \boldsymbol{a})$. Geometrically, these constraints imposed by *strong compatibility* reduce the valid parameter space of the marginal transition probabilities $\big(\eta_1, \eta_2, \eta_{1,2}\big)$ from the cube $[0,1]^3$ down to a pyramid P_{GABD}, as depicted in Figure 4.2.

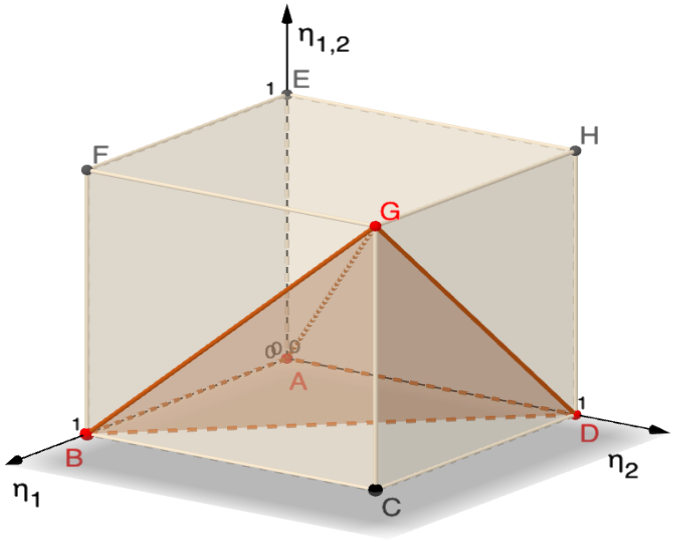


Figure 4.2: An illustration of the parameter space of the marginals modelled by the Hierarchical Marginal Model in the case of two chains with binary states. The coordinates are chosen as $\big(\eta_1, \eta_2, \eta_{1,2}\big)$. The original parameter space for these three marginal transition probabilities is the unit cube $[0,1]^3$, shown in gray. Imposing the *strong compatibility* condition reduces this space to a pyramid P_{GABD}, highlighted in red, defined by the vertices G, A, B, and D.

- The lower boundary surfaces S_{ABD} and S_{GBD} correspond to $\eta_{1,2} = 0$ and $\eta_{1,2} = \eta_1 + \eta_2 - 1$, respectively.

- The upper boundary surfaces S_{ABG} and S_{ADG} correspond to $\eta_{1,2} = \eta_2$ and $\eta_{1,2} = \eta_1$, respectively.

Importantly, imposing the *strong compatibility* condition does not reduce the dimensionality of the parameter space, as both the original cube and the constrained pyramid remain three-dimensional. To sum up, only those marginal probabilities η lying within the pyramid P_{GABD} fulfill the *strong compatibility* condition, and hence can be obtained from a valid joint transition probability matrix P.

These constraints implied by *strong compatibility* can also be viewed through the marginal indicator matrix M. In the forward map $\mathcal{P} \to \mathcal{H}$ defined by Equation 4.5, with $\mathcal{P}$ and $\mathcal{H}$ denote the parameter spaces for joint transition matrix P and the marginal transition matrix H respectively, the operation is straightforward since M consists only of 1 and 0, ensuring valid marginal transition probabilities. However, for the inverse map $\mathcal{H} \to \mathcal{P}$ defined by Equation (4.6), the calculation involves the pseudo-inverse $(M^\top)^{-1}$ of the marginal indicator matrix, which contains negative entries -1. Consequently, the inverse map can yield negative elements in the resulting joint transition matrix P, indicating no joint distribution matches these marginals, and thus implying these marginals are not strongly compatible.

In the general case with K chains and s possible states, the constraints imposed by *strong compatibility* can also be constructed hierarchically:

- For univariate marginals η_k with $k \in [K]$, only the simplex constraints in Equation (4.2) applies and no additional minimum or maximum constraints are needed.

- For bivariate marginals $\eta_{\{k_1,k_2\}}$ with $k_1, k_2 \in [K]$, the simplex constraints in Equation (4.3) still apply, but additional boundaries induced by the univariate marginals

are required to ensure compatibility. This involves minimization and maximization and can be explicitly written as:

$$\max\big[0,\ \eta_{k_1}(b_{k_1} \mid \boldsymbol{a}){+}\eta_{k_2}(b_{k_2} \mid \boldsymbol{a}){-}1\big] \ \leq\ \eta_{1,2}\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big) \ \leq\ \min\big[\eta_{k_1}(b_{k_1} \mid \boldsymbol{a}),\ \eta_{k_2}(b_{k_2} \mid \boldsymbol{a})\big]$$

for each $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$ and $b_{k_1}, b_{k_2} \in \mathcal{S}^\star = \mathcal{S} \setminus \{1\}$. Note that the constraints in Equation (4.7), demonstrated in the binary two-chain example, are a special case of this general form.

- Consider the general i-variate marginals η_j with $g_j = \{k_1, \ldots, k_i\}$ denoting the corresponding i-variate chain group. The boundaries on the i-variate marginal $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$, induced by the univariate to $(i{-}1)$-variate marginals, can be derived using the Inclusion-Exclusion formula:

  Remark 4.5 (Inclusion-Exclusion formula)**.** [Stanley, 2011] Given a function f defined on all the subsets of a finite partially ordered set S. Let $U \subseteq S$ be a subset and define $\phi\big(f(U)\big)$ as the sum f over all the sets V including U:

$$\phi\big(f(U)\big) = \sum_{V \supseteq U} f(V) \qquad \text{for all } \ U \subseteq S$$

  Then, the Inclusion Exclusion formula states that the inverse ϕ^{-1} exists and it can be explicitly written as:

$$\phi^{-1}\big(f(U)\big) = \sum_{V \supseteq U} (-1)^{|V-U|} f(V) \qquad \text{for all } \ U \subseteq S \tag{4.8}$$

  As shown by Rota [1964], this is a special case of the Möbius inversion formula over the Boolean lattice.

  Consider the i-variate marginal $\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big)$ in our setting. Let the finite partially ordered set S be g_j. For each subset $g_{j'} \subseteq g_j$, define $f(g_{j'})$ as the marginal probability corresponding to the configuration of $\boldsymbol{b}_j$ restricted to the chains $k_{i'} \in g_{j'}$,

conditioned on $\boldsymbol{a}$:

$$f(g_{j'}) = \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) = \mathbb{P}\big(\boldsymbol{Z}_{j'}(t) = \boldsymbol{b}_{j'} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) \quad \text{for any } g_{j'} \subseteq g_j$$

Substituting this definition of f into Equation (4.8) yields:

$$\bigcup_{g_{j'} \in g_j} \mathbb{P}\big(\boldsymbol{Z}_{j'}(t) = \boldsymbol{b}_{j'} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) = \sum_{g_{j'} \subseteq g_j} (-1)^{|g_{j'}|-1}\, \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big)$$

Applying De Morgan's laws then gives:

$$\begin{aligned}\mathbb{P}\big(\bigcap_{k \in g_j} Z_k(t) \neq b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) &= 1 - \sum_{g_{j'} \subseteq g_j} (-1)^{|g_{j'}|-1}\, \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) \\ &= (-1)^{|g_{j'}|}\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) + 1 - \sum_{g_{j'} \subset g_j} (-1)^{|g_{j'}|-1}\, \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big)\end{aligned}$$

– For even i, the lower bound arises from ensuring that the left-hand side probability remains non-negative:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) \geq \sum_{g_{j'} \subset g_j} (-1)^{|g_{j'}|-1}\, \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) - 1$$

Thus, the complete lower bound is given by:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) \geq \max\left[\, 0\,,\ \sum_{g_{j'} \subset g_j} (-1)^{|g_{j'}|-1}\, \eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) - 1 \right] \tag{4.9}$$

On the other hand, consider the inequality that the i-variate marginal is bounded above by all $(i-1)$-variate marginals:

$$\mathbb{P}\big(\bigcap_{k \in g_j} Z_k(t) \neq b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) \leq \min_{\substack{g_{j'} \subset g_j \\ |g_{j'}| = i-1}} \left[\mathbb{P}\big(\bigcap_{k' \in g_{j'}} Z_{k'}(t) \neq b_{k'} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big) \right]$$

Applying the Inclusion-Exclusion formula to the probabilities within the min function yields:

$$\mathbb{P}\Big(\bigcap_{k'\in g_{j'}} Z_{k'}(t) \neq b_{k'} \mid \boldsymbol{Z}(t-1)=\boldsymbol{a}\Big) = 1 - \sum_{g_{j''}\subseteq g_{j'}} (-1)^{|g_{j''}|-1}\,\eta_{j''}\big(\boldsymbol{b}_{j''} \mid \boldsymbol{a}\big)$$

Combining and canceling terms lead to the final upper bound:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) \leq \min_{\substack{g_{j'}\subset g_j \\ |g_{j'}|=i-1}} \left[\, 1\,,\ \sum_{\substack{g_{j''}\subset g_j \\ g_j\setminus g_{j'}\subseteq g_{j''}}} (-1)^{|g_{j''}|-1}\,\eta_{j''}\big(\boldsymbol{b}_{j''} \mid \boldsymbol{a}\big)\right] \tag{4.10}$$

– For the odd i, the corresponding bounds can be obtained by multiplying both sides of the even-case boundary by -1, resulting in:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) \leq \min\left[\, 1\,,\ \sum_{g_{j'}\subset g_j} (-1)^{|g_{j'}|}\,\eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) + 1\right] \tag{4.11}$$

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) \geq \max_{\substack{g_{j'}\subset g_j \\ |g_{j'}|=i-1}} \left[\, 0\,,\ \sum_{\substack{g_{j''}\subset g_j \\ g_j\setminus g_{j'}\subseteq g_{j''}}} (-1)^{|g_{j''}|}\,\eta_{j''}\big(\boldsymbol{b}_{j''} \mid \boldsymbol{a}\big)\right] \tag{4.12}$$

Example 4.6. Consider the trivariate marginal with $i=3$, and without loss of generality, let $g_j=\{1,2,3\}$. For simplicity, we use the abbreviated notation $\eta_{j'}\big(\boldsymbol{b}_{j'} \mid \boldsymbol{a}\big) = \eta_{j'}$ for $g_{j'}\subseteq g_j$. Then, the upper bound for η_j can be derived using Equation (4.11) as:

$$\begin{aligned}\eta_j &\leq \min\left[\, 1\,, 1 - \sum_{|g_{j'}|=1} \eta_{j'} + \sum_{|g_{j'}|=2} \eta_{j'}\right] \\ &= \min\big[\, 1\,, 1-\eta_1-\eta_2-\eta_3+\eta_{\{1,2\}}+\eta_{\{1,3\}}+\eta_{\{2,3\}}\big]\end{aligned}$$

For the lower bound, we take the maximum over subsets $g_{j'}=\{1,2\},\{1,3\},\{2,3\}$. The corresponding collections of $g_{j''}$ for each case are:

- For $g_{j'}=\{1,2\}$: $g_{j''}\in\big\{\{3\},\{1,3\},\{2,3\}\big\}$,
- For $g_{j'}=\{1,3\}$: $g_{j''}\in\big\{\{2\},\{1,2\},\{2,3\}\big\}$,

- For $g_{j'} = \{2,3\}$: $g_{j''} \in \big\{\{1\},\{1,2\},\{1,3\}\big\}$.

Plugging these into Equation (4.12) yields:

$$\eta_j \geq \max\Big[\, 0\ ,\ \eta_{\{1,2\}} + \eta_{\{1,3\}} - \eta_1,\ \eta_{\{1,2\}} + \eta_{\{2,3\}} - \eta_2,\ \eta_{\{1,3\}} + \eta_{\{2,3\}} - \eta_3\Big]$$

With this boundary clearly defined, we can now construct the diffeomorphism between the joint transition probabilities and the marginal transition probabilities. Let $\mathcal{H}$ denote the entire parameter space of the marginal transition matrices H, and define the restricted parameter space $\mathcal{H}_1$ as the subset of $\mathcal{H}$ consisting of marginal transition matrices satisfying *strong compatibility*:

$$\mathcal{H}_1 = \big\{H \in \mathcal{H} : \quad \exists\, P \in \mathcal{P} \quad \text{such that} \quad H = PM^\top\big\} \tag{4.13}$$

where M is the marginal indicator matrix constructed according to Equation (4.4). Then, the linear map $\mathcal{P} \to \mathcal{H}_1$ defined by Equation (4.5) is a diffeomorphism, which is formalized in the following Proposition 4.7:

Proposition 4.7. *The linear transformation $\mathcal{P} \to \mathcal{H}_1$, defined via the marginal indicator matrix M in Equation (4.5), establishes a diffeomorphism between the entire parameter spaces of the joint transition matrices $\mathcal{P}$ and the marginal transition matrices $\mathcal{H}_1$ satisfying strong compatibility.*

Proof. Recall from Equation (4.4) the construction of the marginal indicator matrix M. Each element vector v in this construction is either a one-hot encoding of a particular state b_k, excluding the baseline level, or a vector of ones, and thus these vectors are linearly independent. Consequently, their Kronecker products $m_j(\boldsymbol{b})$, which form the row vectors of M, remain linearly independent. Hence, the marginal indicator matrix M has full row rank, guaranteeing that the linear map $\mathcal{P} \to \mathcal{H}_1$ is bijective. Additionally, as a linear map, it is inherently differentiable with a differentiable inverse. Therefore, the map $\mathcal{P} \to \mathcal{H}_1$ constitutes a diffeomorphism.

□

This diffeomorphism establishes a smooth bijective map between the joint transition probability matrix P and the corresponding marginal transition matrix H satisfying strong compatibility, and will be used to show that the overall transformation from P to the regression coefficients $\boldsymbol{\beta}$ under the Hierarchical Marginal Model is a diffeomorphism in Lemma 4.12.

4.2.3 Encode Dependency Structures

Since the map $\mathcal{P} \to \mathcal{H}_1$ is a diffeomorphism, the marginal transition probabilities η_j within H fully encodes all the information in the joint transition matrix P. As a result, it enables us to impose Granger non-causality and Contemporaneous Independence Assumption in a more convenient fashion to P via the marginal transition probabilities $\eta_{j'}$ through the following proposition:

Proposition 4.8. *We can directly impose arbitrary Granger non-causality and contemporaneous independence relationship through constraining the marginal transition probabilities $\eta_{j'}$ modelled by hierarchical marginal model.*

Proof. We consider the contemporaneous independence and Granger non-causality for each pair of non-intersecting chain groups g_{j_1} and g_{j_2}, (i.e. $g_{j_1}, g_{j_2} \subset [k]$ and $g_{j_1} \cap g_{j_2} = \emptyset$):

- For an arbitrary contemporaneous independence relationship with chain groups g_{j_1} and g_{j_2}:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}(t-1) \tag{4.14}$$

  We have the equivalent product conditions: $\forall \boldsymbol{b}_{j_1} \in \boldsymbol{\mathcal{S}}_{j_1}, \boldsymbol{b}_{j_2} \in \boldsymbol{\mathcal{S}}_{j_2}, \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$:

$$\eta_{j_1}(\boldsymbol{b}_{j_1} \mid \boldsymbol{a}) \times \eta_{j_2}(\boldsymbol{b}_{j_2} \mid \boldsymbol{a}) = \eta_{j_1 \cup j_2}\big((\boldsymbol{b}_{j_1}, \boldsymbol{b}_{j_2}) \mid \boldsymbol{a}\big)$$

  This can be translate into setting $\forall g_{j_1'} \subseteq g_{j_1}, g_{j_2'} \subseteq g_{j_2}$ and $\forall \boldsymbol{b}_{j_1'} \in \boldsymbol{\mathcal{S}}^{\star}_{j_1'}, \boldsymbol{b}_{j_2'} \in \boldsymbol{\mathcal{S}}^{\star}_{j_2'}$:

$$\eta_{j_1'}(\boldsymbol{b}_{j_1'} \mid \boldsymbol{a}) \times \eta_{j_2'}(\boldsymbol{b}_{j_2'} \mid \boldsymbol{a}) = \eta_{j_1' \cup j_2'}\big((\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2}) \mid \boldsymbol{a}\big) \tag{4.15}$$

  where $\boldsymbol{\mathcal{S}}^{\star} = \boldsymbol{\mathcal{S}} \backslash 1$ denote the state space excluding the baseline level 1, as modelled by

the Hierarchical Marginal Model. The baseline level is automatically satisfied due to the simplex constraint on the marginal transition probabilities η_j in Equation (4.2). These constraints in Equation (4.15) can be enforced by requiring that the columns $\eta_{j_1' \cup j'2}\big((\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'}) \mid \cdot\big)$ equal the Hadamard product of $\eta_{j_1'}(\boldsymbol{b}_{j_1'} \mid \cdot)$ and $\eta_{j_2'}(\boldsymbol{b}_{j_2'} \mid \cdot)$, both of which are specified by the marginal transition model. Formally, for each index i in the column vector, the Hadamard product is defined as:

$$\big[\eta_{j_1' \cup j_2'}\big((\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'}) \mid \cdot\big)\big]_i = \big[\eta_{j_1'}(\boldsymbol{b}_{j_1'} \mid \cdot)\big]_i \times \big[\eta_{j_2'}(\boldsymbol{b}_{j_2'} \mid \cdot)\big]_i \qquad \forall i = 1, 2, \ldots, s^K$$

ensuring the equality of all elements across every possible $\forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$

- For an arbitrary Granger non-causality that chain group g_{j_2} does not cause g_{j_1} given as:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t-1) \mid \boldsymbol{Z}_{-j_2}(t-1) \tag{4.16}$$

We remove $\boldsymbol{Z}_{j_2}(t-1)$ from the conditioning in η_{j_1}. Concretely,

$$\eta_{j_1} = \mathbb{P}(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}(t-1)) = \mathbb{P}(\boldsymbol{Z}_{j_1}(t) \mid \boldsymbol{Z}_{-j_2}(t-1))$$

which implies that for all $\boldsymbol{b}_{j_1} \in \boldsymbol{\mathcal{S}}_{j_1}$ and any distinct $\boldsymbol{a}_{j_2}, \boldsymbol{a}'_{j_2} \in \boldsymbol{\mathcal{S}}_{j_2}$,

$$\eta_{j_1}\big(\boldsymbol{b}_{j_1} \mid (\boldsymbol{a}_{j_2}, \boldsymbol{a}_{-j_2})\big) = \eta_{j_1}\big(\boldsymbol{b}_{j_1} \mid (\boldsymbol{a}'_{j_2}, \boldsymbol{a}_{-j_2})\big)$$

Equivalently, it can be translated into enforcing

$$\eta_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid (\boldsymbol{a}_{j_2}, \boldsymbol{a}_{-j_2})\big) = \eta_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid (\boldsymbol{a}'_{j_2}, \boldsymbol{a}_{-j_2})\big)$$

for all $g_{j_1'} \subseteq g_{j_1}$ and $\boldsymbol{b}_{j_1'} \in \boldsymbol{\mathcal{S}}^{\star}_{j_1'}$ where $\boldsymbol{\mathcal{S}}^{\star}_{j_1'}$ denotes the state space for $\boldsymbol{b}_{j_1'}$ excluding the baseline state 1. Practically, this is done by forcing the elements within each columns $\eta_{j_1'}(\boldsymbol{b}_{j_1'} \mid \cdot)$ of H to take the same values for distinct $\boldsymbol{a}_{j_2}$ and $\boldsymbol{a}'_{j_2}$, so long as $\boldsymbol{a}_{-j_2}$ matches.

□

Example 4.9. Following the setting in Example 4.2, let $j_1 = \{1\}$ and $j_2 = \{2, 3\}$:

- Enforcing contemporaneous independence (4.14) is equivalent as setting the following column constraints within H, that is $\forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$:

$$\eta_{\{1\}}(2 \mid \boldsymbol{a}) \times \eta_{\{2\}}(2 \mid \boldsymbol{a}) = \eta_{\{1,2\}}\big((2,2) \mid \boldsymbol{a}\big) \tag{4.17}$$

$$\eta_{\{1\}}(2 \mid \boldsymbol{a}) \times \eta_{\{3\}}(2 \mid \boldsymbol{a}) = \eta_{\{1,3\}}\big((2,2) \mid \boldsymbol{a}\big) \tag{4.18}$$

$$\eta_{\{1\}}(2 \mid \boldsymbol{a}) \times \eta_{\{2,3\}}\big((2,2) \mid \boldsymbol{a}\big) = \eta_{\{1,2,3\}}\big((2,2,2) \mid \boldsymbol{a}\big) \tag{4.19}$$

  Equations (4.17) and (4.18) ensure the element-wise independence, that is $Z_1(t)$ being independent of $Z_2(t)$ and $Z_3(t)$ respectively given $\boldsymbol{Z}(t-1)$. Additionally, including the condition that chains 2 and 3 are not in the baseline state in (4.19) ensures their joint presence does not violate $Z_1(t)$'s independence from them, preserving the contemporaneous independence relationship in the group level.

- Imposing Granger non-causality (4.16) translates into the requirement that, for every possible (a_1, a_2, a_3) and (a_1, a_2', a_3') within column $\eta_{\{1\}}(2 \mid \cdot)$,

$$\eta_{\{1\}}\big(2 \mid (a_1, a_2, a_3)\big) = \eta_{\{1\}}\big(2 \mid (a_1, a_2', a_3')\big) \qquad \forall a_1, a_2, a_2', a_3, a_3' \in \{1, 2\} \tag{4.20}$$

  In other words, $\eta_{\{1\}}\big(2 \mid \boldsymbol{a}\big)$ is invariant for distinct values of (a_2, a_3) and (a_2', a_3'), as long as a_1 stays the same, thereby demonstrating the Granger non-causality.

4.3 Multinomial Logistic Regressions

The previous Section 4.2 showed that reparametrising the joint transition probability p through marginal transitions preserves all dependency structures in P via the bijection $H \mapsto P$, while also providing explicit representations of structures such as Granger non-causality and contemporaneous independence in first-order MMCs. In parallel with the Univariate Marginal Model, we further decompose these marginal transition probabilities within the Hierarchical Marginal Model using the multinomial logistic regression to enhance interpretability.

4.3.1 Formulations of the Multinomial Logistic Regressions

In contrast to the Univariate Marginal Model, where the design matrix varies across regressions due to the increasing number of conditioning terms for the chain groups $g_{\sigma(1)}, \ldots, g_{\sigma(J)}$, the multinomial logistic regressions in the Hierarchical Marginal Model are constructed hierarchically with a fixed design matrix. This approach avoids the exponential growth in response levels but instead requires fitting a larger number of individual regressions, as detailed below.

As discussed in Section 4.2, the Hierarchical Marginal Model builds up from univariate to K-variate marginals. The associated multinomial logistic regressions is structured accordingly, helping to manage the exponential growth in the number of response levels. Consider a first-order MMC $\boldsymbol{Z}$ composed of K chains, with each individual chain shares the same finite state space $\mathcal{S} = \{1, 2, \ldots, s\}$, the corresponding multinomial logistic regression for Hierarchical Marginal Model can be built as:

1. Starting with modelling the univariate marginal, that is for all K different singleton $g_j = \{k\}$, and $\forall b_k \in \mathcal{S}, \boldsymbol{a} \in \boldsymbol{\mathcal{S}}, k \in [K]$:

$$\eta_j(b_k \mid \boldsymbol{a}) = \mathbb{P}(Z_k(t) = b_k \mid \boldsymbol{Z}(t-1) = \boldsymbol{a})$$

We adopt the multinomial link function defined as:

$$\eta_j(b_k \mid \boldsymbol{a}) = \frac{e^{\varphi_j(b_k|\boldsymbol{a})}}{\sum e^{\varphi_j(\,\cdot\,|\boldsymbol{a})}} \tag{4.21}$$

The summation in the denominator is taken over all possible states $b_k \in \mathcal{S}$ for the chain k. The associated log odds, denoted by $\varphi_j(b_k \mid \boldsymbol{a})$, take on s different possible values, one for each distinct state $b_k \in \mathcal{S}$. The vector of the log odds $\boldsymbol{\varphi}_j(b_k \mid \cdot\,)$, consisting $\varphi_j(b_k \mid \boldsymbol{a})$ for all $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, can be modelled as the product of a design matrix $\boldsymbol{\mathcal{X}}$ and a corresponding regression coefficient vector $\boldsymbol{\beta}_{a_k}$:

$$\boldsymbol{\varphi}_j(b_k \mid \boldsymbol{a}) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{b_k} \tag{4.22}$$

$$\begin{pmatrix} \varphi_j(b_k \mid \boldsymbol{a}^{(1)}) \\ \varphi_j(b_k \mid \boldsymbol{a}^{(2)}) \\ \vdots \\ \varphi_j(b_k \mid \boldsymbol{a}^{(s_{\boldsymbol{a}})}) \end{pmatrix}_{s_{\boldsymbol{a}}\times 1} = \begin{pmatrix} \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(1)}) & \longrightarrow \\ \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(2)}) & \longrightarrow \\ & \vdots & \\ \longleftarrow & \boldsymbol{\mathcal{X}}^\top(\boldsymbol{a}^{(s_{\boldsymbol{a}})}) & \longrightarrow \end{pmatrix}_{s_{\boldsymbol{a}}\times s_{\boldsymbol{a}}} \begin{pmatrix} \beta_{b_k}^{(1)} \\ \beta_{b_k}^{(2)} \\ \vdots \\ \beta_{b_k}^{(s_{\boldsymbol{a}})} \end{pmatrix}_{s_{\boldsymbol{a}}\times 1} \tag{4.23}$$

- The vector $\boldsymbol{\varphi}_j(b_k \mid \cdot\,)$ is a column vector that consists all log odds $\varphi_j(b_k \mid \boldsymbol{a})$ for a given b_k and every $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$. Consequently, $\boldsymbol{\varphi}_j(b_k \mid \cdot\,)$ is of length $s_{\boldsymbol{a}} = s^k$, representing the cardinality of the state space for $\boldsymbol{\mathcal{S}}_{\boldsymbol{a}} = \boldsymbol{\mathcal{S}}$

- The $s_{\boldsymbol{a}} \times s_{\boldsymbol{a}}$ squared design matrix $\boldsymbol{\mathcal{X}}$ is constructed by stacking together $s_{\boldsymbol{a}}$ individual components $\boldsymbol{\mathcal{X}}(\boldsymbol{a})$, one for each $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$. The explicit construction of each $\boldsymbol{\mathcal{X}}(\boldsymbol{a})$ through the Kronecker product $\bigotimes$ is given as follows:

$$\boldsymbol{\mathcal{X}}(\boldsymbol{a}) = \bigotimes_{k=1}^{K} v(a_k) \tag{4.24}$$

$$\text{with} \quad v(a_k) = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 1 & \cdots & 0 \end{pmatrix}^\top}_{\substack{\text{First element and the } a_k\text{-th} \\ \text{element are set to be 1}}} \qquad a_k \in \mathcal{S}$$

where the vector $v(a_k)$ can be interpreted as the one-hot encoding of the state a_k, of length s with their first component fixed to 1.

- The column vector $\boldsymbol{\beta}_{b_k}$ represents the coefficients for the multinomial logistic regression model and is of length $s_{\boldsymbol{a}}$. These coefficients provide interpretability of the inter-chain effects. To ensure the identifiability, the coefficient $\boldsymbol{\beta}_1$ corresponding to the baseline category $b_k = 1$ is constrained to zero.

2. modelling the bivariate marginals. Consider the $\binom{K}{2}$ subsets $g_j = \{k_1, k_2\}$, each containing exactly two elements. Without loss of generality, assume $k_1 < k_2$. We focus on the non-baseline states, so for each $b_{k_1}, b_{k_2} \in \mathcal{S}^\star = \mathcal{S} \setminus \{1\}$ and $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$:

$$\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a}) = \mathbb{P}\big(Z_{k_1}(t) = b_{k_1},\, Z_{k_2}(t) = b_{k_2} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big).$$

Rather than modelling all s^2 possible responses jointly via one huge multinomial logistic regression, we partition each of η_{k_1} and build a corresponding multinomial logistic regression on that. Concretely, for each $b_{k_1} \in \mathcal{S}^\star$, we model $\eta_j\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big)$ based on the link function:

$$\eta_j\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big) = \underbrace{\eta_{k_1}\big(b_{k_1} \mid \boldsymbol{a}\big)}_{\text{from univariate marginal}} \cdot \frac{e^{\varphi_j\left((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\right)}}{\sum e^{\varphi_j\left((b_{k_1}, \,\cdot\,) \mid \boldsymbol{a}\right)}} \tag{4.25}$$

The summation in the denominator is taken over all possible states $b_{k_2} \in \mathcal{S}$. Such a construction ensures that the simplex constraint

$$\sum_{b_{k_2}=1}^{s} \eta_j\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big) \quad = \quad \eta_{j'}\big(b_{k_1} \mid \boldsymbol{a}\big).$$

will be automatically satisfied.

As in the case of univariate marginals, the log-odds $\varphi_j\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big)$ for the bivariate marginals are modelled in vector form. Specifically, the vector $\boldsymbol{\varphi}_j\big((b_{k_1}, b_{k_2}) \mid \,\cdot\,\big)$, which stacks $\varphi_j\big((b_{k_1}, b_{k_2}) \mid \boldsymbol{a}\big)$ over all $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, is expressed as the product of the design matrix $\boldsymbol{\mathcal{X}}$ and the corresponding regression coefficient vector $\boldsymbol{\beta}_{b_{k_1}, b_{k_2}}$.

$$\boldsymbol{\varphi}_j\big((b_{k_1}, b_{k_2}) \mid \cdot\big) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{b_{k_1},b_{k_2}} \qquad \forall \quad b_{k_1}, b_{k_2} \in \mathcal{S}^\star \tag{4.26}$$

The design matrix $\boldsymbol{\mathcal{X}}$ is the same as that in Equation (4.24) for modelling univariate marginals, indicating that the conditioning structure remains unchanged. Similarly, the vector of regression coefficients $\boldsymbol{\beta}_{b_{k_1},b_{k_2}}$ follows the same formulation as $\boldsymbol{\beta}_{b_k}$ in Equation (4.23), except that the univariate subscription b_k is replaced by the bivariate pair (b_{k_1}, b_{k_2}), reflecting the response structure of each bivariate group g_j.

That is, instead of fitting a single multinomial logistic regression with s^2 response levels, we fit $s-1$ separate multinomial logistic regressions, each standing for a different $\eta_{k_1}\big(b_{k_1} \mid \cdot\big)$. Each of these multinomial logistic regression involves s levels, including 1 baseline level, representing partitioning $\eta_{k_1}\big(b_{k_1} \mid \cdot\big)$ with respect to chain k_2 to obtain the corresponding marginal transition probability for each of the different bivariate group g_j.

In practice, η_j may be obtained by partitioning η_{k_1} with respect to k_2, as is demonstrated, or otherwise by partitioning η_{k_2} with respect to k_1. The *strong compatibility* condition by Bergsma and Rudas [2002] introduced in Definition 4.1 guarantees that the resulting bivariate marginals will be identical in either case, ensuring that the Hierarchical Marginal Model is free of any permutation as the Univariate Marginal Model does.

3. Extending to tri-variate and higher-order marginals, for each i such that $3 \leq i \leq K$, we consider the $\binom{K}{i}$ possible i-variate groups of chains,

$$g_j = \{k_1, k_2, \ldots, k_i\}$$

with the indices ordered as $k_1 < k_2 < \cdots < k_i$ without loss of generality. We focus on non-baseline state combinations, that is for $b_{k_1}, \ldots, b_{k_i} \in \mathcal{S}^\star$, and define the targeting i-variate marginal transition probabilities for any $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$ as:

$$\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big) = \mathbb{P}\big(Z_{k_1}(t) = b_{k_1}, \ldots, Z_{k_i}(t) = b_{k_i} \mid \boldsymbol{Z}(t-1) = \boldsymbol{a}\big). \tag{4.27}$$

Following the same approach as before, we partition each $\eta_{j'}$ associated with the $(i-1)$-variate group $g_{j'} = k_1, \ldots, k_{i-1}$ from the previous step. Concretely, for each $\big(b_{k_1}, \ldots, b_{k_{i-1}}\big) \in \mathcal{S}^\star$, we model i-variate marginal $\eta_j\big((b_{k_1}, \ldots, b_{k_i}) \mid \boldsymbol{a}\big)$ based on the link function:

$$\eta_j\big((b_{k_1}, \ldots, b_{k_i}) \mid \boldsymbol{a}\big) = \underbrace{\eta_{j'}\big((b_{k_1}, \ldots, b_{k_{i-1}}) \mid \boldsymbol{a}\big)}_{\text{from previous step}} \cdot \frac{e^{\varphi_j\left((b_{k_1}, \ldots, b_{k_i}) \mid \boldsymbol{a}\right)}}{\sum e^{\varphi_j\left((b_{k_1}, \ldots, b_{k_{i-1}}, \,\cdot\,) \mid \boldsymbol{a}\right)}} \tag{4.28}$$

The summation is over all the $b_{k_i} \in \mathcal{S}$ to ensure the satisfaction of the the simplex constraint:

$$\sum_{b_{k_i}=1}^{s} \eta_j\big((b_{k_1}, \ldots, b_{k_{i-1}}, b_{k_i}) \mid \boldsymbol{a}\big) \;=\; \eta_{j'}\big(b_{k_1}, \ldots, b_{k_{i-1}} \mid \boldsymbol{a}\big) \qquad \forall b_{k_1}, \ldots, b_{k_{i-1}} \in \mathcal{S}^\star$$

The corresponding log odds $\varphi_j\big((b_{k_1}, \ldots, b_{k_i}) \mid \boldsymbol{a}\big)$ is then modelled in the same way as Equation (4.22) and (4.26) with the same design matrix $\boldsymbol{\mathcal{X}}$ indicating the invariant conditioning structure and the same vector of regression coefficients $\boldsymbol{\beta}$ subject to a modification of the subscripts. Consequently, rather than having a single multinomial logistic regression with s^i levels, we employ $(s-1)^{i-1}$ multinomial logistic regressions each with s levels together with the $\eta_{j'}$ obtained in the previous step to model the i-variate marginal η_j given in Equation (4.27).

By organizing the multinomial logistic regressions hierarchically, we fix the number of response levels in each regression at s, rather than letting it grow from s^2 up to s^K in the Univariate Marginal Model. While this approach requires fitting a large number of individual regressions, specifically $s^K - 1$, as the compensation, it significantly reduces the overall computational burden during inference compared to non-hierarchical models, as will be discussed later in Section 4.6.2.

Consider the map $\mathcal{H} \to \mathcal{B}$ between the parameter spaces of the marginal transition matri-

ces $H \in \mathcal{H}$ and the regression coefficients $\boldsymbol{\beta} \in \mathcal{B}$. This transformation is a diffeomorphism, as can be verified by the following two-step decomposition:

- The map from the marginal transition matrices H to the log odds $\boldsymbol{\varphi}$ is a diffeomorphism: Specifically, this step is defined via multinomial logistic link functions, as given in Equations (4.21) and (4.28). These link functions are explicitly invertible and continuously differentiable, as are their inverses.

- The transformation from the log odds $\boldsymbol{\varphi}$ to the regression coefficients $\boldsymbol{\beta}$ is also a diffeomorphism: This relationship is established through a matrix multiplication involving the design matrix $\boldsymbol{\mathcal{X}}$, constructed according to Equation (4.24). Due to the construction, each one-hot encoding vector $v(a_k)$ representing state a_k is linearly independent. Consequently, the resulting Kronecker products $\boldsymbol{\mathcal{X}}(\boldsymbol{a})$ are also linearly independent, ensuring that the design matrix $\boldsymbol{\mathcal{X}}$ is of full rank. Therefore, the linear map between $\boldsymbol{\beta}$ and $\boldsymbol{\varphi}$ is a diffeomorphism.

However, as illustrated in Example 4.4, when consider the inverse map $\mathcal{H} \to \mathcal{P}$, not all the marginal transition matrices $H \in \mathcal{H}$ gives you a valid joint transition matrix $P \in \mathcal{P}$. Only H within the subset $\mathcal{H}_1 \subset \mathcal{H}$ defined under *strong compatibility* in Equation (4.13) ensures this inverse map yielding a valid P and is our target. The explicit boundary for the regression coefficients can be obtained by applying the link function in Equation (4.28) to transform the boundaries on the marginal probabilities η (see Equations (4.11) and (4.12) for odd cases and (4.9) and (4.10) for even cases) into the corresponding boundaries on the regression parameters β. The following example continues from Example 4.4, illustrating this transformation in the simple case of two chains with binary states:

Example 4.10. Recall Example 4.4 that the bivariate marginal $\eta_{\{1,2\}}$ is constrained by:

$$\max(0, \eta_1 + \eta_2 - 1) \;\leq\; \eta_{\{1,2\}} \;\leq\; \min(\eta_1, \eta_2)$$

Incorporating the link functions from Equations (4.21) and (4.25), we can derive the bounds on the bivariate regression coefficient $\beta_{\{1,2\}}$ in terms of the univariate coefficients

β_1 and β_2 as:

$$\text{logit}\Big(\max\Big[\,0\,,\ \frac{1-e^{-(\beta_1+\beta_2)}}{1+e^{-\beta_2}}\Big]\Big) \le \beta_{\{1,2\}} \le \text{logit}\Big(\min\Big[\,1\,,\ \frac{1+e^{-\beta_1}}{1+e^{-\beta_2}}\Big]\Big) \tag{4.29}$$

Note that β_1 and β_2 are unrestricted, since the corresponding univariate marginal probabilities η_1 and η_2 span the full interval $[0, 1]$. Geometrically, by choosing $(\beta_1, \beta_2, \beta_{\{1,2\}})$ as the coordinates, the upper and lower bounds in the inequality (4.29) define two surfaces cutting the 3D space. These are shown as the red (upper bound) and green (lower bound) surfaces in Figure 4.3. The feasible region for $\beta_{\{1,2\}}$ is sandwiched between these two surfaces, forming a constrained band in the 3D space.

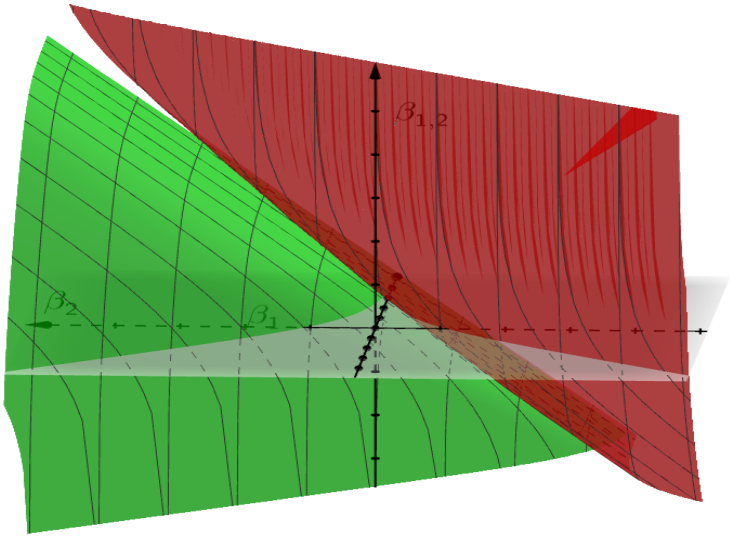


Figure 4.3: A illustration of the upper bound (in red) and lower bound (in green) of the regression coefficients $\beta_{\{1,2\}}$ of the bivariate marginal, as a function of the regression coefficients of the univariate marginals β_1, β_2 when we choose the coordinates to be $(\beta_1, \beta_2, \beta_{\{1,2\}})$.

As a result, we similarly define the subset $\mathcal{B}_1 \subset \mathcal{B}$ to be the subset of the entire parameter space for regression coefficients induced by the *strong compatibility* of marginal transition matrices H, (i.e. the constrained band located in 3D space in previous Example)

$$\mathcal{B}_1 = \big\{\boldsymbol{\beta} \in \mathcal{B} : \quad \exists\, H \in \mathcal{H}_1 \quad \text{such that} \quad \mathcal{F}(H) = \boldsymbol{\beta}\big\}, \tag{4.30}$$

for multinomial logistic transformation $\mathcal{F}$. By construction, the map $\mathcal{H}_1 \to \mathcal{B}_1$ is also a diffeomorphism, which is formalized as the following:

Proposition 4.11. *The map $\mathcal{H}_1 \to \mathcal{B}_1$ between the regression coefficients $\boldsymbol{\beta}$, within the induced parameter space $\mathcal{B}_1$ defined in Equation (4.30), to the marginal transition matrices H consisting of the marginal transition probabilities are strongly compatible within the parameter space defined in Equation (4.13), is a diffeomorphism.*

By combining Proposition 4.7 and 4.11, we yield the following Lemma 4.12 that demonstrates the diffeomorphism between the joint transition matrices $P \in \mathcal{P}$ and the regression coefficients $\boldsymbol{\beta}$ in the induced parameter space $\mathcal{B}_1$ constructed in Equation (4.30):

Lemma 4.12. *The transformation $g_1 : \mathcal{P} \to \mathcal{B}_1$ from the joint transition matrices P within the full parameter space $\mathcal{P}$ to the regression coefficients $\boldsymbol{\beta}$ defined in the induced parameter space $\mathcal{B}_1$ by $\mathcal{H}_1$ under strong compatibility, is a diffeomorphism.*

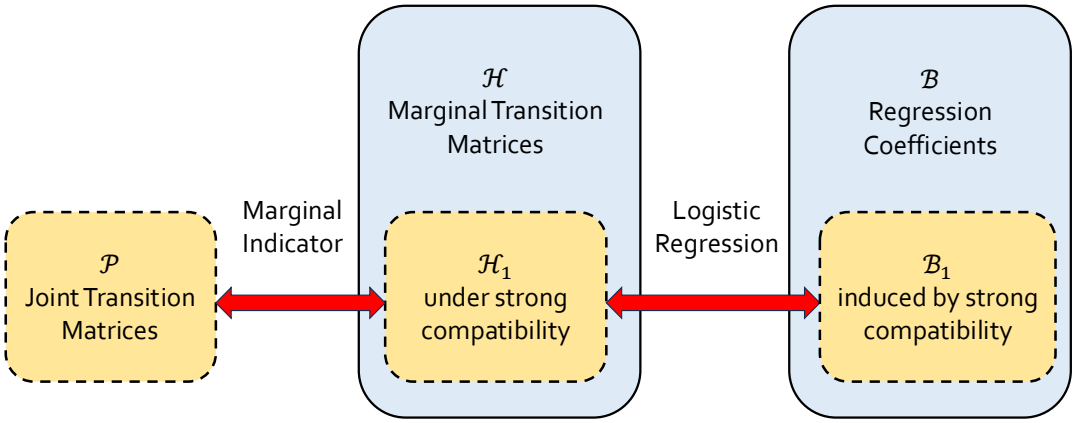


Figure 4.4: An illustration of the transformation of the parameter spaces. Initially, we start with the parameter space $\mathcal{P}$ of the joint transition probabilities. Through marginalization, this parameter space transforms into $\mathcal{H}_1$, the parameter space of marginal transition matrices H satisfying *strong compatibility*. As depicted, $\mathcal{H}_1$ is a subset of the unrestricted parameter space $\mathcal{H}$ for the marginal transition matrices. Subsequently, by incorporating multinomial logistic regression, the parameter space $\mathcal{H}$ can be further transformed into the regression coefficient parameter space $\mathcal{B}$. Within $\mathcal{B}$, a subset $\mathcal{B}_1 \subseteq \mathcal{B}$ establishes a bijection with the parameter space $\mathcal{H}_1$ for strongly compatible marginal transitions, can be identified.

Figure 4.4 illustrates the diffeomorphism between the joint transition probability space $\mathcal{P}$ and the constrained regression coefficient space $\mathcal{B}_1$, as established in Lemma 4.12. We begin with the parameter space $\mathcal{P}$ of the joint transition probabilities. Applying the marginalization map defined in Equation (4.5) through the marginal indicator matrix M yields the image $\mathcal{H}_1$, the space of marginal transition matrices H that satisfy *strong compatibility*. As shown in Figure 4.4, this space $\mathcal{H}_1$ is a subset of the full parameter space $\mathcal{H}$ of the marginal transition matrices, represented as the yellow dotted square nested within the larger blue square. Subsequently, by incorporating multinomial logistic regression, $\mathcal{H}$ is mapped into the parameter space $\mathcal{B}$ of the regression coefficients. Within this space, a subset $\mathcal{B}_1 \subseteq \mathcal{B}$ that form the diffeomorphism with the $\mathcal{H}_1 \subseteq \mathcal{H}$ can be constructed, similarly illustrated as the yellow dotted square nested within the larger blue square in the left.

The red bi-directional arrows in the figure represent the diffeomorphism: from $\mathcal{P}$ to $\mathcal{H}_1$ (Proposition 4.7) and from $\mathcal{H}_1$ to $\mathcal{B}_1$ (Proposition 4.11). Combining these, Lemma 4.12 demonstrates that the map from $\mathcal{P}$ to $\mathcal{B}_1$ is also a diffeomorphism. Analogous to Lemma 3.13 for the Univariate Marginal Model, this diffeomorphism not only resolves the non-identifiability issues but also provides a smooth and invertible reparametrisation, enabling both Frequentist and Bayesian inference on the regression coefficients $\boldsymbol{\beta}$ via the joint transition matrix P. Moreover, it guarantees that the dependency structures encoded in P are fully preserved in $\boldsymbol{\beta}$ under the Hierarchical Marginal Model, as will be discussed in the next section.

4.3.2 Encode Dependency Structures

In this section, we formalize how to encode dependency structures via the regression coefficients $\boldsymbol{\beta}$ in Equations 4.22 in Hierarchical Marginal Model.

Proposition 4.13. *We can encode any desired Granger non-causality and contemporaneous independence relationships by imposing specific structures conveniently through the regression coefficients $\boldsymbol{\beta}$ of η_j.*

Proof. Unlike the Univariate Marginal Model, which enforces both contemporaneous in-

dependence and Granger non-causality by setting the corresponding regression coefficients $\boldsymbol{\beta}$ to zero (see Proposition 3.14), the Hierarchical Marginal Model encodes these dependencies differently. Specifically, contemporaneous independence is imposed by equating certain regressions across marginal transition probabilities, while Granger non-causality is still represented by setting the relevant coefficients to zero.

We consider the contemporaneous independence and Granger non-causality for each pair of non-intersect chain groups g_{j_1} and g_{j_2}, (i.e. $g_{j_1}, g_{j_2} \subset [k]$ and $g_{j_1} \cap g_{j_2} = \emptyset$):

- For an arbitrary contemporaneous independence relationship with chain groups g_{j_1} and g_{j_2}:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t) \mid \boldsymbol{Z}(t-1)$$

  As discussed in Proposition 4.8, we can be translate it into setting $\forall g_{j'_1} \subseteq g_{j_1}, g_{j'_2} \subseteq g_{j_2}$ and $\forall \boldsymbol{b}_{j'_1} \in \boldsymbol{\mathcal{S}}^{\star}_{j'_1}, \boldsymbol{b}_{j'_2} \in \boldsymbol{\mathcal{S}}^{\star}_{j'_2}$:

$$\eta_{j'_1}(\boldsymbol{b}_{j'_1} \mid \boldsymbol{a}) \cdot \eta_{j'_2}(\boldsymbol{b}_{j'_2} \mid \boldsymbol{a}) \;=\; \eta_{j'_1 \cup j'_2}\big((\boldsymbol{b}_{j'_1}, \boldsymbol{b}_{j_2}) \mid \boldsymbol{a}\big) \qquad \forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}} \tag{4.31}$$

  By introducing the link functions defined in Equation (4.21) and (4.25), the constraints in the Equation (4.31) transfers into:

$$\eta_{j'_1}\left(\boldsymbol{b}_{j'_1} \mid \boldsymbol{a}\right) \cdot \frac{e^{\varphi_{j'_2}(\boldsymbol{b}_{j'_2}|\boldsymbol{a})}}{\sum e^{\varphi_{j'_2}(\,\cdot\,|\boldsymbol{a})}}$$

$$= \quad \eta_{j'_1}(\boldsymbol{b}_{j'_1} \mid \boldsymbol{a}) \cdot \eta_{j'_2}(\boldsymbol{b}_{j'_2} \mid \boldsymbol{a}) \qquad \text{Link function (4.21)}$$

$$= \quad \eta_{j'_1 \cup j'_2}\left(\boldsymbol{b}_{j'_1}, \boldsymbol{b}_{j'_2} \mid \boldsymbol{a}\right) \qquad \text{Contemporaneous Independence (4.31)}$$

$$= \quad \eta_{j'_1}\left(\boldsymbol{b}_{j'_1} \mid \boldsymbol{a}\right) \cdot \frac{e^{\varphi_{j'_1 \cup j'_2}(\boldsymbol{b}_{j'_1}, \boldsymbol{b}_{j'_2}|\boldsymbol{a})}}{\sum e^{\varphi_{j'_1 \cup j'_2}(\boldsymbol{b}_{j'_1}, \,\cdot\,|\boldsymbol{a})}} \qquad \text{Link function (4.28)}$$

  This implies that for all $\boldsymbol{b}_{j'_1} \in \boldsymbol{\mathcal{S}}_{j'_1}$, $\boldsymbol{b}_{j'_2} \in \boldsymbol{\mathcal{S}}_{j'_2}$, and $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, the log odds φ should

satisfy:

$$\varphi_{j_2'}(\boldsymbol{b}_{j_2'} \mid \boldsymbol{a}) = \varphi_{j_1' \cup j_2'}(\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'} \mid \boldsymbol{a})$$

Hence, the corresponding regression coefficients $\boldsymbol{\beta}$ must have:

$$\boldsymbol{\beta}_{\boldsymbol{b}_{j_2'}} = \boldsymbol{\beta}_{\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'}}$$

This indicates that under Contemporaneous Independence, the regression coefficients $\boldsymbol{\beta}$ for different groups of chains, for instance $\boldsymbol{b}_{j_2'}$ and the pair $(\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'})$ here, must be identical. Consequently, the number of independent regressions required can be reduced.

- For an arbitrary Granger non-causality that chain group g_{j_2} does not cause g_{j_1} given as:

$$\boldsymbol{Z}_{j_1}(t) \perp\!\!\!\perp \boldsymbol{Z}_{j_2}(t-1) \mid \boldsymbol{Z}_{-j_2}(t-1) \tag{4.32}$$

As discussed in Proposition 4.8, we can be translate it into setting $\forall g_{j_1'} \subseteq g_{j_1}$ and $\forall \boldsymbol{b}_{j_1'} \in \boldsymbol{\mathcal{S}}^{\star}_{j_1'}$ and $\forall \boldsymbol{a}_{j_2}, \in \boldsymbol{\mathcal{S}}^{\star}_{j_2}$:

$$\eta_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid (\boldsymbol{a}_{j_2}, \boldsymbol{a}_{-j_2})\big) = \eta_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid \boldsymbol{a}_{-j_2}\big)$$

which means that the corresponding log odds $\varphi_{j_1'}$ must satisfy:

$$\varphi_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid (\boldsymbol{a}_{j_2}, \boldsymbol{a}_{-j_2})\big) = \varphi_{j_1'}\big(\boldsymbol{b}_{j_1'} \mid \boldsymbol{a}_{-j_2}\big)$$

Hence, the corresponding regression coefficients $\boldsymbol{\beta}_{\boldsymbol{b}_{j_1'}}$ must have:

$$\beta_{\boldsymbol{b}_{j_1'}}(\boldsymbol{a}_j) = 0 \qquad \forall g_{j_1'} \subseteq g_{j_1} \quad \text{and} \quad g_j \cap g_{j_2} \neq \emptyset$$

Here, $\beta_{\boldsymbol{b}_{j_1'}}(\boldsymbol{a}_j)$ refers to the component of the regression coefficient vector $\boldsymbol{\beta}_{\boldsymbol{b}_{j_1'}}$ that accounts for the influence of the past states in chain group g_j. Specifically, g_j includes at least one element from g_{j_2}, the group whose influence is being accounted.

Under the Granger non-causality, this component is set to zero, implying that the past states of g_{j_2} have no effect on the current state of g_{j_1}.

□

Recall that the contemporaneous independence, which describes how current independence among chains, is separate from Granger non-causality, which stipulates the independence from the past time. This distinction is evident in the different ways they constrain the multinomial logistic regression coefficients $\boldsymbol{\beta}$:

- **Contemporaneous independence** imposes relationships among different multinomial logistic regressions, requiring certain regression coefficient vectors to equal the others.
- **Granger non-causality** places restrictions within a single multinomial logistic regression, setting specific coefficients to zero to eliminate any influence from another chain's previous state.

Based on the if and only if relationship between dependency structures and regression coefficients established in Proposition 4.13, we can alternatively define the contemporaneous independence and the Granger non-causality in terms of the regression coefficients in the Hierarchical Marginal Model, in parallel to the Definition 3.15 for the Univariate Marginal Model. This alternative characterization is formalized as follows:

Definition 4.14 (Define Dependency Structures via the Regression Coefficients)**.** *Consider the MMC $\boldsymbol{Z}$ with chains $[K] = 1, 2, \ldots, K$. Given two disjoint groups of chains $g_{j_1}, g_{j_2} \subset [K]$ and the corresponding marginal process $\boldsymbol{Z}_{j_1}$ and $\boldsymbol{Z}_{j_2}$ of $\boldsymbol{Z}$:*

- *$\boldsymbol{Z}_{j_1}$ and $\boldsymbol{Z}_{j_2}$ are said to be contemporaneously independent if and only if the corresponding regression coefficients satisfy:*

$$\boldsymbol{\beta}_{\boldsymbol{b}_{j_2'}} = \boldsymbol{\beta}_{\boldsymbol{b}_{j_1'}, \boldsymbol{b}_{j_2'}}$$

for all $g_{j_1'} \subseteq g_{j_1}$, $g_{j_2'} \subseteq g_{j_2}$ and $\forall \boldsymbol{b}_{j_1'} \in \boldsymbol{\mathcal{S}}_{j_1'}, \boldsymbol{b}_{j_2'} \in \boldsymbol{\mathcal{S}}_{j_2'}$.

- *$\boldsymbol{Z}_{j_1}$ is not Granger caused by $\boldsymbol{Z}_{j_2}$ if and only if the corresponding regression coefficients satisfy:*

$$\beta_{\boldsymbol{b}_{j_1'}}(\boldsymbol{a}_j) = 0$$

for all $g_{j_1'} \subseteq g_{j_1}$ and $g_j \cap g_{j_2} \neq \emptyset$ and $\beta_{\boldsymbol{b}_{j_1'}}(\boldsymbol{a}_j)$ refers to the component of the regression coefficient vector $\boldsymbol{\beta}_{\boldsymbol{b}_{j_1'}}$ that accounts for the influence of the past states in chain group g_j.

It is also important to note that the regression coefficients $\boldsymbol{\beta}$ in the Hierarchical Marginal Model are subject to boundary constraints. As a result, it may occur that certain configurations of $\boldsymbol{\beta}$ specifically those corresponding to Granger non-causality and contemporaneous independence fall outside the feasible region, implying that such dependency structures cannot be imposed under some parameter choices.

However, it can be shown that for any choice of regression coefficients assigned to the $(i-1)$-variate marginals, the coefficients corresponding to the i-variate marginal that encode Granger non-causality and contemporaneous independence will always lie within the feasible region defined by those lower-order marginals. This result is formalized in the following proposition:

Proposition 4.15. *Let g_j denote an i-variate chain group, and let $\boldsymbol{\beta}_{\boldsymbol{b}_j}$ be the corresponding regression coefficients for the marginal probability $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$. Suppose $\boldsymbol{\beta}^{\star}_{\boldsymbol{b}_j}$ denotes the vector of the regression coefficients encoding a desired dependency structure (for instance contemporaneous independence and Granger non-causality). Let $L(\boldsymbol{\beta}_{\boldsymbol{a}_{j'}})$ and $U(\boldsymbol{\beta}_{\boldsymbol{a}_{j'}})$ represent the lower and upper bounds on $\boldsymbol{\beta}_{\boldsymbol{a}_j}$, respectively, induced by regression coefficients of all lower-dimensional marginals $\forall g_{j'} \subset g_j$. Then, the following always holds:*

$$\boldsymbol{\beta}^{\star}_{\boldsymbol{a}_j} \in \Big[\, L(\boldsymbol{\beta}_{\boldsymbol{a}_{j'}}) \,,\, U(\boldsymbol{\beta}_{\boldsymbol{a}_{j'}}) \,\Big]$$

meaning that the coefficients $\boldsymbol{\beta}^{\star}_{\boldsymbol{a}_j}$ representing the desired dependency structure always lie within the feasible region defined by the lower-dimensional marginals. Therefore, such boundaries do not restrict the ability to impose these dependency structures in the Hier-

archical Marginal Model.

Proof. Since the regression parameters β and the marginal transition probabilities η are related by a strictly increasing multinomial logistic link function, it suffices to verify that each $\eta^{\star}_{j}$, representing a dependency structure, lies within the interval defined by its lower-dimensional marginals $\eta_{j'}$, with $g_{j'} \subset g_j$.

Any structural assumption, such as contemporaneous independence and Granger non-causality, can be viewed as a deterministic map from lower-order marginals to higher-order marginals according to Proposition 4.8. This implicitly defines the marginals η, guaranteeing *strong compatibility*. Specifically, as discussed in Section 2.8.2, these dependency structures can be written in terms of the joint transition probabilities p, marginalizing these probabilities yields η, ensuring that the resulting marginal probabilities satisfy *strong compatibility* and thus fall within the feasible region defined by lower-dimensional marginals via the Inclusion-Exclusion formula.

□

The following example provides a numerical demonstration for the simple case with two chains and binary states, followed by an illustration of this property in the 3D plane:

Example 4.16. Continue with Example 4.10 with two chains and binary states:

- Consider the following **Contemporaneous Independence**:

$$Z_1(t) \perp\!\!\!\perp Z_2(t) \mid \boldsymbol{Z}(t-1)$$

  According to Proposition 4.13, this independence corresponds specifically to $\beta^{\star}_{\{1,2\}} = \beta_2$ (equivalently interpreted as conditional independence in this case). Thus, we have to verify the inequalities $\forall \beta_1, \beta_2 \in \mathbb{R}$:

$$\operatorname{logit}\Big(\max\Big[0\ ,\ \frac{1-e^{-(\beta_1+\beta_2)}}{1+e^{-\beta_2}}\Big]\Big) \leq \beta^{\star}_{\{1,2\}} = \beta_2 \leq \operatorname{logit}\Big(\min\Big[1\ ,\ \frac{1+e^{-\beta_1}}{1+e^{-\beta_2}}\Big]\Big)$$

Taking logistic transformations for both sides simplifies these inequalities as:

$$\frac{1-e^{-(\beta_1+\beta_2)}}{1+e^{-\beta_2}} \leq \frac{1}{1+e^{-\beta_2}} \leq \frac{1+e^{-\beta_1}}{1+e^{-\beta_2}}$$

and confirms the validity. This characteristic can also be visualized in the 3D plane, as illustrated in Figure 4.5, where the blue plane, with each point lies on that refers to the $\beta_{12}^{\star}$ standing for the contemporaneous independence, is sandwiched between the lower bound in green and upper bound in red. Hence, for any $\beta_1, \beta_2 \in \mathbb{R}$, the corresponding regression coefficients $\beta_{12}^{\star}$ standing for the contemporaneous independence is guaranteed to lie within the feasible interval induced by β_1, β_2, ensuring a inverse map back to the valid joint transition matrix.

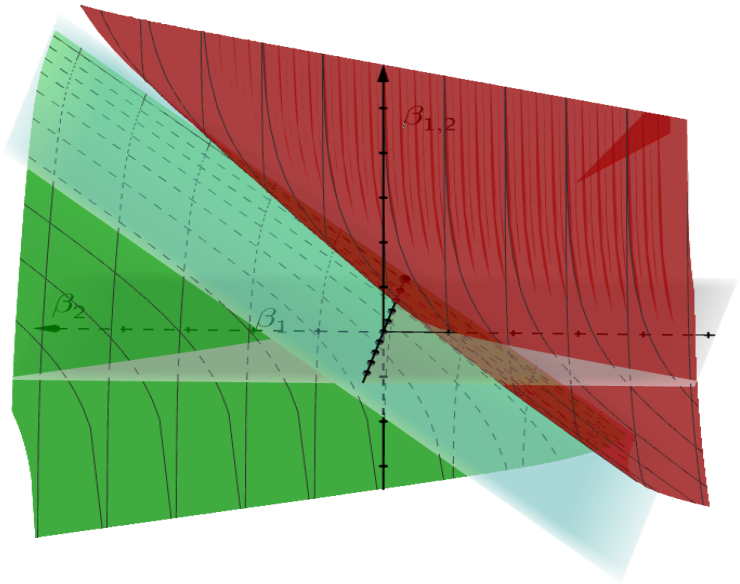


Figure 4.5: An illustration of the feasible region of $\beta_{\{1,2\}}^{\star}$ corresponding to different dependency structures. According to Proposition 4.13, contemporaneous independence corresponds to the blue plane $\beta_{\{1,2\}} = \beta_2$. As shown in the figure, this entire plane lies between the green (lower bound) and red (upper bound) planes induced by the regression coefficients β_1 and β_2. Furthermore, Granger non-causality corresponds specifically to the origin $(\beta_1, \beta_2, \beta_{\{1,2\}}) = (0, 0, 0)$, which also resides within these bounds. This visualization underscores the fact that, for any given regression coefficients β_1 and β_2, it is always possible to identify a suitable $\beta_{\{1,2\}}^{\star}$ that stands for contemporaneous independence or Granger non-causality within the feasible region.

- Consider the following **Granger non-Causality**:

$$Z_3(t) \perp\!\!\!\perp Z_1(t), Z_2(t) \mid \boldsymbol{Z}_{-\{1,2\}}(t)$$

Proposition 4.13 implies that this condition holds if and only if β_1, β_2 and $\beta_{\{1,2\}}$ are simultaneously set to be 0. We therefore need to verify for $\beta_1 = \beta_2 = 0$:

$$\text{logit}\Big(\max\left[\,0\,,\ \frac{1-e^{-(\beta_1+\beta_2)}}{1+e^{-\beta_2}}\right]\Big) \leq \beta^{\star}_{\{1,2\}} = 0 \leq \text{logit}\Big(\min\left[\,1\,,\ \frac{1+e^{-\beta_1}}{1+e^{-\beta_2}}\right]\Big)$$

This verification is straightforward, as the condition $\beta_1 = \beta_2 = 0$ implies that $\beta_{\{1,2\}}$ can take any value in $\mathbb{R}$, clearly satisfying the inequalities. Geometrically, this corresponds to the origin being located within the region bounded by the upper and lower bounds in red and green respectively, as illustrated in Figure 4.5.

4.3.3 Examples on Encoding Dependency Structures

In this section, we present examples on how to encode dependency structures via the Hierarchical Marginal Model. Stick on the same setting in Example 4.9 where we consider an MMC with 3 chains, each taking values in $\mathcal{S} = \{1, 2\}$. Under the Hierarchical Marginal Model, the quantities that need to be specified, as shown in Example 4.2, are:

3 univariate marginals:	$\eta_{\{k\}}\big(2 \mid \boldsymbol{a}\big)$	$\forall k \in \{1,2,3\}$
3 bivariate marginals:	$\eta_{\{k,k'\}}\big((2,2) \mid \boldsymbol{a}\big)$	$\forall k, k' \in \{1,2,3\}$
1 trivariate marginal:	$\eta_{\{1,2,3\}}\big((2,2,2) \mid \boldsymbol{a}\big)$	

Moreover, with a bivariate state space $\mathcal{S} = \{1, 2\}$, the multinomial regression simplifies to a standard logistic regression. We will then construct 7 logistic regressions, with one for each of the marginal quantities above, to explicitly encode conditional independence, contemporaneous independence, and Granger non-causality.

(I) Full Dependency

Recall the full dependency scenario illustrated in Figure 2.2, meaning there are no additional constraints on the logistic coefficients $\boldsymbol{\beta}$. Thus, each coefficient β can vary in $\mathbb{R}$

subject to the *strong compatibility*. The corresponding logistic regression is given as:

$$\text{logit}\big(\boldsymbol{\eta}_j(\boldsymbol{b}_j \mid \cdot)\big) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_j \qquad \forall g_j \subseteq \{1,2,3\}$$

which can be written in matrix form explicitly as:

$$\text{logit}\begin{pmatrix} \eta_j\big(\boldsymbol{b}_j \mid (1,1,1)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (1,1,2)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (1,2,1)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (1,2,2)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (2,1,1)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (2,1,2)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (2,2,1)\big) \\ \eta_j\big(\boldsymbol{b}_j \mid (2,2,2)\big) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_{\emptyset j} \\ \beta_{\{1\}j} \\ \beta_{\{2\}j} \\ \beta_{\{3\}j} \\ \beta_{\{1,2\}j} \\ \beta_{\{1,3\}j} \\ \beta_{\{2,3\}j} \\ \beta_{\{1,2,3\}j} \end{pmatrix} \tag{4.33}$$

Equivalently, such a logistic regression in Equation 4.33 can be written in one equation via the indicator function $\mathbb{I}(\cdot)$ based on the idea from Diggle [2002]:

$$\begin{aligned} &\text{logit}\Big(\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big)\Big) \\ =\ &\underbrace{\beta_{\emptyset j}}_{\text{intercept}} + \underbrace{\mathbb{I}\big(a_1=2\big)\beta_{\{1\}j} + \mathbb{I}\big(a_2=2\big)\beta_{\{2\}j} + \mathbb{I}\big(a_3=2\big)\beta_{\{3\}j}}_{\text{univariate effects}} \\ +\ &\underbrace{\mathbb{I}\big(a_1=2\big)\mathbb{I}\big(a_2=2\big)\beta_{\{1,2\}j} + \mathbb{I}\big(a_1=2\big)\mathbb{I}\big(a_3=2\big)\beta_{\{1,3\}j} + \mathbb{I}\big(a_2=2\big)\mathbb{I}\big(a_3=2\big)\beta_{\{2,3\}j}}_{\text{bivariate effects}} \\ +\ &\underbrace{\mathbb{I}\big(a_1=2\big)\mathbb{I}\big(a_2=2\big)\mathbb{I}\big(a_3=2\big)\beta_{\{1,2,3\}j}}_{\text{trivariate effects}} \end{aligned} \tag{4.34}$$

- $\boldsymbol{\eta}_j(\boldsymbol{b}_j \mid \cdot)$ is an 8×1 column vector that collects $\eta_j(\boldsymbol{b}_j \mid \boldsymbol{a})$ for each $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$.
- $\boldsymbol{\mathcal{X}}$ is the 8×8 design matrix constructed according to Equation (4.24) and is invariant for all of the 7 logistic regressions.

- $\boldsymbol{\beta}_j$ is the 8×1 vector of the logistic regression coefficients for the marginal of chain group g_j. Each term $\beta_{j'j}$ corresponds to the log-odds effect of chain group $g_{j'}$ being in an alternative state at the previous time, on the chain group g_j being in an alternative state at the current time. $\beta_{\emptyset j}$ is the intercept standing for the baseline level.

In the case of full dependency, no structural constraints are imposed on the regression coefficients $\boldsymbol{\beta}_{j'}$, allowing all coefficients to vary freely.

4.3.3.1 (II) Conditional Independence

Recall the Conditional Independence on 3 chains as illustrated in Figure 2.3, where all the blue vertical arrows are missing, indicating all concurrent interactions among the chains are eliminated after conditioning on past states:

$$Z_1(t) \perp\!\!\!\perp Z_2(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1)$$

According to Proposition 4.13, this conditional independence can be incorporated into the logistic coefficients $\boldsymbol{\beta}$ in Equation (4.34) as:

$$\begin{aligned} \boldsymbol{\beta}_{\{1,2\}} &= \boldsymbol{\beta}_{\{2\}}, \qquad \boldsymbol{\beta}_{\{1,3\}} = \boldsymbol{\beta}_{\{3\}} \\ \boldsymbol{\beta}_{\{2,3\}} &= \boldsymbol{\beta}_{\{3\}}, \qquad \boldsymbol{\beta}_{\{1,2,3\}} = \boldsymbol{\beta}_{\{3\}} \end{aligned}$$

Consequently, we only require three logistic regressions for each of the univariate marginal

$$\operatorname{logit}\big(\boldsymbol{\eta}_k(2 \mid \cdot)\big) = \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_k \qquad k = 1, 2, 3$$

instead of 7 logistic regressions under full dependency require explicit modelling. Since the rest 4 regression coefficients $\boldsymbol{\beta}_{\{1,2\}}$, $\boldsymbol{\beta}_{\{1,3\}}$, $\boldsymbol{\beta}_{\{2,3\}}$ and $\boldsymbol{\beta}_{\{1,2,3\}}$ can be deduced from the regression coefficients for 3 univariate marginals. In other words, under conditional independence, we only need the univariate marginals $\boldsymbol{\eta}_{\{k_1\}}(2 \mid \cdot)$ to recover the transition matrix P, which is consistent with what we have discussed in Section 2.4.

4.3.3.2 (III) Contemporaneous Independence

Recall the Contemporaneous Independence on 3 chains as illustrated in Figure 2.6, where the missing blue vertical arrows between $\big(Z_1(t), Z_2(t)\big)$ and $Z_3(t)$ indicating the partially missing concurrent interaction between them:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \quad | \quad \boldsymbol{Z}(t-1) \tag{4.35}$$

According to Proposition 4.13, this contemporaneous independence translates into the following relationships among the coefficients $\boldsymbol{\beta}_{j'}$:

$$\boldsymbol{\beta}_{\{1,3\}} = \boldsymbol{\beta}_{\{2,3\}} = \boldsymbol{\beta}_{\{1,2,3\}} = \boldsymbol{\beta}_{\{3\}} \tag{4.36}$$

Hence, rather than modelling all 7 logistic regressions for the fully connected structure, one only needs 4 logistic regressions under the contemporaneous independence in Equation (4.35), namely as:

$$\begin{aligned}
\text{logit}\Big(\boldsymbol{\eta}_{\{1\}}(2 \mid \cdot)\Big) &= \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{\{1\}} \\
\text{logit}\Big(\boldsymbol{\eta}_{\{2\}}(2 \mid \cdot)\Big) &= \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{\{2\}} \\
\text{logit}\Big(\boldsymbol{\eta}_{\{3\}}(2 \mid \cdot)\Big) &= \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{\{3\}} \\
\text{logit}\Big(\boldsymbol{\eta}_{\{1,2\}}\big((2,2) \mid \cdot\big)\Big) &= \boldsymbol{\mathcal{X}}\boldsymbol{\beta}_{\{1,2\}}
\end{aligned}$$

The remaining three coefficient vectors $\boldsymbol{\beta}_{\{1,3\}}$, $\boldsymbol{\beta}_{\{2,3\}}$, and $\boldsymbol{\beta}_{\{1,2,3\}}$ capturing $\boldsymbol{\eta}_{\{1,3\}}\big((2,2) \mid \cdot\big)$, $\boldsymbol{\eta}_{\{1,2\}}\big((2,2) \mid \cdot\big)$ and $\boldsymbol{\eta}_{\{1,2,3\}}\big((2,2,2) \mid \cdot\big)$ can be derived via Equations (4.36) under Contemporaneous independence (4.35). It is also intuitive to see that the contemporaneous independence imposes one less restriction compared to the contemporaneous independence, as it is less restrictive and retains some of the pairwise or higher-order interactions among the different chains comparing to the conditional independence.

4.3.3.3 (IV) Granger non-Causality

Unlike conditional independence and contemporaneous independence, Granger non-causality specifically limits how a chains previous state influences the current states of other chains. Recall the Granger non-causality on 3 chains as illustrated in Figure 2.8, where the missing black arrows from $Z_3(t-1)$ to $Z_1(t)$ and $Z_2(t)$ represent that chain 3 at previous time does not affect the states of chain 1 or chain 2 at current time. Formally, this is expressed as

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \quad | \quad Z_1(t-1), Z_2(t-1)$$

According to the Proposition 4.13, the Granger non-causality can be encoded within the logistic regression framework as:

$$\beta_{j'j} = 0 \qquad \forall\, g_{j'} \supseteq \{3\} \quad \text{and} \quad g_j \subseteq \{1,2\}$$

meaning that, for each $g_j \subseteq \{1,2\}$, the regression in (4.34) simplifies to

$$\text{logit}\Big(\eta_j\big(\boldsymbol{b}_j \mid \boldsymbol{a}\big)\Big) \;=\; \beta_{\emptyset j} \;+\; \mathbb{I}\big(a_1 = 2\big)\beta_{\{1\}j} \;+\; \mathbb{I}\big(a_2 = 2\big)\beta_{\{2\}j} \;+\; \mathbb{I}\big(a_1 = 2\big)\mathbb{I}\big(a_2 = 2\big)\beta_{\{1,2\}j}$$

It indicates that all the terms corresponding to the effects from chain 3, namely $\beta_{\{3\}j}$, $\beta_{\{1,3\},j}$, $\beta_{\{2,3\},j}$, and $\beta_{\{1,2,3\},j}$, are set to be 0 and consequently eliminates the effects from chain 3 at previous time to the joint of chain 1 and chain 2.

4.4 Generative Model

In this section, we examine the utilization of the Hierarchical Marginal Model as generative model, that is to generate an MMC $\boldsymbol{Z}$ that exhibits desired dependency structures, based on the regression coefficients $\boldsymbol{\beta}$.

As discussed previously in Section 4.2.1, certain constraints have to be imposed on the marginal transition probabilities η to ensure strong compatibility and enable inversion to a valid joint transition probability matrix P. These constraints, in turn, induce restrictions on the corresponding regression coefficients $\boldsymbol{\beta}$, which prevents us from sampling all components of $\boldsymbol{\beta}$ jointly without violating compatibility conditions.

However, as elaborated in Section 4.2.1, these constraints on η can be derived hierarchically via the Inclusion-Exclusion formula. In particular, univariate marginal transition probabilities are unconstrained, while constraints on higher dimensional marginals are determined by the lower ones. This hierarchical structure translates directly into the domain of the regression coefficients $\boldsymbol{\beta}$, allowing us to generate them in a sequential manner. Furthermore, Proposition 4.15 guarantees that these constraints on $\boldsymbol{\beta}$ do not hinder the imposition of the dependency structures such as contemporaneous independence and Granger non-causality. This is because the values of $\boldsymbol{\beta}$ representing such structures always lie within the feasible region defined by the hierarchical constraints. Bringing together these insights, we propose the following Hierarchical Generating Algorithm 2, which utilizes a hierarchical generating strategy (the **For** loop) to generate regression coefficients $\boldsymbol{\beta}$ satisfying their constraints while explicitly incorporating the desired dependency structures.

It is worth noting that, although the Hierarchical Generating Algorithm 2 guarantees that the generated regression coefficients $\boldsymbol{\beta}_\sigma$ always correspond to a valid joint transition matrix P and is able to produce an MMC with arbitrary combinations of dependency structures, it introduces additional computational steps. In particular, the feasible region

for $\boldsymbol{\beta}_\sigma$ must be computed sequentially to ensure the strong compatibility required by the Hierarchical Marginal Model, preventing it from the parallel sampling procedure (as in Algorithm 1 for the Univariate Marginal Model, where all $\boldsymbol{\beta}_\sigma$ can be generated simultaneously). As a result, this generative approach brings extra computational cost.

Algorithm 2 Generating Data from the Hierarchical Marginal Model

Input: the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ for the prior distribution of regression coefficients $\boldsymbol{\beta}$

Output: an MMC $\boldsymbol{Z}$ generated by the Hierarchical Marginal Model

Generate the regression coefficients $\boldsymbol{\beta}_j \sim \mathcal{N}\big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\big)$ of each univariate chain groups g_j
for Regression Coefficients $\boldsymbol{\beta}_j$ for $|g_j| = 2, 3, \ldots, K$ variate marginals **do**
  if impose dependency structures **then**
    Contemporaneous Independence: $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j'}$ (see Proposition 4.13)
    Granger non-Causality: $\boldsymbol{\beta}_j = 0$ (see Proposition 4.13)
  else
    Compute the lower and the upper bound $L(\boldsymbol{\beta}_j)$, $U(\boldsymbol{\beta}_j)$ for $\boldsymbol{\beta}_j$ (see Example 4.16)
    Generate $\boldsymbol{\beta}_j \sim \text{Unif}\big(L(\boldsymbol{\beta}_j), U(\boldsymbol{\beta}_j)\big)$
  end if
end for
Compute the joint transition matrix P by $\boldsymbol{\beta}$ via $g_1(\cdot)$ defined in Lemma 4.12
Generate MMC $\boldsymbol{Z}$ based on the the joint transition matrix P

4.5 Simulation Studies

In this section, we examine the performance of the Hierarchical Marginal Model, in particular the ability in recovering the true underlying regression coefficients, $\boldsymbol{\beta}_\sigma$, as well as detecting the potential dependency structures embedded within the MMC based on the simulated data. We keep the same simulation setup as described earlier in Section 5.4.1 for the Univariate Marginal Model.

The simulation results for the Hierarchical Marginal Model under: (I) Full model, (II) Contemporaneous independence, (III) Granger non-causality, (IV) Mixed dependency structures are presented in the following section 4.5.1.

4.5.1 Simulation Results

Recall from Section 4.3.3, the quantities modelled by the Hierarchical Marginal Model in this setting are:

$$\begin{aligned}
&\text{3 univariate marginals:} && \eta_{\{k\}}\big(2 \mid \boldsymbol{a}\big) && \forall k \in \{1,2,3\} \\
&\text{3 bivariate marginals:} && \eta_{\{k,k'\}}\big((2,2) \mid \boldsymbol{a}\big) && \forall k, k' \in \{1,2,3\} \\
&\text{1 trivariate marginal:} && \eta_{\{1,2,3\}}\big((2,2,2) \mid \boldsymbol{a}\big)
\end{aligned}$$

We will construct 7 logistic regressions, with one for each of the marginal quantities above, to explicitly encode conditional independence, contemporaneous independence, and Granger non-causality.

(I) Full model

Recall the full model illustrated in Figure 4.6. There is no independent relationship among the chains and all the chains are fully connected, indicating no additional restrictions are required to place on the regression coefficients $\boldsymbol{\beta}$.

As an illustration, consider the regression coefficients $\boldsymbol{\beta}_j$ for:

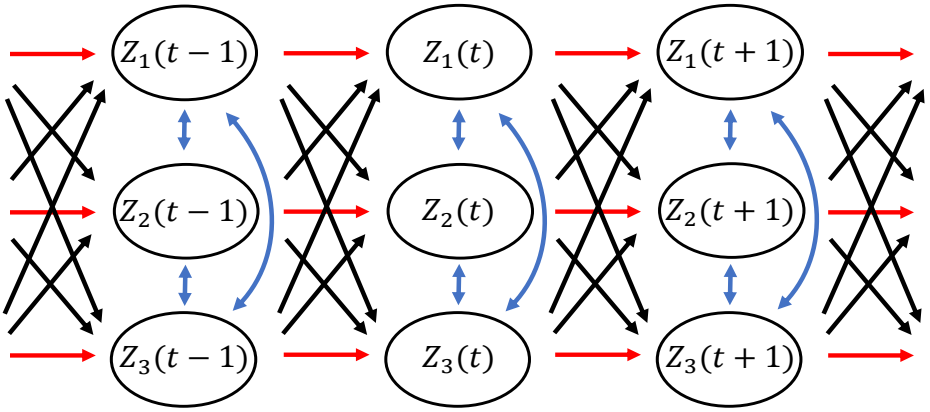


Figure 4.6: An illustration of an MMC with 3 chains under full dependency, with all the chains fully connected with each other.

1. the univariate chain group $g_j = \{1\}$

2. the bivariate chain group $g_j = \{1, 2\}$

For each case, we generate box plots of the estimates for the $2^3 = 8$ regression coefficients

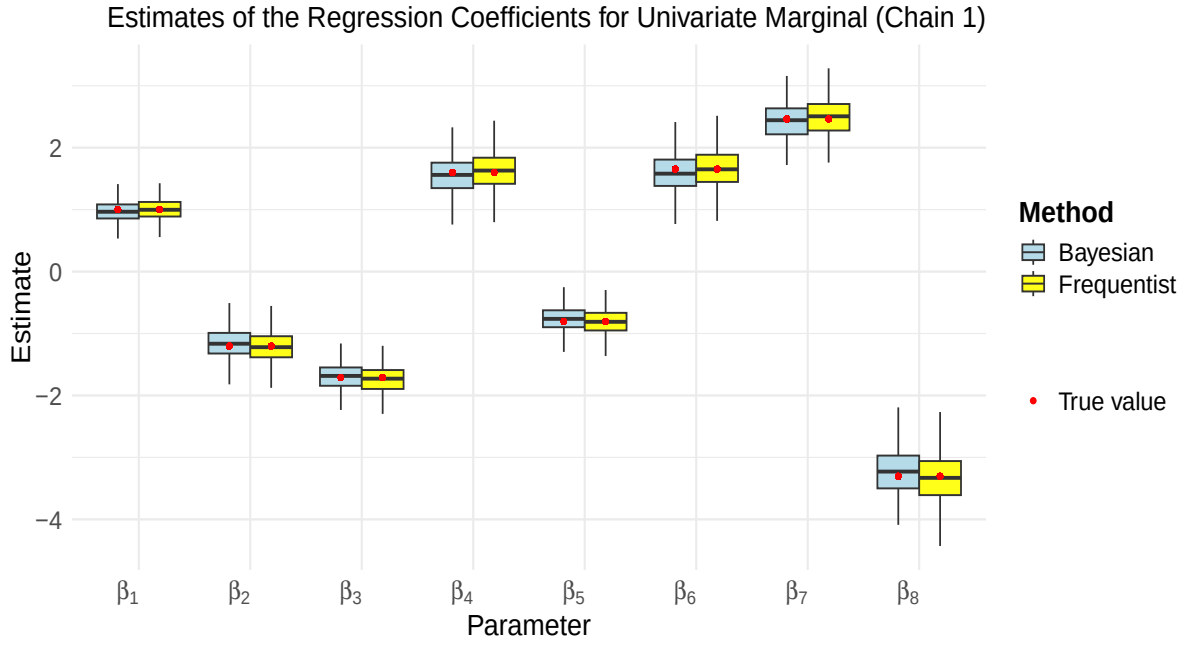


Figure 4.7: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

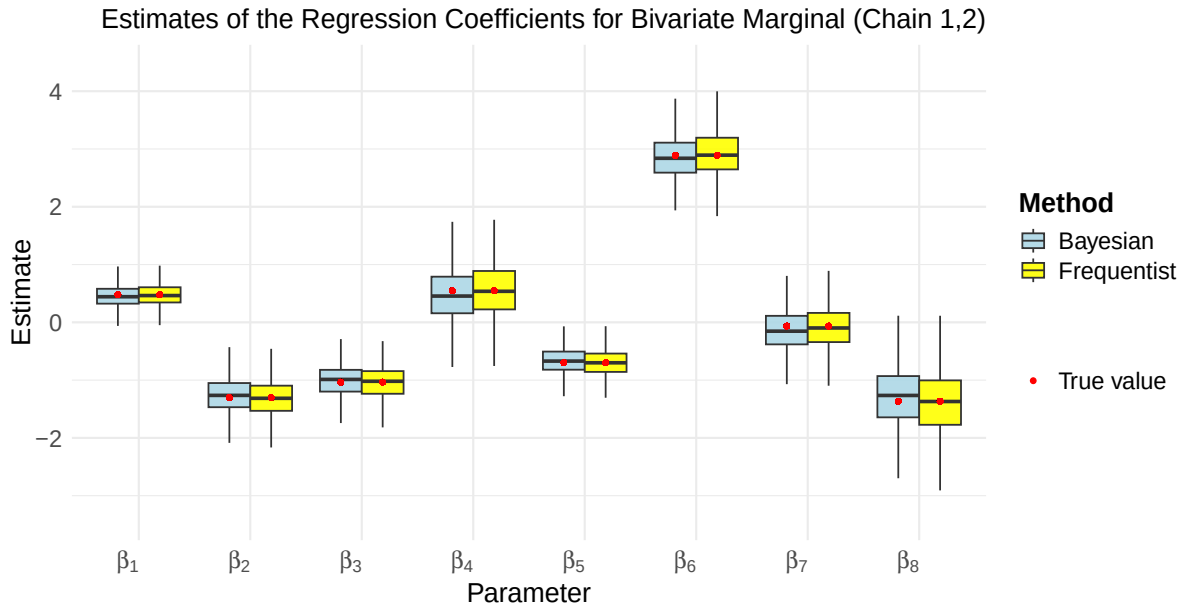


Figure 4.8: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the bivariate chain group $g_j = \{1, 2\}$, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment.

obtained in each repetition, with the Frequentist estimates shown in light blue and the Bayesian estimates of the posterior median in yellow. The true values are indicated by red points in Figures 4.7 and 4.8, respectively.

From the figures, it is evident that both the Frequentist and Bayesian approaches provide accurate estimates of the true regression coefficients $\boldsymbol{\beta}_j$ for the univariate and bivariate marginals, as the true values lie very close to the medians of the box plots.

Although there may be no inherent dependency structures under the full model, we can still perform hypothesis testing on the regression coefficient estimates $\boldsymbol{\beta}$ to investigate each of the potential dependencies among the chains under Hierarchical Marginal model. We consider the null and alternative hypotheses:

$$\begin{aligned} H_0: &\quad Z_k \text{ and } Z_{k'} \text{ are contemporaneously independent or } Z_k \text{ is not Granger caused by } Z_{k'} \\ H_1: &\quad \text{The chains are dependent on each other} \end{aligned} \tag{4.37}$$

For each potential dependency (i.e., each arrow in Figure 4.6), we apply two approaches:

- **Frequentist**: Conduct Hotelling's T^2 test on the MLE estimates from each repetitive experiments and hence compute the corresponding p-value.

- **Bayesian** Compute the contour probabilities (Bayesian p-values) of the posterior medians from each repetitive experiments.

As an example, consider testing whether chain 1 is Granger caused by chain 2 (represented by the arrow from chain 2 to chain 1). According to Proposition 4.13, this is equivalent as testing:

$$\beta_{j\{1\}} = 0 \qquad \text{for} \quad \{2\} \subseteq g_j$$

The p-values obtained from both the Frequentist and Bayesian approaches are agree with each other and can be summarized in Figure 4.9. A lower p-value indicates stronger evidence against the null hypothesis H_0, suggesting a strong dependency between the corresponding chains; such relationships are depicted as red arrows with low transparency. Conversely, a higher p-value indicates insufficient evidence to reject H_0, and the independence between the corresponding chains is retained; these are shown as red arrows with high transparency.

It is not surprising that, under the full model, where the MMC is generated completely at random without imposing any structural constraints on the regression coefficients $\boldsymbol{\beta}$, all p-values from the hypothesis test (4.37) for the corresponding dependency structures, including contemporaneous independence and Granger non-causality, are less than 0.01. This indicates that the three chains are fully connected. The same pattern is reflected in Figure 4.9, where all arrows, representing the inferred p-values, are coloured red, corresponding to $p \leq 0.01$ and confirming the existence of the dependencies. This finding is consistent with the Figure 4.6, which illustrates the complete dependency structure introduced at the beginning of this section.

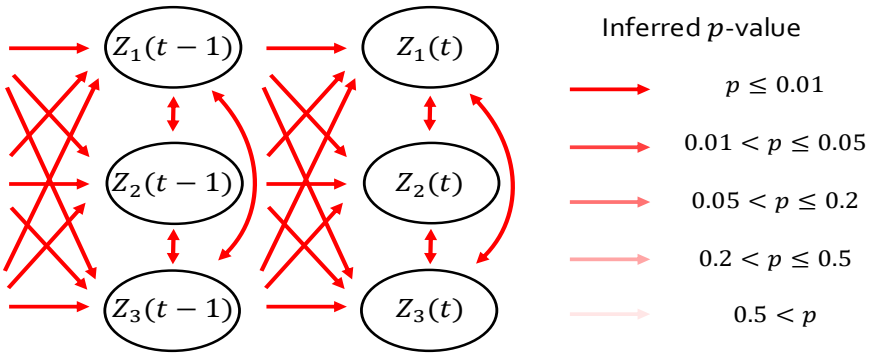


Figure 4.9: An illustration of the inferred p-values from both the Frequentist and Bayesian approaches is shown in red, with varying levels of transparency to indicate the magnitude of the p-values and, consequently, the strength of the inferred dependencies. It is explicit to see that all p-values are notably small ($p \leq 0.01$), leading to rejection of the null hypothesis 4.37 and supporting the existence of dependency structures among all the chains.

(II) Contemporaneous Independence

Recall the contemporaneous independence setting illustrated in Figure 4.10, where chains 1 and 2 are contemporaneously independent from chain 3:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1) \tag{4.38}$$

This independence is visually represented by the absence of the blue vertical arrow connecting $Z_1(t)$ and $Z_2(t)$ with $Z_3(t)$, in contrast to the full model shown in Figure 4.6.

Under contemporaneous independence, the regression coefficients for the chain groups $g_{j_1} = \{3\}$, $g_{j_2} = \{1, 3\}$, $g_{j_3} = \{2, 3\}$, and $g_{j_4} = \{1, 2, 3\}$ must be equal, that is:

$$\boldsymbol{\beta}_{j_1} = \boldsymbol{\beta}_{j_2} = \boldsymbol{\beta}_{j_3} = \boldsymbol{\beta}_{j_4} \tag{4.39}$$

Consequently for illustration, we visualize the estimates of β_1 and β_2, the first two elements of $\boldsymbol{\beta}_j$, across the chain groups $g_j = \{3\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}$ in Figure 4.11.

From Figures 4.11a and 4.11b, it is evident that both the Frequentist (MLE) and Bayesian

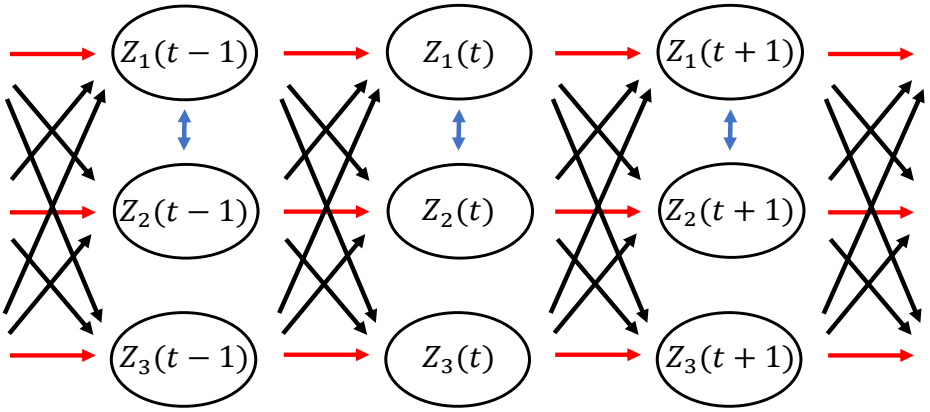


Figure 4.10: An illustration of an MMC with 3 chains under the contemporaneous independence where chain 1 and chain 2 are contemporaneously independent from chain 3.

(posterior median) yield accurate estimates of the true values of β_1 and β_2 under contemporaneous independence (4.38), as the medians of each box lie very close to the true values indicated by the red dashed lines. Moreover, the medians of β_1 and β_2 are nearly identical across the corresponding chain groups, which is consistent with the equality constraint in (4.39).

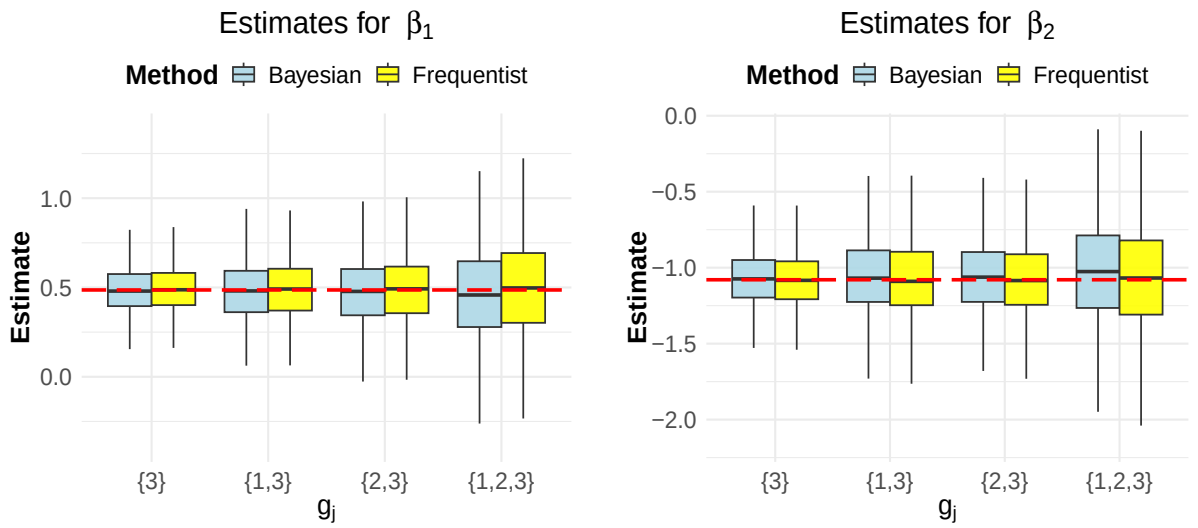
To test the contemporaneous independence in Frequentist approach, we perform a likelihood ratio test within each of the 500 repeated experiments for the following nested model:

$$H_0: \quad Z_1 \text{ and } Z_2 \text{ are contemporaneously independent from } Z_3$$
$$H_1: \quad Z_1 \text{ and } Z_2 \text{ are correlated with } Z_3$$

with the test statistic

$$-2\log\Lambda = 2\left(\max_{\boldsymbol{\beta} \text{ under } H_1} \ell(\boldsymbol{\beta}) - \max_{\boldsymbol{\beta}_0 \text{ under } H_0} \ell(\boldsymbol{\beta}_0)\right) \sim \chi^2_\nu \tag{4.40}$$

where the degree of freedom is $\nu = 8 \times 3 = 24$, since under the restricted model H_0 (contemporaneous independence), four regression coefficients $\boldsymbol{\beta}_j$ across the chain groups



(a) The estimates of β_1

(b) The estimates of β_2

Figure 4.11: An illustration of the posterior medians of the regression coefficients β_1 and β_2 across the chain groups $g_j = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$ from Bayesian (in light blue) and Frequentist (in yellow), with the true value indicated by the red dashed line. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment. Under Contemporaneous Independence, these estimated value of regression coefficients β should agree.

$g_j = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$ are restricted to be the same.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 21 out of 500 experiments (4.2%), which is fully consistent with the random variation under H_0. Furthermore, we record the p-value from each experiment and present their distribution in the histogram as the following Figure 4.12:

According to Lehmann and Romano [2005, pp. 63–65], under the null hypothesis H_0, the p-values should follow a uniform distribution within $(0, 1)$. As shown, the 500 p-values from the repetitive experiments are approximately uniformly distributed over $[0, 1]$, closely matching the Uniform distribution indicated by the red dashed line. This suggests that the test is well-calibrated and that the null hypothesis H_0 of contemporaneous independence is plausible.

Alternatively in Bayesian fashion, we compute the corresponding posterior contour prob-

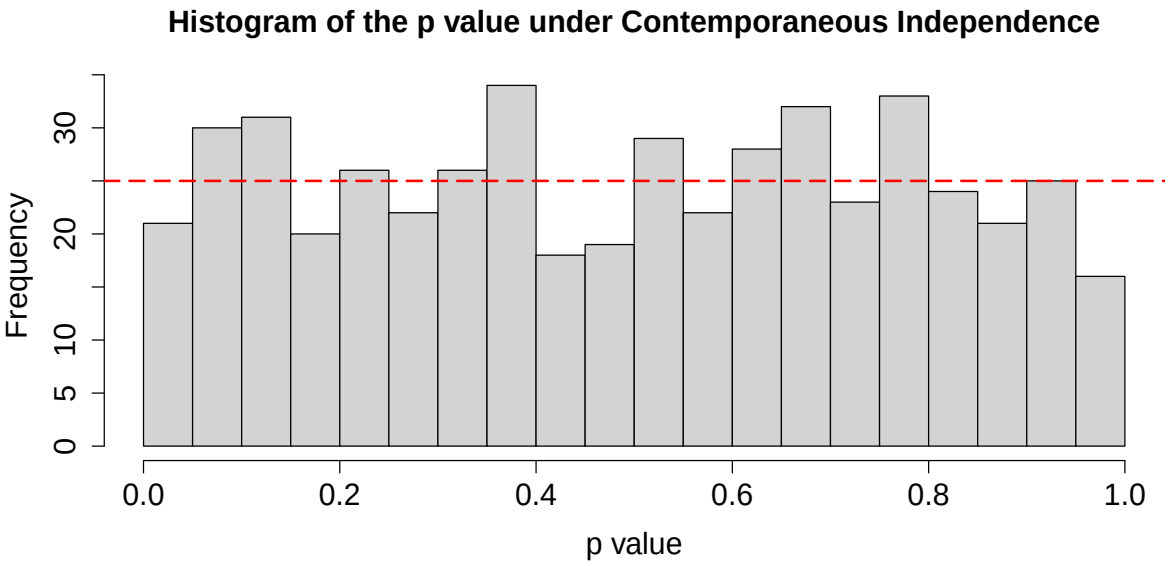


Figure 4.12: The histogram of the p-value from each of the 500 repetitive experiments under contemporaneous independence for the Hierarchical Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

ability for each arrow among the chains based on the inferred regression coefficients $\boldsymbol{\beta}$, and present the results with coloured arrows in Figure 4.13. It is clear that the vertical arrows connecting Z_1, Z_2 and Z_3 are shown in red with high transparency, indicating

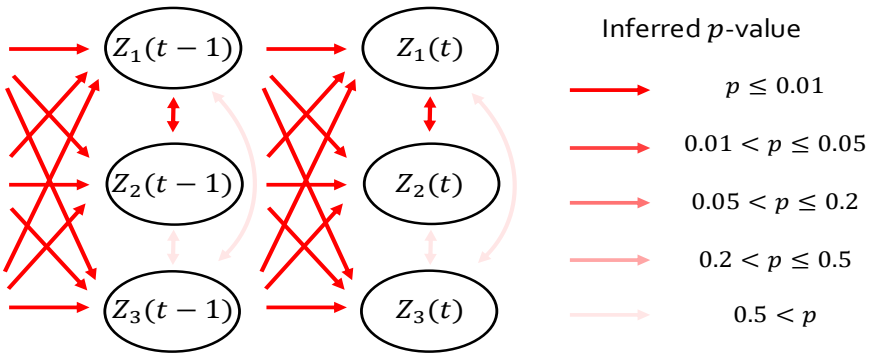


Figure 4.13: An illustration of the inferred p-values from Bayesian posterior contour probabilities, shown in red with varying levels of transparency to reflect the magnitude of the p-values and, correspondingly, the strength of the inferred dependencies. The p-value of the vertical arrows connecting Z_3 with Z_1 and Z_2 are marked in red with high transparency, indicating large corresponding p-values ($p \geq 0.5$) and the inherent contemporaneous independence between them, while the rest arrows are marked with solid red, representing low p-values ($p \leq 0.01$) and the existence of the dependencies.

high p-values and supporting that Z_1 and Z_2 are contemporaneously independent from Z_3. In contrast, the remaining arrows are shown in solid red, corresponding to low p-values and suggesting the presence of dependency structures. Overall, this p-value based depiction of the dependency structure is consistent with Figure 3.5 that illustrates the contemporaneous independence configuration for the simulation.

(III) Granger non-Causality

Recall the Granger non-causality illustrated in Figure 4.14, where chain 1 and chain 2 are not Granger caused by chain 3:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1) \tag{4.41}$$

This independence is characterized by the missing black arrow pointing $Z_1(t)$ and $Z_2(t)$ from $Z_3(t-1)$, in comparison with the full model in Figure 4.6.

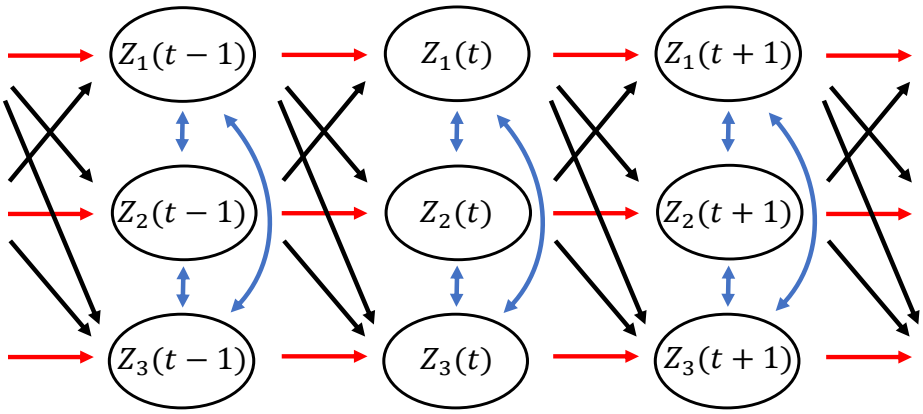


Figure 4.14: An illustration of an MMC with 3 chains under the Granger non-causality where chain 1 and chain 2 are not Granger caused by chain 3.

Under Granger non-causality (4.41), the regression coefficients β accounting for the effects from chain 3, namely β_2, β_4, β_6 and β_8, from $\boldsymbol{\beta}_j$ for chain group $g_j = \{1\}, \{2\}, \{1,2\}$ should

be 0. That is:

$$\boldsymbol{\beta}_j = (\beta_1, 0, \beta_3, 0, \beta_5, 0, \beta_7, 0) \qquad \text{for} \quad g_j = \{1\}, \{2\}, \{1,2\} \tag{4.42}$$

Hence, for illustration purpose, consider the regression coefficients $\boldsymbol{\beta}_j$ for:

1. the univariate chain group $g_j = \{1\}$
2. the bivariate chain group $g_j = \{1, 2\}$

From both figures, it is evident that both the Frequentist (MLE) and the Bayesian (posterior median) yield accurate estimates of the true regression coefficients $\boldsymbol{\beta}_j$ for the univariate and bivariate marginals under the Granger non-causality assumption, as the true values lie very close to the medians of the box plots. Moreover, it is worth noting that the coefficients $\beta_2, \beta_4, \beta_6, \beta_8$, which represent the effects from chain 3, should all be zero within the regression coefficients $\boldsymbol{\beta}_j$ of the chain groups $\{1\}$, $\{2\}$, and $\{1, 2\}$ under the

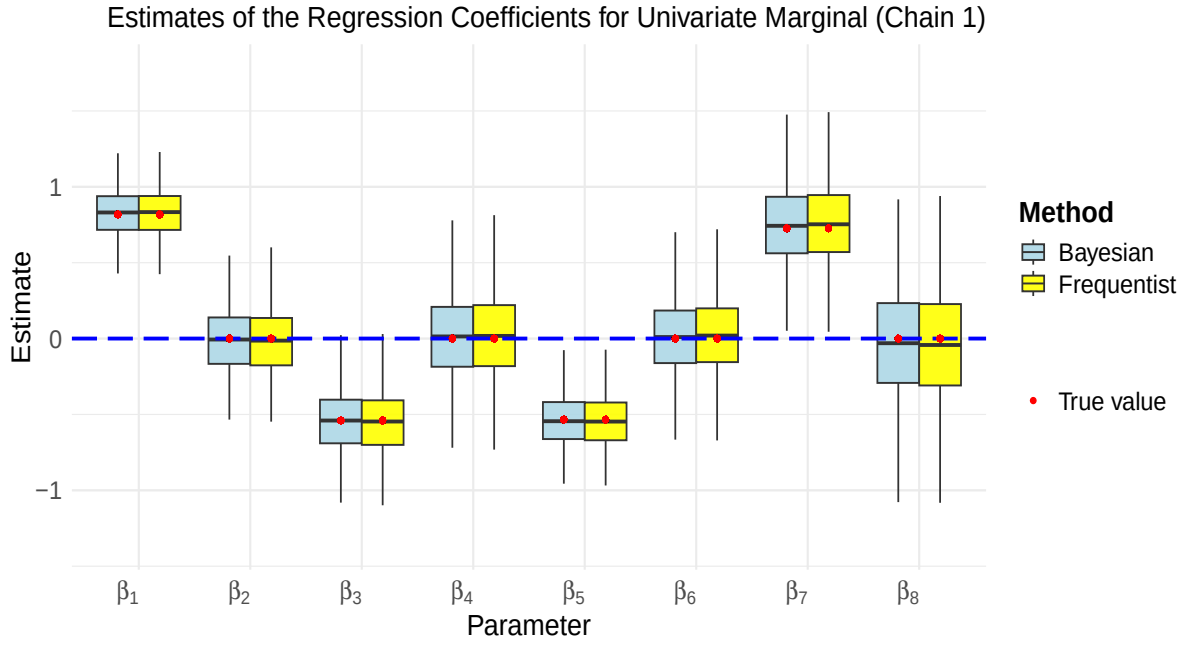


Figure 4.15: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment. Under the Granger non-causality, $\beta_2, \beta_4, \beta_6, \beta_8$ that accounts for the effect from chain 3 should be 0.

Granger non-causality specified in Equation 4.41. This pattern can be explicitly observed as such in Figures 4.15 and 4.16.

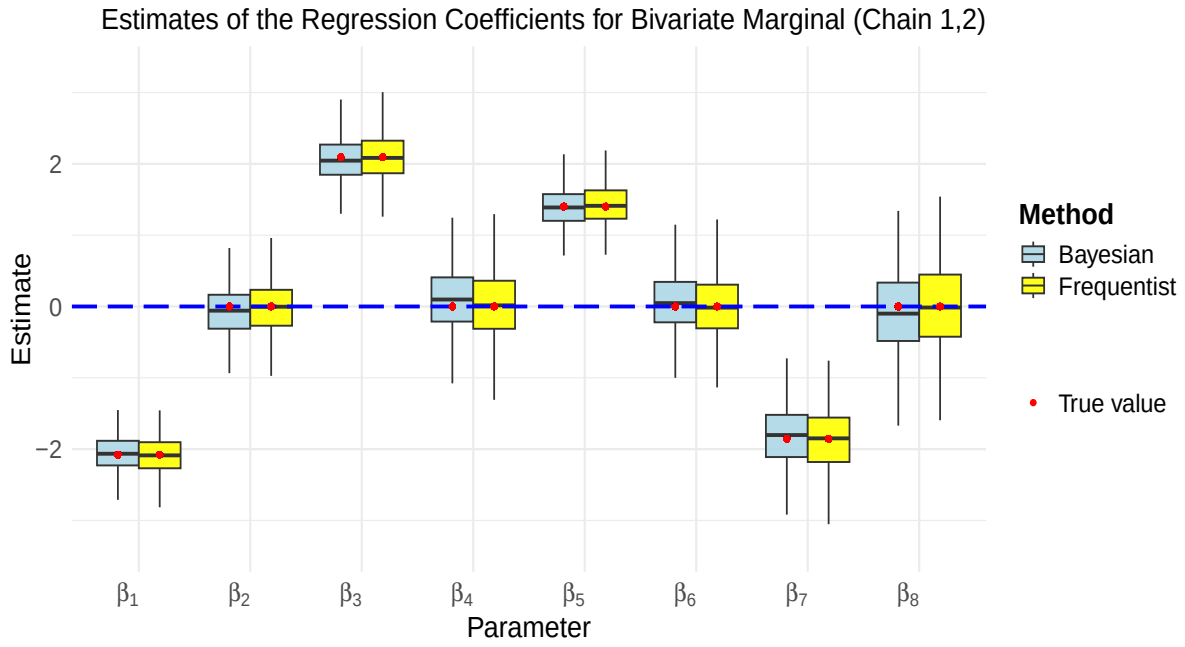


Figure 4.16: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the bivariate chain group $g_j = \{1, 2\}$, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment. Under Granger non-causality, $\beta_2, \beta_4, \beta_6, \beta_8$ that accounts for the effect from chain 3 should be 0.

To test the Granger non-causality in Frequentist approach, we perform a likelihood ratio test within each of the 500 repeated experiments for the following nested model:

$$H_0: \quad Z_1 \text{ and } Z_2 \text{ are not Granger caused by } Z_3$$

$$H_1: \quad Z_1 \text{ and } Z_2 \text{ are Granger caused by } Z_3$$

The test statistic $-2 \log \Lambda$ is the same as is given in Equation (4.40) but with a degree of freedom of $\nu = 4 \times 3 = 12$, since under the restricted model H_0 (Granger non-causality), four coefficients in each $\boldsymbol{\beta}_j$ for $g_j = \{1\}, \{2\}, \{1, 2\}$ are constrained to be zero.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 26 out

of 500 experiments (5.2%), which is fully consistent with the random variation under H_0. Furthermore, we record the p-value from each experiment and present their distribution in the histogram as the following Figure 4.17:

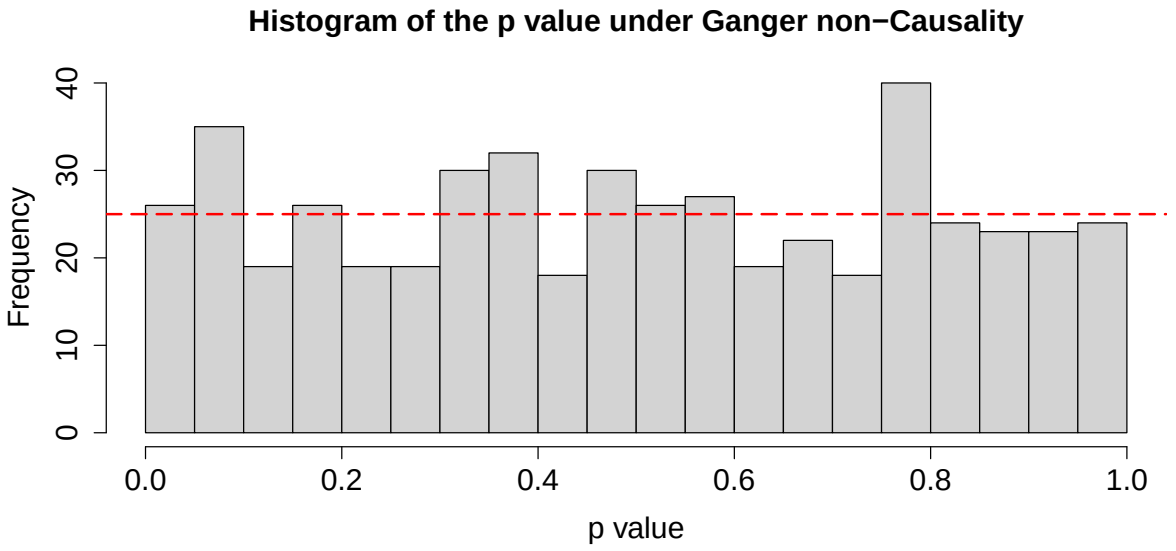


Figure 4.17: The histogram of the p-value from each of the 500 repetitive experiments under Granger non-causality for the Hierarchical Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

As shown, the 500 observed p-values are approximately uniformly distributed over $[0, 1]$, closely matching the Uniform distribution indicated by the red dashed line. This suggests that the test is well-calibrated and that the null hypothesis H_0 of Granger non-causality is plausible.

Alternatively in Bayesian fashion, we compute the corresponding posterior contour probability for each arrow among the chains based on the inferred regression coefficients $\boldsymbol{\beta}$, and present the results with coloured arrows in Figure 4.18. It is clear that the arrows from Z_3 to Z_1 and Z_2 are shown in red with high transparency, indicating high p-values and supporting Granger non-causality from Z_3 to Z_1 and Z_2. In contrast, the remaining arrows are shown in solid red, corresponding to low p-values and suggesting the presence of dependency structures. Overall, this p-value based depiction of the dependency structure

is consistent with Figure 4.14, which illustrates the Granger non-causality configuration for the simulation introduced at the beginning of this section.

(IV) Mixed Dependency

Finally, consider the a mixed dependency structures with both contemporaneous independence and the Granger non-causality, as illustrated in Figure 4.19, where we assume:

- **Contemporaneous Independence**: Chain 1 and chain 2 are contemporaneously independent from chain 3 (see Equation (4.38)), characterized by the missing black arrow pointing $Z_1(t)$ and $Z_2(t)$ from $Z_3(t-1)$, in comparison with the full model.

- **Granger non-Causality**: Chain 1 and chain 2 are not Granger caused by chain 3 (see Equation (4.41)), characterized by the missing blue vertical arrow connecting $Z_1(t)$ and $Z_2(t)$ with $Z_3(t)$, in comparison with the full model in Figure 4.6.

Under Hierarchical Marginal Model, the regression coefficients should satisfy both the constraints from the contemporaneous independence in Equation (4.39) and Granger non-causality (4.42). Hence, for illustration purpose, we choose to visualize:

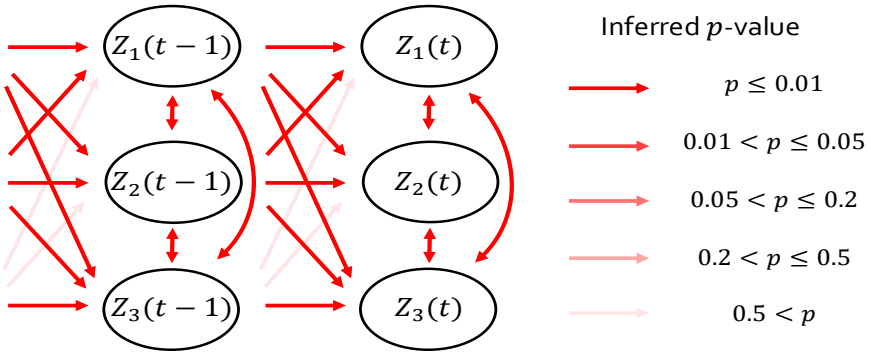


Figure 4.18: An illustration of the inferred p-values from Bayesian posterior contour probabilities, shown in red with varying levels of transparency to reflect the magnitude of the p-values and, correspondingly, the strength of the inferred dependencies. The p-value of the arrows from Z_3 to Z_1 and Z_2 are marked in red with high transparency, indicating large corresponding p-values ($p \geq 0.5$) and the Granger non-causality between them, while the rest arrows are marked with solid red, representing low p-values ($p \leq 0.01$) and the existence of the dependencies.

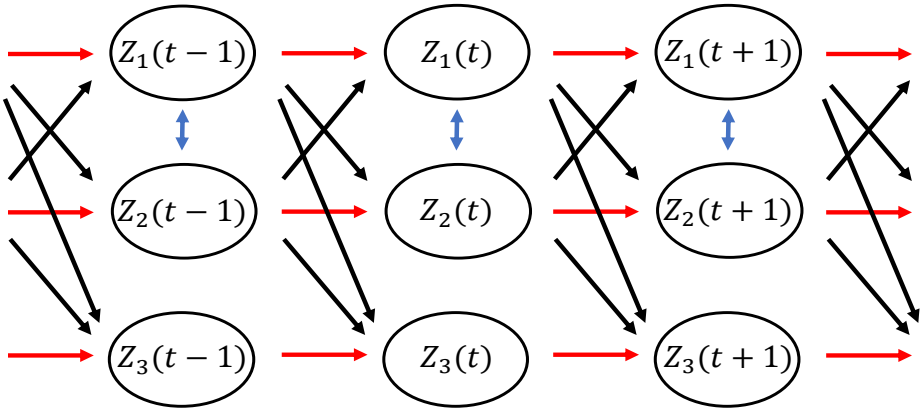


Figure 4.19: An illustration of an MMC with 3 chains under the mixed dependencies including both Contemporaneous Independence where chain 1 and chain 2 are contemporaneously independent from chain 3 and the Granger non-causality where chain 1 and chain 2 are not Granger caused by chain 3.

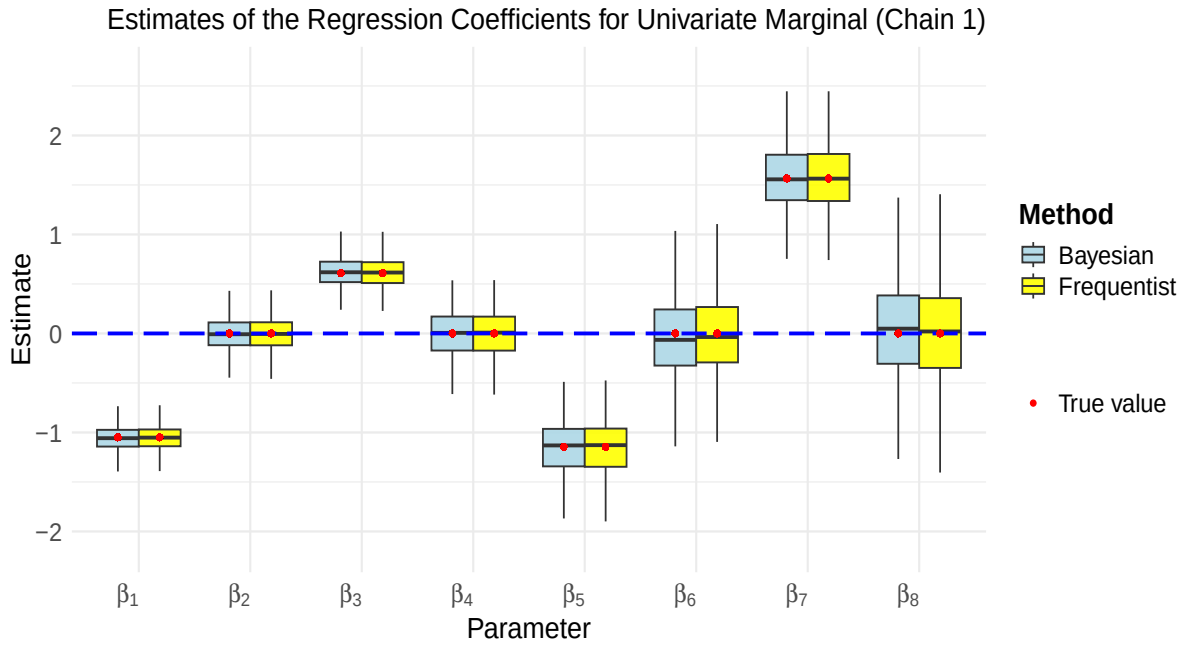
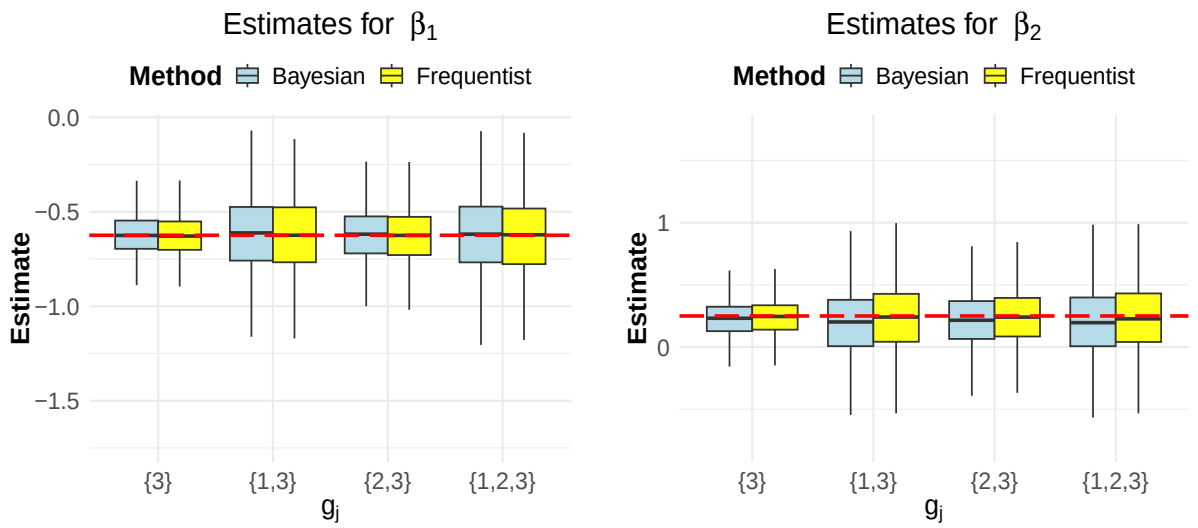


Figure 4.21: An illustration of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$, obtained via both the Bayesian approach (in light blue) and the Frequentist approach (in yellow), with the true values indicated by red points. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment. Under Granger non-causality, $\beta_2, \beta_4, \beta_6, \beta_8$ that accounts for the effect from chain 3 should be 0.

- The estimates of β_1, β_2 (i.e. first two elements in the vector of regression coefficients



(a) The posterior means of β_1

(b) The posterior means of β_2

Figure 4.20: An illustration of the estimated regression coefficients β_1 and β_2 across the chain groups $g_j = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$ from Bayesian (in light blue) and Frequentist (in yellow), with the true value indicated by the red dashed line. The Bayesian estimates are computed as the posterior median for each repeated experiment, while the Frequentist estimates are computed as the MLE for each repeated experiment. Under Contemporaneous Independence, these estimated regression coefficients β should agree in the mean value.

$\boldsymbol{\beta}_j$) across the chain group $g_j = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$ (for Contemporaneous Independence).

- The estimates of $\boldsymbol{\beta}_{\{1\}}$ for the marginal of chain 1 (for Granger non-causality)

From both figures, it is evident that both the Frequentist (MLE) and the Bayesian (posterior median) yield accurate estimates of the true regression coefficients $\boldsymbol{\beta}_j$ for the univariate and bivariate marginals under the mixed dependency, as the true values lie very close to the medians of the box plots. Moreover, it is worth noting that:

- The first and the second coefficients β_1 and β_2 across the chain groups $g_j = \{3\}$, $\{1,3\}$, $\{2,3\}$, $\{1,2,3\}$, as illustrated in Figure 4.20a and 4.20b respectively, almost agrees with each other for the median, which is consistent as that under Contemporaneous Independence (4.39).

- The coefficients $\beta_2, \beta_4, \beta_6, \beta_8$ in Figure 4.21, which represent the effects from chain 3, are close to 0, which is also consistent as that under Granger non-causality (4.42).

To test the mixed dependency structures of contemporaneous independence and the

Granger non-causality in Frequentist approach, we perform a likelihood ratio test within each of the 500 repeated experiments for the following nested model:

$$
\begin{aligned}
H_0: \quad & Z_1 \text{ and } Z_2 \text{ are not Granger caused by } Z_3 \quad \textbf{and} \\
& Z_1 \text{ and } Z_2 \text{ are contemporaneously independent from } Z_3 \\
H_1: \quad & \text{There is no such independence relationships}
\end{aligned}
$$

The test statistic $-2\log\Lambda$ is the same as is given in Equation (4.40) but with a degree of freedom of $\nu = 24 + 12 = 36$ since there are 24 constraints from Contemporaneous Independence and 12 from the Granger non-causality.

Using a fixed significance level of $\alpha = 0.05$, the null hypothesis H_0 is rejected in 14 out of 500 experiments (2.8%). We further record the p-value from each experiment and present their distribution in the histogram as the following Figure 4.22:

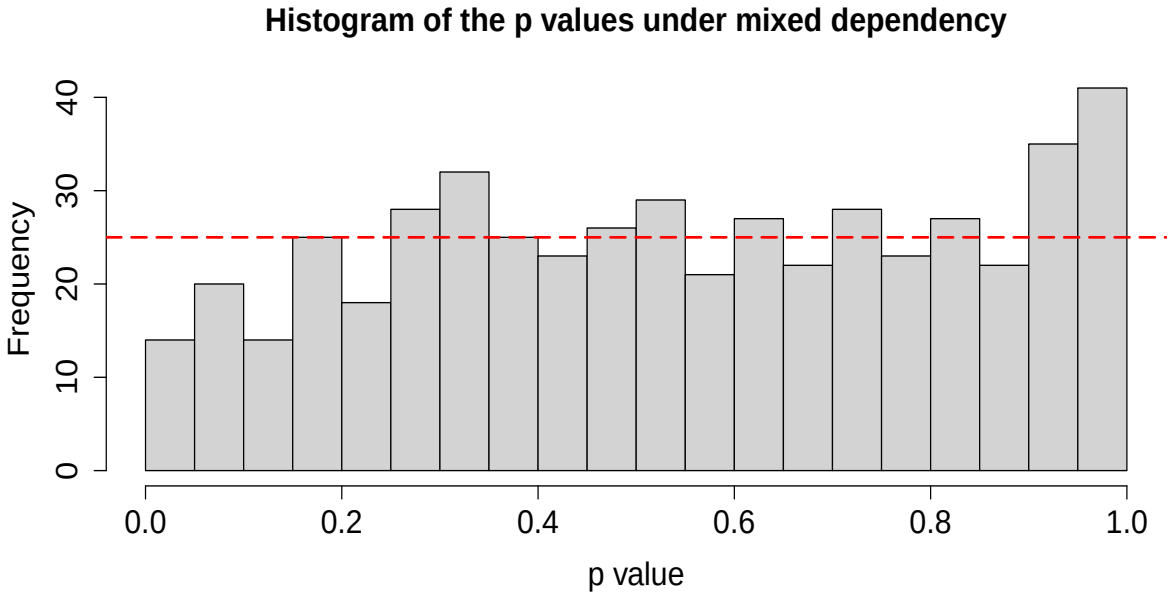


Figure 4.22: The histogram of the p-value from each of the 500 repetitive experiments under mixed dependencies for the Hierarchical Marginal Model, with the red dashed line indicating the uniform distribution between 0 and 1.

As shown, the 500 observed p-values are approximately uniformly distributed over $[0, 1]$

with a slight skewness towards 1, indicating a larger p is more likely to be observed in the experiments and the null hypothesis H_0 for the mixed independency is favoured.

Alternatively, in the Bayesian framework, we compute the posterior contour probability for each arrow among the chains based on the inferred regression coefficients $\boldsymbol{\beta}$, and present the results using coloured arrows in Figure 4.23. The arrows from Z_3 to Z_1 and Z_2, as well as the vertical arrows connecting Z_1, Z_2, and Z_3, are shown in red with high transparency, indicating high corresponding p-values and supporting both the Granger non-causality from Z_3 to Z_1 and Z_2 and the contemporaneous independence between Z_1, Z_2, and Z_3, respectively. In contrast, the remaining arrows appear in solid red, corresponding to low p-values and indicating the presence of dependency structures. Overall, this p-value based depiction of the dependency structure fully recovers the configuration shown in Figure 4.19, which illustrates the mixed dependency structures used in the simulation introduced at the beginning of this section.

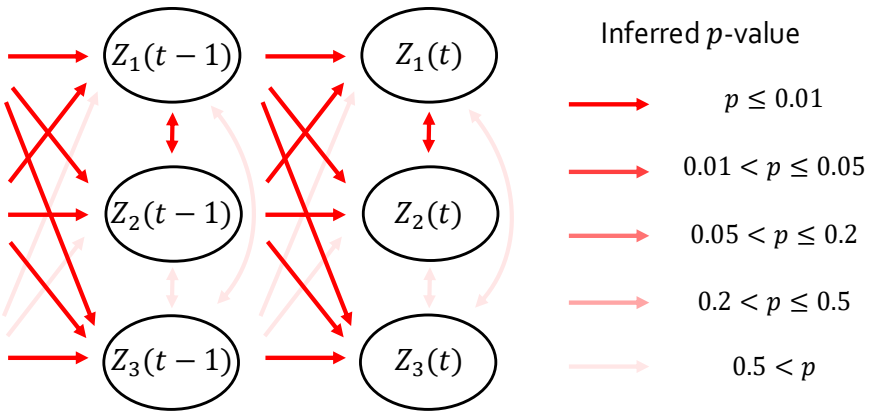


Figure 4.23: An illustration of the inferred p-values from the Bayesian posterior contour probabilities is shown in red, with varying levels of transparency indicating the magnitude of the p-values and, correspondingly, the strength of the inferred dependencies. The arrows from Z_3 to Z_1 and Z_2, as well as the vertical arrows connecting Z_1 and Z_2 with Z_3, are displayed in red with high transparency, indicating large p-values ($p \geq 0.5$) and reflecting Granger non-causality and contemporaneous independence, respectively. In contrast, the remaining arrows are shown in solid red, representing low p-values ($p \leq 0.01$) and indicating the presence of dependencies.

4.6 Discussion

In Chapters 3 and 4, we introduced two novel frameworks for modelling dependency structures among chain groups in an MMC: the *Hierarchical Marginal Model* and the *Univariate Marginal Model*. Figure 4.24 illustrates the reparametrisation considered by these two models. Both frameworks start from the joint transition probabilities p used by the Cartesian Product Model, which naturally captures arbitrary dependency structures at both global and marginal levels. The Univariate Marginal Model factorizes the joint transition probabilities p into auxiliary probabilities ζ_σ following a structure analogous to a Bayesian network, under a specific permutation σ. Proposition 3.9 confirms that the map $\mathcal{P} \to \mathcal{H}_\sigma$ from the full parameter space $\mathcal{P}$ of the joint transition matrices to the parameter space $\mathcal{H}_\sigma$ of auxiliary probabilities ζ_σ under the specified permutation σ is also diffeomorphism, thereby ensuring that dependency structures encoded within the

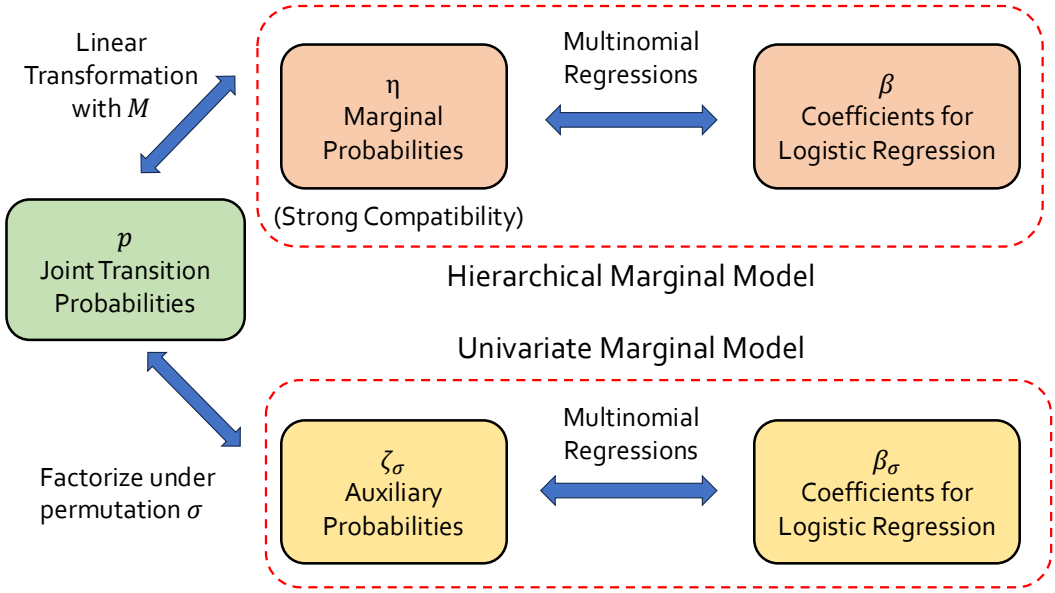


Figure 4.24: An illustration of the parametrisations used in the Hierarchical Marginal Model and the Univariate Marginal Model. Both approaches begin with the joint transition probabilities p. The *Hierarchical Marginal Model*, shown at the top in light red, first derives the marginal transition probabilities η through a linear transformation using the marginal indicator matrix M. It then incorporate the multinomial logistic regression to further decompose these η into regression coefficients β. In contrast, the *Univariate Marginal Model*, shown at the bottom in light yellow, factorizes p into auxiliary probabilities ζ_σ based on a specified permutation σ. Like the Hierarchical Marginal Model, these ζ_σ are subsequently expressed through multinomial logistic regression as the regression coefficients β_σ.

joint transition matrix P are equivalently represented within the auxiliary probabilities ζ_σ. Alternatively, the Hierarchical Marginal Model considers a set of hierarchically constructed marginal probabilities η satisfying *strong compatibility*, and relating them to the joint transition probabilities p through a linear transformation defined by the marginal indicator matrix M. Proposition 4.7 demonstrates that this transformation $\mathcal{P} \to \mathcal{H}_1$ from the full parameter space P of the joint transition matrices to the restricted parameter space $\mathcal{H}_1$ of the marginal transition matrices under strong compatibility is a diffeomorphism, ensuring that all dependency structures embedded in p are fully reflected within the marginals η.

As illustrated in Figure 4.24, we subsequently incorporated multinomial logistic regression frameworks for both the auxiliary probabilities ζ_σ under a specified permutation σ in the Univariate Marginal Model and the marginal transition probabilities η in the Hierarchical Marginal Model. These probabilities are thus further decomposed via a multinomial logistic link function into linear combinations of marginal effects from all the marginal groups. This reparametrisation not only offers clear interpretability of these coefficients and enables straightforward encoding of dependency structures comparing to the Cartesian Product Model mentioned in 2.8.2, but also provides an alternative way to define the dependency structures through the corresponding regression coefficients (see Definition 3.15 and 4.14 for Univariate Marginal Model and Hierarchical Marginal Model respectively). It is worth noting that alternative link functions, such as the probit link, can also be employed, and the multinomial logistic link is merely one modelling choice among many.

Additionally, Chapters 3 and 4 discuss the utilization of two novel marginal models, the Hierarchical Marginal Model and the Univariate Marginal Model, from both generating and inference, as illustrated in Figure 3.1 and 4.1 respectively. For the generative model, a hierarchical construction scheme is developed for the regression coefficients $\boldsymbol{\beta}$ in the Hierarchical Marginal Model. This scheme is based on the idea that the corresponding marginal transition probabilities is sequentially constructed according to the number of

chains in the marginal groups to ensure the strong compatibility. In contrast, the Univariate Marginal Model admits a more straightforward generation procedure, since the regression coefficients $\boldsymbol{\beta}_\sigma$ in this setting are fully unconstrained.

Finally, a comprehensive simulation study demonstrates that the regression coefficients in both the Hierarchical Marginal Model and the Univariate Marginal Model can be accurately recovered under all four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures. The results show that hypothesis tests on the structures of the regression coefficients $\boldsymbol{\beta}$ from the Hierarchical Marginal Model provide a powerful tool for detecting and validating the potential dependency structures embedded within the MMC.

We will discuss below specifically for the Hierarchical Marginal Model introduced in this chapter, highlighting its respective strengths and limitations in Section 4.6.1, followed by an overall model comparison in Section 4.6.2.

4.6.1 Discussion on the Hierarchical Marginal Model

The *Hierarchical Marginal Model* considers marginal transition probabilities and constructs corresponding multinomial logistic regressions hierarchically, offering several key advantages:

- Unlike the Univariate Marginal Model, the Hierarchical Marginal Model does not rely on a predefined permutation σ. Moreover, it requires fitting only a single model to infer all combinations of embedded dependency structures, while the Univariate Marginal Model may necessitate fitting multiple models under different permutations in order to capture the full range of dependencies.

- Each regression coefficient β in the Hierarchical Marginal Model has a clear interpretability: it represents the log odds of the marginal effect of one subgroup of chains being in an alternative state at the previous time on another subgroup of chains being in a particular state at the current time step.

- By employing hierarchically constructed multinomial logistic regressions, Hierarchical Marginal Model avoids the issue of exponentially increasing response levels. Although this approach requires an increasing number of multinomial logistic regressions as a compensation, it significantly reduces computational costs, as demonstrated in Section A.3.2.

- The marginal transition probabilities modelled by Hierarchical Marginal Model are directly utilized in subsequent inference processes, as will be discussed later in Section 5.3.3.

Despite the advantages of the Hierarchical Marginal Model, it does have certain drawbacks. Primarily, it relies on the *strong compatibility* 4.1, which requires that all the marginal transition probabilities modelled by the Hierarchical Marginal Model must be consistent with each other. However, the marginal transition probabilities, obtained by the multinomial logistic regression, are not inherently guaranteed to be compatible under an unconstrained parameter space for the regression coefficients $\boldsymbol{\beta}$. Specifically, certain values of $\boldsymbol{\beta} \in \mathbb{R}^D$ may lead to inconsistencies among the marginal probabilities generated by the multinomial logistic regressions, potentially yielding negative entries in the reconstructed joint transition matrix P. Therefore, when employing the Hierarchical Marginal Model to generate an MMC from $\boldsymbol{\beta}$, it is necessary to determine the corresponding admissible boundaries, in order to guarantee that the resulting η satisfies the strong compatibility.

Such constraints on the regression coefficients $\boldsymbol{\beta}$ can be obtained by transforming the constraints on the marginal transition probabilities η induced by lower dimensional marginals via the multinomial logistic link function. And this restriction on η can be explicitly derived according to the Inclusion-Exclusion formula, as elaborated in Section 4.2.1. Proposition 4.15 further ensures that these constraints do not interfere with the specification of dependency structures. Consequently, we can construct a hierarchical generative model that develops from lower-dimensional to higher-dimensional marginals, to simulate the MMC $\boldsymbol{Z}$ according to the desired dependency structures.

It is also worth noting that, in the inference direction, strong compatibility is automatically satisfied. This is because the states $\boldsymbol{Z}$ necessarily correspond to a valid joint transition matrix P. Any regression coefficients $\boldsymbol{\beta}$ inferred from the data are therefore implicitly induced by marginal transition probabilities η obtained via marginalisation of this valid P, and hence satisfy the strong compatibility constraints by construction.

In practice, with finite observations, point estimates of $\boldsymbol{\beta}$, such as posterior medians, may occasionally lie slightly outside the strongly compatible region due to Monte Carlo variability. Although this issue did not arise in the simulation studies in this thesis, it may occur when some entries of the joint transition matrix P are very small (i.e. close to zero). In such cases, inferred regression coefficients $\hat{\boldsymbol{\beta}}$ can produce corresponding estimates $\hat{p}$ that are slightly negative but close to zero. A simple and practical remedy is to truncate these values at zero, renormalize the remaining entries so that the transition probabilities sum to one, and then transform the adjusted transition matrix back to obtain modified regression coefficients $\tilde{\boldsymbol{\beta}}$ as the final estimates. This procedure preserves the overall structure of the estimates while restoring strong compatibility.

4.6.2 Comparison between Different Models

In this section, we compare the *Hierarchical Marginal Model* and the *Univariate Marginal Model* from the following 8 key perspectives, with a summary of the results provided in Table 4.1.

1. **The response in the multinomial logistic regression:** Both of these models start with the marginals for some $k \in [K]$

$$\eta_k = \mathbb{P}(Z_k(t) \mid \boldsymbol{Z}(t-1))$$

and try to link it with the joint transition probability $p = \mathbb{P}(\boldsymbol{Z}_k(t)|\boldsymbol{Z}_k(t-1))$ with some additional probabilities.

   (a) **Hierarchical Marginal Model** additionally considers the hierarchically con-

structed marginals satisfying *strong compatibility*. For all $g_j = \{k_1, \ldots, k_i\} \in [K]$, the multivariate marginal probability η_j is defined as:

$$\eta_j = \mathbb{P}(Z_{k_1}(t), \ldots, Z_{k_i}(t) \mid \boldsymbol{Z}(t-1))$$

(b) **Univariate Marginal Model** relies on the predefined permutation σ on the groups of chains $[J]$ and considers the auxiliary probabilities $\zeta_{\sigma(j)}$ for $j \in [J]$ with:

$$\zeta_{\sigma(j)} = \mathbb{P}\left(\boldsymbol{Z}_{\sigma(j)}(t) \mid \boldsymbol{Z}_{\sigma(1:(j-1))}(t), \boldsymbol{Z}(t-1)\right)$$

where the additional concurrent terms $\boldsymbol{Z}_{\sigma(1:(j-1))}(t)$ are considered within the conditioning part.

2. **The number of regression and explanatory variables**: Consider the general case where we have K chains and a shared state space $\mathcal{S}$ with cardinality $|\mathcal{S}| = s$

(a) **Hierarchical Marginal Model**: The number of explanatory variables per regression is fixed to be s^K as it involves only $\boldsymbol{Z}(t-1)$ term in the conditioning part. We decompose the response into all possible marginals: for univariate marginals, we have $\binom{K}{1}$ such marginals, each with $s-1$ non-baseline categories. For k-variate marginals, there are $\binom{K}{k}$ combinations, and each has $(s-1)^k$ alternative states excluding the baseline levels as we only model them for the purpose of identifiability. Summing these over k from 1 to K yields number of regressions required by the Hierarchical Marginal Model as:

$$\sum_{k=1}^{K} \binom{K}{k} (s-1)^k \;=\; \big((s-1)+1\big)^K - 1 \;=\; s^K - 1,$$

It is intuitive to see that this model requires need $(s^K - 1) \times s^K$ free regression coefficients to fully specify the $s^K \times s^K$ joint transition matrix, matching the number of independent parameters in the joint transition matrices.

(b) **Univariate Marginal Model**: Only J different multinomial logistic regressions with each corresponding to the marginal of one group of chains are considered. The dimension of the response is set to $s^{|g_j|}$ levels, with $|g_j|$ refers to the number of chains within group g_j. The number of explanatory variables, however, depends on which chain's marginal is being modelled. For $\sigma(1)$, only $\boldsymbol{Z}(t-1)$ is included in the conditioning set, resulting in s^k explanatory variables. For $\sigma(j)$ where $2 \leq j \leq J$, we add an additional concurrent term $\boldsymbol{Z}_{\sigma(1:(j-1))}(t)$, leading to $s^{K+\sum|g_{1:(j-1)}|}$ explanatory variables. Summing across J multinomial logistic regression gives:

$$\begin{aligned}\sum_{j=1}^{J} s^{K+\sum|g_i|} \times (s^{|g_j|}-1) &= s^K\left[s^K - s^{\sum|g_{1:(J-1)}|} + s^{\sum|g_{1:(J-1)}|} - \cdots - 1\right] \\ &= s^K\,(s^K-1)\end{aligned}$$

independent regression coefficients. It exactly matches the dimension required to construct a $s^K \times s^K$ joint transition matrix.

3. **Unconstrained parameter spaces**: In the Hierarchical Marginal Model, the parameter space $\mathcal{B}_1$ for $\boldsymbol{\beta}$ is restricted since it is induced from the $\mathcal{H}_1$ that contains marginal transition probabilities satisfying the *strong compatibility*, in order to recover a valid joint transition matrix. While in the Univariate Marginal Model, the parameter space $\mathcal{B}$ for $\boldsymbol{\beta}$ is fully unconstrained, covering all of $\mathbb{R}$.

4. **Involving permutation:** The Hierarchical Marginal Model is symmetric and permutation-invariant and hence does not involve any permutation in its parametrisation. In contrast, the Univariate Marginal Model requires a carefully chosen permutation σ of the J chain groups to factorize the joint transition matrix P into the corresponding auxiliary probabilities ζ_σ.

5. **Interpretation of the multinomial logistic coefficients $\boldsymbol{\beta}$:** Each model defines its response differently, so the interpretation of $\boldsymbol{\beta}$ varies accordingly:

(a) **Hierarchical Marginal Model:** $\boldsymbol{\beta}$ indicates how one or multiple chains in

the alternative states previously shifts the log-odds of the corresponding multivariate marginal probabilities.

(b) **Univariate Marginal Model:** $\boldsymbol{\beta}$ represents how a specific chain or combination of chains in the past as well as the concurrent ones determined by the permutation being in alternative states shifts the log-odds of the marginal probabilities of one group of chains.

6. **Impose Granger non-causality and contemporaneous independence**:

(a) In **Hierarchical Marginal Model**, Contemporaneous independence is imposed via equating the corresponding regression coefficients among certain regressions, while Granger non-causality is imposed through restricting the corresponding coefficients within a single multinomial logistic regression to zero.

(b) In **Univariate Marginal Model**, we can encode both Granger non-causality or contemporaneous independence by simply setting specific coefficients in $\boldsymbol{\beta}_\sigma$, which would initially be unconstrained under full independency, to zero, under a carefully chosen permutation σ according to the target dependency structures.

7. **Computational Cost for the Polya Gamma Data augmentation**: As reviewed in Section A.3.2, Polson et al. [2013] explicitly presented the computational cost to estimate the regression coefficients $\boldsymbol{\beta}$ for each iteration as:

$$O\big(n\, p^2\, l^2 + p^3\, l^3\big) \tag{4.43}$$

Under the Hierarchical Marginal Model, substituting $n = T - 1$, $l = s$ and $p = s^k$, and noting there are $(s^k - 1)$ such regressions, the total computational cost per iteration is given as:

$$O\big((s^k - 1)\left[(T-1)\, s^{2k}\, s^2 + s^{3k}\, s^3\right]\big) = O((T-1)s^{3k+2} + s^{4k+3})$$

By contrast, in the Univariate Marginal Model, we fit j regressions, one for each group. Each regression with a different number of predictors. Let $|g_j|$ denote the number of chains within the chain group j under identity permutation. Plugging $l = s^{|g_1|}, \ldots, s^{|g_J|}$ and $p = s^k, s^{k+|g_1|}, \ldots, s^{k+\sum_{j=1}^{J}|g_j|}$ for each of the regressions into Equation (4.43) yields:

$$
\begin{aligned}
&O\Big((T-1)\left[s^{2|g_1|}s^{2k} + \cdots s^{2|g_J|}s^{2(k+\sum_{j=1}^{J-1} g_j)}\right] + \left[s^{3|g_1|}s^{3k} + \cdots s^{3|g_J|}s^{3(k+\sum_{j=1}^{J-1} g_j)}\right]\Big) \\
&O\Big((T-1)\left[s^{2|g_J|}s^{2(k+\sum_{j=1}^{J-1} g_j)}\right] + \left[s^{3|g_J|}s^{3(k+\sum_{j=1}^{J-1} g_j)}\right]\Big) \\
=&O\Big((T-1)\, s^{4k} \;+\; s^{6k}\Big)
\end{aligned}
$$

We observe that the computational cost is dominated by the last term in the summation, corresponding to the regression of the final chain group g_J, which involves the most covariates. Notably, the final expression of the computational cost in the Univariate Marginal Model is independent of the permutation σ, as it does not involve any group index j. In general, we have:

$$
O\Big((T-1)s^{3k+2} + s^{4k+3}\Big) < O\Big((T-1)\, s^{4k} \;+\; s^{6k}\Big)
$$

This indicates that inferring the Univariate Marginal Model is more computationally expensive than the Hierarchical Marginal Model. Intuitively, Univariate Marginal Model merges multiple smaller multinomial logistic regressions into a larger one, causing an increase of the number of predictors p that appear in quadratic and cubic form in the Equation (4.43) instead of increasing the number of regressions that appears linearly in the equation, which as a result increasing the overall computational cost.

8. **Incorporation with the *Generalised Individual Forward–Backward* (GIFB) Algorithm**: As will be discussed in Section 5.3.4, the marginal transition probabilities η from the Hierarchical Marginal Model can be directly used in the recursions of the GIFB algorithm, providing an efficient way to infer the hidden chains $\boldsymbol{Z}$ in

the unobserved setting of Chapter 5. In contrast, the auxiliary probabilities ζ_σ from the Univariate Marginal Model cannot be directly incorporated.

Model	Hierarchical Marginal Model	Univariate Marginal Model
The Target	Marginal transition probabilities η_j	Auxiliary probabilities $\zeta_{\sigma(j)}$ under permutation σ
Regressions Response levels Explanatory variables	$s^K - 1$ s fixed to be s^k	J $s^{\lvert g_j \rvert}$ for each $j \in [J]$ $s^{\lvert g_{1:(j-1)} \rvert}$ for each $j \in [J]$
Parameter Space	Constrained	Unconstrained
Permutation Involved	No	Yes
Interpretation of beta	Log odds of the different combinations of multivariate marginals	Log odds of the group-wise marginals conditioned additionally on concurrent states
Granger non-Causality Contemporaneous Independence	Set the corresponding $\boldsymbol{\beta}$ to 0 Equate $\boldsymbol{\beta}$ among different regressions	Set the corresponding $\boldsymbol{\beta}$ to 0 Set the corresponding $\boldsymbol{\beta}$ to 0
Computational Cost for Pólya Gamma	$O\big((T-1)s^{3k+2} + s^{4k+3}\big)$	$O\big((T-1)\, s^{4k} \;+\; s^{6k}\big)$
Incorporate with GIFB	Yes	No

Table 4.1: Comparison between the Hierarchical Marginal Model and the Univariate Marginal Model

Chapter 5

Multivariate Hidden Markov Chain

5.1 Introduction

In the previous chapters, we considered the setting where the realizations of the Multivariate Markov Chain (MMC) are directly observed. In many practical applications, however, the underlying chain is unobserved (hidden), and only an associated process is observed instead. This naturally leads to the framework of Hidden Markov Models (HMMs), where inference on the latent process is carried out through the observed data. Such settings are commonly adopted in various fields in the literature, for instance in speech recognition [Rabiner, 1989], computational biology [Eddy, 1996], and finance [Hamilton, 1989], to name a few.

Building upon the previous analysis of dependency structures in MMCs, in this chapter we extend the framework to the hidden setting. Specifically, we consider Multivariate Hidden Markov Models (MHMMs), where the underlying chain is not directly observed and only data from an *emitted* process are available. Our focus is on developing inference methods, grounded in the marginal models introduced earlier, that allow us to recover the latent process $\boldsymbol{Z}$ and identify the dependency structures embedded within the hidden chains, despite relying solely on noisy and indirect observations.

In literature, the Expectation–Maximization (EM) algorithm is the canonical approach

for parameter estimation in Hidden Markov Models (HMMs). Its origins trace back to the seminal work of Baum and Petrie [1966] and the subsequent development of the *Baum–Welch* algorithm by Baum et al. [1970], which provides an efficient EM-based procedure designed for HMM. While the Baum-Welch algorithm was developed specifically for HMMs, Dempster et al. [1977] subsequently formalized the EM algorithm as a general framework for maximum likelihood estimation with latent variables, under which Baum-Welch can be viewed as a special case. Building on these foundations, Cappé et al. [2005] provide a rigorous analysis of the Baum-Welch algorithm, including its convergence properties, and extend the EM framework to broader settings such as continuous-time models.

Nevertheless, as an EM-based method, Baum–Welch exhibits several drawbacks. It only produces point estimates of the parameters, with no measure of posterior uncertainty. Its convergence can be slow in higher-dimensional models and it is prone to being trapped in local optima, with results sensitive to initialization.

These limitations have motivated the development of Bayesian alternatives based on Gibbs sampling, which iteratively draw samples from the full conditionals of the hidden states and model parameters. Foundational contributions by Chib [1996], Scott [2002], and Fearnhead and Sherlock [2006], among others, have established Gibbs sampling as a standard tool for Bayesian inference in HMMs. Unlike EM, Gibbs sampling enables sampling from the full posterior distribution, thereby providing not only point estimates but also credible intervals, while offering greater flexibility through the incorporation of prior information.

However, in the multivariate setting, Gibbs samplers that rely on the direct extension of the FB algorithm (see Section 5.3.2.4) incur exponentially growing computational costs as the number of hidden chains increases. That is, sampling the hidden states requires implementing the Forward Backward algorithm into a Cartesian Product Space, which scales exponentially with the number of chains (see Section 5.3.2.4). To address this, methods

based on block-wise updates [Spencer et al., 2015] and individual updates [Touloupou et al., 2020] have recently been proposed in the literature. These approaches reduce complexity via sequentially updating chains in a Gibbs manner. However, motivated by epidemic models, these approaches are restricted to the Coupled Hidden Markov Models (CHMMs), where conditional independence among the chains is inherently assumed. In other words, the recursions used in Touloupou et al. [2020] fundamentally rely on the conditional independence (see Assumption 2.2). When this condition does not hold and in the presence of Granger non-causality, the algorithm of Touloupou et al. [2020] cannot be used, thereby limiting its applicability in the general MHMM. This will be elaborated in Section 5.3.3.

To overcome these limitations, we propose the novel *Generalised Individual Forward Backward* (GIFB) algorithm. Building on the Gibbs-like updating schemes of Spencer et al. [2015] and Touloupou et al. [2020], GIFB generalises these ideas to the MHMM framework, where full dependencies are retained. Another key innovation of the proposed Algorithm is it operates on the marginal transition probabilities η rather than joint transition probabilities p within the recursion. This design naturally integrates with the Hierarchical Marginal Model introduced in Section 4.2, which focuses on reparameterising the joint transition probabilities p through the marginals η. Moreover, GIFB reduces the computational cost from exponential to quadratic growth, making inference computationally feasible in high-dimensional settings.

5.1.1 Chapter Layout

Section 5.2 introduces the basic Hidden Markov Model (HMM), starting from the univariate case (Section 5.2.1) to the multivariate extension (Section 5.2.2), under the assumption of an one-to-one emission process (Assumption 5.3). This allows us to focus on the potential dependencies within the hidden process rather than on the emission structure.

Section 5.3 addresses inference for the parameters of interest in MHMMs. We begin

with the likelihood factorization (Section 5.3.1), noting the intractability of the marginal likelihood of the hidden chain. To address this, the Forward–Backward (FB) Algorithm (Algorithm 3) is introduced as an efficient recursive method for the univariate case, though its computational cost grows exponentially in the multivariate setting (Section 5.3.2.4).

Section 5.3.3 presents the *Individual Forward Filtering Backward Sampling* (iFFBS) algorithm of Touloupou [2016], Touloupou et al. [2020], which reduces the cost of the FB algorithm to quadratic order (Section 5.3.3.1). However, iFFBS is limited by its reliance on the conditional independence within the CHMM framework and cannot be applied in the case where model allows for concurrent interactions. This motivates our novel *Generalised Individual Forward Backward* (GIFB) Algorithm (Section 5.3.4), which operates on the general MHMMs involving full dependencies, and allows flexible group-wise updating of hidden states, while maintaining quadratic computational cost (Section 5.3.4.1). Section 5.3.5 summarizes the comparison between FB, iFFBS, and GIFB.

The remaining parameters, including transition probabilities and emission parameters, are estimated via a Bayesian Gibbs sampler (Section 5.3.6). A Pólya–Gamma augmentation scheme is employed to efficiently sample the regression coefficients $\boldsymbol{\beta}$ from the Hierarchical Marginal Model, demonstrating the natural integration of GIFB with this framework. Finally, Section 5.3.7 reviews the Viterbi Algorithm as a method for recovering the optimal hidden path $\boldsymbol{z}^\star$, complementing the full posterior samples from the Gibbs sampler.

In addition to the theoretical development, we conduct simulation studies to assess the performance of GIFB in Section 5.4, with the simulation setup and procedure introduced in Section 5.4.1. We compare the direct extension of the FB algorithm to the multivariate case with the proposed GIFB algorithm, focusing on three criteria: (I) ability to detect the dependency structures among hidden chains (see Section 5.4.3.1), (II) accuracy in recovering the hidden states $\boldsymbol{Z}$ (see Section 5.4.3.2), and (III) computational efficiency (see Section 5.4.3.3).

In Section 5.5, we apply the GIFB algorithm to real Twitter data in the context of UK railways, with the aim of detecting both potential dependency structures among railway company groups and disruption events. We begin with an exploratory background analysis in Section 5.5.1, followed by the formulation of the MHMM in this setting in Section 5.5.2. Using the GIFB algorithm, we then present results on (I) the inferred dependency structures (see Section 5.5.3.1) and (II) the recovered hidden paths, validated against reported news events (see Section 5.5.3.2).

Finally, this chapter concludes with a comprehensive discussion in Section 5.6, which highlights the use of the novel GIFB within the MHMM framework to detect the potential dependency structures and recover the hidden path, and outlines the limitations and directions for future work in Section 5.6.1.

5.2 Multivariate Hidden Markov Model

In this section, we begin with the classical (univariate) Hidden Markov Model (HMM) before extending it into the multivariate setting, which is the Multivariate Hidden Markov Model (MHMM).

5.2.1 Hidden Markov Model (HMM)

According to Baum and Petrie [1966], a Hidden Markov Model consists of two stochastic processes: a latent process Z, which follows the Markov property but is not directly observable, and an observed process X, whose distribution depends on Z through an emission function $f(\theta, Z)$ for some parameter θ. To better illustrate it, we provide the following Trellis Diagram as Figure 5.1 with the arrows indicating the conditional dependencies.

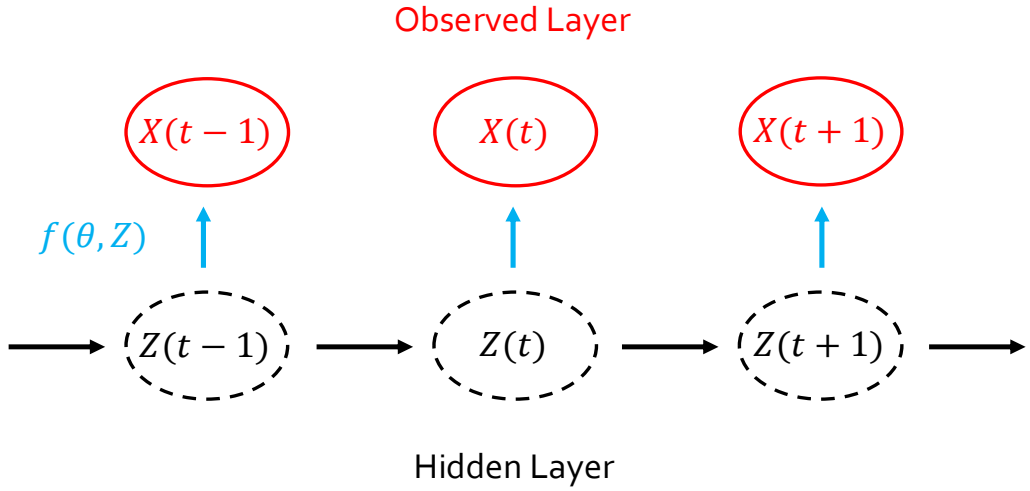


Figure 5.1: Trellis Diagram of an HMM. The hidden states $Z(t)$ are labelled in black dotted circles and follow a first-order Markov property characterized by the black arrows. The observations $X(t)$ are shown as red solid circles, each depending solely on the corresponding hidden state $Z(t)$ through an emission function $f\big(\theta, Z(t)\big)$, represented by the blue vertical arrows.

- $Z = \{Z(1), Z(2), \ldots, Z(T)\}$ denotes the univariate latent Markov chain over discrete time points $t = 1, \ldots, T$. It follows the same Markov setting introduced in Section 2.2, but remains unobserved throughout this chapter.

- $f\big(\theta, Z(t)\big)$ denotes the emission function, which specifies the conditional distribution of the observation $X(t)$ given the current hidden state $Z(t)$, with θ representing the parameters of this distribution, specifically:

$$X(t) \sim f\big(\theta, Z(t)\big)$$

- $X = \{X(1), X(2), \ldots, X(T)\}$ denotes the observed data across the same time points as the latent Markov chain $Z(t)$. The process $X(t)$ itself need not to be Markov.

To simplify the model, we make the following two model assumptions:

Assumption 5.1. *We consider a standard HMM, as illustrated in 5.1, that each $X(t)$ depends only on the corresponding hidden state $Z(t)$ through the emission function $f\big(\theta, Z(t)\big)$.*

Several extensions of the classical HMM have been proposed in the literature. For example, multiple independent hidden chains $\boldsymbol{Z}(t)$ may jointly determine a single observation $X(t)$ [Ghahramani and Jordan, 1995]; each hidden state $Z(t)$ may emit a vector of observations $\boldsymbol{X}(t)$ [Bengio and Frasconi, 1996]; or each hidden state $Z(t)$ may itself correspond to a sub-HMM that generates an entire sequence of observations [Fine et al., 1998]. In this chapter, however, we adopt the simple one-to-one correspondence for the emission, where each hidden state $Z(t)$ generates a single observation $X(t)$. This choice is motivated by our primary focus on modelling the dependency structures within the hidden chain $Z(t)$, rather than on the complexity of the emission process, which could otherwise obscure inference on the hidden dependencies.

Assumption 5.2. *We assume that the emission function is known up to its distributional form, but that the parameters θ are unknown.*

This setup allows us to carry out parametric inference, which can yield efficient estimation from both Frequentist and Bayesian and reduce data requirements when the model is correctly specified. To further simplify the framework, we adopt a Poisson distribution

for the emission function f, which acts as a natural choice for modelling count data:

$$X(t) = Y(t) + \sum_{s=2}^{S} \mathbb{I}\big(Z(t) = s\big) W_s(t) \tag{5.1}$$

where:

- $Y(t) \sim \text{Po}\big(\lambda(t)\big)$ denotes the baseline Poisson random variable with intensity function $\lambda(t)$. We consider two cases:

  1. $\lambda(t) \equiv \lambda$ is **time-invariant**, corresponding to the classical setting of Fischer and Meier-Hellstern [1993], which will be explored in the simulation study in Section 5.4.

  2. $\lambda(t)$ is **time dependent**, which will be applied to the Twitter data on UK railways in Section 5.5.

- $W_s(t) \sim \text{Po}(c_s)$ are state-specific Poisson random variables with the corresponding constant c_s, each associated with a non-baseline hidden state $Z(t) = s$ for $s \geq 2$, thereby introducing the effect of $Z(t)$ into the emission function. For identifiability we assume $0 < c_2 < \cdots < c_S$.

Alternatively, one can consider a non-parametric models for the emission function f, see for instance Kypraios and O'Neill [2018], which are robust to model misspecification and offer greater flexibilities. In this case, a larger sample of data is required to train the model through for instance, Gaussian Process.

Under this setting, the model parameters of interest are:

- The $s \times s$ transition probability matrix P for the hidden Markov process $Z(t)$, with its initial distribution π_0. This is the same as previous MMC setting discussed in Section 2.2.

- The parameters θ within the emission function f, namely $\lambda(t)$ and $c_2, \ldots, c_S$ as given in Equation 5.1

Alongside this, the hidden Markov process Z itself, as under the HMM, is not directly observed and act as the latent variables to be inferred.

The proposed HMM with a Poisson emission function can be regarded as a discretized analogue of the Markov-modulated Poisson process (MMPP) [Fischer and Meier-Hellstern, 1993], a doubly stochastic Poisson process in which the event intensity is driven by a continuous-time Markov chain, when we assume:

- The hidden state, $Z(t)$, is constant within the time intervals $(t-1, t]$ and is only allowed to change value at $t = 1, 2, \ldots, T$.
- The intensity function, $\lambda(t)$, in each time interval $(t-1, t]$ is assumed to be fixed and can be approximate through its integral:

$$\lambda(t) = \int_{t-1}^{t} \lambda(u)du$$

The discretization enables a probabilistic analysis of the hidden Markov chain $Z(t)$ using the Forward Backward Algorithm (see Section 5.3.2). Building on this, the remaining parameters can be inferred within a Bayesian framework using Markov Chain Monte Carlo (MCMC) methods (Section 5.3.6).

5.2.2 Multivariate Hidden Markov Model (MHMM)

While the HMM provides a useful framework for modelling the Markov switching dynamics, in many applications a single hidden chain is insufficient to capture complex dependencies. Real-world phenomena often involve interactions between multiple latent processes, motivating extensions to the Multivariate Hidden Markov Model (MHMM). For instance, Brand et al. [1997] proposed the MHMM to model interacting processes such as human motion or speech gestures. Such multivariate constructions allow richer dependency structures among the latent Markov Chains $\boldsymbol{Z}(t)$, which is the primary focus of this chapter.

By combining multiple Hidden Markov Chains, as illustrated in Figure 5.1, the HMM can be naturally extended to a Multivariate Hidden Markov Model (MHMM). The following Figure 5.2 illustrates this extension in the simple case of an MHMM consisting of two chains:

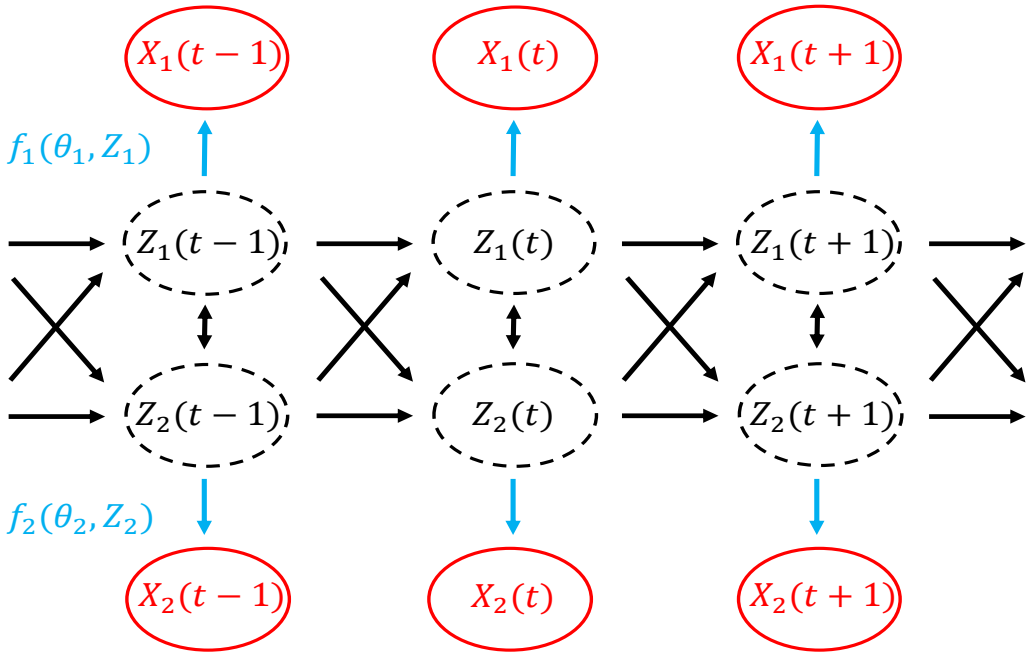


Figure 5.2: An illustration of a MHMM with two hidden chains. The hidden states $Z_1(t)$ and $Z_2(t)$ are shown as black dotted circles and evolve according to the first-order Markov property indicated by black arrows. The observations $X_1(t)$ and $X_2(t)$ are shown as red solid circles, where each observation depends only on its corresponding hidden state through the emission function $f_k(\theta_k, Z_k)$ for each chain $k = 1, 2$, represented by the blue vertical arrows.

- $\boldsymbol{Z}(t) = (Z_1(t), Z_2(t))$ now denotes the bivariate latent Markov chains at time t, represented by the black dotted circles.
- $f_1(\theta_1, Z_1)$ and $f_2(\theta_2, Z_2)$ denote the emission functions associated with Z_1 and Z_2, respectively, allowing each chain to have a different emission distribution.
- $\boldsymbol{X}(t) = (X_1(t), X_2(t))$ denotes the bivariate observations, represented by the red solid circles.

Similar to Assumption 5.1, we make the following assumption of the emitting structure:

Assumption 5.3. *Each observation $X_k(t)$ of chain k depends **only** on its corresponding hidden state $Z_k(t)$ through an emission function $f_k(\theta_k, Z_k)$ with distributional parameter θ_k, as illustrated by the blue vertical arrows.*

In the literature, DeRuiter et al. [2017] employ an MHMM to study the behaviour of blue whales, focusing on dependencies in the emission process while assuming that the hidden Markov chain evolve independently. In contrast, our formulation simplifies the emission structure through the one-to-one correspondence in Assumption 5.3 and concentrates on inferring the dependency structure within the hidden layer rather than the relationship between $\boldsymbol{Z}$ and $\boldsymbol{X}$.

While the one-to-one emission structure simplifies the observation layer, our main novelty lies in the hidden dynamics. In contrast to classical coupled Hidden Markov Models such as those in Brand et al. [1997] and Oliver et al. [2004], which impose conditional independence by omitting the black vertical arrows that represent concurrent interactions between chains, our formulation explicitly retains these dependencies. This enables us to capture the full dependency structures within the hidden multivariate Markov chain.

Under this setting, the parameters of interest are:

- The $s^K \times s^K$ joint transition probability matrix P under Cartesian Product Model (see Section 2.8), which governs the hidden multivariate Markov chain, with the initial distribution π_0 for $\boldsymbol{Z}(1)$.
- The vector of distributional parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ associated with the emission functions $f_1, \ldots, f_K$ for each chain.
- The unobserved states $\boldsymbol{Z}$ of the MMC, are treated as latent variables to be inferred.
- The corresponding joint transition matrix P of $\boldsymbol{Z}$.

In the following section, we will introduce an efficient inference framework for both the parameters of interest and the latent dependency structures, incorporating with the Hierarchical Marginal Model.

5.3 Inference

In this section, we focus on the parameter estimation for the MHMM. We begin with Section 5.3.1 by formulating the likelihood through a factorization of the joint distribution, following the dependency structure illustrated in Figure 5.2. To infer the unobserved states $\boldsymbol{Z}$, we first describe the standard Forward Backward (FB) algorithm [Baum et al., 1970, 1972], highlighting its prohibitive computational cost in the multivariate setting (see Section 5.3.2). To address this, Touloupou et al. [2020] proposed the *individual Forward Filtering Backward Sampling* (iFFBS) algorithm, which updates each hidden chain separately in a Gibbs-like fashion. However, this method is built on the Coupled Hidden Markov Model (CHMM), where conditional independence across chains is inherently assumed (see Section 5.3.3). Without this assumption, the iFFBS recursion breaks down, preventing it from capturing interactions for general MHMMs. To retain the full dependency structure while ensuring computational feasibility, we introduce the *Generalised Individual Forward Backward* (GIFB) algorithm, which operates directly on the MHMM and updates chains in a flexible group-wise fashion (see Section 5.3.4). For the remaining parameters, we present a Bayesian approach based on a Gibbs sampler combined with the proposed GIFB algorithm and the Hierarchical Marginal Model (see Section 5.3.6). Finally, the Viterbi algorithm (Section 5.3.7) is reviewed as a method to recover the optimal hidden path $\boldsymbol{z}^\star$.

5.3.1 Likelihood of the MHMM

According to the dependency structure illustrated in Figure 5.2, the joint likelihood $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}, P)$ can be factorized as

$$\begin{aligned}\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}, P) =& \pi(\boldsymbol{Z} \mid P)\, \pi\big(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{Z}\big) \\ =& \underbrace{\pi_0\big(\boldsymbol{Z}(1)\big) \prod_{t=1}^{T} \mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1), P\big)}_{\text{The latent Markov chains } \boldsymbol{Z}} \underbrace{\prod_{k=1}^{K} \prod_{t=1}^{T} \mathbb{P}\big(X_k(t) \mid \theta_k, Z_k(t)\big)}_{\text{Emissions from } Z_k(t) \text{ to } X_k(t)} \end{aligned} \tag{5.2}$$

- The first term corresponds to the likelihood of the latent Markov chains $\boldsymbol{Z}$, which

is identical to the MMC setting introduced earlier in Equation (A.2).

- The second term corresponds to the emission process from $\boldsymbol{Z}$ to $\boldsymbol{X}$. Assumption 5.3 of one-to-one emission enables the factorization of this joint distribution into the element-wise distributions.

Typically, off-the-shelf data-augmentation algorithms for hidden Markov models proceed by sampling each of the latent states iteratively. For example, to update the hidden path $\boldsymbol{Z}$ involves iteratively drawing $\pi(\boldsymbol{Z}(t) \mid \boldsymbol{X}, \boldsymbol{\theta}, P)$ at time $t = 1, \ldots, T$. The corresponding naive approach to this is through marginalizing over all the rest hidden states $\boldsymbol{Z}$:

$$\pi\big(\boldsymbol{Z}(t) \mid \boldsymbol{X}, \boldsymbol{\theta}, P\big) \propto \pi\big(\boldsymbol{Z}(t), \boldsymbol{X} \mid \boldsymbol{\theta}, P\big) = \sum_{\boldsymbol{Z}(1)} \cdots \sum_{\boldsymbol{Z}(t-1)} \sum_{\boldsymbol{Z}(t+1)} \cdots \sum_{\boldsymbol{Z}(T)} \pi\big(\boldsymbol{Z}, \boldsymbol{X} \mid \boldsymbol{\theta}, P\big) \tag{5.3}$$

However, this requires summing over all $(s^K)^{(T-1)}$ possible hidden states $\boldsymbol{Z}$, where s^K is the size of the joint state space under the Cartesian Product (see Section 2.8) and T is the length of the sequence. Such direct computation quickly becomes infeasible, as the complexity grows exponentially. This intractability motivates the use of recursive algorithms, most notably the Forward Backward (FB) algorithm (the expectation step for the Baum–Welch algorithm [Baum et al., 1970, 1972]), which exploits the Markov property of the hidden chain to compute these probabilities efficiently in polynomial time.

5.3.2 Forward Backward Algorithm

5.3.2.1 Univariate Setting

First formulated by Baum et al. [1970, 1972], Forward Backward (FB) algorithm is an elegant technique to infer the conditional probability $\pi(Z(t) \mid X, \theta, P)$ in HMM under the *univariate* setting. The key concept of the Forward Backward algorithm is that it recursively uses the earlier computations.

Consider the marginal probability $\pi\big(Z(t) \mid X, \theta, P\big)$ of $Z(t)$, it can be split as the product

of the forward probability α times backward probability β:

$$\begin{aligned}\pi\big(Z(t) \mid X, \theta, P\big) &\propto \pi\big(Z(t), X(1{:}t) \mid \theta, P\big)\, \pi\big(X\big((t{+}1){:}T\big) \mid Z(t), \theta, P\big) \\ &\propto \alpha\big(Z(t)\big)\, \beta\big(Z(t)\big)\end{aligned} \tag{5.4}$$

- $\alpha\big(Z(t)\big) = \pi\big(Z(t), X(1{:}t) \mid \theta, P\big)$ is the *forward probability*: the probability of ending up in $Z(t)$ given the first t observations $X\big((t{+}1){:}T\big)$.

- $\beta\big(Z(t)\big) = \pi\big(X\big((t{+}1){:}T\big) \mid Z(t), \theta, P\big)$ is the *backward probability*: the probability of observing the rest data $X\big((t{+}1){:}T\big)$ given the start state $Z(t)$.

The forward probability can be computed recursively for $t = 2, \ldots, T$ as

$$\alpha\big(Z(t)\big) = \mathbb{P}\big(X(t) \mid Z(t), \theta\big) \sum_{z'=1}^{s} \mathbb{P}\big(Z(t) \mid Z(t{-}1) = z', P\big)\, \alpha(z') \tag{5.5}$$

with the initial condition

$$\alpha\big(Z(1)\big) = \pi_0(Z(1))\, \mathbb{P}(X(1) \mid Z(1), \theta)$$

Similarly, the backward probability is computed recursively for $t = T-1, \ldots, 1$ as

$$\beta\big(Z(t)\big) = \sum_{z'=1}^{s} \beta(z')\, \mathbb{P}\big(X(t{+}1) \mid Z(t{+}1) = z', \theta\big)\, \mathbb{P}\big(Z(t{+}1) = z' \mid Z(t), P\big) \tag{5.6}$$

with the initial condition

$$\beta\big(Z(T)\big) = 1 \qquad \forall Z(T) \in \mathcal{S} \tag{5.7}$$

The full Forward–Backward procedure, including both forward and backward recursions, is presented in Algorithm 3.

One could then utilize these forward probabilities $\alpha(\cdot)$ and backward probabilities $\beta(\cdot)$ from the FB algorithm to compute the marginal distribution $\pi\big(Z(t) \mid X, \theta, P\big)$ through Equation (5.4).

Algorithm 3 Forward Backward Algorithm

Input: observations X, parameters θ, initial distribution $\pi_0(\cdot)$, transitions probabilities P, emission function $\mathbb{P}(X(t) \mid Z(t), \theta)$

Output: forward probability $\alpha\big(Z(t)\big)$ and backward probability $\beta\big(Z(t)\big)$ for each t

Forward Process: compute $\alpha\big(Z(t)\big) = \pi\big(Z(t), X(1{:}t) \mid \theta, P\big)$
Initialize $\alpha\big(Z(1)\big) \propto \pi_0\big(Z(1)\big)\, \mathbb{P}\big(X(1) \mid Z(1), \theta\big)$
for $t = 2, \ldots, T$ **do**
  for each state $Z(t) \in \mathcal{S}$ **do**
    Update $\alpha\big(Z(t)\big)$ by Equation (5.5)
  end for
end for
Backward Process: compute $\beta\big(Z(t)\big) = \pi\big(X\big((t{+}1){:}T\big) \mid Z(t), \theta, P\big)$
Initialize $\beta\big(Z(T)\big) = 1$
for $t = T-1, \ldots, 1$ **do**
  for each state $Z(t) \in \mathcal{S}$ **do**
    Update $\beta\big(Z(t)\big)$ by Equation (5.6)
  end for
end for

5.3.2.2 Computational Cost of the Forward Backward Algorithm

The computational cost of the FB algorithm can be derived as follows:

- **Forward recursion:** For each $t = 2, \ldots, T$ and each candidate state $Z(t) \in \mathcal{S}$, the update in Equation (5.5) requires a summation over all s possible states. This costs $O(s)$ per state. Since there are s such states at time t, the total cost is $O(s^2)$ for each time step, and hence $O(Ts^2)$ overall for the forward recursion.

- **Backward recursion:** For each $t = T-1, \ldots, 1$ and each candidate state $Z(t) \in \mathcal{S}$, the update in Equation (5.6) involves a summation over s states, giving $O(s)$ operations per state. Repeating this for all s states yields $O(s^2)$ for each time step, and $O(Ts^2)$ in total for the backward recursion.

Combining the forward and backward recursions, the overall computational cost of the FB algorithm is $O(Ts^2)$. In contrast, the naive approach to the marginal likelihood in Equation (5.3) requires summing over all s^T possible hidden state sequences, leading to

a cost of $O(s^T)$. Generally,

$$O(Ts^2) \ll O(s^T)$$

meaning that the Forward Backward algorithm reduces the exponentially growing computational cost in the naive approach, to a quadratic dependence on the state space size s and linear dependence on the sequence length T.

5.3.2.3 Simulate the Hidden Markov Chain by FB Algorithm

Furthermore, the forward and backward probabilities from the recursion not only yield the marginals, but also allow us to sample the entire hidden sequence $Z = \big(Z(1), \ldots, Z(T)\big)$. By the Markov property, the joint distribution $\pi(Z \mid X, \theta, P)$ can be factorized as:

$$\pi\big(Z \mid X, \theta\big) = \pi\big(Z(1) \mid X, \theta\big) \prod_{t=2}^{T} \pi\big(Z(t) \mid Z(t-1), X, \theta, P\big) \tag{5.8}$$

For the first term $\pi(Z(1) \mid X, \theta)$, we can directly write it as a product of the forward probability and the backward probability:

$$\begin{aligned} \pi\big(Z(1) \mid X, \theta\big) &\propto \pi\big(Z(1), X(1) \mid \theta\big)\, \pi(X(2{:}T) \mid Z(1), \theta) \\ &\propto \alpha(Z(1))\, \beta(Z(1)) \end{aligned} \tag{5.9}$$

The rest terms $\pi\big(Z(t) \mid Z(t-1), X, \theta, P\big)$ throughout $t = 2, \ldots, T$ can be split as:

$$\begin{aligned} &\pi\big(Z(t) \mid Z(t-1), X, \theta, P\big) \\ &\propto \pi(Z(t), Z(t-1) \mid X, \theta, P) \\ &\propto \pi\big(Z(t-1), X(1:t-1) \mid \theta, P\big)\, \mathbb{P}\big(Z(t) \mid Z(t-1), P\big)\, \mathbb{P}\big(X(t) \mid Z(t), \theta\big)\, \pi(X(t{+}1{:}T) \mid Z(t), \theta, P) \\ &\propto \alpha\big(Z(t-1)\big)\, \mathbb{P}\big(Z(t) \mid Z(t-1), P\big)\, \mathbb{P}\big(X(t) \mid Z(t), \theta\big)\, \beta\big(Z(t)\big) \end{aligned} \tag{5.10}$$

Based on Equations (5.9) and (5.10), the hidden Markov chain Z can be sampled given the observations X, the transition matrix P, and the emission function f with parameters θ. This is achieved through the factorization in Equation (5.8), and the procedure is summarized in Algorithm 4.

Algorithm 4 Forward Backward Sampling of of the hidden Markov chain

Input: observations X, parameters θ, initial distribution $\pi_0(\cdot)$, transitions probabilities P, emission function $\mathbb{P}(X(t) \mid Z(t), \theta)$, forward probabilities α and backward probabilities β

Output: sampled hidden Markov Chain Z

Sample $Z(1) \sim \pi\big(Z(1) \mid X, \theta\big)$ according to Equation (5.9)
for $t = 2, \ldots, T$ **do**
  Sample $Z(t) \sim \pi\big(Z(t) \mid Z(t-1), X, \theta, P\big)$ according to Equation (5.10)
end for

5.3.2.4 Extension to the Multivariate

In the context of multivariate cases, a frequently employed approach, originally proposed by Carter and Kohn [1994], involves transforming an MHMM into an equivalent univariate HMM and subsequently applying the standard FB algorithm. This procedure is equivalent as applying the standard Forward Backward algorithm directly on the Cartesian product state space described in Section 2.8. While this approach is straightforward, it becomes computationally infeasible when the number of hidden chains K is large. More precisely, applying the standard Forward Backward recursion to the extended Cartesian state space suffers from a computational cost of

$$O\big(T(s^K)^2\big) = O\big(Ts^{2K}\big)$$

which increases exponentially with the number of chains K. This renders both the forward and backward recursions prohibitively expensive even when K is moderate. To address this computational challenge, it is necessary to develop some alternatives to the standard FB algorithm that reduce the overall computational complexity. In the following sections, we present Gibbs sampler–motivated approaches that update the hidden chains sequentially rather than jointly, as in the standard FB algorithm, to alleviate the prohibitive computational cost. These include:

- iFFBS algorithm by Touloupou et al. [2020] (Section 5.3.3), but is built on the CHMM and relies on the conditional independence.

- Our novel GIFB algorithm (Section 5.3.4), which is suitable for MHMMs and fully captures the dependency structures.

5.3.3 Individual Forward Filtering Backward Sampling (iFFBS) Algorithm

To alleviate the prohibitive computational cost of the Forward Backward algorithm in the multivariate setting, Touloupou [2016] proposed the *individual Forward Filtering Backward Sampling* (iFFBS) algorithm. The key idea of iFFBS is to employ a Gibbs sampling strategy, drawing each hidden chain sequentially conditioning on the others, thereby approximating the joint distribution of all hidden variables. In contrast to directly extending the standard Forward Backward algorithm to sample all chains jointly from $\pi(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}, P)$, iFFBS updates iteratively each chain $k \in [K]$ from the full conditionals $\pi(\boldsymbol{Z}_k \mid \boldsymbol{X}, \boldsymbol{Z}_{-k}, \boldsymbol{\theta}, P)$. Additionally according to the structure of MHMM as illustrated in 5.2, $\boldsymbol{Z}_k$ is independent of $\boldsymbol{X}_{-k}$ when conditioned on $\boldsymbol{Z}_{-k}$, under which we can write the targeting full conditional as:

$$\pi\big(\boldsymbol{Z}_k \mid \boldsymbol{X}, \boldsymbol{Z}_{-k}, \boldsymbol{\theta}, P\big) = \pi\big(\boldsymbol{Z}_k \mid \boldsymbol{X}_k, \boldsymbol{Z}_{-k}, \theta_k, P\big) \tag{5.11}$$

Touloupou et al. [2020] suggested to factorize the density of this full conditional distribution from backwards:

$$\begin{aligned}\pi\big(\boldsymbol{Z}_k \mid \boldsymbol{X}_k, \boldsymbol{Z}_{-k}, \theta_k, P\big) =& \pi\big(Z_k(T) \mid \boldsymbol{X}_k, \boldsymbol{Z}_{-k}, \theta_k, P\big) \\ & \prod_{t=1}^{T-1} \pi\big(Z_k(t) \mid Z_k(t+1), \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big), \boldsymbol{X}_k(1{:}t), \theta_k, P\big)\end{aligned} \tag{5.12}$$

By Bayes' theorem, we can factorize each term in the product for $t = 1, \ldots, T-1$ as:

$$\begin{aligned}&\pi\big(Z_k(t) \mid Z_k(t{+}1), \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big), \boldsymbol{X}_k(1{:}t), \theta_k, P\big) \\ \propto\ &\underbrace{\pi\big(Z_k(t{+}1) \mid Z_k(t), \boldsymbol{Z}_{-k}(t), \theta_k, P\big)}_{\text{Marginal Transition Probabilities}}\ \underbrace{\pi\big(Z_k(t) \mid \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big), \boldsymbol{X}_k(1{:}t), P\big)}_{\text{Modified Conditional Filtered Probability}}\end{aligned}$$

where the first factor is the marginal transition probability and the second factor is the modified conditional filtered probability, which can be initialized at $t = 1$ as:

$$\pi\big(Z_k(1) \mid \boldsymbol{Z}_{-k}(1{:}2), X(1), \theta_k, P\big) \propto \pi_0\big(Z_k(1)\big)\, \mathbb{P}\big(X_k(1) \mid Z_k(1), \theta_k\big) \prod_{k' \neq k} \mathbb{P}\big(Z_{k'}(2) \mid Z_{k'}(1),\, \boldsymbol{Z}_{-k'}(1), P\big) \tag{5.13}$$

For each $t = 2, \ldots, T$, the remaining part of the modified conditional filtering probabilities can be obtained by iteratively performing the following two steps:

1. Compute one-step ahead modified conditional predictive probabilities:

$$\pi\big(Z_k(t) \mid \boldsymbol{Z}_{-k}(1{:}t),\, \boldsymbol{X}\big(1{:}(t{-}1)\big),\, \theta_k, P\big) = \sum_{z \in \mathcal{S}} \Big[\pi\Big(Z_k(t) \mid Z_k(t{-}1) = z,\, \boldsymbol{Z}_{-k}\big(1{:}(t{-}1), P\big)\Big) \times \underbrace{\pi\big(Z_k(t{-}1) = z \mid \boldsymbol{Z}_{-k}(1{:}t),\, \boldsymbol{X}\big(1{:}(t{-}1)\big),\, \theta_k, P\big)}_{\text{from previous recursion}}\Big] \tag{5.14}$$

2. Compute modified conditional filter probabilities:

$$\begin{aligned}\pi\big(Z_k(t) = z \mid \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big),\, \boldsymbol{X}(1{:}t),\, \theta_k, P\big) \;&\propto \pi\big(Z_k(t) = z \mid \boldsymbol{Z}_{-k}(1{:}t),\, \boldsymbol{X}\big(1{:}(t{-}1), P\big),\, \theta_k, P\big) \\ &\times \mathbb{P}\big(X_k(t) \mid Z_k(t) = z,\, \theta_k\big) \\ &\times \prod_{k' \neq k} \mathbb{P}\big(Z_{k'}(t{+}1) \mid Z_k(t) = z, \boldsymbol{Z}_{-k}(t), P\big)\end{aligned} \tag{5.15}$$

   Note that the factorization into the marginals $\mathbb{P}\big(Z_{k'}(t{+}1) \mid Z_k(t) = z, \boldsymbol{Z}_{-k}(t), P\big)$ in the last line requires conditional independence.

Once the modified conditional filtering probabilities are obtained, the hidden states $\boldsymbol{Z}_k$ can be sampled from the backwards for each $t = T - 1, \ldots, 1$ according to:

$$\begin{aligned}&\pi\big(Z_k(t) = z \mid Z_k(t{+}1) = z',\, \boldsymbol{Z}_{-k}(1{:}T),\, \boldsymbol{X}_k(1{:}T),\, \theta_k, P\big) \\ &\propto \pi\big(Z_k(t) = z \mid \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big),\, \boldsymbol{X}_k(1{:}t),\, \theta_k, P\big)\, \pi\big(Z_k(t{+}1) = z' \mid Z_k(t) = z,\, Z_{-k}(t), P\big)\end{aligned} \tag{5.16}$$

This yields the iFFBS that samples the full conditional $\pi\big(\boldsymbol{Z}_k \mid \boldsymbol{X}, \boldsymbol{Z}_{-k}, \boldsymbol{\theta}, P\big)$ in the closed form, hence enabling a Gibbs sampler that iterates over $k = 1, \ldots, K$. The full algorithm can be summarized as the following Algorithm 5:

Algorithm 5 Individual Forward Filtering Backward Sampling Algorithm (iFFBS)

Input: observations $\boldsymbol{X}$, parameters θ, initial distribution $\pi_0(\cdot)$, transition probabilities P, emission function $\mathbb{P}(X(t) \mid Z(t), \theta)$

Output: simulated hidden Markov Chain $\boldsymbol{Z}$

for Each chain $k = 1, \ldots, K$ **do**

Compute $\pi\big(Z_k(1) \mid \boldsymbol{Z}_{-k}(1{:}2), X(1), \theta_k, P\big)$ by Equation (5.13)

for $t = 2, \ldots, T$ **do Forward Filtering:**

Compute $\pi\big(Z_k(t) = z \mid \boldsymbol{Z}_{-k}\big(1{:}(t{+}1)\big), \boldsymbol{X}(1{:}t), \theta_k, P\big)$ by Equation (5.14) and (5.15)

end for

Sample $Z_k(T) \sim \pi\big(Z_k(T) \mid \boldsymbol{Z}_{-k}(1{:}T), \boldsymbol{X}_k(1{:}T), \theta_k, P\big)$ from the last term of the modified filter probability.

for $t = T-1, \ldots, 1$ **do Backward Sampling:**

Sample $Z_k(t) \sim \pi\big(Z_k(t) \mid Z_k(t{+}1), \boldsymbol{Z}_{-k}(1{:}T), \boldsymbol{X}_k(1{:}T), \theta_k, P\big)$ by Equation (5.16)

end for

end for

Up to this point, iFFBS remains an exact sampler under the conditional independence assumption. However, motivated by epidemic models in which within-chain dependence is typically stronger than between-chain dependence, Touloupou et al. [2020] introduced a modified version of iFFBS that incorporates a Metropolis–Hastings correction step to further reduce computational burden. In this scheme, proposals for the hidden chains are generated using the approximation:

$$\mathbb{P}(Z_{k'}(t+1) \mid \boldsymbol{Z}(t), P) \approx \mathbb{P}(Z_{k'}(t+1) \mid \boldsymbol{Z}_{-k}(t), P) \tag{5.17}$$

This approximation effectively neglects part of the between-chain dependence, thereby simplifying the computation by eliminating product terms in Equation (5.15). However, the assumption may be problematic in scenarios where several chains exhibit strong cor-

relations. In this case, the resulting proposal distribution can lead to a high rejection rate in the Metropolis step, which in turn may cause slow mixing and poor convergence of the MCMC algorithm.

Nevertheless, the iFFBS algorithm is built on the Coupled Hidden Markov Model (CHMM), where conditional independence is inherently assumed. This assumption is crucial because it allows the joint transition probabilities in Equation (5.15) to be factorized into marginals. Or otherwise, the factorization does not hold and the recursion becomes invalid. If the objective is to assess potential conditional or contemporaneous independence, one must instead consider a setting where the hidden Markov chains are fully connected rather than restricted by the conditional independence structure of the CHMM. This motivates the introduction of a novel *Generalised Individual Forward–Backward* (GIFB) algorithm, which retains the Gibbs-like iterative sampling over groups of chains while explicitly accounting for concurrent interactions among them.

5.3.3.1 Computational Cost

The computational cost of iFFBS arises primarily from Equations (5.14) and (5.15), and scales as $O(TK^2s^2)$, since:

1. For each chain k and each time t, the filtering step requires $O(s^2)$ operations, leading to a total cost of $O(TKs^2)$.

2. The additional product over the remaining $k-1$ chains increases the overall cost to $O(TK^2s^2)$.

In comparison, recall the standard Forward Backward algorithm applied to the full Cartesian product state space has cost $O\left(T(s^K)^2\right)$ (see Section 5.3.2.4). Since, in general,

$$O(TK^2s^2) \ll O\left(Ts^{2K}\right)$$

meaning that the iFFBS algorithm reduces the exponentially growing computational cost of the standard Forward Backward algorithm in the multivariate setting, to a quadratic

dependence on both the state space size s and the number of chains k. This makes iFFBS computationally feasible for moderate values of k and s.

5.3.4 Generalised Individual Forward Backward (GIFB) Algorithm

Unlike the iFFBS algorithm of Touloupou et al. [2020], which is tailored to Coupled Hidden Markov Models and relies on conditional independence assumptions between chains, our GIFB algorithm is designed for Multivariate Markov Chains with full dependency structures. Additionally, GIFB directly uses the marginal transition probabilities η, as modelled by the Hierarchical Marginal Model (see Section 4.2), rather than the full joint transition probabilities p in the Cartesian Product model (see Section 2.8). By embedding η into the forward and backward recursions, the GIFB algorithm provides an exact and scalable procedure for inferring hidden states in MHMMs while fully preserves the dependencies across chains.

Inspired by Touloupou et al. [2020], we retain the Gibbs-like iterative sampling over chains. However, rather than sampling each individual chain separately as in iFFBS (see Equation (5.11)), we extend the procedure to group-wise updates, where each groups of chains are sampled jointly. Specifically, for each chain group $g_1, g_2, \ldots, g_J$, the target full conditional distribution becomes

$$\pi\big(\boldsymbol{Z}_j \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \qquad j \in [J]$$

Note that when each group g_j is a singleton, the target full conditional remain the same as that of iFFBS in Equation (5.11). Thus, the proposed group-wise formulation can be seen as a natural generalisation of iFFBS, allowing flexible block updates.

In contrast to iFFBS, where the full conditional is factorized in a backward direction to accommodate semi-Markov settings, we instead adopt a forward factorization consistent

with the standard Forward Backward algorithm under the first-order Markov setting:

$$\pi\big(\boldsymbol{Z}_j \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) = \pi\big(\boldsymbol{Z}_j(1) \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \prod_{t=2}^{T} \pi\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \tag{5.18}$$

For each $t = 2, \ldots, T$, the term inside the product in Equation (5.18) can be derived as:

$$\begin{aligned}
&\pi\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \\
\propto &\pi\big(\boldsymbol{Z}_j(t), \boldsymbol{Z}_j(t-1), \boldsymbol{X}_j \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \\
\propto &\pi\big(\boldsymbol{Z}_j(t-1), \boldsymbol{X}_j(1{:}(t-1)) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big)\, \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}, P\big)\, \pi\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big) \\
&\pi\big(\boldsymbol{X}_j((t+1){:}T) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big) \\
\propto &\alpha\big(\boldsymbol{Z}_j(t-1)\big)\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}, P\big)\pi\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big)\beta\big(\boldsymbol{Z}_j(t)\big)
\end{aligned} \tag{5.19}$$

- $\alpha\big(\boldsymbol{Z}_j(t-1)\big) = \pi\big(\boldsymbol{Z}_j(t-1), \boldsymbol{X}_j(1{:}(t-1)) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big)$ is the $(t-1)^{\text{th}}$ term for the *modified forward probability* for the chain group g_j, which conditioned additionally on the rest of the hidden chain $\boldsymbol{Z}_{-j}$ comparing to the standard forward probability in Equation 5.4.

- $\beta\big(\boldsymbol{Z}_j(t)\big) = \pi\big(\boldsymbol{X}_j\big((t+1){:}T\big) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, P\big)$ is the t^{th} term for the *modified backward probability* for the chain group g_j, which conditioned additionally on the rest of the hidden chain $\boldsymbol{Z}_{-j}$ comparing to the standard backward probability in Equation 5.4.

Note that, unlike the standard Forward Backward algorithm discussed previously in Section 5.3.2, here the transition probability in Equation (5.19) takes the form $\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}, P\big)$ which is conditioned additionally on the states of all other chains $\boldsymbol{Z}_{-j}$ instead of $\mathbb{P}(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), P)$ in standard FB Algorithm, which requires further manipulations. Under a first-order Markov chain:

$$\mathbb{P}(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}, P) = \mathbb{P}\Big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}\big((t-1){:}(t+1)\big), P\Big) \tag{5.20}$$

That is, given the states of all other chains, $\boldsymbol{Z}_j(t)$ depends on its own previous state $Z_j(t-1)$, as well as the states of the remaining chains $\boldsymbol{Z}_{-j}\big((t-1){:}(t+1)\big)$ at times $t-1$, t, and $t+1$.

By conditional probability, we can factorize the modified transition probability in Equation (5.20) as:

$$\begin{aligned}
&\mathbb{P}\Big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}\big((t-1):(t+1)\big), P\Big) \\
\propto &\mathbb{P}\big(\boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}(t+1) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}(t), \boldsymbol{Z}_{-j}(t-1), P\big) \\
\propto &\mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1), P\big)\mathbb{P}\big(\boldsymbol{Z}_{-j}(t+1) \mid \boldsymbol{Z}(t), P\big)
\end{aligned} \tag{5.21}$$

Here, both $\mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1), P\big)$ and $\mathbb{P}\big(\boldsymbol{Z}_{-j}(t+1) \mid \boldsymbol{Z}(t), P\big)$ are expressed in terms of the marginal transition probabilities defined previously in Equation (2.4). In other words, this modified transition probability $\mathbb{P}\Big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t-1), \boldsymbol{Z}_{-j}\big((t-1):(t+1)\big), P\Big)$ can be written as a product of the marginal transition probabilities η. Importantly, all these marginal transition probabilities are explicitly modelled within the Hierarchical Marginal Model (see Section 4.2). This allows the GIFB algorithm to be naturally integrated with the Hierarchical Marginal Model, providing an efficient way to sample the hidden states $\boldsymbol{Z}_j$ for each chain group j. Hence, instead of using the joint transition P matrix, we express the transition dynamics through the regression coefficients $\boldsymbol{\beta}$ in the Hierarchical Marginal Model framework, that is,

$$\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1), P\big) \equiv \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}(t-1), \boldsymbol{\beta}\big)$$

We now consider the recursion of the modified forward probability $\alpha\big(\boldsymbol{Z}_j(t)\big)$ for chain

group g_j. For each $t = 2, \ldots, T$, it can be derived as:

$$\begin{aligned}
\alpha\big(\boldsymbol{Z}_j(t)\big) =& \pi\big(\boldsymbol{Z}_j(t), \boldsymbol{X}_j(1{:}t) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=& \sum_{\boldsymbol{Z}_j(t-1)} \pi\big(\boldsymbol{Z}_j(t), \boldsymbol{Z}_j(t{-}1), \boldsymbol{X}_j(1{:}t) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=& \sum_{\boldsymbol{Z}_j(t-1)} \pi\big(\boldsymbol{Z}_j(t), \boldsymbol{Z}_j(t-1), \boldsymbol{X}_j\big(1{:}(t{-}1)\big), \boldsymbol{X}_j(t) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=& \sum_{\boldsymbol{Z}_j(t-1)} \pi\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t{-}1), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)\, \pi\big(\boldsymbol{Z}_j(t{-}1), \boldsymbol{X}_j\big(1{:}(t{-}1)\big) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=& \sum_{\boldsymbol{Z}_j(t-1)} \pi\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t{-}1), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)\, \alpha\big(\boldsymbol{Z}_j(t{-}1)\big) \qquad (5.22)
\end{aligned}$$

Here, the summation is taken over all $\boldsymbol{Z}_j(t{-}1)$ in its Cartesian Product state space $\boldsymbol{\mathcal{S}}_j$. The term $\mathbb{P}\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big)$ denotes the emission probability associated with the hidden chain. Under Assumption 5.3, it can be factorised as the product of the emission probability for each individual chain k within chain group g_j:

$$\mathbb{P}\big(\boldsymbol{X}_j(t) \mid \boldsymbol{Z}_j(t), \boldsymbol{\theta}\big) = \prod_{k \in g_j} \mathbb{P}\big(X_k(t) \mid Z_k(t), \boldsymbol{\theta}\big) \qquad (5.23)$$

The key distinction between the forward recursion in GIFB (Equation (5.22)) and that in the standard FB algorithm (Equation (5.5)) lies in the modified transition probability $\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t{-}1), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)$, which is additionally conditioned on the remaining hidden chains $\boldsymbol{Z}_{-j}$. However, as shown in Equation (5.21), this term can be factorized into a product of the marginal transition probabilities, thereby enabling a natural incorporation with the Hierarchical Marginal Model.

For $t = 1$, the initial term of the modified forward probability $\alpha\big(\boldsymbol{Z}_j(1)\big)$ can be derived as:

$$\begin{aligned}
\alpha\big(\boldsymbol{Z}_j(1)\big) =& \mathbb{P}\big(\boldsymbol{X}_j(1) \mid \boldsymbol{Z}_j(1), \boldsymbol{\theta}\big) \pi\big(\boldsymbol{Z}_j(1) \mid \boldsymbol{Z}_{-j}(1:2), \boldsymbol{\beta}\big) \\
\propto& \mathbb{P}\big(\boldsymbol{X}_j(1) \mid \boldsymbol{Z}_j(1), \boldsymbol{\theta}\big)\, \pi_0\big(\boldsymbol{Z}(1)\big)\, \mathbb{P}\big(\boldsymbol{Z}_{-j}(2) \mid \boldsymbol{Z}(1), \boldsymbol{\beta}\big) \qquad (5.24)
\end{aligned}$$

where $\mathbb{P}\big(\boldsymbol{X}_j(1) \mid \boldsymbol{Z}_j(1), \boldsymbol{\theta}\big)$ denotes the emission probability, $\pi_0\big(\boldsymbol{Z}(1)\big)$ is the predefined initial distribution of the hidden state $\boldsymbol{Z}(1)$, and the final term corresponds to the marginal transition probability η_{-j}.

Analogously, the recursion for the modified backward probability $\beta\big(\boldsymbol{Z}_j(t)\big)$ of chain group g_j can be expressed, for each $t = 1, 2, \ldots, T-1$, as:

$$
\begin{aligned}
&\beta\big(\boldsymbol{Z}_j(t)\big) \\
=&\pi\big(\boldsymbol{X}_j\big((t{+}1){:}T\big) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=&\sum_{\boldsymbol{Z}_j(t+1)} \pi\big(\boldsymbol{X}_j\big((t{+}1){:}T\big), \boldsymbol{Z}_j(t{+}1) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=&\sum_{\boldsymbol{Z}_j(t+1)} \pi\big(\boldsymbol{X}_j(t{+}1) \mid \boldsymbol{Z}_j(t{+}1), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}_j(t{+}1) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)\, \pi\big(\boldsymbol{X}_j\big((t{+}2){:}T\big) \mid \boldsymbol{Z}_j(t{+}1), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\
=&\sum_{\boldsymbol{Z}_j(t+1)} \pi\big(\boldsymbol{X}_j(t{+}1) \mid \boldsymbol{Z}_j(t{+}1), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}_j(t{+}1) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)\, \beta\big(\boldsymbol{Z}_j(t{+}1)\big) \qquad (5.25)
\end{aligned}
$$

To avoid confusion, we use β to denote the backward probability, while $\boldsymbol{\beta}$ denotes the regression coefficients in the Hierarchical Marginal Model. The summation is taken over all $\boldsymbol{Z}_j(t{+}1)$ in its Cartesian Product state space $\boldsymbol{\mathcal{S}}_j$. Under Assumption 5.3, the emission probability $\mathbb{P}\big(\boldsymbol{X}_j(t{+}1) \mid \boldsymbol{Z}_j(t{+}1), \boldsymbol{\theta}\big)$ can be factorized as the product of the emission probabilities for each individual chain k within the group g_j, as in Equation (5.23). Again, the key distinction between the backward recursion in GIFB (Equation (5.25)) and that in the standard FB algorithm (Equation 5.6) lies in the modified transition probability $\mathbb{P}\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t{-}1), \boldsymbol{Z}_{-j}, \boldsymbol{\beta}\big)$, which can be decomposed into the product of the marginal transition probabilities, following Equation (5.21), and again enabling a natural incorporation with the Hierarchical Marginal Model.

For $t = T$, the initial condition for the backward recursion is set to be:

$$\beta\big(\boldsymbol{Z}_j(T)\big) = 1$$

which is the same as that in the standard FB algorithm in Equation (5.7).

Similar as Equation (5.4), the first term $\pi\big(\boldsymbol{Z}_j(1) \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$ in Equation (5.18) can be directly expressed as the product of the forward probability α and the backward probability β:

$$\begin{aligned}\pi\big(\boldsymbol{Z}_j(1) \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) &\propto \pi\big(\boldsymbol{Z}_j(1), \boldsymbol{X}_j(1) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big)\pi\big(\boldsymbol{X}_j(2{:}T) \mid \boldsymbol{Z}_j(1), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\ &\propto \alpha\big(\boldsymbol{Z}_j(1)\big)\beta\big(\boldsymbol{Z}_j(1)\big) \end{aligned} \tag{5.26}$$

The entire process of utilizing the GIFB algorithm to sample the hidden chain $\boldsymbol{Z}$ can be summarized in the following Algorithm 6:

5.3.4.1 Computational Cost

The computational cost of GIFB arises primarily from the forward and backward recursion in Equations (5.22) and (5.25), and can be derived following a same manner as the standard FB algorithm:

- **Forward recursion:** For each $t = 2, \ldots, T$ and each candidate state $\boldsymbol{Z}_j(t) \in \boldsymbol{\mathcal{S}}_j$,the update in Equation (5.22) requires a summation over all $|\boldsymbol{\mathcal{S}}_j| = s^{|g_j|}$ possible states within the Cartesian Product Space for chain group g_j. This costs $O(|\boldsymbol{\mathcal{S}}_j|)$ per state. Since there are $|\boldsymbol{\mathcal{S}}_j|$ such states at time t, the total cost is $O\big(|\boldsymbol{\mathcal{S}}_j|^2\big)$ for each time step, and hence $O\big(T|\boldsymbol{\mathcal{S}}_j|^2\big)$ overall for the forward recursion.

- **Backward recursion:** Similar as the forward recursion, for each $t = T-1, \ldots, 1$ and each candidate state $Z(t) \in \mathcal{S}$, the update in Equation (5.25) involves a summation over all $|\boldsymbol{\mathcal{S}}_j| = s^{|g_j|}$ possible states within the Cartesian Product Space for chain group g_j, giving $O(|\boldsymbol{\mathcal{S}}_j|)$ operations per state. Repeating this for all $|\boldsymbol{\mathcal{S}}_j|$ states yields $O\big(|\boldsymbol{\mathcal{S}}_j|^2\big)$ for each time step, and $O\big(T|\boldsymbol{\mathcal{S}}_j|^2\big)$ in total for the backward recursion.

Algorithm 6 Sampling by Generalised Individual Forward Backward Algorithm

Input: observations $\boldsymbol{X}$, parameters $\boldsymbol{\theta}$, initial distribution $\pi_0(\cdot)$, transition probabilities P, emission functions $\mathbb{P}(X_k(t) \mid Z_k(t), \theta_k)$ for each chain k

Output: sampled hidden Markov Chain $\boldsymbol{Z}$

for each chain group g_j with $j = 1, \ldots, J$ **do**
  Forward Process: compute $\alpha\big(\boldsymbol{Z}_j(t)\big) = \pi\big(\boldsymbol{Z}_j(t), \boldsymbol{X}_j(1{:}t) \mid \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$
  Initialize $\alpha\big(\boldsymbol{Z}_j(1)\big)$ according to Equation (5.24)
  for $t = 2, \ldots, T$ **do**
    for each state $\boldsymbol{Z}_j(t) \in \boldsymbol{\mathcal{S}}_j$ **do**
      Update $\alpha\big(\boldsymbol{Z}_j(t)\big)$ according to Equation (5.22)
    end for
  end for
  Backward Process: compute $\beta\big(\boldsymbol{Z}_j(t)\big) = \pi\big(\boldsymbol{X}_j\big((t{+}1){:}T\big) \mid \boldsymbol{Z}_j(t), \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$
  Initialize $\beta\big(\boldsymbol{Z}_j(T)\big) = 1$
  for $t = T-1, \ldots, 1$ **do**
    for each state $\boldsymbol{Z}_j(t{+}1) \in \boldsymbol{\mathcal{S}}_j$ **do**
      Update $\beta\big(\boldsymbol{Z}_j(t)\big)$ according to Equation (5.25)
    end for
  end for
  Sampling Process: sample $\boldsymbol{Z}_j \sim \pi\big(\boldsymbol{Z}_j \mid \boldsymbol{Z}_{-j}, \boldsymbol{X}_j, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$
  Sample $\boldsymbol{Z}_j(1) \sim \pi\big(\boldsymbol{Z}_j(1) \mid \boldsymbol{Z}_{-j}, \boldsymbol{X}_j, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$ according to Equation (5.26)
  for $t = 2, \ldots, T$ **do**
    Sample $\boldsymbol{Z}_j(t) \sim \pi\big(\boldsymbol{Z}_j(t) \mid \boldsymbol{Z}_j(t{-}1), \boldsymbol{Z}_{-j}, \boldsymbol{X}_j, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$ according to Equation (5.19)
  end for
end for

We repeat the forward and backward recursions for each chain group $g_1, \ldots, g_J$, which yields a computational cost for the GIFB algorithm of

$$O\left(T \sum_{j=1}^{J} |\mathcal{S}_j|^2\right)$$

In the special case where each chain group g_j is a singleton, this cost reduces to that of the iFFBS algorithm, namely $O(Tk^2s^2)$. In practice, the updates can be performed either group-wise or element-wise depending on the data structure. Consequently, GIFB offers a flexible trade-off between computational cost and convergence speed.

Nevertheless, compared with the Cartesian product extension of the FB algorithm proposed by Carter and Kohn [1994] (see Section 5.3.2.4), the GIFB algorithm achieves a substantial reduction in computational complexity. In particular,

$$O\left(T\sum_{j=1}^{J}|\mathcal{S}_j|^2\right) = O\left(T\sum_{j=1}^{J}(s^{|g_j|})^2\right) \ll O\left(T\prod_{j=1}^{J}(s^{|g_j|})^2\right) = O\big(Ts^{2K}\big)$$

since the dependence on the number of chains enters additively across groups, rather than multiplicatively through the full Cartesian product state space.

5.3.5 Comparison between FB, iFFBS and GIFB Algorithms

In this section, we present a comparative overview of the FB, iFFBS, and GIFB algorithms, highlighting their modelling assumptions, setting, sampling schemes, likelihood expansion, and computational costs, as summarized in following Table 5.1:

The standard **Forward–Backward (FB) Algorithm** by Carter and Kohn [1994] operates directly on MHMMs without imposing independence assumptions, thereby enabling to capture all the dependency structures among the hidden chains. As given in Equation (5.18), it samples all hidden chains jointly and expands the likelihood in the forward direction, in accordance with the first-order Markov property. The algorithm is exact (in Monte Carlo sense), since it explicitly enumerates all possible dependencies via utilizing the joint transition probabilities p over the full Cartesian product of states. However, FB algorithm suffers from an exponentially growing computational cost of $O(Ts^{2K})$, which quickly making it infeasible under the high-dimensional cases.

The **Individual Forward Filtering Backward Smoothing (iFFBS) Algorithm** from Touloupou et al. [2020] is designed for CHMMs under the conditional independence assumption, which prevents it from modelling concurrent interactions between hidden chains. It reduces the computational cost by sampling each chain individually in a Gibbs-

Algorithm	Standard FB	iFFBS	GIFB
Setting	MHMM (No assumptions)	CHMM (Under conditional independence)	MHMM (No assumptions)
Hidden Process	Markov	Markov and Semi-Markov	Markov
Sampling Scheme	Jointly for all the chains	Gibbs-like for each chain	Gibbs-like for each group of chains
Likelihood Factorization	Forward	Backward	Forward
Exactness	Yes	Yes, and can use approximation (5.17) with an extra MH step	Yes
Transition Probabilities Used	Joint transition probabilities p	Parametric Marginal Transition Probabilities	Marginal transition probabilities η
Computational Cost	$O\big(Ts^{2K}\big)$	$O(TK^2s^2)$	$O\left(T\sum_{j=1}^{J}\lvert\mathcal{S}_j\rvert^2\right)$

Table 5.1: Comparison between standard FB, iFFBS and GIFB

like scheme and relies on backward factorization of the likelihood, as demonstrated in Equation (5.12), making it suitable for semi-Markov settings. iFFBS algorithm can be exact, but in practice computational gains typically rely on an approximation (5.17) motivated by the epidemic examples, which neglects some cross-chain dependence. This in return requires an additional Metropolis-Hastings correction step. Moreover, the transition probabilities are specified parametrically, again inspired by epidemic modelling. The resulting computational cost is reduced to $O(Tk^2s^2)$, significantly lower than the standard FB but relies on the conditional independence and hence suffers from the reduced modelling flexibility.

The proposed **Generalised Individual Forward Backward (GIFB) Algorithm** also works directly on the MHMMs without assuming conditional independence, enabling to capture the full dependency structures as the FB algorithm. It introduces a flexible sampling scheme where each group of chains are allowed to be updated simultaneously,

bridging the extremes of fully joint (FB algorithm) and fully individual (iFFBS algorithm) updates. As demonstrated in Equation (5.19), GIFB algorithm expands the exact full conditionals $\pi\big(\boldsymbol{Z}_j \mid \boldsymbol{X}_j, \boldsymbol{Z}_{-j}, \boldsymbol{\theta}, \boldsymbol{\beta}\big)$ in the forward direction like standard FB algorithm through utilizing the first-order Markov property. In contrast, GIFB utilize the marginal transition probabilities η, as specified in Equation (5.21) rather than the joint transition probabilities in the standard FB algorithm, which allowing a natural incorporation with the Hierarchical Marginal Model that targetting at the marginal transition probabilities. Furthermore, the GIFB algorithm avoids the exponential growth in computational cost of the FB algorithm while retaining full dependence. Its computational cost is $O\Big(T\sum_{j=1}^{J}|\mathcal{S}_j|^2\Big)$, which scales quadratically in the state space size of each group, making it much more scalable than FB algorithm. Additionally, unlike iFFBS, which updates chains element-wise, GIFB allows a combination of group-wise and element-wise updates, providing the flexibility to trade-off between computational cost and convergence speed.

5.3.6 Gibbs Sampler for MHMMs

We now turn our attention to fitting an MHMM in which the transition dynamics of the hidden process are parametrised through the marginal transition probabilities η, which are explicitly modelled via the regression coefficients $\boldsymbol{\beta}$ under the Hierarchical Marginal Model (see Section 4.2), rather than the joint transition probabilities p. This choice is motivated by the fact that the GIFB algorithm (see Section 5.3.4) operates directly on η, making it a natural and efficient parametrisation for inference. Additionally, by working with η, dependency structures among the hidden chains can be detected and tested in a straightforward manner (see Proposition 4.13), whereas working with the joint transition probabilities p would require an additional transformations.

Inspired by the data–augmentation strategy of Fearnhead and Sherlock [2006], we adopt a Gibbs sampler that iteratively samples from the full conditional of the hidden chain given the parameters, and then from the conditional distribution of the parameters given the hidden chain, as an alternative to the EM algorithm. In contrast to the original use

of the standard FB algorithm, we replace it with the GIFB algorithm (Section 5.3.4) and incorporate it with the Hierarchical Marginal Model (Section 4.2) to efficiently infer the hidden states under the multivariate setting.

Start with the joint posterior: $\pi(\boldsymbol{Z}, \boldsymbol{\theta}, P \mid \boldsymbol{X})$

By Lemma 4.12, under *strong compatibility* (see Definition 4.1), the map between the joint transition matrix P and the regression coefficients $\boldsymbol{\beta}$ is a diffeomorphism. Furthermore, since $\boldsymbol{\beta}$ not only directly parametrise the marginal transition probabilities η used inside GIFB but also encodes the dependency structures explicitly, we opt to reparametrise the posterior as:

$$\pi(\boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \boldsymbol{X})$$

and work with $\boldsymbol{\beta}$ in place of P.

The Gibbs sampler generates samples from this joint posterior distribution by iteratively updating the hidden states $\boldsymbol{Z}$, the emission parameters $\boldsymbol{\theta}$, and the transition regression coefficients $\boldsymbol{\beta}$. Specifically:

- $\pi(\boldsymbol{Z} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{X})$: The hidden states $\boldsymbol{Z}$ are sampled using the GIFB algorithm, where the chains are updated group-wise through a nested Gibbs scheme. This is feasible because the GIFB algorithm operates with the marginal transition probabilities η, which are directly parametrised by the regression coefficients $\boldsymbol{\beta}$ via the multinomial logistic regression.

- $\pi(\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{X})$: Once we know the distributional form of the emission process, under the one-to-one emission Assumption 5.3, the likelihood $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ of the observed data can be factorized as Equation (5.2). Then by Bayes' Theorem, the

full conditional can be computed as:

$$\begin{aligned}\pi\big(\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{X}\big) &\propto \pi(\boldsymbol{\theta})\, \mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\beta}) \\ &\propto \pi(\boldsymbol{\theta}) \prod_{k=1}^{K} \prod_{t=1}^{T} \mathbb{P}\big(X_k(t) \mid \theta_k, Z_k(t)\big) \end{aligned} \tag{5.27}$$

The posterior can be written as a product of some prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and the emission density. Depending on the emission family, $\boldsymbol{\theta}$ can be sampled efficiently via a conjugate posterior update or, for more complex emission models, via a Metropolis–Hastings step.

- $\pi\big(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{X}\big) = \pi\big(\boldsymbol{\beta} \mid \boldsymbol{Z}\big)$: Conditioned on the hidden states $\boldsymbol{Z}$, the regression coefficients $\boldsymbol{\beta}$ are independent of both the emission parameters $\boldsymbol{\theta}$ and the observed data $\boldsymbol{X}$. This reduces the problem to the Bayesian inference on MMC described in Section 3.5.1. In this case, we adopt the Pólya-Gamma sampling scheme introduced in Section A.3.1, which allows direct sampling from the full conditional of $\boldsymbol{\theta}$ through augmentation with auxiliary Pólya-Gamma variables.

The entire Gibbs sampling procedure for inferring these parameters in the MHMM is summarized in the following Algorithm 7.

Unlike the EM algorithm, which might suffer from the intractable computational costs in the multivariate setting, the proposed Gibbs sampler effectively updates each parameter from its exact full conditional distribution and naturally incorporates the GIFB algorithm, which operates on the marginal transition probabilities modelled by the Hierarchical Marginal Model. In the next section, we introduce the Viterbi algorithm, which provides the optimal state sequence given the estimated (posterior median) regression coefficients $\hat{\boldsymbol{\beta}}$ and emission parameters $\hat{\boldsymbol{\theta}}$ from the Gibbs sampler 7.

Algorithm 7 Gibbs Sampler for the Parameter Inference in MHMM

Input: observations $\boldsymbol{X}$, the prior $\pi(\boldsymbol{\theta})$ for the emission parameters, the prior parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ for regression coefficients $\boldsymbol{\beta}$, the number of iterations N

Output: A MCMC sequence of the sampled hidden Markov Chain $\boldsymbol{Z}$, emission parameters $\boldsymbol{\theta}$ and the regression coefficients $\boldsymbol{\beta}$

Initialize the emission parameters $\boldsymbol{\theta}^{(0)} \sim \pi(\boldsymbol{\theta})$ and the regression coefficients $\boldsymbol{\beta}^{(0)}$ based on $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ via Algorithm 2

for each iteration $n = 1, \ldots, N$ **do**

1. **Sample** $\boldsymbol{Z}^{(n)} \sim \pi\big(\boldsymbol{Z} \mid \boldsymbol{\theta}^{(n-1)}, \boldsymbol{\beta}^{(n-1)}, \boldsymbol{X}\big)$ by via GIFB Algorithm 6

2. **Sample** $\boldsymbol{\theta}^{(n)} \sim \pi\big(\boldsymbol{\theta} \mid \boldsymbol{Z}^{(n)}, \boldsymbol{\beta}^{(n-1)}, \boldsymbol{X}\big)$ using either a conjugate posterior distribution, or an Metropolis-Hastings step (Algorithm 12), from Equation (5.27).

3. **Sample** $\boldsymbol{\beta}^{(n)} \sim \pi\big(\boldsymbol{\beta} \mid \boldsymbol{Z}^{(n)}\big)$ by Pólya Gamma via Algorithm 9

end for

5.3.7 Viterbi Algorithm

In many applications, it is of interest to recover a single optimal hidden path $\boldsymbol{Z} = \boldsymbol{z}^\star$ rather than relying on the full set of posterior samples obtained from the Gibbs sampler. To this end, we employ the Viterbi algorithm by Viterbi [2003], which efficiently computes the most likely path $\boldsymbol{z}^\star$ of the hidden Markov chain. Given the posterior medians of the emission parameters $\hat{\boldsymbol{\theta}}$ and the regression coefficients $\hat{\boldsymbol{\beta}}$ from the Gibbs sampler (Algorithm 7), the optimal path is obtained by solving

$$\boldsymbol{z}^\star = \arg\max_{\boldsymbol{Z}} \ \pi(\boldsymbol{Z} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}) \tag{5.28}$$

The key idea of the Viterbi algorithm is to exploit the conditional structure of the problem, reducing the global maximization into a sequence of smaller sub-problems. The following Remark 5.4 provides the fundamental principle underlying the recursion:

Remark 5.4. If $f(a) \geq 0 \quad \forall a \quad$ and $\quad g(a,b) \geq 0 \quad \forall a, b$, then

$$\max_{a,b} f(a)g(a,b) = \max_{a}\{f(a) \max_{b} g(a,b)\} \tag{5.29}$$

This follows directly by $f(a)$ is independent of b.

We begin with the objective

$$\max_{\boldsymbol{Z}} \pi(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\beta})$$

For convenience, define

$$\mu_t\big(\boldsymbol{Z}(t)\big) = \max_{\boldsymbol{Z}(1:(t-1))} \pi\big(\boldsymbol{Z}(1{:}t), \boldsymbol{X}(1{:}t) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\big)$$

that is, the maximum joint probability of the partial path $\boldsymbol{Z}(1{:}t)$ and observations $\boldsymbol{X}(1{:}t)$ ending in state $\boldsymbol{Z}(t)$. By factorization of the likelihood in Equation (5.2), we obtain

$$\begin{aligned}\mu_t\big(\boldsymbol{Z}(t)\big) &= \max_{\boldsymbol{Z}(1:(t-1))} \pi\big(\boldsymbol{Z}(1{:}t), \boldsymbol{X}(1{:}t) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\big) \\ &= \max_{\boldsymbol{Z}(1:(t-1))} \Big\{\mathbb{P}\big(\boldsymbol{X}(t) \mid \boldsymbol{Z}(t), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1), \boldsymbol{\beta}\big)\, \pi\big(\boldsymbol{Z}\big(1{:}(t-1)\big), \boldsymbol{X}\big(1{:}(t-1)\big) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\big)\Big\}\end{aligned}$$

Applying the decomposition trick in Remark 5.4 with $a = \boldsymbol{Z}(t-1)$ and $b = \boldsymbol{Z}(1{:}(t-2))$, we obtain the recursive form for $t = 2, \ldots, T$:

$$\mu_t\big(\boldsymbol{Z}(t)\big) = \max_{\boldsymbol{Z}(1:(t-1))} \Big\{\mathbb{P}\big(\boldsymbol{X}(t) \mid \boldsymbol{Z}(t), \boldsymbol{\theta}\big)\, \mathbb{P}\big(\boldsymbol{Z}(t) \mid \boldsymbol{Z}(t-1), \boldsymbol{\beta}\big)\, \mu_{t-1}\big(\boldsymbol{Z}(t-1)\big)\Big\} \tag{5.30}$$

The recursion is initialized at $t = 1$ by:

$$\mu_1\big(\boldsymbol{Z}(1)\big) = \pi\big(\boldsymbol{Z}(1{:}t), \boldsymbol{X}(1{:}t) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\big) = \pi_0\big(\boldsymbol{Z}(1)\big)\, \mathbb{P}\big(\boldsymbol{X}(1) \mid \boldsymbol{Z}(1), \boldsymbol{\theta}\big) \tag{5.31}$$

Equations (5.31) and (5.30) together define the recursion for the Viterbi algorithm. In practice, the recursion is typically implemented on the log scale to avoid numerical underflow.

Note that this recursion ends up with:

$$\mu_T\big(\boldsymbol{Z}(T)\big) = \max_{\boldsymbol{Z}(1:(T-1))} \pi\big(\boldsymbol{Z}(1{:}T), \boldsymbol{X}(1{:}T) \mid \boldsymbol{\theta}, \boldsymbol{\beta}\big)$$

To achieve the final objective, we additionally maximize over the terminal state $\boldsymbol{Z}(T)$, yielding

$$\boldsymbol{z}^\star = \max_{\boldsymbol{Z}(T)} \mu_T\big(\boldsymbol{Z}(T)\big) = \max_{\boldsymbol{Z}} \ \pi(\boldsymbol{Z} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$$

which corresponds to the target in Equation (5.28). For the corresponding optimal path $\boldsymbol{z}^\star$, it suffices to record at each recursion step the maximizing predecessor state $\boldsymbol{Z}(t-1)$ for each candidate $\boldsymbol{Z}(t)$.

The entire process of the Viterbi algorithm to recover the most likely hidden path is summarized in Algorithm 8.

Algorithm 8 Viterbi Algorithm

Input: observations $\boldsymbol{X}$, parameters $\boldsymbol{\theta}$, initial distribution $\pi_0(\cdot)$, transition probabilities P, emission functions $\mathbb{P}(X_k(t) \mid Z_k(t), \theta_k)$ for each chain k

Output: the optimal hidden path $\boldsymbol{z}^\star$ maximizes the likelihood.

Initialize $\mu_1\big(\boldsymbol{Z}(1)\big)$ according to Equation (5.31)
for $t = 2 \ldots, T$ **do**
  Compute $\mu_t\big(\boldsymbol{Z}(t)\big)$ according to Equation (5.30)
  Record the arguments of the maxima $\boldsymbol{z}^\star\big(1{:}(t-1)\big)$
end for
Compute $\boldsymbol{z}^\star = \max_{\boldsymbol{Z}(T)} \mu_T\big(\boldsymbol{Z}(T)\big)$

In the next section, we present simulation studies illustrating how the Gibbs sampler 7 can be employed to estimate the emission parameters $\boldsymbol{\theta}$ and the regression coefficients $\boldsymbol{\beta}$, enabling tests of potential dependency structures within the MHMM. Furthermore, we demonstrate how the hidden states $\boldsymbol{Z}$ can be recovered through these parameter estimates by computing the optimal path $\boldsymbol{z}^\star$ using the Viterbi Algorithm 8.

5.4 Simulation Study

Having established the theoretical framework for parameter inference in MHMMs, we now conduct simulation studies to evaluate and compare two algorithms:

- The standard FB algorithm (see Section 5.3.2), which operates on the joint transition probabilities p
- The proposed GIFB algorithm (see Section 5.3.4), which operates on the marginal transition probabilities η, with each group set as a singleton so that chains are updated individually (i.e. resemble the setting in iFFBS in Section 5.3.3)

The comparison is conducted with respect to:

1. and the ability in detecting dependency structures within $\boldsymbol{Z}$;
2. the accuracy in recovering the hidden states $\boldsymbol{Z}$;
3. the computational cost.

In other words, we wish to demonstrate the advantage of the GIFB algorithm in retaining the accuracy of the FB algorithm in recovering hidden states and detecting dependency structures, while maintaining an affordable computational cost.

Since the GIFB algorithm operates directly on the marginal transition probabilities η rather than the joint transition probabilities p, it naturally integrates with the Hierarchical Marginal Model (see Section 4.2), which explicitly models η. For this reason, we focus exclusively on reparametrising p via η when applying GIFB, as this provides a direct and efficient framework for detecting dependency structures. In contrast, working with p would require an additional transformation into η and subsequently into the regression coefficients $\boldsymbol{\beta}$ in order to identify potential dependency structures.

Additionally, we do not include the iFFBS algorithm of Touloupou et al. [2020] (see Section 5.3.3) in our comparison, since it has been designed for the CHMM framework

where the conditional independence is inherently assumed, and thus it cannot capture or infer the dependency structures, especially for the concurrent one embedded within MHMM.

5.4.1 Simulation Setup

For simplicity of this particular simulation study, we consider an MHMM with 3 chains and 2 states $\mathcal{S} = \{1, 2\}$. We define the state-dependent emission process of the observation $X(t)$ as follows:

$$X(t) \mid Z(t) = \begin{cases} Y(t) & \text{if} \quad Z(t) = 1 \\ Y(t) + W(t) & \text{if} \quad Z(t) = 2 \end{cases} \tag{5.32}$$

where:

- $Y(t) \sim \text{Po}\big(\lambda(t)\big)$ denotes the baseline Poisson random variable with intensity function $\lambda(t)$. We further assume that $\lambda(t) \equiv \lambda$ is **time-invariant** in the simulation study, corresponding to the classical setting of Fischer and Meier-Hellstern [1993].
- $W(t) \sim \text{Po}(c)$ for some $c > 0$ are state-specific Poisson random variables with the corresponding constant c, associated with the non-baseline hidden state $Z(t) = 2$, thereby introducing the effect of $Z(t)$ into the emission function.

This construction can be interpreted as follows: when the hidden state is at the baseline $Z(t) = 1$, the observation $X(t)$ follows a Poisson distribution with constant rate λ. If the hidden state switches to the alternative $Z(t) = 2$, an additional number of counts following Poisson(c) is introduced additively to the baseline Poisson(λ), making the total observation following a equivalent Poisson distribution with rate $\lambda + c$.

Having specified the model setting, we now outline the simulation procedure as follows:

1. Similar as Section 3.5, we focus on both *contemporaneous independence* and *Granger non-causality*. Specifically, we generate the regression coefficients $\boldsymbol{\beta}$ in the Hierarchical Marginal Model using Algorithm 2, under:

(a) $Z_1(t)$ and $Z_2(t)$ are contemporaneously independent from $Z_3(t)$:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1) \tag{5.33}$$

Recall that this is exactly the one illustrated in Figure 2.6 when we first introduce the contemporaneous independence and the one in Equation 3.25 that is used in Simulation study of MMC in Section 3.5.

(b) $Z_1(t)$ and $Z_2(t)$ are not Granger-caused by $Z_3(t)$:

$$Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1) \tag{5.34}$$

Recall that this is exactly the one illustrated in Figure 2.8 when we first introduce the Granger non-causality and the one in Equation 3.26 that is used in Simulation study of MMC in Section 3.5.

2. Compute the corresponding transition matrix P based on $\boldsymbol{\beta}$ under Hierarchical Marginal Model through Lemma 4.12.

3. Simulate an MMC $\boldsymbol{Z}$ of length 2000, according to this transition matrix P.

4. Based on the simulated $\boldsymbol{Z}$, as well as the emission function f and the predefined parameters $\boldsymbol{\theta}$, we generate the observations $\boldsymbol{X}$.

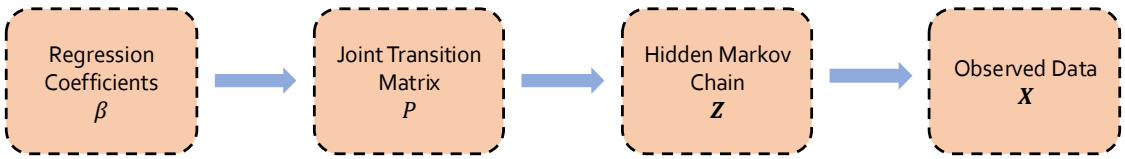


Figure 5.3: An illustration of simulation process from step 1 to step 4

5. A **full model** with full dependency structures is fitted to $\boldsymbol{X}$ via

- A Gibbs sampler with standard FB on joint transition probabilities p
- A Gibbs sampler with GIFB on marginal transition probabilities η

with 2000 MCMC iterations.

6. Following the setup in Section 3.5, we assess the recovery of the true regression coefficients $\boldsymbol{\beta}$ by examining box plots of the posterior samples from:

   - $\boldsymbol{\beta}$ transformed from p obtained via the Gibbs sampler with the standard FB
   - $\boldsymbol{\beta}$ obtained via the Gibbs sampler with GIFB.

   We further compute the corresponding contour probabilities (see Section A.2.2) of these (transformed) posterior samples $\boldsymbol{\beta}$ to evaluate the detection of dependency structures.

7. We measure the accuracy of the recovered hidden path $\boldsymbol{z}^\star$ by Viterbi algorithm using the confusion matrix.

8. We compare the computational efficiency of the two algorithms by recording the runtime required to complete 2000 MCMC iterations.

9. We repeat Steps 3 to 8 for 200 independent experiments, each generated from the same transition matrix P, for both Granger non-causality and contemporaneous independence.

For completeness, we also present MCMC diagnostics for the emission parameters $\boldsymbol{\theta}$ and the regression coefficients $\boldsymbol{\beta}$ in Appendix B.2.2, in order to assess convergence and stationarity of the MCMC chains via visual inspection.

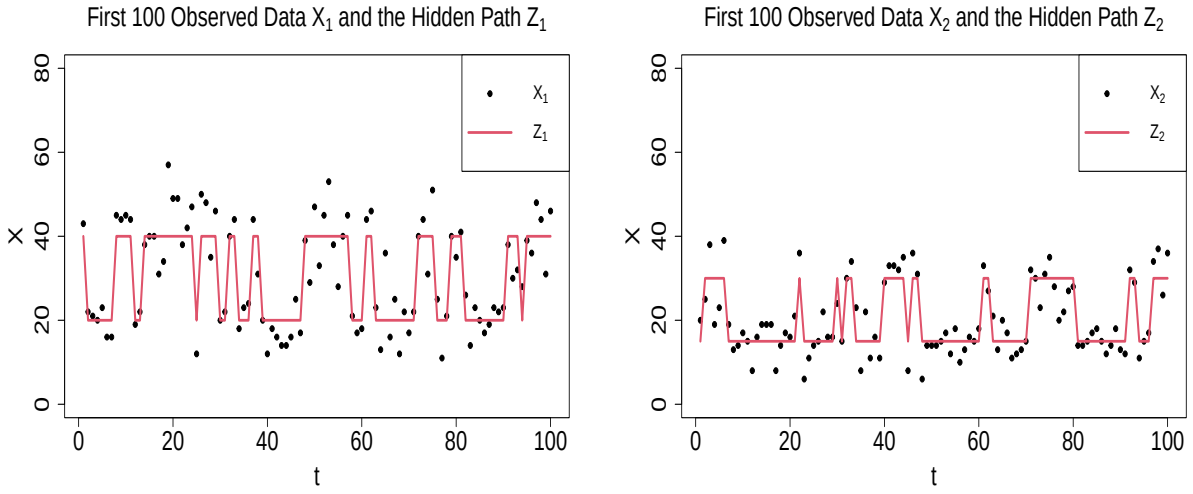
5.4.2 Simulated Data

In this section, we present simulated data generated under the setup 5.4.1, following the procedure illustrated in Figure 5.3. The emission parameters used in the simulation are summarized in Table 5.2.

Dimension	1	2	3
λ	20	15	15
c	20	15	20

Table 5.2: Emission parameters $\boldsymbol{\theta} = (\lambda, c)$ used in the simulation

The following Figures 5.4a and 5.4b display the first 100 simulated observations $\boldsymbol{X}$ (black points) together with the corresponding hidden states $\boldsymbol{Z}$ (red lines) for the first chain Z_1 and the second chain Z_2, respectively, from the first experiment under the contemporaneous independence (5.33).

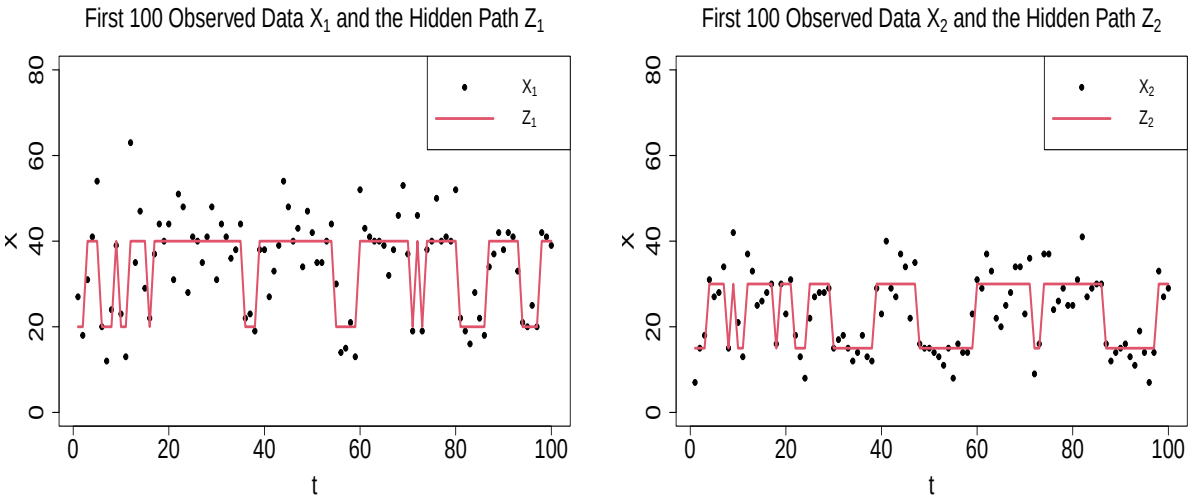


(a) The simulated $X_1(t)$ and $Z_1(t)$ (b) The simulated $X_2(t)$ and $Z_2(t)$

Figure 5.4: The simulated observed data $\boldsymbol{X}$ and the corresponding hidden process $\boldsymbol{Z}$ is given for chain 1 and chain 2 under **contemporaneous independence**. The hidden states $\boldsymbol{Z}$ should only take values in $\mathcal{S} = \{1, 2\}$, for visualization purposes we plot $20 \times Z_1(t)$ for dimension 1 and $15 \times Z_2(t)$ for dimension 2 to align with the corresponding Poisson intensities

Similarly, Figures 5.5a and 5.5b show the first 100 simulated observations $\boldsymbol{X}$ (black points) along with the corresponding hidden states $\boldsymbol{Z}$ (red lines) for the first chain Z_1 and the second chain Z_2, respectively, from the first experiment under the Granger non-causality (5.33).

In principle, the hidden states $\boldsymbol{Z}$ take values in $\mathcal{S} = \{1, 2\}$, but for visualization we plot $20 \times Z_1(t)$ and $15 \times Z_2(t)$ to align them with the corresponding Poisson intensities. As can be seen from the figures, when Z_1 and Z_2 alternate between states 1 and 2, the cor-



(a) The simulated $X_1(t)$ and $Z_1(t)$ (b) The simulated $X_2(t)$ and $Z_2(t)$

Figure 5.5: The simulated observed data $\boldsymbol{X}$ and the corresponding hidden process $\boldsymbol{Z}$ is given for chain 1 and chain 2 under **Granger non-causality**. The hidden states $\boldsymbol{Z}$ should only take values in $\mathcal{S} = \{1, 2\}$, for visualization purposes we plot $20 \times Z_1(t)$ for dimension 1 and $15 \times Z_2(t)$ for dimension 2 to align with the corresponding Poisson intensities

responding observations X_1 and X_2 follow Poisson distributions with rates specified in Table 5.2.

From both Figures 5.4 and 5.4, the dependency structures in the data are not directly observable. As the result, our aim is to recover the latent state $\boldsymbol{Z}$, infer the underlying dependency structures, and estimate the emission parameters $\boldsymbol{\theta}$ from the observed data $\boldsymbol{X}$. We assess these tasks by comparing the standard FB algorithm with the proposed GIFB algorithm.

5.4.3 Results

In this section, we compare the performance of the standard FB algorithm and the proposed GIFB algorithm from three perspectives: (I) ability to detect potential dependency structures, (II) accuracy in recovering the hidden chain $\boldsymbol{Z}$, and (III) computational time, under both contemporaneous independent and Granger non-causality, as outlined at the beginning of Section 5.4. For completeness, the prior and posterior of the emission parameters $\boldsymbol{\theta}$, are presented in Appendix B.2.1, and the corresponding inference results,

including posterior medians and MCMC diagnostic trace plots, are reported in Appendix B.2.2.

5.4.3.1 Detect Dependency Structures

Firstly, we assess the ability of both FB and GIFB to detect the contemporaneous independence (5.33) and Granger non-causality (5.33), through the corresponding regression coefficients $\boldsymbol{\beta}$. Thanks to the Pólya Gamma data augmentation [Polson et al., 2013], the Gibbs sampler with GIFB operates directly on $\boldsymbol{\beta}$ within the Gibbs sampler, allowing posterior samples to be obtained immediately. In contrast, the Gibbs sampler with FB operates on the joint transition probabilities p, requiring an additional transformation into $\boldsymbol{\beta}$ via the deterministic map in Lemma 4.12 to obtain posterior samples.

Contemporaneous Independence

We first present the estimates of the first regression coefficient vector $\boldsymbol{\beta}_1$ for the univariate marginal $g_1 = \{1\}$ under the contemporaneous independence. These estimates are obtained as posterior medians from each of the 200 repeated experiments and are shown in the following Figure 5.6.

It is evident that the regression coefficients estimated by FB and GIFB are in close agreement, further confirming that GIFB functions as an exact Gibbs sampler, updating from the true full conditionals for each chain. Additionally, both algorithms provide accurate estimates of the true regression coefficients $\boldsymbol{\beta}_j$ under the contemporaneous independence, since the true values indicated by the red dots lying very close to the medians of the box plots.

Recall from Section 4.5.1, under the contemporaneous independence (5.33), the regression coefficients vector for the chain groups $g_j \in \{\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$ should take the same value (see Equation (4.39)). Hence for illustration, we present the first two regression coefficients, β_1, β_2 across these chain groups, as shown in the following Figure 5.7.

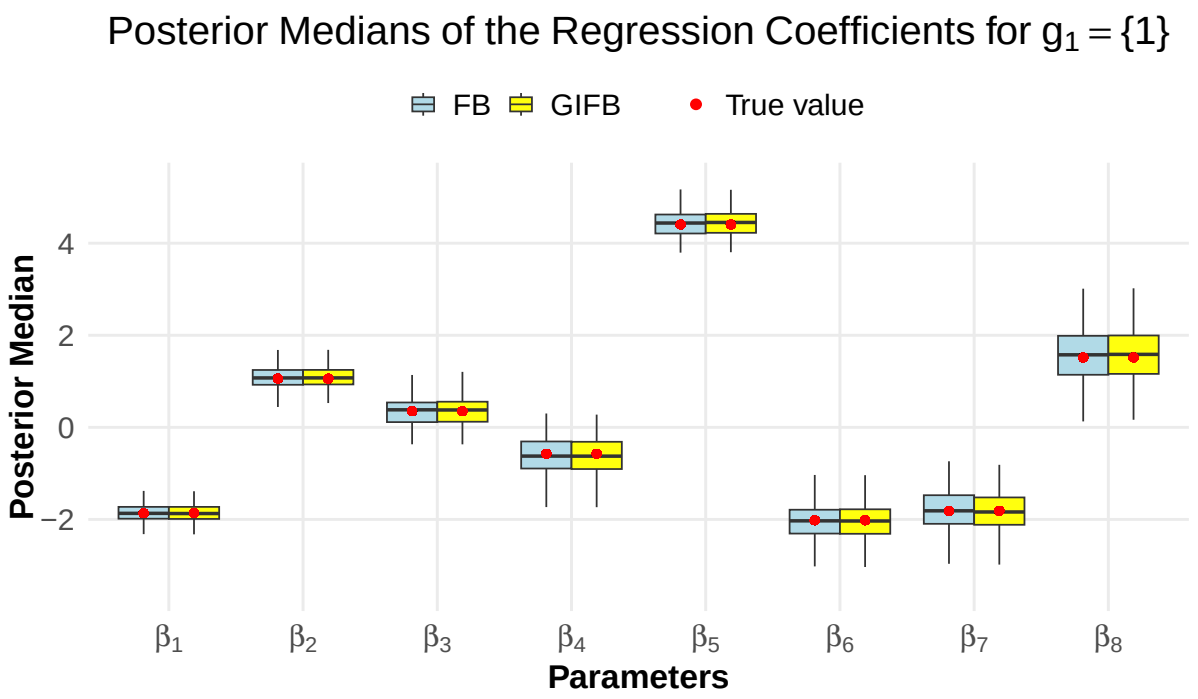


Figure 5.6: Box plots of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$ under the **contemporaneous independence**, obtained from the Gibbs samplers using FB (light blue) and GIFB (yellow) across 200 repeated experiments. The true parameter values are indicated by red points.

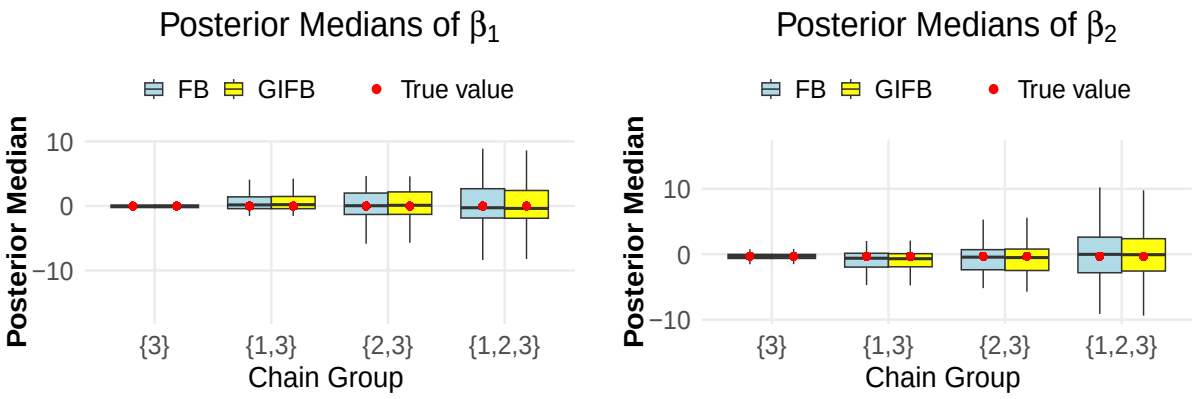
It can be seen from Figure 5.7 that the posterior mean estimates of the regression coefficients β_1 and β_2 are similar across the chain groups $g \in \{\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$, which is consistent with the contemporaneous independence specified in Equation 4.39.

However, a large variability is observed in β_1 and β_2 corresponding to the chain group $\{1,2,3\}$. This can be attributed to the relatively small number of observations in which chains 1, 2, and 3 are simultaneously in the alternative state, leading to increased uncertainty due to limited data support.

Now consider the following hypotheses:

$$H_0: \quad Z_1 \text{ and } Z_2 \text{ are contemporaneously independent from } Z_3$$

$$H_1: \quad Z_1 \text{ and } Z_2 \text{ are correlated with } Z_3$$



(a) The posterior medians of β_1

(b) The posterior medians of β_2

Figure 5.7: The box plots of the posterior medians of the first two regression coefficient β_1 and β_2 across 200 independent simulation experiments, for chain groups $g = \{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}$, obtained via both Gibbs sampler with FB (in light blue) and with GIFB (in yellow). Under the contemporaneous independence (5.33), β_1 and β_2 should be invariant across these chain groups, and therefore all box plots are expected to be centred at the same value.

We then compute the posterior contour probabilities p for the regression coefficients $\boldsymbol{\beta}_j$ associated with the chain groups $g_j \in \{\{3\}, \{1,3\}, \{2,3\}, \{1,2,3\}\}$, all centred at the same point. Following the procedure described in Section 4.5.1, this is achieved by examining whether the differences $\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'}$ are centred at zero for all pairs $g_j, g_{j'}$ within this set.

In accordance with Figure 5.7, it is not surprising to find the resulting posterior contour probabilities are all greater than 0.5, indicating no evidence against the null hypothesis H_0. This supports the presence of contemporaneous independence and demonstrates the ability of both methods to correctly identify the underlying dependency structures.

It is also worth noting that the variance of the estimates increases progressively from the chain groups $\{3\}$, $\{1,3\}$, and $\{2,3\}$ to $\{1,2,3\}$. This increase can be attributed to the diminishing amount of data for the regressions of higher-order marginals. To be specific, observations corresponding to the chain group $\{1,2,3\}$ require all three chains to be simultaneously in state 2, that is, $\boldsymbol{Z} = (2,2,2)$, which occurs far less frequently than events involving only a single chain, such as $Z_3 = 2$. The resulting scarcity of relevant observations leads to greater uncertainty and, consequently, higher variance in the corresponding estimates.

Granger non-causality

As for the Granger non-causality (5.34), the regression coefficients capturing effects from chain 3, namely β_2, β_4, β_6, and β_8 within $\boldsymbol{\beta}_j$ for the chain groups $g_j \in \{\{1\}, \{2\}, \{1,2\}\}$ should be 0 (see Equation (4.42)). Hence for illustration, we focus on the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$, as displayed in Figure 5.8.

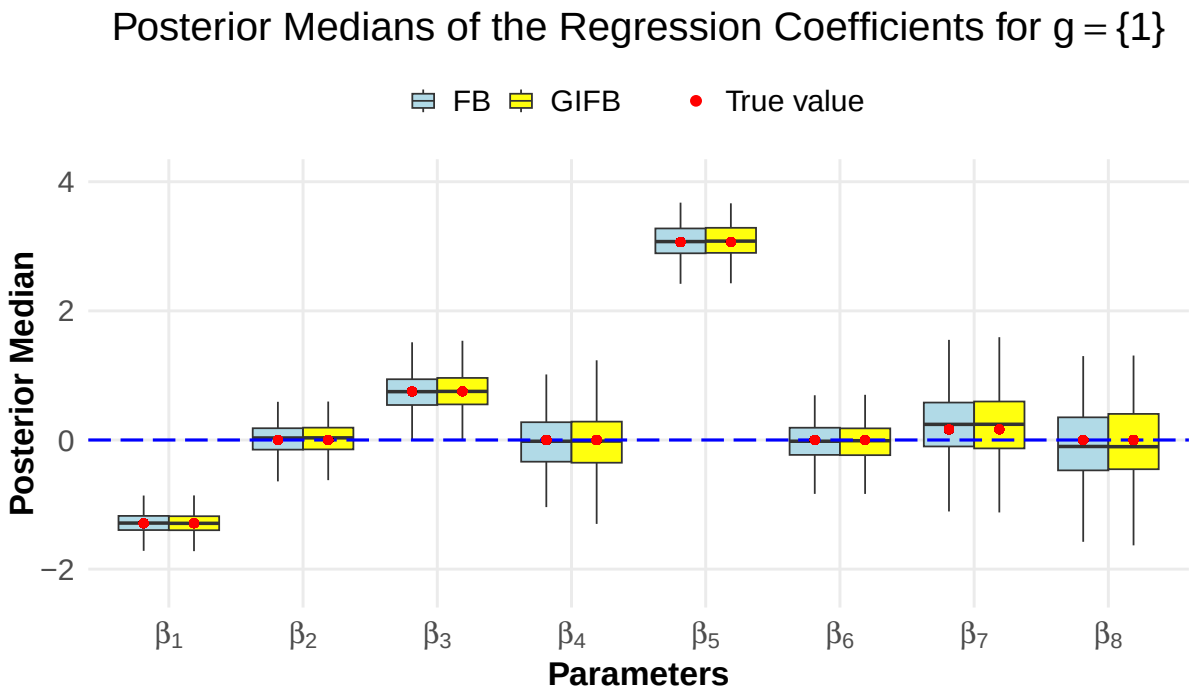


Figure 5.8: Box plots of the posterior medians of the regression coefficients $\boldsymbol{\beta}_j$ for the univariate chain group $g_j = \{1\}$ under **Granger non-causality**, obtained from the Gibbs samplers using FB (light blue) and GIFB (yellow) across 200 repeated experiments. The true parameter values are indicated by red points. Under the Granger non-causality, $\beta_2, \beta_4, \beta_6, \beta_8$ that accounts for the effect from chain 3 should be 0.

It is evident that the regression coefficients estimated by FB and GIFB are in close agreement, further confirming that GIFB functions as an exact Gibbs sampler, updating from the true full conditionals for each chain. Additionally, both algorithms provide accurate estimates of the true regression coefficients $\boldsymbol{\beta}_j$ under the Granger non-causality, with true values lying very close to the medians of the box plots. Importantly, the coefficients $\beta_2, \beta_4, \beta_6, \beta_8$, which represent the effects from chain 3, should be zero in the regression coefficients $\boldsymbol{\beta}_j$ under the Granger non-causality, which is consistent as that given in Equa-

tion 4.42.

Now consider the following hypotheses:

$$H_0: \quad Z_1 \text{ and } Z_2 \text{ are not Granger caused by } Z_3$$
$$H_1: \quad Z_1 \text{ and } Z_2 \text{ are Granger caused by } Z_3$$

Similarly, it is not surprising that the posterior contour probabilities p for these four regression coefficients (as well as the corresponding four coefficients for $g_j \in \{\{2\}, \{1,2\}\}$, omitted for brevity) are centred at zero with probabilities exceeding 0.5 under both FB and GIFB. This indicates no evidence against the null hypothesis H_0, thereby supporting the presence of contemporaneous independence and further demonstrating the ability of both methods to correctly identify the underlying dependency structures.

5.4.3.2 Recovery of the Hidden Chain

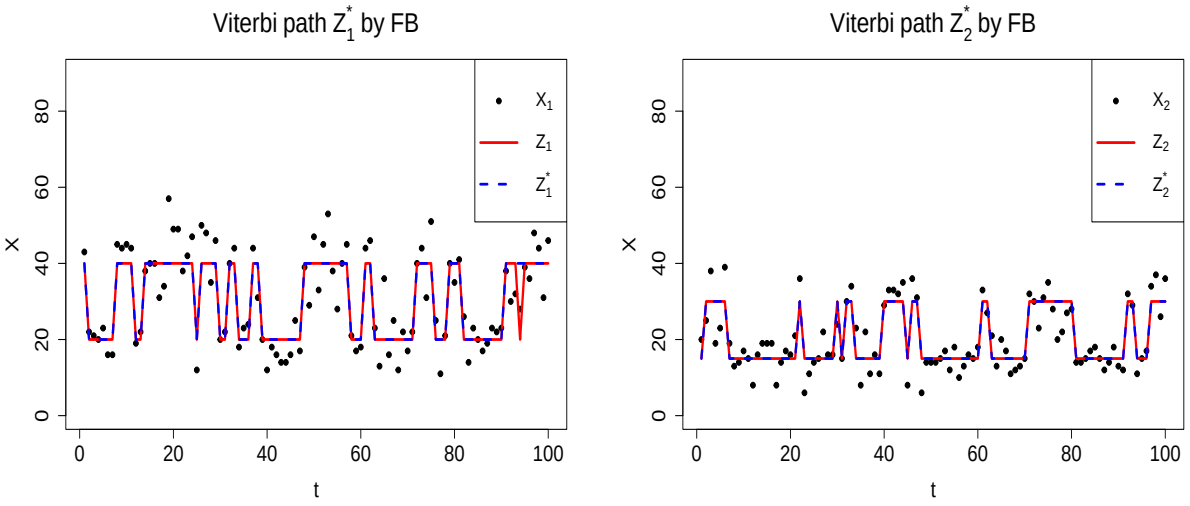
Next, we examine the recovery of the hidden states $\boldsymbol{Z}$ using both the FB and GIFB algorithms. To obtain a single optimal hidden path $\boldsymbol{Z}^\star$, rather than manipulating on the full set of posterior samples, we apply the Viterbi algorithm reviewed in Section 5.3.7. Specifically, we compute

$$\boldsymbol{Z}^\star = \arg\max_{\boldsymbol{Z}}; \pi(\boldsymbol{Z} \mid \boldsymbol{X}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}),$$

where the posterior median estimators $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}$ are obtained from the Gibbs sampler. We perform this procedure twice: once using $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ from the Gibbs sampler incorporating the standard FB algorithm, yielding $\boldsymbol{Z}^\star_{\text{FB}}$, and once using $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}})$ from the Gibbs sampler incorporating GIFB, yielding $\boldsymbol{Z}^\star_{\text{GIFB}}$. The accuracy of hidden state recovery is then assessed by comparing $\boldsymbol{Z}^\star_{\text{FB}}$ and $\boldsymbol{Z}^\star_{\text{GIFB}}$ against the true latent path.

Contemporaneous Independence

We start with the contemporaneous independence setting. The recovery of the hidden states using the FB algorithm is illustrated in Figure 5.9. The plot displays the Viterbi paths $\boldsymbol{Z}^{\star}_{\text{FB}}$ in blue dashed lines against the true hidden states $\boldsymbol{Z}$ in red solid lines for the first 100 time points of $Z_1(t)$ and $Z_2(t)$ in the first experiment, respectively (see Figure 5.4). As can be seen, the Viterbi paths closely follow the true trajectories, demonstrating that $\boldsymbol{Z}^{\star}_{\text{FB}}$ achieve good recovery of the latent process.

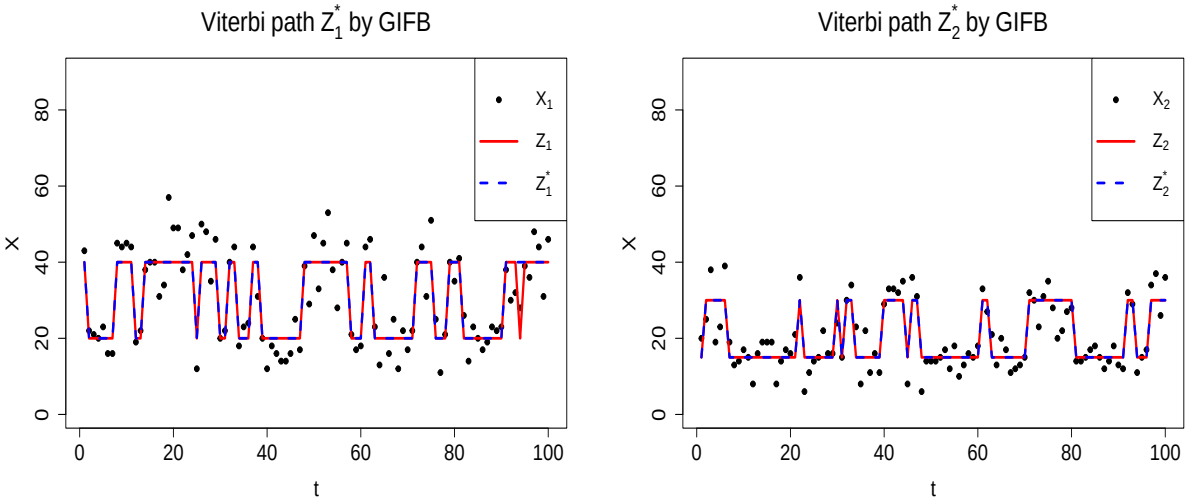


(a) Viterbi path $\boldsymbol{Z}^{\star}_1$ by FB (b) Viterbi path $\boldsymbol{Z}^{\star}_2$ by FB

Figure 5.9: The Viterbi paths $\boldsymbol{Z}^{\star}_{\text{FB}}$ (blue dashed) compared with the true hidden states $\boldsymbol{Z}$ (red solid) for the first 100 time points of $Z_1(t)$ and $Z_2(t)$ under the contemporaneous independence (5.33).

In comparison, Figure 5.10 shows the Viterbi path $\boldsymbol{Z}^{\star}_{\text{GIFB}}$ obtained from the GIFB algorithm. Again, the Viterbi paths $\boldsymbol{Z}^{\star}_{\text{GIFB}}$ in blue dashed lines are plotted against the true states $\boldsymbol{Z}$ in red solid lines for $Z_1(t)$ and $Z_2(t)$ respectively. Similar to $\boldsymbol{Z}^{\star}_{\text{FB}}$, $\boldsymbol{Z}^{\star}_{\text{GIFB}}$ also aligns closely with the truth, indicating that the proposed GIFB retains the high accuracy in recovering the hidden states while operating under reduced computational cost.

It is worth noting that for both Viterbi paths, $\boldsymbol{Z}^{\star}_{\text{FB}}$ and $\boldsymbol{Z}^{\star}_{\text{GIFB}}$, the recovery of chain 1 deviates from the truth at $t = 94$ (see Figures 5.9a and 5.10a). At this time point, the true hidden state is $Z_1(t) = 1$ (shown in the red path), while the observed count $X_1(t)$



(a) Viterbi path $\boldsymbol{Z}_1^\star$ by GIFB (b) Viterbi path $\boldsymbol{Z}_2^\star$ by GIFB

Figure 5.10: The Viterbi paths $\boldsymbol{Z}^\star_{\text{GIFB}}$ (blue dashed) compared with the true hidden states $\boldsymbol{Z}$ (red solid) for the first 100 time points of $Z_1(t)$ and $Z_2(t)$ under the contemporaneous independence (5.33).

falls closer to the intensity associated with $Z_1(t) = 2$. Consequently, the most likely path fail to recover the truth, illustrating how randomness in the emission process can lead to occasional mismatches between the true and recovered hidden states.

Additionally, for each experiment, we compare the Viterbi paths obtained from FB and GIFB with the true hidden state trajectories. By aggregating the classification results over all time points and across the 200 repeated experiments ($2000 \times 200 = 400000$ observations), we construct two aggregated confusion matrices, one for FB (Figure 5.11) and one for GIFB (Figure 5.12), which summarize the overall state recovery performance of the two methods.

Each confusion matrix provides a comparison between the true joint hidden states of the three chains (rows) and the corresponding states inferred by the algorithm (columns). The joint states are denoted as $(1,1,1), (2,1,1), \ldots, (2,2,2)$, representing all possible combinations of the joint states from a binary state space $\mathcal{S} = \{1, 2\}$ with 3 chains.

The entries of the matrix record the number of time points for which a particular true

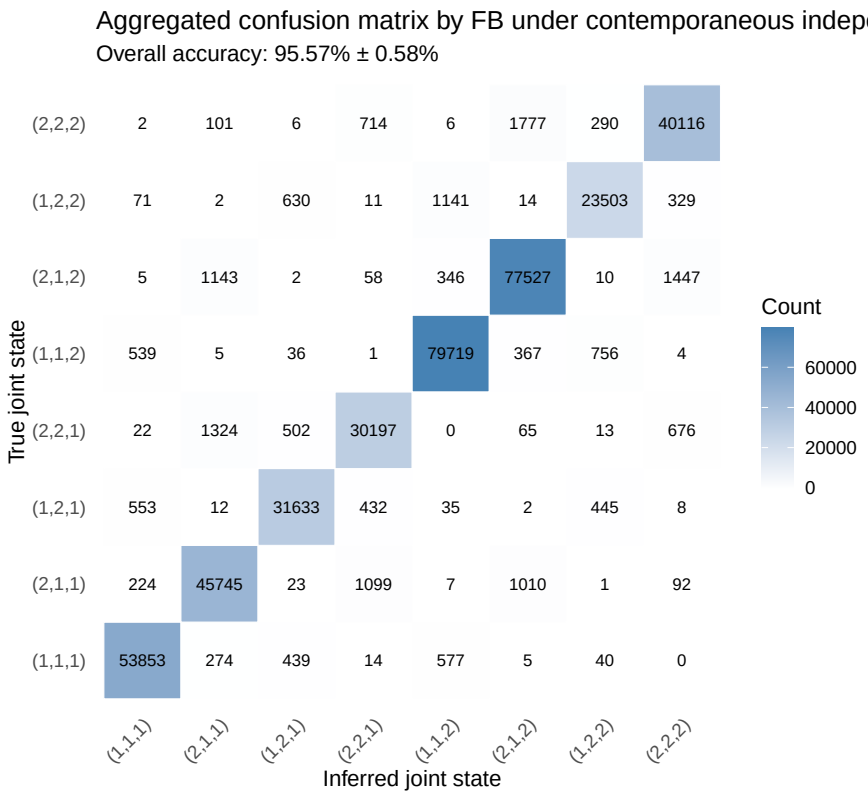


Figure 5.11: Confusion matrix for $\boldsymbol{Z}^{\star}_{\mathrm{FB}}$ under the contemporaneous independence (5.33)

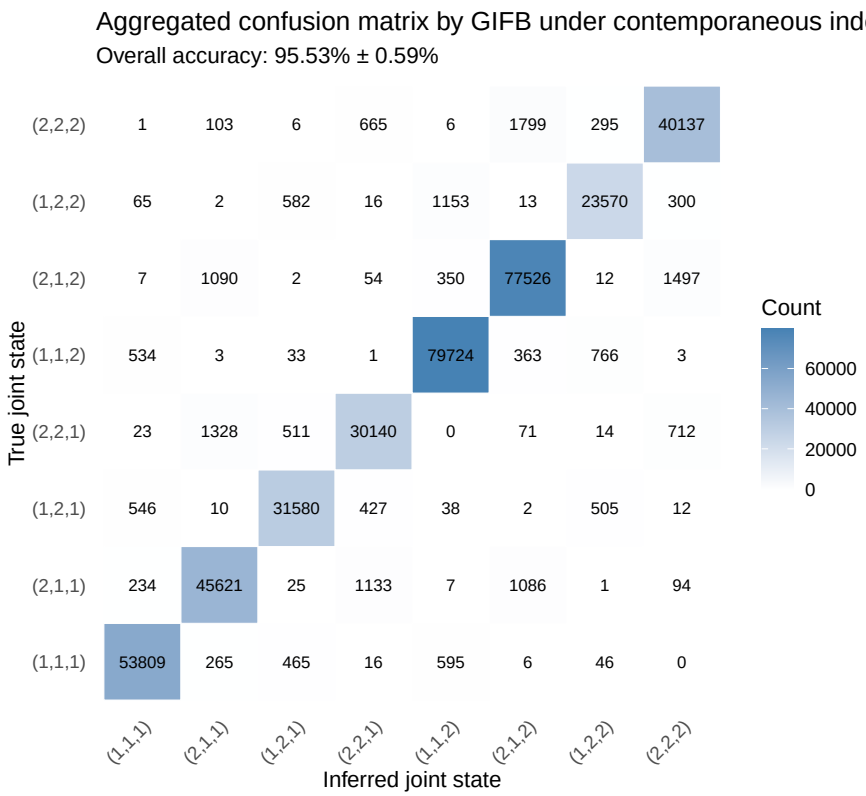


Figure 5.12: Confusion matrix for $\boldsymbol{Z}^{\star}_{\mathrm{GIFB}}$ under the contemporaneous independence (5.33)

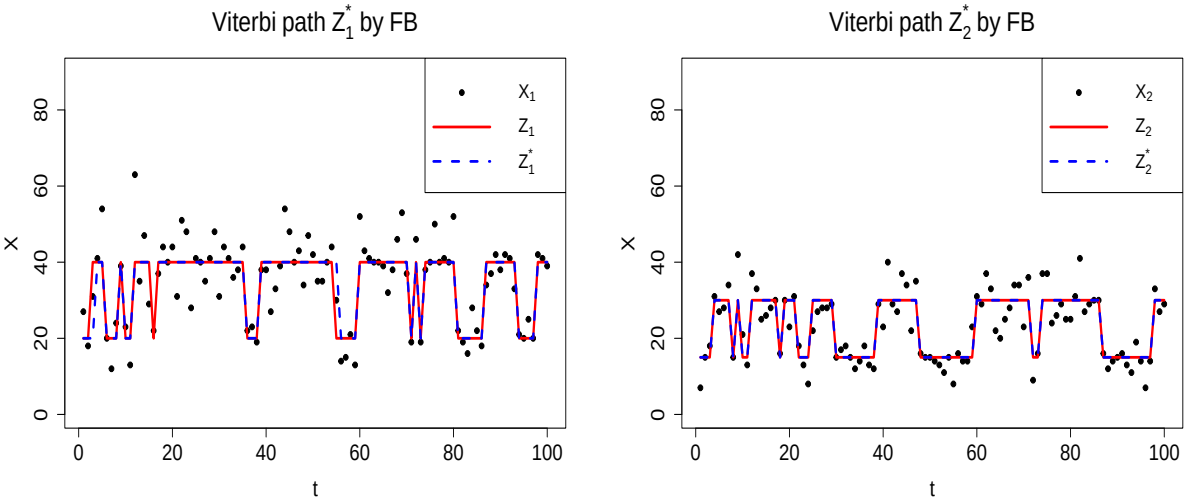
state (row) was classified into a given inferred state (column) with darker shading reflecting higher counts. For instance, in Figure 5.11, the bottom-left cell at row $(1, 1, 1)$ and column $(1, 1, 1)$ contains a value of 53853, showing that the state $(1, 1, 1)$ was correctly identified for 53853 times. It follows that diagonal cells correspond to correct classifications, whereas off–diagonal cells indicate misclassification. The predominance of dark-coloured cells along the diagonal indicates that the majority of joint states are correctly identified.

Specifically, as reported at the top of the confusion matrices, the Viterbi paths obtained using FB achieve an overall accuracy of 95.57% with a standard deviation of 0.58%, while those obtained using GIFB achieve an overall accuracy of 95.53% with a standard deviation of 0.59%. The differences among the corresponding confusion matrix entries are negligible and can be attributed to Monte Carlo error. These results further confirm that the proposed GIFB preserves the high accuracy of FB and validate that GIFB is an exact Gibbs sampler, updating each chain according to its true full conditional distribution.

Granger non-Causality

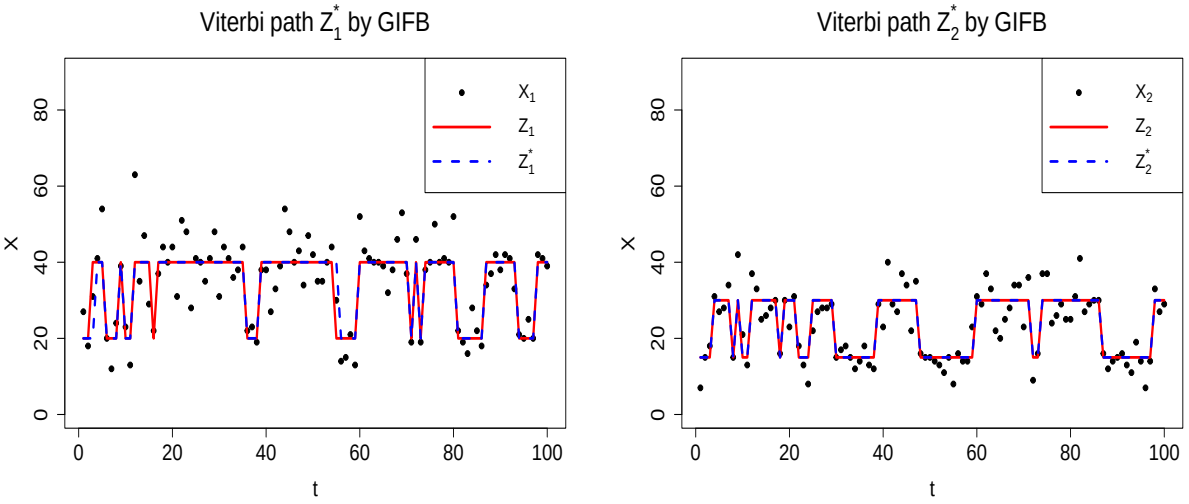
As in the case of contemporaneous independence, we also illustrate the recovery of the hidden state paths using the Viterbi algorithm with posterior median estimates obtained from both FB and GIFB under the Granger non-causality. The recovered paths for the first two dimensions, Z_1 and Z_2 for the first 100 observations in the first experiment (see Figure 5.5), are shown in Figures 5.13 for FB and 5.14 for GIFB.

As in the previous case, both $\boldsymbol{Z}^{\star}_{\text{FB}}$ and $\boldsymbol{Z}^{\star}_{\text{GIFB}}$ closely align with the true hidden state path $\boldsymbol{Z}$, indicating that both algorithms retain high accuracy in recovering the latent states. Minor deviations from the true paths are observed for both methods, for example, in $Z_1(t)$ at $t = 3$ in Figures 5.13a and 5.14a. This again demonstrates that the randomness in the emission process may lead to mismatches between the true and recovered hidden states.



(a) Viterbi path $\boldsymbol{Z}_1^\star$ by FB (b) Viterbi path $\boldsymbol{Z}_2^\star$ by FB

Figure 5.13: The Viterbi paths $\boldsymbol{Z}_{\text{FB}}^\star$ (blue dashed) compared with the true hidden states $\boldsymbol{Z}$ (red solid) for the first 100 time points of $Z_1(t)$ and $Z_2(t)$ under the Granger non-causality (5.34).



(a) Viterbi path $\boldsymbol{Z}_1^\star$ by GIFB (b) Viterbi path $\boldsymbol{Z}_2^\star$ by GIFB

Figure 5.14: The Viterbi paths $\boldsymbol{Z}_{\text{GIFB}}^\star$ (blue dashed) compared with the true hidden states $\boldsymbol{Z}$ (red solid) for the first 100 time points of $Z_1(t)$ and $Z_2(t)$ under the Granger non-causality (5.34).

Similarly, two aggregated confusion matrices are also presented, comparing the Viterbi paths obtained from FB and GIFB with the true hidden state trajectories across 200 repeated experiments, as shown in the Figures 5.15 and 5.16, respectively.

As reported at the top of the confusion matrices, both algorithms achieve an overall

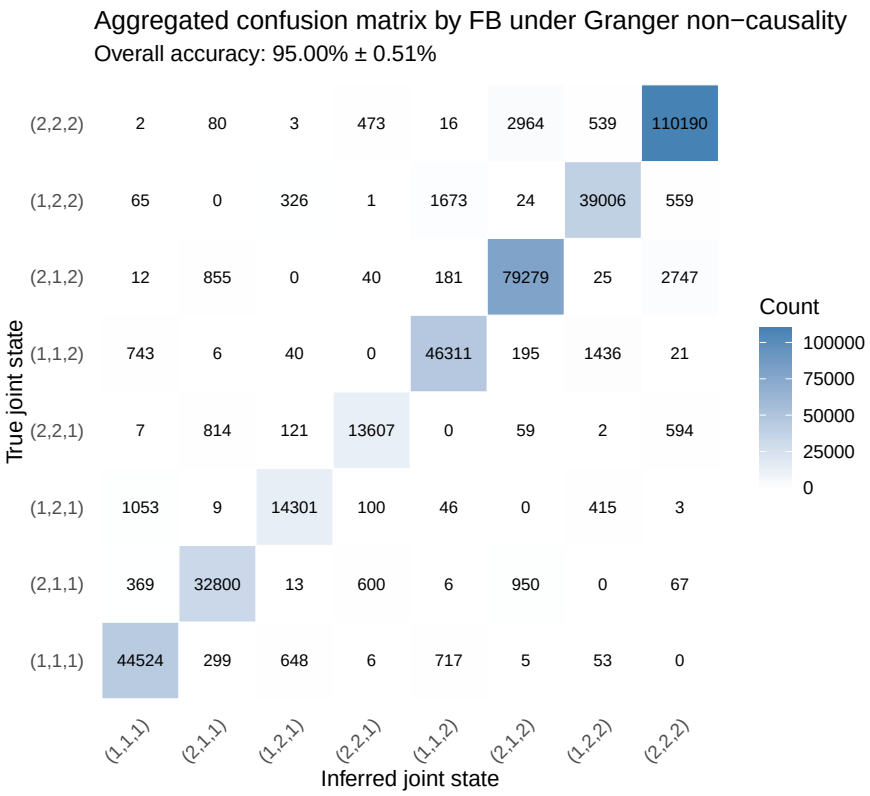


Figure 5.15: Confusion matrix for $\boldsymbol{Z}^{\star}_{\mathrm{FB}}$ under the Granger non-causality (5.34)

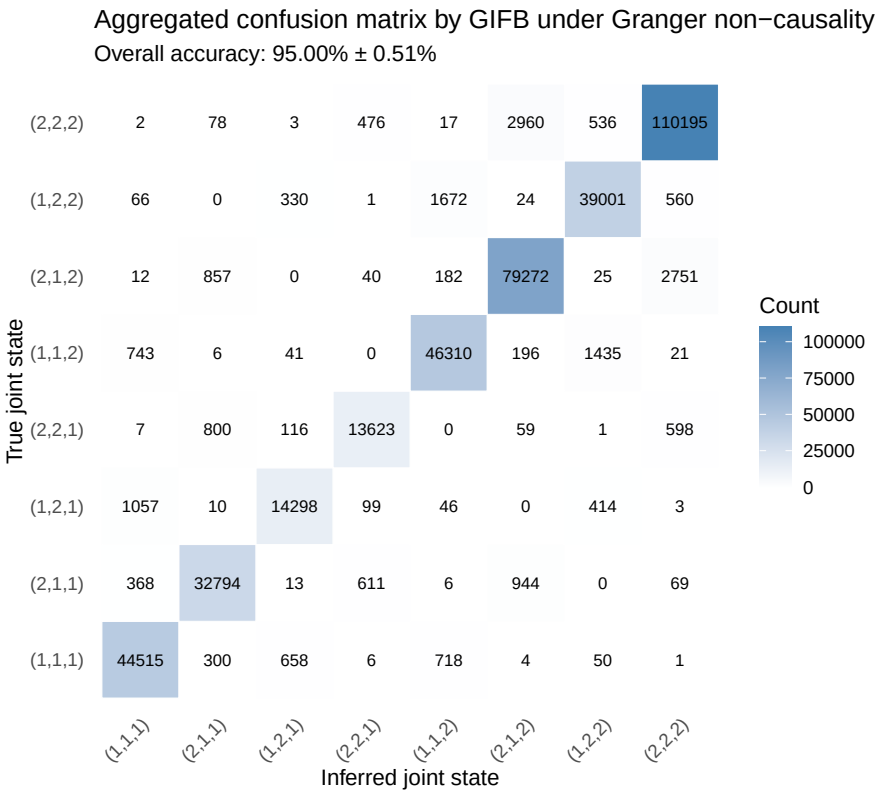


Figure 5.16: Confusion matrix for $\boldsymbol{Z}^{\star}_{\mathrm{GIFB}}$ under the Granger non-causality (5.34)

classification accuracy of 95.00% with a standard deviation of 0.51%. Minor differences among the matrix entries can be observed, which are attributable to Monte Carlo error. These results further confirm that the proposed GIFB preserves the high accuracy of FB and validate that GIFB is an exact Gibbs sampler, updating each chain according to its true full conditional distribution.

5.4.3.3 Computational Cost

Lastly, we examine the computational cost. The main computational cost in the Gibbs sampler (Algorithm 7) arises from the forward–backward recursion used to update the hidden states $\boldsymbol{Z}$ from their full conditional distribution $\pi(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}, \boldsymbol{\beta})$. Table 5.3 reports the average and total computational times of the standard FB algorithm and the proposed GIFB algorithm, averaged across both the contemporaneous independence and Granger non-causality settings, over 2000 MCMC iterations and 200 repeated experiments.

Algorithm	FB	GIFB
Average time per iteration	2.58 s ± 0.21 s	1.13 s ± 0.09 s
Total time (2000 iterations)	1.43 h ± 0.12 h	0.63 h ± 0.05 h

Table 5.3: Computational time for FB and GIFB, with time reported in seconds (s) and hours (h).

In each MCMC iteration, the FB algorithm requires on average 2.58 seconds to jointly update all hidden chains, with a standard deviation of 0.21 seconds. By contrast, GIFB updates the chains sequentially, reducing the average computational cost per iteration to 1.13 seconds with a standard deviation of 0.09 seconds. Over 2000 iterations, GIFB on average completes in 0.63 hours, compared with 1.43 hours for FB, representing a reduction of more than half. Essentially, this reduction reflects the theoretical complexity gap: GIFB scales quadratically with the number of chains, $O(Ts^2k^2)$, whereas FB scales exponentially, $O(Ts^{2k})$, as discussed in Section 5.3.4.1.

5.5 An application to data from Twitter/X on the UK Railways

5.5.1 Context

Having validated our proposed methodology through controlled simulation studies, we now turn to a real-world application in order to further assess its practical utility. In particular, we focus on Twitter data related to the railway system.

Social media are ubiquitous in our lives and affect societies significantly. Among platforms such as Facebook, Instagram, and YouTube, Twitter/X is one of those that offers a social networking service for people to send and read approximately 100-character text messages called "tweets" [Wikipedia, 2022]. Each tweet is constituted by date-time (the time of sending the tweet) and the content (text data). Numerous studies have leveraged this platform for event detection, including earthquakes [Sakaki et al., 2010], terrorist attacks [Imran et al., 2015], and COVID-19 outbreaks [Singh et al., 2020]. Twitter's appeal stems not only from its transparency—most tweets are publicly accessible—but also from the vast amount of information it generates daily, with millions of tweets produced worldwide.

Railway is a widely used mode of travel, with over 1.7 billion passengers reported to have travelled by train in the UK throughout 2024, according to the ORR [2025a]. However, disruptions in the railway system, such as strikes or accidents, are common occurrences. In a report by ORR [2025b], it was stated that approximately 13.7% of trains failed to arrive punctually, and 3.4% of trains were cancelled for various reasons during the first quarter of 2025. Hence, detecting these disruptive events in a timely manner is crucial to minimize the associated losses and inconveniences.

Hence, in this section, we fit an MHMM (see Section 5.2.2) and by utilising the GIFB algorithm (see Section 5.3.4) to detect disruptions in the UK National Railway system using the content and volume of tweets that reference delays and disruptions. The intu-

ition is that users are inclined to express their emotions by tweets, to be specific, they would convey positive sentiment if the railway brings them to their destination on time and express negative sentiment if the trains are delayed. As a result, when a disruption occurs, we expect to observe an additional number of delayed tweets, which naturally serves as the observed process $\boldsymbol{X}$ in our MHMM framework, while the underlying railway's disruption status is represented by the latent chain $\boldsymbol{Z}$. We present two tweets with contrasting attitudes in Table 5.4.

Date / Time	Tweets	Railway Company	Sentiment
Jan 21 2015 17:26:57	It's not quite Hogwarts express but nevertheless. Thanks for the ride	@eastcoastuk	Positive
Jan 21 2015 17:45:30	why are your trains always delayed from Brighton so I miss my connection at Clapham Junction to Milton Keynes!	@SouthernRailUK	Negative

Table 5.4: An example of tweets with positive and negative sentiments

To implement this idea, we employ sentiment analysis [Anber et al., 2016] to classify tweets into two categories: non-delayed tweets with positive sentiment and delayed tweets with negative sentiment. The preprocessing steps leading to this classification are outlined in Appendix C.1. The delayed tweets are then aggregated into counts across the day for every 15 minutes, resulting in a time series data shown in Figure 5.17.

It can be seen that the counts for delayed tweets mainly experience two peaks: one around 9 am and the other around 6 pm. This is intuitive as they correspond to the morning and evening rush hours respectively. This pattern motivates us to utilize a time dependent intensity function $\lambda(t)$, which we will specify in the following Section 5.5.2. Additionally, Figure 5.17 shows that few delayed tweets are recorded between midnight (12 a.m.) and early morning (6 a.m.), reflecting reduced activity. Hence, our modelling will focus on

the rest active part of the day, with time divided into 15-minute intervals.

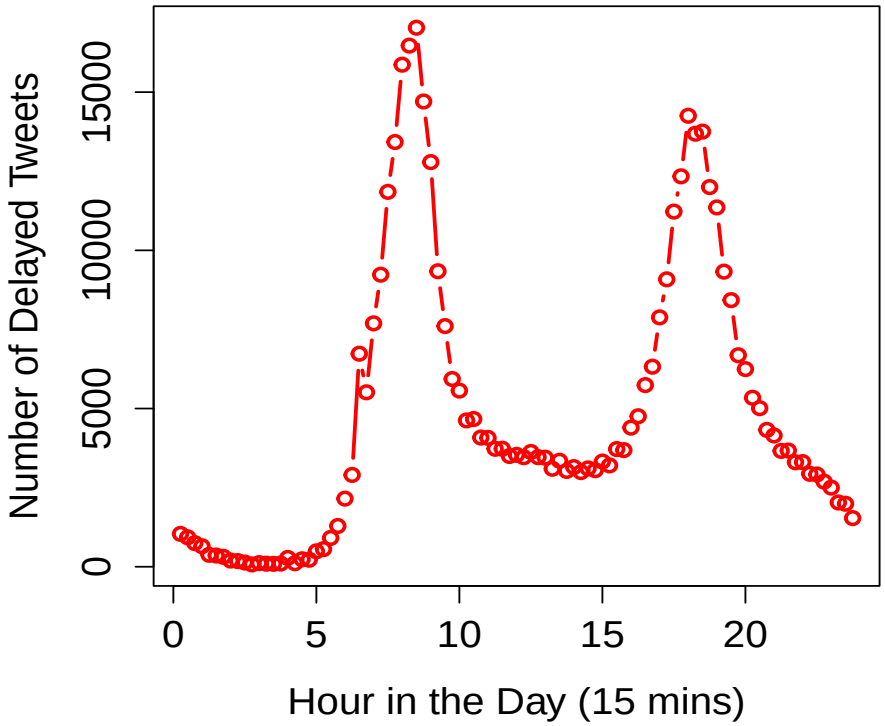


Figure 5.17: Total counts of delayed tweets aggregated across all railway companies and summed over all available days, reported at each 15-minute interval of the day

5.5.2 Model Setup

Motivated by the temporal behaviour of delayed tweets illustrated in Figure 5.17, we assume a time-dependent intensity function $\boldsymbol{\lambda}(t)$. To simplify the setting, we consider two hidden chains, each representing a group of railway companies:

- Chain (group) 1 involves the railway companies to the north of London including Virgin Trains (Avanti West Coast), West Midland Trains, Chiltern Railways, East Midlands Railways, Great Northern Rails, Hull Trains and London North Eastern Railway.

- Chain (group) 2 involves the railway companies to the south of London including Southern Railway, South Western Railway, Great Wester Railway, South Eastern Railway and c2c Railway.

For each group $k = 1, 2$, the delayed counts $X_k(t)$ are obtained by aggregating the delayed tweet counts across all railway companies belonging to that group. The corresponding hidden state $Z_k(t) \in \mathcal{S} = \{1, 2\}$ is then defined as follows:

- $Z_k(t) = 1$: no major disruption in railway group k at time t.
- $Z_k(t) = 2$: a major disruption in railway group k at time t.

In line with the emission setting introduced in Section 5.2.1, we assume that when $Z_k(t) = 2$, an additional number of delayed tweets following a Poisson distribution with fixed rate c (shifting constant) is observed. This allows us to expand the observed counts $X(t)$ as Equation (5.1).

Assume that the counts for delayed tweets experience morning and evening peaks according to Figure 5.17, we model the time-dependent intensity $\lambda_k(t)$ for each chain $k = 1, 2$ by a bimodal function of the form

$$\lambda_k(t) = c_{k,1} \cdot f_k(t) + c_{k,2} \cdot g_k(t) + c_{k,3} \tag{5.35}$$

where $f_k(t)$ and $g_k(t)$ represent the morning and evening rush-hour patterns, with coefficients $c_{k,1}$ and $c_{k,2}$ capturing their magnitudes, while $c_{k,3}$ accounts for normal level of the traffic flow. We further assume that $f_k(t)$ and $g_k(t)$ as

$$f_k(t) = \varphi_{(\mu_{k,1}, \sigma_{k,1})}(t)$$
$$g_k(t) = \varphi_{(\mu_{k,2}, \sigma_{k,2})}(t)$$

Where $\varphi_{(\mu_{k,1}, \sigma_{k,1})}(t)$ and $\varphi_{(\mu_{k,2}, \sigma_{k,2})}(t)$ refers to the probability density function of Normal with mean $\mu_{k,1}, \mu_{k,2}$ and standard deviation $\sigma_{k,1}, \sigma_{k,2}$ respectively.

For each $\lambda_k(t)$, the primary parameters to infer are:

- $\mu_{k,1}$ and $\mu_{k,1}$ corresponding to the peak time for morning and evening rush hours

- Rate c implying how many extra counts are expected to observe when disruption happens
- Coefficients $c_{k,1}, c_{k,2}$ of the morning and evening rush hours as well as the constant $c_{k,3}$ for the baseline flow.

Throughout this work, we assume that $\boldsymbol{\sigma} = (\sigma_1, \sigma_2)$ to be known through the optimizer.

In summary, given the observed counts $\boldsymbol{X}$ and the distributional form of $\lambda_k(t)$ in Equation 5.35, our objective is to employ the Gibbs sampler with GIFB (Algorithm 7) to infer:

- the latent Markov chain $\boldsymbol{Z}$ that model state of the railway;
- the regression coefficients $\boldsymbol{\beta}$ that captures the potential dependency structures;
- the emission parameters $\boldsymbol{\theta}$ that parametrise the intensity function $\boldsymbol{\lambda}$.

5.5.3 Results

For the Twitter data analysis, we run the MCMC for 5000 iterations (which took approximately 45 minutes) and present the results from two main perspectives:

- Inferred regression coefficients and the corresponding dependency structures, assessed through posterior contour probabilities (see Section A.2.2).
- The recovered Viterbi path $\boldsymbol{Z}^\star$, validated against the reported news to confirm the detection of disruptions.

The uninformative priors have been used throughout this section and their full specifications are presented in Appendix C.1.1 and C.1.2. For completeness, additional results such as posterior medians and MCMC diagnostic trace plots are given in Appendix C.1.4.

5.5.3.1 Regression Coefficients

In this setting with two chains and a binary state space, the model involves three groups of regression coefficients $\boldsymbol{\beta}_j$, corresponding to the chain groups 1, 2, and $1, 2$. These are illustrated explicitly in Figures 5.18, 5.19, and 5.20.

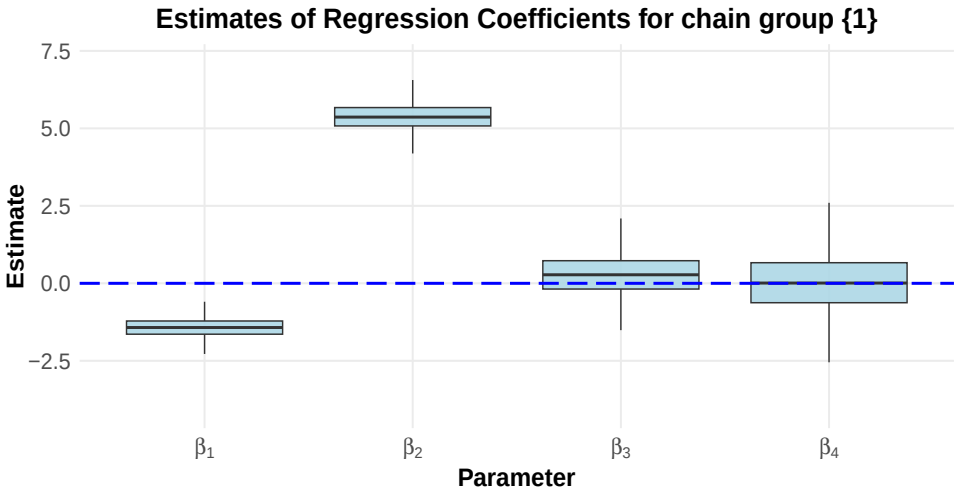


Figure 5.18: Estimated Regression Coefficients for $g_j = \{1\}$

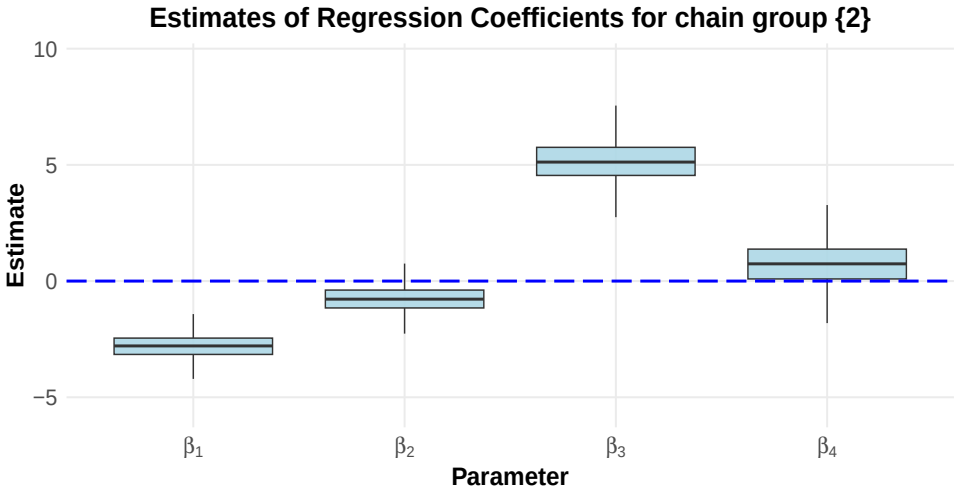


Figure 5.19: Estimated Regression Coefficients for $g_j = \{2\}$

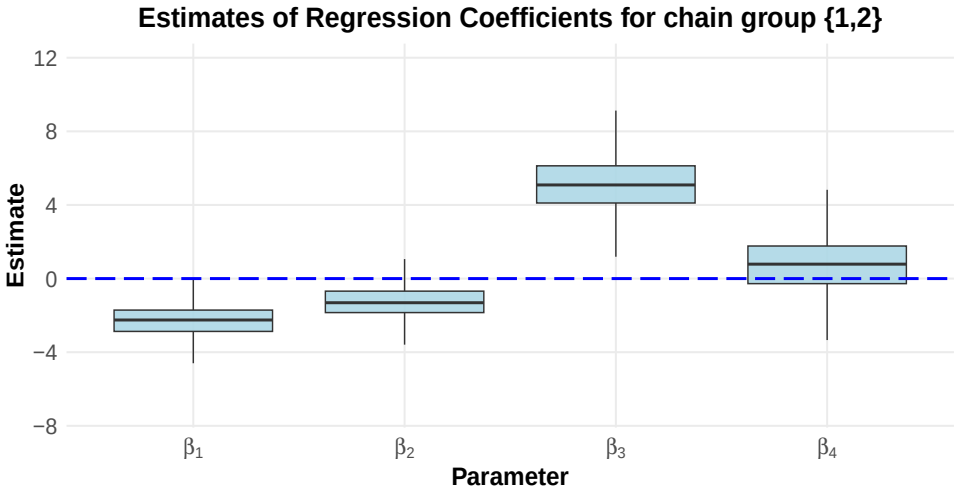


Figure 5.20: Estimated Regression Coefficients for $g_j = \{1, 2\}$

- From Figure 5.18, it can be seen that β_3 and β_4, which correspond to the effect of chain 2 on chain 1, are very close to zero. According to Proposition 4.13, this suggests the absence of a causal relationship from chain 2 to chain 1.
- Similarly, Figure 5.19 shows that β_2 and β_4, which correspond to the effect of chain 1 on chain 2, are near zero. This implies the possible absence of a causal relationship from chain 1 to chain 2.
- Finally, Figure 5.20 demonstrates that $\boldsymbol{\beta}_{1,2}$ is very similar to $\boldsymbol{\beta}_2$ in Figure 5.19. Based on Proposition 4.13, this indicates a potential contemporaneous independence between chains 1 and 2.

Recall that in Section 4.5.1, we formally assessed the dependency structures by computing posterior contour probabilities for the following hypotheses with $k = 1, 2$:

H_0 : Z_k and $Z_{k'}$ are contemporaneously independent or Z_k is not Granger caused by $Z_{k'}$

H_1 : The chains are dependent on each other

This allowed us to evaluate each potential dependency among the chains under the Hierarchical Marginal Model. The results are summarized in Figure 5.21.

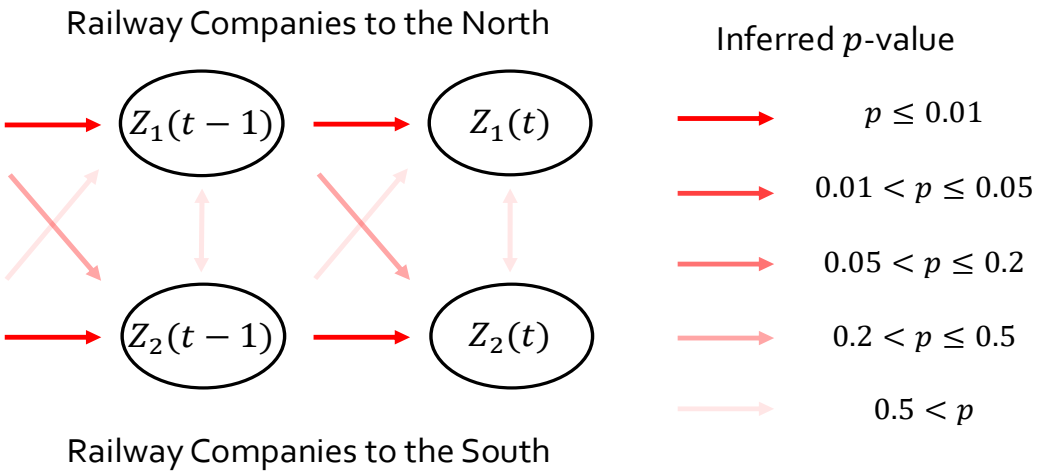


Figure 5.21: The inferred dependency structures among the railway companies via the p-values from Bayesian posterior contour probabilities.

Recall from Section 4.5.1 that lower p-values indicate stronger dependencies between chains, represented as red arrows with low transparency, whereas higher p-values indicate weaker or absent dependencies, shown as red arrows with high transparency. Consistent with the regression coefficient patterns in Figures 5.18, 5.19, and 5.20, we observe a large p-value ($p > 0.5$) for both the contemporaneous independence between chain 1 and chain 2 and the Granger causality from Z_2 to Z_1. In addition, the Granger causality from Z_1 to Z_2 yields a moderately large p-value ($0.2 < p \leq 0.5$). Together, these results indicate that the overall dependency between Z_1 and Z_2 is relatively weak.

Intuitively, in the context of the UK railway system, operations are highly privatized and organized through separate railway franchises, where each company typically controls specific routes and holds exclusive rights to transport passengers on them [Wikipedia, 2023]. This structure naturally reduces interactions among different railway companies.

5.5.3.2 Inferring the unobserved state of the railway

In this section, we present the estimates for the hidden states $\boldsymbol{Z}$. Similarly to Section 5.4.3.2, we apply the Viterbi algorithm reviewed in Section 5.3.7 based on the posterior medians of the emission parameter $\hat{\boldsymbol{\theta}}$ as well as the regression coefficients $\hat{\boldsymbol{\beta}}$ to obtain a single optimal hidden path $\boldsymbol{Z}^\star$.

In contrast to the simulated setting, we do not have the ground truth hidden path $\boldsymbol{Z}$ for direct comparison in the real data. Instead, we validate our inferred results by examining whether the estimated Viterbi path $\boldsymbol{Z}^\star$ aligns with known events reported in the news.

As an example, we focus on 17 February 2015. On this day, the following disruption was reported by GWR [2015]:

"Due to a person being hit by a train between Bristol Temple Meads and Bath Spa, services through these stations may be delayed by up to 60 minutes." (8:58 PM, 17 February)

Figure 5.22 displays the observed counts of delayed tweets for the railway group serving the south of London as blue points (chain 2 introduced previously in Section 5.5.2). The inferred Viterbi path $Z_2^\star$ is shown in green, together with the corresponding intensity function $\hat{\lambda}_2(t)$ constructed by the posterior median estimates in red. As in previous Figures 5.4 and 5.5, the Viterbi path $Z_2^\star$ is scaled by a factor of 20 for visualization purposes. Notably, the disruption reported in the news corresponds closely to the last event detected at approximately 9:30 PM, as indicated by the orange arrow.

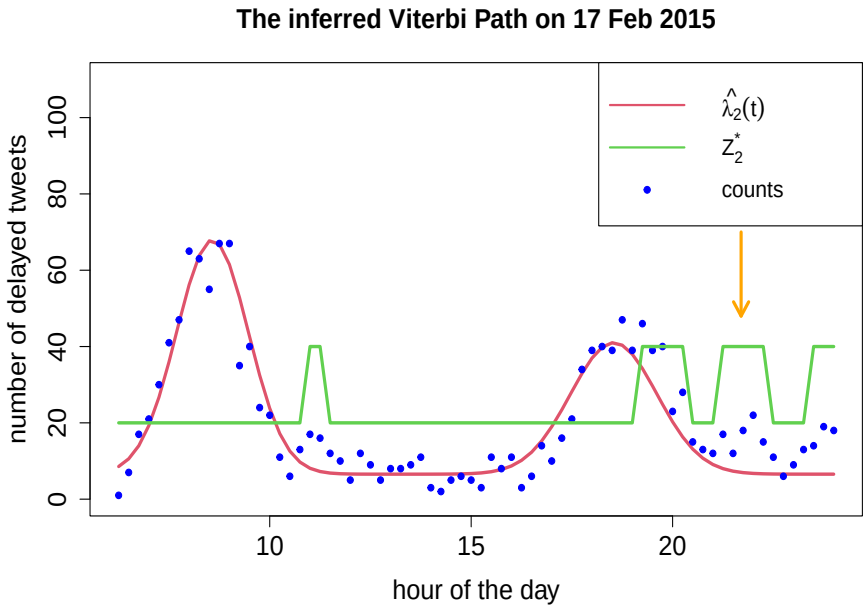


Figure 5.22: The overall picture of the inferred Viterbi path on 17 February 2015. The blue dots represent the observed counts of delayed tweets within each discrete time interval. The red line shows the inferred intensity function $\hat{\lambda}_2(t)$ for chain group 2 (railway groups serving the south of London), constructed from the posterior median estimates. The green line denotes the inferred Viterbi path. The orange arrow highlights the final event detected by the Viterbi path, which corresponds to the disruption reported by GWR in the news.

5.5.4 Comments

In this section, we demonstrated using Twitter data that the proposed model when fitted using the proposed GIFB can assess for different dependency structures of the unobserved MMC Z between railway companies in the north and the south of London. Consistent with the privatization of the UK railway system, where companies typically control distinct routes with limited overlap, our model suggested independence between the two

railway groups, as illustrated in Figure 5.21. Furthermore, the model successfully detected disruptions in line with reported news, for example, the event on 17 February 2015 documented by GWR [2015] as presented in Figure 5.22.

It is important to note, however, that our results are sensitive to the preprocessing of delayed tweets described in Appendix C.1. Alternative preprocessing schemes would yield different observed data $\boldsymbol{X}$, potentially altering both the inferred dependency structures and the recovered hidden path $\boldsymbol{Z}$.

Finally, our analysis considers a simplified case with two chains, representing railway companies to the north and south of London, and two states. With more extensive data that include additional railway companies, the framework could naturally be extended to more complex MHMMs with more chains, allowing the investigation of more intricate dependency structures.

5.6 Discussion

In Chapters 2, 3, and 4, we developed an effective reparametrisation framework to detect and infer dependency structures in MMCs under the assumption that the states are directly observable. In more realistic settings, however, the underlying MMC is latent, and only indirect observations generated through an emission process are available. This motivates a shift in focus in the present chapter to MHMMs, where our primary objective remains the inference of dependency structures, but with the additional challenge of recovering the hidden chains.

We began in Section 5.2 by briefly introducing the structure of a MHMM, illustrated in Figure 5.2. An MHMM consists of a latent process $\boldsymbol{Z}$ that follows the first-order Markov property, as studied in the previous chapters, a set of observed data $\boldsymbol{X}$, and an emission function f that links the hidden process to the observations. To simplify the setting, Assumption 5.2 specifies that the emission function is known up to its distributional form,

while Assumption 5.3 imposes a one-to-one correspondence between hidden states and emission distributions, which allows us to focus on the potential dependency structures embedded in the hidden process, rather than on the complexities of the emission mechanism.

For the inference introduced in Section 5.3, we began by expanding the likelihood according to the structure of the MHMM. A key new challenge that arises in this setting is the inference of the hidden process $\boldsymbol{Z}$. A naive and straightforward approach is to apply off-the-shelf data augmentation, where each latent state is sampled iteratively. However, this requires marginalizing over an exponentially growing number of possible paths to obtain $\pi\big(\boldsymbol{Z}(t) \mid \boldsymbol{X}, \boldsymbol{\theta}\big)$, which quickly becomes computationally intractable.

As introduced in Section 5.3.2, the Forward Backward (FB) algorithm [Baum et al., 1970, 1972] act as an elegant technique to infer this the marginal by recursively uses the earlier computations. However, as discussed in Section 5.3.2.4, its application in the multivariate setting is computationally prohibitive. Specifically, incorporating the standard FB algorithm requires transforming an MHMM into an equivalent univariate HMM by aggregating all possible joint states into an extended state space via the Cartesian product (see Section 2.8). As a result, both the forward and backward recursions need to sum over this exponentially growing state space, which contributes to an infeasible computational cost.

To mitigate this issue, Touloupou [2016] proposed the *individual Forward Filtering Backward Sampling* (iFFBS) Algorithm, which behaves as a Gibbs sampler by updating each individual chain iteratively rather than updating all the chains jointly as in FB algorithm. This modification reduces the computational burden substantially, replacing the originally exponentially dependence to the number of chains in FB algorithm, into a quadratic dependence, as shown in Section 5.3.3.1. However, motivated by the epidemic setting, this algorithm is built under the conditional independence assumption, where all

the concurrent interactions among the chains are ignored. Without this assumption, the forward filtering recursion in Equation 5.15 is no longer applicable. Consequently, iFFBS fails to recover hidden paths in the general MHMM setting. Touloupou et al. [2020] also proposed an extension of iFFBS that incorporates a Metropolis–Hastings step, based on the proposal distribution under Assumption 5.17. However, in the absence of conditional independence, this proposal tends to suffer from a high rejection rate, which in turn leads to slow mixing and poor convergence of the Markov chain.

As a result, we proposed our novel *Generalised Individual Forward Backward* (GIFB) algorithm in Section 5.3.4. Similarly to Touloupou [2016], Touloupou et al. [2020], GIFB follows the Gibbs–like strategy of updating groups of chains iteratively and achieves a substantially lower computational cost, scaling quadratically in the number of states within each chain group (see Section 5.3.4.1). In contrast to iFFBS, GIFB expands the likelihood forwardly in accordance with the FB algorithm, which allows the exact computation of the full conditionals for each group of chains. This makes GIFB an exact Gibbs sampler, enabling the accurate recovery of hidden paths and fully capturing all dependency structures in general MHMMs without conditional independence. Additionally, unlike iFFBS, which updates chains element-wise, GIFB allows a combination of group-wise and element-wise updates, providing the flexibility to trade-off between computational cost and convergence speed. Lastly, GIFB operates on the marginal transition probabilities, rather than the joint transition probabilities as in FB. This setting naturally integrates with the previously introduced Hierarchical Marginal Model (see Section 4.2) and enables the efficient Gibbs sampler described in Algorithm 7, which samples all parameters of the MHMM from their exact full conditionals.

In Section 5.4, we elaborated the advantages of the GIFB algorithm through a simulation study against the standard FB algorithm. The box plots of the regression coefficients $\boldsymbol{\beta}$ in Figures 5.6 and 5.8 together with the posterior contour probabilities confirmed that GIFB successfully identified both the contemporaneous independence and the Granger

non-causality imposed in the data. The confusion matrices in Figures 5.12 and 5.16 further showed that GIFB accurately recovers the hidden chains, thereby verifying its exactness when compared with FB. At the same time, the computational cost in Table 5.3 highlighted a substantial reduction in computational time for GIFB comparing to FB, reflecting its improved scalability.

Finally, we applied the GIFB algorithm by fitting a MHMM to the real data on Twitter/X data to detect disruption in the UK's railway. To simplify the setting, we grouped the railway companies into two categories according to the regions and routes they operate, to the north and the south of London. The model indicated near-independence between these two groups, which can be explained by the privatized nature of the UK railway system, where different companies typically operate on disjoint routes. In addition, several disruptions were detected, and their validity was confirmed by comparison with reported news events.

5.6.1 Limitations and Future Work

One major limitation is that, for the emission function specified in Section 5.2.1, we adopt a parametric Poisson form as specified in Assumption 5.2. While this choice simplifies the inference procedure, it restricts the flexibility of the model. In particular, the Poisson assumption enforces equality between the mean and the variance, which may not hold in many applications where overdispersion is common. A natural extension for future work is to consider more flexible emission models, such as the negative binomial distribution, which can accommodate overdispersion in count data, or even non-parametric approaches such as Gaussian Processes, which could capture more complex and temporal dynamics in the emission process.

Another limitation arises from Assumption 5.3, which enforces a one-to-one emission process and reduces the model flexibility. A possible extension would be to apply the proposed GIFB algorithm to more complex emission structures studied in the literature.

Furthermore, unlike iFFBS, which can accommodate semi-Markov processes, the current GIFB is designed only for the standard Markov setting, limiting its flexibility. Extending GIFB to handle semi-Markov dynamics is therefore a natural direction for the future work.

Additionally, the simulation study in Section 5.4 focuses on the idealized setting in which the observed process $\boldsymbol{X}$ is fully observed. In practice, however, missing data are common, and $\boldsymbol{X}$ may only be partially observed. As future work, we plan to extend our experiments to settings with missing observations in order to assess the robustness of the proposed model and inference procedures. Furthermore, we will investigate scenarios with limited data availability, where fewer time points are observed (i.e., smaller T), to evaluate the performance of the methods under short time series. Finally, additional simulations with smaller differences between intensity levels (i.e., smaller values of c in Table 5.2) will be conducted. Such settings correspond to noisier emission processes and will allow us to further examine the robustness of our inference procedures under weaker signal conditions.

For the Twitter data analysis, the inferred results depend heavily on the observed counts of the delayed tweets, which are in turn determined by the performance of the classification algorithm. Employing more accurate and efficient sentiment classification methods would likely yield more reliable recovery of dependency structures. Another natural extension is to allow for a time inhomogeneous transition matrix $P(t)$, or equivalently, time inhomogeneous regression coefficients $\boldsymbol{\beta}(t)$, since it is reasonable to expect higher disruption probabilities during rush hours compared to off-peak periods (see, for instance, Damásio and Nicolau [2024]). We can start with the simple case where P and $\boldsymbol{\beta}$ are only allowed to switch between several states, for instance one for the peak and the other for off-peak under this setting. Finally, with access to larger and more recent datasets, the MHMM could be extended to include more groups of railway companies, enabling a richer analysis of the intricate dependency structures among them.

Chapter 6

Summary and Reflections

6.1 Summary for the Thesis

The main objectives achieved in this thesis were as follows:

(a) propose new parametrisations that explicitly capture the dependence structure among the components of $\boldsymbol{Z}$, with parameters interpretable as measures of correlation between individual chains.

(b) leverage these parametrisations to design efficient computational algorithms for inferring $\boldsymbol{Z}$ in high-dimensional settings where the chains are not directly observable, upon the existing work on Touloupou et al. [2020].

In Chapter 2, we reviewed the dependency structures of *Conditional Independence*, *Contemporaneous Independence*, and *Granger non-causality* from a novel geometric perspective. Through this visualization, we interpreted how these abstract independence constraints act on the joint transition matrix P via graphical representations. In particular, we showed that Conditional and Contemporaneous Independence, which describe concurrent interactions, correspond to projections within each row of P onto lower-dimensional spaces, while Granger non-causality, which reflects effects from past states, corresponds to projections across different rows onto the same lower-dimensional space. Since the marginal transition probabilities naturally emerge as the key quantities in these interpretations, we were motivated to use them directly to capture dependency structures.

In Chapter 3, we introduced our first novel reparametrisation, the *Univariate Marginal Model*. This model was built from the univariate marginal transition probabilities and connects them to the joint transition probabilities through a product with auxiliary probabilities specified under a permutation σ. To enhance interpretability, we further decomposed these marginal and auxiliary probabilities using multinomial logistic regressions, yielding the regression coefficients. The mapping from the joint transition probabilities to the regression coefficients was shown to be a diffeomorphism, ensuring that the coefficients are able to capture all dependency structures encoded in the joint transition matrix. In particular, dependency structures can be imposed directly by setting the corresponding regression coefficients to zero. Moreover, we demonstrated that this framework serves as a convenient and efficient generative model, capable of producing joint transition matrices and hence MMCs directly from the regression coefficients for the desired dependencies.

In Chapter 4, we introduced our second novel reparametrisation, the *Hierarchical Marginal Model*. Like the Univariate Marginal Model, it was built from the univariate marginal transition probabilities, but connected it to the joint transition probabilities via a set of hierarchically constructed higher-order marginals. As before, we decomposed these probabilities using multinomial logistic regressions, yielding interpretable regression coefficients. The mapping from the joint transition probabilities to the regression coefficients was shown to be a diffeomorphism under *strong compatibility*, ensuring that the coefficients capture all dependency structures encoded in the joint transition matrix. Dependency structures can again be imposed directly through the regression coefficients: Granger non-causality is enforced by setting the corresponding coefficients to zero, while contemporaneous independence is imposed by equating the corresponding regression vectors. Finally, we demonstrated that the Hierarchical Marginal Model can also efficiently generate MMCs with specified dependency structures through a hierarchical generating scheme.

Lastly, in Chapter 5, we shifted our focus to the hidden setting, where the Markov pro-

cess is unobserved and only the emitted data are available. To efficiently infer this hidden process, we proposed a novel Gibbs-like algorithm, the *Generalised Individual Forward Backward* (GIFB). Unlike the direct extension of the standard FB algorithm, which jointly updates all chains and incurs exponential computational cost in the number of chains, our GIFB updates groups of chains iteratively, reducing the cost to the quadratic order. By expanding the likelihood forwardly, we avoided the conditional independence assumption required by the iFFBS of Touloupou et al. [2020], allowing GIFB to operate under general MHMMs. Moreover, since GIFB is formulated in terms of marginal transition probabilities, which are modelled within the Hierarchical Marginal Model, we naturally incorporated them into a Gibbs sampler (Algorithm 7) for the full inference of the parameters within MHMMs. Finally, we applied this algorithm to real Twitter/X data in the context of UK railways to detect dependency structures and disruptive events, demonstrating its effectiveness in practice.

6.2 Limitations and Future Work

The major limitation of the Univariate Marginal Model, as discussed in Section 3.6.1, lies in its reliance on a predefined permutation. Under each permutation, only a subset of the dependency structures is explicitly modelled, while the remainder can only be obtained implicitly through additional manipulations. This limitation is manageable if prior information on the dependency structures is available; otherwise, multiple models under different permutations need to be fitted to capture the full range of dependencies explicitly.

For the Hierarchical Marginal Model, as discussed in Section 4.6.1, the main limitation arises from the inherently assumed *strong compatibility*, which ensures that the regression coefficients can be mapped back to a valid joint transition matrix. This assumption introduces the additional step of computing the feasible region induced by strong compatibility when generating data, thereby increasing the computational burden, especially as the dimensionality grows.

As another direction for future work, additional simulation studies could be conducted to further assess the consequences of mis-specifying dependency structures. For example, in Chapters 3 and 4, data could be generated under the full dependence model while inference is performed using models that impose independence assumptions, allowing the resulting bias in parameter estimates to be systematically evaluated. Similarly, in Chapter 5, a complementary simulation study could consider data generated from a fully dependent hidden process but analysed under models that assume certain independence structure, thereby quantifying the impact of incorrect independence assumptions on state and hence dependency inference. Such simulations would further strengthen the motivation for the proposed framework.

Two natural extensions of the current GIFB framework arise. First, unlike iFFBS, which can accommodate semi-Markov processes, GIFB is restricted to the standard Markov setting; extending it to semi-Markov dynamics would enhance its flexibility. Second, since GIFB relies on the first-order Markov property when expanding the likelihood, it is not directly applicable to higher-order chains. Developing a formulation that relaxes this restriction would allow GIFB to handle more general dependency structures.

Then, several extensions are possible for the application to the Twitter data. As discussed in Section 5.6, the inference results are highly sensitive to the observed tweet counts, which in turn depend on the classification algorithm used to identify delayed-service tweets. Improving the classification accuracy would therefore lead to more reliable inference. Moreover, the current model employs a parametric Poisson emission function for the intensity process, motivated by the exploratory data analysis. Future work could consider more flexible, non-parametric emission models, such as Gaussian processes, to capture non-periodic or more complex temporal patterns. In addition, the transition dynamics could be extended to allow time inhomogeneity, either through a time-varying transition matrix $P(t)$ or, equivalently, time-varying regression coefficients $\boldsymbol{\beta}(t)$. This extension is particularly relevant in railway settings, where disruption probabilities are

expected to vary across time, for example between rush hours and off-peak periods. Lastly, due to data limitations, the railway companies were aggregated into only two groups (see Section 5.5.2). With access to richer and more recent datasets, the proposed framework could be extended to MHMMs with a larger number of chains, allowing for the exploration of more complex dependency structures among multiple railway operators.

Lastly, all models and methods developed in this thesis are formulated in discrete time. An important direction for future research is to extend the proposed framework to continuous-time settings. Such an extension would allow the model to more accurately reflect the underlying dynamics of real-world processes, particularly in applications where events occur continuously rather than at fixed time intervals. Developing continuous-time MHMMs with efficient inference algorithms analogous to GIFB would therefore further broaden the applicability of the proposed methodology.

Bibliography

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182–198, 2009.

Subil Abraham and Suku Nair. Cyber security analytics: a stochastic model for security quantification using absorbing markov chains. *Journal of Communications*, 9(12):899–907, 2014.

Hana Anber, Akram Salah, and AA Abd El-Aziz. A literature review on twitter data analysis. *International Journal of Computer and Electrical Engineering*, 8(3):241, 2016.

Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204, 2010.

Leonard E Baum and Ted Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

Leonard E Baum et al. An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, 3(1):1–8, 1972.

Tim Bedford and Roger M. Cooke. Vines—a new graphical model for dependent random variables. *The Annals of Statistics*, 30(4):1031–1068, 2002.

Yoshua Bengio and Paolo Frasconi. Input-output hmms for sequence processing. *IEEE Transactions on Neural Networks*, 7(5):1231–1249, 1996.

Wicher P Bergsma and Tamas Rudas. Marginal models for categorical data. *The Annals of Statistics*, 30(1):140–159, 2002.

Mauro Bernardi, Antonello Maruotti, and Lea Petrella. Multivariate markov-switching models and tail risk interdependence. *arXiv preprint arXiv:1312.6407*, 2013.

José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009. See Section 3.2.5, pp. 133–136 Some Particular Multivariate Distributions.

Julian Besag, Peter Green, David Higdon, and Kerrie Mengersen. Bayesian computation and stochastic systems. *Statistical science*, pages 3–41, 1995.

George E. P. Box and George C. Tiao. Bayesian inference in statistical analysis / george e.p. box and george c. tiao., 1973.

M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 994–999, 1997. doi: 10.1109/CVPR.1997.609450.

Olivier Cappé, Eric Moulines, and Tobias Rydén. *Elements of Markov Chain Theory*, pages 513–564. Springer New York, New York, NY, 2005. ISBN 978-0-387-28982-3. doi: 10.1007/0-387-28982-8_14. URL https://doi.org/10.1007/0-387-28982-8_14.

Chris K Carter and Robert Kohn. On gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 1994.

Siddhartha Chib. Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, 75(1):79–97, 1996.

Wai-Ki Ching and Michael K Ng. Markov chains. *Models, algorithms and applications*, 650, 2006.

Wai Ki Ching, Eric S Fung, and Michael K Ng. Higher-order markov chain models for categorical data sequences. *Naval Research Logistics (NRL)*, 51(4):557–574, 2004.

Roberto Colombi and Sabrina Giordano. Graphical models for multivariate markov chains. *Journal of Multivariate Analysis*, 107:90–103, 2012.

Roberto Colombi, Sabrina Giordano, et al. Marginal models for multivariate hidden markov processes. In *Advances in Latent Variables-Methods, Models and Applications*, 2013.

Roberto Colombi, Sabrina Giordano, and Maria Kateri. Hidden markov models for longitudinal rating data with dynamic response styles. *Statistical Methods & Applications*, 33(1):1–36, 2024.

Bruno Damásio and João Nicolau. Time inhomogeneous multivariate markov chains: Detecting and testing multiple structural breaks occurring at unknown dates. *Chaos, Solitons & Fractals*, 180:114478, 2024.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Stacy L. DeRuiter, Roland Langrock, Tomas Skirbutas, Jeremy A. Goldbogen, John Calambokidis, Ari S. Friedlaender, and Brandon L. Southall. A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, 11(1):362 – 392, 2017. doi: 10.1214/16-AOAS1008. URL `https://doi.org/10.1214/16-AOAS1008`.

Peter Diggle. *Analysis of longitudinal data*. Oxford university press, 2002.

Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

Paul Fearnhead and Chris Sherlock. An exact gibbs sampler for the markov-modulated

poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):767–784, 2006.

Shai Fine, Yoram Singer, and Naftali Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

Wolfgang Fischer and Kathleen Meier-Hellstern. The markov-modulated poisson process (mmpp) cookbook. *Performance evaluation*, 18(2):149–171, 1993.

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368, 1922.

Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press, 1925.

Nicholas J Foti, Jason Xu, Dillon Laird, and Emily B Fox. Stochastic variational inference for hidden markov models. *Advances in neural information processing systems*, 27, 2014.

Andrew Gelman, John B Carlin, Hal S Stern, and Donald B Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 1995. Section 3.4: Multinomial model for categorical data.

Christian Genest and Jana Nešlehová. A primer on copulas for count data. *Astin Bulletin*, 37(2):475–515, 2007.

Zoubin Ghahramani and Michael Jordan. Factorial hidden markov models. *Advances in neural information processing systems*, 8, 1995.

Shameek Ghosh, Jinyan Li, Longbing Cao, and Kotagiri Ramamohanarao. Septic shock prediction for icu patients via coupled hmm walking on sequential contrast patterns. *Journal of Biomedical Informatics*, 66:19–31, 2017. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2016.12.010. URL `https://www.sciencedirect.com/science/article/pii/S1532046416301848`.

Garique FV Glonek and Peter McCullagh. Multivariate logistic models. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):533–546, 1995.

C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969a. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912791.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969b.

GWR, 2015. URL https://twitter.com/GWRHelp/status/567790218049503232. [Online; accessed 1-Nov-2022].

James D Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the econometric society*, pages 357–384, 1989.

Leonhard Held. Simultaneous posterior probability statements from monte carlo output. *Journal of Computational and Graphical Statistics*, 13(1):20–35, 2004. doi: 10.1198/1061860043083. URL https://doi.org/10.1198/1061860043083.

Leonhard Held and Chris C. Holmes. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1(1):145 – 168, 2006. doi: 10.1214/06-BA105. URL https://doi.org/10.1214/06-BA105.

Harold Hotelling. The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378, 1931.

Ronald A Howard. *Dynamic Probabilistic Systems, Volume I: Markov Models*, volume 1. Courier Corporation, 2012.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. Processing social media messages in mass emergency: A survey. *ACM computing surveys (CSUR)*, 47 (4):1–38, 2015.

Harry Joe. *Dependence Modeling with Copulas*. Chapman and Hall/CRC, 2014.

Theodore Kypraios and Philip D O'Neill. Bayesian nonparametrics for stochastic epidemic models. *Statistical Science*, 33(1):44–56, 2018.

E. L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2 edition, 1998. See Section 1.8, pp. 58–61 for Delta Method and Section 6.5, pp. 461–468 for Multivaiate MLE Ergodic theorem.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005. See Chapter 3, Section 3.3, p-values.

Monia Lupparelli, Giovanni M Marchetti, and Wicher P Bergsma. Parameterizations and fitting of bi-directed graph models to categorical data. *Scandinavian journal of Statistics*, 36(3):559–576, 2009.

Andrey A. Markov. Extension of the law of large numbers to dependent quantities. *Izvestiya Fiziko-Matematicheskogo Obschestva pri Kazanskom Universitete*, 15:135–156, 1906.

James S Martin, Ajay Jasra, Sumeetpal S Singh, Nick Whiteley, Pierre Del Moral, and Emma McCoy. Approximate bayesian computation for smoothing. *Stochastic Analysis and Applications*, 32(3):397–420, 2014.

Peter McCullagh and John A Nelder. *Multivariate Regression Models*, pages 219–225. Springer, 1989.

Pradeep Natarajan and Ramakant Nevatia. Coupled hidden semi markov models for activity recognition. In *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*, pages 10–10, 2007. doi: 10.1109/WMVC.2007.12.

Ara V Nefian, Luhong Liang, Xiaobo Pi, Xiaoxing Liu, and Kevin Murphy. Dynamic bayesian networks for audio-visual speech recognition. *EURASIP Journal on Advances in Signal Processing*, 2002(11):783042, 2002.

Joao Nicolau. A new model for multivariate markov chains. *Scandinavian Journal of Statistics*, 41(4):1124–1135, 2014.

J. R. Norris. *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1997. See Chapter 1, Section 1.5, pp. 52–56 Ergodic theorem.

Nuria Oliver, Ashutosh Garg, and Eric Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding*, 96(2):163–180, 2004.

ORR. Passenger rail usage. `https://dataportal.orr.gov.uk/statistics/usage/passenger-rail-usage/`, 2025a. Accessed: 2025-09-25.

ORR. Passenger rail performance. `https://dataportal.orr.gov.uk/statistics/performance/passenger-rail-performance`, 2025b. Accessed: 2025-09-25.

Jennifer Pohle, Roland Langrock, Mihaela van der Schaar, Ruth King, and Frants Havmand Jensen. A primer on coupled state-switching models for multiple interacting time series. *Statistical Modelling*, 21(3):264–285, 2021.

Nicholas G Polson, James G Scott, and Jesse Windle. Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349, 2013.

L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989. doi: 10.1109/5.18626.

Adrian E Raftery. A model for high-order markov chains. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 47(3):528–539, 1985.

Vasanthan Raghavan, Greg ver Steeg, Aram Galstyan, and Alexander G. Tartakovsky. Coupled hidden markov models for user activity in social networks. In *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6, 2013. doi: 10.1109/ICMEW.2013.6618397.

Thomas Richardson. Markov properties for acyclic directed mixed graphs. *Scandinavian Journal of Statistics*, 30(1):145–157, 2003. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4616754.

Gian Carlo Rota. On the foundations of combinatorial theory i. theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2(4):340–368, 1964. doi: 10.1007/BF00531932. URL https://doi.org/10.1007/BF00531932.

Tamás Rudas and Wicher Bergsma. Marginal models: An overview. *Trends and challenges in categorical data analysis: Statistical modelling and interpretation*, pages 67–115, 2023.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860, 2010.

Aliza Sarlan, Chayanit Nadam, and Shuib Basri. Twitter sentiment analysis. In *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pages 212–216, 2014. doi: 10.1109/ICIMU.2014.7066632.

Lawrence K Saul and Michael I Jordan. Mixed memory markov models: Decomposing complex stochastic processes as mixtures of simpler ones. *Machine learning*, 37:75–87, 1999.

Steven L Scott. Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American statistical Association*, 97(457):337–351, 2002.

Cosma Shalizi. Maximum likelihood estimation for markov chains. https://www.stat.cmu.edu/~cshalizi/462/lectures/06/markov-mle.pdf, January 2009.

Chris Sherlock, Tatiana Xifara, Sandra Telfer, and Mike Begon. A coupled hidden markov model for disease interactions. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 62(4):609–627, 2013.

Lisa Singh, Shweta Bansal, Leticia Bode, Ceren Budak, Guangqing Chi, Kornraphop Kawintiranon, Colton Padden, Rebecca Vanarsdall, Emily Vraga, and Yanchen Wang. A first look at covid-19 information and misinformation sharing on twitter. *arXiv preprint arXiv:2003.13907*, 2020.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Simon EF Spencer, Thomas E Besser, Rowland N Cobbold, and Nigel P French. 'super'or just 'above average'? supershedders and the transmission of escherichia coli o157: H7 among feedlot cattle. *Journal of the Royal Society Interface*, 12(110):20150446, 2015.

Richard P Stanley. Enumerative combinatorics volume 1 second edition. *Cambridge studies in advanced mathematics*, 2011. See Section 2.1, pp. 64–67 Inclusion-Exclusion.

Panayiota Touloupou. *Bayesian inference and model selection for partially observed stochastic epidemics.* PhD thesis, University of Warwick, 2016.

Panayiota Touloupou, Bärbel Finkenstädt, and Simon EF Spencer. Scalable bayesian inference for coupled hidden markov and semi-markov models. *Journal of Computational and Graphical Statistics*, 29(2):238–249, 2020.

Gian Marco Visani, Alexandra Hope Lee, Cuong Nguyen, David M Kent, John B Wong, Joshua T Cohen, and Michael C Hughes. Approximate bayesian computation for an explicit-duration hidden markov model of covid-19 hospital trajectories. In *Machine Learning for Healthcare Conference*, pages 567–613. PMLR, 2021.

Andrew Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 2003.

Wikipedia. Twitter — Wikipedia, the free encyclopedia, 2022. URL `https://en.wikipedia.org/w/index.php?title=Twitter&oldid=1094657535`. [Online; accessed 29-June-2022].

Wikipedia. Passenger rail franchising in great britain, 2023. URL https://en.wikipedia.org/wiki/Passenger_rail_franchising_in_Great_Britain. [Online; accessed 03-Oct-2023].

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Jesse Bennett Windle. Forecasting high-dimensional, time-varying variance-covariance matrices with high-frequency data and sampling pólya-gamma random variates for posterior distributions derived from logistic likelihoods. *PhD Thesis for University of Texas at Austin*, 2013.

Jie Yang and Ahm Mehbub Anwar. Social media analysis on evaluating organisational performance a railway service management context. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress*, pages 835–841, 2016. doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2016.143.

Yoel G. Yera, Rosa E. Lillo, Bo F. Nielsen, Pepa Ramírez-Cobo, and Fabrizio Ruggeri. A bivariate two-state markov modulated poisson process for failure modeling. *Reliability Engineering & System Safety*, 208:107318, 2021. ISSN 0951-8320. doi: https://doi.org/10.1016/j.ress.2020.107318. URL https://www.sciencedirect.com/science/article/pii/S0951832020308115.

Yao Zhai, Wei Liu, Yunzhi Jin, and Yanqing Zhang. Variational bayesian variable selection for high-dimensional hidden markov models. *Mathematics*, 12(7):995, 2024.

Shi Zhong and Joydeep Ghosh. A new formulation of coupled hidden markov models. Technical report, Department of Electrical and Computer Engineering, The University of Texas at Austin, 2001.

Appendix A

Review of relevant statistical methods

A.1 Frequentist inference

In this section, we introduce maximum likelihood estimation (MLE), a standard Frequentist approach for inference, applied to both the Hierarchical Marginal Model and the Univariate Marginal Model under our alternative parameterizations.

A.1.1 Maximum Likelihood Estimation

Maximum Likelihood Estimation (MLE) was first proposed by Fisher [1922] as a foundational method for parameter estimation within the Frequentist framework. He introduced the concept of the likelihood function and advocated estimating the parameters by maximizing this function with respect to the unknown parameters.

Consider a Multivariate Markov Chain (MMC) observed through its states $\boldsymbol{Z}(t)$ over discrete time points $t = 1, \ldots, T$. Our inference procedure begins by estimating the joint transition probabilities p, progresses through the marginal and auxiliary probabilities η and ζ, and concludes with the estimation of the regression coefficients β through the diffeomorphism map introduced in Lemma 4.12 and 3.13. To find the MLE of the joint

transition probability $\hat{p}_{\boldsymbol{ab}}$, we impose simplex constraints ensuring that each row sums up to 1:

$$\sum_{\boldsymbol{b}\in\boldsymbol{\mathcal{S}}} p_{\boldsymbol{ab}} = 1 \qquad \forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}} \tag{A.1}$$

We first write the likelihood function $\mathcal{L}(p)$ of the MMC $\boldsymbol{Z}$ in terms $p_{\boldsymbol{ab}}$ as:

$$\begin{aligned}\mathcal{L}(p) =& \mathbb{P}\big(\boldsymbol{Z}(1) = \boldsymbol{z}(1)\big) \prod_{t=2}^{T} \mathbb{P}(\boldsymbol{Z}(t) = \boldsymbol{z}(t) \mid \boldsymbol{Z}(t-1) = \boldsymbol{z}(t-1)\big) \\ =& \pi_0(\boldsymbol{z}(1)) \prod_{t=2}^{T} p_{\boldsymbol{ab}}^{n_{\boldsymbol{ab}}}\end{aligned} \tag{A.2}$$

where $\pi_0(\cdot)$ denotes the initial distribution of the state at time $t = 1$, and $n_{\boldsymbol{ab}}$ is the observed count of transitions from state $\boldsymbol{a}$ to state $\boldsymbol{b}$, defined as:

$$n_{\boldsymbol{ab}} = \#\Big\{t : 1 < t \leq T, \quad \boldsymbol{Z}(t-1) = \boldsymbol{a} \text{ and } \boldsymbol{Z}(t) = \boldsymbol{b}\Big\}.$$

Taking the logarithm yields the log-likelihood $\ell(p)$ as follows:

$$\ell(p) = \log\big(\ \pi_0(\boldsymbol{z}(1))\ \big) + \sum_{t=2}^{T} n_{\boldsymbol{ab}} \log\big(p_{\boldsymbol{ab}}\big) \tag{A.3}$$

To handle the simplex constraint, we eliminate one parameter by expressing one $p_{\boldsymbol{ab}}$ by the others through the simplex constraints in Equation (A.1). By taking the first derivative, we obtain the following:

$$\frac{\partial \ell(p)}{\partial p_{\boldsymbol{ab}}} = \frac{n_{\boldsymbol{ab}}}{p_{\boldsymbol{ab}}} - \frac{n_{\boldsymbol{ab'}}}{p_{\boldsymbol{ab'}}} \qquad \forall \boldsymbol{a} \in \boldsymbol{\mathcal{S}}$$

Set the first derivative of the log-likelihood equal to 0, the closed-form MLE for the joint transition probabilities can be computed as:

$$\hat{p}_{\boldsymbol{ab}} \ = \ \frac{n_{\boldsymbol{ab}}}{\sum_{\boldsymbol{b}} n_{\boldsymbol{ab}}}$$

For $\hat{p}_{\boldsymbol{ab}}$ to exist, we have to assume the positive recurrent for the MMC so that $\sum_{\boldsymbol{b}} n_{\boldsymbol{ab}} > 0$

for a significantly large time T, or in other words, each state $\boldsymbol{a}$ has been well-explored. Intuitively, the MLE $\hat{p}_{\boldsymbol{ab}}$ can be interpreted as the ratio of the count of transitions from $\boldsymbol{a}$ to $\boldsymbol{b}$ to the total count of transitions from $\boldsymbol{a}$.

A.1.2 Fisher Information

First formalized by Fisher [1925], the Fisher Information Matrix plays an essential role in the asymptotic theory of the MLE. To fully explore this, we now derive the Fisher Information Matrix.

Remark A.1 (Fisher Information Matrix)**.** Consider the vector of K unknown parameters $\boldsymbol{\theta} = \big(\theta_1, \ldots, \theta_K\big)$ and a set of observations $\boldsymbol{X}$ with the probability density function $f(\boldsymbol{X},\ \boldsymbol{\theta})$. Then, the Fisher Information Matrix $\mathcal{I}(\boldsymbol{\theta})$ with respect to the unknown parameter $\boldsymbol{\theta}$ is a $K \times K$ matrix defined to be the variance of the score function $s(\boldsymbol{\theta})$:

$$\mathcal{I}(\boldsymbol{\theta}) = \mathrm{Var}\big(s(\boldsymbol{\theta})\big) = \mathrm{Var}\big[\nabla \log\big(f(\boldsymbol{X},\ \boldsymbol{\theta})\big)\big]$$

where ∇ denotes the gradient operator.

Under regularity conditions, each element of the Fisher Information Matrix can also be expressed as the negative expected second derivative of the log-likelihood:

$$\big[\mathcal{I}(\boldsymbol{\theta})\big]_{k,k'} = \mathbb{E}\big[\frac{\partial^2}{\partial\theta_k\ \partial\theta_{k'}} \log\big(f(\boldsymbol{X},\ \boldsymbol{\theta})\big) \mid \boldsymbol{\theta}\big]$$

In our setting, let $\mathcal{I}(\boldsymbol{p})$ denote the Fisher Information Matrix for the joint transition probability vector $\boldsymbol{p} = \mathrm{vec}(P^\top)$. It is straightforward to see $\mathcal{I}(\boldsymbol{p})$ is block-diagonal,

reflecting the independence among the rows of the joint transition matrix P:

$$\mathcal{I}(\boldsymbol{p}) = \begin{pmatrix} \mathcal{I}_{\boldsymbol{a}^{(1)}} & 0 & \cdots & 0 \\ 0 & \mathcal{I}_{\boldsymbol{a}^{(2)}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathcal{I}_{\boldsymbol{a}^{(s^K)}} \end{pmatrix} \tag{A.4}$$

where each block $\mathcal{I}_{\boldsymbol{a}}$ corresponds to the Fisher Information for the row $P_{\boldsymbol{a}\cdot}$.

Specifically, each submatrix $\mathcal{I}_{\boldsymbol{a}}$ corresponds to the Fisher information for the multinomial distribution with parameters $\boldsymbol{p}_{\boldsymbol{a}\cdot}$, representing the probabilities of transitions originating from state $\boldsymbol{a}$ to all possible subsequent states. Define the corresponding vector of transition counts $\boldsymbol{n}_{\boldsymbol{a}\cdot}$ as follows:

$$\boldsymbol{n}_{\boldsymbol{a}\cdot} = \left(n_{\boldsymbol{a}\boldsymbol{b}^{(1)}}, n_{\boldsymbol{a}\boldsymbol{b}^{(2)}}, \ldots, n_{\boldsymbol{a}\boldsymbol{b}^{(s^K)}}\right) \tag{A.5}$$

This Fisher Information Matrix $\mathcal{I}_{\boldsymbol{a}}$ can be derived by computing the covariance of the score function $s(\boldsymbol{p}_{\boldsymbol{a}\cdot})$ of the transition probability vector $\boldsymbol{p}_{\boldsymbol{a}\cdot}$:

$$\begin{aligned} \mathcal{I}_{\boldsymbol{a}} =& \operatorname{Var}\left(s(\boldsymbol{p}_{\boldsymbol{a}})\right) = \operatorname{Var}\left(\nabla \ell(p)\right) \\ =& \operatorname{Var}\left((\frac{n_{\boldsymbol{a}\boldsymbol{b}^{(1)}}}{p_{\boldsymbol{a}\boldsymbol{b}^{(1)}}}, \ldots, \frac{n_{\boldsymbol{a}\boldsymbol{b}^{(1)}}}{p_{\boldsymbol{a}\boldsymbol{b}^{(s^K)}}})^\top\right) = \operatorname{Var}\left((\frac{\boldsymbol{n}_{\boldsymbol{a}\cdot}}{\boldsymbol{p}_{\boldsymbol{a}\cdot}})\right) \end{aligned}$$

Since $\boldsymbol{n}_{\boldsymbol{a}\cdot} \sim \text{Multinomial}(n_{\boldsymbol{a}}, \boldsymbol{p}_{\boldsymbol{a}\cdot})$, with $n_{\boldsymbol{a}} = \sum_{\boldsymbol{b}\in\boldsymbol{\mathcal{S}}} \boldsymbol{n}_{\boldsymbol{a}\boldsymbol{b}}$, its covariance matrix is:

$$\operatorname{Var}(\boldsymbol{n}_{\boldsymbol{a}\cdot}) = n_{\boldsymbol{a}}\left[\operatorname{diag}(\boldsymbol{p}_{\boldsymbol{a}\cdot}) - \boldsymbol{p}_{\boldsymbol{a}\cdot}\boldsymbol{p}_{\boldsymbol{a}\cdot}^\top\right]$$

Finally,

$$\begin{aligned}\mathcal{I}_{\boldsymbol{a}} =& n_{\boldsymbol{a}} \operatorname{diag}(\frac{1}{\boldsymbol{p}_{\boldsymbol{a}\cdot}}) \left[\operatorname{diag}(\boldsymbol{p}_{\boldsymbol{a}\cdot}) - \boldsymbol{p}_{\boldsymbol{a}\cdot}\boldsymbol{p}_{\boldsymbol{a}\cdot}^{\top}\right] \operatorname{diag}(\frac{1}{\boldsymbol{p}_{\boldsymbol{a}\cdot}}) \\ =& n_{\boldsymbol{a}} \left[\operatorname{diag}(\frac{1}{\boldsymbol{p}_{\boldsymbol{a}\cdot}}) - \mathbf{1}\mathbf{1}^{\top}\right]\end{aligned}$$

And the explicit matrix form is given as:

$$\mathcal{I}_{\boldsymbol{a}} = n_{\boldsymbol{a}} \cdot \begin{bmatrix} \frac{1}{p_{\boldsymbol{a}\boldsymbol{b}^{(1)}}} - 1 & -1 & \cdots & -1 \\ -1 & \frac{1}{p_{\boldsymbol{a}\boldsymbol{b}^{(2)}}} - 1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & \frac{1}{p_{\boldsymbol{a}\boldsymbol{b}^{(s^K)}}} - 1 \end{bmatrix}$$

Each diagonal element of the $\mathcal{I}_{\boldsymbol{a}}$ is given by $n_{\boldsymbol{a}}\big(\frac{1}{p_{\boldsymbol{a}\boldsymbol{b}^{(1)}}} - 1\big)$, while each off-diagonal entry is fixed to $-n_{\boldsymbol{a}}$. Moreover, the submatrix $\mathcal{I}_{\boldsymbol{a}}$ is symmetric, positive semi-definite, and singular with rank $s^K - 1$, reflecting the simplex constraint in Equation (A.1). In practice, we consider the observed Fisher Information $\mathcal{I}(\hat{\boldsymbol{p}})$ by plugging the maximum likelihood estimator $\hat{\boldsymbol{p}}$.

A.1.3 Consistency of MLE

Fisher [1922] established that, under appropriate regularity conditions, the MLE is asymptotically normally distributed with covariance matrix equal to the inverse of the Fisher Information Matrix. Formally, for parameter θ and its maximum likelihood estimator $\hat{\theta}$, this asymptotical normality is given by:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left((0\ ,\ \frac{1}{\mathcal{I}(\theta)}\right) \tag{A.6}$$

where $\xrightarrow{d}$ denotes convergence in distribution, n is the number of the data observed and $\mathcal{I}(\theta)$ is the Fisher Information discussed previously.

When consider the joint transition probability vector $\boldsymbol{p}$ for MMC as the target parameter and adapt A.6 it into the multivariate case yields:

$$\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}) \xrightarrow{d} \mathcal{N}\big(\mathbf{0}\ ,\ \mathcal{I}^{-1}(\boldsymbol{p})\big) \tag{A.7}$$

Here, $n = T - 1$ is the number of transitions within the MMC. $\mathcal{I}^{-1}(\boldsymbol{p})$ is the inverse of the Fisher Information Matrix representing the Cramér–Rao lower bound, the smallest possible variance achievable by any unbiased estimator. Fisher [1925] further proved that the MLE reaches this bound asymptotically, establishing its asymptotic efficiency.

In the context of MMCs, these regularity conditions are naturally associated with the Ergodicity. Specifically, the prerequisites of the Ergodic Theorem, including irreducibility, aperiodicity, and positive recurrence, ensure that the regularity conditions are satisfied and thereby guarantee the consistency of the MLE $\hat{p}$. Norris [1997] provided the Ergodic Theorem for the univariate Markov chain and we will adapt that into the multivariate setting as the following:

Remark A.2 (Ergodic Theorem for MMCs)**.** Consider an MMC $\boldsymbol{Z}$ with n step transition probability $p_{\boldsymbol{ab}}^{(n)} = \mathbb{P}(\boldsymbol{Z}(t+n) = \boldsymbol{b} \mid \boldsymbol{Z}(t) = \boldsymbol{a})$ being:

- **Irreducible:** For any $\boldsymbol{a}, \boldsymbol{b} \in \boldsymbol{\mathcal{S}}$, there exists some $n \geq 1$ such that $p_{\boldsymbol{ab}}^{(n)} > 0$.
- **Aperiodic:** The greatest common divisor of all n for which $p_{\boldsymbol{aa}}^{(n)} > 0$ is 1, for every $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$.
- **Positive Recurrent:** The expected return time to any state $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$ is finite.

Then, the Ergodic Theorem states that there exists a unique stationary distribution $\boldsymbol{\pi} = (\pi_{\boldsymbol{a}})_{\boldsymbol{a} \in \boldsymbol{\mathcal{S}}}$ satisfying

$$\frac{1}{T-1} \sum_{t=1}^{T-1} \mathbb{I}(\boldsymbol{Z}(t) = \boldsymbol{a}) \xrightarrow{\text{a.s.}} \pi_{\boldsymbol{a}} \qquad \text{as} \quad T \to \infty$$

Furthermore, for arbitrary bounded function $f : \boldsymbol{\mathcal{S}} \to \mathbb{R}$, we have

$$\frac{1}{T-1}\sum_{t=1}^{T-1} f(\boldsymbol{Z}(t)) \xrightarrow{\text{a.s.}} \sum_{\boldsymbol{a}\in\boldsymbol{\mathcal{S}}} \pi_{\boldsymbol{a}}\, f(\boldsymbol{a}) \qquad \text{as} \quad T \to \infty \tag{A.8}$$

where $\xrightarrow{\text{a.s.}}$ denotes converge almost surely.

Based on the Ergodic Theorem, Shalizi [2009] provided the proof to show the consistency of the MLE $\hat{p}$ to the true value p. Let $N_{\boldsymbol{ab}}$ denote the counts $n_{\boldsymbol{ab}}$ as a random variable, we can write the MLE $\hat{p}_{\boldsymbol{ab}}$ as the function of the observed data:

$$\hat{p}_{\boldsymbol{ab}} = \hat{p}_{\boldsymbol{ab}}(\boldsymbol{Z}) = \frac{N_{\boldsymbol{ab}}}{\sum_{\boldsymbol{b}} N_{\boldsymbol{ab}}}$$

Both numerator and the denominator can be written as a sum of the indicator functions:

$$\begin{aligned} N_{\boldsymbol{ab}} &= \sum_{t=1}^{T-1} \mathbb{I}\{\boldsymbol{Z}(t) = \boldsymbol{a},\, \boldsymbol{Z}(t+1) = \boldsymbol{b}\} \\ \sum_{\boldsymbol{b}} N_{\boldsymbol{ab}} &= \sum_{t=1}^{T-1}\sum_{\boldsymbol{b}} \mathbb{I}\{\boldsymbol{Z}(t) = \boldsymbol{a},\, \boldsymbol{Z}(t+1) = \boldsymbol{b}\} = \sum_{t=1}^{T-1} \mathbb{I}\{\, \boldsymbol{Z}(t) = \boldsymbol{a}\} \end{aligned}$$

On the other hand, if we choose $f\big(\boldsymbol{Z}(t)\big) = \mathbb{I}\big(\boldsymbol{Z}(t) = \boldsymbol{a}\big)$ for each state $\boldsymbol{a} \in \boldsymbol{\mathcal{S}}$, Equation (A.8) in the Ergodic Theorem transforms into

$$\frac{1}{T-1}\sum_{t=2}^{T} \mathbb{I}\{\boldsymbol{Z}(t) = \boldsymbol{a}\} \xrightarrow{\text{a.s.}} \pi_{\boldsymbol{a}} \qquad \text{as} \quad T \to \infty \tag{A.9}$$

Equation (A.9) can be interpreted as the fraction of time spent in each state $\boldsymbol{a}$ converges to the equilibrium distribution $\pi_{\boldsymbol{a}}$. Consequently, for sufficiently large T, by the fact that the expectation of an indicator function is a probability, we yields:

$$\begin{aligned} \frac{1}{T-1} N_{\boldsymbol{ab}} &\xrightarrow{\text{a.s.}} \mathbb{E}\big[\, \mathbb{I}\{\boldsymbol{Z}(t-1) = \boldsymbol{a}\}\, \mathbb{I}\{\boldsymbol{Z}(t) = \boldsymbol{b}\}\, \big] = \mathbb{P}\big(Z(t-1) = \boldsymbol{a},\, Z(t) = \boldsymbol{b}\big) \\ \frac{1}{T-1}\sum_{\boldsymbol{b}} N_{\boldsymbol{ab}} &\xrightarrow{\text{a.s.}} \mathbb{E}\big[\, \mathbb{I}\{\boldsymbol{Z}(t-1) = \boldsymbol{a}\}\, \big] = \mathbb{P}\big(Z(t-1) = \boldsymbol{a}\big) \end{aligned}$$

Taking the ratio shows that

$$\hat{p}_{\boldsymbol{ab}} = \frac{N_{\boldsymbol{ab}}}{\sum_{\boldsymbol{b}} N_{\boldsymbol{ab}}} \xrightarrow{\text{a.s.}} \frac{\mathbb{P}\big(Z(t-1)=\boldsymbol{a},\, Z(t)=\boldsymbol{b}\big)}{\mathbb{P}\big(Z(t-1)=\boldsymbol{a}\big)} = p_{\boldsymbol{ab}}$$

thus proving that the MLE $\hat{p}_{\boldsymbol{ab}}$ is consistent for the true transition probabilities.

A.1.4 Delta Method

In this section, we will show the consistency of the MLE $\hat{\boldsymbol{\beta}}$ for the regression coefficients $\boldsymbol{\beta}$ under both Multivariate marginal model and Univariate marginal model. We achieve this via implementing the the Delta Method on the MLE.

Remark A.3 (Multivariate Delta Method)**.** [Lehmann and Casella, 1998] Let the vector $\hat{\boldsymbol{\theta}}$ denote the MLE for the target parameters $\boldsymbol{\theta}$. Then, the Delta Method states that for arbitrary differentiable function $g(\cdot)$ defined on the parameter space $\boldsymbol{\Theta}$ of the parameter $\boldsymbol{\theta}$, the asymptotic distribution of $g(\hat{\boldsymbol{\theta}})$ is still normal and is given as:

$$\sqrt{n}\,\big(g(\hat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\big) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0}\,,\ \nabla g(\boldsymbol{\theta})\,\mathcal{I}^{-1}(\boldsymbol{\theta})\,\nabla g(\boldsymbol{\theta})^{\top}\Big) \tag{A.10}$$

where $\nabla g(\boldsymbol{\theta}) = \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is the gradient of the differentiable function $g(\cdot)$ with respect to the vector $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{p}}$ denote the MLE of the joint transition probability vector $\boldsymbol{p}$ for the MMC under ergodicity, ensuring consistency and asymptotic normality of $\hat{\boldsymbol{p}}$ as stated in Equation (A.7). Applying the Delta Method from Equation (A.10) with $\hat{\boldsymbol{p}}$, we obtain:

$$\sqrt{n}\big(g(\hat{\boldsymbol{p}}) - g(\boldsymbol{p})\big) \xrightarrow{d} \mathcal{N}\Big(\mathbf{0},\ \nabla g(\boldsymbol{p})\mathcal{I}^{-1}(\boldsymbol{p})\nabla g(\boldsymbol{p})^{\top}\Big).$$

Recall the diffeomorphic transformations $g_1(\cdot)$ (see Lemma 4.12) and $g_2(\cdot)$ (see Lemma 3.13), which maps $\boldsymbol{p}$ to regression coefficients $\boldsymbol{\beta}$ satisfying strong compatibility in the Hierarchical Marginal Model , and to $\boldsymbol{\beta}_\sigma$ under permutation σ in the Univariate Marginal Model, respectively. Thus, by the Delta Method, the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_\sigma$ are consistent and

asymptotically normal as follows:

$$\begin{aligned}\boldsymbol{\beta} &\sim \mathcal{N}\big(g_1(\hat{\boldsymbol{p}}),\ \frac{1}{n}\nabla g_1(\hat{\boldsymbol{p}})\,\mathcal{I}^{-1}(\hat{\boldsymbol{p}})\,\nabla g_1(\hat{\boldsymbol{p}})^\top\big)\\ \boldsymbol{\beta}_\sigma &\sim \mathcal{N}\big(g_2(\hat{\boldsymbol{p}}),\ \frac{1}{n}\nabla g_2(\hat{\boldsymbol{p}})\,\mathcal{I}^{-1}(\hat{\boldsymbol{p}})\,\nabla g_2(\hat{\boldsymbol{p}})^\top\big)\end{aligned}$$

where $\mathcal{I}^{-1}(\hat{\boldsymbol{p}})$ denotes the inverse of the observed Fisher Information Matrix specified in Equation (A.4) by plugging the MLE $\hat{\boldsymbol{p}}$.

The asymptotic normality of the regression coefficients $\boldsymbol{\beta}$ enables the utilization of the Hotelling's T^2 statistic for hypothesis testing. This allows us to assess the potential Granger non-causality and contemporaneous independence embedded in the MMC by testing the relevant components of $\boldsymbol{\beta}$, as elaborated in the following section.

A.1.5 Hotelling's T^2 Test

In the literature, a widely used method for testing mean vectors with multiple correlated variables is the Hotelling's T^2 test, proposed by Hotelling [1931]. It generalises the univariate Student's t-test to the multivariate case and is commonly applied in both one-sample and two-sample settings when the variance Σ of the data is unknown and approximate by the sample variance $\mathbf{S}$:

- **One-sample Hotelling's T^2 test**: This is to test whether the mean of a multivariate normal equals a specified mean vector. Let $\bar{\boldsymbol{X}}$ and $\mathbf{S}$ be the sample mean and sample covariance matrix from a random variable $\boldsymbol{X}$ of size n from m-variate normal distribution $\mathcal{N}_m(\boldsymbol{\mu}, \Sigma)$. Consider the null and alternative hypothesis:

$$H_0 : \bar{\boldsymbol{X}} = \boldsymbol{\mu}_0 \qquad H_1 : \bar{\boldsymbol{X}} \neq \boldsymbol{\mu}_0$$

  The corresponding test statistic T^2 is given by:

$$T^2 = n(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top \mathbf{S}^{-1}(\bar{\boldsymbol{X}} - \boldsymbol{\mu}_0)$$

Under H_0, the test statistic T^2 can be transformed into an F-distribution:

$$\frac{n-m}{m(n-1)}\, T^2 \sim F_{m,\, n-m}$$

- **Two-sample Hotelling's T^2 test**: This test evaluates whether the mean vectors of two multivariate populations are equal. Consider two sets of random samples $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ independently drawn from m-variate normal distributions:

$$\boldsymbol{X}_1 \sim \mathcal{N}(\mu_1, \Sigma_1) \qquad \boldsymbol{X}_2 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

with sample sizes n_1 and n_2, respectively. Let $\bar{\boldsymbol{X}}_1$ and $\bar{\boldsymbol{X}}_2$ denote the sample means, and let $\mathbf{S}_1$ and $\mathbf{S}_2$ be the sample covariance matrices of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Consider the null and alternative hypotheses:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \qquad H_1 : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$$

Additionally assuming homogeneity of covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ among $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$, the corresponding test statistic T^2 is given by:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)^\top \mathbf{S}_p^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$$

where $\mathbf{S}_p$ denotes the pooled covariance matrix and can be expressed as the weighted average of the sample variance $\mathbf{S}_1$ and $\mathbf{S}_2$ for $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively:

$$\mathbf{S}_p = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}$$

Under H_0, the test statistic T^2 can be transformed into an F-distribution:

$$\frac{(n_1 + n_2 - m - 1)}{m(n_1 + n_2 - 1)}\, T^2 \sim F_{m,\, n-m}$$

This enables us to compute p-values for testing equality of the relevant coefficient vectors

in both the Hierarchical Marginal Model and the Univariate Marginal Model, and thus for learning the dependence structure in the MMC. Specifically:

- **Hierarchical Marginal Model**: test $\boldsymbol{\beta}_j = \mathbf{0}$ for Granger non-causality via the one-sample Hotelling's T^2 test; test $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j'}$ with $g_{j'} \subseteq g_j$ for contemporaneous independence via the two-sample Hotelling's T^2 test.
- **Univariate Marginal Model**: test $\boldsymbol{\beta}_\sigma = \mathbf{0}$ for both Granger non-causality and contemporaneous independence via the one-sample Hotelling's T^2 test.

A.1.6 Likelihood Ratio Test

However, Hotelling's T^2 test relies on the assumption that the regression coefficients follow a multivariate normal distribution. In the two-sample case, it additionally requires that the covariance matrices of the two groups are equal, so that a pooled covariance matrix $\mathbf{S}_p$ can be used. This assumption may not hold in practice. An alternative way in literature is the likelihood ratio test proposed by Wilks [1938], which can also be utilized to examine potential dependency structures in MMC. It is straight forward to see that the structures, such as contemporaneous independence and Granger non-causality, define a nested model by imposing additional restrictions on the regression coefficients $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_\sigma$ in both the Hierarchical Marginal Model and the Univariate Marginal Model. This nesting structures fulfills the prerequisite of the likelihood ratio test.

Formally, consider two nested models:

- A full model with parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$
- A constrained model $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ with parameters restricted to the subset $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$

and the following hypothesis test:

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \qquad \text{versus} \qquad H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$$

The likelihood ratio Λ is defined as the ratio of the maximum likelihood under the re-

stricted parameter space $\boldsymbol{\Theta}_0$ to the maximum likelihood under the full (unrestricted) parameter space $\boldsymbol{\Theta}$:

$$\Lambda = \frac{\max\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta_0}} \mathcal{L}(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\theta})}$$

Under H_0, Wilks [1938] shown that the likelihood ratio statistic, $-2\log\Lambda$, asymptotically follows a chi-squared distribution for the sufficiently large sample size n, given regularity conditions:

$$-2\log\Lambda \sim \chi^2_\nu$$

with ν stands for the degree of freedom, computed as the difference between the number of independent parameters within the full model and the nested model.

In our setting, the null hypothesis H_0 corresponds to the restricted model, where certain dependency structures are present within the MMC. The alternative hypothesis H_1 represents the full model, in which the MMC is fully connected. We consider a full model with no constraints on the regression coefficients $\boldsymbol{\beta}$, and a nested model that imposes specific constraints on $\boldsymbol{\beta}$ reflecting the dependency structures of interest. Fit both of models via maximizing the likelihood and compute the likelihood ratio Λ, followed by testing the targeting dependency structure by comparing the likelihood ratio statistic $-2\log\Lambda$ with the upper tail critical value $\chi^2_{\nu,\alpha}$ with significance level α.

A.2 Bayesian Inference

A.2.1 Conjugate Dirichlet Distribution

In Bayesian setting, the typical way to model the joint transition probability vector $\boldsymbol{p}$, which observes the multinomial distribution, is through the conjugate Dirichlet Distribution [Bernardo and Smith, 2009, Gelman et al., 1995]. Specifically, for K-variate the random vector $\boldsymbol{x} = \big(x_1, \ldots, x_K\big)$ follows the Dirichlet Distribution with parameters $\boldsymbol{\alpha} = \big(\alpha_1, \ldots, \alpha_K\big)$, the corresponding probability density function $\mathrm{Dir}\big(\boldsymbol{x} \mid \boldsymbol{\alpha}\big)$ can be

written as:

$$\mathrm{Dir}(\boldsymbol{x} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{Beta}(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

where $\mathrm{Beta}(\boldsymbol{\alpha})$ is the normalizing constant following the multivariate beta function:

$$\mathrm{Beta}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}$$

In our setting, the random vector $\boldsymbol{x}$ stands for a row of the joint transition matrix $\boldsymbol{p_{a\cdot}}$ and the parameter $\boldsymbol{\alpha}$ is set as a vector of 1, indicating the uniform non-informative prior. It follows the posterior is a still Dirichlet distribution given as:

$$\pi(\boldsymbol{p_{a\cdot}} \mid \boldsymbol{Z}) \propto \prod_{t=2}^{T} p_{\boldsymbol{ab}}^{n_{\boldsymbol{ab}}} \prod_{\boldsymbol{b} \in \boldsymbol{\mathcal{S}}} p_{\boldsymbol{ab}}^{\alpha_{\boldsymbol{ab}} - 1} \;\propto\; \mathrm{Dir}\big(\boldsymbol{n_{a\cdot}} + \boldsymbol{\alpha_{a\cdot}}\big) \tag{A.11}$$

where $\boldsymbol{n_{a\cdot}}$ is defined in Equation (A.5) and $\boldsymbol{\alpha_{a\cdot}}$ is defined in a similar way as:

$$\boldsymbol{\alpha_{a\cdot}} = \big(\alpha_{\boldsymbol{ab}^{(1)}}, \alpha_{\boldsymbol{ab}^{(2)}}, \ldots, \alpha_{\boldsymbol{ab}^{(s^K)}}\big)$$

The conjugate Dirichlet distribution offers an explicit and straightforward way to obtain the posterior distribution of the joint transition probabilities p directly from the observed counts. Analogous to the transformation described in Section A.1.4, posterior samples of p can be mapped through g_1 and g_2 to obtain corresponding posterior samples of the regression coefficients $\boldsymbol{\beta}$, thereby yielding their posterior distribution.

In contrast, Pólya-Gamma logistic regression Polson et al. [2013] enables direct Bayesian inference of $\boldsymbol{\beta}$ by introducing auxiliary Pólya-Gamma variables. This approach will be explored in the following section.

A.2.2 Posterior Contour Probability

In the Bayesian setting, we would like to test the certain structures within the vectors of the regression coefficients $\boldsymbol{\beta}$ and hence induce the potential dependency structures embedded within the MMC. This can be regarded as making simultaneous probability statements in multivariate inferential problems based on the posterior samples. One popularly used concept to quantitatively identify that is the posterior contour probabilities, which is also known as posterior p value, introduced by Held [2004].

Followed by Box and Tiao [1973], given a specific point estimation of parameter of interest $\boldsymbol{\theta}_0$ and the observed dataset $\boldsymbol{x}$, it lies inside or outside Highest Posterior Density (HPD) region with some predefined simultaneous credible level $1-\alpha$ if and only if:

$$\mathbb{P}\big(p(\boldsymbol{\theta} \mid \boldsymbol{x}) > p(\boldsymbol{\theta}_0 \mid \boldsymbol{x}) \mid \boldsymbol{x}\big) \leq 1-\alpha$$

It can easily be reversed to define the posterior contour probability $\mathbb{P}\big(\boldsymbol{\theta_0} \mid \boldsymbol{x}\big)$ of the interested parameter $\boldsymbol{\theta}_0$ as 1 minus the HPD region:

$$\mathbb{P}\big(\boldsymbol{\theta_0} \mid \boldsymbol{x}\big) = 1 - \mathbb{P}\big(p(\boldsymbol{\theta} \mid \boldsymbol{x}) > p(\boldsymbol{\theta}_0 \mid \boldsymbol{x}) \mid \boldsymbol{x}\big) = \mathbb{P}\big(p(\boldsymbol{\theta} \mid \boldsymbol{x}) \leq p(\boldsymbol{\theta}_0 \mid \boldsymbol{x}) \mid \boldsymbol{x}\big) \tag{A.12}$$

When given the posterior samples $p(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{x})$ and the explicit form of the density $p(\boldsymbol{\theta} \mid \boldsymbol{x})$, we can estimate the posterior contour probability by the proportion of posterior samples with density value smaller or equal $p(\boldsymbol{\theta}_0 \mid \boldsymbol{x})$ through the indicator function $\mathbb{I}$:

$$\hat{\mathbb{P}}\big(\boldsymbol{\theta_0} \mid \boldsymbol{x}\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big[p(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{x}) \leq p(\boldsymbol{\theta}_0 \mid \boldsymbol{x})\big] \tag{A.13}$$

which can be regarded as the Monte Carlo version of previous Equation (A.12).

For the case where the explicit form the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\nu})$ is not tractable according to some nuisance parameter $\boldsymbol{\nu}$, Held [2004] suggested to estimate $p(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{\nu})$ via the posterior

median, which ends up with the modified Rao-Blackwell estimate of Equation (A.12):

$$\hat{\mathbb{P}}_{\mathrm{RB}}\big(\boldsymbol{\theta_0} \mid \boldsymbol{x}\big) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big\{\mathrm{med}_j\big[p(\boldsymbol{\theta}^{(i)} \mid \boldsymbol{x}, \boldsymbol{\nu}^{(j)})\big] \leq \mathrm{med}_j\big[p(\boldsymbol{\theta}_0 \mid \boldsymbol{x}, \boldsymbol{\nu}^{(j)})\big]\big\} \tag{A.14}$$

In our case, the data $\boldsymbol{x}$ corresponds to the observed realization of the MMC $\boldsymbol{Z}$, and the parameter $\boldsymbol{\theta}$ represents the associated regression coefficients $\boldsymbol{\beta}$. Since there are no nuisance parameters $\boldsymbol{\nu}$, we directly utilize the posterior probability estimate $\hat{\mathbb{P}}(\boldsymbol{\theta_0} \mid \boldsymbol{x})$ from Equation (A.13) to assess the significance of hypotheses regarding the regression coefficients within both the Hierarchical Marginal Model and the Univariate Marginal Model under the Bayesian framework. Specifically:

- In the **Hierarchical Marginal Model**, we test $\boldsymbol{\beta}_j = \mathbf{0}$ for Granger non-causality, and $\boldsymbol{\beta}_j = \boldsymbol{\beta}_{j'}$ with $g_{j'} \subset g_j$ for contemporaneous independence, which can be transferred via testing $\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'} = 0$.

- In the **Univariate Marginal Model**, we test $\boldsymbol{\beta}_\sigma = \mathbf{0}$ for both Granger non-causality and contemporaneous independence.

Furthermore, the Rao-Blackwell estimate $\hat{\mathbb{P}}_{\mathrm{RB}}$ defined in Equation (A.14) will be employed in the next chapter, where nuisance parameters $\boldsymbol{\nu}$ are introduced.

A.3 Pólya Gamma Distribution

In order to infer the regression coefficients $\boldsymbol{\beta}$ within the logistic regression in a fully Bayesian fashion directly, we introduce the Pólya-Gamma augmentation proposed by Polson et al. [2013]. Essentially, the Pólya–Gamma random variable ω is a specially constructed auxiliary variable that equals in distribution to an infinite sum of Gammas distribution. Precisely, if $\omega \sim PG(b, c)$,

$$\omega \overset{D}{=} \frac{1}{2\pi^2}\sum_{i=1}^{\infty} \frac{g_i}{(i-1/2)^2 + c^2/(4\pi^2)}$$

where $\stackrel{D}{=}$ refers equal in distribution and $g_i \sim \text{Ga}(b, 1)$ are independent gamma random variables.

It follows the key integral identity comes with the Pólya–Gamma distribution for $b > 0$:

$$\frac{(e^{\psi})^a}{(1+e^{\psi})^b} = 2^{-b}\, e^{\kappa\psi} \int_0^{\infty} e^{-\frac{1}{2}\omega\psi^2} p(\omega) d\omega \tag{A.15}$$

where $\kappa = a - \frac{b}{2}$ and $\omega \sim \text{PG}(b, 0)$ is the Pólya–Gamma random variable. The integral identity in Equation (A.15) leads to two useful properties:

- If $\psi = x^{\top}\boldsymbol{\beta}$ is a linear function of the predictors, the resulting likelihood with respective to $\boldsymbol{\beta}$ is Gaussian.

- The conditional distribution of the latent variable ω given ψ is still of Pólya–Gamma.

As a result, the introduction of Pólya–Gamma latent variables ω for each observation renders the logistic likelihood conditionally conjugate to a normal prior on the regression coefficients, enabling straightforward Gibbs updates.

A.3.1 Formulate Pólya Gamma Multinomial Logistic Regression

Consider the multinomial sample $y_i = \{y_{ij}\}$ that denotes the count for category j in observation i for $j = 1, 2, \ldots, J$. n_i is the total number of the responses for i-th observation such that:

$$n_i = \sum_{j=1}^{J} y_{ij} \qquad \text{for} \quad i = 1, 2, \ldots, N$$

The multinomial logistic link function specifies the probability of choosing category j in observation i using a softmax function applied to the log odds ψ:

$$p_{ij} = \frac{e^{\psi_{ij}}}{\sum_{k=1}^{J} e^{\psi_{ik}}}$$

where ψ_{ij} is the log odds of the category j in observation i and it is modeled by $x_i^T \boldsymbol{\beta}_j$. For identifiability, we constrained the baseline level $\boldsymbol{\beta}_1$ to be 0.

Held and Holmes [2006] computed the partial likelihood for $\boldsymbol{\beta}_j$ conditioned on all the rest $\boldsymbol{\beta}_{-j}$ as:

$$\mathcal{L}(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, y) = \prod_{i=1}^{N} \Big(\frac{e^{c_{ij}}}{1+e^{c_{ij}}} \Big)^{y_{ij}} \Big(\frac{1}{1+e^{c_{ij}}} \Big)^{n_{ij}-y_{ij}}$$

where

$$c_{ij} = x_i^\top \boldsymbol{\beta}_j - d_{ij} \qquad \text{with} \qquad d_{ij} = \log \sum_{k \neq j} e^{x_i^\top \boldsymbol{\beta}_k} \tag{A.16}$$

By letting $\kappa_{ij} = y_{ij} - \frac{n_i}{2}$ and taking out , we yields:

$$\mathcal{L}(\boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, y) = \prod_{i=1}^{N} e^{\kappa_{ij} c_{ij}} \Big(\frac{e^{\frac{\eta_{ij}}{2}}}{1+e^{\eta_{ij}}} \Big)^{n_i}$$

Choose $a = \frac{1}{2} n_i$, $b = n_i$ and apply the integral identity in Equation (A.15), we obtain:

$$\mathcal{L}(\omega_j, \boldsymbol{\beta}_j \mid \boldsymbol{\beta}_{-j}, y) = \prod_{i=1}^{N} 2^{-b} e^{\kappa_{ij} c_{ij}} \int_0^\infty e^{-\frac{1}{2}\omega_{ij} c_{ij}^2} \, p(\omega_{ij} \mid n_i, 0) d\omega_{ij} \tag{A.17}$$

where $\omega_{ij} \sim \mathrm{PG}(n_i, 0)$ is the Pólya–Gamma latent variables introduced. We further notice that the integrand is unnormalized joint density of ω_{ij} and $\boldsymbol{\beta}$, which is by definition $\omega_{ij} \sim PG(b, c_{ij})$. This gives us the conjugate posterior of ω_{ij}:

$$\omega_{ij} \mid \boldsymbol{\beta}, y \sim \mathrm{PG}(n_i, c_{ij})$$

On the other hand, by dropping out the irrelevant terms, we can simplify the likelihood with respect to $\boldsymbol{\beta}_j$ in Equation (A.17) as:

$$\mathcal{L}(\boldsymbol{\beta}_j \mid \omega_j, \boldsymbol{\beta}_{-j}, y) \propto \prod_{i=1}^{N} \exp\{\kappa_{ij} c_{ij} - \tfrac{1}{2}\omega_{ij} c_{ij}^2\}$$

Stacking over $i = 1, 2, \ldots, N$ and representing it in the matrix form gives

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}_j \mid \omega_j, \boldsymbol{\beta}_{-j}, y) &\propto \exp\{\boldsymbol{\kappa}_j^\top \boldsymbol{c}_j - \tfrac{1}{2}\boldsymbol{c}_j^\top \Omega \boldsymbol{c}_j\} \\ &\propto \exp\{\boldsymbol{\kappa}_j^\top (X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j) - \tfrac{1}{2}(X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j)^\top \Omega (X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j)\}\end{aligned}$$

where $\boldsymbol{\kappa}_j = (\kappa_{1j}, \ldots, \kappa_{Nj})^\top$, $\boldsymbol{c}_j = (c_{1j}, \ldots, c_{Nj})^\top$, $\boldsymbol{d}_j = (d_{1j}, \ldots, d_{Nj})^\top$ and $\Omega_j = \operatorname{diag}(\omega_{1j}, \ldots, \omega_{Nj})$.

By employing the conjugate multivariate normal prior $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}_{0j}, \Sigma_{0j})$, we can compute the log posterior density as:

$$\begin{aligned}\ell(\boldsymbol{\beta}_j \mid \omega_j, \boldsymbol{\beta}_{-j}, y) &\propto -\tfrac{1}{2}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{0j})^\top \Sigma_{0j}^{-1}(\boldsymbol{\beta}_j - \boldsymbol{\mu}_{0j}) + \boldsymbol{\kappa}_j^\top (X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j) - \tfrac{1}{2}(X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j)^\top \Omega (X^\top \boldsymbol{\beta}_j - \boldsymbol{d}_j) \\ &\propto -\tfrac{1}{2}\boldsymbol{\beta}_j^\top \big(\Sigma_{0j}^{-1} + X^\top \Omega_j X\big)\boldsymbol{\beta}_j + \boldsymbol{\beta}_j^\top \Big[\Sigma_{0j}^{-1}\boldsymbol{\mu}_{0j} + X^\top \big(\boldsymbol{\kappa}_j + \Omega_j \boldsymbol{d}_j\big)\Big]\end{aligned}$$

This log posterior density corresponds to another multivariate normal as:

$$\boldsymbol{\beta}_j \mid \omega_j, \boldsymbol{\beta}_{-j}, y \sim N(\boldsymbol{\mu}_j, \Sigma_j)$$

where

$$\Sigma_j = \big(\Sigma_{0j}^{-1} + X^\top \Omega_j X\big)^{-1} \qquad \boldsymbol{\mu}_j = \Sigma_j \Big[\Sigma_{0j}^{-1}\boldsymbol{\mu}_{0j} + X^\top \big(\boldsymbol{\kappa}_j + \Omega_j \boldsymbol{d}_j\big)\Big] \tag{A.18}$$

The following Algorithm 9 outlines the Gibbs sampling steps for the Pólya–Gamma augmentation applied to multinomial logistic regression in each iteration. In our setting, y_i denotes the observed realization of a subgroup g_j of the states $\boldsymbol{Z}_j(t)$, treated as a multinomial response. The covariate matrix X is constructed by row-binding the design matrices defined in Equations 4.24 and 3.20, corresponding to the Hierarchical Marginal Model and the Univariate Marginal Model, respectively. These design matrices are constructed based on the previous state $\boldsymbol{Z}(t-1)$ at each time step.

Algorithm 9 Pólya Gamma Augmentation for Multinomial Logistic Regression

Inputs: The multinomial observations y_i and the total number of the responses n_i for each observation i. The corresponding covariates X as well as the parameters $\boldsymbol{\mu}_{0j}, \Sigma_{0j}$ for the prior of $\boldsymbol{\beta}_j$ for each class j.

Outputs: Samples of the regression coefficients $\boldsymbol{\beta}_j$ from the posterior.

for each non-baseline class $j = 2, \ldots, J$ **do**
  for each observation $i = 1, \ldots, N$ **do**
    Compute d_{ij} and c_{ij} according to Equation (A.16)
    Sample $\omega_{ij} \sim \mathrm{PG}(n_i, c_{ij})$
  end for
  Compute $\boldsymbol{\mu}_j$ and Σ_j according to Equation (A.18)
  Sample $\boldsymbol{\beta}_j \sim N(\boldsymbol{\mu}_j, \Sigma_j)$
end for

A.3.2 Computational Cost

Polson et al. [2013] explicitly presented the computational cost to estimate the regression coefficients $\boldsymbol{\beta}$. For the case with n observations, l response levels, and p predictors, the computational costs are decomposed as follows:

1. Sampling the latent Polya–Gamma variables ω_i with computational cost $O(nl)$

2. Sampling $(l-1) \times p$ regression coefficients $\boldsymbol{\beta}$: Constructing a $(l-1)\,p \times (l-1)\,p$ covariance matrix with $O\big(n\,p^2\,(l-1)^2\big)$ followed by inverting it with $O\big(p^3\,(l-1)^3\big)$.

Hence, each iteration requires the computational cost:

$$O\big(n\,p^2\,l^2 + p^3\,l^3\big) \tag{A.19}$$

A.4 Logistic Regressions

Starting with a regular logistic regression, we model a single binary response Y via

$$\log \frac{p(Y=2 \mid \boldsymbol{x})}{p(Y=1 \mid \boldsymbol{x})} \;=\; \beta^T \boldsymbol{x}.$$

Extending this idea naively to a multivariate categorical response $(Y_1, Y_2, \ldots, Y_K)$ with K covariates could involve a large multinomial logistic or log-linear model for the joint:

$$P(Y_1 = y_1,\, Y_2 = y_2,\, \ldots,\, Y_K = y_K \mid \boldsymbol{x})$$

However, such a model can become cumbersome and obscure on how marginal and partial log-odds or interactions are structured.

Instead, a widely adopted multinomial logistic regression framework by McCullagh and Nelder [1989] (Chapter 6, p. 219) and Glonek and McCullagh [1995] recovers the joint distribution from the marginal logistic parameters. Specifically, Glonek and McCullagh [1995] introduce the marginal log-odds contrasts η to capture how subsets of the variables deviate from a baseline.

Consider a set of categorical variables $\{Y_1, \ldots, Y_K\}$. Let $Y_i = 1$ denote the baseline category for each i. In the binary case, the marginal contrast η_U in GM logistic regression measures the effect of all variables in the non-empty subset $U \subseteq [K]$ jointly taking non-baseline categories. It can be written as an inclusion-exclusion sum:

$$\eta^U \;=\; \sum_{\forall V \subseteq U} (-1)^{|U \setminus V|}\, \log\Big(\mathbb{P}\big(Y_i = 2 \quad \text{for} \quad i \in V, \quad Y_{i'} = 1 \quad \text{for} \quad i' \in U \setminus V\big)\Big).$$

This construction ensures η^U extracts the pure log-odds effect for the subset U. Then, the corresponding regression model is built on these marginal log-odds contrasts η as:

$$\eta^U = \mathcal{X}\boldsymbol{\beta}$$

for the design matrix $\mathcal{X}$ and a column vector of the regression coefficient $\boldsymbol{\beta}_U$ of subset U.

Example A.4. Consider three categorical response variables A, B, and C, each taking two levels 1 and 2. Under this setting, GM logistic regression model on log-odds contrasts can be constructed as follows:

1. Define the joint probabilities $\pi_{abc} = \mathbb{P}(A = a, B = b, C = c)$ for $a, b, c \in \{1, 2\}$. Let $\boldsymbol{\pi}$ denote the vector of these joint probabilities. The non-singular linear transformation that maps $\boldsymbol{\pi}$ to a vector of marginal probabilities $\boldsymbol{\gamma}$ can be constructed as:

$$\boldsymbol{\gamma} = M\boldsymbol{\pi}$$

where M is the marginal indicator matrix consisting only zeroes and ones and the vector $\boldsymbol{\gamma}$ collects all univariate, bivariate, and trivariate marginals:

$$\boldsymbol{\gamma} = \Big(\underbrace{\pi_{a\cdot\cdot} \quad \pi_{\cdot b\cdot} \quad \pi_{\cdot\cdot c}}_{\substack{\text{univariate}\\ \text{marginal}}} \quad \underbrace{\pi_{ab\cdot} \quad \pi_{a\cdot c} \quad \pi_{\cdot bc}}_{\substack{\text{bivariate}\\ \text{marginal}}} \quad \underbrace{\pi_{abc}}_{\substack{\text{trivariate}\\ \text{marginal}}} \Big)$$

with $\pi_{a\cdot\cdot} = \mathbb{P}(A = a)$ referring to the marginal probability of A being in level a.

2. Then, the log-odds contrasts are obtained via:

$$\boldsymbol{y} = \boldsymbol{C} \log \boldsymbol{\gamma}$$

where $\boldsymbol{C}$ is a contrast matrix, and $\boldsymbol{y}$ is a vector of log-odds contrasts. For three binary responses, it can be explicitly written in the matrix form as:

$$\boldsymbol{y} = \begin{pmatrix} y_a \\ y_b \\ y_c \\ y_{ab} \\ y_{ac} \\ y_{bc} \\ y_{abc} \end{pmatrix} = \begin{pmatrix} \log \pi_{1\cdot\cdot} - \log \pi_{2\cdot\cdot} \\ \log \pi_{\cdot 1\cdot} - \log \pi_{\cdot 2\cdot} \\ \log \pi_{\cdot\cdot 1} - \log \pi_{\cdot\cdot 2} \\ \log \pi_{11\cdot} - \log \pi_{12\cdot} - \log \pi_{21\cdot} + \log \pi_{22\cdot} \\ \log \pi_{1\cdot 1} - \log \pi_{1\cdot 2} - \log \pi_{2\cdot 1} + \log \pi_{2\cdot 2} \\ \log \pi_{\cdot 11} - \log \pi_{\cdot 12} - \log \pi_{\cdot 21} + \log \pi_{\cdot 22} \\ \log\big((\pi_{111}\pi_{221}\pi_{122}\pi_{212})/(\pi_{121}\pi_{211}\pi_{112}\pi_{222})\big) \end{pmatrix}$$

3. Finally, each contrast y is modelled as a linear function of covariates $\boldsymbol{x}$ through

coefficients $\boldsymbol{\beta}$:

$$y_a(\boldsymbol{x}) = \boldsymbol{\beta}_a^T \boldsymbol{x}, \qquad y_b(\boldsymbol{x}) = \boldsymbol{\beta}_b^T \boldsymbol{x}, \qquad y_c(\boldsymbol{x}) = \boldsymbol{\beta}_c^T \boldsymbol{x} \tag{A.20}$$

$$y_{ab}(\boldsymbol{x}) = \boldsymbol{\beta}_{ab}^T \boldsymbol{x}, \qquad y_{ac}(\boldsymbol{x}) = \boldsymbol{\beta}_{ac}^T \boldsymbol{x}, \qquad y_{bc}(\boldsymbol{x}) = \boldsymbol{\beta}_{bc}^T \boldsymbol{x} \tag{A.21}$$

$$y_{abc}(\boldsymbol{x}) = \boldsymbol{\beta}_{abc}^T \boldsymbol{x} \tag{A.22}$$

Equations (A.20), (A.21) and (A.22) demonstrate the logistic regression of the response y on the covariates $\boldsymbol{x}$ through the coefficient $\boldsymbol{\beta}$ for univariate, bivariate and trivariate interactions respectively.

One simpler non-trivial model can be obtained by setting (A.21) and (A.22) as 0:

$$y_{ab}(\boldsymbol{x}) = y_{ac}(\boldsymbol{x}) = y_{bc}(\boldsymbol{x}) = y_{abc}(\boldsymbol{x}) = 0 \tag{A.23}$$

The interpretation of Equation (A.23) is that three responses A, B, C are mutually independent, corresponding to the Conditional Independence Assumption.

Note that instead of modelling the joint distribution π, our primary interest lies in the joint transition probability p, and their derived marginal transitions η and auxiliary probabilities ζ. Diggle [2002] adapt the multinomial logistic regression framework from categorical data analysis to Markov chains. This method is applicable to univariate Markov chains of arbitrary order by fitting separate logistic regressions for each possible history. For instance, in a second-order Markov chain with a binary state space, four logistic regressions are fitted, one for each history (Y_{t-2}, Y_{t-1}), using the corresponding regression coefficients $\boldsymbol{\beta}_{00}, \boldsymbol{\beta}_{01}, \boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{11}$:

$$\begin{aligned}
\text{logit}\left(\mathbb{P}(Y_t = 2 \mid Y_{t-2} = 1, Y_{t-1} = 1)\right) &= \boldsymbol{x}^\top \boldsymbol{\beta_{11}} \\
\text{logit}\left(\mathbb{P}(Y_t = 2 \mid Y_{t-2} = 1, Y_{t-1} = 2)\right) &= \boldsymbol{x}^\top \boldsymbol{\beta_{12}} \\
\text{logit}\left(\mathbb{P}(Y_t = 2 \mid Y_{t-2} = 2, Y_{t-1} = 1)\right) &= \boldsymbol{x}^\top \boldsymbol{\beta_{21}} \\
\text{logit}\left(\mathbb{P}(Y_t = 2 \mid Y_{t-2} = 2, Y_{t-1} = 2)\right) &= \boldsymbol{x}^\top \boldsymbol{\beta_{22}}
\end{aligned}$$

which can be alternatively combined into one equation:

$$\begin{aligned}
&\text{logit}\big(\mathbb{P}(Y_t = 2 \mid Y_{t-1} = y_{t-1}, Y_{t-2} = y_{t-2})\big) \\
&= \boldsymbol{x_t}^\top \Big[\boldsymbol{\beta} + \mathbb{I}(y_{t-1} = 2)\boldsymbol{\alpha_1} + \mathbb{I}(y_{t-2} = 2)\boldsymbol{\alpha_2} + \mathbb{I}(y_{t-1} = 2,\ y_{t-2} = 2)\boldsymbol{\alpha_3}\Big] \qquad \text{(A.24)}
\end{aligned}$$

$$\text{with} \qquad \boldsymbol{\beta_{11}} = \boldsymbol{\beta}, \quad \boldsymbol{\beta_{12}} = \boldsymbol{\beta} + \boldsymbol{\alpha_1}, \quad \boldsymbol{\beta_{21}} = \boldsymbol{\beta} + \boldsymbol{\alpha_2}, \quad \boldsymbol{\beta_{22}} = \boldsymbol{\beta} + \boldsymbol{\alpha_1} + \boldsymbol{\alpha_2} + \boldsymbol{\alpha_3}$$

In Diggle [2002]'s framework, conditions based on past states, such as those at times $t-1, t-2, \ldots$, can be incorporated into the joint probabilities. Our approach resembles this extension but differs in structure: In their model, the conditioning variables are time-lagged states of the same chain, while in contrast, for a first-order MMC, we build covariates using the states from time $t-1$ across all chains, capturing marginal effects from different chain subgroups.

Hence, instead of a single covariate vector $\boldsymbol{x}$, we use a design matrix $\boldsymbol{\mathcal{X}}$, where each row corresponds to a covariate vector associated with a particular previous joint state. This setup has been adopted in modelling multivariate Markov chains by Colombi and Giordano [2012] and Colombi et al. [2013], based on marginal contrasts in the generalised marginal logistic regression framework.

However, under the GM logistic framework, as argued by Lupparelli et al. [2009], the logistic coefficients $\boldsymbol{\beta}$ resides in an unconstrained parameter space $\mathbb{R}^{s^k \times (s^k - 1)}$ only when

there are two chains ($k = 2$). When it turns to 3 or more chains, there is no closed-form inverse mapping from marginal contrasts η back to joint transition matrix P. For instance, in a toy example with $k = 3$ chains and a binary state space $\mathcal{S} = \{1, 2\}$:

$$\boldsymbol{\eta} = (\eta^{\{1\}}, \eta^{\{2\}}, \eta^{\{3\}}, \eta^{\{1,2\}}, \eta^{\{1,3\}}, \eta^{\{2,3\}}, \eta^{\{1,2,3\}}) = (0, 0, 0, 2, 2, 2, 2),$$

admits no nonnegative joint transition probabilities p corresponding to those marginal contrasts, which implies that this marginal contrasts setting by [Glonek and McCullagh, 1995] generally does not offer a fully unconstrained parameter spaces. In contrast, for both Univariate Marginal and Hierarchical Marginal models, which we will show later, the parameter space $\boldsymbol{\beta}$ is fully unconstrained, covering all of $\mathbb{R}^{s^k \times (s^k - 1)}$. This is a crucial reason why we choose to use these two models as an alternative to marginal contrasts.

Appendix B

Simulated Data

B.1 Simulation Study in Chapter 4

B.1.1 Univariate Marginal Model

In this section, we present trace plots for the estimation of the first regression coefficient β_1, associated with the first auxiliary probability

$$\zeta_1 = \eta_1 = \mathbb{P}\big(Z_1(t) \mid \boldsymbol{Z}(t-1)\big)$$

in the first repetition of the experiment (i.e. 5000 MCMC iterations) as an example. The trace plots are shown under four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures.

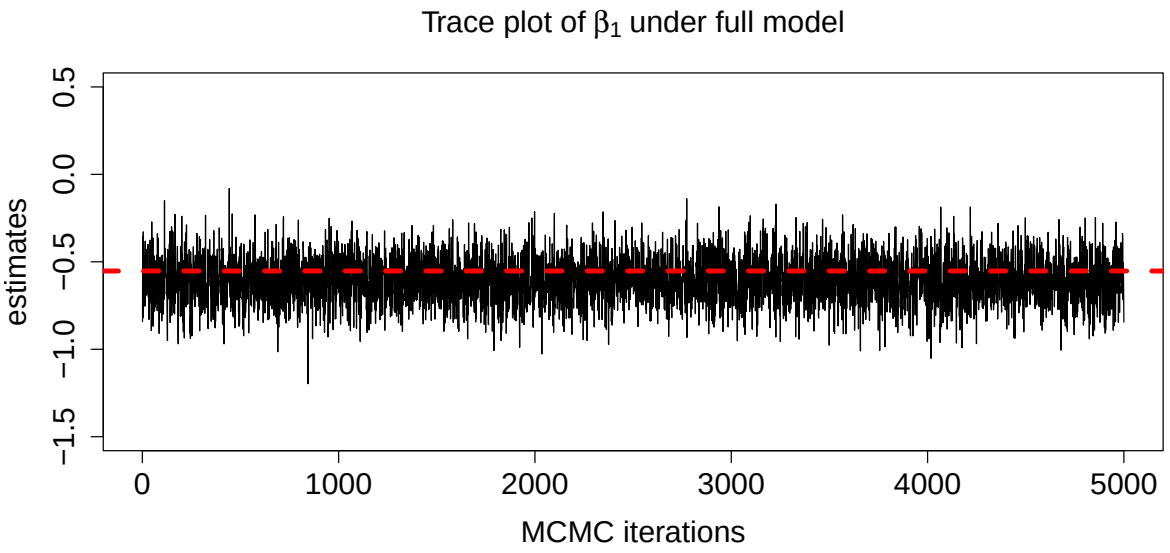


Figure B.1: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the full model, with the true value indicated by the red dotted line.

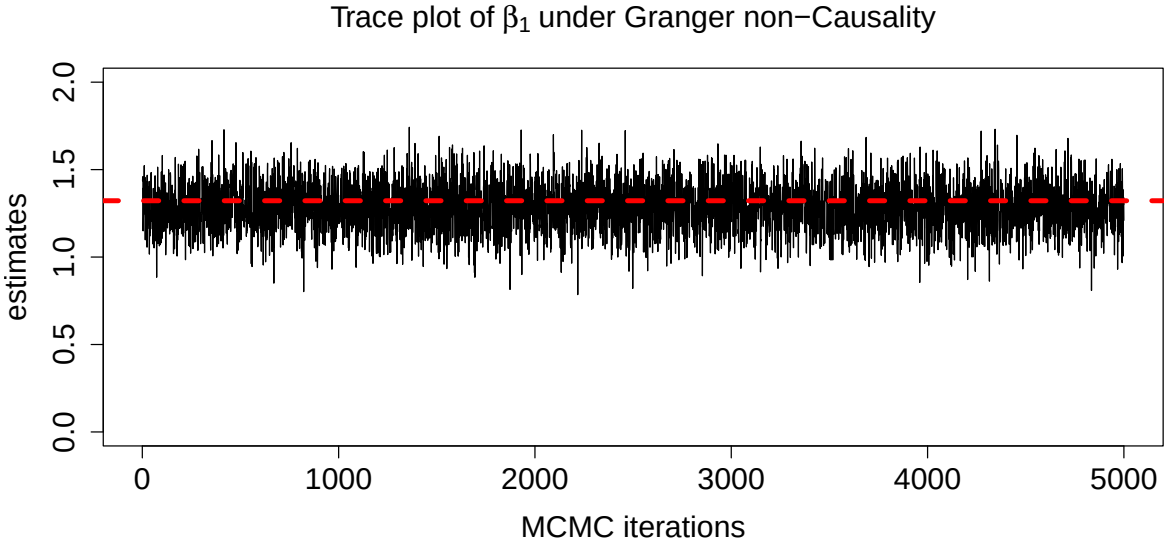


Figure B.3: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the Granger non-Causality $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1)$, with the true value indicated by the red dotted line.

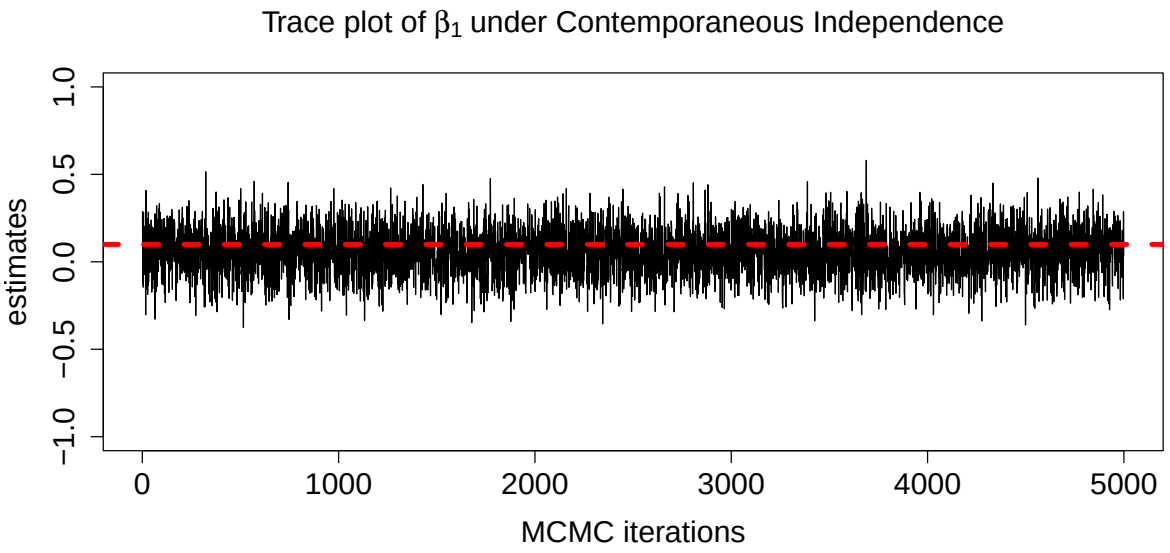


Figure B.2: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the contemporaneous independence $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$, with the true value indicated by the red dotted line.

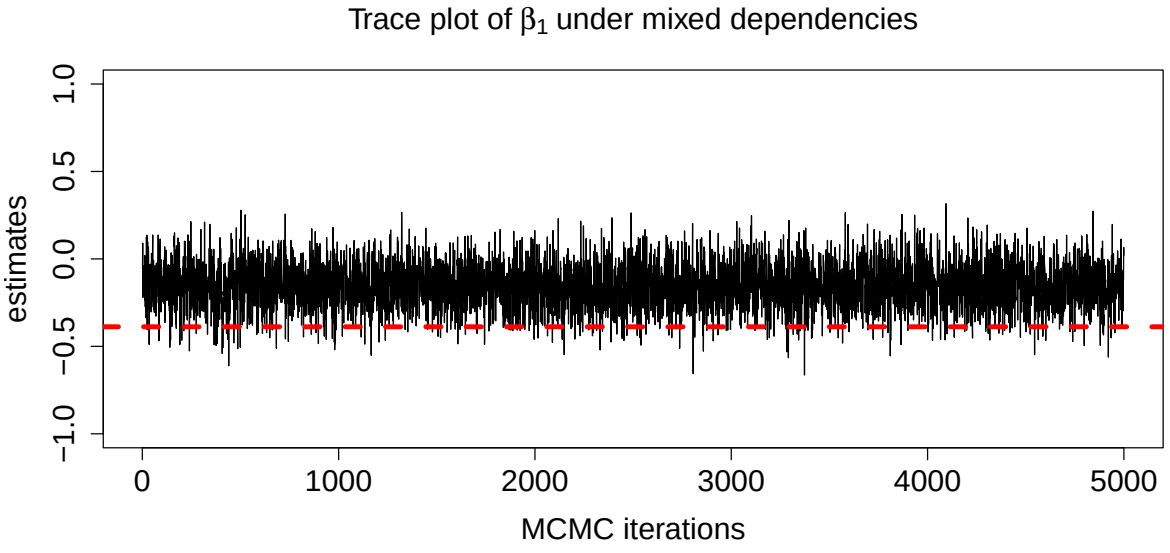


Figure B.4: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the mixed dependencies $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$ (contemporaneous independence) and $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1)$ (Granger non-Causality), with the true value indicated by the red dotted line.

B.1.2 Hierarchical Marginal Model

Similar as previous, we present trace plots for the estimation of the first regression coefficient β_1, associated with the marginal transition probability

$$\eta_1 = \mathbb{P}\big(Z_1(t) \mid \boldsymbol{Z}(t-1)\big)$$

in the first repetition of the experiment (i.e. 5000 MCMC iterations) as an example. The trace plots are shown under four scenarios: (I) full model, (II) contemporaneous independence, (III) Granger non-causality, and (IV) mixed dependency structures.

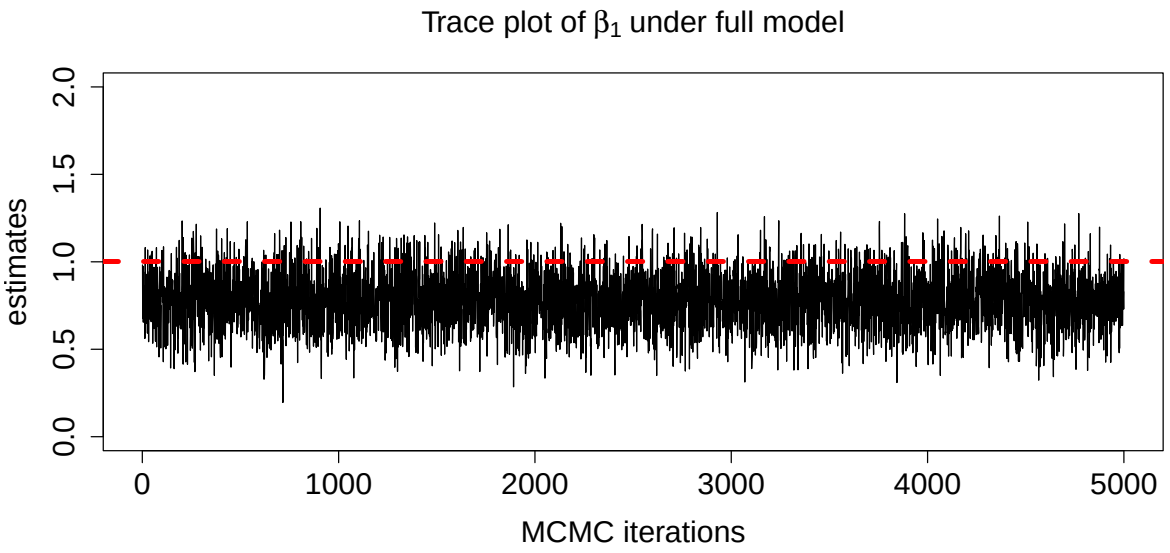


Figure B.5: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the full model, with the true value indicated by the red dotted line.

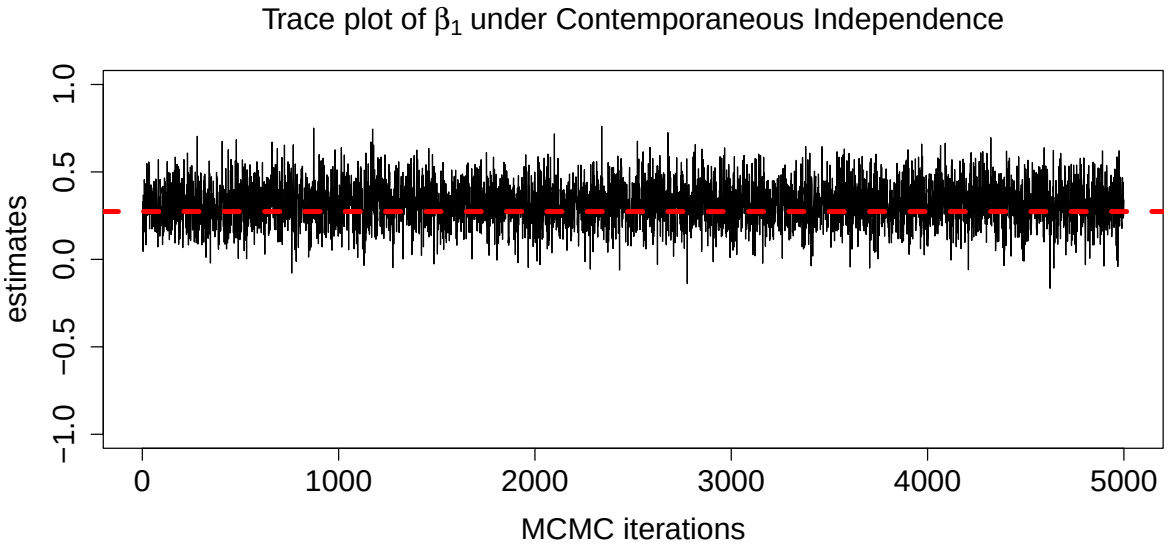


Figure B.6: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the contemporaneous independence $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$, with the true value indicated by the red dotted line.

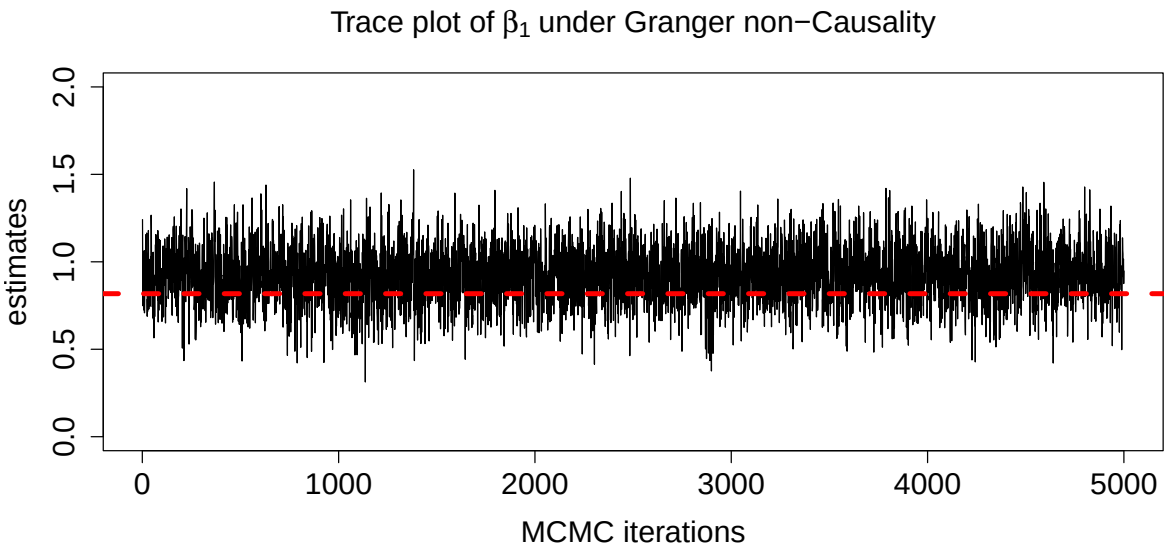


Figure B.7: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the Granger non-Causality $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1)$, with the true value indicated by the red dotted line.

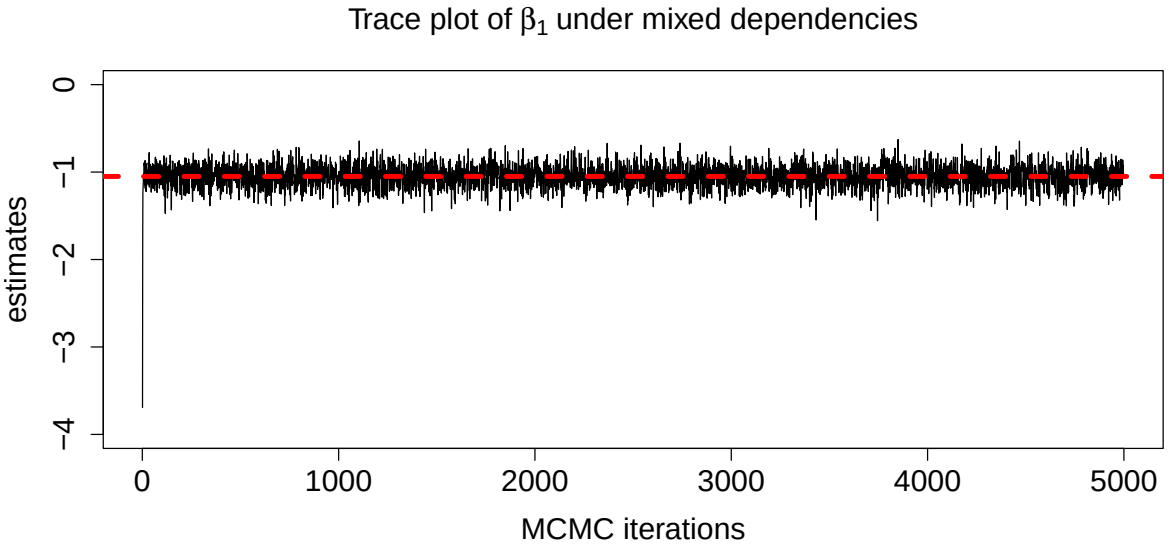


Figure B.8: Trace plot of the regression coefficients $\boldsymbol{\beta}_1$ for η_1 over 5000 MCMC iterations under the mixed dependencies $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t) \mid \boldsymbol{Z}(t-1)$ (contemporaneous independence) and $Z_1(t), Z_2(t) \perp\!\!\!\perp Z_3(t-1) \mid Z_1(t-1), Z_2(t-1)$ (Granger non-Causality), with the true value indicated by the red dotted line.

B.2 Simulation Study in Chapter 5

B.2.1 Prior and Posterior for Emission Parameters

For the emission parameters $\theta_k = (\lambda_k, c_k)$ governing the intensity, we assign independent conjugate Gamma priors with shape parameter 20 and rate parameter 1. Specifically, for each $k = 1, 2, 3$:

$$\lambda_k \sim \Gamma(20, 1) \qquad\qquad c_k \sim \Gamma(20, 1)$$

With the resulting conjugate Gamma distribution for the posterior:

$$\lambda_k \sim \Gamma\Big(20 + \sum_{t:Z_k(t)=1} X_k(t),\ 1 + \sum_{t=1}^{T} \mathbb{I}(Z_k(t) = 1)\Big)$$
$$c_k \sim \Gamma\Big(20 + \sum_{t:Z_k(t)=2} X_k(t),\ 1 + \sum_{t=1}^{T} \mathbb{I}(Z_k(t) = 2)\Big)$$

B.2.2 Estimation of Emission Parameters

The following Table B.1 summarize the estimated emission parameters$\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{c})$ by posterior median across all the repetitive experiments against the ground truth specified in Table 5.2:

	λ_1	λ_2	λ_3	c_1	c_2	c_3
Truth	20	15	15	20	15	20
FB	20.09	14.94	14.85	19.78	14.94	20.13
GIFB	20.09	14.94	14.86	19.80	14.90	20.11

Table B.1: Estimated Emission Parameters $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{c})$ against the Truth

Here, λ_k and c_k denote the emission parameters for chain k. The Table B.1 shows that most estimates are close to the true values, demonstrating the ability of both FB and GIFB to recover the emission parameters.

As an illustration, we take λ_1 and c_1 as examples and present their trace plots from the Gibbs samplers under both FB and GIFB across 2000 MCMC iterations within the first experiment, under the contemporaneous independence.

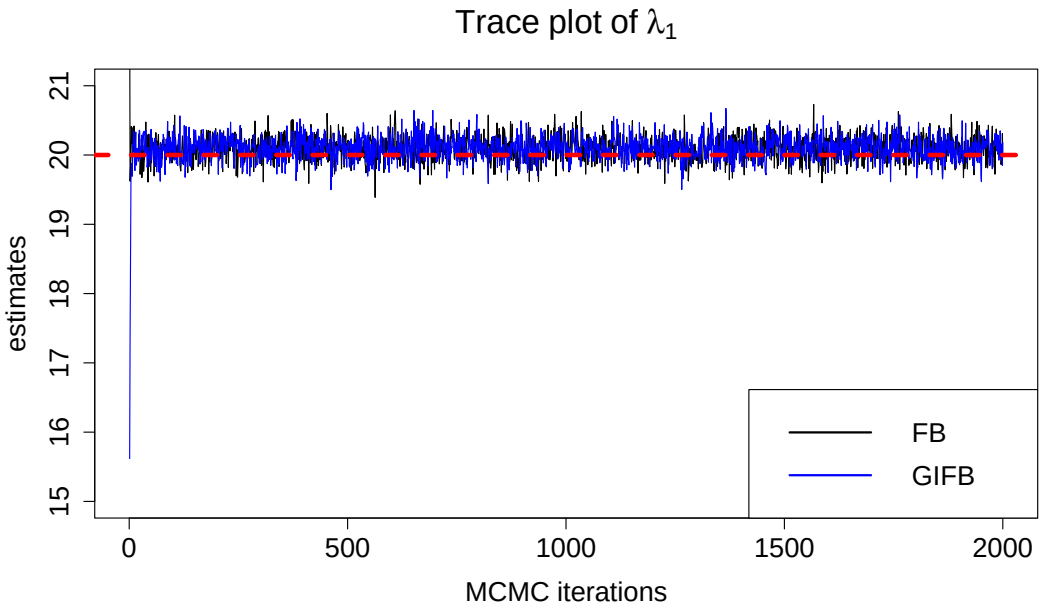


Figure B.9: Trace plot of λ_1 by FB and GIFB

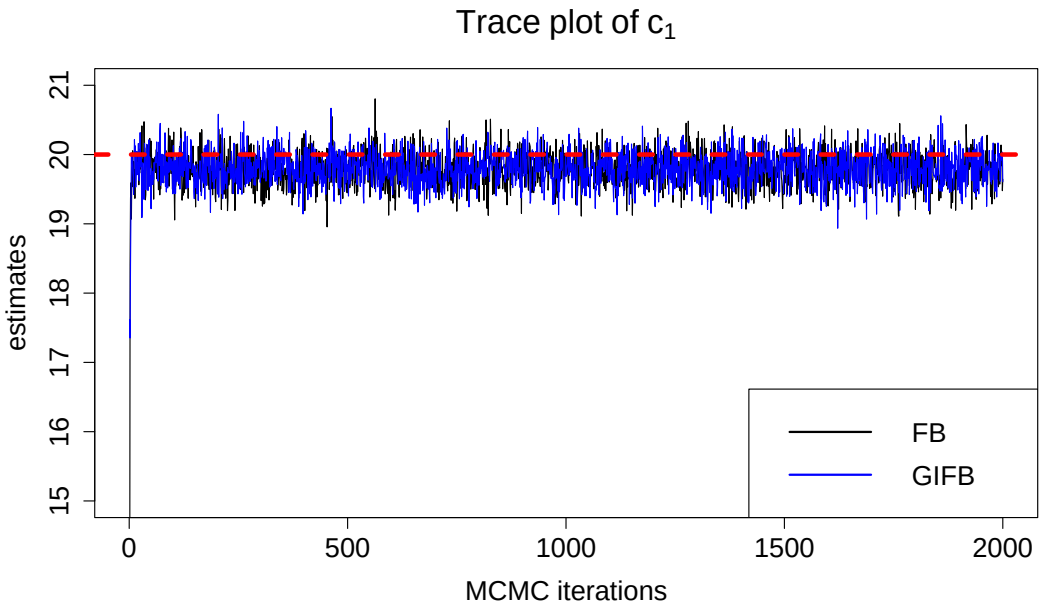


Figure B.10: Trace plot of c_1 by FB and GIFB

B.2.3 Regression Coefficients

Similarly, we take the first regression coefficient β_1 within $\boldsymbol{\beta}_1$ for the univariate marginal of chain group $g = \{1\}$ as an example, and present its trace plots from the Gibbs samplers under FB and GIFB across 2000 MCMC iterations, within the first experiment:

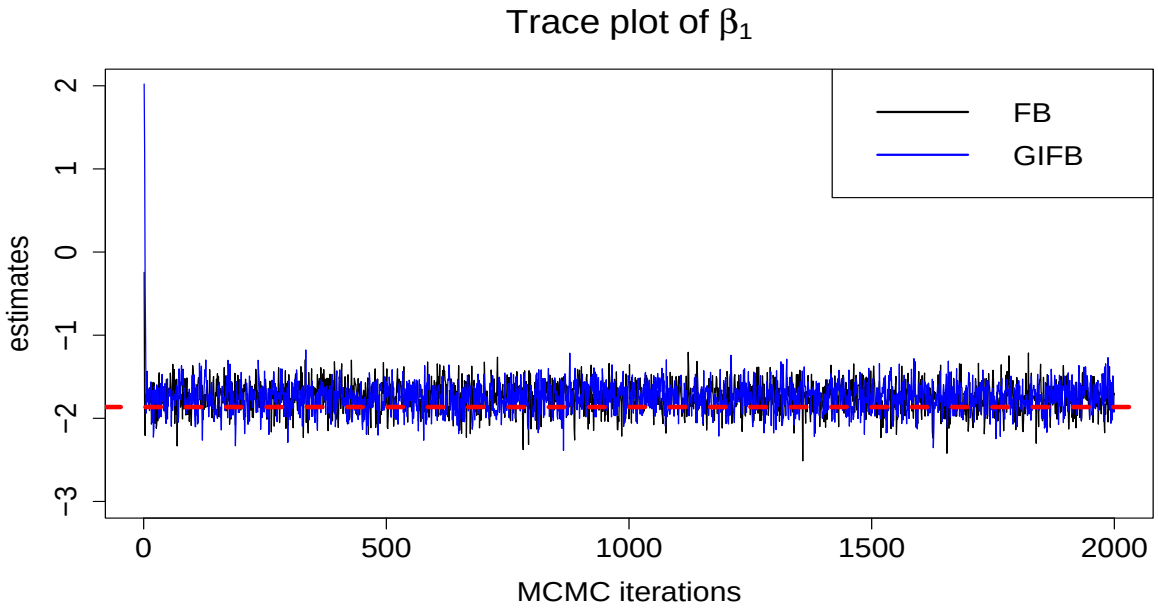


Figure B.11: Trace plot of β_1 by FB and GIFB

Appendix C

Twitter Data

C.1 Preprocess Twitter Data

As is shown in table 5.4, each tweet provides us with both textual (what has been said in text) as well as temporal (when it has been sent) information. Usually, users are inclined to express their emotions through tweets, to be specific, they would convey their gratitude if the railway brings them to their destination on time and show aversion if the trains are delayed. Consequently, we can categorize the emotions expressed in tweets into two polarities: positive and negative.

As mentioned in introduction, what really interests us is the counts of the delayed tweets. The key step to obtain the counts of delayed tweets is through the sentiment classifier. Before classifying tweets, we first manipulate the data with a preprocessing step, that is to erase the irrelevant information contained in tweets [Yang and Anwar, 2016]. Our preprocessing mainly involves three parts:

1. General preprocessing: remove hashtags, links, single characters, and Twitter handlers and substitute multiple spaces with a single space.

2. Remove stop words: filter out some insignificant words according to the dictionary provided in Word Net [Baccianella et al., 2010]

3. Lemmatization: transform the words into their original form, which refers to the

present tense for a verb and singular form for a noun

After that, we classify the tweets into these two genres according to the Lexicon-based approach, which uses the predefined list of words associated with the specific sentiment [Sarlan et al., 2014]. We specify the domain focused Lexicon DF in our railway delay context as follows:

$$DF = \{\text{not moving, standstill, slow, disrupt, stuck, delay, wait, late, cancel, crowd, ...}\}$$

Next, We compute the Lexicon score for each tweet using algorithm 10. By comparing the resulting score of each tweet to the threshold, we can filter out the delayed tweets effectively.

Algorithm 10 Compute Lexicon Score for the Tweets

Inputs: A Preprocessed Tweet T in text with the words $\{w_1, w_2, \dots\}$ and a Domain-Focused Lexicon L_{DF} related to railway delays with score S_{DF}

Outputs: The score s of the tweet according to the L_{DF}

Initialize the score s
for all the words $w_i \in T$ such that $w_i \in SWN \cup DF$ **do**
  if $w_i \in L_{\text{DF}}$ **then**
    $s = s + S_{DF}(w_i)$
  end if
end for

C.1.1 Priors for the Emission Parameters

In the Bayesian framework, inference begins with specifying prior distributions. We assign the following priors to the model parameters (for notational convenience, we omit the subscript k for chain 1 and 2; identical priors are used for both chains):

1. We set the prior of the shifting constant c to be Exponential with rate β as:

$$\pi(c) = \beta e^{-\beta c}$$

2. Based on μ_1 and μ_2 being the peak for the morning and evening respectively, we choose the prior of them as the ordered statistics of two independent uniform distributions as:

$$U_1, U_2 \sim U(0, 24)$$
$$\mu_1 = U_{(1)}, \qquad \mu_2 = U_{(2)}$$

Where $U_{(1)}$ and $U_{(2)}$ are the ordered statistic of U_1 and U_2

With the joint distribution $\pi(\mu_1, \mu_2)$.

$$\pi(\mu_1, \mu_2) = \frac{1}{288} I_{\{0 \leq \mu_1 \leq 24\}} I_{\{0 \leq \mu_2 \leq 24\}} I_{\{\mu_1 < \mu_2\}}$$

Where $I_{\{\cdot\}}$ is the indicator function. The first two come from the uniform distribution and the last one comes from the order statistics since we require $\mu_1 < \mu_2$.

3. According to the average tweets per day throughout the exploratory analysis, we choose the prior of c_1, c_2, and c_3 as three independent uniform distribution as:

$$c_1, c_2 \sim U(0, 150)$$
$$c_3 \sim U(0, 50)$$

C.1.2 The Posterior Distributions

Once we get the prior and likelihood (See Equation 5.2), it will be fairly easy to work out the posterior distribution according to the Bayesian formula as:

$$\begin{aligned}\pi(\boldsymbol{\theta} \mid \boldsymbol{Z}, \boldsymbol{X}, \boldsymbol{\beta}) \propto & \pi(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\ \propto & \pi(\boldsymbol{X} \mid \boldsymbol{Z}, \boldsymbol{\theta})\pi(c)\pi(c_1, c_2, c_3)\pi(\mu_1, \mu_2)\end{aligned}$$

We construct a Gibbs sampler that iteratively samples from each of the emission parameters from their full conditionals. The skeleton of this is given in following Algorithm 11:

Algorithm 11 Gibbs Sampler for the Emission Parameters

Inputs: The hidden states $\boldsymbol{Z}$, observations $\boldsymbol{X}$ and the emission function f up to the distributional form in Equation 5.35, number of iterations n

Outputs: The samples of the emission parameters $\boldsymbol{\theta}$

for $i = 1, 2, \ldots, n$ **do**
  for each chain $k = 1, 2$ **do**
    Sample $c \sim \pi(c \mid X_k, Z_k, \boldsymbol{\theta}_{-c})$
    Sample $\mu_1, \mu_2 \sim \pi(\mu_1, \mu_2 \mid X_k, Z_k, \boldsymbol{\theta}_{-\mu_1,\mu_2})$
    Sample $c_1, c_2, c_3 \sim \pi(c_1, c_2, c_3 \mid X_k, Z_k, \boldsymbol{\theta}_{-c_1,c_2,c_3})$
  end for
end for

For all parameter updates, we employ the random walk Metropolis–Hastings , as summarized in Algorithm 12.

C.1.2.1 Full conditional distribution of c

In this case, we know all the parameters that governing the intensity function $\lambda(t)$, and we are supposed to find the shifting constant c. If we could set

$$T_1 = \{t : Z(t) = 1\}$$

$$T_2 = \{t : Z(t) = 2\}$$

We obtain the full conditional as:

$$\begin{aligned}\pi(c \mid X_k, Z_k, \boldsymbol{\theta}_{-c}) \propto & \pi(c)\ \pi(X_k \mid Z_k, \boldsymbol{\theta}) \\ \propto & e^{-\beta c} \prod_{t_2 \in T_2} (\lambda(t_2) + c)^{X(t_2)} e^{-(\lambda(t_2)+c)} \prod_{t_1 \in T_1} \lambda(t_1)^{X(t_1)} e^{-\lambda(t_1)} \\ \propto & e^{-(|T_2|+\beta)c} \prod_{t_2 \in T_2} (\lambda(t_2) + c)^{X(t_2)}\end{aligned}$$

Where $|T_1|$ stands for the number of elements in set T_1

C.1.2.2 Full conditional distribution of $\boldsymbol{\mu} = (\mu_1, \mu_2)$

In this section, we would like to find the full conditional of $\boldsymbol{\mu} = (\mu_1, \mu_2)$. Specifically for this section, for $t = 1, \ldots, T$, we let

$$\lambda_t(\mu_1, \mu_2) = \int_{t-1}^{t} \lambda(u; \mu_1, \mu_2, c_1, c_2, c_3) du$$

$$\lambda'_t\big(\mu_1, \mu_2, Z_k(t)\big) = \lambda_t(\mu_1, \mu_2) + c \cdot \mathbb{I}\big(Z_k(t) = 2\big)$$

We are able to obtain the full conditional of $\boldsymbol{\mu}$ as:

$$\pi(\boldsymbol{\mu} \mid X_k, Z_k, \boldsymbol{\theta}_{-\boldsymbol{\mu}}) \propto I_{\{0 \le \mu_1 < \mu_2 \le 24\}} \left[e^{-\sum_{t=1}^{T} \lambda'_t\left(\mu_1, \mu_2, Z_k(t)\right)} \prod_{t=1}^{T} \lambda'_t\big(\mu_1, \mu_2, Z_k(t)\big)^{X_k(t)} \right]$$

C.1.2.3 Full conditional distribution of c_1, c_2, c_3

Follow the same recipe in section C.1.2.2, we obtain the posterior of c_1, c_2 and c_3 as:

$$\pi(c_1, c_2 \mid X_k, Z_k, \boldsymbol{\theta}_{-c_1,c_2}) \propto I_{\{0 \le c_1 < c_2 \le 150\}} \left[e^{-\sum_{t=1}^{T} \lambda'_t\left(c_1, c_2, Z_k(t)\right)} \prod_{t=1}^{T} \lambda'_t\big(c_1, c_2, Z_k(t)\big)^{X_k(t)} \right]$$

$$\pi(c_3 \mid X_k, Z_k, \boldsymbol{\theta}_{-c_1,c_2}) \propto I_{\{0 \le c_3 \le 50\}} \left[e^{-\sum_{t=1}^{T} \lambda'_t\left(c_3, Z_k(t)\right)} \prod_{t=1}^{T} \lambda'_t\big(c_3, Z_k(t)\big)^{X_k(t)} \right]$$

We sample this posterior by Random-Walk Metropolis-Hastings Algorithm 12. Note that we use the bivariate normal proposal for $\boldsymbol{\mu} = (\mu_1, \mu_2)$ and (c_1, c_2).

C.1.3 Random-Walk Metropolis-Hastings Algorithm

For our unimodal full conditional, we use the normal proposal:

$$q(c^\star \mid c_{i-1}) \sim N(c_{i-1}, \sigma)$$

Where $N(c_{i-1}, \sigma)$ is normal with mean c_{i-1} and standard deviation σ. This σ is computed by the inverse of the Hessian. If we consider random-walk Metropolis-Hastings where $q(c^\star \mid c_{i-1}) = q(c_{i-1} \mid c^\star)$ and

$$\alpha = \frac{\pi\big(c^\star \mid X_k, Z_k, \lambda(t)\big)\; q(c^\star \mid c^{i-1})}{\pi\big(c^{i-1} \mid X_k, Z_k, \lambda(t)\big)\; q(c^{i-1} \mid c^\star)} = \frac{\pi\big(c^\star \mid X_k, Z_k, \lambda(t)\big)}{\pi\big(c^{i-1} \mid X_k, Z_k, \lambda(t)\big)} \tag{C.1}$$

As for $\boldsymbol{\mu}$, we use a bivariate normal proposal as:

$$q(\boldsymbol{\mu}^\star \mid \boldsymbol{\mu}_{i-1}) \sim N_2(\boldsymbol{\mu_{i-1}}, \Sigma)$$

Where $N_2(\boldsymbol{\mu_{i-1}}, \Sigma)$ is bivariate normal with mean $\boldsymbol{\mu}_{i-1}$ and standard deviation Σ. The same bivariate proposal is utilized for (c_1, c_2) as well.

Algorithm 12 Metropolis Hastings Algorithm

Inputs: The full condition $\pi(c \mid \cdot)$, some suitable proposal $q(\cdot \mid \cdot)$ and the number of iterations n_{iter}

Outputs: A Markov Chain $c_1, c_2, \ldots, c_{n_{\text{iter}}}$

Select the initial value c_0
for $i = 1, 2 \ldots, n_{\text{iter}}$ **do**
  Draw new candidate $c^\star \sim q(c^\star \mid c_{i-1})$
  Compute the ratio α in equation C.1
  Draw $u \sim \mathrm{U}(0, 1)$
  if $\alpha > u$ **then**
    Accept $c^\star$ and set $c_i = c^\star$
  else
    Reject $c^\star$ and set $c_i = c_{i-1}$
  end if
end for

C.1.4 Estimates of the Emission Parameters

The posterior median estimates of the emission parameters, computed after a burn-in of the first 500 samples, are reported in Table C.1.

Estimated Parameters	μ_1	μ_2	c_1	c_2	c_3	c
k = 1	8.49	18.55	117.75	56.98	14.23	23.98
k = 2	8.49	18.32	76.93	57.43	10.34	41.37

Table C.1: Estimates of the Emission Parameters

We also present the trace plot for c, c_1 and μ_1 of the first chain for instance, for visual inspection of the convergence of MCMC.

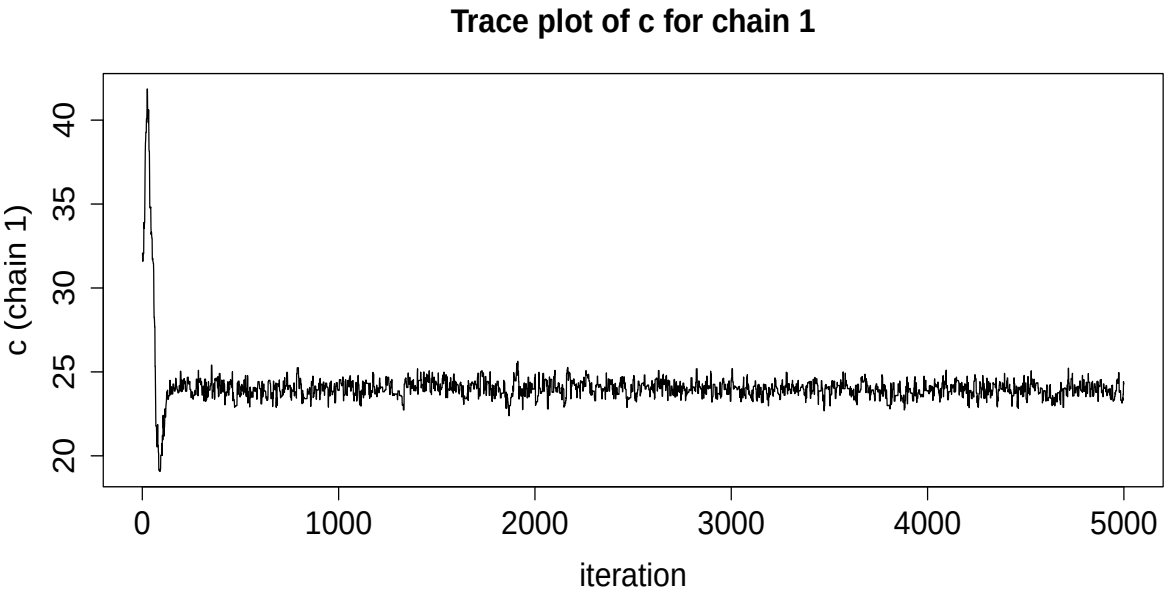


Figure C.1: Trace plot of c for chain 1

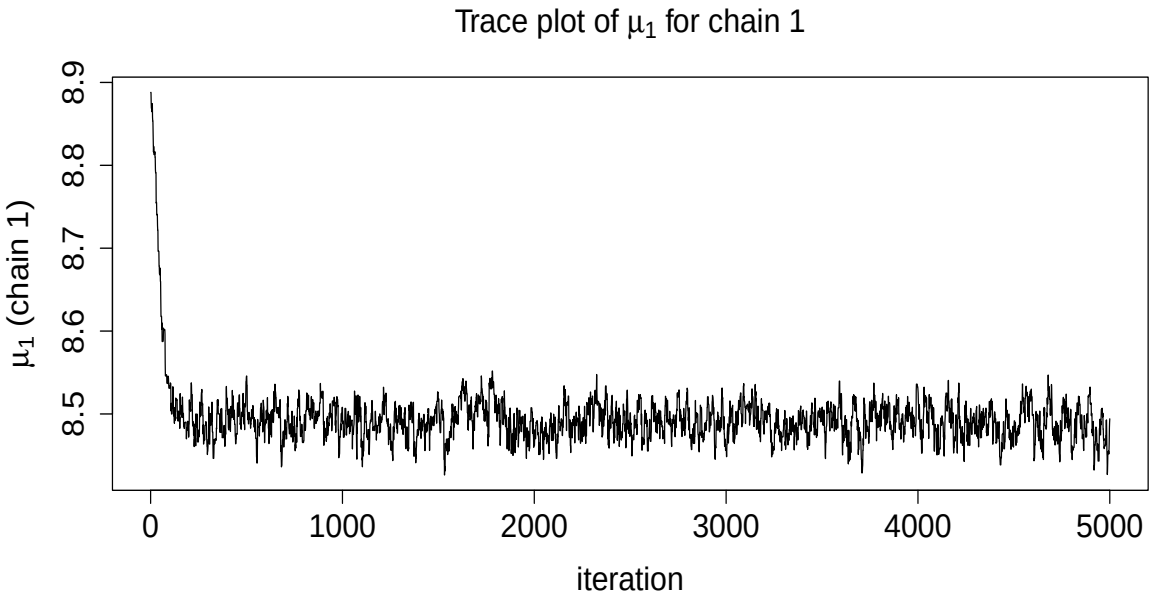


Figure C.2: Trace plot of μ_1 for chain 1

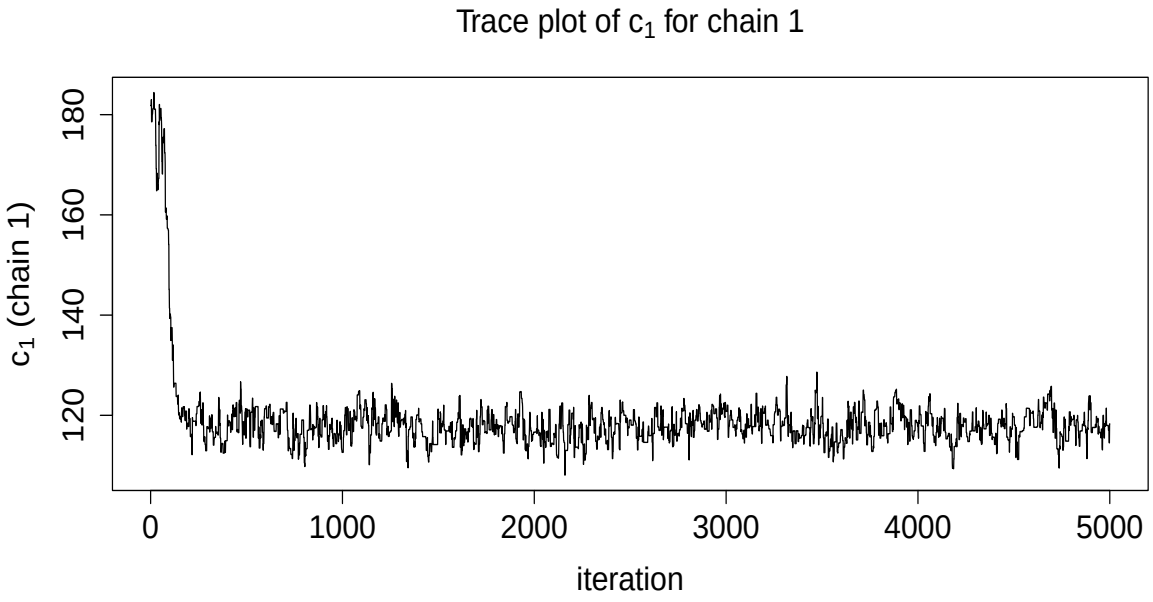


Figure C.3: Trace plot of c_1 for chain 1