Bayesian Perspectives on Statistical Modelling

Nicholas G. Polson

Thesis submitted to the University of Nottingham for

the degree of Doctor of Philosophy

October 1988

To my parents.

Acknowledgements

I would like to thank my supervisor, Professor A.F.M. Smith, for all his advice and encouragement throughout this thesis. This research was financed by a grant from the Science and Engineering Research Council, for which I am grateful.

ABSTRACT

This thesis explores the representation of probability measures in a coherent Bayesian modelling framework, together with the ensuing characterisation properties of posterior functionals.

First, a decision theoretic approach is adopted to provide a unified modelling criterion applicable to assessing prior-likelihood combinations, design matrices, model dimensionality and choice of sample size. The utility structure and associated Bayes risk induces a distance measure, introducing concepts from differential geometry to aid in the interpretation of modelling characteristics.

Secondly, analytical and approximate computations for the implementation of the Bayesian paradigm, based on the properties of the class of transformation models, are discussed.

Finally, relationships between distance measures (in the form of either a derivative of a Bayes mapping or an induced distance) are explored, with particular reference to the construction of sensitivity measures.

CONTENTS

		PAGE
Chapter 1	INTRODUCTION	1
	(1.1) The "what if" principle	2
	(1.2) Choice of utility function	3
	(1.3) Choice of class $\mathscr{C}$	4
	(1.4) Nonparametric versus parametric modelling	6
	(1.5) Model elaboration	6
	(1.6) Differential geometry	7
	(1.7) Profile likelihoods	8
	(1.8) Modelling characteristics	8
	(1.9) Outline of the thesis	9
Chapter 2	UTILITY STRUCTURES FOR PROBABILITY BELIEFS	10
	(2.1) Decision problems on $\mathscr{P}$	10
	(2.1.1) Reporting the prior-posterior pair	11
	(2.1.2) Approximating P by Q	12
	(2.1.3) Simultaneous approximation of the pair (F, G)	12
	(2.1.4) Variational solution to choice of Q	13
	(2.1.5) A non-local utility structure	17
	(2.2) Differential geometry in statistics	18
	(2.2.1) α-connections and α-geodesics	19
	(2.2.2) Geometry of the space of distributions	20
	(2.2.3) Geometry of the probability simplex	21
	(2.3) Discussion	23
Chapter 3	BAYESIAN MODEL CHOICE : A DECISION THEORETIC CRITERION	24
	(3.1) Asymptotic information gain	26
	(3.1.1) Information on a cost scale with application to selection of dimensionality	27
	(3.1.2) Risk under an incorrect model specification	28
	(3.1.3) Example : The location problem	29

(3.1.4) Further examples 33
(i) Normal and double exponential 34
(ii) Logistic 35
(iii) Application to the scale case 35
(iv) Compact parameter spaces 36
(v) Finite mixture models 37
(3.2) Calculation of the Bayes risk for a modelling framework 38
(3.2.1) Asymptotic information gain 39
(3.2.2) Determination of $p(\lambda)$ 39
(3.2.3) Examples : Scale mixtures of normality 40
(i) Justification of the t-family 40
(ii) Justification of the double exponential 41
(iii) Justification of the exponential power family 42
(3.2.3) Discrete case 43
(3.2.4) Infinite discrete case 44
(3.2.5) Asymptotic information gain, orthogonality and independence 45
(3.2.6) Comparison of experiments 46
(3.3) Application of Ressel for the construction of model elaborations 46
(3.4) Connecting two distributions 49
(3.4.1) Utility structure 50
(3.4.2) Application to prior elaboration 51
(3.4.3) Application to the location-scale family 51
(3.5) Approximating statistical models with flexible families 52
(3.5.1) Decision-theoretic setting for projecting $\mathcal{P}_\Omega$ onto $\mathcal{P}_\Lambda$ 52
(3.5.2) Model choice on $\mathbb{R}^+$ 53
(3.5.3) Examples of power transformations 55
(3.5.4) Construction of a family of transformations 55
(3.6) A decision-theoretic approach to the design of experiments 56
(3.6.1) A- and D- optimality from a decision-theoretic perspective 57
(i) A-optimality 57
(ii) D-optimality 58
(iii) Design criterion under a non-local utility structure induced by a loss structure 59
(3.6.2) Reference priors and the design of experiments 59
(3.7) Discussion 61

Chapter 4 ELIMINATION OF NUISANCE PARAMETERS; REFERENCE PRIORS 63

(4.1) Profile likelihoods 64
(4.1.1) Group transformation models 65
(4.1.2) Justification of the modified profile likelihood 67
(4.1.3) Bayesian paradigm 68
(4.1.4) Approximate computations 68
(4.1.5) Examples 69
(4.2) Reference priors 69
(4.2.1) Examples : (i) A moment class 69
(4.2.2) Strong inconsistencies and amenability 70
(4.2.3) Examples
(i) Exponential connection 70
(ii) Hierarchical models 71
(4.3) Discussion 72

Chapter 5 POSTERIOR FUNCTIONALS CHARACTERISING PRIOR MEASURES IN THE EXPONENTIAL FAMILY 73

(5.1) Identities in the exponential family 73
(5.2) Functional minimisation of the Bayes risk 75
(5.2.1) Links with inference for a location parameter 76
(5.2.2) Prior elaboration 78
(5.2.3) Representation of Fisher's information 80
(5.2.4) Likelihood elaboration 81
(5.3) Examples with a normal likelihood 83
(5.3.3) Asymptotic behaviour of expectations 86
(5.4) Characterisations involving moment generating functions 86
(5.4.1) Application of Ralescu and Ralescu 88
(5.4.2) Application of Diaconis and Ylvisaker 88
(5.4.3) Results of Goldstein 90
(5.5) Associated differential equations for moments and priors 91
(5.6) Linear regression in X 93
(5.7) Further constraints on posterior moments characterising priors 94
(5.7.1) Characterisations involving ratios 96
(5.8) Application to other likelihoods 97

(5.9) Justification of differentiation under the integral sign 99
(5.10) Discussion 100

Chapter 6 DERIVATIVES; DISTANCES; SENSITIVITY 102

(6.1) Derivative of the prior-posterior map 102
(6.2) Relationship with Bayes factors for nested models 103
(6.3) Application to posterior functionals and model choice 104
(6.3.1) Application to model choice and nuisance parameters 105
(6.4) An inequality between discrimination information and variational distance 105
(6.5) A survey of results concerning distance measures 107
(6.6) Application to influence and outlyingness 107
(6.7) Application in decision theory 109
(6.8) Application to moments of Bayes factors 110
(6.9) Discussion 112

REFERENCES

Chapter 1 : Introduction

In de Finetti's subjectivist account of Probability theory, the concept of probability is not one that is independent of the observer but one that quantifies the observer's personal uncertainty about events in a complex world. Nevertheless, the observer must have an operational framework for its measurement. This thesis addresses the problem of the specification of *a priori* beliefs and related aspects of a Bayesian statistical modelling framework.

A foundational result is the de Finetti representation theorem allowing the study of the interrelationship between different observers' world views. The general theorem (see, for example Hewitt and Savage (1955), Diaconis and Freedman (1986a), Ressel (1985)) explains how coherent Bayesians act when confronted with an infinitely exchangeable sequence of observable random variables, denoted by $\boldsymbol{X} = \{X_1, X_2, \ldots\}$. The symmetry condition of infinite exchangeability expresses the belief that, for each n, the joint probability measures concerning $\boldsymbol{X}$ are invariant under the action of S_n, the symmetric group on n letters. The concept was initiated by Haag (1924) and formalised by de Finetti (1931, 1937). The theorem determines the structural nature of the beliefs about X as a mixture decomposition by means of a unique measure $\mu(\cdot)$ as follows,

$$P_\mu(A \times B) = \int_A F^\infty(B)\mu(dF)$$

where P_μ denotes the joint beliefs under mixing measure $\mu(\cdot)$ over the space of distributions, $\mathcal{P}$, such that conditional on $F \in \mathcal{P}$ they are independent.

A special case of the above, formulating the parametric modelling framework, is obtained by restricting the measure μ to a finite dimensional subset of $\mathcal{P}$. Consider a family of probabilities, $\{Q_\theta : \theta \in \Theta\}$ indexed by $\Theta \in \mathbb{R}^k$, together with a prior measure μ over Θ. Let Q_θ^∞ denote the infinite product measure on X^∞ for which $\{X_1, X_2, \ldots\}$ are independent with common distribution Q_θ. The general de Finetti representation becomes

$$P_\mu(A \times B) = \int_A Q_\theta^\infty(B)\mu(d\theta) \tag{1.1.1}$$

that is, all observers agree on the same conditional model Q_θ^∞. Dawid (1982, 1986) defines this as the conditional I-model interpreted within the notion of an intersubjective model. The de Finetti theorem thus justifies the specification of a statistical model (i.e. the joint measure $P_\mu(\cdot)$) in terms

of the conventional separation into prior-likelihood combinations, where it is *as if* the X_i are an independent sample, conditional on Q_θ. Moreover, if the predictive distribution for $(X_{n+1}, X_{n+2},\ldots)$ given $(X_1,\ldots, X_n)$, denoted by $P_\mu{}^n$, is considered, then Bayes theorem emerges from the decomposition,

$$P_\mu{}^n(B) = \int_A Q_\theta^\infty(B)\mu_n(d\theta)$$

where the measure $\mu_n(\cdot)$ denotes the posterior measure, as determined by Bayes theorem.

Clearly the structure imposed via exchangeability on the joint distribution depends heavily on the structure of the sample space, the intersubjective model Q_θ^∞ sometimes having a definite functional form: for example, {0, 1} exchangeable random variables determine the Bernoulli model. However, the de Finetti representation theorem holds for quite abstract topological sample spaces; the caveat, from a pragmatic specification perspective, is that in general the mixing measure $\mu(\cdot)$ assigns its mass over the whole space of probability measures, $\mathcal{P}$.

Sometimes careful interpretation of the respective components of the unobservables allows us to identify parameters of interest that possess a physical meaning. These parameters will be termed extrinsic (Dawid (1985)). The specification of *a priori* beliefs for such a parameter then directly represents our belief about the physical quantity, θ (Lindley (1972)). Thus, throughout this thesis, we shall assume that the class $\mathcal{C}$ is a subset of $\mathcal{P}$ comprised of the set of measures, $\mu(\cdot,\cdot)$, defined on the parameter and modelling spaces. Often this will be further interpretable as

$$Data = Structure \; o \; Noise ,$$

where o denotes the operation quantifying the interaction between structure and noise (Smith (1986)). Clearly, when selecting a possible family (or parameterisation) of measures, an attempt should be made to do so by means of extrinsic parameters, thus allowing a reasonable approximation to their *a priori* beliefs to be assessed. Furthermore, existing physical theory may yield insight into the necessary functional forms to be employed for the structure component. However, for a wide class of problems; for example, specifying priors for hyperparameters or specifying the error component, pragmatic choices inevitably have to be made and it seems necessary to be able to adopt a formal unified approach to the choice of representation. The following introductory sections give an overview of the formal approaches explored in this thesis.

(1.1) The "what if" principle

In general, to obtain a formal guide for the handling of such a representation in a purely statistical modelling context we will adopt the "what if" principle (Diaconis and Freedman (1986a)) together with a concept of approximation in the form of flexibility and tractability of the ensuing

statistical model.

First, the concept of infinite exchangeability must itself be judged as an idealised approximation. In a sense we can only feel confident with adopting the consequent mixture representation if the results holds, at least approximately, under a relaxation to a more realistic assumption of finite or partial exchangeability. Happily, this is the case for a wide range of statistical models; for example, exponential families with uniformly bounded fourth moments (Diaconis and Freedman (1986c)). An interesting technical result is that the representation in the finitely exchangeable case is exact if we allow the possibility of a signed mixing measure (Jaynes (1982a)).

The "what if" principle encourages *a priori* assessments induced by the implied modelling characteristics, for example, consistency. Moreover, it encompasses retrospective judgements in the form of sensitivity measures, for example, derivatives of posterior functionals with respect to the prior. Careful judgements of *a priori* assumptions in the light of the data are allowable on the grounds of approximation, for the original prior should only be viewed as an approximation (sometimes a poor one) to a true prior (Diaconis and Freedman (1986a)), suggesting that the continuity (in some sense) of the Bayes map and corresponding risks should be explored. For results in this direction, in the parametric case see Berk (1966), in the nonparametric case see Dalal and Hall (1980).

The "what if" principle suggests various interesting techniques for the assessment of statistical models and their ensuing modelling characteristics. For example,

(i) The application of techniques from decision theory together with the principle of maximum expected utility to induce modelling criteria.

(ii) Model elaboration techniques, including the modification of *a priori* inputs in the light of inconsistencies or sensitivity assessments.

(iii) The extent to which posterior functionals characterise prior measures.

(1.2) Choice of utility function

A formal application of the "what if" principle will be adopted to identify an interesting statistical modelling criterion via the decision problem of reporting the posterior distribution of the parameter of interest θ. Such a procedure requires the specification of a utility structure over the space of *a priori* beliefs, denoted by $\mathscr{C} \subset \mathscr{P}$. Thus the associated expected utility is proposed as a criterion to judge the choice of space $\mathscr{C}$. The principle of maximum expected utility applies to yield interesting choices of sub-families of measures contained in $\mathscr{C}$. Moreover, the Bayes risk can be viewed as quantifying the concept of approximation between two measures, for example, prior

or likelihood spaces, thus quantifying the notion of approximation and allowing the assessment of the playoff between a formal mathematical representation and a tractable statistical model.

Concerned with the choice of a suitable utility function, Bernardo (1979a) introduces the concept of a local and honest utility structure. This is sufficient to characterise the logarithmic utility function, an appealing choice in itself since it establishes a link between Information theory and Bayesian decision theory, thus allowing concepts like the code length of a string to be interpreted in a Bayesian setting (Rissanen (1987), Wallace and Freedman (1987)). The characterisation of the logarithmic utility structures and other well-known measures including the Rényi-α distance (Rényi (1961)) are considered. By interpreting the expected utility as inducing a quasi-distance on the relevant space of measures, concepts from differential geometry can be employed to assess the statistical model: for example, the midpoint, shortest line (geodesic) between two families of probability measures.

The utility structure can be adapted to include the possibility of the model being incorrect. A key result for the assessment of the model via its consistency is that the notion of convergence can be quantified in terms of a Kullback-Leibler distance (Berk (1966)). Berk's theorem states that under an incorrect model the posterior measure asymptotically converges to a subset of the parameter space, known as the asymptotic carrier, which minimises the Kullback-Leibler distance between the assumed family of measures and the true model. Thus a further insight is obtained for the induced Bayes risk under a logarithmic utility structure. The unified modelling criterion needs no notion of asymptotic normality (where the asymptotic carrier is a point) for the posterior density. In the setting of a consistent estimator the asymptotic information gain can be interpreted as the missing information concerning the parameter θ. Otherwise, for smooth models, it represents the possible amount of information to be gained from the model as a whole and cannot be interpreted as a function of the parameter of interest alone. The behaviour of the asymptotic information gain is examined in Chapter 3 using results of Ibramigov and H'asminsky (1973). The ensuing decomposition quantifies, on an information scale, the following components; dimensionality k of the parameter of interest, prior information, likelihood curvature and design matrix, through the expected logarithm of Fisher's information and the entropy functional of the prior. Thus a unified modelling criterion is obtained for the assessment of the relevant modelling components. Rényi (1961) shows that the information gain can be placed on a cost scale, this is of central importance in the interpretation of the criterion as a Bayesian risk.

(1.3) Choice of class $\mathscr{C}$

The "what if" principle implies that the choice of $\mathscr{C}$ must be judged by its implications for future modelling decisions. First, tractability of the class is paramount, for the updating of prior to

posterior via Bayes theorem must be feasible. Rather surprisingly, the updating can be carried out for a variety of large classes of measures; for example, Dirichlet priors, tail free priors, neutral to the left and right priors. An elegant theory for such processes was first proposed by Ferguson (1974) and consequently extended in many directions (see for example Antoniak (1974), Diaconis and Freedman (1983)). The flexibility of such a setting is apparent by virtue of the fact that, under a suitable topology, *any* prior measure can be approximated by a mixture of Dirichlet priors (Dalal and Hall (1980)), a completeness property for such prior measures. Secondly, the "what if" principle requires the property of consistency and the sensitivity of such a model to be addressed. It is here where the fully nonparametric approach appears to falter, for it has been shown that mixtures of Dirichlets can lead to inconsistencies (Diaconis and Freedman (1986b)). However, the class of Dirichlet priors are consistent and under mild conditions on the underlying density, for example, symmetry, the Bayesian formalism is consistent for a surprisingly large class of models. However, rather more alarmingly in the nonparametric framework, and to a lesser extent in the parametric case, is the possibility of an apparently innocuous prior dominating the effect of the data on the posterior, leading to a non-robust inference. Here an application of the "what if" principle in the form of a sensitivity or induced risk assessment of the modelling components quantifies dominant features. A host of problems can occur in an ill-specified parametric setting; for example, improper priors (Efron (1973), Stone and Dawid (1972)), curvature problems with likelihoods (e.g. exponential regression (Mitchell (1967)). Thus any modelling framework should take account of these possibilities and provide a warning mechanism for such unwanted properties. Two possible directions are as follows:

First, one could view the asymptotic information gain, by definition, as quantifying the relative domination of some components of the statistical model over others. The minimisation of the gain would then identify the least sensitive modelling input to the data, the maximisation to the beliefs where most is expected to be learnt from the data sequence, thus allowing a general framework for defining notions of sensitivity and robustness.

Secondly, one could explore the consequences, under a coherent modelling framework, for posterior functionals; for example, the mean and variance, as quantitative measures, or score functions, in order to induce *a priori* modelling assessments. Thus the examination of *a posteriori* assumptions that characterise aspects of statistical models is of interest. One such characterisation is for exponential families where linearity of the posterior mean of the natural parameter determines the class $\mathcal{C}$ as that of the conjugate family of prior measures (Diaconis and Ylvisaker (1985), Goel and DeGroot (1980)).

(1.4) Nonparametric versus parametric modelling

The bridge between the general (nonparametric) representation and one involving a finite dimensional conditioning parameter θ can be built by a variety of techniques. For example:

(i) In the spirit of the original de Finetti theorem it may be natural to impose further invariances, for example, random centred symmetry, to obtain a more precise characterisation of the mixture measure $\mu(\cdot,\cdot)$ (Smith (1981), Ressel (1985), Diaconis and Freedman (1987)).

(ii) The "what if" principle in the form of an induced expected utility constraint quantifies the process of learning in a probabilistic framework, thus formalising the concept of the domination of particular features of the statistical model in the learning process. An interesting consequence of the induced modelling criterion is that it gauges the rate at which we can expect to learn about additional parameters in a statistical model. For smooth likelihood surfaces the proposed rate is $O(n/\log n)$ where n is the sample size, coincidentally the same rate as proposed by the prime number theorem.

(iii) A formal application of the principle of maximum expected utility over a nonparametric class $\mathscr{C}$ can be applied to yield optimal sub-families of measures, for example, the location family made famous in Huber's (1964) fundamental paper on robust estimation.

(1.5) Model elaboration

The preceding discussion places us in a model elaboration framework (Box (1980), Smith (1983)), a methodology based on the assumption that it is better to begin with elementary building blocks and then to proceed by cementing them together at a rate gauged by the experience gained in the light of the data. The unified modelling criterion proposes such an elaborating framework by virtue of its three components; assessing likelihood, prior and cost of experimentation, respectively. Thus to proceed with such a framework, a tractable elaboration in the form of a one parameter embedding is adopted. The Bayes paradigm naturally lends itself to the reporting of such a model via *a posteriori* summaries in the form of marginal or conditional posterior densities.

Hence the decision theoretic framework determines interesting I-models over $\mathscr{C}$ upon which all observers will eventually agree. It further establishes techniques for the embedding and connecting of two possible families of models, $\{P_\theta\}$ and $\{Q_\theta\}$ say, that are under consideration, thus quantifying the process of moulding together the foundational bricks in an elaborated framework.

The unification obtained from the modelling criterion based on the concept of missing information is illustrated in detail for the location-scale family. The elementary building blocks consist of familiar families including the normal and double exponential densities. Techniques for

embedding and connecting these families of models in an optimal sense are determined, the solutions being interpretable in a differential geometric setting. An appealing technical result is the characterisation, via random centred symmetry, of the class of location-scale mixtures of normality. Choices of mixture measures in this class are considered.

A further desirable consideration for a statistical model is its flexibility, in that assigning zero *a priori* probability to a large class of measures can lead to inconsistent results (Cromwell's rule (Lindley (1972)). Note that given an unintended zero assignment, Bayes theorem alone cannot warn of such a incorrect specification, since the *a posteriori* probability is also zero. A framework for model criticism and elaboration is thus required (Box (1980), Smith (1983)). One possible viewpoint (Dawid (1982)) regarding the assessment of the adequacy of a given model is that a model is adequate if it does not assign a zero or near-zero probability to any prespecifiable event which then occurs.

The concepts of parameter orthogonality and *a priori* independence can be viewed as decision-theoretic solutions to well-posed decision problems involving information measures, the former being related to a relative information gain. However, note that the logarithmic utility structure is characterised by the fact that the asymptotic gain is invariant under a one-to-one transformation of the parameter (Good (1969), Amari (1982a)).

(1.6) Differential geometry

Differential geometry has a natural potential role in the Bayesian framework in the context of understanding the structure of a space of probability measures when we require a notion of distance. One such notion of distance is induced by considering the Bayes risk of the required decision problem (Diaconis and Ylvisaker (1985)). This is constructed by an application of the "what if " principle in the form of the Bayes risk associated with the missing information of the model.

The use of geometrical ideas in classical statistics was made apparent in Rao (1945) and has been exploited since by many authors, primarily, Efron (1975), Amari (1982a) and Cencov (1972) A recent review is contained in Cox, Reid and Barndorff-Nielsen (1982). Fisher's information is of central importance in assessing the properties of the parameter space, and results of these authors are directly applicable to the Bayesian methodology by virtue of the equivalence of modelling criterion and Fisher's information. A striking characterisation showing the central importance of Fisher's information in a differential framework is due to Cencov (1972), who shows that it is the only invariant Riemannian metric under symmetry conditions. This is a very appealing result to a Bayesian since it parallels the notion of exchangeability and the construction of model itself. In the Bayesian framework it is natural to consider the geometry of the space of measures and not the underlying parameter space. However, by virtue of the decomposition of the modelling criterion,

Fisher's information still has an important role to play. Furthermore, locally, the geometry under any invariant induced Bayes risk on $\mathcal{P}$ is equivalent to that of Fisher's information (Amari (1982a)). Further unification is obtained for the notion of approximation and coding length in a differential geometric setting (Campbell (1985)).

There are further applications of ideas from differential geometry within the "what if" framework, including the use of quantitative curvature measures for nonlinear models (Bates and Watts (1980)). However, we will adopt a more formal Bayesian approach to constructing such quantitative measures, taking the form of derivatives of mappings, the norm of the latter quantifying the "what if" sensitivity assessment and itself leading to a model selection criterion allowing interpretation of least sensitive inputs.

(1.7) Profile likelihoods

A central problem in the parametric statistical modelling framework is that of the elimination of a generally vector nuisance parameter. The classical approaches based on the likelihood function have adopted sufficiency and maximisation techniques, whereas the Bayesian approach is one of an averaging process with respect to the *a priori* beliefs. However, these contrasting techniques can be compared within the class of statistical model that possess some form of group structure. The latter framework provides a starting point for extending the classical framework, although equivalences with the Bayesian approach now take the form of approximations where agreement is only for large samples.

(1.8) Modelling characteristics

One consequence of the representation (1.1.2) is the need to specify the structure component of a statistical model involving, for example, the choice of design or a transformation to simplify the operation o (appearing in the context of "Data = Structure o Noise") or the specification of the noise component. In a sense, it is possible to view *a priori* specifications under the modelling criterion (see (3.1.2)) by considering contours of equal missing information. Thus the latter quantifies the playoff between the choice of design, specification of measures, transformation etc. It will be shown that the reference prior has a central role to play in the selection of an optimal design. Moreover, the concept of approximation is applicable in the form of projecting one family of measures onto a simplified class of measures, itself possibly related to a transformation of the parameter of interest. Thus the minimisation of the associated risk leads to a natural selection criterion for the specification of the latter.

(1.9) Outline of thesis

The application of the "what if" principle via formal decision theoretic criterion and quantitative proposals for the investigation of the robustness of proposed models are the main concerns of the thesis. These are discussed in the following chapters.

Chapter 2 reviews the role of distance (or discrimination) measures as candidates for Bayesian utility measures on function spaces, focusing primarily on the logarithmic utility function. The relationship with differential geometry is studied, allowing notions of curvature and distance, together with concepts of geodesics, to be applied in a Bayesian modelling framework.

Chapter 3 applies the principle of maximum expected utility in a decision theoretic approach in order to identify a Bayesian model selection criterion, the risk being interpreted as inducing a distance on the relevant space of measures, thus providing a link up with ideas in differential geometry. The central concept is the decision problem of reporting the posterior measure, the risk being the asymptotic information gain (or missing information) concerning the parameter θ. Hence, a unified approach to the selection of likelihood-prior combinations, incorporating the choice of design matrix, is obtained.

Chapter 4 discusses the nature of the selection of the prior measure under the criterion in Chapter 3 (Bernardo (1979b), Good (1969)). The application of such measures to a wide class of problems possessing some form of symmetry (invariance with respect to a group) are reviewed. The links between the elimination of nuisance parameters via a (modified) profile likelihood and marginal Bayes posteriors are reviewed.

Chapter 5 considers quantitative notions applicable in a modelling framework and illustrates these mainly by reference to the exponential family and the location-scale family. The characterisation properties of posterior functionals such as the mean are determined. The quantitative behaviour of means and variances, applicable in a robustness setting, is described for flexible classes of *a priori* input in the form of a scale mixture of normality.

Chapter 6 views more formal, possible, quantitative measures, via the concepts of differential geometry developed in Chapter 2, in the form of derivatives of possible Bayes mappings. These are primarily useful as *a posteriori* summaries, in the form of curvature measures, warning of possible departures from the current model in the form of curvature measures. The relationship of such curvature measures with the model selection criterion of Chapter 3 and profile likelihoods in Chapter 4 are discussed. Furthermore, a survey of the relationships, in the form of inequalities and local approximations, between the information measures viewed as Bayesian utility measures are explored, with a view to quantifying possible sensitivity departures.

Chapter 2 : Utility structures for reporting probability beliefs

In order to formalise procedures to aid us in the specification of interesting choices of beliefs, as represented by probability distributions, the principle of maximum expected utility (Lindley (1972), Jaynes (1982b), Good (1950)) over a subset $\mathscr{C}$ of the space of all distributions $\mathscr{P}$ is adopted. Thus it is required to specify a utility function on $\mathscr{C}$, that is $u : \mathscr{C} \to \mathbb{R}$. In most applications this can be viewed in the more general setting of a utility for a pair of distributions P and Q: for example, prior and posterior pairs, or a pair (P, Q) where Q is an approximation to the less tractable belief P (de Finetti (1979)). The latter decision problem has an associated Bayes risk, $E_P(u(P, Q))$, denoted by $R(P, Q)$. In general $R(P, Q)$ will not be symmetric in P and Q.

This chapter reviews the literature on the characterisation of such utility structures, the logarithmic scoring rule playing a central role as the only utility being local and honest for the reporting of beliefs (Bernardo (1979a)), thus establishing a link with the Kullback-Leibler divergence between two probability measures (Kullback and Leibler (1951)). Further characterisations are studied in the context of the decision problem of simultaneously approximating two distributions. It is shown that other well-known measures of distance between probability measures, for example the α-distance, can be viewed as Bayesian utility structures. A direct application is to the connecting of two families of measures, itself applicable in a variety of statistical problems: for example; reporting an expert opinion, building a model elaboration in a hierarchical fashion, or comparing competing models (Box (1980), Smith (1983)).

One natural way of viewing the class $\mathscr{C}$ is itself defined in terms of a utility, or distance, constraint. Here we consider variational techniques for the determination of measures under the principle of maximum expected utility over such classes $\mathscr{C}$, and we can appeal to notions from differential geometry to interpret the nature of the solutions: for example; midpoints, shortest lines or geodesics.

(2.1) Decision problems on $\mathscr{P}$

In this section we view possible specifications and characterisations of utility functions for the class of decision problems of reporting and approximating distributions. Let $P, Q \in \mathscr{C}$, then the utility of the pair (P, Q) is a function $u : \mathscr{C} \times \mathscr{C} \to \mathbb{R}$ and the expected utility, under the belief P, is denoted by $E_P(u(P, Q))$.

(2.1.1) Reporting the prior-posterior pair

Let θ denote the parameter of interest and x the data. It is required to construct a utility function for the decision problem of reporting our beliefs as represented by $(p(\theta), p(\theta|x))$. The following characterisation emerges from Bernardo (1979a) and is closely related to Seidler (1959).

Theorem (2.1.1) : The logarithmic scoring rule is characterised by the assumptions that it is local, honest and updates in a linear fashion.

(i) $u(\cdot)$ is local,

$$u(p(\cdot), p(\cdot|x)) = u(p(\theta), p(\theta|x)) .$$

(ii) linear updating, that is a gain in utility of the form

$$u(p(\theta), p(\theta|x)) = u^*(p(\theta|x)) - u^*(p(\theta))$$

for some utility function $u^* : \mathbb{R} \to \mathbb{R}$.

(iii) $u^*(\cdot)$ is honest, that is $P = Q$ is a solution to

$$\max_{\mathcal{P}} E_P(u(Q) - u(P)) .$$

Proof : Consider the calculus of variations of the functional

$$B(p(\theta)) = E_{\theta|x}(u^*(p(\theta|x)) - u^*(p(\theta)))$$

the solution to which are given by the Euler-Lagrange equations (see Hildebrand (1965)),

$$\frac{d}{d\theta}\left(\frac{\partial B}{\partial p'} \right) - \frac{\partial B}{\partial p} = \lambda$$

for some constant λ. Let p_x denote the posterior measure, then by virtue of the local assumption on u^*, $p_x = p$ is a solution to the above equation. Therefore evaluating at p,

$$-p\left(\frac{\partial u^*(p_x)}{\partial p_x} \frac{\partial p_x}{\partial p}(p) - \frac{du^*}{dp} \right) = \lambda .$$

However, from Diaconis and Freedman (1986) (see also (6.1.1)) we have

$$\frac{\partial p_x}{\partial p}(p) = 0 .$$

Hence,

$$\frac{du^*}{dp} = \frac{\lambda}{p}$$

which on integration yields

$$u^*(p) = A \log p + B \, .$$

for some constants $A, B \in \mathbb{R}$, as required.

(2.1.2) Approximating P by Q

Consider the problem of approximating a distribution P by another, Q say. Here theorem (2.1.1) applies to characterise the logarithmic utility. Denote the corresponding Bayes risk by

$$I(P, Q) = E_P(u(P) - u(Q)) = E_P\left(\log \frac{P}{Q}\right)$$

which is the Kullback-Leibler divergence between P and Q (Kullback and Leibler (1951)).

(2.1.3) Simultaneous approximation of the pair (F, G)

Suppose F, G $\in \mathcal{P}$ and it is required to approximate the pair (F, G) by a single distribution Q. Of primary concern will be the construction of the mid-point and shortest line through the two families F and G. The following notation will be used; let $d(Q\ ;$ (F, G)) denote the expected utility of the distribution Q with respect to the pair (F, G).

An extension of the ideas in (2.1.2) is to select $d(\cdot)$ to be of the form

$$d(Q\ ;\ (\mathrm{F}, \mathrm{G})) = u^*(\ R(Q, \mathrm{F}), R(Q, \mathrm{G})\ ) \qquad (2.1.1)$$

comprising the expected utilities of Q approximating F and G separately.

A natural choice of the utility $u^*(\cdot)$ which enables the mid-point and shortest line to be determined is to choose the utility distance for all $f, g \in \mathbb{R}$ to be

$$u^*(f, g) = \max\ (f, g)\ .$$

Combining this with (2.1.1), under a logarithmic scoring rule gives the measure

$$d_I(Q) = \max\ (\ I(Q, \mathrm{F}), I(Q, \mathrm{G})\ )\ . \qquad (2.1.2)$$

Note that $d_I(\cdot)$ is symmetric under interchange of F and G. In general it is not symmetric under interchange with Q, in which case the measure

$$d_I^*(Q) = \max\ (\ I(\mathrm{F}, Q), I(\mathrm{G}, Q)\ ) \qquad (2.1.3)$$

is obtained.

(2.1.4) Variational solution to choice of Q

The distance (2.1.3) and its variants can now be implemented in a range of decision problems for selecting the approximation Q. The technique required is that of the calculus of variations (see Hildebrand (1965)). Suppose we have a linear updating utility, then,

Theorem (2.1.2) : Let the distance $d_u(Q)$ be defined by

$$d_u(Q) \;=\; \max\,\big(\, E_Q(u(Q)-u(\mathrm{F})),\, E_Q(u(Q)-u(\mathrm{G}))\,\big)$$

and correspondingly define $d_u^{\,*}(Q)$. Assume that as $x \to \infty$ the following smoothness condition holds,

$$Qu(Q) \to 0\ .$$

Then the following solutions are obtained, where $\lambda_1, \lambda_2 \in \mathbb{R}$.

(i) $\inf\limits_{Q}\ d_I^{\,*}(Q)$ satisfies

$$u(Q) \;=\; (1-\lambda_1)u(\mathrm{F}) + \lambda_1 u(\mathrm{G}) + \lambda_2\ . \qquad (2.1.4)$$

(ii) $\inf\limits_{Q}\ d_u^{\,*}(Q)$ satisfies

$$((1-\lambda_1)\mathrm{F} + \lambda_1\mathrm{G})u(Q) \;=\; \lambda_2 Q\ . \qquad (2.1.5)$$

(iii) A corollary to (2.1.5) is that $\inf\limits_{Q}\ d_I^{\,*}(Q)$ satisfies

$$(1-\lambda_1)\mathrm{F} + \lambda_1\mathrm{G} \;=\; Q\ .$$

Proof of (i) : First we recast the problem of determining $\inf\limits_{Q} d_u(Q)$ as the equivalent variational problem,

$$\inf_{Q} E_Q(u(Q)-u(\mathrm{F}))\ \text{ subject to }\ E_Q(u(Q)-u(\mathrm{F})) \;=\; E_Q(u(Q)-u(\mathrm{G}))\ .$$

As Q is a density, an equivalent variational problem is

$$\inf_{Q} E_Q(u(Q)-u(\mathrm{F}))\ \text{ subject to }\ E_Q(u(\mathrm{G})-u(\mathrm{F})) = 0\ \text{ and }\ \int Q = 1\ .$$

Define the functional $R(\cdot)$, for $\lambda_1, \lambda_2 \in \mathbb{R}$, by

$$R(Q) \;=\; E_Q(u(Q)-u(\mathrm{F})) \;-\; \lambda_1 E_Q(u(\mathrm{G})-u(\mathrm{F})) \;-\; \lambda_2\Big(\int Q \;-\; 1\Big)\,. \tag{2.1.6}$$

Consider one parameter variations of $R(Q)$ given by the Euler-Lagrange equations (see Hildebrand (1965))

$$\frac{d}{dQ}(Qu(Q)) \;-\; u(\mathrm{F}) \;-\; \lambda_1(u(\mathrm{G})-u(\mathrm{F})) \;=\; \lambda_2\,,$$

on integration,

$$Qu(Q) \;=\; \lambda_2 Q \;+\; (1-\lambda_1)u(\mathrm{F})Q \;+\; \lambda_1 u(\mathrm{G})Q \;+\; A$$

for some $A \in \mathbb{R}$. By the smoothness assumptions, $Qu(Q) \to 0$ and $Q \to 0$ as $x \to \infty$, therefore $A = 0$. Hence,

$$u(Q) \;=\; (1-\lambda_1)u(\mathrm{F}) \;+\; \lambda_1 u(\mathrm{G}) \;+\; \lambda_2$$

as required.

Proof of (ii) : Proceeding as in (i) the required differential equation is

$$((1-\lambda_1)\mathrm{F} \;+\; \lambda_1\mathrm{G})\frac{du(Q)}{dQ} \;=\; \lambda_2 \tag{2.1.7}$$

which on integration yields,

$$((1-\lambda_1)\mathrm{F} \;+\; \lambda_1\mathrm{G})u(Q) \;=\; \lambda_2 Q + A\,. \tag{2.1.8}$$

But G, F, $Q \to 0$ as $x \to \infty$, implying $A = 0$.

Proof of (iii) : Using (2.1.7) with $u(Q) = \log Q$ gives

$$(1-\lambda_1)\mathrm{F} \;+\; \lambda_1\mathrm{G} \;=\; \lambda_2 Q\,,$$

but due to the fact that we are dealing with probability measures, $\lambda_2 = 1$, giving the required result.

The linear updating structure of theorem (2.1.4) lends itself to many applications and the geometries associated with it will be discussed in Section (2.2). Here we list some alternatives which prove useful in certain instances.

Theorem (2.1.3) : Assume that the conditions for theorem (2.1.2) hold, then

(i) Under the L^2 and L^1-norms for the risk $R(\cdot,\cdot)$ the linear connection is the shortest line

$$Q \;=\; (1-\lambda_1)F \;+\; \lambda_1 G\,.$$

(ii) Under Rényi's α-distance for the risk $R(\cdot,\cdot)$ the shortest line is represented by the family,

$$Q\ \alpha\ (\ \mu_1 F^{1/(\alpha-1)} + \mu_2 G^{1/(\alpha-1)}\ )^{\alpha-1}\ .$$

Proof : Consider the case of the L^2-norm, the solution follows as in (2.1.6). The functional is now

$$R(\cdot)\ =\ \int ((\ Q-F\ )^2 + \lambda_1(\ (Q-F)^2\ -\ (Q-G)^2\ ) + \lambda_2 Q)\ .$$

The Euler-Lagrange equations yield

$$(Q-F)\ -\ \lambda_1(Q-G)\ =\ \lambda$$

for some $\lambda \in \mathbb{R}$, rearranging gives the linear connection. The proof for the L^1-norm follows from the triangle inequality and that for Rényi's distance follows by the same argument as the L^2-norm.

The structure (2.1.3) can be used to characterise further well-known utilities; for example, the power utility is a solution to the following variational argument.

Theorem (2.1.4) : Suppose we have the following decision problem. Assume $u(\cdot)$ is local and honest, where honest is defined by the constraint that P is the solution to

$$\max_{P^\dagger}\ \Big(\ \int\ P^\dagger u(P)\ \Big)\ \text{ subject to }\ \int\ P u(P^\dagger) = \text{const}\ .$$

Then $u(\cdot)$ is necessarily of the form

$$u(P)\ =\ \mu_1 P^\lambda + \mu_2$$

for some $\mu_1, \mu_2, \lambda \in \mathbb{R}$.

Proof : Define the functional $R(\cdot)$ by

$$R(u)\ =\ \int\ (P^\dagger u(P) + \lambda_1 P u(P^\dagger) - \lambda_2 P^\dagger)$$

A one parameter variation of R implies, by the Euler-Lagrange equations, that

$$u(P^\dagger) - \lambda_1 P^\dagger \frac{\partial u}{\partial P^\dagger}\ =\ \lambda_2$$

which on integration gives a solution of the form (2.1.7), as required.

Corollaries : (i) The same solution is attained for the problem,

$$\max_{P^\dagger}\ \Big(\max\ \Big(\ \int\ P^\dagger u(P), \int\ P u(P^\dagger)\ \Big)\Big)\ .$$

(ii) The logarithmic scoring rule is a special case of the above, as $\lim_{\lambda \to 0} \dfrac{x^\lambda - 1}{\lambda}\ =\ \log x$.

The above can be extended to justify the α-distance (Rényi (1961)) as a Bayesian utility measure when formally reporting beliefs (Good (1968, 1969)). The required constraint is that of a ratio updating for the gain of information, parallelling the linear updating structure of (2.1.1).

Theorem (2.1.5) : Suppose we have a local and honest utility structure that can be decomposed as a ratio, that is, there exists $u^*(\cdot)$ such that,

$$u(P^\dagger, P) = \frac{u^*(P^\dagger)}{u^*(P)} \tag{2.1.8}$$

for all $P^\dagger, P \in \mathcal{P}$. Then, for some $C, \lambda \in \mathbb{R}$, $u(\cdot)$ is necessarily of the form,

$$u(P^\dagger, P) = CP^\dagger\left(\frac{P^\dagger}{P}\right)^\lambda .$$

Proof : The Bayes risk is given by

$$R(P^\dagger) = E_{P^\dagger}\big(u(P^\dagger, P)\big) ,$$

by (2.1.8),

$$R(P^\dagger) = E_{P^\dagger}\left(\frac{u^*(P^\dagger)}{u^*(P)}\right) .$$

A one parameter variation, with respect to $P^\dagger$, given by the Euler-Lagrange equations, is

$$\frac{u(P^\dagger)}{u(P)} + \frac{\partial u}{\partial P^\dagger}\frac{P^\dagger}{u(P)} = \lambda , \tag{2.1.9}$$

but $P^\dagger = P$ is a solution to (2.1.9) by the honesty of $u(\cdot)$. Hence,

$$\frac{\partial u}{\partial P^\dagger}\frac{P^\dagger}{u(P^\dagger)} = \lambda - 1 .$$

Therefore, on integration

$$u(P^\dagger) = C(P^\dagger)^{\lambda-1}$$

as required.

Note that division by $P^\dagger$ is required, containing the implicit assumption that $P^\dagger$ is absolutely continuous with respect to P. Thus, in instances where this fails, for example, truncated parameter spaces, the choice of particular utility (or distance) requires caution.

(2.1.5) A non-local utility structure

One clear direction for the extension of the utility structures analysed so far is to construct non-local utility structures. Previous techniques based on invariance have been used to characterise such structures (Amari (1982a), Rényi (1961), Good (1966a)). A direct application is to the possibility of requiring a utility structure for reporting beliefs $p(\theta|x)$, but where a further decision problem is specified for which the posterior is to be used. The technique employed will be to induce the non-local utility for the former decision problem with respect to the loss structure involved in the latter.

Consider the "pure" inference decision problem specified by utility $u : \mathscr{C} \rightarrow \mathbb{R}$ and posterior measure $p_{\theta|x}(\cdot)$, noting the fact that it is a member of the class $\mathscr{C}$. Then the measures, $U_n^{\theta|x}$ and $E_x(\, U_n^{\theta|x}\,)$, will be useful as *a posteriori* and *a priori* risks, respectively, for assessing statistical questions. For instance, define

$$U_n^{\theta|x} \;=\; \int u(\, p_{\theta|x}(\cdot)\,)p(\theta|x)d\theta\,.$$

Furthermore, assume that a loss structure $L(\theta, \theta')$ exists, where $(\theta, \theta') \in \Theta\times\Theta'$, thus quantifying the loss in reporting θ' when θ is the true value. In order to incorporate this into the structure of the former decision problem, an invariance argument (Good (1969)) leads us to consider the class of utilities of the form,

$$u(\, p_{\theta|x}(\cdot)\,) \;=\; U\left(\, E_{\theta'}(\, U^{-1}\,(\, L(\theta, \theta'))\,)\right)$$

for some function $U : \mathbb{R} \rightarrow \mathbb{R}$. Hence the curvature of the loss function is taken into account. Note that if $L(\cdot,\cdot)$ collapses to the line $\theta = \theta'$, then (2.1.5) is precisely the assumption of a local utility. A further interesting characterisation of non-local utilities in the form of mixtures, using Abel's theorem, is given in Good (1966a), Lindley (1972). The above framework has many applications in statistical modelling; for example, quantifying the influence of observations, selecting the sample size, and choosing the design matrix (see section (3.3.1)).

Consider the following choice of logarithmic utility structure for $U(\cdot)$,

$$U(\cdot) \;=\; \frac{1}{\alpha}\log(\cdot)\,,$$

so that the non-local utility becomes,

$$U_\alpha^{\theta|x} \;=\; \frac{1}{\alpha}\int p(\theta|x)\log\Big(\int \exp(\,\alpha\, L(\theta, \theta')\,)p(\theta')d\theta'\Big)d\theta\,.$$

The following result of Good (1969) involving the choice of loss structure will be applicable in Chapter 3, where an equivalent result with the asymptotic information gain under a local utility

structure will be made apparent. Consider the limit as α tends to zero,

$$\lim_{\alpha \to 0} U_{\alpha}^{\theta|x} = \int p(\theta)\log\left(\frac{p(\theta)}{\Delta(\theta)}\right)d\theta ,$$

where $\Delta(\theta)$ is related to the curvature of the loss function,

$$\Delta(\theta) = \frac{\partial^2 L(\theta, \theta')}{\partial\theta\partial\theta'} .$$

Thus the solution to the reference prior under the above combination of decision problems can be deduced from Good (1969) to have the form

$$\pi(\theta) \propto \Delta(\theta) .$$

To obtain a link with the local utility structure and the concept of missing information, consider the natural loss structure for the reporting of the likelihood structure as given by

$$L(\theta, \theta') = \int f(x|\theta)\log\left(\frac{f(x|\theta)}{f(x|\theta')}\right)dx .$$

Then (Good (1969)),

$$\Delta(\theta) = (\, I(\theta)\,)^{\frac{1}{2}}$$

where $I(\cdot)$ denotes Fisher's information. Thus the reference prior reduces to Jeffrey's prior (Jeffreys (1961)).

(2.2) Differential geometry in statistics

Consider a class $\mathcal{C}$ of measures with densities with respect to some dominating measure μ denoted by $p(\theta)$. Let $l(\cdot)$ denote the log density. Typical examples for the nature of the class $\mathcal{C}$ are: the prior-posterior space, likelihood space or predictive space, depending on the nature of the required inference.

There are many ways of considering the geometry of a statistical model. Primarily we will concentrate on the geometry of the parameter space and of the space of distributions $\mathcal{P}$. First, the geometry associated with the parameter space can be determined via the Fisher's information associated with the family of measures (Rao (1945), Jeffreys (1961)), allowing curves and surfaces to be defined. Secondly, and central to the Bayesian methodology, the geometry of the space of distributions is considered. Here the geometry will be induced via the Bayes risk of the decision problem at hand, the risk being interpreted as a measure of distance (or discrimination) between two probability measures. Such associated geometries are discussed in Amari (1982b) and were implicit in Jeffreys (1961) for constructing invariant prior measures. The modelling criterion (see

(3.1.1)) can be viewed as such a discrimination measure on $\mathcal{P}$. The criterion of maximum expected utility for the selection of prior, likelihood and design inputs can be viewed in this geometric setting.

(2.2.1) α-connections and α-geodesics

In the context of a model elaboration with parameter λ consider the induced smooth curve $p(\theta|\lambda)$ in $\mathcal{C}$ together with $l(\theta|\lambda)$. The parameterisation by λ allows us to discuss the associated geometries of the model. A function of central importance is the score function with respect to λ, denoted by $\dot{l}$, where

$$\dot{l} \;=\; \frac{d}{d\lambda} l(\theta|\lambda)\;,$$

this itself has previously been proposed as a quantitative measure for judging model robustness via the marginal beliefs for the data (Box (1980)). It satisfies

$$E_{\theta|\lambda}\left(\frac{d}{d\lambda} l(\theta|\lambda)\right) \;=\; 0\;.$$

Under suitable smoothness conditions we can define the tangent space, $T_{p(\cdot)}$, of $\mathcal{C}$ at $p(\cdot)$ by

$$T_p \;=\; \{\, g(\theta) \mid E_\theta(\, g(\theta)\,) = 0 \,\}\;.$$

A tangent vector can then be viewed as a linear mapping from the set of smooth functions to the real line. Amari (1982b) discusses the use and interpretation of T_p in a statistical context. Furthermore T_p can be endowed with an inner product structure by

$$<\, f, g \,> \;=\; E_\theta(\, f(\theta) g(\theta)\,)\;.$$

The information metric then becomes,

$$\|\dot{l}\|^2 \;=\; E_{\theta|\lambda}\left(\left(\frac{d}{d\lambda} l(\theta|\lambda)\right)^2\right) \;=\; I_{\theta|\lambda}(\cdot)$$

that is, Fisher's information for the one-parameter family of distributions $p(\theta|\lambda)$.

The α-connection and α-geodesic can now be defined. For a parameter α, the curve $s(\theta, \lambda)$ is a parallel displacement with respect to the α-connection along the curve if it satisfies

$$\dot{s} \;+\; \tfrac{1}{2}(1-\alpha)\, s\, \dot{l} \;+\; \tfrac{1}{2}(1+\alpha) E_{\theta|\lambda}(\, s\dot{l}\,) \;=\; 0\;. \qquad (2.2.1)$$

If the tangent vectors $\dot{l}$ are parallel displacements then the curve is called an α-geodesic. By (2.2.1) they satisfy

$$\ddot{l} + \tfrac{1}{2}(1-\alpha)\dot{l}^2 + \tfrac{1}{2}(1+\alpha)I(\cdot) = 0 .$$

Note that the ($\alpha = 1$)-geodesic connects the two measures via the exponential embedding while the ($\alpha = -1$)-geodesic is the linear, or mixture, connection, thus neatly parallelling their corresponding roles in the decision-theoretic framework of Chapter 3. The case $\alpha = 0$ leads to a differential equation via (2.2.1) similar to that obtained for the minimisation of Fisher's information as required in Chapter 3.

(2.2.2) Geometry of the space of distributions

Consider now the intrinsic geometry of the space of distributions, $\mathcal{P}$. Dawid (1977) explains this by embedding $\mathcal{P}$ in a Hilbert space as follows. Let Q be the set of finite measures of which $\mathcal{P}$ is a subset. Let μ be a σ-finite measure such that every member of $\mathcal{P}$ is equivalent to: the map $Q \to 2\sqrt{dQ/d\mu}$ embeds Q into an L^2-space, the image of $\mathcal{P}$ being the set

$$\{\, g \mid \|g\| = 2 \text{ and } g > 0 \,\},$$

that is, a portion of a sphere radius 2. Thus the geometry of the space of distributions will resemble that of a sphere, where geodesics are parts of great circles.

The induced distance between two measures p and q is the Hellinger ($\alpha = \frac{1}{2}$) distance, defined by

$$\tfrac{1}{2}(\,\rho(p, q)\,) = 1 - \int \sqrt{pq}\, d\mu .$$

Note that locally the geometry for a parameterised family is equivalent to the Riemannian geometry as given by Fisher's information, due to the fact that, evaluated at $\lambda = \theta$,

$$\frac{d}{d\lambda}\,\rho(\,P_\theta, P_{\theta+\lambda}\,) = (\,I(\theta)\,)^{\frac{1}{2}} .$$

A generalisation of the equivalence of the local geometry of any local, honest and ratio updating utility structure with the Riemannian geometry of Fisher's information is given in Amari (1982a).

The construction of the geodesic between two measures is given in Dawid (1977), Amari (1982a). The geodesic curve can be parameterised by the elaboration parameter λ, for $0 \leqslant \lambda \leqslant 1$, and is given by

$$p(\theta|\lambda) = c(\lambda)\,(\,\sqrt{p(\theta)} + \lambda(\sqrt{q(\theta)} - \sqrt{p(\theta)})\,)^2 \qquad (2.2.2)$$

for some suitable normalising constant $c(\cdot)$. Thus (2.2.2) can be viewed as a candidate for a continuous model elaboration of the measures p and q. The interpretation, however, of the model parameter λ is hard to identify. The Bayes risk associated with such a solution is given by the

geodesic distance between p and q, denoted by λ_H,

$$\lambda_H = 2\cos^{-1}\int\sqrt{pq}\,d\mu\ .$$

Thus the Bayes risk for the Hellinger utility can be interpreted as the geodesic distance between the two measures. Furthermore the derivative of the prior to posterior map can be defined in such a topology (Eplett (1985)) but again a natural interpretation of the resulting measures is lacking. To overcome this we instead adopt the variational norm, noting that this is an equivalent metric due to the inequality

$$\tfrac{1}{2}\rho(p,q) \leqslant V^2 \leqslant \rho(p,q)\ .$$

Thus from a geometric viewpoint, the two topologies will appear the same, although the corresponding distance measures will take different forms. Moreover, the total variational norm readily lends itself to interpretation as a utility function and the corresponding measures have easily interpretable forms in a robustness setting; see Chapter 6.

(2.2.3) Geometry of the probability simplex

Consider a space of measures concentrating on the probability simplex, S^{n-1}, associated with a discrete probability vector $\boldsymbol{p}$ of dimension n. Cencov (1972), Campbell (1985) consider the geometry of this simplex in a statistical context. Note that it possesses wide application; for example, when we wish to consider the geometry of a discrete prior or posterior space, or the geometry of a discrete likelihood surface parameterised by a vector $\theta = (\theta_1,\ldots,\theta_k) \in \mathbb{R}^k$. The latter will induce a geometry on $\mathbb{R}^k$ through the metric defined on S^{n-1}. For convenience assume that

$$S^{n-1} = \left\{ p \mid p_i > 0,\ \sum_{i=1}^{n} p_i = 1 \right\}.$$

The simplex can be embedded in the positive cone, $\mathbb{R}_n^+$, by dropping the condition that it forms a density. Both sets can be viewed as differentiable manifolds and thus possess geometrical structure. Here we review the work of Cencov (1972), Campbell (1985), Amari (1982b), but with emphasis on the Bayesian decision theoretic nature of the corresponding metrics. In Chapter 3, a Bayes risk will be proposed to assess the statistical model itself inducing a notion of distance.

To each point $\boldsymbol{p} \in \mathbb{R}_n^+$ we can associate a tangent space, itself a vector space, denoted by $T_{\boldsymbol{p}}$. The vectors in $T_{\boldsymbol{p}}$ define directional derivatives. A possible set of basis vectors for the space $T_{\boldsymbol{p}}$ is given by the coordinate vectors $\dfrac{\partial}{\partial p_i}$.

A Riemannian metric on $\mathbb{R}^+$ is an inner product defined on the tangent space in such a way that when evaluated at $\boldsymbol{p}$, $< \cdot,\cdot >_{\boldsymbol{p}}$, is a C^∞ function of $\boldsymbol{p}$. In a similar fashion to the assumption

of exchangeability, we will be interested in characterising Riemannian metrics that possess special symmetry properties. The latter condition will be one of invariance under the set of Markov mappings. To define the latter consider a mapping which takes the n dimensional cone to an m dimensional cone, denoted by $\phi : \mathbb{R}_n^+ \rightarrow \mathbb{R}_m^+$, which will induce a mapping, $\phi^* : T_p \rightarrow T_q$ on the corresponding tangent spaces. The function $\phi^*(\cdot)$ is known as an isometry if it leaves the inner product invariant.

Theorem (Cencov (1972)) : Consider the Riemannian metric defined by on S^{n-1} by

$$\left\langle \frac{\partial}{\partial p_i}, \frac{\partial}{\partial p_j} \right\rangle_p = \frac{\delta_{ij}}{p_i} \tag{2.2.3}$$

where δ_{ij} is the Kronecker delta. Then (2.2.3) defines the only Riemannian metric (up to a multiplicative constant) that is invariant under all Markov isometries.

Proof : See Theorem (11.1) of Cencov (1972). The generalisation to $\mathbb{R}_n^+$ is given in Campbell (1986).

The metric (2.2.3) now defines the Riemannian distance on the simplex as

$$(ds)^2 = \sum_{i=1}^{n} \frac{(dp_i)^2}{p_i} , \tag{2.2.4}$$

thus it is natural to adopt (2.2.4) to explain the geometries of a discrete statistical problem by virtue of the characterisation result (2.2.3). Campbell (1985) uses the above structure to unify some of the central concepts in statistical modelling methodologies.

First, a link with Fisher's information can be established. In this case the model consists of a parameterised likelihood, $p_i(\theta)$. The induced Riemannian metric on $\mathbb{R}^k$, generated by the inner product between tangent vectors $\frac{\partial}{\partial \theta_i}$ and $\frac{\partial}{\partial \theta_j}$ is given by (2.2.4) is

$$\left\langle \frac{\partial}{\partial \theta_i}, \frac{\partial}{\partial \theta_j} \right\rangle_\theta = \sum_{k=1}^{n} \frac{1}{p_k} \frac{\partial p_k}{\partial \theta_i} \frac{\partial p_k}{\partial \theta_j}$$

the associated matrix being Fisher's information.

Secondly, in a Bayesian context, the procedure of minimum information divergence or projection (Csiszár (1975)) as a notion of approximation can be explained. Amari (1982b) considers this in the concept of an α-ancillary subspace, Campbell (1985) proves an orthogonality property with respect to the metric (2.2.2) when a minimum divergence solution is required under a moment constraint. Note that it is possible to view the modelling criterion (3.1.2) in such a setting, for it is a divergence between the prior and posterior measures.

Thirdly, the coding theory notion of minimum description, or coding, length has an interpretation under the distance (2.2.4) (Campbell (1985)).

(2.3) Discussion

Utility structures for reporting beliefs on the space of distributions, $\mathcal{P}$, have been explored with a view to constructing model elaborations in a Bayesian statistical framework (see Chapter 3). The latter requires the application of the principle of maximum expected utility, the ensuing solutions are obtained via techniques from the calculus of variations and are generally interpretable as the shortest line or midpoint between two measures. The specification of the utility structure itself has been explored, the main results being the characterisation of the logarithmic and power utility functions.

Differential geometry provides a means for describing the properties of the space of distributions. Such a framework unifies statistical concepts and suggests further directions to explore, for example, curvature measures and geodesics. A notion of distance is required and the Riemannian metric generated by Fisher's information is mathematically natural by virtue of its invariance properties. In Chapter 3 the distance measure will be one induced via the Bayesian risk of reporting the posterior measure. However, an equivalence with Fisher's information on the parameter space will be established, thus allowing results from differential geometry to apply.

One interesting link requiring further study is with that of the differential geometry of finite state spaces in Physics (see Ingarden (1981)).

Chapter 3 : Bayesian model choice : A decision theoretic criterion

Statistical modelling requires a high degree of flexibility in specifying interesting functional forms for subjective *a priori* beliefs. This is reflected in Bayesian methodology by the general version of the de Finetti representation theorem (Hewitt and Savage (1955)) which establishes that the coherent modelling and analysis of an infinitely exchangeable sequence requires the specification of prior measures over spaces of distribution functions. Concerned with such a representation, Smith (1984) remarked that "the task of translating actual beliefs into the required mathematical form of measures over function spaces seems - to say the least - a daunting prospect".

On the sole basis of mathematical convenience we often proceed as if the prior measure concentrates its weight on a particular finite dimensional family, for example, the normal. In itself this immediately poses a vast collection of possible elaboration questions; for example, the type of functional form, how many parameters, which parameterisation, specification of prior beliefs in such a parameterisation, etc ?

The methodology adopted will be that of a model elaboration formalisation (Smith (1986)). In order to quantify, in a unified manner, the process of model elaboration we adopt a formal decision theoretic framework under a utility structure, $u(\cdot)$, defined on the prior space, $\mathscr{C}$. The principle of maximum expected utility then characterises interesting classes of simplified measures contained in $\mathscr{C}$. Our analysis proceeds as if these measures are our natural beliefs, with the tenet that any inferences drawn will serve as a good approximation to a full, in general nonparametric, analysis over $\mathscr{C}$. Reliable guidelines on elicitation of prior beliefs and interpretation of reported beliefs are lacking in the nonparametric setting, although an elegant theory exists for the handling of large tractable spaces of prior measures via Bayes theorem (Ferguson (1974), Antoniak (1974)).

The application of a fully automatic procedure is far removed from the Bayesian methodology. The role, however, of a formal decision theoretic approach as a mathematical artefact to aid in identifying *interesting* directions for exploration in a model building framework seems inevitable. Subjectivists might adopt such a criterion as a working approximation, while others might find the objectivity an attractive feature. Furthermore, interpreting the Bayes risk as inducing a concept of distance on the space $\mathscr{C}$ allows notions from differential geometry to aid us in the understanding of the underlying structure of the set of prior measures. For example, equipped with a suitable topology, the space of probability measures, $\mathscr{P}$, can be viewed as a sphere (Dawid (1977)).

The specification of beliefs now takes the form of a class $\mathscr{C}$ and a utility structure defined over $\mathscr{C}$. Consider the generation of interesting classes $\mathscr{C}$. Two possible directions can be taken:

First, from a foundational viewpoint, the nature of $\mathscr{C}$ will typically be one of a symmetry constraint for the joint beliefs $p(\boldsymbol{X})$. For example, partial exchangeability or random centred symmetry, the latter necessarily leading to the class of scale mixtures of normality. Ressel (1985) exhibits conditions for the representation of measures in terms of mixtures. It is noted that such representations depend heavily on the underlying structure of the sample space; for example, the original de Finetti characterisation theory leads to mixtures of binomials. Moreover, there are close links with the notion of sufficiency and partial sufficiency (Diaconis and Freedman (1981)). The notion of sufficiency playing a central role in the modelling criterion under a logarithmic utility for the Kullback-Leibler measure is invariant with respect to sufficiency (Kullback-Leibler (1951)). Thus we need only view the assessment of the model via the associated sufficient statistics, in a similar manner to the theory of stochastic complexity (Rissanen (1987)).

Secondly, it seems appealing to generate (in some optimal way) the class $\mathscr{C}$ from initial building blocks, for example normality. One such elaboration is simply to consider the ε-contaminated class of normals. This construction can be shown to be optimal, in the sense that it is generated by the geodesic between the normal family and an arbitrary measure H, under the variational utility structure.

The asymptotic Bayes risk for reporting the posterior distribution can, in general, be interpreted as the missing information concerning the parameter θ. It will be shown that this yields a unified modelling criterion for the selection of; dimensionality, likelihood and prior combinations, design matrix. In so doing it unifies, and extends, previously proposed criterion of Schwarz (1978), Bernardo (1979b), Rissanen (1987) and is closely related to that of Dawid (1984).

The appropriate elaboration is characterised by principle of maximum expected utility (Lindley (1972)), In the parametric setting with a known likelihood and elaboration, the procedure, if formally applied, reduces to the determination of the reference prior of the (vector) parameter θ (Bernardo (1979b)).

A fruitful class of problems suggested via the "what if" principle is the choice of a univariate elaboration parameter λ from a decision theoretic procedure, this requires the specification of the current modelling position and a class $\mathscr{C}$ of possible departures. Then the minimisation of the induced Bayesian risk with respect to $\mathscr{C}$ yields an optimal one parameter variation, indexed by the modelling parameter λ, through our current model. A full Bayesian analysis would require a probability measure over the whole space $\mathscr{C}$, but here we select a measure that concentrates with probability one on the one parameter variation. The model flexibility is determined via the prior measure

on λ, denoted by $p(\lambda)$.

The flexibility of the approach lies with the input class $\mathcal{C}$, which can either be infinite dimensional; for example (ε-contaminated by an arbitrary measure H) or specified by a finite dimensional indexing parameter. In some sense the one parameter variation can be viewed as the most robust departure from the current model with respect to the class $\mathcal{C}$. For simplicity, suppose the current model is given by the normal family parameterised by $\theta = (\mu, \sigma^2)$. The following elaborations are exhibited as one parameter variations with respect to some $\mathcal{C}$; the Huber family (Huber (1964)), the t-family and the pair (λ_D, λ_U) denoting the double exponential and uniform distributions.

(3.1) Asymptotic information gain

Ibragimov and H'asminsky (1973) obtained regularity conditions for the Shannon information between the prior and posterior. Here we briefly review their results which calculate the missing information about a parameter θ. Consider a family of measures $\{P_\theta \mid \theta \in \Theta\}$ where $\Theta \subset \mathbb{R}^k$. Let $f(x|\theta)$ denote the corresponding densities with respect to some σ-finite measure μ. Furthermore, assume that θ is a random variable with density $p(\theta)$. Define the gain in information in θ contained in the sequence X by

$$I_n^\theta = \int_\Theta \int_X p(\theta) f(x|\theta) \log\left(\frac{f(x|\theta)}{p(x)}\right) dx d\theta .$$

Consider the process $Z_n(\alpha)$ defined by

$$Z_n(\alpha) = \prod_{i=1}^{n} \frac{f(x_i|\theta + \phi(n)\alpha)}{f(x_i|\theta)}$$

for a suitable normalising constant $\phi(n)$. By the martingale convergence theorem the above process converges to a limiting process as $n \to \infty$, denoted by $Z(\alpha)$. Therefore,

$$I_n^\theta = -\log \phi(n) - \int p(\theta) \log p(\theta) - E_\theta\left(\log \int \frac{p(\theta + \phi(n)\alpha)}{p(\theta)} Z_n(\alpha) d\alpha\right) .$$

Under suitable smoothness conditions on the prior, the dominated convergence theorem implies that, as $n \to \infty$,

$$\lim_{n \to \infty} \left(I_n^\theta + \log \phi(n)\right) = -\int p(\theta) \log p(\theta) - E_\theta\left(\log \int Z(\alpha) d\alpha\right) . \qquad (3.1.1)$$

It will be shown that (3.1.1) provides a unified modelling criterion. Moreover it yields a general version of theorem (4.2) in Rissanen (1987) where the information gain is interpreted as a coding length. Note that the asymptotic gain is in fact infinite to learn about a continuous parameter. However, if the cost of experimentation is allowed for the gain can be stabilised (see (3.1.3)).

Consider two possibilities for the behaviour of $\phi(n)$:

(i) In the smooth case, $\phi(n) = 1/\sqrt{n}$, the limiting process is Gaussian,

$$Z(\alpha) = \exp\left(\left(I^{\frac{1}{2}}(\theta)\alpha\right)^T Z - \tfrac{1}{2}(I(\theta)\alpha)^T\alpha\right)$$

where $\boldsymbol{Z}$ is a standard multivariate normal random variable and $I(\cdot)$ is Fisher's information. Therefore, the asymptotic information gain (3.1.1) decomposes as

$$\lim_{n\to\infty}\left(I_n^{\theta} - \frac{k}{2}\log\left(\frac{n}{2\pi e}\right)\right) = \int p(\theta)\log\left(\frac{|I(\theta)|^{\frac{1}{2}}}{p(\theta)}\right)d\theta\,. \qquad (3.1.2)$$

Thus, from a formal decision theoretic perspective, for a smooth k dimensional model, equation (3.1.1) quantifies the sensitivity of the model to the following choices; dimensionality k, sample size n, prior $p(\theta)$, likelihood $f(\boldsymbol{x}\,|\,\theta)$ and design matrix (via the Fisher information), thus yielding a unified modelling criterion.

A direct application of the above is to the location family where the necessary regularity conditions for the two cases $1/\phi(n) = \sqrt{n}$, n are given in Ibragimov and H'asminsky (1973).

(ii) Consider the non-regular situation for $\phi(n)$. Depending on the smoothness of $f(\boldsymbol{x}\,|\,\theta)$ the following possibilities occur for $1/\phi(n)$; n^{γ} for $\gamma > \frac{1}{2}$, $n \log n$. (see Ibragimov and H'asminsky (1973)).

(3.1.1) Information on a cost scale with application to selection of dimensionality.

The missing information as defined by (3.1.1) leads to a unified modelling criterion. The amount of missing information concerning a continuous parameter will be infinite. The first term quantifies, as $\frac{1}{2}k \log n$, the rate at which we can expect to learn about our vector θ, parallel to previous model choice criterion; for example, Schwarz (1978), Rissanen (1987), Dawid (1984). Clearly, the cost of experimentation must be included in the unified modelling criterion, for otherwise the optimal solution would be to sample to infinity and almost surely determine θ. Rényi (1961) proves that the information measure is comparable to a cost scale, so if c is the cost of experimentation on an additive scale the criterion (3.1.2) gives risk

$$\lim_{n\to\infty}\left(I_n^{\theta} - \frac{k}{2}\log\left(\frac{n}{2\pi e}\right) + cn\right) = \int p(\theta)\log\left(\frac{|I(\theta)|^{\frac{1}{2}}}{p(\theta)}\right)d\theta\,. \qquad (3.1.3)$$

Clearly a natural choice for the rate at which we can hope to learn about parameters under such a risk structure is given by

$$k = \frac{2cn}{\log\left(\frac{n}{2\pi e}\right)}$$

for this stabilises the asymptotic information gain to the (generally finite) Kullback-Leibler distance between Jeffrey's prior and the actual beliefs $p(\theta)$. The latter explains the nature of Jeffrey's prior as one of an approximation to a proper belief $p(\theta)$ and quantifies the behaviour of individual modelling characteristics.

Note that if (3.1.1) is viewed as an automatic approximation for all n, then a continuous elaboration is beneficial only for $n \geq 16$, otherwise a simple discrete mixture seems appropriate. For large sample size it is apparent that (3.1.1) gauges the optimal rate at which we can expect to learn about unobservables as $O(n/\log n)$, allowing us to increasingly elaborate as the data increases, an idea previously suggested by many authors, for example, Huber (1973). The term $\frac{1}{2}k \log (2\pi e)$ can be interpreted as the entropy of a k dimensional standard multivariate normal.

(3.1.2) Risk under an incorrect model specification

First, an implicit assumption in the framework of defining the missing information for θ is that the process generating the data is an element, $f(\boldsymbol{x}|\theta_0)$ say, of the modelling space, for then the posterior converges almost surely to θ_0. Clearly, in such an instance the divergence between prior and posterior measures the missing information about θ.

Consider the scenario where the sequence $\boldsymbol{X}$ is generated by the measure $G(x) \notin P_\theta$. It is then known that the posterior converges to the asymptotic carrier, A_0, of the set Θ (Berk (1966)). The decision problem now requires a specification of the loss associated with the family $\{P_\theta\}$. Suppose our loss structure is additive and defined by the distance of an element of $\{P_\theta\}$ to the density $g(y,\cdot)$ after observing the data $\boldsymbol{x}$. Thus, an intuitive construction for the Bayes risk for the vector (θ, P_θ) is given by

$$I_n^{(\theta,\, f(y|\cdot))}(\, g(\cdot)) \;=\; I_n^{\theta} \,+\, E_\theta\left(\int g(y|\boldsymbol{x})\log\left(\frac{g(y|\boldsymbol{x})}{f(y|\theta)}\right)dy\right)$$

where $f(y,\cdot) \in P_\theta$. That is, the measure of gain in information concerning θ plus the risk in approximating the inference from the true measure $G(\cdot)$ by that of the family $\{P_\theta\}$ after observing data $\boldsymbol{x}$. Note that, asymptotically, the final term is (see Berk (1966))

$$\sup_{\theta \in \Theta} \int g(y|\theta)\log\left(\frac{g(y|\theta)}{f(y|\theta)}\right)dy\,,$$

in the case $g(\cdot) \in P_\theta$ it is zero. Moreover, it is zero if $g(\boldsymbol{x})$ possess enough symmetry, for example, centred symmetry if $\{P_\theta\}$ is normal, to be decomposed as a mixture of the family $\{P_\theta\}$. In which case $g(y|\theta) = f(y|\theta)$, and the models are equivalent after observing the sequence $\boldsymbol{X}$. In a sense the model $f(\boldsymbol{x}|\theta)$ is robust for all possible mixtures of itself.

(3.1.3) Example : The location problem

Suppose that our model consists of a location family induced via the representation (1.1.1), giving rise to a density denoted by $f(x-\theta)$ and prior measure $p(\theta)$. Fisher's information, $I(\theta)$, is independent of θ, thus write $I(\theta) = I_{f(\cdot)}(\cdot)$ noting the dependence on $f(\cdot)$. Under suitable regularity conditions (Ibragimov and H'asminsky (1973)), the modelling criterion (3.1.2) with $k = 1$ yields,

$$\lim_{n \to \infty} \left(I_n^{\theta} - \frac{1}{2}\log\left(\frac{n}{2\pi e}\right)\right) = \frac{1}{2}\log I_{f(\cdot)}(\cdot) - H(p(\theta)) ,$$

thus quantifying the missing information concerning the location parameter θ.

Hence, as in the discrete case, the influence of the prior on the Bayes risk is determined by the entropy functional, $H(p(\theta))$. Note that maximising this over the space $\mathcal{P}$ yields the improper lebesgue measure for the prior $p(\theta)$. Moreover attention can be focused on the modelling structure, $f(\cdot)$ where the necessary functional minimisation is that of $I_{f(\cdot)}(\cdot)$. Thus a link has been established between a formal decision theoretic proposal and previous *ad hoc* suggestions in a classical robustness context involving Fisher's information. Note that, Huber (1981) has adopted a sceptical stance where he states "Bayesian statistics ... confounds the subject with admissible estimation in an *ad hoc* supermodel, and still lacks reliable guidelines on how to select the supermodel and prior so we end up with robustness".

The formal criterion of minimisation of the risk given in (3.1.1) is adopted to select the Bayesian model. In a sense this can be viewed as a Bayesian definition of robustness; for the consequent modelling structure is by construction least sensitive to data input, thus guarding against aberrant observations. Note the parallel concept underlying the reference prior as that of the most sensitive prior, that is most to be learnt, to the data input. The equivalence exhibited in (3.1.1) with the minimisation of Fisher's information allows results from classical robustness to be applied in the Bayesian model choice context. In the latter we are concerned with determining interesting forms of intersubjective models which eventually observers will agree on, thus the regularity conditions on $\mathcal{C}$ for the uniqueness are required. These are achieved via the equivalence with Fisher's information where existence and uniqueness are given by a variational argument (Huber (1981)); existence is established for vaguely compact sets $\mathcal{C}$, uniqueness follows if $\mathcal{C}$ and the set where the density of the minimising distribution are both convex.

The following theorem can be applied to exhibit the Huber family as a formal Bayesian decision theoretic solution to a well-posed modelling problem.

Theorem (Huber (1964)) : Consider the class of contaminations of the distribution function G, denoted by $\mathscr{C}_G = \{(1-\varepsilon)G + \varepsilon H \mid H \text{ arbitrary}\}$. Then the variational problem,

$$\min_{\mathscr{C}_G} I_{f(\cdot)}(\cdot)$$

has solution given by

$$\begin{aligned} f_0(x) &= (1-\varepsilon)g(x_0)\exp(k(x-x_0)) && x \leqslant x_0 \\ &= (1-\varepsilon)g(x) && x_0 \leqslant x \leqslant x_1 \\ &= (1-\varepsilon)g(x_1)\exp(-k(x-x_1)) && x \geqslant x_1 , \end{aligned} \tag{3.1.3}$$

where $x_0 < x_1$ are the endpoints of the interval $|g'/g| \leqslant k$, k itself being a function of ε. The following theorem can now be deduced.

Theorem : The solution to the formal Bayesian decision problem,

$$\min_{\mathscr{C}_G} I_{\infty}^{\theta}$$

is given by the Huber family as defined by (3.1.3).

Proof : By virtue of the decomposition (3.1.2), the following equivalence of holds,

$$\min_{\mathscr{C}_G} I_{\infty}^{\theta} \equiv \min_{\mathscr{C}_G} I_{f(\cdot)}(\cdot) ,$$

the solution to which is given by (3.1.3).

A direct application is to a possible elaboration from normality, where G is the normal distribution function conditional on μ and σ, denoted Φ. Then a one parameter elaboration from normality, as determined by the criterion (3.1.2) with solution (3.1.3), is given by

$$f_H(x|\lambda) = \frac{1-\varepsilon}{\sqrt{2\pi}}\exp(-u_\lambda(x)) ,$$

where, suppressing μ and σ, $u_\lambda(x) = \frac{1}{2}x^2$ for $|x| \leqslant \lambda$, $u_\lambda(x) = \lambda|x| - \frac{1}{2}\lambda^2$ for $|x| \geqslant \lambda$, the so-called Huber family. The indexing parameter λ is defined in terms of ε by

$$\frac{2\phi(\lambda)}{\lambda} - 2\Phi(-\lambda) = \frac{\varepsilon}{1-\varepsilon} .$$

By definition of $u_\lambda(\cdot)$ we see that as λ varies the double exponential and normal families are obtained as limiting cases. Thus the Huber family can be viewed as a model elaboration connecting these two families, contrasting the exponential embedding (see section (3.4)).

Example : The location-scale family

Consider an infinitely exchangeable sequence, $\boldsymbol{X}$, for which we shall impose a further symmetry constraint. Let $\mathcal{C}_S$ denote the class of distribution functions F that are symmetric about zero. In the following we will be interested in mixtures of the form

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = \int_{\Theta}\int_{X} \prod_{i=1}^{n} F(x_i - \theta)\, \mu(dF, d\theta)\,, \tag{3.1.4}$$

where the measure $\mu(\cdot,\cdot)$ represents the *a priori* beliefs over $\mathcal{C}_S \times \mathbb{R}$, corresponding to the modelling structure $f(\cdot)$ and the parameter θ. For a multivariate generalisation in the setting of an array of partially exchangeable random variables, see Aldous (1981).

First note that the necessary symmetry property to characterise a mixture of the form (3.1.4) is the notion of a conditionally location symmetric process (Freedman and Diaconis (1982)). The concept is a natural one given the "what if" principle where we require to report a function of the data sequence which in turn will be interpreted as a consistent estimator of the location parameter. It naturally extends an idea of Gauss (1821) who originally proposed the mean as an interesting summary statistic which also happened to be appropriate for the special case of the normal family (Huber (1972)).

A further decomposition of the representation (3.1.4) for the joint density, $F_{\boldsymbol{X}}(\cdot)$, is as an independent and identical mixture of a location-scale family, denoted by $F(\cdot)$, given by

$$F_{\boldsymbol{X}}(x_1,\ldots, x_n) = \iint \prod_{i=1}^{n} F\left(\frac{x_i-\theta}{\lambda\sigma}\right) \mu(\theta, \lambda, \sigma, dF)\,.$$

In order to obtain interesting representations of the measure μ, the properties of such a representation will be assessed by the Bayes risk of the decision problem of reporting the parameter of interest, either θ or σ.

For the location problem the implicit symmetry involved decomposes the modelling criterion (3.1.2) naturally as a sum of terms, one involving the function $F(\cdot)$ the other a function of the *a priori* assumptions. The optimal sub-family of measures is obtained by minimising the Bayes risk over a space $\mathcal{C}$ defined, generally, by moment constraints.

Furthermore, the criterion (3.1.2) proposes Fisher's information as a nonparametric roughness measure, previously justified on grounds of approximation via a Kullback-Leibler measure. (Good and Gaskins (1971)). Such a roughness penalty leads to the theory of exponential splines.

The representation (1.1.2) allows flexible modelling of the data distribution through the measure $\mu(\cdot)$. Here (μ, σ) will be interpreted as extrinsic parameters whereas λ, together with $F(\cdot)$,

The representation (1.1.2) allows flexible modelling of the data distribution through the measure $\mu(\cdot)$. Here (μ, σ) will be interpreted as extrinsic parameters whereas λ, together with $F(\cdot)$, represents model flexibility. One possible scenario in the parametric setting is to assume a particular functional for the part of μ that assigns weight to dF. The flexibility of the model then depends on the measure $\mu(\theta, \sigma, \lambda)$, the pure modelling component coming from the specification of $p(\lambda|\theta, \sigma)$.

If we assume that $F(\cdot)$ is normal, then by varying the prior on λ we obtain the space of scale mixtures of normality, including the exponential power family (Box and Tiao (1973), West (1987)), the t-family (Fraser (1976), Relles and Rogers (1977)) and the logistic distribution (Andrews and Mallows (1974)). For a survey of applications for Bayesian robustness, see Dempster (1975), Smith (1983).

By virtue of the fact that the flexibility is determined via the *a priori* assumption $p(\lambda|\mu, \sigma)$, the decision theoretic procedure will reduce to the determination of reference-type priors, the information gain and corresponding constraints. However, care must be taken with the interpretation of extrinsic parameters as λ varies (Simar (1983)).

The asymptotic information gain can be decomposed in a number of ways depending on the nature of the *a priori* input, for example, under independence and orthogonality (see (3.1.6)) we obtain

$$I^{\theta,\lambda} = I^{\lambda} + I^{\theta}$$

Furthermore, from the modelling criterion (3.1.2), under a parameterisation for which Jeffrey's prior is lebesgue measure, the asymptotic information gain is purely a function of the prior entropy under that parameterisation.

Example : The scale parameter

Fisher's information for the scale can be decomposed using the transformation $x \rightarrow \frac{x}{\lambda}$,

$$I(\lambda) = \frac{1}{\lambda^2} I(1) ,$$

equivalently, in terms of a derivative with respect to x,

$$I(\lambda) = \int \left(\frac{x}{\lambda} \frac{d}{dx} \log p(x) \right)^2 p(x) dx - \frac{1}{\lambda^2} .$$

Let $G(p)$ denote the functional

$$G(p) = \int \left(x \frac{d}{dx} \log p(x) \right)^2 p(x) dx ,$$

therefore, the asymptotic gain in information about λ is

$$I^{\lambda} = \tfrac{1}{2} \log G_p(\cdot) - \int (\log \lambda) p(\lambda) d\lambda - H(p(\lambda)) .$$

Thus the missing information about λ, which induces a distance on the space of possible choices for the measure $\mu(\cdot)$, shows that we need only consider the properties of the functional $G_p(\cdot)$. The optimal selection for μ will then assign its weight to the $p(\cdot)$ that minimises G over the space $\mathcal{C}$.

Example : Solution for a moment class

Let the space $\mathcal{C}$ be defined by the constraint that, conditional on μ and λ, the first and second moments of $p(\cdot)$ are known, which, by convention can be achieved by the definitions of μ and λ. The following lemma determines the gamma distribution as the optimal selection for $p(\cdot)$.

Lemma (3.1) : In the class of measures with given first and second moments, $\min_{p(\cdot)} G_p(\cdot)$, is attained by the gamma family.

Proof : An algebraic proof is contained in Kagan, Linnik and Rao (1973) (theorem 13.1.2). Here we give a variational argument where the extremal of the above calculus of variations problem is obtained from the Euler-Lagrange equation,

$$\frac{d}{dx}\left(x^2 \frac{d}{dx} \log p(x) \right) + \frac{1}{2}\left(x \frac{d}{dx} \log p(x) \right)^2 = \alpha_1 x - \alpha_2 x^2 .$$

Clearly, a solution on $\mathbb{R}^+$ for some $\alpha, \beta \in \mathbb{R}$ is

$$x \frac{d}{dx} \log p(x) = \alpha - \beta x ,$$

therefore identifying $p(x)$ as the gamma family

$$p(x) \propto x^{\alpha} \exp(-\beta_2 x) .$$

(3.1.4) Further examples

The nature of the model elaboration clearly depends on the structure of the input class $\mathcal{C}$. Thus it is of interest to determine solutions to a wide range of possible subjective assessments of the class $\mathcal{C}$; for example, a moment, a quantile or a distance constraint (e.g. variational norm).

First, suppose that $\mathscr{C}$ is determined by the natural constraint that the variational distance of any element is at most ε from the normal; denote this class by $\mathscr{C}_{VN}$. The following result, from classical robustness, aids in the construction of a formal Bayesian elaboration.

Theorem (Huber (1964)) : Consider the symmetric variational contaminated neighbourhood of normality, denoted by $\mathscr{C}_{VN} = \left\{ F \mid \sup_x |F(x) - \Phi(x)| \leqslant \varepsilon \text{ , } f \text{ symmetric} \right\}$. Then the solution to the variational problem, $\min_{\mathscr{C}_{VN}} I_{f(\cdot)}(\cdot)$, is given by the symmetric density,

$$\begin{aligned} f_0(x) &\propto (\cos\tfrac{1}{2}ax)^2 && 0 \leqslant x \leqslant a \\ &\propto \phi(x) && a \leqslant x \leqslant b \\ &\propto \exp(-b(x-b)) && x \geqslant b \end{aligned} \tag{3.1.5}$$

for suitable a, b. (see also Bickel (1981)). For the case $\varepsilon \geqslant 0.3$, and the handling of quantile classes see Sacks and Ylvisaker (1972).

The form of $f_0(x)$ is appealing for large values of a, for then the density is uniform in the centre, followed by the normal and finally possesses exponential tails. This adds further support to the pragmatic use of the computationally convenient three point mixture class, $\{\lambda_D, \lambda_N, \lambda_U\}$ (Spiegelhalter (1981)). It should mimic the density (3.1.5) closely, and from a subjective viewpoint it readily lends itself to *a priori* and *a posteriori* interpretation.

Secondly, suppose that the class $\mathscr{C}$ is purely defined by a constraint on the first and second moments of $f(\cdot)$, the first moment being implicit by the definition of the location problem. In a sense, ensuing solutions can be viewed as an initial model for the basis of experimentation.

(i) Normal and double exponential families

The following properties of Fisher's information can be deduced from the Euler-Lagrange equations for its functional minimisation (cf. (3.1.6)). Alternatively, they hold by virtue of the fact that the normal and double exponential are special cases of the Huber family.

(i) Let $\mathscr{C}_\theta = \left\{ F \mid \int x dF(x) = 0, f \text{ symmetric} \right\}$, then the variational problem $\min_{\mathscr{C}_\theta} I_\infty^\theta$ is attained by the double exponential distribution

(ii) Let $\mathscr{C}_\sigma = \left\{ F \mid \int x dF(x) = 0, \int x^2 dF(x) = \sigma^2 \right\}$, then the variational problem $\min_{\mathscr{C}_\sigma} I_\infty^\theta$. has solution given by the normal family.

Thus the proposed methodology naturally yields the pair $\{\lambda_D, \lambda_N\}$ as possible elementary building blocks for initial input into an analysis. By imposing further structure on the class $\mathscr{C}$, elaborations of such a framework; for example, the exponential connection, the Huber family, the class $\{\lambda_D, \lambda_N, \lambda_U\}$ are obtained.

(ii) Logistic distribution

Consider the problem of minimising Fisher's information over the space of all probability measures $\mathscr{P}$. The required functional minimisation is that of

$$B(\cdot, p, p') = \int \left(\left(\frac{p'}{p}\right)^2 p + \alpha(p-1) \right) \tag{3.1.6}$$

where α is a Lagrange multiplier. The Euler-Lagrange equations,

$$\frac{d}{dx}\left(\frac{\partial B}{\partial p'}\right) - \frac{\partial B}{\partial p} = 0 ,$$

yield the following differential equation for the score function $u(x) = \frac{d}{dx}\log p(x)$,

$$\frac{du}{dx} + \frac{1}{2}u^2 = \frac{1}{2}\alpha .$$

By direct substitution a solution for the whole of $\mathbb{R}$, with $\alpha = 1$, is given by

$$u(x) = \tanh\left(\frac{x}{2}\right)$$

so that the density $p(x)$ is logistic,

$$p(x) = \frac{e^{-x}}{(1+e^{-x})^2} .$$

Furthermore, as α varies, the logistic distribution with any mean and variance can be obtained. Thus the modelling criterion yields the logistic error structure as a formal Bayesian decision-theoretic solution. A further interesting property of the above density is that it is a scale mixture of normality (Andrews and Mallows (1974)).

(iii) Application to the scale case

Lemma (3.1) shows that the gamma family can be considered as an elementary building block for the study of a scale parameter. It is possible to input further structure into the minimising class $\mathscr{C}$ as it the location case. The solutions can be determined by virtue of the fact that if one requires the estimation of the scale parameter of a random variable X then this is equivalent to the

estimation of the location parameter for the quantity $Y = \log X^2$. Thus the results for the location case directly apply. Suppose that the initial model for X is again normal. In order to interpret the location results for the scale case we note that the corresponding density for Y is given by

$$f_Y(u) = \frac{1}{\sqrt{2\pi}}\exp\left(-\tfrac{1}{2}e^u + \tfrac{1}{2}u\right)$$

thus Bayesian solutions will take the form of a Huberised extreme value distribution for contaminated classes.

(iv) Compact parameter spaces

The techniques of the previous section can be applied to the compact parameter case and generalised to the multivariate case. Here we review some of the classical results of Huber (1974), Bickel (1981) concerning minimisation properties of Fisher's information. Then, under criterion (3.1.1), these will yield Bayesian model elaborations.

Theorem : The solution to $\min_{f(\cdot)} I_{f(\cdot)}(\cdot)$ such that $f(\cdot)$ concentrates on the interval $[-1,1]$ is,

$$u_1(x) = \cos^2\left(\frac{\pi}{2}x\right) \quad |x| \leqslant 1$$
$$= 0 \qquad \text{otherwise} \qquad (3.1.7)$$

Furthermore, the extremal value for Fisher's information is,

$$I_{u_1(\cdot)}(\cdot) = \pi^2 .$$

Proof : The Euler-Lagrange variational equation is given by

$$\frac{(\sqrt{u_1})''}{\sqrt{u_1}} = -\frac{\pi^2}{4}$$

By direct substitution (3.1.7) satisfies the necessary condition.

Multivariate case

Theorem : The distribution $u_{1p}(\cdot)$ that uniquely minimises $I_{f(\cdot)}(\cdot)$ among all spherically symmetric $f(\cdot)$ concentrating on the unit sphere is given by

$$u_{1p}(\|x\|) = c_p \|x\|^{-2t} J_t^2(\|x\|\gamma_t) \qquad \|x\| \leqslant 1 \qquad (3.1.8)$$

where $t = \frac{1}{2}p - 1$ if p is odd or divisible by 4, $t = -(\frac{1}{2}p - 1)$ if p is even and not divisible by 4. Here J_t denotes the Bessel function of the first kind of order t and γ_t denotes its first zero.

Furthermore,

$$I_{u_{1,p}}(\cdot) \;=\; 4\gamma_t^2 \;.$$

Proof : Using the rotational symmetry, transform to polar coordinates and re-express the variational problem as

$$\min_{f(\cdot)} \int_0^\infty r^{p-1}\frac{(f'(r))^2}{f(r)}dr \;\text{ subject to }\; \int_0^\infty r^{p-1}f(r)\,dr$$

since the relevant part of the Jacobian is r^{p-1}. Thus the Euler-Lagrange equations yield the following differential equation for the score function ψ',

$$2\frac{d}{dr}(r^{p-1}\psi') - \gamma_t^2 r^{p-1}(\psi')^2 \;=\; 0\,.$$

It is noted that for odd p the expression (3.1.8) can be re-expressed in terms of rational and elementary functions (Whittaker and Watson (1927)). For example, the case $p = 3$ gives the solution,

$$u_{13}(r) \;=\; \frac{1}{2\pi}\frac{\sin^2(\pi r)}{r^2} \qquad 0 < r < 1$$

$$\;=\; 0 \qquad\qquad\qquad\qquad \text{otherwise}$$

where $\gamma_{\frac{1}{2}} \;=\; \pi$.

Hence it is possible to explore optimal properties of the asymptotic information gain on a compact parameter space. The elementary functions, cosine and sine, appear to play a central role. The characterisations require extension to include well-known families, for example, the von Mises-Fisher distribution.

(v) Finite mixture models

Instead of applying modelling criterion (3.1.2) to scale mixtures of normality to obtain flexible models, consider the case where we require to elaborate on the location structure itself. This then allows the possibility of multimodal densities in our model elaboration procedure, useful in the analysis of outliers and clusters. Clearly, by virtue of the lack of symmetric shapes the solutions will be correspondingly hard to compute. Here we outline one possible scenario.

Consider the class of location mixtures of normality with discrete weight α at the origin denoted by

$$\mathcal{C}_L{}^{\alpha} \;=\; \left\{ f(\cdot) \mid f(x) = \int_{-\infty}^{\infty} \phi(x-u)\,dG(u),\; dG(0) \;=\; \alpha \right\}.$$

Here the weight on the original model is α and $G(\cdot)$ represents our mixing measure, such a class being useful in the event of possible outliers. The parameter α and criterion (3.1.2) will induce a finite mixture model elaboration of the standard normal distribution. The necessary variational minimisation is,

$$\min_{\mathcal{C}_L^{\alpha}} I_{f(\cdot)}(\cdot) \ .$$

An analytical solution is not available at present although the conjectured solution (Mallows (1978)) is a finite mixture model with geometrical weights on the location family of the form,

$$f(x) \ = \ \sum_j p_j \phi(x - g_j)$$

where, $p_j = cp^j$, $g_j = jg$, $p_0 = \alpha$, thus giving a symmetric mixture whose weights are uniformly spread and decrease geometrically. There are clearly close links here with the Bose-Einstein distribution, itself a maximum entropy solution.

(3.2) Calculation of the Bayes risk for a modelling framework.

Consider the "pure inference" problem of reporting beliefs $p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{x})$, so that the Bayes risk under a local and honest utility structure is given by (Bernardo (1979b)),

$$I^{\boldsymbol{\theta},\boldsymbol{\lambda}} \ = \ \int p(\boldsymbol{x}) \left(\int \int p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{x}) \log \left(\frac{p(\boldsymbol{\theta}, \boldsymbol{\lambda}|\boldsymbol{x})}{p(\boldsymbol{\theta}, \boldsymbol{\lambda})} \right) d\boldsymbol{\lambda} d\boldsymbol{\theta} \right) d\boldsymbol{x} \ .$$

After algebraic manipulation, we obtain,

$$I^{\boldsymbol{\theta},\boldsymbol{\lambda}} \ = \ \int p(\boldsymbol{x}) \int p(\boldsymbol{\lambda}|\boldsymbol{x}) \log p(\boldsymbol{\lambda}|\boldsymbol{x}) d\boldsymbol{\lambda} d\boldsymbol{x} \ - \ \int p(\boldsymbol{\lambda}) \log p(\boldsymbol{\lambda}) d\boldsymbol{\lambda} \ + \ \int p(\boldsymbol{\lambda}) I^{\boldsymbol{\theta}|\boldsymbol{\lambda}}(p(\boldsymbol{\theta}|\boldsymbol{\lambda})) d\boldsymbol{\lambda}$$

which leads to a decision theoretic interpretation of the corresponding formulas in Borth (1975), Perrichi (1984) where,

$$I^{\boldsymbol{\theta}|\boldsymbol{\lambda}} \ = \ \int p(\boldsymbol{x}|\boldsymbol{\lambda}) \int p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{x}) \log \left(\frac{p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{x})}{p(\boldsymbol{\theta}|\boldsymbol{\lambda})} \right) d\boldsymbol{\theta} d\boldsymbol{x}$$

$$= \ \int p(\boldsymbol{\theta}|\boldsymbol{\lambda}) \log \left(\frac{\exp\left(\int p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\lambda}) \log p(\boldsymbol{\theta}|\boldsymbol{\lambda}, \boldsymbol{x}) \right)}{p(\boldsymbol{\theta}|\boldsymbol{\lambda})} \right) d\boldsymbol{\theta} d\boldsymbol{x} \ . \qquad (3.2.3)$$

In terms of conditional information gains (3.2.3) can be rewritten as,

$$I^{\boldsymbol{\theta},\boldsymbol{\lambda}} \ = \ I^{\boldsymbol{\lambda}} \ + \ E_{\boldsymbol{\lambda}} \left(I^{\boldsymbol{\theta}|\boldsymbol{\lambda}} \right) ,$$

expressed as a sum of model information gain about $\boldsymbol{\lambda}$ and an expected within model information gain about $\boldsymbol{\theta}|\boldsymbol{\lambda}$.

(3.2.1) Asymptotic information gain

The techniques employed in section (3.1) can be used to define the concept of the amount of missing information about (θ, λ) by determining the asymptotic information gain. The information gain concerning the model parameter λ, will describe the behaviour of the model elaboration.

First, define the entropy function, $H(\cdot)$, by

$$H(p(\lambda|x)) \;=\; -\int p(\lambda|x)\log\, p(\lambda|x)\,d\lambda dx\;.$$

The following limits are required as $x \to \infty$. There are two possibilities:

(i) Suppose $\lambda \in \Lambda$ is discrete then (Rényi (1964), see (3.2.4)),

$$H(p(\lambda|x)) \to 0\;.$$

(ii) Suppose Λ is finite dimensional, under suitable regularity conditions (Ibragimov and H'asminsky (1973), Bernardo (1979b)),

$$\int p(x|\lambda)H(p(\lambda|x))\,d\lambda \;\to\; -\log\left((I(\lambda))^{\frac{1}{2}}\right)$$

where $i(\lambda)$ is Fisher's information. Hence the following limit emerges,

$$\int p(\lambda)\Big(\int p(x|\lambda)H(p(\lambda|x))\,dx\Big)d\lambda \;\to\; 0 \;\text{ if }\; \Lambda \;\text{ discrete}$$

$$\to \;\; -\int p(\lambda)\log\left((I(\lambda))^{\frac{1}{2}}\right)d\lambda \;\text{ if }\; \Lambda \;\text{ continuous}\;.$$

The above limit now defines the missing information about (θ, λ) by

$$\lim_{n\to\infty} I^{\theta,\lambda} \;=\; \int p(\lambda)\log\left((I(\lambda))^{\frac{1}{2}}\right)d\lambda \;-\; \int p(\lambda)\log\, p(\lambda)\,d\lambda \;+\; \int p(\lambda)I^{\theta|\lambda}d\lambda\;. \qquad (3.2.1)$$

(3.2.2) Determination of $p(\lambda)$

Consider the problem of selecting a reference prior for the model parameter λ by maximising the asymptotic information gain given by equation (3.2.1). The calculus of variations yields,

$$\log\left((I(\lambda))^{\frac{1}{2}}\right) \;+\; I^{\theta|\lambda} \;-\; \log\, p(\lambda) \;=\; \alpha \qquad (3.2.2)$$

for some $\alpha \in \mathbb{R}$. The asymptotic conditional information gain $I^{\theta|\lambda}$ has the equivalent representation,

$$\lim_{n\to\infty} I^{\theta|\lambda} \;=\; \int p(\theta|\lambda)\log\left(\frac{(I(\theta|\lambda))^{\frac{1}{2}}}{p(\theta|\lambda)}\right)d\theta$$

therefore,

$$\lim_{n \to \infty} I^{\theta|\lambda} = \int p(\theta|\lambda)\log\left((I(\theta|\lambda))^{\frac{1}{2}}\right)d\theta - \int p(\theta|\lambda)\log\left(p(\theta|\lambda)\right)d\theta . \qquad (3.2.3)$$

Assume that the parameter of interest θ to be *a priori* independent of λ (i.e. $p(\theta|\lambda) = p(\theta)$). Thus equation (3.2.4) as a function of λ becomes,

$$\lim_{n \to \infty} I^{\theta|\lambda} \propto \int p(\theta|\lambda)\log\left((I(\theta|\lambda))^{\frac{1}{2}}\right)d\theta$$

leading to the general solution,

$$p(\lambda) \propto (I(\lambda))^{\frac{1}{2}}\exp\Big(\int p(\theta)\log\left((I(\theta|\lambda))^{\frac{1}{2}}\right)d\theta\Big) .$$

(3.2.3) Examples : Scale mixtures of normality

Suppose that our beliefs about the infinitely exchangeable sequence $\boldsymbol{X}$ possess orthogonal symmetry. Thus the joint density can be represented as a mixture of normals (Smith (1981), Ressel (1985)). Let θ, the location parameter, be the parameter of interest and σ^2 denote the scale. The above structure written in terms of conditional distributions gives

$$p(x|\lambda, \mu, \sigma) = N(\mu, \lambda\sigma^2) . \qquad (3.2.4)$$

Marginalising with respect to $p(\lambda|\mu, \sigma)$ gives

$$p(x|\mu, \sigma) = \int p(x|\lambda, \mu, \sigma)p(\lambda|\mu, \sigma)d\lambda . \qquad (3.2.5)$$

Consider the *a priori* assumption that λ is independent of μ and σ (*i.e.* $p(\lambda|\mu, \sigma) = p(\lambda)$). Then a decision problem, specifying the aim of the model elaboration, will be defined in order to characterise the nature of $p(\lambda)$ which, via (3.2.5), induces the required flexibility in the functional form of $p(x|\mu, \sigma)$. The setting will be that of reporting selected posteriors for the focus of the elaboration considered as a "pure inference" question. To model the elaboration process, the minimisation of the Bayes risk will be carried out, subject to risk and moment constraints on the modelling parameter λ leading, in a sense, to a quantitative robustness with respect to the elaboration.

(i) Justification of the t-family

Under the framework of (3.1), define the constraint class C_t by

$$\mathcal{C}_t = \left\{ p(\lambda) \mid \lim_{n \to \infty} I^{\lambda} = \alpha_1, E_\lambda(\lambda) = \lambda_N \right\} ,$$

where $\lambda_N = 1$, the index for the normal family, the moment constraint modelling the current model status of the normal family. If θ is the parameter of interest then a clearly sensible criterion, over $\mathscr{C}_t$, for the decision problem of reporting the marginal $p(\lambda|\mu, \sigma)$, is given by

$$\max_{\mathscr{C}_t}\left(E_\lambda(I^{\theta|\lambda}) \right) .$$

The Euler-Lagrange equations yield the following solution for $p(\lambda)$, where m and β represent the Lagrange multipliers for the constraints in $\mathscr{C}_t$,

$$\log\left((I(\lambda))^{\frac{1}{2}}\right) - \log p(\lambda) + m\, I^{\theta|\lambda} = \mu - \lambda\beta .$$

Hence,

$$p(\lambda) \propto (I(\lambda))^{\frac{1}{2}}\exp(m\, I^{\theta|\lambda} - \lambda\beta) .$$

Suppose we have a uniform (reference) prior for θ, thus from the normal framework we obtain,

$$(I(\theta|\lambda))^{\frac{1}{2}} \propto \lambda^{\frac{1}{2}}$$

$$(I(\lambda))^{\frac{1}{2}} \propto \lambda^{\frac{1}{2}} .$$

Hence substituting into (3.2) gives the solution for $p(\lambda)$ as the gamma family

$$p(\lambda) \propto \lambda^{\frac{1}{2}(m-1)}\exp(-\lambda\beta) .$$

By marginalising out λ we can act as if the likelihood has a t-distribution with the degrees of freedom depending on the Lagrange multiplier m.

(ii) Justification of the double exponential

Alternatively, suppose the space of possible mixing measures is $\mathscr{C}_D$, where

$$\mathscr{C}_D = \left\{ p(\lambda) \mid \lim_{n \to \infty} I^\lambda = \alpha_1, E_\lambda(\lambda) = \lambda_N = 1 \right\} .$$

The solution for the mixing measure is from (3.3),

$$p(\lambda) \propto \exp(-\beta\lambda) .$$

By Andrews and Mallows (1974), the relevant marginalised likelihood corresponds to the double exponential distribution.

If the class $\mathscr{C}$ is defined via an inverse moment constraint the stable distribution of index a half is obtained, a member of the inverse Gaussian family whose use for modelling long-tailed data on $\mathbb{R}^+$ was suggested by Kingman (1978).

(iii) Justification of the exponential power family

Suppose *a priori* that the infinitely exchangeable sequence X possesses orthogonal symmetry. The joint density $p(X)$ can be represented as a scale mixture of normals (see Ressel (1985)), a result attributed to Schöenburg (1938).

In a hierarchical fashion it seems natural to consider the prior mixing measure to also admit such a representation, itself requiring a mixing measure over the normal family. Clearly, repeating the above process yields a class of measures that can be represented as EE...EEM's (elaborated ... etc.) under the given symmetry condition. Then by marginalising with respect to the final prior measure the corresponding likelihood structure is formed.

An interesting application of such a procedure is given by the three parameter exponential power family, for it is a scale mixture of normality with mixing measure given by a stable distribution (Kanter (1975), West (1987)). Moreover, the stable family of distributions is closed under repeated scale mixing of normality, as the stable distribution of index a is a scale mixture with respect to that of index $\frac{1}{2}a$.

The methodology for section (3.1.2) is applied to determine the reference prior $p(\lambda)$. The form of which highlights the two special cases λ_D and λ_U which have previously been proposed by Spiegelhalter (1981) as a model elaboration with particular application to small sample sizes. In view of the general modelling criterion (3.1.2) it seems sensible to increase the model elaboration at the rate $k = O(n/\log n)$. The class of Spiegelhalter (that is $\{ \lambda_D, \lambda_N, \lambda_U \}$) corresponding to $k = 3$ should therefore perform well for $n \leqslant 15$. By assumption, the likelihood is given by

$$\log p(x|\lambda, \mu, \sigma) = \log a(\lambda) - \tfrac{1}{2}\left|\frac{x-\mu}{\sigma}\right|^{\frac{2}{1+\lambda}} - \log \sigma ,$$

with this definition, the Fisher information of σ given λ is,

$$I(\sigma|\lambda)^{\frac{1}{2}} = \frac{1}{\sigma}\left(\frac{1-\lambda}{1+\lambda}\right)^{\frac{1}{2}} .$$

The solution for $p(\lambda)$ becomes,

$$p(\lambda) \propto (I(\lambda))^{\frac{1}{2}}\left(\frac{1-\lambda}{1+\lambda}\right)^{\frac{1}{2}}$$

yielding the two special cases $\lambda = -1, 1$, corresponding to the λ_U and λ_D distributions, respectively.

(3.2.4) Discrete case

Consider two discrete probability measures, P and Q, with finite probability vectors, denoted by $(p_1,\ldots, p_n)$ and $(q_1,\ldots, q_n)$. The asymptotic behaviour of the information gain concerning θ is well behaved. Rényi (1974) proves that the missing information about a discrete parameter θ is precisely the prior entropy, $H(p(\theta))$, furthermore the rate of convergence is exponential with sample size, n, gauged by a constant, $\lambda < 1$, in turn related to the α-distance. The result is as follows,

Theorem (Rényi (1964)) : Let θ be a discrete random variable. Let $X_1,\ldots,X_n,\ldots$ be discrete and conditionally independent given θ. Then the following holds,

$$0 \leqslant I_n^{\theta} - H(\theta) \leqslant A\lambda^n \tag{3.2.6}$$

for some constant A, where

$$\lambda = \min_{\alpha} \max_{i \neq j} \int p_i^{\alpha} p_j^{1-\alpha} d\mu ,$$

where p_i corresponds to the likelihood when $\theta = \theta_i$. Then $0 < \lambda < 1$ and therefore

$$I_{\infty}^{\theta} = H(\theta) ,$$

independent of the process generating the data. Thus the decision theoretic criterion of maximising the missing information about θ in the discrete case reduces to the maximum entropy principle (Jaynes (1982b)).

Further properties of decision problems can be inferred via the bound (3.2.6). For example, the rate of convergence together with an application of the Borel-Cantelli lemmas proves that if we select the highest posterior probability the correct decision will be made with probability one.

A problem that has received little attention is that of the behaviour of the discrete posterior weights, for example, in hypothesis testing and the analysis of mixture models. The constraint of coherence via the updating mechanism of Bayes theorem imposes constraints on the possible posterior weights, sometimes restrictive, depending on the nature of the prior assessments. Formally, the behaviour of the Bayes map is rather surprisingly related to the Kullback-Leibler distance (Matsubara (1976)).

It is also possible to explore probability models with optimal properties with respect to the asymptotic information gain as in (3.1.2). One such result concerning the maximisation of entropy in a class of probabilistic models is the following.

Theorem (Rényi (1964)) : For every $T \geq 0$, among all homogeneous point processes with a given rate $\lambda > 0$, the Poisson process has the greatest entropy in the interval $(0, T)$.

(3.2.5) Infinite discrete case

Consider two countably infinite discrete probability measures P and Q which are of the form $P = \{ p_1, p_2, \ldots \}$ and $Q = \{ q_1, q_2, \ldots \}$ where $(p_i, q_i > 0)$. The risk of approximating P by Q as defined by the Kullback-Leibler distance is given by

$$I(P, Q) = \sum_{i=1}^{\infty} p_i \log \left(\frac{p_i}{q_i} \right).$$

Here Q might denote the posterior associated with the prior P. Unfortunately, a general result for the asymptotic convergence of the posterior measure and asymptotic gain in information is not available. In fact the behaviour of the Kullback-Leibler distance is highly non-regular. Here we note a result of Csiszár (1967b) showing one such adverse property. Let $N(P, \varepsilon)$ denote the ε-neighbourhood of P under the Kullback-Leibler distance measure. Then, for $\varepsilon > 0$, there exists $Q \in N(P, \varepsilon)$ such that for any $\varepsilon_1 > 0$ there exists $R \in N(Q, \varepsilon_1)$ for which $I(R, P) = \infty$, clearly an undesirable feature for a Bayesian risk. In the continuous case, however, the Bayes risk for a parameterised family does possess continuity properties, Berk (1966), Loh (1984). However, by following Rissanen (1983), we can construct (under suitable smoothness conditions on P) a measure Q that stabilises the Bayes risk for all $\mathcal{P}$.

Theorem : Let P satisfy the following regularity conditions (Rissanen (1983)),

(i) $p_i < 1$ for all i, such that there exists M such that $p_{i+1} \leq p_i$ for $i > M$.

(ii) $H(P) = \infty$.

Then there exists Q such that, for all P,

$$\lim_{n \to \infty} \sum_{i=1}^{n} p_i \log \left(\frac{p_i}{q_i} \right) < \infty$$

where Q is given by

$$q_i = 2^{-\log^*(i)},$$

where for $i \geq 1$, $\log^*(i) = \log i + \log \log i + \ldots$ up to the last non negative term.

Proof : A direct application of Appendix 2 (Rissanen (1983)).

Note that Q defines a proper probability measure and is the universal modelling prior suggested in Rissanen (1983) by virtue of its minimum description length properties and the fact that it satisfies Kraft's inequality. However, its interpretation here is solely within a Bayesian decision theoretic framework of reporting the posterior distribution in order that the risk is well behaved, and is strongly related to the concept of a reference prior (Bernardo (1979b)).

(3.2.6) Asymptotic information gain, orthogonality and independence

Consider a model of fixed dimension parameterised by the vector (θ, ϕ), ϕ denoting the nuisance parameter. The modelling criterion (3.1.1) proposes the assessment of the full model via the asymptotic information gain $I_\infty^{(\theta,\phi)}$. Suppose it is required to select a suitable parameterisation for the nuisance parameter ϕ. Note that the full information gain is invariant to reparameterisation whereas the conditional information gain, $I_\infty^{\theta|\phi}$, is not. Thus the latter can be applied to assess the choice of ϕ. It possesses a natural interpretation via the relative asymptotic information gain,

$$I_\infty^{(\theta,\,\phi)} - I_\infty^{\theta} = E_\phi\left(I_\infty^{\theta|\phi}\right).$$

If the orthogonal parameterisation exists then it is an optimal solution by virtue of the fact that it maximises the relative asymptotic information gain as follows, by definition,

$$I_\infty^{\theta|\phi} = \int p(\theta|\phi)\log\left(\frac{|I_{\theta|\phi}|^{\frac{1}{2}}}{p(\theta|\phi)}\right)d\theta\,,$$

the conditional Fisher's information, $I_{\theta|\phi}$, defined by

$$I_{\theta|\phi} = I_{\theta\theta} - I_{\theta\phi}I_{\theta\theta}^{-1}I^{\theta\phi}\,.$$

However, in the sense of positive semi-definiteness (Amari (1982)),

$$I_{\theta|\phi} \leqslant I_{\theta\theta}$$

with equality if and only if $I_{\theta\phi} = 0$, that is ϕ is orthogonal to θ. Thus the conditional asymptotic information gain is maximised under such a parameterisation as

$$I_\infty^{\theta|\phi} = \int p(\theta|\phi)\log\left(\frac{|I_{\theta\theta}|^{\frac{1}{2}}}{p(\theta|\phi)}\right)d\theta\,.$$

Furthermore,

$$I_\infty^{\theta|\phi} = I_\infty^{\theta\theta} \quad\Leftrightarrow\quad p(\theta|\phi) = p(\theta)\,.$$

Under orthogonality and *a priori* independence the relative information gain decomposes as

$$I_\infty^{(\theta,\,\phi)} - I_\infty^{\phi} = I_\infty^{\theta}$$

thus representing the least sensitive parameterisation.

To examine the existence of an orthogonal parameterisation we follow Amari (1982a). Theorem (8.2) of the latter reduces the required condition to that of the integrability of a partial differential equation. The corollary that for a scalar parameter of interest an orthogonal parameterisation always exists is of fundamental importance as it gives the concept an operational definition for an arbitrary statistical model. The solution, however, of the associated partial differential equation is often highly intractable (Hills (1987a)). Sensible approaches suggested on the grounds of approximation are available by only requiring local orthogonality. The Riemannian geometry of the underlying parameter space and the existence of the stronger condition of a covariance stablising transformation is discussed in Kass (1981). Again, however, such a condition fails for some simple parametric models (Holland (1973)).

(3.2.7) Comparison of experiments

A fruitful area of application of the decision criterion (3.1.2) is to the comparison of information in experiments. This has received little attention in a Bayesian framework since the foundational work of Lindley (1956), Stone (1959). However, the classical literature (Torgersen (1976), (1981)) contains numerous calculations of distance measures which by virtue of the characterisation results in Chapter 2 can be interpreted in a Bayesian setting.

First, consider the problem of combining information from two sources. Suppose the vector parameter (θ, ϕ) is such that the data decomposes as $z = (x, y)$ with joint density factorising as

$$p(x, y|\theta, \phi) = p(x|\theta)p(y|\phi) .$$

By virtue of the linearity of Bayes theorem for the logarithmic utility the combination of information can be quantified by the decomposition of information gain,

$$I^{\theta}_{x,y} = I^{\theta}_{y|x} + I^{\theta}_{x} ,$$

where $I^{\theta}_{y|x}$ represents the additional information gain about θ from the component y after observing data x.

An intuitive result is that *a priori* independence should be characterised by the property that no additional information is learnt about θ in the light of observing y. Formally, if the family P_{ϕ} is complete, the following characterisation, holds (Stone and Springer (1965))

$$I^{\theta}_{y|x} = 0 \quad \Leftrightarrow \quad p(\theta,\phi) = p(\theta)p(\phi)$$

that is θ and ϕ are *a priori* independent.

(3.3) Application of Ressel (1985) for the construction of model elaborations

Suppose that our current modelling framework is denoted by M_0, corresponding to a joint belief $p(x|M_0)$ for our infinitely exchangeable sequence X. It is required to build an elaborated model $p(x|M_0, \lambda)$ that collapses to the current model when $\lambda = \lambda_0$. The joint beliefs concerning X are determined, by marginalisation, with respect to a measure for λ, denoted by $p(\lambda)$.

$$p(x) = \int p(x|M_0, \lambda)p(\lambda)d\lambda .$$

In order to model the embedding of the current model we will assume that the following smoothness properties in the form of moment constraints hold, for the prior measure on λ

$$\int \lambda p(\lambda) = \lambda_0 \tag{3.3.1}$$

$$p(x|M_1) = \phi_n(t(x)) \tag{3.3.2}$$

for an arbitrary function ϕ_n and some statistic $t(x)$.

The constraint (3.3.2) asserts that the ensuing estimation, via $p(x|M_1)$, for the model parameter is smooth as a function of the statistic $t(x)$. A similar technique is employed in Goldstein (1974) where the space of possible estimation functions are polynomial in x (see section (5.4)). The possible modelling families defined by assumptions (3.3.1) and (3.3.2) are extensive. First, note that the latter induce a constraint on the moments of $p(X)$ given by

$$E_x(\phi_n(t(x))) = E_x(p(x|M_1)) = E(\lambda) = \lambda_0 . \tag{3.3.3}$$

Consider the decision problem of reporting beliefs $p(x)$ relative to the current model, $p(x|M_0)$ with respect to a local and honest utility structure, with Bayes risk equal to the Kullback-Leibler divergence between the two models. The model elaboration is determined by minimising the Bayes risk under the constraint (3.3.3). The required calculus of variations minimisation is,

$$\min_{p(x)} \int p(x)\log\left(\frac{p(x|M_0)}{p(x)}\right)dx \text{ subject to } E_x(\phi_n(t(x))) = \lambda_0 .$$

Let α_1, α_2 be Lagrange multipliers and consider the functional,

$$\int p(x)\left(\log\left(\frac{p(x|M_0)}{p(x)}\right) - \alpha_1\phi_n(t(x)) - \alpha_2\right)dx .$$

By the Euler-Lagrange equations, the extremal necessarily satisfies,

$$\log\left(\frac{p(x|M_0)}{p(x)}\right) + \alpha_1\phi_n(t(x)) = \alpha_2 .$$

Therefore,

$$p(x) = p(x|M_0)\psi_n(t(x)) \tag{3.3.4}$$

describes the space of possible structures for $p(x)$ where ψ_n is arbitrary and $t(x)$ a given statistic.

The following theorem is given in Ressel (1985) which proves that a constraint of the form (3.3.4) is sufficient to characterise the space of possible measures for the functional form of $p(x|M_0, \lambda)$ as a mixture class. The technique applied is that of harmonic analysis on abelian semi-groups. Here we state an abbreviated version of the full theorem.

Theorem : Let the measure P have the property that

$$P(x) = \prod_{j=1}^{n} \beta(x_j)\phi(\sum_{j=1}^{n} t(x_j)) \tag{3.3.5}$$

for all $n \geq 1$ and all $x \in X$, where $\phi(0) = 1$. Then ϕ has a unique representing measure that concentrates on the (relatively compact Borel) set,

$$W_\beta = \{ \rho \mid \sum \beta(x_j)\rho(t(x_j)) = 1 \} . \tag{3.3.6}$$

Conversely, for each measure μ on W_β, the function $\phi(s) = \int \rho(s)d\mu(\rho)$ defines a probability measure via (3.3.5).

Proof : See Ressel p.907.

The representation of $\phi(\cdot)$, via (3.3.6), yields an integral representation for the marginal beliefs about x given by

$$P(x) = \int \prod_{j=1}^{n} \beta(x_j)\rho(t(x_j))d\mu(\rho) .$$

The following examples exhibit the power of the approach where the above theorem is applied in the context of equation (4.4), identifying suitable forms for $\beta(\cdot)$ and $\phi(\cdot)$ (see Ressel p.908).

Examples :

(i) **Mixtures of Poissons.** Let the current model $p(x|M_0)$ be i.i.d. Poisson random variables with mean one.

$$p(x|M_0) = \frac{1}{\prod_{j=1}^{n} x_j!} .$$

If the statistic $t(x) = \sum_{j=1}^{n} x_j$ then the family of mixtures of Poissons are characterised by (3.3.4). Similarly mixtures of binomials and inverse binomials can be characterised (Freedman (1962)).

(ii) **Mixtures of uniforms.** Let the sample space be $\mathbb{N}$ and statistic $t(x) = x_{max}$. If the current model is lebesgue measure, then the criterion characterises discrete mixtures of the uniform distribution. A further characterisation is given in Dawid (1982).

(iii) **Mixtures of exponentials and normals.** The characterisation of mixtures of these families can be viewed in a unified manner via symmetry conditions on the associated characteristic function of the joint density.

The current model can be interpreted, via the statistic $t(x)$, as a conditional distribution by the equation, see Ressel ,

$$P\left(X_1 = x_1, \ldots, X_n = x_n \,\middle|\, \sum_{j=1}^{n} t(X_j) = s\right) = p(x \,|\, M_0)\chi_{\sum_{j=1}^{n} t(X_j) = s}(s) ,$$

where $\chi(\cdot)$ denotes an indicator function, thus establishing a link with the concept of sufficiency, which formally interprets the role of $t(x)$ as one of reducing the elaborated model to the current model.

(3.4) Connecting two distributions

Suppose our current modelling framework consists of two plausible families of measures denoted by $\mathcal{P}_A$ and $\mathcal{P}_B$, the corresponding densities, with respect to some dominating measure μ, being given by $f(x \,|\, \alpha)$ and $g(x \,|\, \beta)$. Three typical statistical scenarios are,

(i) Discrimination between possible error structures, for example, normal and double exponential on $\mathbb{R}$, log Normal and exponential on $\mathbb{R}^+$, Poisson and Geometric on $\mathbb{N}$.

(ii) Different choices of functional forms for the regression structure, for example, parabolic versus piecewise linear.

(iii) Existence of changepoints whose number might be unknown.

In order to elaborate on this framework we might find it useful to build a model, which is in some sense is optimal, embedding $\mathcal{P}_A$ and $\mathcal{P}_B$. This hierarchical model then allows us to examine possible departures from one family to the other, using the posterior distribution for the model parameter, λ, obtained by Bayes' theorem (Smith (1983)). The elaboration parameter λ can be discrete, or continuous, depending on the required application.

To quantify the sense in which the connection between the families is optimal we cast the problem in a decision-theoretic setting, adopting the logarithmic utility function which arises naturally in this context as the only "pure inference" utility (Bernardo (1979a)). The risk function is

then interpretable as a Kullback-Leibler distance, defining a risk associated with an arbitrary measure Q, induced via the families $\mathcal{P}_A$ and $\mathcal{P}_B$.

The application of the above decision theoretic procedure leads to a connection, or "shortest" line, between the two families. By varying the possible decision problem the general solution is obtained, the exponential embedding playing a central role as being related to the Kullback-Leibler distance (see Loh (1984)). This has previously been suggested by virtue of mathematical convenience of the additivity of the log-likelihood (Cox (1961, 1962), Atkinson (1970), Brown (1971)).

The procedure for comparing models in a Bayesian framework relies on the associated Bayes factor, itself a measure of distance between the two families. In general the two families, $f(x|\alpha)$ and $g(x|\beta)$ will be separate families. That is, for an arbitrary parameter value α_0, the measure $f(x|\alpha_0)$ cannot be approximated arbitrarily closely by a member of the family $\mathcal{P}_B$. The sense of the approximation can be quantified by the Bayes risk which, in turn, is related to the Bayes factor.

The elaboration technique is applied both to the prior and likelihood spaces. For simplicity, the location-scale problem is used to exhibit statistical properties of the elaborated structure. Let the families $f(x|\alpha)$ and $g(x|\beta)$ be the normal and double exponential families, respectively. The ensuing estimator of location, being adaptive in nature, performs well in relation to previous proposals (Loh (1984)).

(3.4.1) Utility structure

Let $Q \in \mathcal{P}$, and $F \in \mathcal{P}_A$, $G \in \mathcal{P}_B$. Suppose that the families $\mathcal{P}_A$, $\mathcal{P}_B$ are absolutely continuous with respect to some measure μ yielding finitely parameterised densities f, g, respectively. The results from Chapter 2 are directly applicable where the induced distance between elements of $\mathcal{P}_A$ and $\mathcal{P}_B$ is given by

$$d(Q^*; (F, G)) = \min_{Q} \max (I(Q, F), I(Q, G)) . \qquad (3.4.1)$$

Following Loh (1984), a solution exists, and $Q^* \in \mathcal{P}$ whose densities, with respect to μ, take the form,

$$p(x|\alpha, \beta, \lambda) = k_\lambda (f(x|\alpha))^\lambda (g(x|\beta))^{1-\lambda} \quad \lambda \in [0,1] \qquad (3.4.2)$$

for some suitable normalising constant k_λ.

A useful geometrical interpretation for Q^* is that, if the mid-point between the families exists, it is precisely Q^* (Loh (1984)). Hence it is natural to use the density Q^* as a compromise between f and g. By varying λ from zero to one, p_λ traces out the shortest line (or geodesic), which under the decision problem (3.4.1) gives an optimal one parameter model elaboration

through f and g.

It is natural to attempt to justify the linear connection as a decision- theoretic solution. By theorem (2.1) it is the shortest line under the total variation utility measure, and will be termed the ε-contaminated connection. It has a further minimisation property with respect to the logarithmic utility measure by virtue of the Pythagorean interpretation for such a utility see Cencov (1972), Csiszár (1975). Thus the widely used class of discrete mixtures can be viewed as a geodesic surface with respect to a Kullback-Leibler measure (Cencov (1972)).

(3.4.2) Application to prior elaboration

Consider two possible prior densities $p(\theta)$ and $\pi(\theta)$, then the Bayes risk induce a distance, $D(\cdot,\cdot)$ between the priors as follows,

$$D_x(p(\theta), \pi(\theta)) = \int p(\theta)(u(p(\theta|x), p(\theta)) - u(\pi(\theta|x), \pi(\theta)))d\theta .$$

By looking at the asymptotic gain in information the above distance will reduce to the difference in gain of information from the two *a priori* inputs.

$$\lim_{n \to \infty} D_x(p(\theta), \pi(\theta)) = \int p(\theta)\log\left(\frac{I_f(\theta)^{\frac{1}{2}}}{p(\theta)}\right)d\theta - \int p(\theta)\log\left(\frac{I_f(\theta)^{\frac{1}{2}}}{\pi(\theta)}\right)d\theta .$$

Therefore,

$$D(p(\theta), \pi(\theta)) = \int p(\theta)\log\left(\frac{\pi(\theta)}{p(\theta)}\right)d\theta$$

thus inducing the natural measure on the prior space of the implications of approximating one prior by the other.

(3.4.3) Application to location-scale families

Suppose that $f(x|\alpha)$ and $g(x|\beta)$ are, respectively, the normal and double exponential families. The exponential connection given by (3.4.2) between these families can be parameterised as follows,

$$p(x|\theta, \sigma, \tau) = k(\sigma,\tau)\exp(-\tfrac{1}{2}\sigma^2(x-\theta)^2 - \tau|x-\theta|) \qquad (3.4.3)$$

where, for $\sigma, \tau > 0$,

$$k(\sigma,\tau) = \frac{\sigma\phi\left(\frac{\tau}{\sigma}\right)}{2\Phi\left(\frac{-\tau}{\sigma}\right)}$$

and $\phi(\cdot)$, $\Phi(\cdot)$ denote the standard normal density and cumulative distribution functions. The appropriate definition is given to $k(\cdot,\cdot)$ when either of σ or τ are zero, corresponding to λ_D and λ_N, respectively.

By virtue of the updating mechanism in a Bayesian analysis, the model (3.4.3) together with a measure $p(\theta, \sigma, \tau)$ yields a Bayesian analogue of the *ad hoc* adaptive procedures due to Hogg (1972, 1974) to estimate a location parameter. Heuristically, the maximum likelihood estimators are the median and mean for λ_D and λ_N respectively. Thus for smooth priors the posterior mean will adapt between these two measures.

(3.5) Approximating statistical models with flexible families.

The statistical process of summarisation of complex underlying mechanisms in terms of simplified interpretable models, via subjective input, lies at the centre of the Bayesian methodology (Smith (1983), (1986)). As a first step, without oversimplification, it might be of use to represent

$$Data \;=\; Structure \;+\; Complex\ Noise \tag{3.5.1}$$

but without any interpretation, or formal suggestion, of the form of *complex noise* we may wish to approximate (3.6.1) by

$$Transformed\ Data \;=\; Transformed\ Structure \;+\; Simplified\ Noise\ . \tag{3.5.2}$$

In general there will exist constraints relating the old structure (3.5.1) to the new (3.5.2), for example, we may wish to retain the same interpretation of the location parameter. Choices of the form of *simplified noise* might incorporate the normal, or double exponential distributions. Nevertheless, the reporting of such oversimplified structures requires some degree of sensitivity analysis, warning against possible departures.

(3.5.1) Decision theoretic setting for projecting $\mathcal{P}_\Omega$ onto $\mathcal{P}_\Lambda$.

Consider the two families of probability measures $\mathcal{P}_\Omega$, $\mathcal{P}_\Lambda$ indexed by the sets Ω, Λ, which usually will be finite dimensional and parameterised in $\mathbb{R}^p$. In order to define the notion of approximating, or projecting, one family onto another we adopt a decision-theoretic approach, thus, allowing us to induce a distance on the set $\Omega\times\Lambda$ via the Bayes risk. Consider the distance, denoted by $d(\omega, \lambda)$, generated the logarithmic function (Bernardo (1979a))

$$d(\omega, \lambda) \;=\; \int p_\omega \log\left(\frac{p_\omega}{p_\lambda}\right)$$

for all $(\omega, \lambda) \in (\Omega, \Lambda)$. The closest member of $\mathcal{P}_\Lambda$, given by index λ^*, to an element of $\mathcal{P}_\Omega$ can be defined by

$$d(\omega, \lambda^*) = \inf_{\Lambda} d(\omega, \lambda) .$$

Therefore,

$$d(\omega, \lambda^*) = \inf_{\Lambda} \int (p_\omega \log p_\omega - p_\omega \log p_\lambda) .$$

Hence, λ^* attains,

$$\inf_{\Lambda} \int (-p_\omega \log p_\lambda) . \tag{3.5.3}$$

A formalisation of such an approach to the approximation of probability measures in the notion of an I-projection (Csiszár (1975)), here formally interpreted in a Bayesian decision-theoretic setting.

(3.5.2) Example : Model choice on $\mathbb{R}^+$

In order to exhibit a possible scenario for moving from (3.5.1) to (3.5.2) suppose our data can be modelled as observations from a distribution on $\mathbb{R}^+$. For simplicity assume that the structure is determined by a location parameter μ. Concentrating on the noise term, suppose that

$$\textit{Complex Noise} = \textit{generalised gamma} \tag{3.5.4}$$

$$\textit{Simplified Noise} = \textit{lognormal} . \tag{3.5.5}$$

In the sense of an approximation to the true density we would hope that (3.5.4), denoted by $\mathcal{P}_\Omega$, is sufficient. The corresponding density function is given by

$$p_\omega(x) \propto x^{\kappa\beta - 1} \exp\left(-\left(\frac{x}{\alpha} \right)^\beta \right) . \tag{3.5.6}$$

Unfortunately, the model parameters (α, β, κ) do not lend themselves to a easily interpretable form, making specification of *a priori* beliefs difficult. Therefore we might tentatively assume the structure (3.5.5), with the transformation denoted by ϕ. Denoting this class by $\mathcal{P}_\Lambda$ which is parameterised by $\lambda = (\mu, \sigma, \phi)$ where $\phi \in \Phi$, the set of all, increasing, one-to-one differentiable transformations, the corresponding density function is,

$$p_\lambda(x) \propto \frac{1}{\sigma} \exp\left(-\tfrac{1}{2} \left(\frac{\phi(x) - \mu}{\sigma} \right)^2 \right) \frac{d\phi}{dx} \tag{3.5.7}$$

that is transformed so that the error structure is normal; the lognormal occurs as a special case when ϕ is logarithmic. The model $\mathcal{P}_\Lambda$ now has parameters in an interpretable form, practically allowing elicitation of prior beliefs with some degree of confidence. This contrasts with the class $\mathcal{P}_\Omega$ where, due to the nature of the parameters a careful sensitivity analysis of any prior input would be required.

The methodology from (3.5.3) can now be applied, projecting the family $\mathcal{P}_\Lambda$ onto $\mathcal{P}_\Omega$. Suppose our objective is an appropriate choice of transformation, ϕ, reflecting the fact that $\mathcal{P}_\Lambda$ is an approximation to $\mathcal{P}_\Omega$. Hence we require to determine the behaviour of $d(\omega, \lambda)$ as a function of ϕ. Let $\phi^*(\cdot)$ denote the optimal choice with regard to criterion (3.5.3). Thus, neglecting irrelevant constants,

$$d(Q\ ;\ (\mathrm{F},\ \mathrm{G})) \ = \ \inf_{\Phi}\left(-\int x^{\beta-1}\exp(-x^{\beta})\left\{(\phi(x)-\mu)^2-\log\frac{d\phi}{dx}\right\}dx\right). \tag{3.5.8}$$

If the parameter of interest is μ, we have the following constraint linking the two models,

$$E_{p_\omega(x)}(\phi(x)) \ = \ \mu\ .$$

The following lemma determines the required solution,

Lemma (3.5.1) : Define the functional $B(\phi)$ by

$$B(\phi) \ = \ \int p_\omega(x)\left(\frac{1}{2\sigma^2}(\phi(x)-\mu)^2-\log\frac{d\phi}{dx}\right)dx\ . \tag{3.5.9}$$

Then ϕ^*, defined by $B(\phi^*) = \inf_{\Phi} B(\phi)$, satisfies,

$$\frac{d^2\phi}{dx^2} \ = \ \left(\frac{d\phi}{dx}\right)^2\left(\frac{\phi-\mu}{\sigma^2}\right) + \frac{d\phi}{dx}\frac{d}{dx}\log p(x)\ .$$

Proof : One parameter variations with respect to ϕ are given by the Euler-Lagrange equation,

$$\frac{d}{dx}\left(\frac{\partial B}{\partial \phi'}\right) - \frac{\partial B}{d\phi} \ = \ 0\ .$$

By definition of B,

$$\frac{d}{dx}\left(\frac{p}{\phi'}\right) + p\left(\frac{\phi-\mu}{\sigma^2}\right) \ = \ 0\ .$$

Hence,

$$\frac{d\phi}{dx}\frac{d}{dx}\log p(x) - \frac{d^2\phi}{dx^2} + \left(\frac{d\phi}{dx}\right)^2\left(\frac{\phi-\mu}{\sigma^2}\right) \ = \ 0 \tag{3.5.10}$$

and rearranging gives the required result.

Thus the solution for the transformation ϕ only depends on the score function of $p(x)$, showing the importance of the latter in assessing a transformation. In a sense this establishes a link with concepts from robustness where again it is the score function that quantitatively defines the sensitivity of an estimation problem. Hence, in certain instances, we can hope to achieve robustness via an appropriate choice of transformation ϕ.

Applying lemma (3.5.1) in the context of the previous example leads to the calculation of the score function,

$$\frac{d}{dx}\log p_\omega(x) \;=\; \frac{\kappa\beta-1}{x} \;-\; \beta x^{\beta-1}\,. \tag{3.5.11}$$

By direct substitution, it can be shown that $\phi(x) = x^\lambda$ is a solution to (3.5.8) with the score function of $p_\omega(x)$ defined by (3.5.11). Thus the Box-Cox transformation family (Box and Cox (1964)) can be seen as an one parameter variation for the above decision problem.

(3.5.3) Examples on power transformations

The risk (3.5.3) can be applied to select power transformations to normality. Consider projecting the density of $\frac{1}{\lambda}(x^\lambda - 1)$ onto normality. The following examples are given in Hernandez and Johnson (1979),

(i) Let $X \sim$ Gamma, then criterion (3.5.3) yields the approximate solution $\lambda^* = 1/3$, originally suggested by Wilson and Hilferty (1931).

(ii) Let $X \sim$ Inverse Gaussian, then (3.5.3) gives $\lambda^* = 0$, the logarithmic transformation suggested in Whitmore and Yalovsky (1978).

(iii) Let $X \sim$ Pareto, then (3.5.3) has solution $\lambda^* = 1/\sqrt{2}$.

The above setting allows us to unify previously proposed transformations with a formal decision theoretic framework. Further suggestions for the binomial family are contained in Bernardo (1985b).

(3.5.4) Construction of a family of transformations

Suppose we now restrict the class of possible transformations by imposing model constraints, in the form of moments, before minimising the functional $B(\phi)$. In the context of projecting a density onto the normal family with risk given by $B(\phi)$, as defined by (3.5.9), the following moment constraint has an intuitive appeal,

$$E_{p_\omega(x)}((\phi(x)-\mu)^2) \;=\; \sigma^2\,, \tag{3.5.12}$$

thus matching the corresponding moment for the parameters of interest.

Consider now a possible one parameter variation through the above constraint as defined by $\inf_{\Phi} B(\phi)$. By definition of $B(\cdot)$, (3.5.9) gives,

$$B(\phi) = \frac{1}{2} - \int p_\omega(x)\log\left(\frac{d\phi}{dx}\right)dx .$$

A possible variation, by the Euler-Lagrange equations, is given by

$$\frac{d}{dx}\left(\frac{p}{\phi'}\right) = 0 .$$

Therefore,

$$\phi(x) = \int^x p(y)dy . \qquad (3.5.13)$$

For example, suppose the density for X has a Beta(α, β) distribution. Then a space of possible normalising transformations to search is given by (3.5.13) which includes the logit and inverse sine transformations.

(3.6) A decision theoretic approach to the design of experiments

A unified approach, using concepts from decision theory, is adopted in order to select appropriate design criterion. The properties of the statistical problem, generally that of reporting beliefs, are used to define a decision theoretic criterion. However the criterion can be adapted to incorporate prediction or control as required.

Lindley (1956) proposed the use of maximising expected gain in Shannon information between prior and posterior, denoted by I_n^θ, as a formal criterion for assessing design of experiments. There is a wide range of selection criterion and equivalence theorems in the context of classical design (see, for example, Whittle (1973), Federov (1972), Silvey (1980)). Much importance is placed on Fisher's information in such a framework. Due to the equivalence between the asymptotic information gain and Fisher's information as exhibited in the modelling criterion (3.1.2) the Bayesian can expect that most classical results to be of direct relevance. For example, under a normal linear hierarchical model analytical computations are available leading to Bayesian information design criterion based on posterior variance (Stone (1959), Smith and Verdinelli (1980)). In turn under weak *a priori* information these reduce to D- optimal designs, a result reviewed in section (3.6.1). Further Bayesian results for the linear model are contained in Chaloner (1984) and for the nonlinear situation in Zacks (1977).

To exhibit the power of the approach we consider two possible applications of the criterion (3.1.2) with respect to the design. First, A- and D- optimal designs are viewed as Bayesian decision theoretic procedures. Secondly, the use of modelling criterion (3.1.2) and the reference prior are discussed in a general modelling framework. Further areas of application include the optimal selection of sample size previously viewed in a decision-theoretic setting by Lindley (1956),

Antelman (1965), the Bayesian procedure generalising previous *ad hoc* proposals.

(3.6.1) A- and D- optimality from a decision theoretic perspective

The design of experiments, although containing a vast selection of different criterion, has concentrated mainly on two criteria; A- and D- optimal designs. An equivalence result between such designs and the minimisation of Bayesian risk will be shown.

Consider the normal linear regression model. By assumption our observation, y, is generated by the process

$$y = N(X\theta, V)$$

for some p dimensional parameter θ, which itself is a random variable, with prior density,

$$p(\theta) = N(\theta_0, V_0) .$$

Following Lindley and Smith (1972), application of Bayes theorem gives the posterior distribution,

$$p(\theta|y) = N(Bb, B) , \tag{3.6.1}$$

where,

$$B^{-1} = X^T V^{-1} X + V_0^{-1}$$

$$b = X^T V^{-1} y + V_0^{-1} \theta_0 .$$

(i) A-optimality

Consider the symmetrised Kullback-Leibler distance between prior and posterior, J_n^θ, as quantifying the gain in information of the experiment, that is

$$J_n^\theta = \int p(y) \int (p(\theta|y) - p(\theta)) \log \left(\frac{p(\theta|y)}{p(\theta)} \right) d\theta dy . \tag{3.6.2}$$

By virtue of of normality of the posterior, the inner integral in (1.4) can be determined as,

$$-p + \frac{1}{2}(tr\,(V_0^{-1}B + B^{-1}V_0) + (Bb - \theta_0)^T (V_0^{-1} + B^{-1})(Bb - \theta_0)) .$$

By linearity of the $tr(\cdot)$ operator,

$$E_\lambda\,(tr\,(\cdot)) = tr\,(E_\lambda\,(\cdot)) .$$

Hence,

$$J_n^{\theta} = -p + \frac{1}{2}tr\,(V_0^{-1}B + B^{-1}V_0) + \frac{1}{2}tr\,((V_0^{-1} + B^{-1})V_y(Bb)) .$$

Finally, the variance term can be calculated as follows,

$$V_y(Bb) = BV_y(b)B^T .$$

By definition of b,

$$V_y(Bb) = BX^TV^{-1}(V_y(y))V^{-1}XB^T$$

$$V_y(Bb) = BX^TV^{-1}XB^T + BX^TV^{-1}XV_0X^TV^{-1}XB^T . \qquad (3.6.3)$$

Suppose now that our prior beliefs are vague, in the sense that, we can assume $V_0 = \sigma^2 I_p$ as $\sigma^2 \to \infty$, where I_p is the $p \times p$ identity matrix. Substituting into (3.6.3) yields,

$$V_y(Bb) \to B + V_0 .$$

Hence,

$$\frac{1}{\sigma^2}J_n^{\theta} \to -p + \frac{1}{2}tr\,(B^{-1}I) + \frac{1}{2}tr\,((V_0^{-1} + B^{-1})(B + V_0))$$

therefore,

$$\frac{1}{\sigma^2}J_n^{\theta} \to -p + tr\,(X^TV^{-1}X) + \frac{1}{2}tr\,(I_p + I_p + B^{-1}I_p)$$

therefore,

$$\lim_{\sigma^2 \to \infty} J_n^{\theta} = tr\,(X^TV^{-1}X) .$$

Thus maximising the gain in information with respect to the design matrix X is equivalent to an A-optimal design.

(ii) D-optimality

Consider the information gain of an experiment as defined by the measure, I_n^{θ}, where

$$I_n^{\theta} = \int p(y)\int p(\theta|y)\log\left(\frac{p(\theta|y)}{p(\theta)}\right)d\theta dy , \qquad (3.6.4)$$

that is the Kullback-Leibler directed divergence between $p(\theta|y)$ and $p(\theta)$.

For the normal linear regression model a calculation similar to (3.6.2) gives,

$$I_n^{\theta} = \frac{1}{2}\log|\, I + V_0X^TV^{-1}X \,| \qquad (3.6.5)$$

therefore,

$$I_n^{\theta} = \frac{1}{2}\log |V_0| + \log | V_0^{-1} + X^T V^{-1} X | .$$

Under vague prior knowledge, $V_0^{-1} \to 0$, so the criterion of maximising the missing information as given by (3.6.5) yields the D-optimal design,

$$\max_X | X^T V^{-1} X | .$$

A rather more interesting discussion of the general case (5.3.1) is given in Stone (1959). A question of clear importance is how many design points are necessary to be able to achieve the maximum information gain? The result is a generalisation of Chernoff's theorem and is given in Stone (1959), the required number of design points is $\frac{1}{2}k(k+1+2q)$ where k is the dimension of the parameter of interest and q is the number of nuisance parameters. Further results and examples are exhibited in Stone (1959) where a Bayesian interpretation of Wald's design criterion of generalised variance is obtained.

(iii) Design criterion under a non-local utility structure induced by a loss structure

In many applications of the design of experiments involve a direct purpose, for example, reporting the LD50 dose. Clearly in such instances the Bayesian is concerned with the posterior measure but further feels it necessary in adopting the loss structure for the quantity to be reported to induce a non-local utility structure for assessing the possible design matrices.

From the utility structure defined in (2.1.5) and the posterior and prior normality as given by the equations (3.6.1) the risk is analytically computable. To give any indication of the functional form under a non-local utility structure consider a loss function that is quadratic associated with the utility $U_{\alpha}^{\theta|x}$. Therefore,

$$L(\theta, \theta') = (\theta-\theta')^T D^{-1} (\theta-\theta')$$

after algebraic manipulation (Good (1969)) we obtain,

$$U_{\alpha}^{\theta|x} = -\frac{1}{2\alpha}\log |I + \alpha V_0 D^{-1}| - \frac{1}{2\alpha} tr \left(Bb(V_0-\alpha^{-1}D)^{-1} \right) - \frac{1}{2\alpha}(Bb-V_0)^T (V_0+\alpha^{-1}D)^{-1}(Bb-V_0) .$$

Thus the maximisation of the above with respect to the design leads to the optimal choice.

(3.6.2) Reference priors and the design of experiments

Let $\{P_{\theta}\}$ denote a family of measures where the information gain for the experiment is defined by

$$I_n^{\theta} = \int p(y) \int p(\theta|y) \log \left(\frac{p(\theta|y)}{p(\theta)} \right) d\theta dy \, .$$

The asymptotic information gain, $\lim_{n \to \infty} I_n^{\theta}$, can be interpreted as the amount of missing information about the parameter of interest θ. Under suitable regularity conditions (see (3.1.2)) this has the form,

$$\lim_{n \to \infty} I_n^{\theta} = \int p(\theta) \log \left(\frac{|I_X(\theta)|^{\frac{1}{2}}}{p(\theta)} \right) d\theta \qquad (3.6.6)$$

where $I_X(\cdot)$ is Fisher's information.

In certain instances, when our *a priori* beliefs are weak we might find it illuminating to use a reference prior (Bernardo (1979b)), which, by definition, maximises the missing information about the parameter of interest, in this case θ.

Hence from (3.6.6) the solution to maximising the missing information, over the space $\mathcal{P}$, is Jeffrey's invariance prior.

$$\pi(\theta) \propto |I_X(\theta)|^{\frac{1}{2}} \, . \qquad (3.6.7)$$

Thus the reference prior can depend on the design X.

Suppose our prior beliefs are such that (3.6.6) exists, then the measure (3.6.6) can be interpreted as the Kullback-Leibler distance between our beliefs $p(\theta)$ and the reference or Jeffrey's prior. The optimal design is to select the design such that the reference prior is closest to our beliefs $p(\theta)$, that is the design for which most is to be learnt from our *a priori* beliefs. Note that criterion (3.6.6) is equivalent to an optimal design criterion of the form,

$$\max_X \left(\lim_{n \to \infty} I_n^{\theta} \right) = \max_X \left(\int p(\theta) \log \left(|I_X(\theta)|^{\frac{1}{2}} \right) d\theta \right) . \qquad (3.6.8)$$

If we employ the reference prior as an approximation to weak *a priori* beliefs then substituting into (3.6.6) gives the criterion,

$$\lim_{n \to \infty} I_n^{\theta} = \int \pi(\theta) \log \left(\int |I_X(\theta)|^{\frac{1}{2}} d\theta \right) d\theta \, ,$$

that is the prior expectation of the logarithm of the averaged Fisher's information over the parameter space.

An interesting example that exhibits the necessity of investigating the dominance of an apparently harmless likelihood specification is that of exponential regression. Here the possibility of improper posterior densities arises. The use of reference priors overcomes such difficulties and

leads to the design criterion (3.6.8).

Here we outline some examples for the application of modelling criterion (3.1.2). Note that the optimal design and the error structure can both be viewed under the unified selection criterion. Designs accounting for the possible incorrect specification of the model have previously been discussed under a quadratic loss structure, quantifying the estimation procedure (see Federov (1972), Atkinson and Federov (1975)). By virtue of the decision problem the criterion (3.1.2) is easier to handle and we shall see that the reference prior froms a central role in the determination of the optimal design via (3.6.8).

The class of nonlinear models represents an interesting application for the criterion (3.6.8) primarily because the Fisher's information matrix depends heavily on the choice of design matrix affecting (3.6.7). Box and Lucas (1959) contains a selection of analytic calculations useful for a wide range of nonlinear models. An interesting result concerning the relationship between the reference prior and the identifiability of modelling parameters in such a setting is contained in Hills (1987b).

Two particular examples of interest are logistic and polynomial regression. First, under an information theoretic criterion Smith and Verdinelli (1980) construct the optimal design for polynomial regression, applicable in (3.6.8). Moreover, the error structure can be assessed under the same criterion, in a similar manner to Rissanen (1987). The optimal design with discrimination between such models via the Bayes factor can also be viewed in such a setting Smith and Spiegelhalter (1982). Secondly, for an application to logistic regression models see Larntz and Chaloner (1986). Again note that the logistic error structure and the design criterion are decision theoretic solutions.

The handling of nuisance parameters requires care and attention. However decompositions of the relevant parts of the modelling criterion do exist for a range of such models, for example, partially nonlinear models (see Hill (1980)). Further simplification arises under (approximate) orthogonality of the parameters.

(3.7) Discussion

This Chapter develops a unified modelling criterion via the Bayesian decision problem of reporting beliefs. The logarithmic utility function is adopted, leading to an information gain which quantifies on a cost scale the relevant modelling components of prior-likelihood combinations, design matrices, model dimensionality and sample size. Three cases arise for analysing the behaviour of the asymptotic information gain; a continuous, a finite discrete or a countably infinite discrete parameter of interest.

The main concern of this Chapter is that of a model elaboration (Smith (1983)), carried out via a sequence of one parameter elaborations that are built up from initial oversimplified families. The techniques adopted involve the calculus of variations and two natural possibilities arise; either letting the class $\mathcal{C}$ of possible measures depend on a parameter and then to determine the model elaboration which maximises the missing information in $\mathcal{C}$, or to adopt a flexible model elaboration indexed by a modelling parameter λ and to determine a selection of reference-type priors for λ. Examples of the modelling criterion are given for the class of scale mixtures of normality where optimal decision-theoretic solutions are discussed in detail. Furthermore, the Huber family is exhibited as a formal Bayesian decision-theoretic model elaboration.

Other applications of the decision-theoretic setting include the selection of data transformations and the design of experiments.

Clearly, there are numerous further applications of the modelling criterion (3.1.2), of particular interest are:

(i) Determining contours of equivalent information gain with respect to the modelling components, for example, the selection of the order and error structure of a time series (see Rissanen (1979)).

(ii) The construction and behaviour of Hypothesis tests (Good (1966b), Bernardo (1980)).

(iii) How the modelling criterion varies with the parameter of interest, for example, prediction in a linear modelling framework (San Martini and Spezzaferri (1984), Geisser and Eddy (1979)). Another problem of interest is to establishing equivalences between risk structures; for example, prediction in a gamma-gamma hierarchical model and estimation in a t-family model elaboration of normality.

Other areas for future work are: characterisation properties of Fisher's information, with particular reference to hierarchical and multivariate statistical models (for example, characterisations in the class of scale mixtures of the multivariate normal including the hyperbolic and generalised inverse Gaussian distributions (Barndorff-Nielsen and Halgreen (1977)); the construction of well-posed decision problems in the form of distance constraints leading to skewed or multi-modal model elaborations; the application to *a priori* specification with the aid of an imaginary training sample for models of different dimensionality (Smith and Spiegelhalter (1982), Box and Kanemasu (1973)).

Chapter 4 : Elimination of nuisance parameters; reference priors

The elimination of nuisance parameters plays a central role in any statistical methodology. A fruitful area of application is to the model elaboration framework discussed in Chapter 3. The latter scenario consists of a parameterised family of measures, denoted by $\{Q_\omega \mid \omega \in \Omega\}$ with indexing set Ω that can be decomposed as $\Theta \times \Lambda$ whose elements denote the parameter of interest and nuisance parameter, respectively.

The following sections will contain an overview of the existing methodologies for the elimination of the nuisance parameter. These will be illustrated primarily within the class of statistical models possessing some form of group structure (see, for example, Fraser (1964), Dawid *et al* (1973), Wijsman (1986), Barndorff-Nielsen and Jupp (1988), Bondar and Milnes (1981)). Such a setting allows a unified approach in which to contrast statistical techniques and concepts; for example, marginal and profile likelihoods, reference priors and inconsistencies, analytic and approximate inferences.

First, the Bayesian methodology is clear cut; integrate out the nuisance parameter with respect to the prior $p(\lambda \mid \theta)$ to obtain a marginal posterior inference. For a wide class of problems a reference prior will be adopted as an approximation to a weak *a priori* specification (Bernardo (1979b)). The interpretation of such an assessment under the unified modelling criterion (3.1.1) will be discussed. However caution must be adopted when employing improper prior measures, for incoherence and strong inconsistency or marginalisation paradoxes might arise (Stone (1976), Dawid *et al* (1973)). Furthermore, the problem will be apparent in any form of sensitivity analysis carried out via the "what if" principle for the assessment of an approximation for a high dimensional prior. This notion will be quantified by virtue of the decomposition of the information gain for the vector (θ, λ) and explored in Chapter 6. For illuminating examples concerning improper priors and the associated properties as judged via the "what if" principle see, for example, Stone and Dawid (1972), Dawid *et al* (1973), Bernardo (1979b).

Secondly, at first sight, the elegant classical procedures based on sufficiency for handling nuisance parameters seem far removed from the Bayesian setting—the methodologies certainly are. The former requires an automatic procedure to determine a function solely of the parameter of interest from which the required inference can be drawn. Clearly, any attempt to define such an *automatic* process is fraught with danger, see, for example, Neyman and Scott (1948). However,

by first considering the class of models that possess some form of group structure, either that of a pure or composite transformation model, a mathematically natural function appears based on the distribution of the maximal invariant statistic (Anderson (1982), Barndorff-Nielsen (1983)). The above procedure determines the modified profile likelihood, denoted by $L_{MP}(\theta)$, a parameterisation invariant function of the parameter of interest. Moreover, under such a group transformation structure analytical and approximate representations of the modified profile likelihood exist aiding in analytical computations for marginal likelihoods with reference priors, see, for example, Lindsay (1980), Kalbfleisch and Sprott (1970), Barndorff-Nielsen (1983).

The following sections will review such procedures indicating the application to reference priors and model elaboration.

(4.1) Profile likelihoods

A number of approaches to the elimination of nuisance parameters in a classical context via the likelihood function have been discussed by Kalbfleisch and Sprott (1970). Such methods have been extended in various ways (Cox (1975), Cox and Reid (1987), Barndorff-Nielsen (1983, 1988)). A central function is the profile likelihood, $L_P(\theta)$, defined by

$$L_P(\theta) \;=\; \sup_{\lambda|\theta} f(x|\theta,\lambda) \;=\; f(x|\theta,\hat{\lambda}_\theta)$$

whose primary application is to large sample size situations.

For the moment a heuristic discussion is adopted. The substitution of $\lambda = \hat{\lambda}$ is, in a sense, pessimistic for hopefully there exists information concerning the nuisance parameter through the parameter of interest. It also has the undesirable feature of possibly leading to inconsistencies even for simple models (Neyman and Scott (1948)). Conversely, the substitution of $\lambda = \hat{\lambda}_\theta$ is rather optimistic by virtue of the fact that, in general, there is uncertainty involved in the parameter θ. The Bayesian approach, however, is an averaging process with respect to the prior beliefs concerning θ, thus taking account of the curvature of the likelihood. The modified profile likelihood also quantifies the latter for it is defined as a weighted profile likelihood by

$$L_{MP}(\theta) \;=\; \left|\frac{\partial\hat{\lambda}}{\partial\hat{\lambda}_\theta}\right| |\hat{i_\theta}|^{-\frac{1}{2}} L_P(\theta)\,.$$

The first weighting factor is $1+O(n^{-1})$ under an orthogonal parametrisation, whereas the second factor is the observed information for fixed θ related to a variance stabilising transformation. Alternatively, under transformation,

$$L_{MP}(\theta) \;=\; \left|\frac{\partial^2 \log f(\theta,\hat{\lambda}_\theta)}{\partial\lambda\partial\hat{\lambda}}\right|^{-1} |I_{\lambda\lambda}(\theta,\hat{\lambda}_\theta)|^{\frac{1}{2}} L_P(\theta)\,, \qquad (4.1.1)$$

of primary use in small samples under weak information content where the profile likehood is inappropriate.

A formal approach will be adopted (see Barndorff-Nielsen (1983)) for the justification of the above construction in section (4.2.1).

It is interesting to note that from a geometrical stance the classical approach mimics the Bayesian one in the sense that for a defined parameter of interest, here θ, the profile likelihood induces a geometrical structure. Barndorff-Nielsen and Jupp (1988) explore such structures in great depth. A key concept involved in the explanation of the likelihood structure is that of an L-sufficient statistic for the parameter of interest (Rémon (1984)).

The following sections contain an overview of some of the existing methodologies and related theorems and examples of transformation models. One such survey is contained in Kass (1979).

(4.1.1) Group transformation models

Consider a statistical model defined by the triple $\{ X, f(x|\Omega), \Omega \}$. Suppose that a group, G, acts on the sample space, X. Furthermore assume that the action has the property,

$$f(gx|\omega) = f(x|\, g^{-1}\omega) \qquad (4.1.2)$$

for all $g \in G$, $\omega \in \Omega$. Thus the action of G on X induces an action on the parameter space Ω via the identity (4.1.2). Such a statistical model is termed a group transformation model.

Maximal invariants

A maximal invariant is a function constant on orbits taking different values on each orbit where the G-orbit is defined by

$$Gx = \{gx|\, g \in G\}$$

Let X/G denote the space of orbits. Dawid *et al* (1973) explains how the nuisance parameter can be identified in such a decomposition.

Amenability

A natural topological smoothness condition for a group is that of amenability see, for example, Bondar and Milnes (1981). Basically, this allows the right invariant Haar measure to be approximated, in some sense, by a sequence of proper priors. Stone (1979) gives a review and interpretation of this concept in relation to improper priors. Emerson and Greenleaf (1967), Bondar

and Milnes (1981) contain equivalent conditions and examples of amenable groups.

Although such mathematical assumptions provide a framework for exploring the interplay between the Bayesian and classical methodologies, the regularity conditions are severe in that straightforward model assumptions can lead to irregularities, for example, in multivariate analysis the group $GL^n_{\mathbf{R}}$ is not amenable (Wijsman (1986)).

A further possibility is that of a composite transformation model. Here the group structure is is only involved with the nuisance parameter and not the full vector. The invariance property is

$$f(gx|\omega) = \chi(g, x)f(x|\theta, g^{-1}\lambda)$$

for all $g \in G$. where $\chi(\cdot, \cdot)$ is known as the multiplier of the group G. Barndorff-Nielsen and Jupp (1988) consider approximation formulae for integral decompositions in such models. The technique is to apply Laplace's approximation to a ratio of likelihoods of the form $f(x|\theta, \lambda)/f(x|\theta_0, \lambda_0)$ for some θ_0, λ_0, thus aiding in the regularity conditions (see Berk (1966)).

Marginal likelihood

For group models the following integral representation theorem holds. Heuristically, by virtue of the fact that the nuisance parameter can be identified with the group structure, an analytical integration can be performed for the right invariant Haar measure on the nuisance parameter.

Theorem (Barndorff-Nielsen (1983)) : Consider a statistical model $\{ X, f(x|\theta, g), \Theta \times G \}$ with densities relative to invariant measure μ on X. Let $G = HK$ be a left coset factorisation of G such that $f(x|\theta, k) = f(x|\theta, e)$ for all x and $k \in K$. Then the distribution of the maximal invariant is given by

$$f(u|\theta) = \int f(x|\theta, h)\Delta(h)^{-1}d\nu(h) \qquad (4.1.3)$$

where ν is invariant measure on H, and Δ is the modular function of the group.

Moreover, it is possible to examine equivalences between *a posteriori* statements and classical procedures under the restrictive condition of a free group action. It is possible to obtain equivalence results between the Neyman-Pearson coverage probability, denoted by $\beta(\theta)$, and the Bayes credible probability associated with the posterior calculated under the right invariant prior, denoted by α_x. The following results exploring such equivalences are contained in Bondar (1977).

Theorem (Bondar (1977)) : Suppose that a statistical transformation model satisfies mild regularity conditions. Furthermore assume that the action of the group is free. Then if the credible region is exact (that is, $\alpha_x = \alpha$) then

$$\inf_{\theta \in \Theta} \beta(\theta) \leqslant \alpha \leqslant \sup_{\theta \in \Theta} \beta(\theta) \,. \tag{4.1.4}$$

The equality in (4.1.4) can be established for an equivariant set C, that is $gC_x = C_{gx}$ for all x and if such a set exists one possible candidate is the highest posterior density (HPD) region (Bondar (1977)).

However, a general theory for such equivalences does not exist and even apparently innocuous models, for example, the Behrens-Fisher problem, disobey certain assumptions (Bondar (1977)).

(4.1.2) Justification of the modified profile likelihood

To mimic the construction of the modified profile likelihood for the class of group transformation models we require a statistic $u(\cdot)$ such that its distribution depends only on θ. Furthermore, assume that the score function $\frac{\partial}{\partial \theta} \log p(x \,|\, \theta, \hat{\lambda}_\theta)$ depends on the observation x only through the statistic $u(\cdot)$. Such a statistic is said to be L-sufficient for θ (see, for example, Barndorff-Nielsen (1978), Rémon (1984), Barndorff-Nielsen and Jupp (1988)).

$$p(u, \hat{\lambda} \,|\, \theta, \lambda) \;=\; p(\hat{\lambda} \,|\, \theta, \lambda, u) p(u \,|\, \theta, \lambda)$$

By assumption, $p(u \,|\, \theta, \lambda) \;=\; p(u \,|\, \theta)$, therefore,

$$p(u \,|\, \theta) \;=\; \frac{p(u, \hat{\lambda}_\theta \,|\, \theta, \lambda)}{p(\hat{\lambda} \,|\, \theta, \lambda, u)}$$

$$p(u \,|\, \theta) \;=\; \frac{p(\hat{\theta}, \hat{\lambda} \,|\, \theta, \lambda) \left| \dfrac{\partial(\hat{\theta}, \hat{\lambda})}{\partial(u, \hat{\lambda})} \right|}{p(\hat{\lambda}_\theta \,|\, \theta, \lambda, u) \left| \dfrac{\partial \hat{\lambda}}{\partial \hat{\lambda}_\theta} \right|} \,. \tag{4.1.5}$$

In order to further explore the above structure, the exact or approximate distribution of the relevant densities is required. A general formula (Barndorff-Nielsen (1983)) for the distribution of the maximum likelihood estimator is given by

$$p(\hat{\lambda} \,|\, \lambda, a) \;=\; c \,|\hat{i}\,|^{\frac{1}{2}} \bar{L} \,,$$

where $\bar{L} = \dfrac{L(\lambda)}{L(\hat{\lambda})}$ the normalised likelihood function, $\hat{i}$ is observed Fisher's information, a is an ancillary statistic and c is a suitable normalising constant. This can be used in (4.1.5) to obtain an exact or approximate inference. The resulting approximation giving the modified profile likelihood as defined in (4.1.1).

Note that the maximal invariant statistic in a composite transformation model is an L-sufficient statistic for θ (Barndorff-Nielsen and Jupp (1988)), thus the concept is a natural extension of that encountered in the transformation group structure.

(4.1.3) Bayesian paradigm

Smith (1983) explores the use of the Bayesian paradigm in the context of a (nuisance) modelling parameter, λ. In such a setting inference for the parameter of interest θ is drawn from the posterior,

$$\begin{aligned} p(\theta|x) &\propto \int p(\theta, \lambda|x)d\lambda \\ &\propto \int f(x|\theta, \lambda)p(\theta, \lambda)d\lambda \\ &\propto \left(\int f(x|\theta, \lambda)p(\lambda|\theta)d\lambda\right)p(\theta) , \end{aligned} \tag{4.1.5}$$

thus analytical, or approximate, representations of the inner integral of (4.1.5) are necessary. The former are possible if the measure $p(\lambda|\theta)$ is chosen to be the right invariant Haar measure, by virtue of theorem (4.1.3) (see Polson (1987)). In general, approximate calculations can be performed for large sample sizes by adopting some form of expansion of (4.1.5). The next section briefly discusses Laplace's method.

(4.1.4) Approximate computations

Laplace's approximation can be used to explore the behaviour of the marginal posterior with respect to a measure $p(\lambda|\theta)$ (Sweeting (1987)). In a sense this is just a restatement of the law of stable measurement (Savage (1954)) for the approximation is only valid for large n and the profile likelihood is in close agreement with that of the reference posterior. The calculation is,

$$p(\theta|x) = \int \exp(\log f(x|\theta, \lambda))p(\lambda|\theta)d\lambda$$

which on application of Laplaces method (Sweeting (1987), Tierney and Kadane (1986)) yields,

$$p(\theta|x) \sim (2\pi)^{\frac{1}{2}k} p(\hat{\lambda}_\theta|\theta)|\hat{i_\theta}|^{-\frac{1}{2}} L_P(\theta)p(\theta) . \tag{4.1.6}$$

Under an orthogonal nuisance parameter,

$$\left|\frac{\partial\hat{\lambda}}{\partial\hat{\lambda}_\theta}\right| = 1+O(n^{-1})$$

and (4.1.6) reduces to,

$$p(\theta|x) \sim (2\pi)^{\frac{1}{2}k} p(\hat{\lambda}_\theta|\theta)L_{MP}(\theta) .$$

(4.1.5) Examples

In this section a brief review of existing computations for families of statistical models is given. A selection of examples are given in Barndorff-Nielsen (1983) including, the von Mises-Fisher model, the hyperboloid model and matched pairs in contingency tables. Nonlinear regression models are considered in Kalbfleisch and Sprott (1972). Further examples are contained in Wijsman (1986), Kalbfleisch and Sprott (1970), Dawid *et al* (1973), Fraser (1968). To illustrate the wide application, consider the class of generalised linear models as defined by

$$f(x|\psi,\kappa) \;=\; b(x,\kappa)\exp(\,\alpha(\kappa)\phi(x,\psi)\,)$$

Suppose that κ is the parameter of interest, then (Barndorff-Nielsen (1983)) the modified profile likelihood for κ is given by

$$L_{MP}(\kappa) \;=\; |\alpha(\kappa)|^{-\frac{1}{2}d} b(x,\kappa)\exp(\alpha(\kappa)\phi(x,\hat{\psi})\,)$$

where $\hat{\psi}$ is an L-sufficient statistic for ψ.

The above formula can be applied in the Bayesian context in a number of ways, see for example, computing sensitivity measures (6.1.1), or approximate posterior distributions (4.1.6).

(4.2) Reference priors

Consider a statistical model defined by the triple $\{\, X, f(x|\theta,\lambda), \Theta\times\Lambda \,\}$. Suppose now that the focus of attention is the specification of the beliefs concerning θ, denoted by $p(\theta)$. The Bayes risk governing the information gain given by the modelling criterion (3.1.2) decomposes as,

$$\lim_{n\to\infty}\left(I_n^{\theta} - \tfrac{1}{2}k\log\left(\frac{n}{2\pi e}\right)\right) \;=\; \int p(\theta)\log\left(\frac{|I(\theta)|^{\frac{1}{2}}}{p(\theta)}\right)d\theta\,. \qquad (4.2.1)$$

The necessary regularity conditions being related to the curvature of the likelihood surface and not directly to asymptotic posterior normality. Suppose that our beliefs belong to a space $\mathcal{C}_p$, itself possibly indexed by a further hyperparameter. Thus the decision-theoretic solution to the representation of $p(\theta)$ of maximising (4.2.1) over the space $\mathcal{P}$ is given by Jeffrey's prior, $\pi(\theta)$, where

$$\pi(\theta) \;\propto\; |I(\theta)|^{\frac{1}{2}}\,.$$

In general the solution is the projection onto Jeffrey's prior under the constraint $\mathcal{C}_p$.

(4.2.1) Examples : (i) Moment class

Suppose that the class $\mathcal{C}_p$ is given by the moment constraint

$$\mathcal{C}_p = \left\{ p(\theta) \mid \int \phi(\theta) p(\theta) d\theta = 0 \right\}.$$

Then an application of the calculus of variations yields the solution, for some Lagrange multiplier α_1 is

$$\pi_\phi(\theta) \propto |I(\theta)|^{\frac{1}{2}} \exp(-\alpha_1 \phi(\theta)) .$$

Note that in general the solution is related to the I-projection of the prior onto Jeffrey's prior, existence and examples of which are contained in Csiszár (1975).

(4.2.2) Strong inconsistency and amenability

Stone (1979) contains a review of the results applicable to improper priors. Two such results are given in Bondar and Milnes (1981) and Dawid *et al* (1973) where strong inconsistency disappears under an amenable group structure and marginalisation paradoxes are avoided for the right invariant Haar measure. Other examples can be interpreted in such a setting, for example, Stone (1976) considers the free group on two generators, F_2, a nonamenable group leading to a possible strong inconsistency.

(4.2.3) Examples :

(i) Exponential Connection

Consider two possible location-scale families with densities denoted by $f((x-\mu)/\sigma)$ and $g((x-\mu)/\sigma)$. A possible embedding of the two families is the exponential connection (see (3.4.1)). The log-likelihood is given by

$$\log p(x|\mu, \sigma, \lambda) = \log c_\lambda + \lambda f\left(\frac{x-\mu}{\sigma}\right) + (1-\lambda)\, g\left(\frac{x-\mu}{\sigma}\right).$$

Reference priors and sensitivity measures are determined by Fisher's information matrix, which for this three parameter likelihood takes the form

$$I(\mu, \sigma, \lambda) = \begin{pmatrix} b_\lambda & \frac{a_\lambda}{\sigma} & a_\lambda \\ \frac{a_\lambda}{\sigma} & d_\lambda + \frac{1}{\sigma^2} & 0 \\ a_\lambda & 0 & \frac{d^2 \log c_\lambda}{d\lambda^2} \end{pmatrix}$$

where the constants a_λ, b_λ are computable in terms of the functions $f(\cdot)$ and $g(\cdot)$.

(ii) Hierarchical models

First, consider a binomial sampling framework where the reference prior is given by

$$\pi(\theta) \;=\; \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \,.$$

Suppose that an elaboration of a conjugate prior, Beta(α, β) say, is required. One possibility is to generate the class $\mathscr{C}_p$ via the exponential connection between the reference prior and the conjugate prior as given by (4.2.1), leading to the class

$$\mathscr{C}_p \;=\; \left\{ \int_{\Lambda} p(\theta|\lambda)p(\lambda)d\lambda \;|\; p(\theta|\lambda) = \text{Beta}(\lambda\alpha+\kappa,\, \lambda\beta+\kappa) \right\}$$

where $\kappa \;=\; \frac{1}{2}(\lambda-1)$. Thus a one parameter mixture of conjugate priors is obtained.

Similar techniques can be applied to other exponential families and conjugate priors, for example to a scale parameter with an exponential prior to generate the inverse chi-squared family.

The reference prior for the parameter (α, β) can be determined as follows:

$$f(x|\alpha,\beta) \;=\; \int f(x|\theta)p(\theta|\alpha,\beta)d\theta \,,$$

thus on marginalising with respect to the conjugate beta prior, the log-likelihood becomes,

$$\log f(x|\alpha,\beta) \;=\; \log\left(\,\Gamma(\alpha+x)\Gamma(1+\beta-x)\,\right) - \log\left(\,\Gamma(\alpha)\Gamma(\beta+1) + \Gamma(\alpha+1)\Gamma(\beta)\,\right).$$

Fisher's information matrix for (α, β) is computable in terms of the digamma function, denoted by $\psi(\cdot)$, as follows,

$$I(\alpha,\beta) \;=\; \begin{pmatrix} \psi'(\alpha+1)-\psi'(\alpha)+\dfrac{1}{(\alpha+\beta)^2} & \dfrac{1}{(\alpha+\beta)^2} \\ \dfrac{1}{(\alpha+\beta)^2} & \psi'(\beta+1)-\psi'(\beta)+\dfrac{1}{(\alpha+\beta)^2} \end{pmatrix}$$

however, this simplifies due to the identity,

$$\psi'(\alpha+1) \;-\; \psi'(\alpha) \;=\; -\frac{1}{\alpha^2}$$

therefore the reference prior for the parameter (α, β) is given by

$$\pi(\alpha,\beta) \;=\; |I(\alpha,\beta)|^{\frac{1}{2}} \,.$$

Algebraic manipulation gives, for $\alpha > 0, \beta > 0$,

$$\pi(\alpha,\beta) \;\propto\; \frac{1}{\sqrt{\alpha\beta}(\alpha+\beta)} \,.$$

Note that for fixed β a proper prior is obtained, thus avoiding over-shrinkage at the origin, a clearly desirable feature. This is typical of second stage reference priors, a fruitful area of application in that the ensuing estimators tend to be close to previously proposed shrinkage estimators, see, for example, Takada (1979), Akaike (1980), Bondar (1987).

(4.3) Discussion

This Chapter contains a survey of existing literature as a basis for future work and discussion on the equivalence between classical and Bayesian approaches to the elimination of nuisance parameters. Analytical and approximate computations to aid in the implementation of the Bayesian paradigm are explored, based mainly on the properties of group transformation models. The modified profile likelihood and its relation to marginal Bayesian inferences and the approximation of posterior measures are explored. Further areas for future work include the computation of sensitivity measures (see (6.1.1)) and the approximation of Bayes factors (Lindley (1961)).

A class of models that has received little attention is that of non-group models which can be embedded in a group model by addition of further modelling parameters (Dawid (1975)).

The limitation and equivalences of the methods adopted throughout this Chapter needs further investigation, for example, to large dimensional parameter spaces and to non-group transformation models.

A fruitful area for calculations is that of references priors for hyperparameters in hierarchical models, where it appears that the solutions are usually integrable (at least on compacts) and, via marginalisation, explain previous functional forms of shrinkage-type estimators. Their behaviour might be explained by considering inequalities associated with information measures in hierarchical models (Goel and DeGroot (1981), Haitovsky and Zidek (1986)).

Chapter 5 : Posterior expectations characterising prior distributions in the exponential family

Consider a random variable X whose density belongs to the exponential family through μ i.e. $dP_\theta(X) = \exp(X\theta - M(\theta))d\mu(X)$ where $\theta \in \Theta$. In the context of Bayesian statistics we are interested in estimating functionals of θ, denoted by $\psi(\theta)$, via posterior expectations $E(\psi(\theta)|X)$.

We consider approaches to the problem of assessing the properties of our prior distribution $p(\theta)$ induced by constraints on the form of $E(\psi(\theta)|X)$. The central characterisation result is that if $E(M'(\theta)|X)$ is linear in X, then the prior distribution is uniquely identified as a conjugate prior (Diaconis and Ylvisaker (1979, 1985)). Furthermore corresponding results for the location family, $f(X-\theta)$, are reviewed (Goldstein (1975), Diaconis and Ylvisaker (1985)).

First we review useful identities that arise in such models (Hudson (1978), Masreliez (1975)). Secondly, motivated by a problem of Diaconis and Ylvisaker (1985) concerning characterisations when the functionals are polynomials, we look at examples arising in the special case of a $N(\theta,1)$ likelihood, including the class of multimodal priors of the form "polynomial times a normal".

In the location problem the implications for prior moments take the form of a set of recurrence relations (Goldstein (1975)). Then, in certain instances, Carlemann's theorem (Kagan, Linnik and Rao (1973)) characterises $p(\theta)$. For the exponential family, properties of the moment generating function reduce the problem either to the determination of $E(\theta|X)$ or to a differential equation for $p(\theta)$ (Sampson (1975), Diaconis and Ylvisaker (1979), Ralescu and Ralescu (1981)).

Finally, we consider the possible extension of examples outside the normal family and to possibilities for future work. For convenience, $E_{\theta|X}(\cdot)$ will denote the expectation operator when it is required to suppress the conditioning.

(5.1) Identities in the exponential family

A number of natural identities exist for a continuous random variable in the exponential family with support $\mathbb{R}$ (Hudson (1978)). In a Bayesian setting these can be used for a number of purposes, primarily for simplification of Bayes risks with regard to shrinkage estimation. In this section we review these identities for posterior expectations allowing partial insight into the forms

of functionals ψ that arise naturally in this context.

Consider the exponential family through μ in its natural parameterisation, where θ has support $\mathbb{R}$. Thus the posterior will be of the form,

$$p(\theta|X) \propto p(\theta)e^{X\theta - M(\theta)} . \tag{5.1.1}$$

Define

$$e^{B(X)} = \int_{-\infty}^{\infty} p(\theta)e^{X\theta - M(\theta)}d\theta .$$

Rewriting (5.1.1) gives

$$p(\theta|X) = p(\theta)e^{-M(\theta)}e^{X\theta - B(X)} ,$$

a density for θ lying in the exponential family. The following identities hold (Hudson (1978)).

(i) Let $g(\theta)$ be any absolutely continuous function on $\mathbb{R}$, such that $E_{\theta|X}|g'(\theta)| < \infty$, then

$$E_{\theta|X}\left(\frac{d}{dX}\log p(\theta|X)g(\theta)\right) = -E_{\theta|X}(g'(\theta)) .$$

(ii) Suppose θ is a discrete parameter taking values in $\{0,1,2,...\}$. Then for $g(\theta)$ satisfying $E_{\theta|X}|g(\theta)| < \infty$

$$e^{X}E_{\theta|X}(g(\theta)) = E_{\theta|X}(t(\theta)g(\theta-1)) ,$$

where

$$t(\theta) = \frac{p(\theta-1)}{p(\theta)}e^{M(\theta)-M(\theta-1)} .$$

(iii) Let $\mu = E(\theta|X)$, then the normal, gamma and inverse chi-squared cases of the exponential family have the property that there exists a function $a(\theta)$ such that for all absolutely continuous functions $g(\theta)$ the following holds

$$E_{\theta|X}((\theta-\mu)g(\theta)) = E_{\theta|X}(a(\theta)g'(\theta)) .$$

Due to the fact that the posterior is in the exponential family, the posterior mean is given by

$$E(\theta|X) = B'(X) .$$

By definition of the score (i.e. minus the derivative of the logarithm) function of the posterior with respect to X,

$$E(\theta|X) = \theta - \frac{d}{dX}\log p(\theta|X) .$$

Hence the posterior variance can be calculated as

$$V(\theta|X) = \int p(\theta|X)\left(\frac{d}{dX}\log p(\theta|X)\right)^2 d\theta$$

therefore,

$$V(\theta|X) = I_{\theta|X}(\cdot) ,$$

where $I(\cdot)$ is Fisher's information, thus the associated expected Bayes risk, $R(\cdot)$ can be expressed as

$$R(\cdot) = E_X(V(\theta|X)) = E_X((B'(X)-\theta)^2) ,$$

by properties of the exponential family we have,

$$R(\cdot) = B''(X) ,$$

therefore,

$$R(\cdot) = \int p(X)\frac{d^2}{dX^2}\log p(X)dX + \int p(X)S''(X)dX$$

where $S''(X) = -\frac{d^2}{dX^2}\log f(X|\theta)$. Hence,

$$R(\cdot) = \int p(X)S''(X)dX - I_X(\cdot) \qquad (5.1.2)$$

a functional of $p(X)$ which is determined by Fisher's information and a moment constraint.

(5.2) Functional minimisation of the Bayes risk

Consider the decision problem defined by the family of measures P_θ and risk function given by (5.1.2). Suppose our prior beliefs for $\theta \in \Theta$ are only specified by the constraint that $p(\theta) \in \mathcal{C}$, for some subset $\mathcal{C}$ of the space of possible probability measures over Θ. This then induces a subset $\mathcal{C}_X$ of the space of possible probability measures over X such that $p(X) \in \mathcal{C}_X$. A full Bayesian analysis would require a further measure over $\mathcal{C}$, to average over the set $\mathcal{C}$. If our primary object is to report the posterior mean the associated risk from (5.1.2) can be used to induce a distance on $\mathcal{C}$. The ensuing distance allowing a choice of measure over $\mathcal{C}$, given by concentrating on the particular prior, $p(\theta)$, which attains

$$\min_{\mathcal{C}_x} R(\cdot) .$$

Consider the problem of minimising R (·) over the space of all distributions, $\mathcal{P}$. This can be solved by finding the extremal solutions, using the Euler-Lagrange equations, of the following calculus of variations problem,

$$\int F(p, p') \text{ subject to } \int p = 1$$

where

$$F(p, p') = pS'' - \frac{(p')^2}{p} .$$

The Euler-Lagrange equation being,

$$\frac{d}{dX}\left(\frac{\partial F}{\partial p'}\right) - \frac{\partial F}{\partial p} = \alpha$$

where α is a Lagrange multiplier. Let $u(X) = \frac{d}{dX}\log p(X)$, then we have a first order differential equation for $u(\cdot)$ given by

$$\frac{du}{dX} + \frac{1}{2}u^2 = \frac{1}{2}(\alpha + S'') . \tag{5.2.1}$$

There are close links with the Riccati and Schrödinger equations (see also (3.1.6)) which have no general solution, but for given forms of $S''(X)$ solutions for $u(\cdot)$ can often be identified.

For example, suppose the family $\{P_\theta\}$ is the exponential family with mean $1/\theta$. By direct substitution, if $p(X)$ is a gamma density equation (5.2.1) holds. Hence the set $\mathcal{C}_X$ is the family of gamma densities, which implies that the original prior space is also the set of gamma densities (as the likelihood is exponential). Thus, in this setting, an optimal choice of prior measure over $\mathcal{C}$ is one that concentrates its mass on the gamma family, which is indexed by two hyperparameters. The same characterisation holds when the likelihood is also gamma.

(5.2.1) Links with inference for a location parameter

Suppose now that our observation X is generated from a location family with scale unity, giving rise to a density $f(X-\theta)$. This is apparent for the normal family and certain members of the exponential family after a suitable transformation (LeJeune and Faulkenberry (1982)).

By exploring analytic forms for posterior means and variances we obtain guidance to forms of departures encountered in model elaboration (Box (1980), Smith (1983)). Under a normal prior, the relevant identities for assessment of posterior means and variances are given by the following theorem.

Theorem (Masreliez (1975)) : Let $g(X) = -\frac{d}{dX}\log p(X)$ and $G(X) = g'(X)$. Then under a normal prior $p(\theta) = N(m, c^2)$, and a bounded likelihood (see Appendix (5.10)),

$$E(\theta|X) = m + c^2 g(X) \tag{5.2.2}$$

$$V(\theta|X) = c^2 - c^4 G(X) \; . \tag{5.2.3}$$

For a discussion of their interpretation in a robustness setting see Smith (1983), also O'Hagan (1979). Similar relations hold for scale parameters and gamma priors (West (1984)).

For the moment we pursue this direction further by noting that $g(X)$ can be expressed as the posterior expectation of the likelihood score. Thus providing a direct link with corresponding notions of M-estimation and appropriate choice of score function (Ramsey and Novick (1980), Huber (1981), Marazzi (1980)).

Lemma : Under the conditions of Masreliez's theorem,

$$E(\theta|X) = m + c^2 E_{\theta|X}\left(\frac{d}{d\theta}\log f(X-\theta)\right) \; .$$

Proof : By definition,

$$E_{\theta|X}\left(\frac{d}{d\theta}\log f(X-\theta)\right) = \frac{1}{p(X)}\int_0^\infty p(\theta)\frac{d}{d\theta}f(X-\theta)d\theta \; .$$

Interchanging the derivative with respect to θ with that of X, together with the integral sign yields

$$E_{\theta|X}\left(\frac{d}{d\theta}\log f(X-\theta)\right) = -\frac{d}{dX}\log p(X) \; ,$$

by Masreliez theorem we have the desired result.

This generalises to a sample of size n, showing that under a normal prior the score function for X_i quantitatively describes the behaviour of the posterior mean given by the following lemma.

Lemma : Suppose our prior is normal and we have a random sample of size n, then

$$E(\theta|X) = m + c^2 \sum_{i=1}^{n} E_{\theta|X}\left(\frac{d}{d\theta}\log f(X_i|\theta)\right) \; .$$

Note that if we choose a bounded score function then the posterior mean will also be bounded, for all X.

The Bayesian methodology naturally allows, and encourages, flexible inputs for prior and likelihood combinations in order to study robustness in the form of a model elaboration (Box (1980), Smith (1983)). One such class that arises naturally is the set of all scale mixtures of normality, the addition of hyperparameters giving flexibility in the underlying structure imposed on the posterior for the parameter of interest θ. This class incorporates a very wide selection of possible shapes including the student, logistic and exponential power families. We now look at the consequences for posterior moments.

(5.2.2) Prior elaboration

First consider a prior elaboration, for known σ, having the representation,

$$p(\theta|\sigma) = \int_0^\infty p(\theta|c, \sigma)p(c|\sigma)dc$$

where, $p(\theta|c, \sigma) = N(m, c^2\sigma^2)$, $p(c|\sigma) = p(c)$. For simplicity, take $m = 0$, $\sigma^2 = 1$. Then, by Fubini,

$$E(\theta|X) = \frac{1}{p(X)} \int_0^\infty \left(\int_{-\infty}^\infty \theta f(X-\theta)p(\theta|c)d\theta \right) p(c)dc$$

The inner integral can be written using (5.2.2),

$$\int_{-\infty}^\infty \theta f(X-\theta)p(\theta|c)d\theta = -c^2 \frac{d}{dX} p(X|c)$$

hence,

$$E(\theta|X) = -\frac{1}{p(X)} \int_0^\infty \frac{d}{dX} p(X|c)c^2 p(c)dc \ . \qquad (5.2.4)$$

Define the new density $p^*(c) = c^2p(c)/A$ for suitable normalising constant A and correspondingly define $p^*(X)$. Then we can write

$$E(\theta|X) = -\frac{A}{p(X)} \frac{d}{dX} p^*(X)$$

$$= -A \frac{\frac{d}{dX} p^*(X)}{\frac{d}{dX} p(X)} \frac{d}{dX} \log p(X)$$

therefore,

$$E(\theta|X) = -D(X)\frac{d}{dX}\log p(X) \qquad (5.2.5)$$

which is of the same form as (5.2.2) in the original theorem except for the introduction of a down-weighting factor $D(X)$. In the special case where $p(c)$ is gamma, then so is $p^*(c)$ and hence $p(\theta)$ has a t-distribution.

Alternatively, (5.2.5) can be rearranged in terms of the second moment of the hyperparameter c, as follows,

$$E(\theta|X) = -\frac{1}{p(X)}\frac{d}{dX}(p(X)E(c^2|X))$$

$$= \frac{d}{dX}E(c^2|X) - E(c^2|X)\frac{d}{dX}\log p(X) .$$

The equivalent to (5.2.1) can be determined by evaluating $E(\theta^2|X)$. Unfortunately, this term does not combine neatly with $(E(\theta|X))^2$. The following expression for the variance is obtained

$$V(\theta|X) = A\frac{p^*(X)}{p(X)} - D^*(X)\frac{d}{dX}\log p(X) - D(X)^2\frac{d}{dX}\log p(X) ,$$

where A^* is the normalising constant for prior $c^4p(c)$, and

$$D^*(X) = A^*\frac{\frac{d^2}{dX^2}p^{**}(X)}{\frac{d}{dX}p(X)} .$$

Calculation of the Bayes risk, R (·)

A useful identity for the Bayes risk for the posterior mean under a quadratic loss structure was established in Brown (1971). Define, $\delta(X) = E(\theta - X|X)$.

By definition,

$$R(\cdot) = E_X(V(\theta|X))$$

$$R(\cdot) = E_{(X,\theta)}((\theta - X)^2) - E_X(\delta(X)^2) .$$

By hypothesis $\mathrm{Var}(X) = 1$, giving

$$R(\cdot) = 1 - \int p(X)\delta(X)^2 dX .$$

Hence, depending on the choice of prior we have different forms for $\delta(X)$. For example, under a normal prior, by Masreliez theorem we obtain

$$R(\cdot) = 1 - I_X(\cdot) \; . \tag{5.2.6}$$

Hence, as in the exponential family, the Bayes risk is a functional of the Fisher information of the marginal beliefs about X.

Fisher's information can be further expressed in terms of a Kullback-Leibler distance, using DeBruijn's identity (see Barron (1985)), by the following theorems.

(5.2.3) Representation of Fisher's information

Suppose the conditions for Masreliez's theorem hold, thus (5.2.2) and (5.2.3) are valid. In order to relate the decision problem of reporting the posterior mean with risk, $R(\cdot)$, with that of reporting the marginal beliefs about X, we apply DeBruijn's identity to represent the Bayes risk, $R(\cdot)$, in terms of a Kullback-Leibler distance. Such a distance can itself be interpreted as a Bayes risk to a pure inference problem comprising a local utility function which necessarily has to be logarithmic (Bernardo (1979a)).

Theorem : Under the conditions of Masreliez's theorem,

$$\frac{d}{dc^2}\left(\int p(X)\log p(X)dX\right) \;=\; -\tfrac{1}{2}I_X(\cdot) \; . \tag{5.2.7}$$

Proof : This is DeBruijn's identity (see Blachman (1965), Barron (1985)) written in a Bayesian context.

The following theorem links the Bayes risk, $R(\cdot)$, with the decision problem of projecting $p(X)$ onto $\phi(X)$.

Theorem : Let $\phi(X) \sim N(0,\tau)$, then the Bayes risk has representation

$$R(\cdot) \;=\; 2\tau^2 \frac{d}{d\tau}\left(\int p(X)\log\left(\frac{p(X)}{\phi(X)}\right)dX\right) \; .$$

Proof : By (5.2.7), the Bayes risk is given by

$$R(\cdot) \;=\; 2\tau^2\left(\frac{1}{2\tau} + \frac{d}{d\tau}\left(\int p(X)\log p(X)dX\right)\right) \; . \tag{5.2.8}$$

By definition of $\phi(X)$,

$$\int p(X)\log \phi(X)dX \;=\; -\tfrac{1}{2}\log\tau + A$$

for some constant A. Hence,

$$\frac{d}{d\tau}\left(\int p(X)\log\phi(X)dX\right) = -\frac{1}{2\tau}.$$

Substituting into (5.2.8) gives,

$$R(\cdot) = 2\tau^2\frac{d}{d\tau}\left(\int p(X)\log\left(\frac{p(X)}{\phi(X)}\right)dX\right)$$

as required. The risk is thus represented as a derivative of the loss when approximating the marginal beliefs about X by a normal distribution with variance τ.

Minimax solutions can be obtained, as in the exponential family, by minimising (5.2.6) (see Bickel (1981)).

(5.2.4) Likelihood elaboration

Suppose instead that our likelihood is a scale mixture of normality, with mixing measure indexed by λ. Reversing the roles of θ and X leads to a similar expression to (5.2.5) for the posterior mean as follows. By definition, absorbing irrelevant constants,

$$E(\theta|X) = \frac{1}{p(X)}\int_{-\infty}^{\infty}\int_{0}^{\infty}\frac{\theta}{\lambda}\exp\left(-\frac{1}{2}\left(\frac{X-\theta}{\lambda}\right)^2\right)p(\lambda)p(\theta)d\lambda d\theta\,,$$

where it is assumed our prior beliefs about λ and θ are independent. Making the substitution $u = X-\theta$ gives

$$E(\theta|X) = X + \frac{1}{p(X)}\int_{0}^{\infty}\int_{-\infty}^{\infty}\lambda p(\lambda)p(X-u)dh(u)d\lambda\,,$$

where $h(u) = \exp\left(-\frac{1}{2}\left(\frac{u}{\lambda}\right)^2\right)$. In a similar manner to before, integration by parts and reversal of derivative and integral sign yields

$$E(\theta|X) = X + \frac{1}{p(X)}\int_{0}^{\infty}p(\lambda)\lambda^2\frac{d}{dX}p(X|\lambda)d\lambda \qquad (5.2.9)$$

with the previous notation we obtain

$$E(\theta|X) = X + \frac{Ap^*(X)}{p(X)}\frac{d}{dX}\log p^*(X)\,.$$

Thus in the setting of such an elaborated likelihood, our posterior mean has the functional form of X plus a Bayes factor times a score function, the main quantity of interest being the marginal beliefs about X under the elaborated likelihood with mixing measure $\lambda^2 p(\lambda)$.

Rearranging equation (5.2.9) leads to expressions either in terms of moments of the mixing parameter, or the likelihood score function. First, as in (5.2.2),

$$E(\theta|X) = X + \frac{d}{dX}E(\lambda^2|X) + E(\lambda^2|X)\frac{d}{dX}\log p(X) .$$

Secondly, in terms of a posterior expectation with respect to $p^*(\lambda|X)$, we have

$$E(\theta|X) = X - A\, E^*_{\lambda|X}\left(\frac{d}{dX}\log p(X|\lambda)\right)$$

where, $A = E_\lambda(\lambda^2)$. This generalises (5.2.2) as

$$E(\theta|X) = X - A\, E^*_{\lambda|X}(g(X|\lambda))$$

where $g(X|\lambda)$ is the score function calculated under the first stage normality assumption. The robustifying mechanism being that of an adaptive averaging of such score functions by the posterior $p^*(\lambda|X)$. For results concerning outlier proneness for members of the exponential power family see O'Hagan (1979).

A straightforward application of this last result occurs if we assume our prior beliefs are normal, mean zero and scale c^2. For then $p(X|\lambda)$ is directly computable as a $N(0, c^2+\lambda^2)$ density. Correspondingly, the score function $g(X|\lambda) = X/(c^2+\lambda^2)$. An adaptive shrinkage estimator of the form

$$E(\theta|X) = \left(1 - A\, E^*_{\lambda|X}\left(\frac{1}{c^2+\lambda^2}\right)\right) X \qquad (5.2.10)$$

is obtained. In terms of the original posterior,

$$E(\theta|X) = \left(1 - E_{\theta|X}\left(\frac{\lambda^2}{c^2+\lambda^2}\right)\right) X ,$$

invoking approximations to the inner expectation will then lead to expressions resembling known forms of shrinkage type estimators (Efron and Morris (1971)). Furthermore, due to positivity of the inner expectation it shows that for the elaboration of scale mixtures of normality the posterior mean is always shrunk towards the prior mean.

The corresponding formula for $V(\theta|X)$ does not simplify for the likelihood elaboration, but for completeness it is given by

$$V(\theta|X) = A\frac{p^*(X)}{p(X)} - A^{**}E^{**}_{\lambda|X}\left(\frac{1}{p(X|\lambda)}\frac{d^2}{dX^2}p(X|\lambda)\right) - A^2\left(E^*_{\lambda|X}(g(X|\lambda))\right)^2 .$$

Although the variance function does not have a direct interpretation, it still plays a central role in model elaboration. The dispersion properties of mixtures have many applications in statistics and probability (Finucan (1971), Shaked (1980)).

(5.3) Examples with a $N(\theta, 1)$ likelihood

Suppose our likelihood is normal with mean θ and variance unity. Throughout we will be concerned with *a posteriori* constraints, for all X, of the form

$$E(\psi(\theta)|X) = \eta(X) \ . \tag{5.3.1}$$

In this section we list five specific examples, where the functionals ψ and η are polynomials, together with the corresponding prior distributions. Finally we note the asymptotic behaviour of $E(\psi(\theta)|X)$ as $X \to \infty$ (Meeden and Isaacson (1977), Dawid (1973)). Sometimes these conditions allow us to show that no such prior exists.

Examples :

(i) Consider a constraint on the nth posterior central moment (i.e. $\psi(\theta) = \theta^n$) where n is odd.

Lemma : $E(\theta^n|X) = X$ can be solved with prior given by

$$p(\theta) \propto \exp\left(\frac{1}{2}\left(\theta^2 - 2\frac{\theta^{n+1}}{n+1}\right)\right) . \tag{5.3.2}$$

Proof : See Section (5.6)

(ii) The following example resolves a conjecture of Diaconis and Ylvisaker (1985) as to the existence of non-normal priors with ψ and η both polynomials.

Lemma : Suppose $p(\theta) \propto \theta^2 e^{-\frac{1}{2}\theta^2}$, then we obtain

$$E(2\theta^3 - 3\theta|X) = \frac{X^3}{4} + 3X \tag{5.3.3}$$

$$E(\theta^4 - 4\theta^2|X) = \frac{X^4}{16} + 3\frac{X^2}{4} - \frac{9}{4} \ . \tag{5.3.4}$$

Proof : By Bayes theorem the posterior is given by

$$p(\theta|X) \propto \theta^2 e^{-(\theta - \frac{1}{2}X)^2} \ . \tag{5.3.5}$$

Integration by parts leads to the following recurrence relation for the posterior moments.

$$E(\theta^i|X) = \frac{i+1}{2}E(\theta^{i-2}|X) + \frac{X}{2}E(\theta^{i-1}|X) \tag{5.3.6}$$

we also have (e.g. Diaconis and Ylvisaker (1985))

$$E(\theta|X) = \frac{X^3+6X}{2X^2+4} \ . \tag{5.3.7}$$

Algebraic manipulation of (5.3.6) for i=2,3,4, and (5.3.7) implies the equations (5.3.2) and (5.3.3). As such there exist solutions for certain cubics and quartics with a non-normal prior.

(iii) Class of priors of the form, $\phi(\theta)e^{-\frac{1}{2}\theta^2}$, ϕ = polynomial

We outline a method for examining the constraints involved under such a model structure. In theory analytical computations can be carried out reducing condition (5.3.1) to that of a linear system of equations as follows. By Bayes' theorem the posterior is given by,

$$p(\theta|X) \propto \phi(\theta)e^{-(\theta-\frac{1}{2}X)^2} \ .$$

The posterior expectation of $\psi(\theta) = \sum_{j=0}^{m} b_j\theta^j$, if analytically determined, would take the form

$$\left(\sum_{i=0}^{n} a_iC_i(X)\right)E(\psi(\theta)|X) = \sum_{i=0}^{n}\sum_{j=0}^{m} a_ib_jC_{i+j}(X) \tag{5.3.8}$$

where $C_i(X)$ is a polynomial of degree i in X. First note that for (5.3.1) to have a solution we need ψ and η to have the same degrees. Hence let

$$\eta(X) = \sum_{j=0}^{m} d_jX^j \ .$$

Constraint (5.3.1) now reduces to the polynomial identity in X given by

$$\left(\sum_{j=0}^{m} d_jX^j\right)\left(\sum_{i=0}^{n} a_iC_i(X)\right) = \sum_{i=0}^{n}\sum_{j=0}^{m} a_ib_jC_{i+j}(X) \tag{5.3.9}$$

yielding constraints for the possible coefficients in the polynomials.

The main advantages of this class of priors are that the posterior distribution and its moments are analytically tractable, also it contains a varying selection of 'shapes' ranging from normality to multi-modality for prior input.

(iv) ψ and η both quadratics

Suppose that ψ and η are both quadratics i.e. without loss of generality, assume that

$$E(\theta^2+b_1\theta|X) = d_0+d_1X+d_2X^2 \ . \tag{5.3.10}$$

We have the following results:

(i) For a solution to exist d_2 has to be greater than zero.

(ii) If we restrict our prior to the class considered in Section 4, then (5.3.10) has a solution in this class if and only if $p(\theta)$ is normal (i.e. ϕ = constant).

Proof of (i) : First, the following lemma restricts the possibilities for the polynomials ϕ and ψ.

Lemma (5.3) : Suppose $\psi(\theta)$ has complex roots. Then if (5.3.1) has a solution, $\eta(X)$ must have complex roots also.

Proof : As $\psi(\theta)$ has complex roots, $\psi(\theta) \geq 0$, hence l.h.s. (5.3.1) is ≥ 0. Therefore $\eta(X) \geq 0$ i.e. $\eta(X)$ has complex roots.

Corollary (5.3) If ψ is even and η is odd then no solution exists.

In order to prove (i), rewrite (5.3.10) as

$$E(\theta^2+b_1\theta+b_0|X) = d_0+b_0+d_1X+d_2X^2 \qquad (5.3.11)$$

for any b_0. Choose b_0 such that

$$b_1{}^2 < 4b_0 \text{ and } d_0+b_0 > 0 .$$

Then $\psi(\theta)$ has complex roots. Hence, the r.h.s. of (5.3.11) has complex roots and as $d_0+b_0 > 0$ we must have $d_2 > 0$.

Proof of (ii) : Suppose that there exists a solution to (5.3.10) in the class polynomial times normal where degree $\phi \geq 2$ Hence applying the polynomial identity (5.3.9) we obtain

$$d_0 = b_0 - \frac{(2n+1)b_2}{2}$$

therefore (5.3.10) now becomes

$$E(b_2\theta^2+b_1\theta|X) = -\frac{(2n+1)b_2}{2}+\frac{b_1}{2}X+\frac{b_2}{4}X^2 . \qquad (5.3.12)$$

Now choose $c_0 = \frac{b_1{}^2}{4b_2}$ and add c_0 to both sides of (5.3.12). By choice of c_0, the polynomial on the l.h.s. has complex roots. After algebraic manipulation the discriminant of polynomial on the r.h.s. is $2b_2{}^2(2n+1) \geq 0$. Hence it has real roots, contradicting lemma (5.3). Therefore the only solution in this class can be the normal.

It is possible to examine different tail behaviour in relation to posterior expectations, for example, let $p(\theta) = e^{-\frac{1}{2}\theta^4}$ (i.e. light tails) then it will be shown that (see section (5.6))

$$E(2\theta^3+\theta|X) = X\ .$$

(5.3.1) Asymptotic Behaviour of $E(\psi(\theta)|X)$ as $X \to \infty$

Using techniques of Laplace approximation it is possible to examine the behaviour of $\psi(\theta)$ as $X \to \infty$ (Meeden and Isaacson (1977)). This is clearly relevant to establishing conditions for $\eta(X)$. Furthermore, it is appealing from a robustness viewpoint when considering aberrant observations (see also Dawid (1973)). Rather more can be said about the approximate posterior, it is in fact approximately normal. Concentrating on expectations, the following results from Meeden and Isaacson are useful.

(i) Priors θ^{-c}, $e^{-\theta^{3/2}}$, lead to posterior means with behaviour like $X+\sqrt{X^2-A}+B$ as $X \to \infty$. The prior $p(\theta) \propto \exp(-c\theta^{\alpha}+\theta^2)$, $\alpha > 2$, leads to a posterior mean with behaviour like $X^{-1/(\alpha-1)}$

(ii) For a likelihood in the exponential family the behaviour of $E(\psi(\theta)|X)$ for polynomial ψ and large X is related to that of $E(\theta|X)$ via theorem 4 of Meeden and Isaacson, which states that

$$\lim_{X \to \infty} \frac{E(\psi(\theta)|X)}{\psi(E(\theta|X))} = 1\ .$$

(5.4) Characterisations involving moment generating functions

A probabilistically appealing tool for characterisations is the moment generating function. For exponential families it is directly related to the posterior mean (Sampson (1975), Goldstein (1977)) It can further be used as a device to reduce the integral equation (5.3.1) to a differential equation for $p(\theta)$ (Ralescu and Ralescu (1981), Diaconis and Ylvisaker (1979)). First, we review the main result of Sampson.

Theorem (Sampson (1975)) : Suppose $\{T_\theta\}$ is a family of scalar random variables whose density lies in the exponential family through μ. Let $g(\theta) = E(T_\theta)$. Then the density of T_θ is uniquely determined via its m.g.f.

$$m_{T,\theta}(s) = \exp\left(\int_{\theta}^{\theta+s} g(w)dw \right)\ .$$

Note that the converse also holds and there exists a multivariate extension (Sampson (1975)).

A number of implications arise in a statistical context. First, to that of assessing the inherent properties of exponential dispersion models through their variance function $V(\theta)$ (Tweedie (1947), Jørgensen (1987)). The variance function is related to the mean value mapping, $M'(\theta)$, via

$$\frac{d}{d\theta}M'^{-1}(\theta) = V^{-1}(\theta)\ . \tag{5.4.1}$$

Following Sampson (1975) no exponential dispersion model exists with a polynomial mean function of degree two or more. This discounts a large class of possible variance functions through equation (5.4.1). For example; power variance functions with power $1-\frac{1}{n}$, $n \in \mathbb{N}$ (a special case of theorem 2, Jørgensen (1987)), certain variances, $V(\theta)$, of the form $\sqrt{a+b\theta}$ as then the mean function is quadratic. (see also Burridge (1987)). However, it should be noted that the variance function can be a polynomial, for example $M(\theta) = 1/(1+e^{-\theta})$ corresponds to a Bernoulli random variable with $V(\theta) = \theta(1-\theta)$. Although we note here that there are only limited possibilities for quadratic variance functions (Morris (1982), Bar-Lev and Stramer (1987)). The required conditions on $M(\theta)$ for the family to be infinitely divisible are contained in (Sampson (1975), Jørgensen (1986)).

Secondly, in the setting of Bayesian posterior distributions the roles of θ and X are reversed, but the posterior is still a member of the exponential family with respect to the prior for θ. There are a number of corresponding results :

(i) The posterior mean is never a polynomial of degree greater than one (Goldstein (1977)).

(ii) The posterior mean is never of the form X^{α}, $\alpha > 1$ (as there is no power variance exponential dispersion model with power between 0 and 1, theorem 2 (Jørgensen (1987))).

Thus we cannot attempt to predict the posterior mean by a polynomial of degree greater than one or power greater than one, a type of shrinkage property for such families. For characterisation purposes the above result becomes

Lemma : $E(\theta|X)$ characterises the posterior, prior pair for a likelihood in the exponential family.
Proof : Directly applying Sampson's theorem implies $E(\theta|X)$ characterises $p(\theta|X)$ for all X . The prior is then determined almost everywhere.
For example, the constraint $E(\theta|X) = \frac{X^3+6X}{2X^2+4}$ characterises the prior $p(\theta)\ \alpha\ \theta^2 e^{-\frac{1}{2}\theta^2}$ under a $N(\theta,1)$ likelihood.

It is possible to recast the integral constraint (5.3.1) into a differential equation for $p(\theta)$. First, when the support of μ is $\mathbb{R}$, we apply the methods of Ralescu and Ralescu (1981). Secondly, we adopt the approach of Diaconis and Ylvisaker (1979) which also has application when the support of μ is the nonnegative integers.

(5.4.1) Application of Ralescu and Ralescu (1981)

Concerned with admissible estimation in the one parameter exponential family with support $\mathbb{R}$ taking the form

$$E(\psi(\theta)|X) = \frac{aX+b}{cX+d} .$$

Ralescu and Ralescu (1981) determined a first order differential equation for $p(\theta)$ (see also Ghosh and Meeden (1977)). In the context of constraint (5.3.1) we have the following

Theorem : Suppose equation (5.3.1) is satisfied with $\eta(\theta)$ a polynomial. Then there exists a differential equation of order degree η that the prior $p(\theta)$ must satisfy.

Proof : Suppose equation (5.3.1) holds, let $\beta = e^{-M}$. Then by definition of $E(\psi(\theta)|X)$ the following holds

$$\eta(X) \int_{-\infty}^{\infty} e^{\theta X}\beta p = \int_{-\infty}^{\infty} \psi\beta p e^{\theta X} . \tag{5.4.2}$$

Now due to the fact that η is a polynomial $\int_{-\infty}^{\infty} \eta(X)e^{\theta X}\beta p$ can be rearranged, by parts, as

$$\int_{-\infty}^{\infty} \eta(X)e^{\theta X}\beta p = \eta(D(\beta p)) ,$$

where D denotes the differential operator (e.g. D^n denotes the nth derivative with respect to θ). Substituting into equation (5.3.1) and using the uniticity of the Laplace transform (as a function of X) gives

$$\eta(D(\beta p)) = \psi\beta p ,$$

a differential equation of order degree η for $p(\theta)$, in theory yielding a solution for the prior, which then must be checked to be a proper prior.

(5.4.2) Application of Diaconis and Ylvisaker (1979)

In the previous section (5.4.1) the support of μ was $\mathbb{R}$, hence integration by parts does not involve endpoint evaluations. Suppose instead that X is discrete and the support of μ is the nonnegative integers. For this setting, Θ is an unbounded open interval to the left. Here it will be assumed that $\Theta = (-\infty,\theta_0)$, where $\theta_0 < \infty$. As such the technique in (5.4.1) is not directly applicable, instead (5.3.1) is rewritten in integral form as in Diaconis and Ylvisaker (1979).

In this discrete setting, the required moment characterisation property is that a signed measure on a compact interval having all moments zero must in fact be zero.

Let $p(\theta) = d\tau(\theta)$, then constraint (5.3.1) can be written as an integral equation,

$$\int_{-\infty}^{\theta_o} e^{X\theta}\psi(\theta)e^{-M(\theta)}d\tau(\theta) = \eta(X)p(X) \ .$$

Without loss of generality assume that $\eta(0) = 0$, then rewriting the l.h.s. gives

$$\int_{-\infty}^{\theta_o} \psi(\theta)\left(\int_{-\infty}^{\theta} Xe^{Xy}dy\right)e^{X\theta}d\tau(\theta) = \int_{-\infty}^{\theta_o} Xe^{Xy}\left(\int_{y}^{\theta_o} \psi(\theta)e^{-M(\theta)}d\tau(\theta)\right)dy \ .$$

Now the inner integral can be rewritten, using constraint (5.3.1) when $X = 0$, as

$$\int_{y}^{\theta_o} \psi(\theta)e^{-M(\theta)}d\tau(\theta) = \int_{-\infty}^{y} \psi(\theta)e^{-M(\theta)}d\tau(\theta) \ .$$

Substituting back, interchanging θ and y, we obtain,

$$X\int_{-\infty}^{\theta_o} e^{X\theta}\left(\int_{-\infty}^{\theta} \psi(y)e^{-M(y)}d\tau(y)\right)d\theta = \eta(X)p(X) \ .$$

Let $F(\theta)$ denote the inner integral, then

$$\int_{-\infty}^{\theta_o} e^{X\theta}F(\theta)d\theta = \frac{\eta(X)}{X}\int_{-\infty}^{\theta_o} e^{X\theta}e^{-M(\theta)}d\tau(\theta) \ .$$

Let $t = e^{\theta}$, note that the support becomes the compact interval $[0, e^{\theta_o}]$. Suppose $\eta(X) = X$, then the r.h.s. of (5.4.3) reduces to a set of moment constraints for which the above characterisation theorem yields,

$$F(\theta)d\theta = e^{X\theta}d\tau(\theta) \ .$$

Note that differentiating once with respect to θ yields a first order differential equation for $p(\theta)$, as in Diaconis and Ylvisaker (1979).

Reapplying the procedure yields

$$\int_{-\infty}^{\theta_o} e^{X\theta}F(\theta)d\theta = \int_{-\infty}^{\theta_o} Xe^{Xy}\left(\int_{y}^{\theta_o} F(\theta)d\theta\right)dy \ ,$$

the inner integral can be expressed as

$$\int_{-\infty}^{\theta_0} Xe^{Xy}\left(\int_{y}^{\theta_0} F(\theta)d\theta\right)dy \ .$$

Again, the inner integral can be expressed as

$$\int_{y}^{\theta_0}\left(\int_{-\infty}^{\theta} \psi(t)e^{-M(t)}d\tau(t)\right)d\theta = \int_{-\infty}^{y} \psi(t)e^{-M(t)}\left(\int_{y}^{t} d\theta\right)d\tau(t) \ .$$

Hence, on interchange of θ and y, constraint (5.3.1) becomes

$$\int_{-\infty}^{\theta_0} e^{Xy}\int_{-\infty}^{y}(t-y)\psi(t)e^{-M(t)}d\tau(t)dy = \frac{\eta(X)}{X^2}\int_{-\infty}^{\theta_0} e^{X\theta}e^{-M(\theta)}d\tau(\theta) \ .$$

Suppose $\eta(X) = X^2$, for then the same characterisation argument applies yielding,

$$\int_{-\infty}^{\theta}(y-\theta)\psi(t)e^{-M(t)}d\tau(t) = e^{-M(\theta)}d\tau(\theta) \ .$$

Differentiating this expression twice with respect to θ will yield a second order differential equation for $p(\theta)$ involving derivatives of ψ and M. Note that the characterisation step will still hold if X is in fact continuous.

(5.4.3) Results of Goldstein (1975)

By employing linear functionals Johnson (1957, 1967) and Ericson (1969) deduced constraints on prior moments. These results were extended by Goldstein (1975) reducing the problem to a set of recurrence relations for the prior moments. The technique can clearly be applied in the case where ψ and ϕ are polynomials, unfortunately these seem tractable only when ψ and η are quadratic (or less), for which we have the following example.

Example : (i) The constraint $E(\theta^2|X) = X^2$ has no solution for prior with finite variance.
(ii) Suppose that either $E(\theta^3) = 0$ or $E(\theta) = 0$. Then for $0 < c < 1$ the constraint

$$E(\theta^2|X) = c + c^2X^2 \qquad (5.4.4)$$

characterises a N(0, τ^2) prior, where $\tau^2 = c/(1-c)$.

Proof : From (5.4.4), we have for all n,

$$cE(X^{n-2}) + c^2E(X^n) = E(E(X^{n-2}\theta|X)) = E(\theta E(X^{n-2}|\theta))$$

This leads to recurrence relations, for $n \geq 2$, of the form

$$c\sum_{r=0}^{n-2}\binom{n-2}{r}m_{n-2-r}\bar{v}_r+c^2\sum_{r=0}^{n}\binom{n}{r}m_{n-r}\bar{v}_r=\sum_{r=0}^{n-2}\binom{n-2}{r}m_{n-2-r}\bar{v}_{r+2}\ , \qquad (5.4.5)$$

where $\bar{v}_r$ is the rth central moment of the prior and m_i is defined by,

$$m_i = E((X-\theta)^i|\theta)\ .$$

Now suppose likelihood is symmetric (i.e. $m_i = 0$). Letting $n = 2$ in (5.4.5) uniquely determines $\bar{v}_2$ as, $\bar{v}_2 = c/(1-c)$, which implies $0 < c < 1$ for a solution to exist. If n is now even in (5.4.5), we see recursively that $\bar{v}_r$, r even, are uniquely determined. Similarly the odd moments are uniquely determined in terms of $\bar{v}_3$, or $\bar{v}_1$. Suppose $\bar{v}_3 = 0$, from recurrence relations this implies $\bar{v}_r = 0$, r odd. It is easy to check that (5.4.4) is satisfied for a $N(0, \tau^2)$ prior, where $\tau^2 = c/(1-c)$. Hence a prior satisfying (5.4.4) has the same moments as a $N(0, \tau^2)$ distribution. We now show that this is the only possibility.

Computing even moments of the above normal prior gives,

$$(\bar{v}_{2n})^{-\frac{1}{2n}} = \frac{1}{\sqrt{2\pi}(n-\frac{1}{2})!}2^{n+\frac{1}{2}}\ .$$

Estimating this we obtain,

$$(\bar{v}_{2n})^{-\frac{1}{2n}} > \frac{1}{n-\frac{1}{2}},$$

therefore,

$$\sum_{n=1}^{\infty}(\bar{v}_{2n})^{-\frac{1}{2n}} = \infty\ .$$

Theorem (Carlemann) : A distribution F is uniquely determined by its moments, $\bar{v}_n$ if

$$\sum_{n=1}^{\infty}(\bar{v}_{2n})^{-\frac{1}{2n}} = \infty\ .$$

Applying this we see that (5.4.4) characterises a normal prior, as required.

A general theorem for the linear case, combining the use of characteristic functions and Carlemann's theorem, is given in Diaconis and Ylvisaker (1985), extending the characterisation of Goldstein (1975).

(5.5) Associated differential equations for moments and priors

Consider the exponential family through μ in its natural parameterisation, from (5.1.1) the posterior will be of the form,

$$p(\theta|X) = p(\theta)e^{X\theta - M(\theta) - B(X)} .$$

By interchanging derivatives and integrals, with the required regularity conditions (see section (5.9)), a differential equation in X can be formed for posterior moments as follows (see also Goldstein (1977)).

Theorem : Let $f(X) = E(\psi(\theta)|X)$, under suitable regularity conditions,

$$\frac{df}{dX} = E(\theta\psi(\theta)|X) - f(X)E(\theta|X) .$$

Proof : For any polynomial $\psi(\theta)$ we have

$$E(\psi(\theta)|X) = \frac{1}{B(X)} \int_{-\infty}^{\infty} \psi(\theta)p(\theta)e^{X\theta - M(\theta)}d\theta .$$

Differentiate partially with respect to X, as $\psi(\theta)$ is a polynomial this is valid (see Appendix (5.9) and Weir (1973) p.118, 256). Let $f(X) = E(\psi(\theta)|X)$, then

$$\frac{df}{dX} = \frac{1}{B(X)} \int_{-\infty}^{\infty} \theta\psi(\theta)p(\theta)e^{X\theta - M(\theta)}d\theta - \frac{B'(X)}{B(X)^2} \int_{-\infty}^{\infty} \psi(\theta)p(\theta)e^{X\theta - M(\theta)}d\theta .$$

Hence,

$$\frac{df}{dX} = E(\theta\psi(\theta)|X) - E(\psi(\theta)|X)E(\theta|X)$$

therefore,

$$\frac{df}{dX} = E(\theta\psi(\theta)|X) - f(X)E(\theta|X) . \qquad (5.5.1)$$

Corollaries : (i) Applying (5.5.1) in the case $\psi(\theta) = \theta$ gives

$$\frac{df}{dX} = V(\theta|X) .$$

(ii) If $\psi(\theta)$ is a quadratic, substituting into (5.5.1) and letting $y = f(X)$ we obtain a differential equation of the form

$$\frac{dy}{dX} = \bar{\eta}(X) - (y^2 - b_1 y) .$$

(5.6) Linear regression in X

In the case where $\eta(X)$ is linear in X, it is possible to determine solutions to the *a posteriori* constraint (5.3.1) for the exponential family. The main result aiding in characterisations is theorem 4 of Diaconis and Ylvisaker (1985). It states,

Theorem (Diaconis and Ylvisaker (1985)) : Let X be from the family $\{P_\theta\}$, where the support of μ contains an open interval in $\mathbb{R}^d$. Then if

$$E(M'(\theta)|X) = aX+b \tag{5.6.1}$$

for some non zero constant a and vector b, then, $a > 0$ and $p(\theta)$ is absolutely continuous with respect to lebesgue measure with density

$$p(\theta) \propto \exp(a^{-1}(b\theta-(1-a)M(\theta))) \ . \tag{5.6.2}$$

Appropriate versions for discrete data are given in Diaconis and Ylvisaker (1979). The above characterisation (5.6.1) can be viewed as a definition of the conjugate family. Furthermore the characterisations are invariant to diffeomorphic transformations of the parameter space (Diaconis and Ylvisaker (1985)).

From a robustness viewpoint, as in the location case, we note that (5.6.1) can be rewritten in terms of the score function of the likelihood as

Corollary : Suppose that

$$E_{\theta|X}\left(\frac{d}{d\theta}\log f(X|\theta)\right) = a^*X+b \ , \tag{5.6.3}$$

then $a^* < 1$, and $p(\theta)$ is given by the conjugate family (5.6.2).

Proof : By definition of the density for X, the likelihood score function is

$$\frac{d}{d\theta}\log f(X|\theta) = X - M'(\theta)$$

hence (5.6.3) becomes,

$$E(M'(\theta)|X) = (1-a^*)X + b \ ,$$

the result follows from (5.6.1). Note that, the condition on a^* is a type of shrinkage property for such families.

Consider the example of linear regression in X, with a $N(\theta,1)$ likelihood where uniqueness is only apparent if ψ is also linear. That is, suppose $E(\psi(\theta)|X) = \eta(X) = X$, then there are two

possibilities,

(i) Suppose that $\psi(\theta)$ is linear, see Diaconis and Ylvisaker (1985). Here the solution is unique.

(ii) Consider priors of the form $p(\theta) \propto e^{-\frac{1}{2}S(\theta)}$, where S is a polynomial such that $p(\theta)$ forms a density. The following lemma establishes a relation between S and ϕ.

Lemma : Given any odd polynomial ϕ (assumed monic) by choosing $S(\theta)$ such that

$$S'(\theta) = 2(\psi(\theta)-\theta) , \tag{5.6.4}$$

then $E(\psi(\theta)|X) = X$ for the prior $p(\theta) \propto e^{-\frac{1}{2}S(\theta)}$.

Note that, as ϕ is of odd degree and monic the solution to (5.6.4) satisfies required integrability for the prior. A solution will not exist if ϕ is of even degree (see Corollary (5.3)).

Proof : Consider the prior $p(\theta)\ \alpha\ e^{-\frac{1}{2}S(\theta)}$. Then, by Bayes theorem, we have posterior,

$$p(\theta|X) \propto e^{-\frac{1}{2}(S(\theta)+\theta^2-2X\theta)} .$$

Define

$$C(X) = \int_{-\infty}^{\infty} e^{-\frac{1}{2}(S(\theta)+\theta^2-2X\theta)} d\theta ,$$

which exists for all X if $p(\theta) \in L^1$. Hence, we have,

$$E(\psi(\theta)-X|X) = \frac{1}{C(X)} \int_{-\infty}^{\infty} (\psi(\theta)-X) e^{-\frac{1}{2}(S(\theta)+\theta^2-2X\theta)} d\theta . \tag{5.6.5}$$

If we now choose $S(\theta)$ such that it satisfies (5.6.4) then the r.h.s. of (5.6.5) is directly integrable by parts, giving zero. Hence,

$$E(\psi(\theta)|X) = X$$

as required. Unfortunately, the class of priors used in (iii) are not analytically tractable in contrast to (5.3.8), so posterior moments for this class have to be obtained numerically.

(5.7) Further constraints on posterior moments characterising priors

Consider the generalisation of the problem in (5.3.1), namely that for all X the *a posteriori* assumption that,

$$E(\psi(\theta, X)|X) = 0 . \tag{5.7.1}$$

First, we prove a lemma that is useful in ensuing characterisations.

Lemma (5.7) : Let $\psi_2(\theta) > 0$ a.e. on the range of θ . Let $\hat{q}(\theta) = \psi_2(\theta)p(\theta)$. Assume $p(\theta)$ is such

that $\hat{q}(\theta) \in L^1$. Let E_p denote the posterior expectation under the prior $p(\theta)$. Then,

$$E_p(\psi_1\psi_2) = E_q(\psi_1)E_p(\psi_2) \ . \tag{5.7.2}$$

Proof : By assumption $\int \hat{q}$ exists, clearly it is not zero as $\psi_2 > 0$ and $\int p = 1$. Similarly, $\int qf$ exists and is not zero.

By Bayes' theorem we can compute the l.h.s. (5.7.2) as follows,

$$E_p(\psi_1\psi_2) = \frac{\int \psi_1\psi_2 pf}{\int pf} = \frac{\int \psi_1 \hat{q}f}{\int \hat{q}f}\frac{\int \hat{q}f}{\int pf}$$

$$= \frac{\int \psi_1 qf}{\int qf}\frac{\int \psi_2 pf}{\int pf}$$

$$= E_q(\psi_1)E_p(\psi_2) \ .$$

Combining this lemma with the Diaconis and Ylvisaker characterisation (see (5.6.1)) gives the following theorem from which many interesting characterisations can be obtained.

Theorem : Let $\psi_1(\theta) = E_\theta(X)$. Suppose the conditions of lemma (5.7) are satisfied. Then

$$E_p(\psi_1\psi_2 - (aX+b)\psi_2) = 0$$

characterises the prior

$$p(\theta) \propto \frac{q(\theta)}{\psi_2(\theta)} \ .$$

where $q(\theta)$ is the conjugate prior associated with (5.6.2).

Proof : First, apply the characterisation (5.6.1) in the form,

$$E_q(\psi_1) = aX+b \tag{5.7.3}$$

characterises $q \propto \psi_2 p$ as a conjugate prior. Hence as $\psi_2 > 0$, $p(\theta)$ is characterised as required.

Furthermore, as $E_p(\psi_2) > 0$, lemma (5.7) shows that (5.7.3) is equivalent to

$$E_p(\psi_1\psi_2 - (aX+b)\psi_2) = 0 \ . \tag{5.7.4}$$

Therefore (5.7.4) characterises $p(\theta)$, as required.

Examples : (i) t-distributions: From (5.7.4), with a $N(\theta,1)$ likelihood

$$p(\theta) = \frac{1}{1+\theta^2}e^{-\frac{1}{2}\theta^2}$$

is characterised by

$$E_p((1+\theta^2)(2\theta - X)) = 0 \,. \tag{5.7.5}$$

Proof : The posterior only depends on product $f(X|\theta)p(\theta)$ thus if $p(\theta)$ is characterised for likelihood $f(X|\theta)$ then $p(\theta)e^{\frac{1}{2}\theta^2}$ is characterised for likelihood $f(X|\theta)e^{-\frac{1}{2}\theta^2}$. Hence (5.7.5) characterises a Cauchy prior for a $N(2\theta, 2)$ likelihood. Similarly a t_m distribution can be characterised.

(ii) The prior $p(\theta) \propto e^{-S(\theta)}$ in section (5.6): Here $p(\theta)$ is characterised by

$$E_p(\exp(S(\theta) - \tfrac{1}{2}\theta^2)(\theta - \tfrac{1}{2}X)) = 0 \,.$$

(5.7.1) Characterisations involving ratios

Some previous attention has been give to characterisations through ratios (e.g. Bilkidar and Patil (1968)). Here, as $E_p(\psi_2) > 0$, lemma (5.7) can be rewritten

$$\frac{E_p(\psi_1\psi_2)}{E_p(\psi_2)} = E_q(\psi_1) \tag{5.7.6}$$

therefore,

$$\frac{E_p(\psi_1\psi_2)}{E_p(\psi_2)} = aX + b \tag{5.7.7}$$

characterises $\psi(\theta)$ as $\dfrac{q(\theta)}{\psi_2(\theta)}$, where $q(\theta)$ is the corresponding conjugate prior from (5.6.2).

Example : Suppose that θ is restricted to be positive. Then (5.7.7) applies with $\psi_2 = \theta$. Conditions for $p(\theta)$ are that it has finite first moment and $\int qf$ exists, under these conditions

$$\frac{E_p(\psi_1\theta)}{E_p(\theta)} = aX + b$$

characterises $p(\theta) = \dfrac{1}{\theta}\hat{p}(\theta)$.

Characterisations involving variance

We have from (5.5.1) that

$$V(\theta|X) = \frac{d}{dX}E(\theta|X) \,.$$

Hence assuming $V(\theta|X)$ is a known function of X, thus $E(\theta|X)$ is determined up to a constant. Then by Sampson's theorem the prior will be characterised. This has an intuitive robustness appeal, as one might want $V(\theta|X)$ to reflect uncertainty when X is large.

(5.8) Application to other likelihoods.

All the methods used in previous sections can be applied to likelihoods in the exponential family in its natural parameterisation as long as $X \in \mathbb{R}$. This is not typical of the well known members, so we view some of them separately. We concentrate mainly on the on the case of linear regression.

Examples : (i) Binomial

In its natural parameterisation we have likelihood,

$$f(X|\theta) \propto \exp\left(X \log\left(\frac{\theta}{1-\theta}\right) - \log(1-\theta)\right) .$$

The natural parameter is

$$\bar{\theta} = \log\left(\frac{\theta}{1-\theta}\right) .$$

Unfortunately, X takes only a finite number of values. Hence some proofs, where $X \to \infty$, are not applicable. In the case of linear regression in X, the prior

$$p(\theta) \propto (1-\theta) e^{-\int \psi(\theta) d\theta}$$

is integrable for $\theta \in (0,1)$, and gives

$$E_{\theta|X}\left(\psi\left(\log\left(\frac{\theta}{1-\theta}\right)\right) \right) = X .$$

Characterisations from section (5.7)

Returning to the usual binomial parameterisation, then $\theta > 0$. We now list some of the characterisations which are direct consequences of lemma (5.7) and the result (see Diaconis and Ylvisaker (1985)) that $E_p(\theta) = \frac{1}{3}(1+X)$ characterises the prior $p(\theta) = U(0,1)$.

The reference prior, $p(\theta) \propto \frac{1}{\theta^{\frac{1}{2}}(1-\theta)^{\frac{1}{2}}}$, is characterised by

$$E_p((\theta - \frac{1}{3}(1+X))\theta^{\frac{1}{2}}(1-\theta)^{\frac{1}{2}}) = 0 .$$

Furthermore, it is possible to characterise the family of conjugate priors via ratios of posterior moments as follows.

Lemma : The prior $p(\theta) \propto \frac{1}{\theta^n}\hat{p}(\theta)$ is characterised by

$$\frac{E_p(\theta^{n+1})}{E_p(\theta^n)} = aX+b\ , \tag{5.8.1}$$

where $\hat{p}$ is characterised by $E_{\hat{p}}(\theta) = aX+b$, that is $\hat{p}$ is a Beta distribution.

Hence (5.8.1) characterises the set of conjugate priors and improper priors. Note that, we only really need $\int qf$ not to vanish in lemma (5.7).

(ii) Poisson

Here the likelihood is given by

$$f(X\,|\,\theta) \propto \exp(\, X\log\,\theta-\theta\,)$$

The natural parameter is $\bar{\theta} = \log\theta$. Consider, for example, the case of linear regression. The relevant prior would have to be

$$p(\theta) \propto \exp(e^{\theta}-\int \psi(\theta)d\theta)$$

which is not integrable on $\mathbb{R}^+$ for polynomial $\psi(\theta)$.

(iii) Gamma

In this case the natural parameter is $\theta \in (0,\infty)$. Hence the results of Section (5.4.1) are not directly applicable. The Diaconis and Ylvisaker characterisation (see (5.6.1)) is

$$E_{\hat{p}}(\theta^{-1}) = aX+b\ .$$

Hence, lemma (5.7) applies, for example,

(i) The prior $p(\theta) \propto \theta\,\hat{p}(\theta)$ is characterised by

$$\frac{E_p(\theta^{-2})}{E_p(\theta^{-1})} = aX+b\ .$$

(ii) The prior $p(\theta) \propto \theta^{-2}\hat{p}(\theta)$ is characterised by

$$E_p(\theta-(aX+b)\theta^2) = 0$$

again showing the polynomial nature of the characterisations.

(iv) Normal likelihood with known mean, unknown variance

Without loss of generality suppose that the mean is zero, then the likelihood is of the form

$$f(y\,|\,\tau) \propto \exp(\, \tau y+\tfrac{1}{2}\log\,\tau\,)$$

where $\tau > 0$, $y < 0$. The conjugate prior is

$$p(\theta) \propto \tau^{\alpha} \exp(\alpha\tau) \; . \tag{5.8.2}$$

The results obtained are similar to the gamma likelihood, for example, ratios of posterior moments characterise the set of priors (5.8.2).

If we now consider a normal likelihood where the variance is a function of the mean, for example, $N(\theta, \theta^{-2})$. Then the class of priors, $p(\theta) \propto e^{-S(\theta)}$ where S is a polynomial, then we have equations of the form (5.7.1),

$$E(\psi(\theta, X)|X) = 0$$

where $\psi(\theta, X)$ is a polynomial in θ and X.

(v) Uniform likelihood, $U(0,\theta)$

If $p(\theta)$ is a gamma distribution, then after computation we obtain equations of the form

$$E(\psi(\theta)|X) = \frac{r(X)}{s(X)} \; ,$$

where r and s are polynomials such that deg r - deg s = deg ψ.

Thus likelihoods outside the exponential family admit such polynomial expressions, although characterisation results are unknown.

(5.9) Appendix : Justification of differentiation under the integral sign

In order for validity of theorem (5.1) we give regularity conditions for the function $I(X)$ to have derivatives of all orders, where

$$I(X) = \int_{-\infty}^{\infty} \psi(\theta) p(\theta) e^{X\theta - M(\theta)} d\theta$$

for polynomials ψ. Suppose $p(\theta)$ is bounded and that $M(\theta)$ is continuous with $M(\theta) > a\theta^2$ for sufficiently large θ and some constant a (w.l.o.g. take $a = 1$).

Let

$$f(X, \theta) = \psi(\theta) p(\theta) e^{X\theta - M(\theta)} \; .$$

Therefore

$$\frac{d}{dX} f(X, \theta) = \theta\psi(\theta) p(\theta) e^{X\theta - M(\theta)} \; .$$

In order for differentiation under the integral sign to hold, we require a dominating function,

independent of X, for $|\frac{d}{dX}f(X, \theta)|$ (see Weir (1973) p.118, p.256). Unfortunately, there does not exist a dominating function which works for all $X \in \mathbb{R}$.

However, every point in $\mathbb{R}$ lies in an open interval $(-x, x)$; so it will be enough to prove that for each positive real x, $\frac{d}{dX}f(X, \theta)$ is dominated by a function in L^1. Now, for $|X| < x$,

$$\left|\frac{d}{dX}f(X, \theta)\right| \leqslant A|\psi(\theta)|e^{-M(\theta)-|\theta|} \tag{5.9.1}$$

where $|p(\theta)| \leqslant A$ by hypothesis. Let $g(\theta)$ equal the r.h.s. of (5.9.1). Clearly, $g(\theta)$ is continuous, therefore integrable on compact intervals and hence on $(-1-x, 1+x)$. For $|\theta| \geqslant 1+x$ the condition on $M(\theta)$ implies $M(\theta) \geqslant (1+x)|\theta|$. This implies that in this region $g(\theta)$ is dominated by a polynomial times $e^{-|\theta|}$. Therefore $g(\theta) \in L^1$. Hence $I(X)$ possesses a first derivative and by reapplication has derivatives of all orders.

In order for the validity of Masreliez theorem, assume that the likelihood is bounded. We require to be able to differentiate $\int_{-\infty}^{\infty} f(X-\theta)p(\theta)$ twice, where the prior is normal. By a linear transformation this is equivalent to $\int_{-\infty}^{\infty} f(\theta)p(X-\theta)$. Due to the fact that p is normal the conditions for $M(\theta)$ hold and together with the boundedness of the likelihood implies that the above holds and we have the desired result. The boundedness condition on the likelihood is necessary as the theorem is invalid for the function $\frac{1}{\sqrt{|X-\theta|}}\exp(-(X-\theta)^2)$.

(5.10) Discussion

This Chapter explores the behaviour of posterior functionals, primarily the mean and variance, in the class of exponential families and scale mixtures of normality. Representation properties of posterior moments help understand the flexibility of the model as do their characterisation properties. Quantitative measures (for example, the score function) appear in such representations and give a guide to further interpretations. Clearly, one possible extension is to explore other properties of the posterior, for example, unimodality (see Andrews *et al* (1973)). The characterisation results for the exponential family rely on theorems that can be arrived at by a variety of techniques; for example, moment generating functions, differential and integral equations.

Fisher's information again plays a central role in describing the behaviour of the posterior variance in the exponential family, thus allowing results from Chapter 3 characterising prior-likelihood combinations to be applicable.

A further area of application is to a quantitative robustness setting where the determination of the behaviour of posterior functionals over classes of measures is required (Huber (1973), Berger (1984)).

Many of the examples considered require uniqueness of the solutions to be established.

Chapter 6 : Derivatives; distances; sensitivity

Here we review the machinery of Diaconis and Freedman (1986a) for the computation of derivatives of posterior functionals with respect to the modelling components; for example, prior, likelihood or utility structure. To define a suitable notion of derivative for probability measures we require a topological structure and a measure of distance.

(6.1) Derivative of the prior-posterior map

Consider the set of probabilities as a subset of the space of all signed measures equipped with the variational norm as a measure of distance between measures μ and ν, that is

$$\|\mu - \nu\| = \int \left| \frac{d\mu}{d\lambda} - \frac{d\nu}{d\lambda} \right| d\lambda$$

where λ is any σ-finite dominating measure, thus endowing the set of measures with a Banach space structure.

Consider a family of measures $\{Q_\theta : \theta \in \Theta\}$ dominated with respect to some σ-finite measure λ so that all the Q_θ are absolutely continuous with respect to λ with density $f(x|\theta)$. Suppose $f(\cdot,\cdot)$ is measurable and $\sup_\theta f(x|\theta) < \infty$ for every x. Let the Bayes mapping, $B : \mu \to B_\mu$, be defined as

$$B_\mu(d\theta) = \frac{N_\mu(d\theta)}{D_\mu},$$

where

$$N_\mu(d\theta) = f(x|\theta)\mu(d\theta) \text{ and } D_\mu = \int f(x|\theta)\mu(d\theta).$$

Thus N_μ and D_μ are linear functions of the prior measure μ, giving B_μ a structure similar to that of a Möbius transformation. Note that B_μ is only defined for $D_\mu > 0$, but this is a set of P_μ-measure one. Let $\dot{B}_\mu$ denote the derivative of the map B, itself a map of measures, defined by

$$B_{\mu+\delta} = B_\mu + \dot{B}_\mu(\delta) + o(\|\delta\|) \tag{6.1.1}$$

as $\|\delta\| \to 0$, where δ is a signed measure with signed mass zero. The norm of $\dot{B}_\mu$ quantifies the sensitivity of the posterior to small changes in the prior μ. By definition,

$$\| \dot{B}_\mu \| \;=\; \sup_{\| \delta \| \,=\, 1} \| \dot{B}_\mu(\delta) \| \;. \tag{6.1.2}$$

The following theorem gives forms of (6.1.1) and (6.1.2),

Theorem (Diaconis and Freedman (1986) : Let $\sup_0 f(x|\theta) \;=\; \sup_\theta \{ f(x|\theta) : \mu\{\theta\} = 0 \}$, then the derivative of the Bayes map and its norm are given by

$$\dot{B}_\mu(\delta)(\cdot) \;=\; \frac{N_\delta(\cdot)}{D_\mu} - \frac{D_\mu N_\mu(\cdot)}{D_\mu{}^2} \tag{6.1.3}$$

$$\frac{\sup_0 f(x|\theta)}{D_\mu} \;\leqslant\; \| \dot{B}_\mu \| \;\leqslant\; \frac{\sup_\theta f(x|\theta)}{D_\mu} \;; \tag{6.1.4}$$

note that for many priors the upper and lower bounds in (6.1.4) are identical.

Analytical computation is clearly possible for exponential families under a conjugate prior assumption. Approximate computation in the presence of a nuisance parameter is briefly discussed in (6.1.3). It has application in defining an outlier-prone model specification by considering

$$\lim_{x \,\to\, \infty} \| \dot{B}_\mu \| \;,$$

which if an unbounded, leads to a non-robust inference.

(6.2) Relationship with Bayes factors for nested models

Let M_0 and M_1 denote nested parametric models. Consider a family of measures $\{Q_\theta : \theta \in \Theta\}$ with densities with respect to a dominating measure λ given by $f(x|\theta)$. Let μ denote the prior measure. On the basis of the observed data x we require to compare the competing models. Let $p(x|M_0)$ and $p(x|M_1)$ denote the corresponding posteriors. Jeffreys proposes the odds ratio or Bayes factor, $B_{01}(x)$, defined as

$$B_{01}(x) \;=\; \frac{p(x|M_0)}{p(x|M_1)} \;.$$

Consider the chain of maps given by

$$\theta \;\to\; \mu \overset{S}{\to} p(\theta|x, M_0) \overset{R}{\to} p(\theta|x, M_1) \;,$$

where R denotes the inclusion mapping, such that,

$$R(\; p(\theta|x, M_0)\;) \;=\; p(\theta|x, M_1) \;.$$

Define the composite map T by the rule

$$T(\mu) \;=\; R(S(\mu))) \;.$$

The norm of a composite map satisfies

$$\|RS\| \;\leqslant\; \|R\|\,\|S\| \;.$$

Therefore,

$$\|R\| \;\geqslant\; \frac{\|T\|}{\|S\|} \;.$$

Furthermore, by the nested property

$$\sup_{\theta \in M_0} f(x|\theta) \leqslant \sup_{\theta \in M_1} f(x|\theta) \tag{6.2.1}$$

hence, by (6.1.4),

$$\|R\| \;\geqslant\; \frac{\sup_{\theta \in M_1} f(x|\theta)}{p(x|M_1)} \times \frac{p(x|M_0)}{\sup_{\theta \in M_0} f(x|\theta)} \;. \tag{6.2.2}$$

By (6.2.1)

$$\|R\| \;\geqslant\; \frac{p(x|M_0)}{p(x|M_1)} \;=\; B_{01}(x) \;,$$

thus giving an inequality between the norm of the inclusion mapping R and the Bayes factor between models. When comparing models with a large difference in dimensionality (6.2.2) should be employed for it shows how the maximised likelihood affects the derivative.

(6.3) Application to posterior functionals and model choice

First, consider the problem of establishing the sensitivity of the posterior mean with respect to the prior specification. The necessary functional derivative is again in Diaconis and Freedman (1986).

Theorem (Diaconis and Freedman (1986)) : Let the mapping M take $\mu(\cdot)$ to the posterior mean,

$$M : \mu \;\rightarrow\; \frac{\int \theta f(x|\theta)\mu(d\theta)}{D_\mu}$$

Then the derivative is given by

$$\dot{M}_\mu(\cdot) = \frac{N_1(\cdot)}{D_\mu} - \frac{N_1(\mu)}{D_\mu^2} D(\cdot) ,$$

with corresponding norm

$$\| \dot{M}_\mu \| = \frac{1}{2} \text{range} \left((\theta - M_\mu) \frac{f(x|\theta)}{p(x)} \right) .$$

Clearly this can be extended to any posterior functional of θ.

As in (6.1.1) this has application in the exponential family and can be used to warn of possible outlier problems when the inferential problem is that of reporting the posterior mean.

(6.3.1) Application to model choice and nuisance parameters

Consider the parametric modelling framework as specified by the family $\{ X, f(x|\theta, \lambda), \Theta \times \Lambda \}$. In order to assess the behaviour of the nuisance parameter consider the mapping, $B_{\lambda|\theta}$, such that,

$$B_{\lambda|\theta} : p(\lambda|\theta) \rightarrow p(\lambda|x, \theta) .$$

Then by (6.1.4) the norm of the derivative of this map is given by

$$\| \dot{B}_{\lambda|\theta} \| = \frac{\sup_{\lambda|\theta} f(x|\theta, \lambda)}{p(x|\theta)} .$$

Thus from Chapter 4 we see that this is in turn related to the (modified) profile likelihood as follows,

$$\| \dot{B}_{\lambda|\theta} \| = \frac{L_P(\theta)}{p(x|\theta)}$$

$$\| \dot{B}_{\lambda|\theta} \| = \frac{L_{MP}(\theta)}{|\hat{i_\theta}|^{\frac{1}{2}} p(x|\theta)} . \qquad (6.3.1)$$

Note that under a reference prior (6.3.1) simplifies. Furthermore, Patefield (1977) contains formulae to aid in the calculation of the numerator, LeJeune and Faulkenberry (1982) show that under certain transformation properties the derivative can be made unity for some members of the exponential family, leading to a least sensitive inference.

(6.4) An inequality between discrimination information and variation distance

Given two probability measures P and Q, the discrimination information, $I(P, Q)$, arises naturally as a Bayes risk to the pure inference problem of approximating P by Q (Bernardo

(1979a)), the corresponding utility function necessarily being the logarithmic function. However, the variation distance, $V(P, Q)$, seems more natural as it has the direct application in explaining possible sensitivity departures for credible regions and can be used to bound risks for certain smooth decision problems. Here we review the inequalities between I and V, sharpening known inequalities when V is close to 2.

Let P and Q be absolutely continuous with respect to μ, then define,

$$I(P, Q) = \int p \log\left(\frac{p}{q}\right) d\mu \tag{6.4.1}$$

$$V(P, Q) = \int |p-q| d\mu \, . \tag{6.4.2}$$

Considerable attention has been directed to determining a lower bound for I in terms of V, for example, Volkonskij and Rozanov (1959), Pinsker (1961), Csiszár (1967a), Kullback (1969). More recently, Toussaint (1975) proved that,

$$I \geqslant \frac{1}{2}V^2 + \frac{1}{36}V^4 + \frac{1}{288}V^6 \, . \tag{6.4.3}$$

Correspondingly, a bound which works well for V near 2 is given by Vajda (1970),

$$I \geqslant \log\left(\frac{2+V}{2-V}\right) - \frac{2V}{2+V} \, . \tag{6.4.4}$$

This can be used to show that as $V \to 2$, then $I \to \infty$.

Under a restrictive constraint, this upper bound can be replaced by (Toussaint (1975)),

$$I \geqslant \frac{V}{2}\log\left(\frac{4}{4-V^2}\right) . \tag{6.4.5}$$

The following theorem applies results from Kraft (1955), Wolfowitz and Hoeffding (1958) to determine a sharper bound than (6.4.4), but not in the special case of (6.4.5).

Theorem : Suppose I and V are given by (6.4.1) and (6.4.2) respectively, then $V \in [0,2]$ and,

$$I \geqslant \log\left(\frac{4}{4-V^2}\right) .$$

Proof : Following Wolfowitz and Hoeffding,

$$-\frac{1}{2}I = \int p \log\left(\frac{p}{q}\right)^{\frac{1}{2}} d\mu \leqslant \log\left(\int (pq)^{\frac{1}{2}} d\mu\right) .$$

Therefore,

$$1 - \left(\int (pq)^{\frac{1}{2}} d\mu\right)^2 \leqslant 1 - \exp(-I) \tag{6.4.6}$$

but we have, see for instance Kraft (1955) lemma 1,

$$\frac{1}{4}V^2 \leqslant 1 - \left(\int (pq)^{\frac{1}{2}} d\mu\right)^2 . \tag{6.4.7}$$

Combining (6.4.6) and (6.4.7) leads us to,

$$\exp(-I) \leqslant 1 - \frac{1}{4}V^2 . \tag{6.4.8}$$

Therefore,

$$I \geqslant \log\left(\frac{4}{4-V^2}\right)$$

as required. After some elementary algebra it can be shown that (6.4.8) is a sharper bound than (6.4.4) for V near 2.

(6.5) A survey of results concerning distance measures

Here we give a brief review of some of the existing literature concerning the properties of the Bayesian risk, $R(\cdot,\cdot)$, defined on the space of distributions, $\mathcal{P}$. In general, consider a risk in the form of a f-divergence, that is

$$R(P, Q) = \int p\, f\left(\frac{p}{q}\right) d\mu$$

where $f : \mathbb{R} \to \mathbb{R}$ and a corresponding definition for the risk in the discrete case. A comprehensive review of such measures is contained in Csiszár (1977), for a more recent unified approach see Burbea and Rao (1982). For further results see Vajda (1972) and Osterreicher (1972).

Csiszár (1984, 1985) develops the notion into one of a generalised I-projection and discusses the existence and applications in a probabilistic framework. Other useful results concerning special cases of (5.2.1), applicable in a Bayesian framework, are: Blahut (1972) where convexity properties of the Kullback-Leibler measure in a discrete setting are established, Abrahams (1982) where the applications of f-divergences are contrasted and Burbea and Rao (1984) who view the properties in a differential geometric setting.

(6.6) Application to influence and outlyingness

In order to examine changes in *a posteriori* statements resulting from deletion of specific observations from the data set, the following decision theoretic measures have been proposed (Pettit

and Smith (1985), Bernardo (1985)).

$$O(S) = p(\boldsymbol{x}(\bar{S})|\boldsymbol{x}(S)) = \int p(\boldsymbol{x}(S)|\boldsymbol{x}(\bar{S}))p(\psi|\boldsymbol{x}(\bar{S}))d\psi \tag{6.6.1}$$

$$J(S) = \int p(\psi|\boldsymbol{x})\log\left(\frac{p(\psi|\boldsymbol{x}(\bar{S}))}{p(\psi|\boldsymbol{x})}\right)d\psi\,. \tag{6.6.2}$$

where $\boldsymbol{x}(S)$ denotes the elements of the data $\boldsymbol{x}$ whose labels are in $S \in \{\,1,\ldots,n\,\}$ and ψ is the parameter of interest for which we wish to report $p(\psi|\boldsymbol{x})$. For example, if $\psi = x_{n+1}$ we are concerned with the sensitivity of the predictive density $p(x_{n+1}|\boldsymbol{x})$ with respect to $\boldsymbol{x}(S)$.

It is show that $J(S)$ leads to an upper bound to the sensitivity of a Bayesian credible region, irrespective of the dimensionality of ψ. A further measure, $K(S)$, is introduced which measures the relative information gain in S, which has the advantage that it is symmetric in S and is directly related to $O(S)$.

Theorem : Let $O(S)$ and $J(S)$ be defined by (6.6.1) and (6.6.2), respectively. Define $K(S)$, by

$$K(S) = E_{\psi|\boldsymbol{x}(\bar{S})}\left(\log\left(\frac{p(\psi|\boldsymbol{x})}{p(\psi|\boldsymbol{x}(S))}\right) - \log\left(\frac{p(\psi|\boldsymbol{x}(\bar{S}))}{p(\psi)}\right)\right).$$

Then, $K(S)$ is symmetric in S and is given by

$$K(S) = K(\bar{S}) = \log\left(\frac{p(\boldsymbol{x}(S))}{O(S)}\right).$$

Furthermore, for any set A,

$$|P(A|\boldsymbol{x}) - P(A|\boldsymbol{x}(\bar{S}))|^2 \leq 1 - \exp(-J(S))\,. \tag{6.6.3}$$

Proof : By Bayes's theorem,

$$\frac{p(\psi|\boldsymbol{x})}{p(\psi|\boldsymbol{x}(S))} = \frac{p(\boldsymbol{x}(\bar{S})|\psi)}{p(\boldsymbol{x}(\bar{S})|\boldsymbol{x}(S))}$$

therefore,

$$\frac{p(\psi|\boldsymbol{x})}{p(\psi|\boldsymbol{x}(S))} = \frac{p(\boldsymbol{x}(\bar{S})}{p(\boldsymbol{x}(\bar{S})|\boldsymbol{x}(S))}\frac{p(\psi|\boldsymbol{x}(\bar{S}))}{p(\psi)}\,.$$

Hence, by definition of $K(S)$,

$$K(S) = \log\left(\frac{p(\boldsymbol{x}(\bar{S})}{p(\boldsymbol{x}(\bar{S})|\boldsymbol{x}(S))}\right) = \log\left(\frac{p(\boldsymbol{x}(S))}{p(\boldsymbol{x}(S)|\boldsymbol{x}(\bar{S}))}\right)$$

therefore,

$$K(S) = \log\left(\frac{p(x(S))}{O(S)}\right).$$

By applying inequality (6.6.3) we see that

$$|P(A|x) - P(A|x(\bar{S}))|^2 \leqslant \sup_A |P(A|x) - P(A|x(\bar{S}))|^2 \leqslant 1 - \exp(-J(S)) .$$

For small deviations, we use the bound (6.4.3), which neglecting terms of order V^3 gives, for any set A,

$$|P(A|x) - P(A|x(\bar{S}))|^2 \leqslant \tfrac{1}{2}J \tag{6.6.3}$$

as required.

(6.7) Application in decision theory

Consider a decision problem specified by a family of probability measures $\{P_\theta \mid \theta \in \Theta\}$ and a loss function L. Let $\delta(x)$ and $\delta(x(\bar{S}))$, respectively, denote the Bayes rules associated with or without using $x(S)$. It is natural in this context, following a suggestion of Diaconis and Ylvisaker (1985), to define a distance between the two procedures by the difference in the associated risks, hence inducing a measure of influence on the set S given by

$$I(S) = \int p(\psi|x)L(\psi, \delta(x))d\psi - \int p(\psi|x(\bar{S}))L(\psi, \delta(x(\bar{S})))d\psi . \tag{6.7.1}$$

Assuming that our loss function satisfies suitable smoothness conditions we show that a rough bound exists between this natural measure $I(S)$ and $J(S)$.

Theorem : Suppose the loss function L is bounded and is Lipschitz continuous in its second argument i.e. there exist constants A, B such that,

$$|L(\theta, \delta)| \leqslant A$$

$$|L(\theta, \delta_1) - L(\theta, \delta_2)| \leqslant B|\delta_1 - \delta_2| \tag{6.7.2}$$

for all rules δ, δ_1, δ_2, and $\theta \in \Theta$. Then the following inequality holds

$$|I(S)| = A\sqrt{2J} + B|\delta(x) - \delta(x(\bar{S}))| .$$

Proof : By definition (6.7.1) and smoothness condition (7.4.2),

$$I(S) = E_{\psi|x(\bar{S})}(L(\psi, \delta(x)) - L(\psi, \delta(x(\bar{S})))) + \int (p(\psi|x) - p(\psi|x(\bar{S})))L(\psi, \delta(x))d\psi$$

therefore,

$$|I(S)| \leqslant E_{\psi|x(\bar{S})}(|L(\psi, \delta(x)) - L(\psi, \delta(x(\bar{S})))|) + A \int |p(\psi|x) - p(\psi|x(\bar{S}))| d\psi .$$

Using (6.7.2) and (6.4.3) gives

$$|I(S)| \leqslant A\sqrt{2J} + B\,|\delta(x) - \delta(x(\bar{S}))| \tag{6.7.3}$$

as required.

The above result shows that the natural measure $I(S)$ can be bounded by a sum of two terms; the first involving the measure of influence J, the second the absolute difference in the Bayes rules.

The bound (6.7.3), although being too coarse for practical application, gives a quantitative meaning to the possible sizes of departures to be expected in a sensitivity analysis concerning outliers. For a more rigorous bound we could apply a result due to LeCam (1982), Birge (1980) which bounds the Hellinger distance which in turn is related to J via inequality (6.4.7).

(6.8) Application to moments of Bayes factors

In a statistical analysis we are forever elaborating on our current modelling framework, whether through more data or a model elaboration, albeit with necessary care and attention (Box (1980), Smith (1983, 1986)). In order to carry out such a procedure we often entertain, *a priori*, the plausibility of two competing models. *A posteriori*, these models are then compared on the basis of a Bayes Factor $B_{01}(x)$. The moments of the Bayes factor can be related to Rényi's α-distance between p and q (Good (1984)) and properties of these moments can then be expressed in terms of the distance measure $I(P, Q)$ via an inequality similar to (6.4.6). Furthermore, under suitable regularity conditions, properties of the α-moment with respect α can be studied, complementing the results of Good (1984).

In the notation of (6.4), define Rényi's α-distance, $I_\alpha(p, q)$, by

$$I_\alpha(p, q) = \int \left(\frac{p}{q}\right)^\alpha p d\mu$$

The following lemmas examine the properties of $I_\alpha(p, q)$ with respect to α and the Kullback-Leibler distance between p and q.

Lemma : Suppose p and q are two probability densities, then

$$I_\alpha(p, q) \geqslant \exp\left(\alpha \int p \log\left(\frac{p}{q}\right) d\mu\right) .$$

Proof : By the convexity of the logarithmic function we have,

$$\alpha \int p \log\left(\frac{p}{q}\right) d\mu \;=\; \int p \log\left(\left(\frac{p}{q}\right)^{\alpha}\right) d\mu \leqslant \log \int \left(\frac{p}{q}\right)^{\alpha} p d\mu \;.$$

Therefore,

$$I_{\alpha}(p, q) \geqslant \exp\left(\alpha \int p \log\left(\frac{p}{q}\right) d\mu\right) . \tag{6.8.1}$$

Define the distance measure $k_{\alpha}(p, q)$, by

$$k_{\alpha}(p, q) \;=\; (I_{\alpha}(p, q))^{\frac{1}{\alpha}} .$$

If we impose mild regularity conditions on p and q, allowing interchange of derivative and integral, then the measure possesses a derivative with respect to α given by

Lemma : Suppose p and q satisfy the required smoothness constraints, then

$$\frac{d}{d\alpha}(k_{\alpha}(p, q)) \;=\; (I_{\alpha}(p, q))^{1-\alpha} \frac{1}{\alpha} \int p \log\left(\frac{p}{q}\right)\left(\frac{p}{q}\right)^{\alpha} d\mu \;.$$

Proof : The smoothness conditions on p and q allow interchange of derivative and integral. By definition,

$$\frac{d}{d\alpha}(k_{\alpha}(p, q)) \;=\; \frac{1}{\alpha}\left(\int \left(\frac{p}{q}\right)^{\alpha} p d\mu\right)^{1/\alpha - 1} \int p \log\left(\frac{p}{q}\right)\left(\frac{p}{q}\right)^{\alpha} d\mu \;.$$

Hence,

$$\frac{d}{d\alpha}\left((I_{\alpha}(p, q))^{\frac{1}{\alpha}}\right) \;=\; (I_{\alpha}(p, q))^{1-\alpha} \frac{1}{\alpha} \int p \log\left(\frac{p}{q}\right)\left(\frac{p}{q}\right)^{\alpha} d\mu \tag{6.8.2}$$

as required.

The preceding lemmas allow us to show that $k_{\alpha}(p, q)$ is an increasing function of α and by continuity with respect to α, inequality (6.8.1) yields,

$$\lim_{\alpha \to 0} k_{\alpha}(p, q) \;=\; \exp\left(\int p \log\left(\frac{p}{q}\right) d\mu\right) .$$

For the discrete case, and an algebraic proof of the above see Good (1984).

A related derivative which sometimes proves useful in calculation the two asymmetric Kullback-Leibler distances between p and q is,

$$\frac{d}{d\alpha}\left(I_{\alpha}(p, q) \right) \;=\; \int p \log\left(\frac{p}{q}\right)\left(\frac{p}{q}\right)^{\alpha} d\mu \;.$$

Note that evaluating at $\alpha = 0,1$ yields the two Kullback-Leibler distances between the measures p and q. (see Loh (1984)). Hence providing a useful analytical trick for simultaneously calculating both Kullback-Leibler measures if Rényi's α-distance is computable; for example, the normal and t-family, normal and double exponential, exponential families.

(6.9) Discussion

This Chapter explores possible sensitivity measures, either being the norm of a derivative of a Bayes mapping, or an induced Bayesian risk from a decision problem concerning the parameter of interest. Computations, approximations and relationships in the form of inequalities for the above distance measures have been discussed.

The properties of information measures aid in the understanding of Bayesian risks, for example, leading to convexity which in turn opens up the possibility of reducing the calculus of variation constraints in Chapter 3 to one of inequality rather than equality.

Clearly, the computation of sensitivity measures for a wide range of problems is required, one such area is the connection between influence measures and reference priors, especially in situations where Fisher's information matrix depends heavily on the design matrix.

The concept of approximation requires further study, for results in this direction see Crain (1977) and Brockett *et al* (1985). A further area of interest is that of the theory of large deviations (Sanov (1957)), where it appears that some results have an interpretation in a Bayesian decision-theoretic framework.

REFERENCES

Abrahams, J. (1982). On the selection of measures of distance between probability distributions. *Info. Science*, 109-113.

Akaike, H. (1980). Ignorance prior distribution of a hyperparameter and Stein's estimator. *Ann. Inst. Statist. Math.*, **32**, 171-178.

Aldous, D. (1981). *Exchangeability and related topics*. Springer Lecture notes in Math. no.1196.

Amari, S. (1982a). *Differential-geometrical methods in statistics*. Springer-Verlag, Berlin.

Amari, S. (1982b). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.*, **10**, 357-387.

Anderson, S. A. (1982). Distributions of maximal invariants using quotient measures. *Ann. Statist.*, **10**, 955-961.

Andrews, D. F., Arnold, J. C., and Krutchkoff, R. G. (1972). Shrinkage of the posterior measure in the normal case. *Biometrika*, **59**, 693-695.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normality. *J. R. Statist. Soc. B*, **36**, 99-102.

Antelman, G. R. (1965). Insensitivity to non-optimal design in Bayesian decision theory. *J. Amer. Statist. Ass.*, **60**, 584-601.

Antoniak, C. (1974). Mixtures of Dirichlet processes with application to Bayesian non-parametric problems. *Ann. Statist*, **2**, 1152-1174.

Atkinson, A. C. (1970). A method for discriminating between models. *J. R. Statist. Soc. B*, **32**, 323-353.

Atkinson, A. C. and Federov, V. V. (1975). The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57-70.

Bar-Lev, S. K. and Stramer, O. (1987). Characterisations of natural exponential families with power variance functions by zero regression properties. *Prob. Theory and related fields*, **76**, 509-523.

Barndorff-Nielsen, O. E. and Halgreen, C. (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Z. Wahr. verw. Geb.*, **38**, 309-311.

Barndorff-Nielsen, O. E. (1978). *Information and exponential families in Statistical theory*. Wiley.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343-365.

Barndorff-Nielsen, O. E. (1985). Properties of the modified profile likelihood. *Tech. Rep., Aarhus.*

Barndorff-Nielsen, O. E. and Jupp, P. E. (1988). Differential geometry, profile likelihood, L-sufficiency and composite transformation models. *Ann. Statist*, (To appear).

Barron, A. R. (1985). Entropy and the central limit theorem. *Ann. Prob.*, **14**, 336-342.

Bates, P. M. and Watts, D. G. (1980). Relative curvature measures of non-linearity. *J. R. Statist. Soc. B*, **40**, 1-25.

Berger, J. O. (1982). The robust Bayesian viewpoint. *Tech. Rep., Univ. of Purdue.*

Berk, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Ann. Statist.*, **37**, 51-58.

Bernardo, J. M. (1979a). Expected information as expected utility. *Ann. Statist.*, **7**, 686-690.

Bernardo, J. M. (1979b). Reference posterior distributions for Bayesian inference (with Discussion). *J. R. Statist. Soc. B*, **41**, 113-148.

Bernardo, J. M. (1980). A Bayesian analysis of classical hypothesis testing. In: *Bayesian Statistics I* (ed. by J. M. Bernardo *et al*), 605-618.

Bernardo, J. M. (1985a). Comment to Pettit and Smith. In: *Bayesian Statistics II* (ed. by Bernardo *et al.*), 492.

Bernardo, J. M. (1985b). A decision-theoretic approach to approximation in statistics. *Tech. Rep., Valencia.*

Bickel, P. J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.*, **9**, 1301-1309.

Bildikar, S. and Patil, G. P. (1968). Multivariate exponential-type distributions. *Ann. Math. Statist.*, **39**, 1316-1326.

Birge, L. (1980). In: *Approximation dans les espaces métriques et theorie de l'estimation; inegalités de Cramér-Chernoff et théorie asymptitique des tests.* Thesis, Univ. of Paris.

Blachman, N. M. (1965). The convolution inequality for entropy powers. *I.E.E.E. Trans. Inform. Theory*, **IT-11**, 267-271.

Blahut, R. (1972). *Computation of information measures.* Ph.D. thesis, Univ. of Cornell.

Bondar, J. V. (1977). A conditional confidence principle. *Ann. Statist.*, **5**, 881-891.

Bondar, J. V. and Milnes, P. (1981). Amenability: a survey for statistical applications of Hunt-Stein and related conditions on groups. *Z. Wahr. verw. Geb.*, **57**, 103-128.

Bondar, J. V. (1987). How much improvement can a shrinkage estimator give? In: *Foundations of Stat. Inf.* (ed. by I. MacNiell and G. Umphrey), 93-103.

Borth, D. M. (1975). A total entropy criterion for the dual problem of model discrimination and parameter estimation. *J. R. Statist. Soc. B*, **37**, 77-87.

Box, G. E. P. and Lucas, H. L. (1959). Design of experiments in nonlinear situations. *Biometrika*, **46**, 77-90.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc. B*, **26**, 211-243.

Box, G. E. P. and Kanemasu, H. (1973). Posterior probabilities of candidate models in model discrimination. *Tech. Rep. no. 322, Univ. of Wisconsin.*

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian inference in statistical analysis.* Reading, Mass. Addison-Wesley.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling (with Discussion). *J. R. Statist. Soc. A*, **143**, 383-430.

Brockett, P., Charnes, A., and Paick, K. (1983). Information-theoretic approach to unimodal density estimation. *Tech. Rep., Univ. of Texas.*

Brown, L. D. (1971). Non-local asymptotic optimality of appropriate likelihood ratio tests. *Ann. Math. Statist.*, **42**, 1206-1240.

Burbea, J. and Rao, C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. *J. Multi. Var. Analys.*, **12**, 575-596.

Burbea, J. and Rao, C. R. (1984). Differential metrics in probability spaces. *Prob. and Math. Stats.*, **3**, 241-258.

Burridge, J. (1987). In Discussion of Jørgensen. *J. R. Statist. Soc. B*, **49**, 159.

Campbell, L. L. (1985). The relation between information theory and the differential geometry approach to statistics. *Info. Sci.*, **35**, 199-210.

Campbell, L. L. (1986). An extended Cencov characterization of the information metric. *Amer. Math. Soc.*, **98**, 135-141.

Cencov, N. N. (1972). *Statistical decision rules and optimal inference (in Russian)*. Nauka, Moscow; translated in English AMS (1982).

Chaloner, K. (1984). Optimal Bayesian experimental design for linear models. *Ann. Statist.*, **12**, 283-300.

Chambers, E. A. and Cox, D. R. (1967). Discrimination between alternative binary response models. *Biometrika*, **54**, 573-578.

Chernoff, H. (1953). Locally optimum designs for estimating parameters. *Ann. Math. Statist.*, **24**, 586-602.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proc. 4th Berkeley Symp.*, **1**, 105-123.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc. B*, **24**, 406-424.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269-276.

Cox, D. R., Reid, N., and Barndorff-Nielsen, O. E. (1982). The role of differential geometry in statistical theory. *Int. Statist. Rev.*, **54**, 83-96.

Cox, D. R. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with Discussion). *J. R. Statist. Soc. B*, **49**, 1-40.

Crain, B. R. (1977). An information-theoretic approach to approximating a probability distribution. *Siam J. Appl. Math.*, **32**, 339-346.

Csiszár, I. (1967a). Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, **2**, 299-318.

Csiszár, I. (1967b). On topological properties of f-divergence. *Studia. Sci. Math. Hungar.*, **2**, 329-339.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Prob.*, **3**, 146-158.

Csiszár, I. (1977). Information measures: a critical survey. *Trans VIIth Prague conf. on info. theory*, 73-86.

Csiszár, I. (1984). Sanov property, generalised I-projection and a conditional limit theorem. *Ann. Prob.*, **12**, 768-793.

Csiszár, I. (1985). An extended maximum entropy principle. In: *Bayesian Statistics II* (ed. by Bernardo *et al*), 83-98.

Dalal, S. R. and Hall, G. (1980). On approximating parametric Bayes models by nonparameteric Bayes models. *Ann. Statist*, **8**, 664-672.

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian inference (with Discussion). *J. R. Statist. Soc. B*, **35**, 189-233.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, **60**, 664-666.

Dawid, A. P. (1975). On concepts of sufficiency and ancillarity in the presence of nuisance parameters. *J. R. Statist. Soc. B*, **37**, 248-258.

Dawid, A. P. (1977). Further comments on a paper by Bradley Efron. *Ann. Statist.*, **5**, 1249.

Dawid, A. P. (1982). Intersubjective statistical models. In: *Exchangeability in Probability and Statistics* (ed. by G. Koch and F. Spizzichino), 217-232.

Dawid, A. P. (1984). Present position and potential developments : Some personal views. *J. R. Statist. Soc. A*, **147**, 278-292.

Dawid, A. P. (1985). Probability, Symmetry and Frequency. *Brit. J. Phil. Sci.*, **36**, 107-128.

Dawid, A. P. (1986). A Bayesian view of statistical modelling. In: *Bayesian inference and decision techniques* (ed. by P. Goel and A. Zellner), 391-404.

DeFinetti, B. (1931). Funzioni caratteristica dinn fenomeno alentorio. *Atti della R. Academia Nazionale dei Lincei Ser 6, Memoie Classe di Scienze Fisiche, Mathematiche e Naturali*, **4**, 251-299.

DeFinetti, B. (1937). Foresight: its logical laws, its subjective sources. In: *Studies in subjective probability (1964)* (ed. by H. Kyburg and H. Smokler), 93-158.

DeFinetti, B. (1979). In Discussion of Bernardo (1979b). *J. R. Statist. Soc. B*, **41**, 135.

Dempster, A. P. (1975). A subjectivist look at robustness. *Tech. Rep. S-33, Harvard Univ.*

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.*, **7**, 269-281.

Diaconis, P. and Freedman, D. (1981). Partial exchangeability and sufficiency. In: *Statistics: Applications and new directions*, 205-236.

Diaconis, P. and Freedman, D. (1983). Frequency properties of Bayes' rules. In: *Scientific Inference, Data analysis and robustness* (ed. by G. Box, T. Leonard, C. Wu). Academic, New York.

Diaconis, P. and Ylvisaker, D. (1985). Quantifying prior opinion. In: *Bayesian Statistics 2* (ed. by Bernardo *et al*), 133-156.

Diaconis, P. and Freedman, D. (1986). A finite version of de Finetti's theorem for exponential families with uniform asymptotic estimates. *Tech. Rep., Univ. of California.*

Diaconis, P. and Freedman, D. (1986a). On the consistency of Bayes estimates. *Ann. Statist.*, **14**, 1-68.

Diaconis, P. and Freedman, D. (1986b). On inconsistent Bayes estimates of location. *Ann. Statist.*, **14**, 68-87.

Diaconis, P. and Freedman, D. (1987). A dozen de Finetti-style results in search of a theory. *Ann. Inst. Henri Poincaré*, **23**, 397-423.

Efron, B. and Morris, C. N. (1971). Limiting the risk of Bayes and empirical Bayes estimates. Part I—the Bayes case. *J. Amer. Statist. Ass.*, **66**, 807-815.

Efron, B. (1973). In Discussion of Dawid *et al. J. R. Statist. Soc. B*, **35**, 219.

Efron, B. (1975). Defining the curvature of a Statistical problem (with applications to second order efficiency). *Ann. Statist.*, **5**, 1189-1242.

Emerson, W. R. and Greenleaf, F. P. (1967). Covering properties and Følner conditions for locally compact groups. *Math. Zeitschr.*, **102**, 370-384.

Eplett, W. J. R. (1985). Influence functions and local minimax theory for Bayesian estimation. *Tech. Rep., Univ. of Oxford.*

Ericson, W. A. (1969). A note on the posterior mean of a population mean. *J. R. Statist. Soc. B*, **31**, 332-334.

Federov, V. V. (1972). In: *The theory of optimum experiments* (ed. by Studden and Klinko). Academic Press, New York..

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *Ann. Statist.*, **2**, 615-629.

Finucan, H. M. (1971). Posterior precision for non-normal distributions. *J. R. Statist. Soc. B*, **33**, 95-97.

Fraser, D. A. S. (1968). *The structure of inference*. Wiley, New York.

Fraser, D. A. S. (1976). Necessary analysis and adaptive inference. *J. Amer. Statist. Ass.*, **71**, 99-113.

Freedman, D. (1962). Invariants under mixing which generalise de Finetti's theorem. *Ann. Math. Statist.*, **33**, 916-923.

Freedman, D. and Diaconis, P. (1982a). de Finetti's theorem for symmetric location families. *Ann. Statist.*, **10**, 184-189.

Gauss, C. F. (1821). Göttingische gelehrte Anzeigen. In: *Werke Bd.*, 321-327.

Geisser, S. and Eddy, W. F. (1979). A predictive approach to model selection. *J. Amer. Statist. Ass.*, **74**, 153-160.

Ghosh, M. and Meeden, G. (1977). Admissibility of linear estimators in the one parameter exponential family. *Ann. Statist.*, **5**, 772-778.

Goel, P. K. and DeGroot, M. (1980). Only normal distributions have linear posterior expectations in linear regression. *J. Amer. Stat. Ass.*, **75**, 895-900.

Goel, P. K. and DeGroot, M. H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Ass.*, **76**, 140-147.

Goldstein, M. (1974). *Aspects of linear statistical inference*. D. Phil. Thesis, Univ. of Oxford.

Goldstein, M. (1975). Uniqueness relations for linear posterior expectations. *J. R. Statist. Soc. B*, **37**, 402-405.

Goldstein, M. (1977). On contractions of Bayes estimators for exponential family distributions. *Ann. Statist.*, **5**, 1235-1239.

Good, I. J. (1950). *Probability and the weighing of evidence*. Griffin, London.

Good, I. J. (1966a). A derivation of the probabilistic explication of information. *J. R. Statist. Soc. B*, **28**, 578-581.

Good, I. J. (1966b). How to estimate probabilities. *J. Inst. Maths. Applics.*, **2**, 364-383.

Good, I. J. (1968). Utility of a distribution. *Nature*, **219**, 1392.

Good, I. J. (1969). What is the use of a distribution? In: *Multivariate Analysis II* (ed. by P. R. Krishnaiah), 183-203.

Good, I. J. and Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika*, **58**, 255-277.

Good, I. J. (1984). Monotonic properties of the moments of a Bayes factor and the relationships with measures of divergence. *J. Stat. Comp.*, **19**, 320-325.

Haag, J. (1924). Sur une question de probabilités. *C. R. Acad. Sci. Paris*, **178**, 1140-1142.

Haitovsky, Y. and Zidek, J. V. (1986). Approximating hierarchical normal priors using a vague component. *J. Mult. Anal.*, **19**, 48-67.

Hernandez, F. and Johnson, R. A. (1979). Behaviour of transformations to normality. *J. Amer. Statist. Ass.*, **75**, 855-861.

Hewitt, E. and Savage, L. J. (1955). Symmetric measures on Cartesian products. *Trans. Amer. Math. Soc.*, **80**, 470-501.

Hildebrand, F. B. (1965). *Methods of applied mathematics*. Prentice-Hall.

Hill, P. D. H. (1980). D-optimal designs for partially nonlinear regression models. *Technometrics*, **22**, 275-276.

Hills, S. (1987a). In Discussion of Cox and Reid. *J. R. Statist Soc. B*, **49**, 23.

Hills, S. (1987b). Reference priors and identifiability problems in non-linear models. *The Statistician*, **36**, 235-240.

Hogg, R. V. (1972). More light on kurtosis and related statistics. *J. Amer. Statist. Ass.*, **67**, 422-424.

Hogg, R. V. (1974). Adaptive robust procedures. *J. Amer. Statist. Ass.*, **69**, 909-927.

Holland, P. (1973). Covariance stabilising transformations. *Ann. Statist.*, **1**, 84-92.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **36**, 1753-1758.

Huber, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.*, **43**, 1041-1067.

Huber, P. J. (1973). The use of Choquet capacities in statistics. *Bull. Inst. Internat. Statist.*, **45**, 181-191.

Huber, P. J. (1974). Fisher information and spline interpolation. *Ann. Statist.*, **2**, 1029-1034.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Hudson, H. M. (1978). A natural identity for exponential families with applications in multiparameter estimation. *Ann. Statist.*, **6**, 473-484.

Ibragimov, I. A. and H'asminsky, R. Z. (1973). On the information contained in a sample about a parameter. *2nd Int. Symp. on Info. Theory*, 295-309.

Ingarden, R. S. (1981). Information geometry in function spaces of classical and quantum finite statistical systems. *Intern. J. Engrg. Sci.*, **19**, 1609-1633.

Jørgensen, B. (1986). Some properties of exponential dispersion models. *Scand. J. Statist.*, **13**, 187-197.

Jørgensen, B. (1987). Exponential dispersion models. *J. R. Statist. Soc. B*, **49**, 127-163.

Jaynes, E. (1982a). Some applications and extensions of the de Finetti representation theorem. In: *Bayesian Inference and Decision Techniques with application* (ed. by P. Goel and A. Zellner).

Jaynes, E. (1982b). On the rationale of maximum entropy methods. *Proc. of I.E.E.E., Special Issue on Spectral Estimation*, **70**, 939-952.

Jeffreys, H. (1961). *Theory of Probability (3rd edition)*. Clarendon Press, Oxford.

Johnson, N. L. (1957). Uniqueness of a result in the theory of accident proneness. *Biometrika*, **44**, 530-531.

Johnson, N. L. (1967). Note on a uniqueness relation in certain accident proneness models. *J. Amer. Statist. Ass.*, **62**, 288-289.

Kagan, A. M., Linnik, Y. V., and Rao, C. R. (1973). *Characterization problems in mathematical statistics*. Wiley, New York.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Applications of likelihood methods to models involving large numbers of parameters (with Discussion). *J. R. Statist. Soc. B*, **32**, 175-208.

Kalbfleisch, J. D. and Sprott, D. A. (1972). Marginal and conditional likelihoods. *Sankhya*, **A**, 311-328.

Kanter, M. (1975). On stable densities under a change in scale and total variation inequalities. *Ann. Prob.*, **3**, 697-707.

Kass, R. E. (1979). The orbit integral representation of the density of a maximal invariant statistic. *Tech. Rep., Univ. of Chicago.*

Kass, R. E. (1981). The geometry of asymptotic inference. *Tech. Rep. no. 215, Carnegie-Mellon University, Pittsburgh.*

Kingman, J. F. C. (1978). In Discussion of Folks and Chhikara. *J. R. Statist. Soc. B*, **40**, 281.

Kraft, O. (1955). Some conditions for consistency and uniform consistency of statistical procedures. *Univ. of California Publications in Stats.*, **2**, 125-142.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Statist.*, **22**, 79-86.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Kullback, S. (1969). A lower bound for discrimination information in terms of variation. *I.E.E.E. Trans. Info. Theory*, **IT-13**, 126-127.

Larntz, K. and Chaloner, K. (1986). Optimal Bayesian design applied to logistic regression experiments. *Tech. Rep. no. 483, Univ. of Minnesota.*

LeCam, L. (1982). On the risk of Bayes estimates. In: *Statistical Decision Theory and Related Topics, III* (ed. by J. Berger and S. Gupta), 121-138.

LeJeune, M. and Faulkenberry, G. D. (1982). A simple predictive density function. *J. Amer. Statist. Assoc.*, **77**, 654-659.

Lindley, D. V. (1956). On the measure of information provided by an experiment. *Ann. Statist.*, **27**, 986-1005.

Lindley, D. V. (1961). On the use of prior probability distributions in statistical inference and decisions. *Proc. of IVth Berkeley Symp.*, **1**, 436-468.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model (with Discussion). *J. R. Statist. Soc. B*, **34**, 1-41.

Lindley, D. V. (1972). *Bayesian Statistics—A Review*. SIAM, Philadelphia.

Lindsay, B. G. (1980). Nuisance parameters, mixture models, and efficiency of partial likelihood estimators. *Philos. Trans. Roy. Soc. London Ser. A*, **296**, 639-662.

Loh, W. Y. (1984). Partially-adaptive robust estimators of location via exponential embedding. *Comm. Statist.*, **13**, 2549-2570.

Loh, W. Y. (1985). A note on the geometry of Kullback-Leibler information numbers. *Comm. Statist.*, **14**, 895-904.

Mallows, C. L. (1978). Minimising an integral. *Siam review*, **20**, 183.

Marazzi, A. (1980). Robust Bayesian estimation for the linear model. *Research Report no.27, Zürich.*

Masreliez, C. J. (1975). Approximate non-Gaussian filtering with linear state and observation relations. *I.E.E.E. Trans. Aut. Control*, **AC-20**, 107-110.

Matsubara, N. (1976). Bayes theorem, information numbers and behaviour of posterior distributions. *Ann. Inst. Stat. Math.*, 125-144.

Meeden, G. and Isaacson, D. (1977). Approximate behaviour of the posterior distribution for a large observation. *Ann. Statist.*, **5**, 899-908.

Mitchell, A. M. F. (1967). In Discussion of Good. *J. R. Statist. Soc. B*, **29**, 423.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *Ann. Statist.*, **10**, 65-80.

Neyman, J. and Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrika*, **16**, 1-32.

O'Hagan, A. (1979). On outlier rejection phenomena in Bayes inference. *J. R. Statist. Soc. B*, **41**, 358-367.

Osterreicher, F. (1972). An information type measure of the difference of probability distributions based on test. *Proc. of European stat., Budapest*, 593-600.

Patefield, W. (1977). On the maximised likelihood function. *Sankhya B*, **39**, 92-96.

Perrichi, L. R. (1984). An alternative to the standard Bayesian procedure for discrimination between noraml linear models. *Biometrika*, **71**, 575-586.

Pettit, L. and Smith, A. F. M. (1985). Outliers and influential observations in linear models. In: *Bayesian Statistics II* (ed. by Bernardo *et al.*), 473-494.

Pinsker, M. S. (1960). *Information and information stability of random variables and processes.* Moscow.

Polson, N. G. (1987). In Discussion of Cox and Reid. *J. R. Statist. Soc. B*, **49**, 24.

Rémon, M. (1984). On a concept of partial sufficiency : L-sufficiency. *Int. Statist. Rev.*, **52**, 127-136.

Rényi, A. (1961). On measures of entropy and information. *Proc. IVth Berkeley Symp.*, **1**, 547-561.

Rényi, A. (1964). On an extremal property of the Poisson process. *Ann. Inst. Stat. Math.*, **16**, 129-133.

Rényi, A. (1974). On the amount of missing information and the Neyman-Pearson lemma. *Research papers in Statistics*, 281-288.

Ralescu, D. and Ralescu, S. (1981). A class of nonlinear admissible estimators in the one-parameter exponential family. *Ann. Statist.*, **9**, 177-183.

Ramsey, J. O. and Novick, M. R. (1980). PLU robust Bayesian decision theory. *J. Amer. Statist. Ass.*, **75**, 901-907.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.*, **37**, 81-91.

Relles, P. A. and Rogers, W. H. (1977). Statisticians are fairly robust estimators of location. *J. Amer. Statist. Ass.*, **72**, 77-111.

Ressel, P. (1985). de Finetti type theorems: an analytic approach. *Ann. Prob.*, **13**, 898-922.

Rissanen, J. (1979). Shortest data description and consistency of order estimates in ARMA-processes. *Int. Symp. on systems optimisation and anal.*, 92-98.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, **11**, 416-431.

Rissanen, J. (1987). Stochastic complexity (with Discussion). *J. R. Statist. Soc. B*, **49**, 223-240.

Sacks, J. and Ylvisaker, D. (1972). A note on Huber's robust estimator of a location parameter. *Ann. Math. Statist*, **43**, 1068-1075.

Sampson, A. R. (1975). Characterizing exponential family distributions by moment generating functions. *Ann. Statist.*, **3**, 747-753.

SanMartini, A. and Spezzaferri, F. (1984). A predictive model selection criterion. *J. R. Statist. Soc. B*, **46**, 296-303.

Sanov, I. N. (1957). On the probability of large deviations of random variables. In: *Sel. Transl. Math. Statist. Prob. (1961)*, 213-244.

Savage, L. J. (1954). *The Foundations of Statistics*. Wiley, New York.

Schöenburg, I. J. (1938). Metric spaces and positive definite functions. *Trans. Amer. Math. Soc.*, **44**, 522-536.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461-464.

Seidler, J. (1959). Relationships between information theory and decision functions theory. In: *Trans. of 2nd Prague conf. on info. theory*, 579-592.

Shaked, M. (1980). On mixtures for exponential families. *J. R. Statist. Soc. B*, **42**, 192-199.

Silvey, S. D. (1980). *Optimal design*. Chapman and Hall, London and New York.

Simar, L. (1983). Protecting against gross errors: the aid of Bayesian methods. In: *Specifying statistical models* (ed. by J. Florens *et al*), 1-13.

Smith, A. F. M. and Verdinelli, I. (1980). A note on Bayes designs for inference using a hierarchical linear model. *Biometrika*, **67**, 613-619.

Smith, A. F. M. (1981). On random sequences with centred spherical symmetry. *J. R. Statist. Soc. B*, **43**, 208-209.

Smith, A. F. M. and Spiegelhalter, D. J. (1982). Bayes Factors for linear and log-linear models with vague prior information. *J. R. Statist. Soc. B*, **44**, 377-387.

Smith, A. F. M. (1983). Bayesian approaches to outliers and robustness. In: *Specifying statistical models* (ed. by J. Florens *et al*), 13-35.

Smith, A. F. M. (1984). Present position and potential developments: some personal views on Bayesian statistics. *J. R. Statist. Soc. A*, **147**, 245-259.

Smith, A. F. M. (1986). Some Bayesian thoughts on modelling and model choice. *The Statistician*, **35**, 97-102.

Spiegelhalter, D. J. (1981). Sampling properties of a finite mixture model. *Tech. Rep., University of Nottingham.*

Stone, M. (1959). Application of a measure of information to the design and comparison of regression experiments. *J. R. Statist. Soc. B*, **21**, 55-70.

Stone, M. (1963). Robustness of non ideal decision procedures. *J. Amer. Statist. Ass.*, **58**, 480-486.

Stone, M. and Springer, B. G. F. (1965). A paradox involving quasi prior distributions. *Biometrika*, **52**, 623-627.

Stone, M. and Dawid, A. P. (1972). Un-Bayesian implications of improper Bayes inference in routine statistical problems. *Biometrika*, **59**, 369-375.

Stone, M. (1976). Strong inconsistency from uniform priors (with Discussion). *J. Amer. Statist. Ass.*, **71**, 114-125.

Stone, M. (1979). A review and analysis of some inconsistencies related to improper priors and finite additivity. *Proc. of sixth International congress of logic, methodology and Philosophy of Science*, 413-426.

Sweeting, T. (1987). In Discussion of Cox and Reid. *J. R. Statist. Soc. B*, **49**, 20.

Takada, Y. (1979). Stein's positive part estimator and Bayes estimation. *Ann. Inst. Statist. Math.*, **31 A**, 177-183.

Tierney, L. and Kadane, J. B. (1986). Accurate appoximations for posterior moments and marginal densities. *J. Amer. Stat. Ass.*, **81**, 82-86.

Torgersen, E. N. (1976). Deviations for total information and for total ignorance as measures of information. *Tech. Rep., Univ. of Oslo.*

Torgersen, E. N. (1981). Measures of information based on comparison with total information and with total ignorance. *Ann. Statist.*, **9**, 638-657.

Toussaint, G. T. (1975). Sharper lower bounds for discrimination information in terms of variation. *I.E.E.E. Trans. Info. Theory*, **IT-21**, 99-100.

Tsai, C. L. (1983). *Contributions to the design and analysis of non-linear models*. Ph.D. Thesis, Univ. of Minnesota.

Tweedie, M. C. K. (1947). Functions of a statistical variate with given means, with special reference to Laplacian distributions. *Proc. Cambridge Phil. Soc.*, **49**, 41-49.

Vajda, I. (1970). Note on discrimination information and variation. *I.E.E.E. Trans. Info. Theory*, **IT-16**, 771-773.

Vajda, I. (1972). On the f-divergence and singularity of probability measures. *Per. Math. Hung.*, **2**, 223-234.

Volkonskij, V. A. and Rozanov, J. A. (1959). Some limit theorems for random functions—I. *Theory Prob. Appl.*, **4**, 178-197.

Wallace, C. S. and Freeman, P. R. (1987). Estimation and inference by compact coding (with Discussion). *J. R. Statist. Soc. B*, **49**, 240-257.

Weir, A. J. (1973). *Lebesgue integration and measure*. C.U.P., London.

West, M. (1984). *Aspects of recursive Bayesian estimation*. Ph.D. thesis, Univ. of Nottingham.

West, M. (1987). On scale mixtures of normality. *Biometrika*, **74**, 694-697.

Whitmore, G. A. and Yalovsky, M. (1978). A normalizing logarithmic transformation for inverse Gaussian random variables. *Technometrics*, **20**, 207-208.

Whittaker, E. T. and Watson, C. N. (1927). *A course of modern analysis*. CUP, Cambridge.

Whittle, P. (1973). Some general points in the theory of optimum experimental design. *J. R. Statist. Soc. B*, **35**, 123-130.

Wijsman, R. A. (1986). Global cross sections as a tool for factorisation of measures and distribution of maximal invariants. *Sankhya*, **48 A**, 1-42.

Wilson, E. B. and Hilferty, M. M. (1931). The distribution of chi-square. *Proceedings of the National Academy of Sciences (U.S.A.)*, **17**, 684-688.

Wolfowitz, J. and Hoeffding, W. (1958). Distinguishability of sets of distributions. *Ann. Math. Statist.*, **29**, 700-718.

Zacks, S. (1977). Problems and approaches in design of experiments for estimation and testing in non-linear problems. In: *Multivariate Analysis IV* (ed. by P. R. Krishnaiah), 209-233.