EXPONENTIAL FAMILIES AND THEORETICAL INFERENCE

Bent Jørgensen Rodrigo Labouriau

August, 2012

Contents

Preface

These notes are a translation of notes in Portuguese (Jørgensen and Labouriau, 1992). We are grateful to Sonia Mazzi, Victor Espinoza-Balderas, Xue-Kun Song, Norman Phillips and David Peterson for their help with the translation and editing of the manuscript. Also, thanks to Bertrand Clarke for some useful comments on Chapter 4.

The initial thrust of the translation of parts of the manuscript was done using our favourite spellchecker, which was a unique experience! There are probably still traces of this approach left in the text. As mentioned in the preface to the Portuguese edition, most of the material in Chapters 1 to 3 originate in Danish lecture notes from Aarhus University. Readers who, besides English, know these two "secret" languages (Danish and Portuguese) might have fun comparing the English and Portuguese versions of the notes with the original lecture notes from Aarhus.

Vancouver and Foulum, June 1995

Bent Jørgensen Rodrigo Labouriau

Preface to the Portuguese edition

Much recent statistical research has been dedicated to the study of specific models and techniques, leaving aside the development of general theory. We believe, however, that general theory, besides its interest in itself, is extremely useful, both for critical analysis of existing techniques, and for the development of new procedures. Thus, we attempt to bring to the fore, in these notes, the discussion of some general principles.

In the times of Pearson and Fisher, the debates around the philosophy and fundamental principles of statistics were quite fierce. With time the discussions became more technical, but still accentuated. The disputes continue today, even if the divergences are not always out in the open. We shall here expound an essentially Fisherian line of statistical thinking, whose origin we attempt to clear up in the following, without, however, pretending to give an account of the diverse exiting currents.

Sir Ronald Fisher is the precursor of many fundamental statistical concepts. ("*Hvad Fader gør er altid det Rigtige*", Andersen, 1866) The basic notions of likelihood, sufficiency, consistency and efficiency, were first defined by him. On the other hand, the notion of fiducial probability, one of the ideas that he defended vigorously, is today considered with doubt by many statisticians ("*Kejserens nye Klæder*", Andersen, 1866). Furthermore, the imprecise manner with which he often presented his ideas caused, in a certain way, some difficulty in the development of the Fisherian theory. Among the various books of Fisher, the only one that treats theoretical statistics as such (Fisher, l956) is more dedicated to attacking his predecessors and adversaries, than to explaining his ideas. The exposition there is incomplete, and often the arguments are based on examples. All these aspects left Fisher very exposed to critique, leaving to his successors the task of developing and extending his ideas.

Among the followers of the Fisherian thinking, the names of D.R. Cox and O.E. Barndorff-Nielsen are perhaps most prominent. A good exposition of the ideas of Cox can be encountered in the book Cox and Hinkley (1974). This book contains also a presentation of diverse other currents of statistical thinking as, for example, the Bayesian theory and the frequentist theory of Neyman and Pearson, even if the emphasis is essentially Fisherian. The book of Barndorff-Nielsen (1978) is probably the one that explains the more advanced development of the Fisherian ideas, principally with respect to the notions of sufficiency and ancillarity and the technique of inference in the presence of nuisance parameters ("inferential separation").

We treat in this book of some aspects of the theory of statistical inference, that are essentially derived from the Fisherian ideas developed by Cox and Barndorff-Nielsen. In Chapter 1, we expound the classical theory of exponential families of distributions, that is

one of the richer sources of statistical models, and which will serve as basis for constructing many examples in the following chapters. In Chapter 2 we present the classical theory of sufficiency and ancillarity, and also some relations of this theory with that of exponential families. Chapter 3 is dedicated to the theory of inferential separation, that is, we develop notions of sufficiency and ancillarity for models with nuisance parameters. Finally, in Chapter 4, we study the theory of inference functions (or estimating equations), including both the classical theory of optimality, and some notions of sufficiency and ancillarity adopted to this context.

A major part of this book was written during a curse of theoretical statistical inference, at the PhD level, given at IMPA in 1990. The two first chapters are based on notes written in the Department of Statistics of the University of Aarhus, where Professor Ole Barndorff-Nielsen works, except for Section 2.3, which was inspired by a paper of Lehmann (1981). Thus, Chapter 1 relies on notes written by Jørgen Granfeldt Pedersen, Preben Blæsild and Geert Schou, and Chapter 2 on notes of Preben Blæsild, Geert Schou and Eva Bjørn Jensen. We should emphasize that these notes of the University of Aarhus suffered various modifications by various authors during their use. Furthermore, they are in turn based on older notes written by Professor Ole Barndorff-Nielsen. In this way, it is difficult to allocate with precision the authorship of this material. Chapter 3 was written by Bent Jørgensen, inspired by notes of Preben Blæsild, Geert Schou, Eva Bjørn Jensen and Jens Ledet Jensen, including, however, some original formulations, such as the concept of I-nonformation. Chapter 4 was written by Rodrigo Labouriau, the basic reference being a sequence of papers by Godambe and the works of McLeish and Small, among others. We include some extensions of this work, making a reformulation of the theory of sufficiency and ancillarity of McLeish and Small.

The mathematical prerequisites for the reader of this book are a knowledge of probability theory and basic measure theory. It is also necessary to have some familiarity with elements of functional analysis for the last parts of Chapter 4.

We would like to thank Michael Sørensen and Jørgen Hoffmann-Jørgensen, for discussions regarding a preliminary version of Chapter 4, to Renée Xavier de Menezes for helping in the work of editing and revising and the Rogério Dias Trindade for the excellent typing of the manuscript.

Rio de Janeiro, June 1992

Bent Jørgensen Rodrigo S. Labouriau

Chapter 1

EXPONENTIAL FAMILIES

Exponential families are without any doubt among the most important statistical models, and include many classical examples. This concept, as well as many other basic notions of statistics, was introduced by Fisher. In honour of him and of some other precursors of the theory, exponential families of distributions are sometimes referred to as families of Fisher-Darmois-Koopman-Pitman type. As we will see, there already exists a well developed theory about these families, which we will study in this chapter. Moreover, we will see in Chapter 2 that there is a close relation between the fundamental concepts of sufficiency and the notion of exponential families.

1.1 Definitions

In this chapter, $(\mathcal{X}, \mathcal{A})$ will be a measurable space, and vectors will always be column vectors.

Definition 1.1 *A family $\mathcal{P}$ of probability measures in $(\mathcal{X}, \mathcal{A})$ is called an exponential family if there exists a σ-finite measure ν in $(\mathcal{X}, \mathcal{A})$, a positive integer k, functions $\alpha : \mathcal{P} \to \mathbb{R}^k$, $a : \mathcal{P} \to \mathbb{R}_+$, $t : \mathcal{X} \to \mathbb{R}^k$ and $b : \mathcal{X} \to \mathbb{R}_+ \cup \{0\}$, where b and t are measurable, such that the density function of $P \in \mathcal{P}$ with respect to ν has the form*

$$\frac{dP}{d\nu}(x) = a(P)b(x)e^{\alpha(P)\cdot t(x)}. \tag{1.1}$$

The notation "$\cdot$" represents the usual inner product in $\mathbb{R}^k$, that is,

$$\alpha(P) \cdot t(x) = \sum_{i=1}^{k} \alpha_i(P)t_i(x),$$

with $t_1, \ldots, t_k$ and $\alpha_1, \ldots, \alpha_k$ as the coordinate functions of t and α, respectively.

Note that if μ is the measure whose density function with respect to ν is the function b (*i.e.*, $(d\mu/d\nu)(x) = b(x)$) then the measures in $\mathcal{P}$ have density functions with respect to μ of

the form

$$\frac{dP}{d\mu}(x) = a(P)e^{\alpha(P)\cdot t(x)}, \tag{1.2}$$

i.e., we can absorb the function b in the dominating measure and obtain representation (1.2) from (1.1).

In the rest of this chapter, X will be a random variable on $\mathcal{X}$, whose distribution has density function with respect to μ given by (1.2). Under these conditions, $T = t(X)$ will be called the *canonical statistic* and $\alpha(P)$ the *canonical parameter*.

We will use the notation

$$X \sim EM(t(X), \alpha(P)) \quad \text{and} \quad \mathcal{P} \sim EM(t(X), \alpha(P)),$$

if the distributions of X have exponential representation with canonical statistic $t(X)$ and canonical parameter $\alpha(P)$. If it is necessary to specify that the exponential representation is with respect to the measure μ, we will write $\mathcal{P} \sim EM(t(X), \alpha(P), \mu)$. Note that t, α and μ determine the density function (1.2), since $\int dP/d\mu(x)d\mu(x) = 1$, so that $a(P) = [\int \exp\{\alpha(P) \cdot t(x)\}d\mu(x)]^{-1}$.

The smallest value of k for which $\mathcal{P}$ has a representation on the form (1.2) is called the *order* of the family, and is denoted $\text{ord}(\mathcal{P})$. Note that the order does not depend on the choice of the measure μ.

Definition 1.2 *The representations (1.1) or (1.2) are called minimal if the following conditions hold:*

i) The functions $1, t_1, \ldots, t_k$ are linearly independent with respect to μ;

ii) The functions $1, \alpha_1, \ldots, \alpha_k$ are linearly independent, where 1 is the constant function 1.

That is,

$$\sum_{i=1}^{k} a_i t_i = a_0 \ [\mu] \Rightarrow a_i = 0, \ \textit{for} \ \ i = 0, 1, \ldots, k \tag{1.3}$$

and

$$\sum_{i=1}^{k} b_i \alpha_i(P) = b_0, \ \forall P \in \mathcal{P} \Rightarrow b_i = 0, \ \textit{for} \ \ i = 0, 1, \ldots, k. \tag{1.4}$$

If the representations (1.1) or (1.2) are minimal, the statistic $t(X)$ will be called a *minimal canonical statistic*, and the parameter $\alpha(P)$ will be called a *minimal canonical parameter*. If the minimal canonical statistic is the identity function on $\mathcal{X}$ then $\mathcal{P}$ is called a *linear exponential family*.

Note that condition (1.4) is equivalent to saying that the set $\Omega = \{\alpha(P) : P \in \mathcal{P}\}$ is not contained in any affine subspace of dimension less than k. The same interpretation is valid for condition (1.3).

If representation (1.2) satisfies (1.3), then

$$P_1 = P_2 \Longleftrightarrow \alpha(P_1) = \alpha(P_2),$$

and, in this case $\mathcal{P}$ can be parametrized by Ω and by the function $\alpha^{-1} : \Omega \to \mathcal{P}$ (see Problem 1.2).

The following theorem states a relation between an arbitrary representation of the family and the minimal representation. Moreover, it shows that the minimal canonical statistic and the minimal canonical parameter are unique except for a non-singular affine transformation.

Theorem 1.3 *Let $\mathcal{P}$ be a family with representation (1.2) and with minimal representation*

$$\frac{dP}{d\mu}(x) = b(P)e^{\beta(P)\cdot u(x)} \tag{1.5}$$

of order $m \leq k$. Then for each $P_0 \in \mathcal{P}$ and $x_0 \in \mathcal{X}$ there exist matrices A and $\overline{A}$ each with dimension $m \times k$, such that:

$$\beta(P) - \beta(P_0) = A\{\alpha(P) - \alpha(P_0)\}, \quad \forall P \in \mathcal{P}$$

and

$$u(x) - u(x_0) = \overline{A}\{t(x) - t(x_0)\} \quad [\mu],$$

where $A\overline{A}^\top = \overline{A}A^\top = I_m$ (I_m is the identity matrix of order m).

Proof: From (1.2) and (1.5) it follows that for all $P, P_0 \in \mathcal{P}$ and $x, x_0 \in \mathcal{X}$

$$\{\beta(P) - \beta(P_0)\} \cdot \{u(x) - u(x_0)\} = \{\alpha(P) - \alpha(P_0)\} \cdot \{t(x) - t(x_0)\}, \ [\mu]. \tag{1.6}$$

Since (1.5) is a minimal representation, there exist $P_1, \ldots, P_m \in \mathcal{P}$ and $x_1, \ldots, x_m \in \mathcal{X}$ such that the m vectors

$$\beta(P_1) - \beta(P_0), \ldots, \beta(P_m) - \beta(P_0)$$

are linearly independent, as well as the vectors

$$u(x_1) - u(x_0), \ldots, u(x_m) - u(x_0).$$

Then the $m \times m$ matrices

$$B_0 = \{\beta_j(P_i) - \beta_j(P_0)\}_{i,j=1,\ldots,m}$$

and

$$U_0 = \{u_j(x_i) - u_j(x_0)\}_{i,j=1,\ldots,m}$$

are invertible. Analogously, we define the following $m \times k$ matrices :

$$A_0 = \{\alpha_j(P_i) - \alpha_j(P_0)\}_{\substack{i=1,\ldots,m \\ j=1,\ldots,k}}$$

and

$$T_0 = \{t_j(x_i) - t_j(x_0)\}_{\substack{i=1,\ldots,m \\ j=1,\ldots,k}}.$$

Using (1.6) for $P = P_i$, $i = 1, \ldots, m$ we obtain

$$B_0\{u(x) - u(x_0)\} = A_0\{t(x) - t(x_0)\} \quad [\mu]$$

and $\overline{A} = B_0^{-1}A_0$.

Using (1.6) for $x = x_i$, $i = 1, \ldots, m$ we have

$$U_0\{\beta(P) - \beta(P_0)\} = T_0\{\alpha(P) - \alpha(P_0)\},$$

which gives $A = U_0^{-1}T_0$.

Finally, using (1.6) for $P = P_i$, $i = 1, \ldots, m$ and $x = x_{i'}$, $i' = 1, \ldots, m$, we have $U_0B_0^T = T_0A_0^T$, which proves the last part of the theorem. □

Note that if the representation (1.1) is minimal, then the order of the family is k (see Problem 1.3).

We say that $\mathcal{P}$ has *kernel* if, for a minimal representation 1.1, $\operatorname{int}\alpha(\mathcal{P})$ is not empty.

In the following, we will define two fundamental concepts, namely, that of an exponential family generated by a canonical statistic $t(X)$ and a measure μ, and that of a full family. As we shall see, we almost always deal with full families generated by a canonical statistic $t(X)$ and a measure μ.

Definition 1.4 *Given a canonical statistic $t(X)$ and a measure μ, we define the family generated by $t(X)$ and μ, $\mathcal{P}(t, \mu)$, as the family of measures whose density function with respect to μ is of the form:*

$$a(\theta)e^{\theta \cdot t(x)}, \quad \theta \in \Theta, \tag{1.7}$$

where

$$1/a(\theta) = c(\theta) = \int_{\mathcal{X}} e^{\theta \cdot t(x)}\mu(dx)$$

and

$$\Theta = \{\theta \in I\!R^k : c(\theta) < \infty\}.$$

The parameter θ is called the *canonical parameter* and Θ the *domain of the canonical parameter*. Note that the domain of the canonical parameter is the largest possible, given the statistic $t(X)$ and the measure μ.

In the following, the probability measure with density function (1.7) will be represented by P_θ. The mean and the variance under P_θ will be represented by E_θ and Var_θ, respectively.

If $\mathcal{P}$ is the family with minimal canonical statistic $t(X)$ and $P \in \mathcal{P}$, then the family generated by $t(X)$ and P, $\mathcal{P}(t, P)$, does not depend on the choice of P or T. Moreover, we have that $\mathcal{P} \subseteq \mathcal{P}(t, P)$. In this way, we will refer to the family $\mathcal{P}(t, P)$ as the *exponential family generated by* $\mathcal{P}$, and we will write $\tilde{\mathcal{P}}$ instead of $\mathcal{P}(t, P)$. Note that $\operatorname{ord}\mathcal{P} = \operatorname{ord}\tilde{\mathcal{P}}$. If $\mathcal{P} = \tilde{\mathcal{P}}$, we say that $\mathcal{P}$ is *full*.

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a minimal canonical parametrization of $\mathcal{P}$. Then there exists a set $\tilde{\Theta} \supseteq \Theta$ such that $\tilde{\mathcal{P}} = \{P_\theta : \theta \in \tilde{\Theta}\}$ is a minimal canonical parametrization of $\tilde{\mathcal{P}}$. In this case, $\mathcal{P}$ is full if and only if $\Theta = \tilde{\Theta}$. A family $\mathcal{P}$ is called *regular* if it is full and Θ is open.

It follows immediately that if $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ has representation (1.7), not necessarily minimal, and $\Theta = \{\theta \in I\!R^k : c(\theta) < \infty\}$, then $\mathcal{P}$ is full. The reciprocal implication is also valid, if the representation is minimal.

The following concepts will be useful. A subset $A \subseteq I\!R^k$ is called *relatively open* if A is open as a subset of the smallest affine subspace in which it is contained. In this case, the *relative interior* of A, represented by ri A, is the interior of A when seen as a subset of the smallest affine space in which it is contained. The *relative boundary* of A is the set $A \setminus$ ri A.

The concept of a regular family can be generalized without mentioning the minimal representation in the following way: A full family, $\mathcal{P}$, is regular if $\{\theta : c(\theta) < \infty\}$ is relatively open. In order to avoid unnecessary complications, from now on we will always assume that Θ is open. In the rest of Section 1.1, we will assume that $1, t_1, \ldots, t_k$ are linearly independent with respect to μ, that is they satisfy (1.3).

The closure of the convex support of the measure $t(\mu)$ (*i.e.*, the measure μ transformed by the function t) plays an important role in exponential families. Let us remember that the *closed the convex support* of the measure $t(\mu)$, represented by $C_{t(\mu)}$, is the smallest closed convex set whose complement has null measure $t(\mu)$, that is,

$$C_{t(\mu)} = \bigcap_{K \in \psi} K,$$

where $\psi = \{K \subseteq I\!R^k : K \text{ is closed, convex and } \mu(t^{-1}(K^c)) = 0\}$. We represent $C_{t(\mu)}$ by C, eliminating T and μ from the notation, since $C_{t(\mu)}$ is the same for any choice of μ. The following theorem shows that C contains the mean of $t(X)$.

Theorem 1.5 *$E_\theta\{t(X)\} \in$ int C, $\forall\ \theta \in$ int Θ.*

In the following section we will show that the expectation $\mathrm{E}_\theta\{t(X)\}$ exists, for any θ in int Θ, so that the theorem above makes sense. To prove Theorem 1.5, we will need the following results:

Theorem 1.6 *(Separation Theorem) Let K be a closed convex set in $I\!R^k$ and $t_0 \notin K$. Then there exists a closed half-space (*i.e.*, a subset bounded by a hyperplane of dimension $k-1$), H_1, such that $K \subseteq H_1$ and $t_0 \notin H_1$.*

Let O be a relatively open convex set and $t_0 \notin O$. Then there exists an open half-space H_2, such that $O \subseteq H_2$ and $t_0 \notin H_2$.

Proof: See Rockafellar (1970, Section 11). Figure 1.1 illustrates the theorem. □

Lemma 1.7 *The closed convex support of $t(\mu)$ can be expressed in the following way:*

$$C = \bigcap_{H \in \psi} H, \tag{1.8}$$

where $\Psi = \{H : H$ is a closed half-space, with $\mu(t^{-1}(H^c)) = 0\}$.

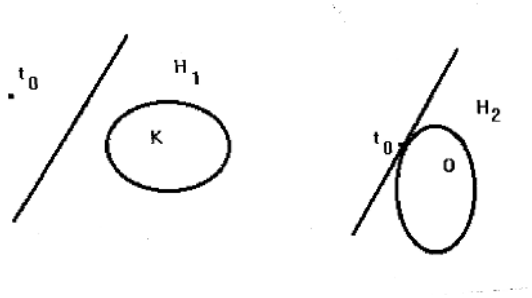

Figure 1.1: Proof of the Separation Theorem

Proof: Let C_1 be the set defined by the right hand side of (1.8). then, C_1 is closed and convex, since it is the intersection of closed and convex sets (half-spaces). Moreover, $\mu(t^{-1}(C_1^c)) = 0$ holds. Thus, $C_1 \supseteq C$.

We will show, by contradiction, that $C_1 \subseteq C$. Let $t_0 \in C_1$, $t_0 \notin C$. By the Separation Theorem there exists a closed half-space H, such that $C \subseteq H$ and $t_0 \notin H$. But $C \subseteq H$ implies $\mu(t^{-1}(H^c)) = 0$, hence $C_1 \subseteq H$, which contradicts the fact that $t_0 \notin H$. □

Proof: (of Theorem 1.5) We define $\tau(\theta) = \mathrm{E}_\theta(t(X))$, $\forall\theta \in \text{int } \Theta$. Given an arbitrary vector w we have that:

$$w \cdot t(X) \leq d \quad [\mu] \Rightarrow w \cdot \mathrm{E}_\theta[t(X)] \leq d.$$

In this way, any closed half-space that contains $t(X)$ $[\mu]$ will also contain $\tau(\theta) = \mathrm{E}_\theta[t(X)]$. Therefore, by Lemma 1.7 $\tau(\theta) \in C$.

We will show that $\tau(\theta) \notin \text{bd } C = C \setminus \text{int } C$, and the theorem will follow. Suppose that $\tau(\theta) \in \text{bd } C$. By the Separation Theorem, there exists a closed half-space H such that int $C \subseteq H$ and $\tau(\theta) \in \text{bd } H$ (note that int C is convex). The half-space H can be described in the following way:

$$H = \{s : w \cdot \{s - \tau(\theta)\} \geq 0\}.$$

Then

$$Z = w \cdot (t(X) - \tau(\theta)) \geq 0 \quad [\mu].$$

Note that $\mathrm{E}_\theta(Z) = 0$, which implies that

$$w \cdot (t(X) - \tau(\theta)) = 0 \quad [\mu],$$

which contradicts the fact that H "separates" int C from $\tau(\theta)$. To be precise, if $Z = 0$ almost surely, $1, t_1, \ldots, t_k$ cannot be linearly independent, contrary to what is assumed. We conclude that $\tau(\theta) \in \text{int } C$. □

In the case that the statistic T does not satisfy (1.3), it can be shown that C is contained in an affine subspace of dimension less than k and, in this case, Theorem 1.5 can be modified in the following way:

$$\mathrm{E}_\theta(t(X)) \in \text{ri } C, \quad \forall \theta \in \text{int } \Theta.$$

Example 1.8 *(The Binomial Distribution)* Let $\mathcal{X} = \{0, 1, \ldots, n\}$, with $n \in I\!N$ fixed. The probability of each point x in $\mathcal{X}$ is given by

$$\binom{n}{x} p^x (1-p)^{n-x} = (1-p)^n \exp\left\{x \log \frac{p}{1-p}\right\} \binom{n}{x}, \quad 0 < p < 1.$$

Putting $t(x) = x$, $\alpha(p) = \log \frac{p}{1-p}$, $a(p) = (1-p)^n$, $b(x) = \binom{n}{x}$ and ν as the counting measure, we see that the probability function above is of the form given by (1.1).

Let us consider now the full exponential family generated by $\mu(\{x\}) = \binom{n}{x}$ and $t(x) = x$. We have that

$$c(\theta) = \sum_{x=0}^{n} e^{\theta x} \binom{n}{x} = (1 + e^\theta)^n$$

and that

$$\Theta = \{\theta \in I\!R : c(\theta) < \infty\} = I\!R.$$

Writing $\theta = \log \frac{p}{1-p}$, we obtain

$$\frac{e^{\theta x}}{(1+e^\theta)^n} = \frac{e^{x \log p/(1-p)}}{(1 + p/(1-p))^n} = p^x (1-p)^{n-x}.$$

If p takes on values in the interval $(0, 1)$, θ will take on values in $I\!R$, which shows that a family of binomial distributions with fixed number of trials n and probability parameter p in $(0, 1)$, is a full exponential family. Since $\Theta = I\!R$ is open, the family is regular. The family is of order 1 and the closed convex support is $C = [0, n]$.

Example 1.9 *(The Gamma Distribution)* Let $\mathcal{X} = I\!R_+$ and $\mathcal{P}$ be the class of distributions with density function with respect to Lebesgue measure ν given by

$$\frac{1}{\Gamma(\lambda)\beta^\lambda} x^{\lambda-1} e^{-x/\beta} = \frac{1}{\Gamma(\lambda)\beta^\lambda} e^{\lambda \log x - \frac{x}{\beta}} \frac{1}{x},$$

where $\lambda > 0$ and $\beta > 0$. Define $t(x) = (x, \log x)^\top$, $\alpha(\lambda, \beta) = (-1/\beta, \lambda)^\top$, $a(\lambda, \beta) = \{\Gamma(\lambda)\beta^\lambda\}^{-1}$. Defining the measure μ as the one having density function $1/x$ with respect to ν, we see that $\mathcal{P}$ is an exponential family of the form (1.2).

Let us consider now the full family generated by μ and $t(x) = (x, \log x)^\top$. We have that

$$\begin{aligned} c(\theta) &= \int_0^\infty e^{\theta_1 x + \theta_2 \log x} (1/x) dx = \int_0^\infty x^{\theta_2 - 1} e^{\theta_1 x} dx \\ &= \begin{cases} \frac{\Gamma(\theta_2)}{|\theta_1|^{\theta_2}}, & \text{if } \theta_1 < 0 \text{ and } \theta_2 > 0 \\ \infty, & \text{otherwise .} \end{cases} \end{aligned}$$

Then $\Theta = \{(\theta_1, \theta_2)^\top \in I\!R^2 : \theta_1 < 0 \text{ and } \theta_2 > 0\}$. Since Θ is open, the full family generated by μ and T is regular.

Writing $\theta = (-1/\beta, \lambda)$, we see that $\mathcal{P}$ is an exponential family generated by μ and $t(x) = (x, \log x)^\top$. The order of the family is 2 and the closed convex support is

$$C = \{(t_1, t_2)^\top \in I\!R^2 : t_1 > 0 \text{ and } t_2 \leq \log t_1\}.$$

Example 1.10 *(The Multinomial Distribution)* Let us consider the family $\mathcal{P}$ of multinomial distributions with number of trials parameter n and parameter of probability contained in the set

$$\Pi = \{(p_1, \ldots, p_k)^T : p_j > 0, j = 1, \ldots, k,\ p_1 + \ldots + p_k = 1\}.$$

Here, $\mathcal{X}$ is the subset of $I\!N_0^k = [I\!N \cup \{0\}]^k$ defined by

$$\mathcal{X} = \{(x_1, \ldots, x_k)^T : x_j \geq 0, j = 1, \ldots, k, x_1 + \ldots + x_k = n\}.$$

Again, we can write the probabilities in the exponential form

$$\binom{n}{x_1 \ldots x_k} p_1^{x_1} \ldots p_k^{x_k} = \binom{n}{x_1 \ldots x_k} \exp\{x_1 \log p_1 + \ldots + x_k \log p_k\}. \tag{1.9}$$

We conclude that $\mathcal{P}$ is an exponential family by writing $t(x) = (x_1, \ldots, x_k)^\top$, $\alpha(p) = (\log p_1, \ldots, \log p_k)^\top$ and

$$\mu(\{x\}) = \binom{n}{x_1 \ldots x_k}.$$

The representation (1.8) is not minimal, since

$$x_1 + \cdots + x_k = n.$$

If we write $x_k = n - x_1 - \cdots - x_{k-1}$ in (1.9), we obtain

$$\begin{aligned}\binom{n}{x_1 \cdots x_k} &\exp\{x_1 \log p_1 + \cdots + x_{k-1} \log p_{k-1} + (n - x_1 - \cdots - x_{k-1}) \log p_k\} \\ &= \binom{n}{x_1 \ldots x_k} p_k^n \exp\left\{\textstyle\sum_{i=1}^{k-1} x_i \log \frac{p_i}{p_k}\right\}.\end{aligned}$$

Thus, we found a new representation of the exponential family (1.9) with

$$t'(x_1, \ldots, x_k) = (x_1, \ldots, x_{k-1})^\top$$

as canonical statistic and canonical parameter

$$\alpha'(p_1, \ldots, p_k) = \left(\log \frac{p_1}{p_k}, \ldots, \log \frac{p_{k-1}}{p_k}\right)^\top.$$

We also have $a(p_1,\ldots,p_k) = p_k^n = (1-p_1-\cdots-p_{k-1})^n$. It is easy to verify that this representation is minimal.

Considering the family generated by t' and μ we obtain in the same way as in the previous examples that

$$\begin{aligned} c(\theta) &= \sum_{x_1+\ldots+x_k=n} e^{\theta\cdot t'(x)} \binom{n}{x_1\ldots x_k} \\ &= (1+e^{\theta_1}+\cdots+e^{\theta_{k-1}})^n, \end{aligned}$$

showing that

$$\Theta = \{\theta \in I\!R^{k-1} : c(\theta) < \infty\} = I\!R^{k-1}.$$

Writing $\theta_i = \log p_i/(1-p_1-\cdots-p_{k-1})$, $i=1,\ldots,k-1$, we see that

$$\frac{1}{c(\theta)} e^{\theta\cdot t'(x)} = p_1^{x_1}\ldots p_{k-1}^{x_{k-1}}(1-p_1-\ldots-p_{k-1})^{n-x_1-\cdots-x_{k-1}}.$$

Since θ takes all values in $I\!R^{k-1}$ when $(p_1,\ldots,p_k)$ varies in Π, $\mathcal{P}$ is the full family generated by t' and μ. $\mathcal{P}$ is regular, and has order $k-1$. The closed convex support is $C = \{(t_1,\ldots,t_{k-1})^T : t_1 \geq 0,\ldots,t_{k-1}\geq 0, n-t_1-\cdots-t_{k-1}\geq 0\}$.

Example 1.11 Let $\mathcal{X} = S^{k-1}$ be the unit sphere of $I\!R^k$, and let $\mathcal{P} = \{P_{(\mu,\lambda)} \mid (\mu,\lambda) \in S^{k-1}\times[0,\infty)\}$ be the family of distributions of von Mises-Fisher in S^{k-1}, given by

$$\frac{dP_{(\mu,\lambda)}}{dP_0}(x) = a(\lambda)e^{\lambda\mu\cdot x},$$

where P_0 is the surface measure of S^{k-1} (Lebesgue measure). μ and λ are parameters that vary independently, $\mu \in S^{k-1}$ and $\lambda \geq 0$. These parameters are called the mean direction, and the concentration respectively.

The normalizing function $a(\lambda)$ depends on λ only, and can be expressed in the following way

$$a(\lambda) = \frac{\lambda^{\frac{k}{2}-1}}{(2\pi)^{\frac{k}{2}} I_{\frac{k}{2}-1}(\lambda)},$$

where I_ν is the modified Bessel function of first kind and order ν.

For $k=2$, we have that

$$a(\lambda) = \frac{1}{2\pi I_0(\lambda)},$$

which gives the so-called distribution of von Mises on the circle. For $k=3$, we have that

$$a(\lambda) = \frac{\lambda}{\sinh\lambda},$$

which gives the so-called Fisher distribution on the sphere.

The family $\mathcal{P}$ is exponential of order k, with minimal canonical parameter given by $\theta = \lambda\mu \in \Theta = I\!R^k$. Note that the mapping $(\mu,\lambda) \to P_{(\mu,\lambda)}$ is not a parametrization since $P_{(\mu,0)} = a(0)P_0$, $\forall\mu \in S^{k-1}$.

Let $\mathcal{P}$ be a full exponential family with minimal representation given by (1.7). The function $\mathcal{K} = \log c$ is said to be *steep* if for each $\theta \in \text{int } \Theta$ and each $\tilde{\theta} \in \Theta \setminus \text{int } \Theta = \text{bd } \Theta$ we have

$$(\tilde{\theta} - \theta) \cdot D\mathcal{K}\,[\,\alpha\theta + (1-\alpha)\tilde{\theta}\,] \stackrel{\alpha \to 0}{\longrightarrow} \infty$$

where $D\mathcal{K}(\cdot) = \frac{\partial \mathcal{K}}{\partial \theta}(\cdot)$. Evidently, either $\mathcal{K}$ is steep for any minimal representation of the family or $\mathcal{K}$ is not steep for any of them. That is, the property of being steep is intrinsic to the family $\mathcal{P}$, and hence we simply say that $\mathcal{P}$ is steep if $\mathcal{K}$ is steep.

Using results of the theory of convex functions one can show the following result.

Theorem 1.12 *If $\mathcal{P}$ is regular then $\mathcal{P}$ is steep.*

Proof: See Barndorff-Nielsen(1978, p. 117).

The converse of this theorem is not always valid, as the following example shows.

Example 1.13 *(The Inverse Gaussian Family)* Let $N^-(\mu, \lambda)$ be a family of continuous distributions with density function of the form

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} x^{-3/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right], \quad x > 0,$$

where $(\mu, \lambda) \in I\!R^2_+$. This family is called the Inverse Gaussian Family and will be represented in this example by $\mathcal{P}$.

We can reparametrize $\mathcal{P}$ defining $\psi = \lambda/\mu^2$ and writing the density function $f(x; \mu, \lambda)$ in the form

$$\frac{1}{\sqrt{2\pi}} x^{-3/2} \sqrt{\lambda} e^{\sqrt{\psi\lambda}} e^{-\psi x/2 - \lambda/(2x)}.$$

We see that $\mathcal{P}$ is not full. The full family generated by $\mathcal{P}$ includes also the case where $\psi = 0$, where we have the density function

$$(\frac{\lambda}{2\pi})^{1/2} x^{-3/2} e^{-\lambda/(2x)}$$

which is the density function of the stable distribution with stability index $1/2$ and scale parameter $1/\lambda$. We represent the full family by $\tilde{\mathcal{P}}$. The cumulant generating function, $\mathcal{K} = \log c$, is given by

$$\mathcal{K}(\psi, \lambda) = -1/2 \log \lambda - \sqrt{\psi\lambda}.$$

It is easy to show that $\tilde{\mathcal{P}}$ is steep but not regular. Note that the mean of these distributions is μ and λ is a kind of concentration parameter.

1.2 Analytical properties of the Laplace transform

In this section we will study some important properties of the function

$$c(\theta) = \int_{\mathcal{X}} e^{\theta \cdot t(x)} \mu(dx) = \int_{I\!R^k} e^{\theta \cdot z} t(\mu)(dz), \tag{1.10}$$

where $t(\mu)$ is the measure μ transformed by the measurable function $t : \mathcal{X} \to I\!R^k$ and $\Theta = \{\theta \in I\!R^k : c(\theta) < \infty\}$ (here we use the notation of Section 1.1). The last term of equation (1.10) shows that $c(\theta)$ is the Laplace Transform of the measure $t(\mu)$. Some elementary results for moment generating functions and characteristic functions, useful for the present section, may be found in Appendix A.

We define the *cumulant generating function* by

$$\mathcal{K}(\theta) = \log c(\theta).$$

Some properties of $c(\cdot)$ will be expressed in terms of $\mathcal{K}(\cdot)$. We recall the notation introduced in Section 1.1, where P_θ is a probability measure given by the density function with respect to μ

$$\frac{dP_\theta}{d\mu}(x) = \frac{1}{c(\theta)} e^{\theta \cdot t(x)} = e^{\theta \cdot t(x) - \mathcal{K}(\theta)}, \quad \forall \theta \in \Theta. \tag{1.11}$$

In this section we will not assume that the functions $1, t_1, \ldots, t_k$ are linearly independent with respect to μ, *i.e.*, we will not assume that the canonical statistic T is minimal. Meanwhile, we will point out the cases where the results here obtained can be improved on such an assumption.

Theorem 1.14 *The set $\Theta = \{\theta \in I\!R^k : c(\theta) < \infty\}$ is convex and the function $\mathcal{K}$ is strictly convex in Θ,* i.e.*,*

$$\mathcal{K}(\alpha\theta_1 + (1-\alpha)\theta_2) \leq \alpha\mathcal{K}(\theta_1) + (1-\alpha)\mathcal{K}(\theta_2), \tag{1.12}$$

for all $\theta_1, \theta_2 \in \Theta$ and $\alpha \in [0,1]$. In the case of equality in (1.12), for some $\alpha \in (0,1)$, then $P_{\theta_1} = P_{\theta_2}$.

Proof: Consider $\theta_1, \theta_2 \in \Theta$ and $\theta = \alpha\theta_1 + (1-\alpha)\theta_2$ for some $\alpha \in (0,1)$. By Hölder's inequality it follows that

$$\begin{aligned} c(\theta) &= \int_{\mathcal{X}} \{e^{\theta_1 \cdot t(x)}\}^\alpha \{e^{\theta_2 \cdot t(x)}\}^{(1-\alpha)} \mu(dx) \\ &\leq \left\{\int_{\mathcal{X}} e^{\theta_1 \cdot t(x)} \mu(dx)\right\}^\alpha \left\{\int_{\mathcal{X}} e^{\theta_2 \cdot t(x)} \mu(dx)\right\}^{(1-\alpha)} \\ &= c(\theta_1)^\alpha c(\theta_2)^{1-\alpha} < \infty. \end{aligned} \tag{1.13}$$

Then $\theta \in \Theta$ for all $\theta_1, \theta_2 \in \Theta$ and $\alpha \in [0,1]$ and hence Θ is convex.

Taking logarithms on both sides of inequality (1.13) we see that $\mathcal{K} = \log c$ is convex. Note that equality in Hölder's inequality holds if and only if

$$e^{\theta_1 \cdot t(x)} = k_0 e^{\theta_2 \cdot t(x)}, \quad [\mu] \tag{1.14}$$

for some constant k_0, which implies that $P_{\theta_1} = P_{\theta_2}$. □

Note that if $1, t_1, \ldots, t_k$ are linearly dependent with respect to μ, the family $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is not parametrized by $\theta \in \Theta$. More precisely, $\mathcal{P}$ is overparametrized by $\theta \in \Theta$ and θ is not identifiable. This is why Theorem 1.14 was formulated in such a careful way. On the other hand, when $1, t_1, \ldots, t_k$ are linearly independent we will obviously have that $P_{\theta_1} = P_{\theta_2}$ implies that $\theta_1 = \theta_2$.

Theorem 1.15 *The function c has Taylor series expansion for any $\theta \in \Theta$ with $\theta \mathcal{P} mh \in \Theta$,*

$$E_\theta [h \cdot t(X)]^n < \infty, \quad \forall n \in I\!N \tag{1.15}$$

and

$$c(\theta + h) = c(\theta) \sum_{n=0}^{\infty} \frac{1}{n!} E_\theta [h \cdot t(X)]^n. \tag{1.16}$$

Proof: Using the Taylor series expansion of the exponential function we have that

$$c(\theta + h) = \int_{\mathcal{X}} \sum_{n=0}^{\infty} \frac{[h \cdot t(x)]^n}{n!} e^{\theta \cdot t(x)} \mu(dx). \tag{1.17}$$

Note that we can interchange the integral and the summation in (1.17) since

$$\sum_{n=0}^{\infty} \frac{|h \cdot t(x)|^n}{n!} = e^{|h \cdot t(x)|} \leq e^{h \cdot t(x)} + e^{-h \cdot t(x)},$$

and $e^{h \cdot t(x)}$ as well as $e^{-h \cdot t(x)}$ are by hypothesis integrable with respect to $e^{\theta \cdot t(x)} \mu(dx)$, since $\theta + h$ and $\theta - h \in \Theta$. Then (1.15) follows. Interchanging the integral and the summation signs in (1.17) we obtain (1.16). □

In the rest of this section we will assume that Θ has non-empty interior, and in this way, we will be able to define the partial derivatives of $\mathcal{K}$. Let $\theta \in \text{int } \Theta$. Define

$$\tau(\theta) = \frac{\partial \log c}{\partial \theta}(\theta) = \frac{\partial \mathcal{K}}{\partial \theta}(\theta)$$

and

$$V(\theta) = \frac{\partial^2 \log c}{\partial \theta^2}(\theta) = \frac{\partial^2 \mathcal{K}}{\partial \theta^2}(\theta).$$

Note that $\tau(\theta)$ is the vector of dimension k, whose coordinates are $\tau_i(\theta) = \frac{\partial \mathcal{K}}{\partial \theta_i}(\theta)$, $i = 1, \ldots, k$ and $V(\theta)$ is the matrix $k \times k$ with entries

$$V_{ij}(\theta) = \frac{\partial^2 \mathcal{K}}{\partial \theta_i \partial \theta_j}(\theta), \quad i = 1, \ldots, k, \quad j = 1, \ldots, k.$$

Theorem 1.16 *For $\theta \in int\ \Theta$ we have*

$$\frac{\partial^n c(\theta_1, \ldots, \theta_k)}{\partial \theta_1^{a_1} \cdot \ldots \cdot \partial \theta_k^{a_k}} = c(\theta) E_\theta [t_1(X)^{a_1} \cdot \ldots \cdot t_k(X)^{a_k}], \tag{1.18}$$

where $a_1 + \ldots + a_k = n$. Moreover,

$$\tau(\theta) = E_\theta \{t(X)\} \tag{1.19}$$

and

$$V(\theta) = Var_\theta \{t(X)\}. \tag{1.20}$$

Proof: Using the multinomial expansion we obtain

$$\mathrm{E}_\theta [h \cdot t(X)]^n = \sum_{a_1 + \ldots + a_k = n} \left(\begin{array}{c} n \\ a_1 \ldots a_k \end{array} \right) \mathrm{E}_\theta \prod_{i=1}^{k} h_i^{a_i} t_i(X)^{a_i}.$$

Substituting the expression above in (1.16) it can be seen that (1.18) holds. From (1.18), (1.19) and (1.20) the theorem follows. □

Theorem 1.17 *The mapping $\tau : int\ \Theta \to I\!R^k$ has the following properties:*

(i) τ is strictly increasing, in the sense that for θ_1 and $\theta_2 \in int\ \Theta$

$$(\theta_1 - \theta_2) \cdot \{\tau(\theta_1) - \tau(\theta_2)\} \geq 0$$

holds, with equality if and only if $P_{\theta_1} = P_{\theta_2}$;

(ii) τ is injective in the sense that $\tau(\theta_1) = \tau(\theta_2) \Rightarrow P_{\theta_1} = P_{\theta_2}$;

(iii) τ is differentiable and $\frac{\partial \tau}{\partial \theta}(\theta) = V(\theta)$;

(iv) The matrix $V(\theta)$ is symmetric and positive definite, in the sense that $h^\top V(\theta) h = 0 \Rightarrow P_\theta = P_{\theta + h}$.

Proof: Let $\theta_1 \neq \theta_2 \in$ int Θ be fixed and $f(z) = \mathcal{K}(\theta_2 + z(\theta_1 - \theta_2))$. Then $f(0) = \mathcal{K}(\theta_2)$, $f(1) = \mathcal{K}(\theta_1)$, and since f is convex $\frac{\partial f}{\partial z}(0) \leq \frac{\partial f}{\partial z}(1)$, we obtain

$$(\theta_1 - \theta_2) \cdot \tau(\theta_2) \leq (\theta_1 - \theta_2) \cdot \tau(\theta_1)$$

with equality if f is linear in $[0, 1]$, which implies by Theorem 1.14 that $P_{\theta_1} = P_{\theta_2}$. This proves (i).

Item (ii) follows immediately from (i), since $\tau(\theta_1) = \tau(\theta_2)$ implies that $(\theta_1 - \theta_2) \cdot [\tau(\theta_1) - \tau(\theta_2)] = 0$.

The differentiability of τ follows from Theorem 1.16, which shows (iii).

Item (iv) can be shown in the following way. If $h^\top V(\theta)h = 0$, then $\mathrm{Var}_\theta[t(X)\cdot h] = 0$. Therefore $h\cdot t(x) = \mathrm{E}_\theta\ h\cdot t(X) = h\mathrm{E}_\theta\ t(X) = h\cdot\tau(\theta)\ [P_\theta]$. Then

$$c(\theta+h) = \int_{\mathcal{X}} e^{(\theta+h)\cdot t(x)}\mu(dx) = c(\theta)e^{h\cdot\tau(\theta)},$$

and

$$\frac{dP_{\theta+h}}{d\mu}(x) = \frac{1}{c(\theta+h)}e^{(\theta+h)\cdot t(x)} = \frac{1}{c(\theta)}e^{\theta\cdot t(x)} = \frac{dP_\theta}{d\mu}(x)\ [\mu]$$

implying that $P_{\theta+h} = P_\theta$. □

If $1, t_1, \ldots, t_k$ are linearly independent with respect to μ, then we can improve items (i), (ii) and (iv) in the previous theorem since, in this case, $(\theta_1-\theta_2)\cdot\{\tau(\theta_1)-\tau(\theta_2)\} = 0$ implies that $\theta_1 = \theta_2$, τ is injective and $V(\theta)$ is positive definite, in the usual sense.

1.3 Estimation in regular exponential families

Consider a regular exponential family generated by the statistic T, with linearly independent functions $1, t_1, \ldots, t_k$ with respect to μ. Since the family is regular, we have that Θ is open.

Let X be the random variable that represents an observation from a distribution of the family and let x be the observed value of X. The likelihood function of θ is given by

$$L(\theta) = \frac{e^{\theta\cdot t(x)}}{c(\theta)}, \quad \theta\in\Theta, \tag{1.21}$$

and the log-likelihood function is

$$\ell(\theta) = \log L(\theta) = -\log c(\theta) + \theta\cdot t(x) = -\mathcal{K}(\theta) + \theta\cdot t(x). \tag{1.22}$$

Since $\mathcal{K}$ is a strictly convex function in Θ then ℓ is strictly concave and hence ℓ cannot have more than one maximum in Θ.

If ℓ has a maximum, it can be found by differentiating ℓ. We then obtain the equation

$$-\tau(\theta) + t(x) = 0.$$

That is,

$$\mathrm{E}_\theta[t(X)] = t(x). \tag{1.23}$$

Equation (1.23) is the standard form of the equation of maximum likelihood for exponential families. In order to simplify the notation we will write T instead of $t(x)$.

By the discussion above, we conclude that there exists a maximum of ℓ if and only if Equation (1.23) has a solution, that is if $t\in\Omega = \tau\{\Theta\}$. The following theorem tells us that $\Omega = \mathrm{int}\ C$ for a regular family.

Theorem 1.18 *(Barndorff-Nielsen) If $\mathcal{P}$ is a regular exponential family with minimal representation, the maximum likelihood estimator exists if and only if $t \in \text{int } C$. If the maximum likelihood estimator exists, it is given by*

$$\hat{\theta} = \tau^{-1}(t).$$

The same conclusion of the theorem above is valid for $\theta \in \text{int } \Theta$ if the family is steep. The proof of this fact uses arguments based on the theory of convex functions and can be found in Barndorff-Nielsen (1978).

Proof: (of Theorem 1.18) Let $t \in \text{int } C$. We will show that for any $\theta_0 \in \Theta$ and any half-line

$$\{\theta_\alpha = \theta_0 + \alpha(\theta - \theta_0) : \alpha \geq 0\},$$

we have that $\ell(\theta_\alpha) \stackrel{\alpha\to\infty}{\longrightarrow} -\infty$. Since the set $\{\theta : \ell(\theta) \geq \ell(\theta_0)\}$ is convex and a convex set that does not contain any unbounded rays is compact, the global maximum of ℓ belongs to $\{\theta : \ell(\theta) \geq \ell(\theta_0)\}$. The strict convexity of $\ell(\theta)$ shows that the maximum is unique. The reciprocal of the likelihood function on the half-line is

$$e^{-\ell(\theta_\alpha)} = \int_{\mathcal{X}} e^{\alpha(\theta-\theta_0)\cdot(t(x)-t)} e^{\theta_0\cdot(t(x)-t)} \mu(dx).$$

We will divide the domain of integration $\mathcal{X}$ of the integral above in three disjoint subsets

$$\begin{aligned} A^+ &= \{x : (\theta - \theta_0) \cdot (t(x) - t) > 0\} \\ A^0 &= \{x : (\theta - \theta_0) \cdot (t(x) - t) = 0\} \\ A^- &= \{x : (\theta - \theta_0) \cdot (t(x) - t) < 0\}. \end{aligned}$$

Note that the integrand $e^{\alpha(\theta-\theta_0)\cdot(t(x)-t)}$ as a function of α is increasing in A^+, constant in A^0 and decreasing in A^-. Since the integrand is positive, the integral on A^- is bounded as a function of α, and hence, by the monotone convergence theorem, $\ell(\theta_\alpha) \to \ell(\theta_{\alpha_0})$ when $\alpha \nearrow \alpha_0$. We will divide the proof in two cases

- Θ is bounded in the direction given by $\theta - \theta_0$ and
- Θ is not bounded in this direction.

Let us consider the first case, *i.e.*, Θ is bounded in the direction given by $\theta - \theta_0$. Let θ_{α_0} be a point in the boundary. In this point we have $c(\theta_{\alpha_0}) = \infty$, and therefore $\ell(\theta_\alpha) \to -\infty$ when $\alpha \nearrow \alpha_0$.

In the case that Θ is not bounded in the direction determined by $\theta - \theta_0$ we will need the condition $t \in \text{int } C$.

If t is an interior point of C, the hyperplane $\{u : (\theta - \theta_0) \cdot (u - t) = 0\}$ divides C through t in two parts, both with positive measure. Specifically

$$\mu(A^+) = \mu(\{x : (\theta - \theta_0) \cdot (t(x) - t) > 0\}) > 0,$$

and, hence,

$$e^{-\ell(\theta_\alpha)} \overset{\alpha\to\infty}{\longrightarrow} \int_{A^0} e^{\theta_0\cdot(t(x)-t)}\mu(dx) + \int_{A^+} \lim_{\alpha\to\infty} e^{\alpha(\theta-\theta_0)\cdot(t(x)-t)} e^{\theta_0\cdot(t(x)-t)}\mu(dx) = \infty,$$

showing that $\ell(\theta_\alpha) \to -\infty$ when $\alpha \to \infty$. Here, we use the fact that the integrand converges monotonically to zero in A^-, implying that the integral on A^- converges to zero. In this way, we have shown that $\hat{\theta}$ exists if $t \in \text{int } C$.

In the case that $t \notin \text{int } C$, there exists a hyperplane $\{u : w \cdot (u - t) = 0\}$ through t, such that

$$\mu(\{x : w \cdot (t(x) - t) > 0\}) = 0.$$

Considering $\ell(\theta_0 + \alpha w)$, we see that ℓ is increasing in α. Since θ_0 is arbitrary, it is evident that the likelihood function does not assume its maximum value in Θ. □

The proof was given by Johansen (1979).

If we have a regular family such that $1, t_1, \ldots, t_k$ are not linearly independent, we can modify Theorem 1.18 in the following way.

Theorem 1.19 *In a regular exponential family the maximum likelihood estimator $\hat{\theta}$ exists if and only if $t \in riC$. The estimator can be found by obtaining a solution of the maximum likelihood equation (1.23). The solution is unique, in the sense that all the solutions represent the same probability measure.*

In the following we present a theorem on maximum likelihood estimation in a full family, but not necessarily steep. The theorem was proved by Ole Barndorff-Nielsen, using the theory of convex functions.

Theorem 1.20 *Let $\mathcal{P}$ be a full exponential family with minimal representation (1.7). In this case, $\hat{\theta}$ exists if and only if $t(x) \in int\ C$. If $t(x) \in \Omega = \tau(int\ \Theta)$, then $\hat{\theta}$ is a unique solution of the maximum likelihood equation*

$$\tau(\theta) = t(x), \quad \theta \in int\ \Theta.$$

If $t(x) \notin int\ C$, the likelihood function does not attain its maximum.

Note that if $t(x) \in \text{int } C \setminus \Omega$, the theorem does not show how to find $\hat{\theta}$, but obviously $\hat{\theta} \in \text{bd } \Theta$.

We conclude this section with some considerations on estimation based on a sample from an exponential family $\mathcal{P}$. Let $X_1, \ldots, X_n$ independent and identically distributed with distribution in $\mathcal{P}$, where $\mathcal{P}$ has a minimal representation (1.7). The joint density function of $X_1, \ldots, X_n$ is

$$\frac{dP_\theta^{\otimes n}}{d\mu^{\otimes n}}(x) = a(\theta)^n \exp\left\{\theta \cdot \sum_{i=1}^n t(x_i)\right\}.$$

Then $(X_1, \ldots, X_n) \sim EM(\sum_{i=1}^n t(X_i), \theta)$. We represent the family of distributions of $X_1, \ldots, X_n$ by $\mathcal{P}_n$. Since $1, t_1(X), \ldots, t_k(X)$ are linearly independent, the same is valid for $1, \sum_{i=1}^n t_1(X_i), \ldots, \sum_{i=1}^n t_k(X_i)$. It is also clear that, if $\mathcal{P}$ is full, with minimal canonical domain Θ, then $\mathcal{P}_n$ is also full, with the same minimal canonical domain.

If $\mathcal{P}$ is steep then $\mathcal{P}_n$ is also steep. Let $\mathcal{K}_n$ and τ_n be the cumulant generating functions with respect to $\mu^{\otimes n}$. Then

$$\mathcal{K}_n(\theta) = n\mathcal{K}(\theta) \tag{1.24}$$

$$\tau_n(\theta) = n\tau(\theta). \tag{1.25}$$

The maximum likelihood equation is

$$\tau_n(\theta) = \sum_{i=1}^n t(x_i),$$

or $\tau(\theta) = \bar{t}$, where

$$\bar{t}(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^n t(x_i).$$

That is, the likelihood equation has the same form in $\mathcal{P}_n$ as for $\mathcal{P}$. Note that the equation has a unique solution in int Θ if and only if $\bar{t} \in$ int C.

1.4 Marginal and conditional distributions

Let us consider the full exponential family given by the density function

$$\frac{dP_\theta}{d\mu}(x) = a(\theta) e^{\theta \cdot t(x)} \quad [\mu],$$

where θ and t are k-dimensional. Let $V = At(X)$ be a linear function of the canonical statistic. We will show that the family of conditional distributions of V is an exponential family and that, under certain restrictions on the canonical parametric space, the family of marginal distributions of the statistic V is also an exponential family.

Let A be an $m \times k$ matrix of full rank m ($m < k$). We choose a $(k-m) \times k$ matrix B of rank $k-m$ such that $AB^\top = 0$. Then the rows of A generate the null space of B, and then $B\theta = B\theta_0 \Leftrightarrow \exists \eta \in I\!R^m$ such that $\theta = \theta_0 + A^\top \eta$. In particular,

$$\{\theta \in I\!R^k : B\theta = B\theta_0\} = \{\theta_0 + A^\top \eta : \eta \in I\!R^m\}.$$

Let $\Theta_0 = \{\theta_0 + A^\top \eta : \eta \in I\!R^m\} \cap \Theta$ and $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta_0\}$.

The hypothesis $\mathcal{P}_0$ is called an *affine hypothesis* for the canonical parameter. The family of marginal distributions of V under $\mathcal{P}$ is represented by

$$V(\mathcal{P}) = \{V(P_\theta) : P_\theta \in \mathcal{P}\}$$

and under $\mathcal{P}_0$ by

$$V(\mathcal{P}_0) = \{V(P_{\theta_0}) : P_{\theta_0} \in \mathcal{P}_0\}.$$

In order to find the distribution of V, we write the density function of P_θ with respect to P_{θ_0},

$$\begin{aligned}
\frac{dP_\theta}{dP_{\theta_0}}(x) &= \frac{a(\theta)}{a(\theta_0)} e^{(\theta-\theta_0)\cdot t(x)} \\
&= \frac{a(\theta)}{a(\theta_0)} e^{(\theta-\theta_0)^\top \begin{pmatrix} A \\ B \end{pmatrix}^{-1} \begin{pmatrix} A \\ B \end{pmatrix} t(x)} \\
&= \frac{a(\theta)}{a(\theta_0)} e^{\{(A^\top : B^\top)^{-1}(\theta-\theta_0)\}^\top \begin{pmatrix} A \\ B \end{pmatrix} t(x)} \\
&= \frac{a(\theta)}{a(\theta_0)} e^{(\phi_1^\top, \phi_2^\top)^\top \cdot (v(x)^\top, u(x)^\top)^\top},
\end{aligned}$$

where $v(x) = At(x)$, $u(x) = Bt(x)$ and $(\phi_1^\top, \phi_2^\top) = (A^\top : B^\top)^{-1}(\theta - \theta_0)$, and ϕ_1 has dimension m. Therefore, the marginal distribution of $V = At(X)$ has density function

$$\begin{aligned}
\frac{dV(P_\theta)}{dV(P_{\theta_0})}(v) &= \mathrm{E}_{\theta_0}\left\{ \frac{a(\theta)}{a(\theta_0)} e^{(\phi_1^\top, \phi_2^\top)^\top \cdot (v(x)^\top, u(x)^\top)^\top} \mid V = v \right\} \\
&= \frac{a(\theta)}{a(\theta_0)} e^{\phi_1 \cdot v} \mathrm{E}_{\theta_0}\left\{ e^{\phi_2 \cdot u(x)} \mid V = v \right\}.
\end{aligned}$$

$V(\mathcal{P})$ is an exponential family only if ϕ_2 is a constant as a function of $\theta - \theta_0$. If θ is of the form $\theta = \theta_0 + A^\top \eta$, we have

$$\begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}^\top = (A^\top : B^\top)^{-1}(\theta - \theta_0) = (A^\top : B^\top)^{-1} A^\top \eta = \begin{pmatrix} \eta \\ 0 \end{pmatrix}^\top.$$

For $\theta \in \Theta_0$, we have

$$\frac{dV(P_\theta)}{dV(P_{\theta_0})}(v) = \frac{a(\theta)}{a(\theta_0)} e^{\eta \cdot v},$$

showing that $V(\mathcal{P}_0)$ is an exponential family.

If $\mathcal{P}$ is full then $V(\mathcal{P}_0)$ is also full, since

$$\begin{aligned}
H &= \left\{ \eta : \int_{I\!R^m} e^{\eta \cdot v} V(P_{\theta_0})(dv) < \infty \right\} \\
&= \left\{ \eta : \int_{\mathcal{X}} e^{\eta \cdot v(x)} P_{\theta_0}(dx) < \infty \right\} \\
&= \left\{ \eta : \int_{\mathcal{X}} e^{(\theta_0 + A^\top \eta) \cdot t(x)} \mu(dx) < \infty \right\} \\
&= \{\eta : c(\theta_0 + A^\top \eta) < \infty\} \\
&= \{\eta : \theta_0 + A^\top \eta \in \Theta\}.
\end{aligned}$$

We also have that H is open if Θ is open, since H is the inverse image of Θ by a continuous transformation. In this way, we have just proved the following result.

Theorem 1.21 *The family of marginal distributions $V(\mathcal{P}_0)$ of $V = At(X)$ is a full and linear family, with canonical parametric space $H = \{\eta : \theta_0 + A^\top \eta \in \Theta\}$. If $\mathcal{P}$ is regular, then $V(\mathcal{P}_0)$ is regular.*

This was the treatment for the case of the marginal distribution of V. Sometimes the interest is concentrated in the affine hypothesis $\mathcal{P}_0$ of $\mathcal{P}$. The proof of Theorem 1.21 shows that if $\mathcal{P}$ is regular, then $\mathcal{P}_0$ is a regular exponential family with canonical statistic V, which is minimal if $t(X)$ is minimal. In Problem 1.13 the likelihood ratio test of $\mathcal{P}_0$ under $\mathcal{P}$ is derived.

Theorem 1.22 *The conditional distribution of X given $V = At(X) = v$ is an exponential family, whose parameter depends on θ only through $B\theta$.*

Proof: The conditional distribution of X given $V = v$ with respect to P_θ will be represented by $P_\theta(\,\cdot\,|V = v)$. The density function of $P_\theta(\,\cdot\,|V = v)$ with respect to $P_{\theta_0}(\,\cdot\,|V = v)$ is

$$\begin{aligned}
\frac{dP_\theta(\,\cdot\,|V=v)}{dP_{\theta_0}(\,\cdot\,|V=v)}(x) &= \frac{\frac{dP_\theta}{dP_{\theta_0}}(x)}{\mathrm{E}_{\theta_0}\{\frac{dP_\theta}{dP_{\theta_0}} \mid V = v\}} \\
&= \frac{e^{(\theta-\theta_0)\cdot t(x)}}{\mathrm{E}_{\theta_0}\{e^{(\theta-\theta_0)\cdot t(x)} \mid V = v\}} \\
&= \frac{e^{(\theta-\theta_0)\,\cdot\, t(x)}}{\int_{\mathcal{X}} e^{(\theta-\theta_0)\,\cdot\, t(x)} P_{\theta_0}(dx|V=v)}, \qquad (1.26)
\end{aligned}$$

which is an exponential representation.

We will now show that, if $B\theta_1 = B\theta_2$, then $P_{\theta_1}(\,\cdot\,|V = v) = P_{\theta_2}(\,\cdot\,|V = v)$. If $B\theta_1 = B\theta_2$, then $\theta_1 - \theta_2 = A^\top \eta$ for some $\eta \in I\!R^m$. Using (1.26) with $\theta = \theta_1$ and $\theta_0 = \theta_2$, we have

$$\frac{dP_{\theta_1}(\,\cdot\,|V=v)}{dP_{\theta_2}(\,\cdot\,|V=v)}(x) = \frac{e^{\eta\,\cdot\,At(x)}}{\mathrm{E}_{\theta_2}\{e^{\eta\,\cdot\,At(x)}|V=v\}} = 1$$

$[P_{\theta_2}(\,\cdot\,|V = v)]$, since $At(x) = v$ $[P_{\theta_2}(\,\cdot\,|V = v)]$. □

In the proof of the first part of Theorem 1.22 we did not use the fact that V is a linear function of $t(X)$, and hence we have that *the conditional distribution of any function of a canonical statistic is an exponential family*.

Generally, the exponential family of conditional distributions given a function V of $t(X)$ of the full exponential family is not necessarily full, since the integral

$$\mathrm{E}_{\theta_0}[e^{(\theta-\theta_0)\cdot t(x)}|V = v]$$

can be finite in a larger set than Θ.

Let $\theta = (\theta^{(1)}, \theta^{(2)})$ and $t = (t^{(1)}, t^{(2)})$ be a partition of θ and t, such that $\theta^{(1)}$ and $t^{(1)}$ are m-dimensional. Considering $V = t^{(1)}(X)$, A and B are determined by $t^{(1)} = At$ and $\theta^{(2)} = B\theta$. According to Theorem 1.21, the family of marginal distributions of $t^{(1)}(X)$ for $\theta^{(2)} = \theta_0^{(2)}$ fixed, is a full exponential family with canonical parameter $\theta^{(1)}$. According to Theorem 1.22 the conditional distributions of X given $t^{(1)}(X) = t_0^{(1)}$ are exponential families, whose parameter only depends on $\theta^{(2)}$.

1.5 Parametrizations

Let $\mathcal{P}$ be a regular exponential family. In Section 1.1 we introduced the canonical parameter θ whose domain is Θ. It happens that the correspondence $\theta \mapsto P_\theta$ is not one-to-one, except if the functions $1, t_1, \ldots, t_k$ are linearly independent with respect to μ. In the following, we will discuss two other parametrizations $\mathcal{P}$, which will be introduced as functions on Θ, but which as functions of $\mathcal{P}$ are one-to-one .

When $\mathcal{P}$ is regular and the representation is minimal, the family $\mathcal{P}$ can be parametrized by the mean $\tau(\theta)$, since, by Theorem 1.17, τ is injective. This parametrization will be called the *parametrization by the mean*. By Theorem 1.18 the domain of τ is int C, if $1, t_1, \ldots, t_k$ are linearly independent with respect to μ. If this is not the case one can show that the domain of τ is ri C.

In the following we will introduce a third concept, that of a mixed parametrization. As in Section 1.4, let A be an $m \times k$ matrix of rank m and B a $(k-m) \times k$ matrix of rank $k-m$ satisfying $AB^\top = 0$, and let $V = v(X) = At(X)$. We define

$$\psi(\theta) = \mathrm{E}_\theta V = \mathrm{E}_\theta\{At(X)\} = A\tau(\theta)$$

$$\sigma(\theta) = B\theta.$$

Theorem 1.23 *The function (ψ, σ) is injective, in the sense that $(\psi, \sigma)(\theta_1) = (\psi, \sigma)(\theta_2) \Rightarrow P_{\theta_1} = P_{\theta_2}$.*

Proof: Let θ_1 and θ_2 be such that $\psi(\theta_1) = \psi(\theta_2)$ and $\sigma(\theta_1) = \sigma(\theta_2)$. From $\psi(\theta_1) = \psi(\theta_2)$ it follows that $A\{\tau(\theta_1) - \tau(\theta_2)\} = 0$. Since $AB^\top = 0$, and B has maximum rank, the rows of B generate the null space of A. Therefore there exists $\delta \in I\!R^{k-m}$, such that $\tau(\theta_1) - \tau(\theta_2) = B^\top\delta$. Now,

$$\begin{aligned}(\theta_1 - \theta_2) \cdot \{\tau(\theta_1) - \tau(\theta_2)\} &= (\theta_1 - \theta_2) \cdot B^\top\delta \\ &= \{\sigma(\theta_1) - \sigma(\theta_2)\}\delta = 0,\end{aligned}$$

using the fact that $\sigma(\theta_1) = \sigma(\theta_2)$. From Theorem 1.17 we obtain that $P_{\theta_1} = P_{\theta_2}$. □

The parameter (ψ, σ) is called the *mixed parameter*.

Theorem 1.24 *In a regular exponential family, the parameters ψ and σ are variation independent, that is, the domain of (ψ, σ) is the product set $\psi(\Theta) \times \sigma(\Theta)$.*

Proof: Let $\theta_0 \in \Theta$ be arbitrary. We consider $\Theta_0 = \{\theta : \sigma(\theta) = \sigma(\theta_0) = \sigma_0\}$. We will show that $\psi(\Theta_0)$ does not depend on σ_0. The hypothesis $\Theta_0 = \{\theta : B\theta = B\theta_0\} \cap \Theta$ is an affine hypothesis. By Theorem 1.21, $V = A\,t(X)$ follows a regular exponential family with canonical parametric domain Θ_0. The mean of V is $\psi(\theta) = A\tau(\theta)$, and the image of τ is the relative interior of the convex support of $V(\mathcal{P}_0)$. Since all the measures of $\mathcal{P}$ are equivalent, the support does not depend on θ_0, and hence $\psi(\Theta_0)$ does not depend on σ_0. □

Example 1.25 *(The Gamma Distribution)* We will illustrate the various given concepts using the gamma distribution, which has density function

$$\frac{dP_{\theta,\lambda}}{d\mu}(x) = \frac{\theta^\lambda}{\Gamma(\lambda)} x^{-1} e^{\lambda \log x - \theta x}, \quad x > 0.$$

We have $c(\theta, \lambda) = \Gamma(\lambda)\theta^{-\lambda}$, $\log c(\theta, \lambda) = \log \Gamma(\lambda) - \lambda \log \theta$, and hence, for $t(x) = (-x, \log x)$,

$$E(-X) = \tau^{(1)}(\theta, \lambda) = -\frac{\lambda}{\theta}$$

$$E \log X = \psi(\lambda) - \log \theta = \tau^{(2)}(\theta, \lambda),$$

where ψ is the digamma function . Then a mixed parametrization is

$$(-\frac{\lambda}{\theta}, \lambda).$$

The interior of C is $\{(x_1, x_2)^T : x_1 < 0 \ \textit{and} \ \log(-x_1) > x_2\}$. Letting $-n\bar{X}_+$ and $n\bar{X}_\sim$ be the canonical statistics for a sample $X_1, \ldots, X_n$, we have the maximum likelihood equations

$$\overline{X}_+ = \frac{\lambda}{\theta}$$

$$\overline{X}_\sim = \psi(\lambda) - \log \theta,$$

which can be written

$$\overline{X}_\sim - \log \overline{X}_+ = \psi(\lambda) - \log \lambda$$

and

$$\theta = \frac{\lambda}{\overline{X}_+}.$$

Finding the solution of the first equation, the second can be solved. We have $(-\overline{X}_+, \overline{X}_\sim) \in$ *int* C with probability 1 in the case $n \geq 2$.

1.6 The multivariate normal distribution

Let X be the random vector with regular multivariate normal distribution,

$$X \sim N_p(\mu, \Sigma),$$

where $\mu \in I\!R^p$, $\Sigma \in S_p$, and S_p is the set of $p(p+1)/2$ vectors representing the set of $p \times p$ symmetric and positive definite matrices. The density function of X with respect to Lebesgue measure is

$$(2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left\{-1/2(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}.$$

In order to find an exponential representation, we will write the exponent in the form $\theta \cdot t$:

$$\begin{aligned} -\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) &= -\frac{1}{2}x^\top\Sigma^{-1}x + x^\top\Sigma^{-1}\mu - \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\ &= \mathrm{tr}(-\frac{1}{2}\Sigma^{-1}xx^\top) + (\Sigma^{-1}\mu)\cdot x - \frac{1}{2}\mu^\top\Sigma^{-1}\mu \\ &= \mathrm{tr}\{\theta^{(2)}t^{(2)}(x)\} + \theta^{(1)}\cdot t^{(1)}(x) + \frac{1}{4}\theta^{(1)^\top}\theta^{(2)^{-1}}\theta^{(1)}, \end{aligned}$$

where $\theta^\top = (\theta^{(1)^\top}, \theta^{(2)^\top}) = (\Sigma^{-1}\mu, -\frac{1}{2}\Sigma^{-1})$ and

$$t^\top(x) = (t^{(1)^\top}(x), t^{(2)^\top}(x)) = (x, xx^\top).$$

Here, we adopt the convention that $t^{(2)}$ and $\theta^{(2)}$ are symmetric matrices, *i.e.*, it is a vector of dimension $p(p+1)/2$ and $tr(\theta^{(2)}t^{(2)})$ denotes the inner product between two matrices, with

$$\mathrm{tr}(AB) = \sum_{i=1}^p\sum_{j=1}^p A_{ij}B_{ji} = \sum_{i=1}^p\sum_{j=1}^p A_{ij}B_{ij}$$

being essentially the inner product between A and B as vectors in $I\!R^{p(p+1)/2}$.

Let M_p be the space of $p \times p$ symmetric matrices . Note that $S_p \subseteq M_p$. We have

$$\begin{aligned} c(\theta) &= (2\pi)^{p/2}|\Sigma|^{1/2}e^{-1/4\theta^{(1)^\top}\theta^{(2)^{-1}}\theta^{(1)}} \\ &= (2\pi)^{p/2}|-\frac{1}{2}\theta^{(2)^{-1}}|^{\frac{1}{2}}e^{-1/4\theta^{(1)^\top}\theta^{(2)^{-1}}\theta^{(1)}} \\ &= \pi^{p/2}|-\theta^{(2)}|^{-\frac{1}{2}}e^{-1/4\theta^{(1)^\top}\theta^{(2)^{-1}}\theta^{(1)}}. \end{aligned}$$

Therefore,

$$\Theta = \{(\theta^{(1)}, \theta^{(2)}) : \theta^{(1)} \in I\!R^p, -\theta^{(2)} \in S_p\}.$$

S_p is an open set in M_p, hence, Θ is open. We also have that if

$$\mathrm{tr}\{\theta^{(2)}xx^\top\} + \theta^{(1)}\cdot x = K \quad [\mu],$$

then $\theta^{(2)} = 0$, $\theta^{(1)} = 0$ and $K = 0$. Hence, the family of normal distributions is a regular family of order $p + p(p+1)/2$.

The parametrization by the mean is

$$\tau^{(1)} = \mathrm{E}_\theta t^{(1)}(X) = \mathrm{E}_\theta X = \mu$$
$$\tau^{(2)} = \mathrm{E}_\theta t^{(2)}(X) = \mathrm{E}_\theta \left(XX^\top\right) = \Sigma + \mu\mu^\top.$$

Since the image of τ is int C, we have

$$\text{int } C = \{(t^{(1)}, t^{(2)}) : t^{(1)} \in I\!R^p, t^{(2)} - t^{(1)}t^{(1)^\top} \in S_p\}$$

and hence

$$C = \{(t^{(1)}, t^{(2)}) : t^{(1)} \in I\!R^p, t^{(2)} - t^{(1)}t^{(1)^\top} \in S_p^0\},$$

where S_p^0 is the set of positive semi-definite, symmetric, $p \times p$ matrices.

The maximum likelihood equations for the independent observations $X_1, \ldots, X_n$ from $N_p(\mu, \Sigma)$ are

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^n X_i \\ \Sigma + \mu\mu^\top &= \frac{1}{n}\sum_{i=1}^n X_i X_i^\top, \end{aligned}$$

whose solution is

$$\begin{aligned} \hat{\mu} &= \overline{X}_+ \\ \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \overline{X}_+\overline{X}_+^\top. \end{aligned}$$

This solution is the maximum likelihood estimator if and only if $(\overline{X}_+, \frac{1}{n}\sum_{i=1}^n X_i X_i^\top) \in$ int C, that is, if and only if $\frac{1}{n}\sum_{i=1}^n X_i X_i^\top - \overline{X}_+\overline{X}_+^\top = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_+)(X_i - \overline{X}_+)^\top$ is positive definite. This happens with probability 1 if $n > p$.

A mixed parametrization is

$$\mathrm{E}_\theta t^{(1)}(X) = \mu, \quad \theta^{(2)} = -\frac{1}{2}\Sigma^{-1},$$

which is essentially the parametrization using the mean and the precision.

1.7 Asymptotic theory

We will consider the regular exponential family, with minimal representation

$$\frac{dP_\theta}{d\mu}(x) = \frac{1}{c(\theta)} e^{\theta \cdot t(x)}, \quad x \in \mathcal{X}, \tag{1.27}$$

where $\theta \in \Theta \subseteq \mathbb{R}^k$. Therefore, Θ is open and $1, t_1, \ldots, t_k$ are linearly independent with respect to μ.

We will study models of the form

$$\frac{dP_\beta}{d\mu}(x) = \frac{1}{c(\theta(\beta))} e^{\theta(\beta)\cdot t(x)}, \tag{1.28}$$

where $\beta \in B \subseteq \mathbb{R}^m$, B is open and convex, and $1 \leq m \leq k$. Here, θ is a function $\theta : B \longrightarrow \Theta$. We can think of (1.28) as a sub-family of a regular exponential family, Almost all the conclusions can be generalized for the case where $\theta \in \text{int } \Theta$, if (1.28) is steep.

We will study the asymptotic theory of estimation and tests for n independent and identically distributed observations, when $n \to \infty$. It will be necessary to impose regularity conditions on the family, which will be conditions exclusively on the function θ. We will assume that $\theta : B \to \Theta$ satisfies the following regularity conditions:

(G1) θ is a homeomorphism (i.e., it is a continuous bijection with continuous inverse);

(G2) θ is twice differentiable and the second derivative is continuous (i.e., it is of class C^2);

(G3) $\frac{\partial \theta^\top}{\partial \beta}$ has rank m.

In fact, our conclusions demand only that θ be differentiable with continuous derivative, but the proofs are simpler if θ is twice differentiable. Condition (G3) prevents us from finding singularities in the surface $\theta(B)$, that is "holes", where the true dimension of $\theta(B)$ is smaller than m.

A model as the one given by (1.28) that satisfies (G1), (G2) and (G3) is called a *smooth* model, or a smooth hypothesis, and m is called *order* of the model or of the hypothesis. Every affine hypothesis in the regular exponential family is a smooth model. Examples of affine hypothesis are the log-linear models for contingency tables.

We will show the asymptotic results for estimation and hypothesis testing for n independent observations from the same smooth family. Hence, it would be easy to generalize the results to a situation where we have independent observations X_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$, the density function of X_{ij} having the form

$$\frac{1}{c_i(\theta_i(\beta))} e^{\theta_i(\beta)\cdot t(x)}, \quad \beta \in B \subseteq \mathbb{R}^m,$$

where the functions $\theta_i : B \to \Theta_i$ satisfy (G1), (G2), (G3). The asymptotic results could be obtained if $n_i \to \infty$, $i = 1, \ldots, k$, in such a way that $n_i/(n_1 + \ldots + n_k)$ would not be close to 0 or 1.

1.7.1 Estimation

The likelihood function for n independent and identically distributed observations is

$$L(\beta) = \frac{1}{c\{\theta(\beta)\}^n} \exp\left\{\theta(\beta) \cdot \sum_{i=1}^{n} t(x_i)\right\}, \quad \beta \in B, \tag{1.29}$$

and the log-likelihood function is

$$\ell(\beta) = \log L(\beta) = n\{\theta(\beta) \cdot \bar{t}_n - \log c[\theta(\beta)]\}, \tag{1.30}$$

where $\bar{t}_n = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$. Therefore, the maximum likelihood equation is

$$\frac{\partial \ell}{\partial \beta}(\beta) = 0 \quad \text{or} \quad \frac{\partial \theta^\top}{\partial \beta}\{\bar{t}_n - \tau[\theta(\beta)]\} = 0. \tag{1.31}$$

Using that

$$\frac{\partial \tau^\top}{\partial \beta}(\beta) = \frac{\partial \theta^\top}{\partial \beta}\frac{\partial \tau^\top}{\partial \theta} = \frac{\partial \theta^\top}{\partial \beta} V\{\theta(\beta)\} \quad ,$$

the equations have the alternative form

$$\frac{\partial \tau^\top}{\partial \beta} V\{\theta(\beta)\}^{-1}\{\bar{t}_n - \tau[\theta(\beta)]\} = 0. \tag{1.32}$$

We have that the columns of the $k \times m$ matrix, $\partial\theta^\top/\partial\beta$ generate the tangent plane to the set $\theta(B)$ at the point $\theta(\beta)$, and the columns of the $k \times m$ matrix $\partial\tau^\top/\partial\beta$ generate the tangent plane to the set $\tau(\theta(B))$ at the point $\tau(\theta(\beta))$. Therefore, the solutions of (1.31) and (1.32) have geometric interpretations. In (1.31) the solution is given by the β such that the vector $\bar{t}_n - \tau(\theta(\beta))$ is orthogonal to the tangent plane to the set $\theta(B)$ at the point $\theta(\beta)$. In (1.32) the solution is given by the β such that the vector $\bar{t}_n - \tau(\theta(\beta))$ is orthogonal to the tangent plane to the set $\tau(\theta(B))$ at the point $\tau(\theta(\beta))$, being the orthogonality with respect to the inner product defined by $V(\theta(\beta))^{-1}$. See Figure 1.2.

In general, the likelihood function does not necessarily have a unique maximum, and in some specific models it is necessary to verify that this happens. We will assume here that, for all $t \in \text{int } C$, there exists a unique maximum:

(G4) There exists a measurable Borel function $g:\ \text{int } C \to B$, such that for all $t \in \text{int } C$, the function

$$f(\beta, t) = \theta(\beta) \cdot t - \log c(\theta(\beta)), \quad \beta \in B$$

has a unique maximum at $\beta = g(t)$.

With this condition we can show the theorem on the existence of the maximum likelihood estimator.

Theorem 1.26 *With probability 1, there exists, for large enough n, a unique maximum likelihood estimator of β, $\hat{\beta} = g(\bar{t}_n)$. This estimator is Fisher consistent* i.e., $g\{\tau[\theta(\hat{\beta})]\} = \hat{\beta}$, $\beta \in B$.

Proof: From the strong law of large numbers it follows that with probability 1

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} t(X_i) \to \mathrm{E}_\theta T = \tau\{\theta(\beta)\}.$$

Since $\tau\{\theta(\beta)\} \in \text{int } C$, $\bar{t}_n \in \text{int } C$, for n large enough. By condition (G4) we have then that $g(\bar{t}_n)$ maximizes the likelihood function.

From the discussion of maximum likelihood estimation of regular exponential families (1.3), we know that for $t \in \text{int } C$

$$\theta \cdot t - \log c(\theta) \leq \hat{\theta} \cdot t - \log c(\hat{\theta}), \quad \forall \theta \in \Theta, \tag{1.33}$$

with equality if and only if, $\theta = \hat{\theta} = \tau^{-1}(t)$.

If $t = \tau_0 = \tau(\theta(\beta_0))$ and θ has the form $\theta = \theta(\beta)$, (1.33) turns into

$$\theta(\beta) \cdot \tau_0 - \log c(\theta(\beta)) \leq \theta(\beta_0) \cdot \tau_0 - \log c(\theta(\beta_0)), \quad \beta \in B.$$

Therefore, the maximum of the left hand side is $\beta = \beta_0$. According to the definition of g we have $g(\tau_0) = \beta_0$, that is,

$$g\{\tau[\theta(\beta_0)]\} = \beta_0,$$

showing that $g(\bar{t}_n)$ is Fisher consistent. □

The conclusion on the existence of the maximum likelihood estimator β in Theorem 1.26 is not very strong. It says that $[P^{\infty}_{\theta(\beta)}]$, for almost all sequences $\{x_i\}_{i=1}^{\infty} \in \mathcal{X}^{\infty}$ we have

$$\frac{1}{n}\sum_{i=1}^{n} t(x_i) \in \text{int } C \quad ,$$

for $n > n_0$, but n_0 depends on the sequence $\{x_i\}_{i=1}$. For all the examples of continuous exponential families that we found until now, it is valid that $\frac{1}{n}\sum_{i=1}^{n} t(x_i) \in \text{int } C$ with probability 1 if and only if n is larger than a fixed n_0, being $n_0 = 2$ for the gamma, normal, and inverse normal distributions, and p for the multivariate normal.

About consistency, recall that a sequence of estimators $\{T_n\}_{n=1}^{\infty}$ for the parameter β is called consistent if

$$T_n \to \beta \qquad [P^{\infty}_{\beta}] \quad .$$

Here, T_n is defined for a sample of size n. This definition does not necessarily imply that T_n is a reasonable estimator for a sample of fixed size. If $\{T_n\}_{n=1}^{\infty}$ is consistent, then $\{\tilde{T}_n\}_{n=1}^{\infty}$, where

$$\tilde{T}_n = \begin{cases} \beta_0, & n < 10^6 \\ T_n, & n \geq 10^6, \end{cases}$$

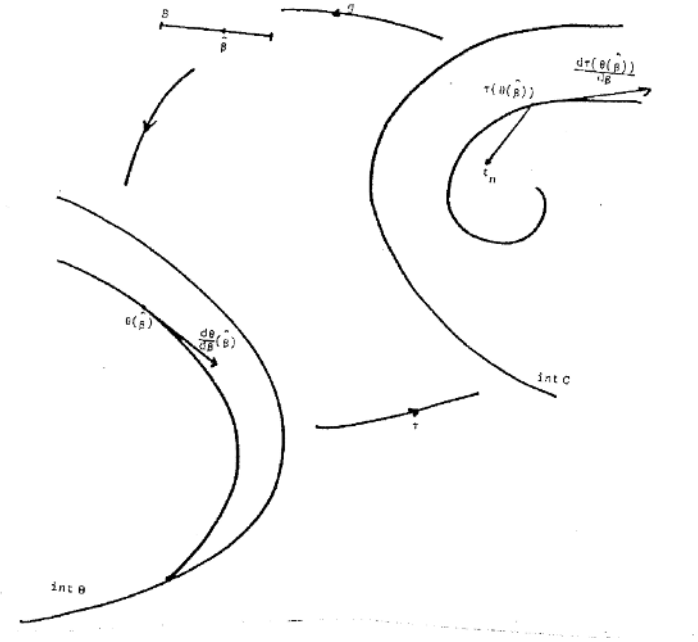

Figure 1.2: Geometrical interpretation of maximum likelihood estimation in a smooth exponential family

is also consistent. Fisher consistency, is however a reasonable property for any sample size. Essentially, this property says that, if the observations fit a given member of the family perfectly, then the estimator must point to this distribution.

We will compute the derivative of g at $\tau\{\theta(\beta)\}$, $\beta \in B$.

Lemma 1.27 *The function g is continuously differentiable, and the derivative at the point $\tau\{\theta(\beta)\}$ is given by*

$$\left.\frac{\partial g^\top}{\partial t}\right|_{t=\tau(\theta(\beta))} = \frac{\partial \theta}{\partial \beta^\top} i(\beta)^{-1},$$

where

$$i(\beta) = \frac{\partial \theta^\top}{\partial \beta} V\{\theta(\beta)\} \frac{\partial \theta}{\partial \beta^\top}$$

is Fisher's information matrix for β.

Proof: Since g maximizes the likelihood function, for $t \in \text{int } C$, $g(t)$ is a solution of the maximum likelihood equation (1.31), and hence

$$h(\beta, t) = \frac{\partial \theta^\top}{\partial \beta}(\beta)\{t - \tau(\theta(\beta))\} = 0, \quad \beta = g(t).$$

Since g is given implicitly by the equation $h\{g(t), t\} = 0$, we have by the implicit function theorem, that g is continuously differentiable, and that $\frac{\partial g^\top}{\partial t}$ is determined by the equation

$$\frac{\partial g^\top}{\partial t}\frac{\partial h}{\partial \beta^\top} + \frac{\partial h^\top}{\partial t} = 0 \quad , \quad \beta = g(t), \tag{1.34}$$

if the matrix $\frac{\partial h}{\partial \beta^\top}$ is invertible. The derivative of h is determined by

$$\begin{aligned}\frac{\partial h}{\partial \beta^\top} &= \sum_{i=1}^{k} \frac{\partial^2 \theta_i}{\partial \beta 1 \partial \beta^\top}\{t_i - \tau_i(\theta(\beta))\} - \frac{\partial \theta^\top}{\partial \beta} V(\theta(\beta)) \frac{\partial \theta}{\partial \beta^\top} \\ &= -\frac{\partial \theta^\top}{\partial \beta} V(\theta(\beta)) \frac{\partial \theta}{\partial \beta^\top} \quad \text{for} \quad t = \tau(\theta(\beta))\end{aligned}$$

and

$$\frac{\partial h^\top}{\partial t} = \frac{\partial \theta}{\partial \beta^\top}.$$

We have $\frac{\partial h}{\partial \beta^\top}|_{t=\tau(\theta(\beta))} = -i(\beta)$, which is invertible, where $V\{\theta(\beta)\}$ invertible, and $\frac{\partial \theta}{\partial \beta^\top}$ has rank m. Inserting in (1.34) we obtain the conclusion. □

Lemma 1.28 *Let X_n be a sequence of random vectors of dimension k, such that*

$$X_n \xrightarrow{P} c,$$

where c is a constant, and

$$b_n(X_n - c) \xrightarrow{\mathcal{D}} Y,$$

where Y is non-degenerate, and $b_n \to \infty$. Let $g : I\!R^k \to I\!R^m$ be differentiable in c. Then

$$b_n\{g(X_n) - g(c)\} - b_n \frac{dg}{dt^\top}(X_n - c) \xrightarrow{P} 0$$

and

$$b_n\{g(X_n) - g(c)\} \xrightarrow{\mathcal{D}} \frac{dg}{dt^\top}(c)Y.$$

Proof: See Problem 1.15.

Theorem 1.29 *The maximum likelihood estimator is asymptotically normal with*

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} N(0, i(\beta)^{-1}).$$

Proof: By the central limit theorem we have that

$$\sqrt{n}\{\overline{T}_n - \tau(\theta(\beta))\} \xrightarrow{\mathcal{D}} N(0, V(\theta(\beta))).$$

Using Lemma 1.28, we have

$$\sqrt{n}\{g(\overline{T}_n) - g\{\tau[\theta(\beta)]\}\} = \sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{D}} N\left(0, \frac{dg}{d\tau^\top V}\{\theta(\beta)\}\frac{dg^\top}{dt}\right).$$

According to Lemma 1.27 we have

$$\frac{\partial g}{\partial \tau^\top} V\{\theta(\beta)\}\frac{\partial g^\top}{\partial \tau} = i(\beta)^{-1}\frac{\partial \theta^\top}{\partial \beta} V(\theta(\beta))\frac{\partial \theta}{\partial \beta^\top} i(\beta)^{-1} = i(\beta)^{-1}.$$

□

Theorem 1.30 *Let $h:$ int $C \to B$ be differentiable and Fisher consistent. Then*

$$\sqrt{n}\{h(\overline{T}_n) - \beta\} \xrightarrow{\mathcal{D}} N(0, V_h),$$

and $V_h - i(\beta)^{-1}$ is positive semi-definite, that is, $\hat{\beta}$ is efficient in the class of the differentiable estimators and Fisher consistent.

Proof: We consider the vector $Y_n^\top = (h(\overline{T}_n)^\top, g(\overline{T}_n)^\top)$. According to Lemma 1.28, $\sqrt{n}(Y_n - (\beta, \beta)^\top) \xrightarrow{\mathcal{D}} N(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} \frac{\partial h}{\partial t^\top} \\ \frac{\partial g}{\partial t^\top} \end{bmatrix} V\{\theta(\beta)\}\left\{\frac{\partial h^\top}{\partial t}\frac{\partial g^\top}{\partial t}\right\} = \begin{bmatrix} V_h & \Sigma_{12} \\ \Sigma_{21} & i(\beta)^{-1} \end{bmatrix}.$$

Hence

$$\Sigma_{12} = \frac{\partial h}{\partial t^\top} V\{\theta(\beta)\}\frac{\partial g^\top}{\partial t} = \frac{\partial h}{\partial t^\top} V\{\theta(\beta)\}\frac{\partial \theta}{\partial \beta^\top} i(\beta)^{-1},$$

and using that h is Fisher consistent, we have $h\{\tau[\theta(\beta)]\} = \beta$, and hence

$$I_m = \frac{\partial h}{\partial \beta^\top} = \frac{\partial h}{\partial t^\top}\frac{\partial \tau}{\partial \theta^\top}\frac{\partial \theta}{\partial \beta^\top} \qquad t = \tau(\theta(\beta)).$$

Since $\frac{\partial \tau}{\partial \theta^\top} = V(\theta)$, we have that $\Sigma_{12} = i(\beta)^{-1}$, and then

$$\Sigma = \begin{bmatrix} V_h & i(\beta)^{-1} \\ i(\beta)^{-1} & i(\beta)^{-1} \end{bmatrix}.$$

Therefore, the asymptotic distribution of $\{h(\overline{T}_n)^\top - g(\overline{T}_n), g(\overline{T}_n)^\top\}$ is

$$\sqrt{n}\{h(\overline{T}_n)^\top - g(\overline{T}_n)^\top, g(\overline{T}_n)^\top\}^\top \xrightarrow{d} N(0, \Gamma),$$

where

$$\Gamma = \begin{bmatrix} I_m & -I_m \\ 0 & I_m \end{bmatrix} \Sigma \begin{bmatrix} I_m & 0 \\ -I_m & I_m \end{bmatrix} = \begin{bmatrix} V_h - i(\beta)^{-1} & 0 \\ 0 & i(\beta)^{-1} \end{bmatrix}.$$

Since $V_h - i(\beta)^{-1}$ is a matrix of variances and covariances, we have that it is positive semi-definite. □

Note that if $h(\overline{T}_n) - g(\overline{T}_n)$ and $g(\overline{T}_n)$ had been normally distributed with covariance $\frac{1}{n}\Gamma$, then these random variables would be independent. Therefore, $h(\overline{T}_n) - g(\overline{T}_n)$ and $g(\overline{T}_n)$ are called *asymptotically independent*.

1.7.2 Hypothesis testing

In this section we will find the asymptotic distribution of likelihood ratio tests for smooth hypothesis. Also, we will give some approximations for the likelihood ratio test.

Frequently, we will test several hypothesis successively. In order to better describe the relations among the hypothesis tests and the hypothesis estimators, we consider simultaneously the following three hypothesis:

H_0: $\theta = \theta(\beta) \quad \beta \in B \subseteq \mathbb{R}^m, \quad m < k$

H_1: $\beta = \beta(\alpha) \quad \alpha \in A \subseteq \mathbb{R}^r, \quad 0 \leq r < m$

H_2: $\alpha = \alpha_0$.

Both H_0 and H_2 are special cases of H_1. We will assume that the function $\beta : A \to B$ satisfies conditions (G1), (G2), (G3) and (G4) of a smooth family.

Let $\alpha_0 \in A$ and let $\beta_0 = \beta(\alpha_0)$, $\theta_0 = \theta(\beta_0)$, $\tau_0 = \tau(\theta_0)$, $V_0 = V(\theta_0)$ and $i_0 = i(\theta_0)$. Let also S be the tangent space to Θ of the curve $\theta(B)$ at the point θ_0, and S_1 be the tangent space to Θ of $\theta(\beta(A))$ at the point θ_0. Evidently,

$$S_1 \subseteq S \subseteq \mathbb{R}^k.$$

See Figure 1.3.

Finally, $\hat{\alpha}$ represents the maximum likelihood estimator of α under H_1, and $\hat{\beta}$ represents the maximum likelihood estimator of β under H_0, and $\hat{\theta}$ represents the maximum likelihood estimator under the full family. See Figure 1.4.

We will need the orthogonal projections onto S and S_1.

Lemma 1.31 *The tangent plane S is generated by the columns of $\frac{d\theta}{d\beta_0^\top} = \frac{d\theta}{d\beta^\top}(\beta_0)$, and the orthogonal projection on S with respect to the inner product V_0 has matrix*

$$P = \frac{\partial \theta}{\partial \beta_0^\top} i_0^{-1} \frac{\partial \theta^\top}{\partial \beta_0} V_0.$$

In the same way, the orthogonal projection on S_1 with respect to V_0 has matrix

$$P_1 = \frac{\partial \theta}{\partial \alpha_0^\top} i_1^{-1} \frac{\partial \theta^\top}{\partial \alpha_0} V_0,$$

where $i_1 = i_1(\alpha_0)$ represents the information under H_1.

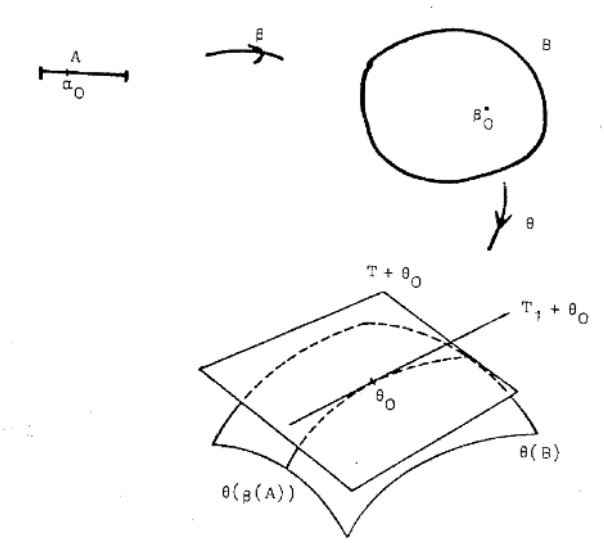

Figure 1.3: Parameter domains and tangent spaces

Proof: For a weighted regression with design matrix $\frac{\partial \theta}{\partial \beta^\top}$ and weights V_0, we have the projection

$$P = \frac{\partial \theta}{\partial \beta_0^\top} \left(\frac{\partial \theta^\top}{\partial \beta_0} V_0 \frac{\partial \theta}{\partial \beta_0^\top} \right)^{-1} \frac{\partial \theta^\top}{\partial \beta_0} V_0.$$

□

Theorem 1.32 *For $\alpha = \alpha_0$, that is under $P_{\theta_0}^\infty$, we have*

$$\sqrt{n} \begin{bmatrix} \hat{\theta} - \theta(\hat{\beta}) \\ \theta(\hat{\beta}) - \theta(\beta(\hat{\alpha})) \\ \theta(\beta(\hat{\alpha})) - \theta_0 \end{bmatrix} \xrightarrow{\mathcal{D}} \begin{bmatrix} I - P \\ P - P_1 \\ P_1 \end{bmatrix} Y, \tag{1.35}$$

where $Y \sim N_k(0, V_0^{-1})$.

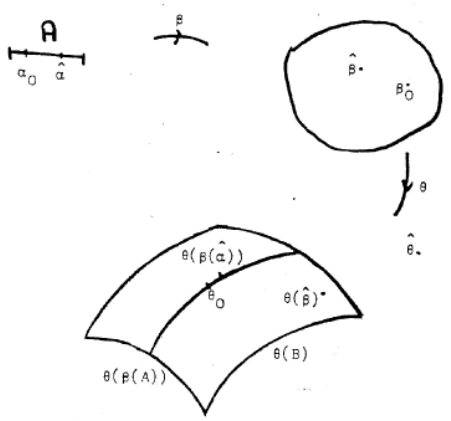

Figure 1.4: Maximum likelihood estimators under the different hypotheses

Proof: Since $\sqrt{n}(\overline{T}_n - \tau_0) \xrightarrow{\mathcal{D}} W$, where $W \sim N_k(0, V_0)$, we obtain

$$\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\tau^{-1}(\overline{T}_n) - \tau^{-1}(\tau_0)) \xrightarrow{\mathcal{D}} V_0^{-1}W = Y.$$

The functions $\theta(\hat{\beta})$ and $\theta(\beta(\hat{\alpha}))$ are differentiable functions of $\hat{\theta}$, hence the conclusion follows finding the derivatives and using Lemma 1.28. In this way,

$$\begin{aligned}\frac{\partial\theta(\hat{\beta})}{\partial\hat{\theta}^\top} &= \tfrac{\partial}{\partial\hat{\theta}^\top}\theta(g(\tau(\hat{\theta}))) \\ &= \tfrac{\partial\theta(\hat{\beta})}{\partial\beta^{1\top}}\tfrac{\partial g(\tau(\hat{\theta}))}{\partial\tau^\top}\tfrac{\partial\tau(\hat{\theta})}{\partial\hat{\theta}^\top}.\end{aligned}$$

For $\hat{\theta} = \theta_0$ and $\hat{\beta} = g(\tau(\hat{\theta})) = \beta_0$ we obtain by Lemmas 1.27 and 1.31, that

$$\frac{\partial\theta(\hat{\beta})}{\partial\hat{\theta}^\top}\Big|_{\hat{\theta}=\theta_0} = \frac{\partial\theta}{\partial\beta_0^\top}i_0^{-1}\frac{\partial\theta^\top}{\partial\beta_0}V_0 = P.$$

Therefore, we have by Lemma 1.28, that

$$\sqrt{n}\{\theta(\hat{\beta}) - \theta_0\} \xrightarrow{\mathcal{D}} PY.$$

In the same way it follows that

$$\frac{d\theta(\beta(\hat{\alpha}))}{d\hat{\theta}^\top}\Big|_{\hat{\theta}=\theta_0} = P_1,$$

and hence

$$\sqrt{n}\{\theta(\beta(\hat{\alpha})) - \theta_0\} \xrightarrow{\mathcal{D}} P_1Y.$$

This completes the proof. □

Since $S \supseteq S_1$, and P and P_1 are V_0-orthogonal on T and T_1, respectively, we have that $I-P$, $P-P_1$ and P_1 are V_0-orthogonal projections on the spaces that are V_0-orthogonal. In terms of matrices we have

$$PP_1 = P_1, \tag{1.36}$$

$$(I-P)^2 = I-P, \quad (P-P_1)^2 = P-P_1, \quad P_1^2 = P_1, \tag{1.37}$$

and

$$V_0(I-P) = (I-P)^\top V_0,\ V_0(P-P_1) = (P-P_1)^\top V_0,\ V_0P_1 = P_1^\top V_0. \tag{1.38}$$

Formula (1.272) shows that we have V_0-orthogonal projections.

These conclusions can be used to show that the three components on the right hand side of (1.33) are independent . We have

$$\begin{bmatrix} (I-P)Y \\ (P-P_1)Y \\ P_1Y \end{bmatrix} \xrightarrow{\mathcal{D}} N_{3k}(0,\Sigma),$$

where Σ is given by

$$\begin{bmatrix} (I-P)V_0^{-1}(I-P)^\top & (I-P)V_0^{-1}(P-P_1)^\top & (I-P)V_0^{-1}P_1^\top \\ (P-P_1)V_0^{-1}(I-P)^\top & (P-P_1)V_0^{-1}(P-P_1)^\top & (P-P_1)V_0^{-1}P_1^\top \\ P_1V_0^{-1}(I-P)^\top & P_1V_0^{-1}(P-P_1)^\top & P_1V_0^{-1}P_1^\top \end{bmatrix}$$
$$= \begin{bmatrix} (I-P)V_0^{-1}(I-P)^\top & 0 & 0 \\ 0 & (P-P_1)V_0^{-1}(P-P_1)^\top & 0 \\ 0 & 0 & P_1V_0^{-1}P_1^\top \end{bmatrix},$$

because it follows from (1.36) and (1.38) that the elements outside the diagonal are zero. It also follows that

$$Y^\top(I-P)^\top V_0(I-P)Y, \quad Y^\top(P-P_1)^\top V_0(P-P_1)Y \text{ and } Y^\top P_1^\top V_0P_1Y$$

have χ^2 distribution with $k-m$, $m-r$ and r degrees of freedom, respectively. Therefore, it has been proved that

Theorem 1.33 *For $\alpha = \alpha_0$, that is under $P_{\theta_0}^\infty$, we have that*

$$\sqrt{n}\{\hat\theta - \theta(\hat\beta)\}, \quad \sqrt{n}\{\theta(\hat\beta) - \theta[\beta(\hat\alpha)]\} \text{ and } \sqrt{n}\{\theta(\beta(\hat\alpha)) - \theta_0\}$$

are asymptotically independent and normal. The quadratic forms

$$\begin{aligned} K_0 &= n\{\hat\theta - \theta(\hat\beta)\}^\top V_0\{\hat\theta - \theta(\hat\beta)\} \\ K_1 &= n\{\theta(\hat\beta) - \theta(\beta(\hat\alpha))\}^\top V_0\{\theta(\hat\beta) - \theta(\beta(\hat\alpha))\} \\ K_3 &= n\{\theta(\beta(\hat\alpha)) - \theta_0\}^\top V_0\{\theta(\beta(\hat\alpha)) - \theta_0\} \end{aligned}$$

are asymptotically independent with χ^2-distributions with $k-m$, $m-r$ and r degrees of freedom, respectively.

Let us consider now the likelihood ratio test, Q, of the hypothesis $H_1 : \beta = \beta(\alpha)$, $\alpha \in A$ on $H_0 : \theta = \theta(\beta)$, $\beta \in B$.

Theorem 1.34 *The statistic $2 \log Q$ has asymptotic distribution $\chi^2(m-r)$, and is asymptotically independent of the maximum likelihood estimator $\hat{\alpha}$ of α.*

Proof: We have

$$\begin{aligned} \log Q &= n\{\theta(\hat{\beta}) - \theta(\beta(\hat{\alpha}))\} \cdot \overline{T}_n - n\{\log c(\theta(\hat{\beta})) - \log c(\theta(\beta(\hat{\alpha})))\} \\ &= n\{\theta(\hat{\beta}) - \theta(\beta(\hat{\alpha}))\} \cdot \tau(\hat{\theta}) - n\{\log c(\theta(\hat{\beta})) - \log c(\theta(\beta(\hat{\alpha})))\}, \end{aligned}$$

since $\overline{T}_n = \tau(\hat{\theta})$ for $\overline{T}_n \in \text{int } C$. Using the Taylor series expansion of $\log c$ we have

$$\log c(\theta_1) - \log c(\theta_2) = (\theta_1 - \theta_2) \cdot \tau(\theta_2) + \frac{1}{2}(\theta_1 - \theta_2)^\top V(\tilde{\theta})(\theta_1 - \theta_2),$$

for some $\tilde{\theta} \in [\theta_1, \theta_2]$. Writing $\theta_2 = \theta(\hat{\beta})$, $\theta_1 = \theta(\beta(\hat{\alpha}))$ we have

$$\begin{aligned} \log Q &= n\{\theta(\hat{\beta}) - \theta(\beta(\hat{\alpha}))\} \cdot \{\tau(\hat{\theta}) - \tau(\theta(\hat{\beta}))\} \\ &\quad + \frac{1}{2} n\{\theta(\hat{\beta}) - \theta(\beta(\hat{\alpha}))\}^\top V(\tilde{\theta})\{\theta(\hat{\beta}) - \theta(\beta(\hat{\alpha}))\}, \end{aligned}$$

where $\tilde{\theta} \in [\theta(\beta(\tilde{\alpha})), \theta(\tilde{\beta}),]$.

Since $\theta(\hat{\beta}) \xrightarrow{P} \theta_0$ and $\theta(\beta(\hat{\alpha})) \xrightarrow{P} \theta_0$; hence $\tilde{\theta} \xrightarrow{P} \theta_0$ and $V(\tilde{\theta}) \xrightarrow{P} V_0$, we have by Theorem 1.32, that

$$\begin{aligned} 2 \log Q &\xrightarrow{d} 2Y^\top (P - P_1)^\top V_0 (I - P)Y + Y^\top (P - P_1)^\top V_0 (P - P_1)Y \\ &= Y^\top (P - P_1)^\top V_0 (P - P_1)Y, \end{aligned}$$

which, according to the comment after Theorem 1.32, follows a $\chi^2(m-r)$ distribution, and is independent of $P_1 Y$, which determines the asymptotic distribution of $\theta(\beta(\hat{\alpha}))$ and hence that of $\hat{\alpha}$. □

Note that in the proof above we obtained that

$$2 \log Q - K_1 \xrightarrow{P} 0,$$

and hence both statistics, which have the same asymptotic distribution, will take on values that are very close with high probability.

Since both H_0 and H_2 are special cases of H_1, we know now how to test successively H_0 under the regular exponential family, H_1 under H_0 and H_2 under H_1. First, we test H_0 under the regular exponential family and, in accordance with Theorem 1.34, the test is asymptotically a $\chi^2(k-m)$ test, and the test is, under the hypothesis, asymptotically independent of the maximum likelihood estimator of β. Next, we test H_1 under H_0 with a test that is asymptotically $\chi^2(m-r)$, and again the maximum likelihood estimator of α is asymptotically independent of the test. Finally, we can test a simple hypothesis on α.

Table 1.1: Table of likelihood ratio tests

Hypothesis	Test	Degrees of Freedom
H_0	$G(\tilde{\theta}) - G(\theta(\hat{\beta}))$	$k-m$
$H_1 \mid H_0$	$G(\theta(\hat{\beta})) - G(\theta(\beta(\hat{\alpha})))$	$m-r$
$H_2 \mid H_1$	$G(\theta(\beta(\hat{\alpha}))) - G(\theta_0)$	r
Total	$G(\tilde{\theta}) - G(\theta_0)$	k

The similarity with successive tests in the classical analysis of variance is clear. Thus, we can show the results in a table similar to an ANOVA table. Putting $G(\theta) = 2\log L(\theta)$, we have the results in Table 1.1.

It is possible to put the conclusions of Theorem 1.32 and 1.33 in terms of the mean τ instead of the canonical parameter θ. Let $\hat{\tau}$ be the maximum likelihood estimator of τ on H_1 and $\hat{\hat{\tau}}$ be the maximum likelihood estimator under H_2, *i.e.*, $\overline{T}_n = \tau(\hat{\theta})$, $\hat{\tau} = \tau(\theta(\hat{\beta}))$ and $\hat{\hat{\tau}} = \tau(\theta(\beta(\hat{\alpha})))$. The quadratic forms in question are

$$\begin{aligned} K_0' &= n\{\overline{T}_n - \hat{\tau}\}^\top V_0^{-1}(\overline{T}_n - \hat{\tau}) \\ K_1' &= n(\hat{\tau} - \hat{\hat{\tau}})^\top V_0^{-1}(\hat{\tau} - \hat{\hat{\tau}}) \\ K_2' &= n(\hat{\hat{\tau}} - \tau_0)^\top V_0^{-1}(\hat{\hat{\tau}} - \tau_0). \end{aligned}$$

Theorem 1.35 *For $\alpha = \alpha_0$, that is under $P_{\theta_0}^\infty$, we have*

$$\sqrt{n}\{\overline{T}_n - \hat{\tau}, \hat{\tau} - \hat{\tau}, (\hat{\tau} - \tau_0)\} \xrightarrow{\mathcal{D}} (I - P, P - P_1, P_1)W, \tag{1.39}$$

where $W \sim N(0, V_0)$. The three components on the right hand side of (1.39) are asymptotically independent and asymptotically normal. Moreover, K_0', K_1' and K_2' are asymptotically independent, and have asymptotic distributions χ^2 with $k-m$, $m-r$ and r degrees of freedom respectively.

Proof: We know that $\sqrt{n}(\overline{T}_n - \tau_0) \xrightarrow{\mathcal{D}} W$ and $\frac{d\tau}{d\theta} = V_0$, and hence we have that

$$\sqrt{n}(\hat{\tau} - \tau_0) = \sqrt{n}\{\tau(\theta(\hat{\beta})) - \tau(\theta_0)\} \xrightarrow{\mathcal{D}} V_0^{-1}P^T V_0 W = PW,$$

and

$$\sqrt{n}(\hat{\hat{\tau}} - \tau_0) = \sqrt{n}\{\tau(\theta(\beta(\hat{\alpha}))) - \tau(\theta_0)\} \xrightarrow{\mathcal{D}} V_0^{-1}P_1^\top V_0 W = P_1^\top W,$$

according to the proof of Theorem 1.32 and the definitions of P and P_1 in Lemma 1.31.

The other conclusions follow as in Theorem 1.33, using that $P^\top$ and $P_1^\top$ are V_0^{-1} orthogonal, which is the case because P and P_1 are V_0-orthogonal. $P^\top$ and $P_1^\top$ are V_0^{-1} orthogonal in the tangent spaces $V_0 S$ and $V_0 S_1$ with respect to respectively $\tau(\theta(B))$ and $\tau\{\theta[\beta(A)]\}$ in τ_0. □

1.8 Problems

Problem 1.1 Let $\mathcal{P}$ be the family with representation (1.2) with respect to the σ-finite measure μ.

(i) Show that, for all $P \in \mathcal{P}$, we have

$$P(A) = 0 \Leftrightarrow \mu(A) = 0 \quad , \quad \forall A \in \mathcal{A}.$$

This means that all measures in $\mathcal{P}$ have the same null sets, since these are the same null sets of μ. We can then write $[\mathcal{P}]$ instead of $[\mu]$. Then as measures in $\mathcal{P}$ they are equivalent. Hence, by Radon-Nikodym's Theorem, they have density function with relation to any other element $P_0 \in \mathcal{P}$. Consider the representation with respect to μ

$$\frac{dP}{d\mu}(x) = a(P)e^{\alpha(P)\cdot t(x)} \quad [\mu] \quad .$$

(ii) Show that

$$\begin{aligned} \frac{dP}{dP_0}(x) &= \frac{dP}{d\mu}(x)\frac{d\mu}{dP_0}(x) \\ &= \frac{a(P)}{a(P_0)}e^{\{\alpha(P)-\alpha(P_0)\}\cdot t(x)} \quad [P_0] \end{aligned}$$

Problem 1.2 Let $\mathcal{P} \sim EM(t(X), \alpha(P))$. Suppose that the components of T are affinely independent. Show that

$$P_1 = P_2 \Leftrightarrow \alpha(P_1) = \alpha(P_2).$$

Problem 1.3 Let $\mathcal{P} \sim EM(t(X), \alpha(P))$, where $t(X)$ and $\alpha(P)$ have dimension k. Show that the representation is minimal if and only if the order of $\mathcal{P}$ is k.

Problem 1.4 This problem shows that a family $\mathcal{P}$ of equivalent probability measures is an exponential family if and only if the corresponding family of log density functions is contained in a finite-dimensional space.
Let X_1 and X_2 be measurable real functions in $\mathcal{X}$. Then X_1 and X_2 are said to be equivalent if

$$X_1 = X_2 \quad [\mathcal{P}] \tag{1.40}$$

(i) Show that (1.40) is an equivalence relation.

(ii) Let V be the family of equivalence classes of (1.40), and let $\langle X \rangle$ be the equivalence class that contains X. Show that V is a real vector space, with the obvious definitions.

(iii) Let $\mathcal{D}$ be the subset of log density functions for $\mathcal{P}$, that is

$$\mathcal{D} = \left\{ \left\langle \log \frac{dP}{dP_0} \right\rangle : P \in \mathcal{P} \right\},$$

where $P_0 \in \mathcal{P}$. Let

$$\overline{\mathcal{D}} = span[< 1 > \cup \mathcal{D}].$$

Show that $\mathcal{P}$ is an exponential family if and only if $\overline{\mathcal{D}}$ is finite-dimensional.

(iv) Let $\mathcal{P}$ be an exponential family. Show that

$$ord\ \mathcal{P} = \dim \overline{\mathcal{D}} - 1.$$

Problem 1.5 Let $\mathcal{P}$ be an exponential family of order k, with minimal canonical domain Θ. Show that if the closed convex support C is bounded, then $\Theta = I\!R^k$, and hence $\mathcal{P}$ is regular.

Problem 1.6 The Poisson distribution Po_μ has density function

$$\frac{dPo_\mu}{d\nu}(x) = \frac{\mu^x}{x!} e^{-\mu}, \quad x = 0, 1, 2, \ldots,$$

with respect to the counting measure on $\mathbf{N_0}$. Let

$$\mathcal{P} = \{Po_\mu : \mu \in (0, \infty)\}.$$

(i) Show that $\mathcal{P}$ is a regular exponential family of order 1.

(ii) Let $X \sim Po_\mu$. Find EX and Var X.

Problem 1.7 The negative binomial distribution $P_{\lambda,p}$ is given by the density function

$$\frac{dP_{\lambda,p}}{dm}(x) = \begin{pmatrix} \lambda + x - 1 \\ x \end{pmatrix} p^x (1-p)^\lambda, \quad x = 0, 1, \ldots$$

with respect to the counting measure on $\mathbf{N_0}$. Let $\mathcal{P}_\lambda = \{P_{\lambda,p} : p \in (0,1)\}$ be the family of negative binomial distributions with fixed index parameter λ.

(i) Show that $\mathcal{P}_\lambda$ is a regular exponential family of order 1.

(ii) Let $X \sim \mathcal{P}_{\lambda,p}$. Find EX and Var X.

Problem 1.8 The logarithmic distribution P_p has density function

$$\frac{dP_p}{dm}(x) = -\frac{1}{\log(1-p)} \frac{p^x}{x} \quad x = 1, 2, \ldots,$$

with respect to the counting measure on $\mathbf{N}$. Let $\mathcal{P} = \{P_p : p \in (0,1)\}$.

(i) Show that $\mathcal{P}$ is a regular exponential family of order 1.

(ii) Let $X \sim P_p$. Find EX and $\mathrm{Var}X$.

Problem 1.9 Let

$$\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in I\!R, \sigma^2 > 0\}$$

be the family of normal distributions in $I\!R$.

(i) Show that $\mathcal{P}$ is a regular exponential family of order 2, with canonical statistic $t(x) = (x, x^2)$.

(ii) Determine Θ and C.

Problem 1.10 Compute EX, $VarX$, EX^{-1} and $VarX^{-1}$, for $X \sim IG(\mu, \lambda)$ (the inverse Gaussian distribution). Show that (μ, λ) provides a mixed parameter for this family.

Problem 1.11 Let $\mathcal{X} = I\!R$ and ν be a measure with density function $e^{-\frac{1}{2}x^2}$ with respect to Lebesgue measure on $I\!R$. We consider the full exponential family $\mathcal{P}$, generated by ν and $t(x) = (x + x^3, x - x^3)^\top$.

(i) Show that $\Theta = \{\theta : c(\theta) < \infty\} = \{(\theta_1, \theta_2)^\top : \theta_1 = \theta_2\}$ and that $\mathcal{P}$ is regular.

(ii) Show that $ord\mathcal{P} = 1$, and find a minimal canonical statistic.

(iii) Show that $\mathcal{P} = \{N(\mu, 1) : \mu \in I\!R\}$, i.e., the family of normal distributions with variance 1.

(iv) Show that $E_\theta t(X)$ exists.

Now define

$$f(x) = \begin{cases} e^{x^2} - 1, & (x > 0) \\ 0, & (x = 0) \\ -(e^{x^2} - 1), & (x < 0). \end{cases}$$

(v) Show that $\mathcal{P}$ is also generated by ν and $\tilde{t}(x) = (x + f(x), x - f(x))^\top$.

(vi) Show that $\tilde{t}(X)$ does not have a mean with respect to P_θ, and that the unique functionals $w \cdot \tilde{t}(X)$, $w \in I\!R^2$, that have mean, are the ones with $w \in \Theta$.

Comment: The problem shows that, in the case where the canonical parameter domain is relatively open, but not open, one should be careful when considering the mean of the canonical statistic. This explains why we demand that Θ is open in a regular exponential family, even if the representation is not minimal.

Problem 1.12 Let μ be a measure on $\mathcal{X}$, and t be a measurable function from $\mathcal{X}$ to $\mathbb{R}^k$, such that $1, t_1, \ldots, t_k$ are linearly independent . Show that the point t_0 in $\mathbb{R}^k$ belongs to the interior of the convex support $C_{t(\mu)}$ if and only if any hyperplane through t_0 partitions $\mathbb{R}^k$ in two sets with $t(\mu)$-positive measures.

Problem 1.13 Let $\mathcal{P}$ be a regular exponential family with minimal representation

$$\frac{dP_\theta}{d\mu}(x) = a(\theta)e^{\theta \cdot t(x)},$$

and consider the notation of Section 1.4. In particular, let $\mathcal{P}_0$ be an affine hypothesis

$$\mathcal{P}_0 = \{P_\theta : \theta \in \Theta \cap \{\theta_0 + A^\top \eta : \eta \in \mathbb{R}^m\}\},$$

where the $m \times k$ matrix A has full rank m. Let t_0 be an observed value of the minimal canonical statistic T.

(i) Show that, if $t_0 \in int\, C_{t(\mu)}$, then $At_0 \in int\, C_{v(\mu)}$, hence, if the maximum likelihood estimator $\hat{\theta}$ of θ exists in the model $\mathcal{P}$, then the estimator of maximum likelihood $\hat{\theta}$ of θ in the model $\mathcal{P}_0$ also exists.

(ii) Find the likelihood ratio test of $\mathcal{P}_0$ under $\mathcal{P}$.

Problem 1.14 Let X_1 and X_2 independent, with binomial distributions $X_1 \sim Bi(n_1, p_1)$ and $X_2 \sim Bi(n_2, p_2)$.

(i) Show that the family of joint distributions of (X_1, X_2) is an exponential family of order 2, with canonical statistic $(X_1, X_2)^\top$, and canonical parameter

$$(\theta_1, \theta_2)^\top = (\log \frac{p_1}{1-p_1}, \log \frac{p_2}{1-p_2})^\top.$$

(ii) Let us use the terminology of Section 1.4, and let us consider the distribution of $V = X_1 + X_2 = (1,1)(X_1, X_2)^\top$. Let $A = (1,1)$ and $B = (1,-1)$. Show that the affine hypothesis $\Theta_0 = \{\theta_0 + A^\top \eta : \eta \in \mathbb{R}\}$ can be written as

$$\frac{p_1}{1-p_1}\frac{1-p_2}{p_2} = \textit{ constant}$$

in terms of the original parameters.

(iii) Find the marginal distributions of $X_1 + X_2$ under the affine hypothesis Θ_0.

(iv) Find the marginal distributions of $X_1 + X_2$ under the hypothesis $\tilde{\Theta}_0 = \{A^\top \eta : \eta \in \mathbb{R}\}$, or

$$\frac{p_1}{1-p_1}\frac{1-p_2}{p_2} = 1 \quad \textit{(that is, } p_1 = p_2\textit{)}.$$

(v) Find the conditional distribution of $(X_1, X_2)^\top$ given

$$X_1 + X_2 = v.$$

Problem 1.15 Show Lemma 1.28.

Problem 1.16 This problem shows some important results on the information functions i and j, relative to a reparametrization and smooth hypothesis.

(i) Let $a = a(x)$ and $b = b(x)$ be k-dimensional vectors and $M = M(x)$ a $k \times k$ matrix, all differentiable with respect to x, where x is d-dimensional. Suppose that

$$a(x) = M(x)b(x) \qquad \forall x \in \Omega \subseteq I\!R^d,$$

where Ω is an open set. Show that

$$\frac{\partial a}{\partial x^\top} = M\frac{\partial b}{\partial x^\top} + \frac{\partial M}{\partial x^\top} \times b,$$

where the product $\times$ is defined by

$$\frac{\partial M}{\partial x^\top} \times b = \begin{bmatrix} \frac{\partial M}{\partial x_1} b \\ \vdots \\ \frac{\partial M}{\partial x_d} b \end{bmatrix}$$

(ii) Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be the statistical model, with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset I\!R^k$. Suppose that $\mathcal{P} \ll \mu$, where μ is a σ-finite measure. Assume the following hypotheses regarding the density function

$$p(x;\theta) = \frac{dP_\theta}{d\mu}(x).$$

(a) $p(x;\theta) > 0$ $\forall x \in \mathcal{X}$, $\forall \theta \in \Theta$.

(b) $p(x;\cdot)$ is twice continuously differentiable $\forall x \in \mathcal{X}$.

(c) $\frac{\partial}{\partial\theta}\int p(x;\theta)\mu(dx) = \int \frac{\partial}{\partial\theta} p(x;\theta)\mu(dx) = 0$.

(d) The Fisher information function

$$i(\theta) = -\int (\frac{\partial}{\partial\theta})^2 \log p(x;\theta) P_\theta(dx)$$

exists $\forall \theta \in \Theta$.

Let j be the observed information function

$$j(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^\top}\ell(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^\top}\log p(x;\theta).$$

Let Ω be an open set in $\mathbb{R}^d$, $d \leq k$, and suppose that $\theta(w)$ is an injective function, two times continuously differentiable from Ω into Θ. Consider the model $(\mathcal{X}, \mathcal{A}, \mathcal{P}_0)$ where $\mathcal{P}_0 = \{P_{\theta(w)} : w \in \Omega\}$. The information functions of the model $\mathcal{P}_0$ will be called i_w and j_w and the ones for the model $\mathcal{P}$, i_θ and j_θ. Show that

$$j_w(w) = \frac{\partial \theta^\top}{dw}(w) j_\theta(\theta(w)) \frac{\partial \theta}{\partial w^\top} - \frac{\partial}{\partial w}\left(\frac{\partial \theta}{\partial w^\top}(w)\right) \times \frac{\partial \ell}{\partial \theta}(\theta(w))$$

and

$$i_w(w) = \frac{\partial \theta^\top}{\partial w}(w) i_\theta(\theta(w)) \frac{\partial \theta}{\partial w^\top}(w).$$

Also show that if $d = k$ and $|\frac{\partial \theta}{\partial w^\top}(w)| \neq 0$ for all $w \in \Omega$, then

$$j_w(\hat{w}) = \frac{\partial \theta^\top}{\partial w}(\hat{w}) j_\theta(\hat{\theta}) \frac{\partial \theta}{\partial w^\top}(\hat{w}),$$

where $\hat{\theta}$ is the maximum likelihood estimator of θ in the model $\mathcal{P}$, and $\hat{w}$ is the estimator of w under $\mathcal{P}_0$.

(iii) Let now $\mathcal{P}$ be a full exponential family, that is

$$p(x;\theta) = a(\theta) b(x) e^{\theta \cdot t(x)}, \quad \theta \in \Theta,$$

and suppose that $\theta(\Omega) \subseteq int\ \Theta$. Show that

$$i_w(w) = \frac{\partial \theta^\top}{\partial w} V(\theta) \frac{\partial \theta}{\partial w^\top}$$

and

$$j_w(w) = i_w(w) - \frac{\partial}{\partial w}\left(\frac{\partial \theta}{\partial w^\top}\right) \times (t - E_\theta T),$$

where $T = T(x)$, and conclude that, in the case $k = d$,

$$j_w(\hat{w}) = i_w(\hat{w}).$$

Problem 1.17 Let $\mathcal{P}$ be the family of bivariate normal distributions defined by

$$\mathcal{P} = \left\{ P_\alpha : P_\alpha = N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1 & \alpha \\ \alpha & 1+\alpha^2 \end{bmatrix} \right), \quad \alpha \in \mathbb{R} \right\}.$$

(i) Show that $\mathcal{P}$ is an exponential family of order 2.

(ii) Show that $\mathcal{P}$ is not regular.

(iii) Consider from now on $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ independent and identically distributed random vectors, with distribution in $\mathcal{P}$. Show that the maximum likelihood estimator $\hat{\alpha}$ of α is given by

$$dom\ \hat{\alpha} = \{(x, y)^\top = (x_1, \ldots, x_n, y_1, \ldots, y_n)^\top \in I\!R^{2n} : \sum_{j=1}^{n} x_i^2 > 0\}$$

and that

$$\hat{\alpha}(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$$

for $(x, y)^\top \in dom\ \hat{\alpha}$.

(iv) The family satisfies the regularity conditions (G1)–(G4) in Section 1.7 (you do not need to show this). Show that, under P_{α_0},

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{\mathcal{D}} N(0, 1)\ \ as\ n \to \infty. \tag{1.41}$$

(v) Suppose that $(X_i, Y_i)^\top \sim P_{\alpha_0}$, $i = 1, \ldots, n$. Show that

$$\begin{aligned}(Y_1, \ldots, Y_n)^\top | (X_1, \ldots, X_n)^\top &= (x_1, \ldots, x_n)^\top \\ &\sim N_n((\alpha_0 x_1, \ldots, \alpha_0 x_n)^\top, I_n),\end{aligned}$$

where I_n is the $n \times n$ identity matrix . Show that

$$\begin{aligned}\sum_{i=1}^{n} X_i Y_i | (X_1, \ldots, X_n)^\top &= (x_1, \ldots, x_n)^\top \\ &\sim N(\alpha_0 \textstyle\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i^2).\end{aligned}$$

Show that this implies that

$$\left. \frac{\sum_{i=1}^n X_i Y_i - \alpha_0 \sum_{i=1}^n X_i^2}{\sqrt{\sum_{i=1}^n X_i^2}} \right| X = x \sim N(0, 1)$$

and that hence

$$\frac{\sum_{i=1}^n X_i Y_i - \alpha_0 \sum_{i=1}^n X_i^2}{\sqrt{\sum_{i=1}^n X_i^2}} \sim N(0, 1). \tag{1.42}$$

(vi) Consider the hypothesis $H : \alpha = \alpha_0$. Show that the test of H based on (1.42) is identical to the asymptotic test of H, obtained using the observed information instead of Fisher's information.

Problem 1.18 The two-dimensional exponential distribution may be defined as follows. Let $\beta > 0$ and $\delta > 0$, and let $Z_1(t;\beta)$, $Z_2(t;\beta)$ and $Z_{12}(t;\delta)$ be independent Poisson processes with parameters as indicated. Consider a system consisting of two components. An event in the process $Z_1(t;\beta)$ causes component no 1 to fail. Similarly, an event in the process $Z_2(t;\beta)$ causes component no 2 to fail, while an event in the process $Z_{12}(t;\delta)$ causes both components to fail. This two-component system is assumed to continue functioning as long as at least one component is functioning. If X and Y denote the time elapsed until failure for respectively component no 1 and component no 2, we say that $(X,Y)^\top$ follows a two-dimensional exponential distribution with parameters β and δ, and we write $(X,Y)^\top \sim E_2(\beta,\delta)$.

1. Show that

$$\begin{aligned}\bar{F}(x,y) &= P_{(\beta,\delta)}(X > x, Y > y)\\ &= e^{-\beta(x+y)-\delta\max(x,y)}.\end{aligned}$$

2. If $(X,Y)^\top \sim E_2(\beta,\delta)$, the probability of the event $\{X = Y\}$ is positive, and is given by

$$P_{(\beta,\delta)}(X = Y) = \frac{\delta}{2\beta+\delta}.$$

(You do not need to show this result.) Let $\mathcal{B}$ denote the Borel σ-algebra in $\mathbf{R}_+^2$ and let μ_2 denote the Lebesgue measure on $(\mathbf{R}_+^2, \mathcal{B})$. From the results above, we find that the probability measure $P_{(\beta,\delta)}$ is not absolutely continuous with respect to μ_2. Let $\Delta = \left\{(x,y)^\top : x = y, 0 < x < \infty\right\}$ and $\upsilon(B) = \mu_1(B \cap \Delta)$, $B \in \mathcal{B}$, where μ_1 denotes the Lebesgue measure on $\mathbf{R}$. Let $\mu = \mu_2 + \upsilon$. Show that $P_{(\beta,\delta)}$ has the following probability density function with respect to μ,

$$\frac{dP_{(\beta,\delta)}}{d\mu}(x,y) = \begin{cases} \beta(\beta+\delta)\bar{F}(x,y) & if \quad (x,y) \notin \Delta \\ \delta\bar{F}(x,x) & if \quad (x,y) \in \Delta. \end{cases}$$

3. Let $\mathcal{P}$ denote the set of two-dimensional exponential distributions,

$$\mathcal{P} = \left\{P_{(\beta,\delta)} : \beta > 0, \delta > 0\right\}.$$

Show that $\mathcal{P}$ is an exponential family of order 3, with minimal representation of the form

$$\frac{dP_{(\beta,\delta)}}{d\mu}(x,y) = \beta(\beta+\delta)e^{-\beta(x+y)-\delta\max(x,y)+\left(\log\frac{\delta}{\beta(\beta+\delta)}\right)1_\Delta(x,y)},$$

where 1_Δ denotes the indicator function for the set Δ.

4. Let P_0 be the probability measure corresponding to $\beta = 1$ and $\delta = 1$, and let

$$T(x,y) = \left\{x+y, \max(x,y), 1_\Delta(x,y)\right\}^\top.$$

Let $\mathcal{Q}$ denote the full exponential family generated by $[P_0, T]$. Write $\mathcal{Q}$ on exponential form, and find the corresponding canonical parameter domain. Show that $\mathcal{P}$ is contained in $\mathcal{Q}$ and that $\mathcal{P}$ does not have open kernel. Find also the expectation and variance for the two failure times X and Y of the two-component system above.

5. Let $(X_1, Y_1)^\top, \ldots, (X_n, Y_n)^\top$ denote n independent and identically distributed observations from the distribution $E_2(\beta, \delta)$. Find the support and the convex support for the minimal canonical statistic

$$T_n = \left\{ \sum_{i=1}^n (X_i + Y_i), \sum_{i=1}^n \max(X_i, Y_i), \sum_{i=1}^n 1_\Delta(X_i, Y_i) \right\}^\top.$$

Finally, describe the set of observations for which the maximum likelihood estimator for the parameters of $\mathcal{Q}^{(n)}$ exists.

Problem 1.19 Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be an exponential family of order 2 with minimal representation

$$\frac{dP_\theta}{d\mu}(x) = a(\theta) b(x) e^{\theta_1 x + \theta_2 T(x)},\ \theta = (\theta_1, \theta_2)^\top \in \Theta \subseteq \mathbf{R}^2,$$

where $x \in \mathbf{R}$ and μ is a dominating σ-finite measure on ($\mathbf{R}$,$\mathcal{B}$).

1. Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \sim P_\theta$, and let $X_+ = \sum_{i=1}^n X_i$. Assume that the distribution for X_+ has probability density function of the form

$$\frac{dP_{\theta X_+}}{d\mu}(y) = a_1(\theta) b_1(y) e^{\theta_1 y + \theta_2 S(y)},\ \theta = (\theta_1, \theta_2)^\top \in \Theta.$$

Show that for $\theta = (\theta_1, \theta_2)^\top \in \Theta$ and $\alpha = (\alpha_1, \alpha_2)^\top \in \Theta$, we have the following conditional probability density function

$$\frac{dP_\theta^n(\sum_{i=1}^n T(X_i)|X_+)}{dP_\alpha^n(\sum_{i=1}^n T(X_i)|X_+)}(v|u) = \frac{a^n(\theta) a_1(\alpha)}{a^n(\alpha) a_1(\theta)} e^{(\theta_2 - \alpha_2)(v - S(u))}.$$

2. Show that, if $\left\{\theta_2 | \exists \theta_1 : (\theta_1, \theta_2)^\top \in \Theta\right\}$ contains an open set, then X_+ and $\sum_{i=1}^n T(X_i) - S(X_+)$ are independent.

Problem 1.20 Let $(\mu, \lambda)^\top \in (0, \infty)^2$, and consider the inverse Gaussian distribution $N^-(\mu, \lambda)$, defined by the probability density function with respect to Lebesgue measure on $\mathbf{R}$

$$f(x; \mu, \lambda) = \left(\frac{\lambda}{2\pi}\right)^{1/2} x^{-3/2} \exp\left[-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right], \quad x > 0.$$

1. Show that the family $\mathcal{P}^{(n)} = \left\{N^-(\mu,\lambda)^{(n)} : (\mu,\lambda)^\top \in (0,\infty)^2\right\}$, corresponding to n independent and identically distributed observations $X_1,\ldots,X_n$, is an exponential family of order 2 with canonical statistic

$$(X_+, S)^\top = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^{-1}\right)^\top.$$

2. Discuss maximum likelihood estimation of the mixed parameter $(\mu,\lambda)^\top$ based on $X_1,\ldots,X_n$. Show that if the maximum likelihood estimator $(\hat{\mu},\hat{\lambda})^\top$ exists, it is given by

$$(\hat{\mu},\hat{\lambda})^\top = \left(\bar{X}_+, \frac{n}{S - n\bar{X}_+^{-1}}\right)^\top,$$

where $\bar{X}_+ = X_+/n$.

3. Show that the Laplace transform of $X \sim N^-(\mu,\lambda)$ is

$$d(t) = e^{\sqrt{\psi\lambda}-\sqrt{(\psi-2t)\lambda}},\ t \le \frac{\psi}{2},$$

where we have $\psi = \lambda/\mu^2$. Use this result to show that $\hat{\mu} \sim N^-(\mu, n\lambda)$.

4. Show, possibly by using the result of Exercise 1.19, that $\hat{\mu}$ and $\hat{\lambda}$ are independent, and that

$$\hat{\lambda}^{-1} \sim (n\lambda)^{-1}\chi^2(n-1).$$

Hint: It may be useful to remember that the Laplace transform for a $\chi^2(n-1)$ distribution is

$$l(\zeta) = (1-2\zeta)^{-\frac{n-1}{2}},\ (\zeta < \frac{1}{2}).$$

Chapter 2

SUFFICIENCY AND ANCILLARITY

In the following we will describe in a detailed way the classical concepts of sufficiency, completeness and ancillarity. Such concepts will be used and extended in the next chapter, on inferential separation, where we will also discuss the statistical interpretation of these concepts in a more detailed way.

In Section 2.1 we will present the classical concepts of sufficiency and completeness. We will use several results on σ-algebras which will be reviewed throughout the section. In Section 2.2 we deal with the concept of ancillarity and we prove Basu's Theorem, which will give us ways to characterize ancillary statistics, as well as a first application of the concept. Section 2.3 deals with the extension of the classical concepts of sufficiency and ancillarity, which will allow us to interpret more clearly the concept of completeness. It will be especially useful to relate the concepts of sufficiency and ancillarity for inference functions defined in Chapter 4 with the classical concepts developed below.

2.1 Sufficiency

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model. Suppose that we wish to make inference on $\mathcal{P}$, based on the observation x. In many cases we use a statistic, *i.e.*, a measurable function $t : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ (where $\mathcal{B}$ is a σ-algebra of $\mathcal{Y}$). Roughly, we say that the statistic $T = t(X)$ is sufficient if it contains "all the information" that x gives about $\mathcal{P}$. In this section we will give a precise, theoretical formulation of this concept.

The concept of sufficiency, in the sense stated above, was introduced by Fisher in the beginning of the 20s. But only at the end of the 40s and the beginning of the 50s did Halmos, Savage and Bahadur make the general formulation and mathematical analysis of this concept. (Halmos and Savage, 1949 ; Bahadur, 1954).

In this chapter we will impose some restrictions compared with the general theory of sufficiency. We will assume that the fundamental σ-algebra, $\mathcal{A}$, is separable, and that the class $\mathcal{P}$ of probability measures is dominated by a σ-finite measure. Such restrictions are not

severe, since they naturally occur in many applications. For the case without a dominating measure see Halmos and Savage (1949) and Bahadur (1954, 1955).

Recall that a σ-algebra is called separable if it has a countable generator (*i.e.*, it is generated by a class of countable sets). We will later show that any separable σ-algebra is generated by a random variable, *i.e.*, a real measurable function. For this, see Theorem 2.16, which we will leave for the end of the section since it is a technical result.

We will assume in the following that for all $P \in \mathcal{P}$ there exists a regular conditional probability of P given T. Since we are dealing with more than one probability measure, we will use an index to determine which probability is being used to define the conditional expectation, for example, $\mathrm{E}_P(X|T)$.

2.1.1 Three lemmas

First we will show some lemmas, that will be useful to develop the theory of sufficiency. In order to fix the notation, let $(\mathcal{X}, \mathcal{A})$ be a given measurable space, where $\mathcal{A}$ is a separable σ-algebra. Let P be a probability measure on $(\mathcal{X}, \mathcal{A})$. We define $\mathcal{N}(P) = \{A \in \mathcal{A} : P(A) = 0\}$ and $\mathcal{C} \vee \mathcal{N}(P)$ as the smallest σ-algebra that contains $\mathcal{C}$ and $\mathcal{N}(P)$, where $\mathcal{C}$ is a sub-σ-algebra of $\mathcal{A}$.

Lemma 2.1 $\mathcal{C} \vee \mathcal{N}(P) = \{A \in \mathcal{A} : I_A = E_P(I_A|\mathcal{C}) \ [P]\}$.

Proof: Let $\mathcal{D} = \{A \in \mathcal{A} : I_A = \mathrm{E}_P(I_A|\mathcal{C}) \ [P]\}$. Note that $\mathcal{D}$ is a σ-algebra. $\mathcal{D}$ contains $\mathcal{C}$ and $\mathcal{N}(P)$. Hence $\mathcal{C} \vee \mathcal{N}(P) \subseteq \mathcal{D}$.

We will show that $\mathcal{C} \vee \mathcal{N}(P) \supseteq \mathcal{D}$. Let $A \in \mathcal{D}$. Define $\phi = \mathrm{E}_P(I_A|\mathcal{C})$. Hence $A = [\{\phi = 1\} \cup \{A \setminus \{\phi = 1\}\}] \setminus [\{\phi = 1\} \setminus A]$. Then $A \in \mathcal{C} \vee \mathcal{N}(P)$, because $\{\phi = 1\} \in \mathcal{C}$, $A \setminus \{\phi = 1\} \in \mathcal{N}(P)$ and $\{\phi = 1\} \setminus A \in \mathcal{N}(P)$. □

Lemma 2.2 *Let $X : \mathcal{X} \to \mathbb{R}$ be a P-integrable random variable. Suppose that $\sigma(X) \subseteq \mathcal{C} \vee \mathcal{N}(P)$. Then $X = E_P(X|\mathcal{C})$, $[P]$.*

Proof: Suppose that $\mathrm{E}_P(X|\mathcal{C})$ is a version of $\mathrm{E}_P(X|\mathcal{C} \vee \mathcal{N}(P))$. Hence $X = \mathrm{E}_P(X|\mathcal{C})$ $[P]$, because X is $\mathcal{C} \vee \mathcal{N}(P)$-measurable by hypothesis. So, it will be enough to show that $\mathrm{E}_P(X|\mathcal{C})$ is a version of $\mathrm{E}_P(X|\mathcal{C} \vee \mathcal{N}(P))$, that is we have to show that:

(i) $\mathrm{E}_P(X|\mathcal{C})$ is $\mathcal{C} \vee \mathcal{N}(P)$-measurable;

(ii) $\int_A \mathrm{E}_P(X|\mathcal{C})dP = \int_A X dP$, $\forall A \in \mathcal{C} \vee \mathcal{N}(P)$.

Part (i) is trivial. Now, the proof of (ii). Let $A \in \mathcal{C} \vee \mathcal{N}(P)$. We have that

$$\begin{aligned} \int_A \mathrm{E}_P(X|\mathcal{C})dP &= \int I_A \mathrm{E}_P(X|\mathcal{C})dP \\ &= \int \mathrm{E}_P(I_A \mathrm{E}_P(X|\mathcal{C})|\mathcal{C})dP \end{aligned}$$

$$
\begin{aligned}
&= \int \mathrm{E}_P(X|\mathcal{C})\mathrm{E}_P(I_A|\mathcal{C})dP \\
&= \int \mathrm{E}_P(\mathrm{E}_P(I_A|\mathcal{C})X|\mathcal{C})dP \\
&= \int \mathrm{E}_P(I_A|\mathcal{C})XdP = \int I_A X dP = \int_A X dP,
\end{aligned}
$$

since $\mathrm{E}_P(I_A|\mathcal{C}) = I_A$ $[P]$, $\forall A \in \mathcal{C} \vee \mathcal{N}(P)$ (by Lemma 2.1). □

Lemma 2.3 *Let $\mathcal{C}$ be a sub-σ-algebra of $\mathcal{A}$. Then there exists a separable σ-algebra $\mathcal{C}_0$ such that $\mathcal{C}_0 \subseteq \mathcal{C} \subseteq \mathcal{C}_0 \vee \mathcal{N}(P)$.*

Proof: Since $\mathcal{A}$ is separable there exists a sequence $\mathcal{A}_0 = \{A_n\}_{n=1}^{\infty}$ such that $\sigma(\mathcal{A}_0) = \mathcal{A}$. Let

$$\mathcal{C}_0 = \sigma\{\mathrm{E}_P(I_{A_n}|\mathcal{C}) : \quad n = 1, 2, \dots\}.$$

We claim $\mathcal{C}_0 \subseteq \mathcal{C}$, since for each n we have that $\mathrm{E}_P(I_{A_n}|\mathcal{C})$ is $\mathcal{C}$-measurable. Evidently $\mathcal{C}_0$ is separable because it is generated by a countable set of random variables. We only have to show that $\mathcal{C} \subseteq \mathcal{C}_0 \vee \mathcal{N}(P)$.

Let $\mathcal{A}_1 = \{A \in \mathcal{A} : \exists$ a $\mathcal{C}_0$-measurable version of $\mathrm{E}_P(I_A|\mathcal{C})\}$. Hence $\mathcal{A}_0 \subset \mathcal{A}_1 \subset \mathcal{A}$. Note that $\mathcal{A}_1$ is a σ-algebra, hence we have $\mathcal{A} = \sigma(\mathcal{A}_0) \subseteq \mathcal{A}_1$ and therefore $\mathcal{A} = \mathcal{A}_1$. To show that $\mathcal{C} \subseteq \mathcal{C}_0 \vee \mathcal{N}(P)$, let us take $A \in \mathcal{C}$, since $\mathcal{A}_1 = \mathcal{A}$ we have that

$$I_A = \mathrm{E}_P(I_A|\mathcal{C}) = \mathrm{E}_P(I_A|\mathcal{C}_0), \quad [P].$$

Using Lemma 2.1 we have that $A \in \mathcal{C}_0 \vee \mathcal{N}(P)$. □

2.1.2 Definitions

We now return to the situation where we have a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, with $\mathcal{A}$ separable, and where the class of probability measures $\mathcal{P}$ is dominated by a σ-finite measure μ. In the case that the conditional probability of P given T is the same for all the members of the class $\mathcal{P}$, it is reasonable to say that T contains all the information about $\mathcal{P}$. We have then the following definition.

Definition 2.4 *A statistic $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ is called a sufficient statistic if there exists a Markov kernel $\pi(A|t)$ from $\mathcal{A} \times \mathcal{Y}$ into $[0, 1]$, such that π is a regular conditional probability of P given T for any $P \in \mathcal{P}$, that is, π satisfies the following conditions:*

(i) $\pi(\cdot|t)$ *is a probability measure,* $\forall t \in \mathcal{Y}$*;*

(ii) $\pi(A|\cdot)$ *is* $\mathcal{B}$*-measurable,* $\forall A \in \mathcal{A}$*;*

(iii) $\int_B \pi(A|t)P_T(dt) = P(A \cap T^{-1}(B))$*,* $\forall A \in \mathcal{A}$*,* $B \in \mathcal{B}$ *and* $P \in \mathcal{P}$*.*

If T is sufficient, all the information about $\mathcal{P}$ is contained in the distribution P_T of T, in the sense that P can be determined using P_T, according to (iii). Note that if T is sufficient, $\mathrm{E}_P(X|T)$ does not depend on P.

At this stage the question of the existence of a minimal sufficient statistic arises, in the sense that a sufficient statistic is minimal if it is a simpler than any other sufficient statistic. In the following we will formalize this concept. To do this we will use the notation

$$\mathcal{N}(\mathcal{P}) = \{A \in \mathcal{A} : P(A) = 0, \quad \forall P \in \mathcal{P}\} \quad \text{and}$$

$$\mathcal{A}_1 \subseteq \mathcal{A}_2[\mathcal{P}],$$

if $\mathcal{A}_1$ and $\mathcal{A}_2$ are two sub-σ-algebras of $\mathcal{A}$ such that $\mathcal{A}_1 \subseteq \mathcal{A}_2 \vee \mathcal{N}(\mathcal{P})$.

Definition 2.5 *A statistic T_0 is called a minimal sufficient statistic if:*

(i) T_0 is sufficient and

(ii) $\sigma(T_0) \subseteq \sigma(T)[\mathcal{P}]$, for all sufficient Ts.

The following result will give us the motivation for the definition above. Let $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ and $T_0 : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}_0, \mathcal{B}_0)$. Hence $\sigma(T_0) \subseteq \sigma(T)[\mathcal{P}]$ if and only if, there exists a measurable function $f : (\mathcal{Y}, \mathcal{B}) \to (\mathcal{Y}_0, \mathcal{B}_0)$ such that $T_0 = f(T)[P]$, $\forall P \in \mathcal{P}$. This result can be proved using Lemma 2.2.

2.1.3 The case of equivalent measures

In the following we will assume that the measures of $\mathcal{P}$ are equivalent, in the sense that $\mathcal{N}(P) = \mathcal{N}(P')$, $\forall P$ and $P' \in \mathcal{P}$, that is, the sets with null measure are the same for any member of the family $\mathcal{P}$. Later on, we will consider a more general situation where we only assume that the members of $\mathcal{P}$ are dominated by a σ-finite measure μ in $(\mathcal{X}, \mathcal{A})$ (recall that such a hypothesis is also valid here).

Let $P_0 \in \mathcal{P}$. We have that all measures $P \in \mathcal{P}$ are P_0-absolutely continuous. We can then refer to dP/dP_0, as a version of the density of P with respect to P_0.

Theorem 2.6 *A statistic T is sufficient if and only if, for all $P \in \mathcal{P}$, there exists a T-measurable version of dP/dP_0.*

Proof: Let T be sufficient and let π be a Markov kernel such that π is a regular conditional probability of P given T, $\forall P \in \mathcal{P}$ (according to the definition). Define $U : \mathcal{X} \to I\!R$ by

$$U(x) = \int_{\mathcal{X}} dP/dP_0(y)\pi(dy|T(x)) = \mathrm{E}_0[dP/dP_0|T](x),$$

where $\mathrm{E}_0 = \mathrm{E}_{P_0}$.

We will show that U is a T-measurable version of dP/dP_0. The fact that U is T-measurable follows from the "usual proof" of measure theory. We will show that $\int_A U dP_0 = P(A)$, $\forall A \in \mathcal{A}$.

For each $A \in \mathcal{A}$ we have that

$$\begin{aligned}
\int_A U dP_0 &= \int_{\mathcal{X}} I_A \mathrm{E}_0(dP/dP_0|T) dP_0 \\
&= \int_{\mathcal{X}} \mathrm{E}_0(I_A \mathrm{E}_0(dP/dP_0|T)|T) dP_0 \\
&= \int_{\mathcal{X}} \mathrm{E}_0(dP/dP_0|T)\mathrm{E}_0(I_A|T) dP_0 \\
&= \int_{\mathcal{X}} \mathrm{E}_0[\mathrm{E}_0(I_A|T)\frac{dP}{dP_0}|T] dP_0 \\
&= \int_{\mathcal{X}} \mathrm{E}_0(I_A|T)\frac{dP}{dP_0} dP_0 = \int_{\mathcal{X}} \mathrm{E}_0(I_A|T) dP = P(A).
\end{aligned}$$

In the last step we used that $\mathrm{E}_0(I_A|T)$ is also a conditional expectation of P given T.

We will assume now that there exists a T-measurable version of dP/dP_0. Let

$$\pi(A|t) = P_0(A|T = t), \quad \forall A \in \mathcal{A}, \quad t \in \mathcal{Y}.$$

Hence, π is a Markov kernel, and if

$$\int_B \pi(A|t) P_T(dt) = P(A \cap T^{-1}(B)), \quad \forall A \in \mathcal{A}, \quad \forall B \in \mathcal{B} \text{ and } \forall P \in \mathcal{P},$$

the theorem will be proved.

Now, as $\frac{dP}{dP_0}$ is T-measurable, we have that

$$\begin{aligned}
\int_B \pi(A|t) P_T(dt) &= \int_{\mathcal{Y}} I_B(t) \pi(A|t) P_T(dt) \\
&= \int_{\mathcal{X}} I_{T^{-1}(B)}(x) P_0(A|T = T(x)) P(dx) \\
&= \int I_{T^{-1}(B)} \mathrm{E}_0(I_A|T)\frac{dP}{dP_0} dP_0 \\
&= \int \mathrm{E}_0[I_{T^{-1}(B)} I_A \frac{dP}{dP_0}|T] dP_0 \\
&= \int I_{A \cap T^{-1}(B)} \frac{dP}{dP_0} dP_0 \\
&= P(A \cap T^{-1}(B)).
\end{aligned}$$

□

Note that if f_P and g_P are versions of dP/dP_0 and $\mathcal{C} = \sigma(f_P : P \in \mathcal{P})$ and $\mathcal{D} = \sigma(g_P : P \in \mathcal{P})$, then it is easy to prove that $\mathcal{C} = \mathcal{D}[\mathcal{P}]$. This is especially interesting in view of the following theorem which gives a characterization of a minimal sufficient statistic.

Theorem 2.7 *(Bahadur) A statistic T_0 is minimal sufficient if and only if, $\sigma(T_0) = \sigma(\frac{dP}{dP_0} : P \in \mathcal{P})$ $[\mathcal{P}]$.*

Proof: Let $\mathcal{C}_0 = \sigma(T_0)$ and $\mathcal{C} = \sigma(f_P : P \in \mathcal{P})$, where f_P is any version of dP/dP_0, $P \in \mathcal{P}$ (f_P fixed).

We will first assume that T_0 minimal sufficient. Since in this case T_0 is sufficient, by Theorem 2.6 there exists a T_0-measurable version of dP/dP_0, $\forall P \in \mathcal{P}$. Therefore, $\mathcal{C} \subseteq \mathcal{C}_0[\mathcal{P}]$.

To prove that $\mathcal{C}_0 \subseteq \mathcal{C}[\mathcal{P}]$ we use that, by Lemma 2.3, there exists a separable σ-algebra $\mathcal{C}_1$, such that

$$\mathcal{C}_1 \subseteq \mathcal{C} \subseteq \mathcal{C}_1 \vee \mathcal{N}(\mathcal{P}).$$

Let T_1 be a statistic such that $\sigma(T_1) = \mathcal{C}_1$ (such a statistic exists, by Theorem 2.16, whose proof uses none of the results of this section). By Lemma 2.2 we have that

$$f_P = \mathrm{E}_0(f_P|T_1), \quad [P_0] \quad \forall P \in \mathcal{P}.$$

Therefore, $\mathrm{E}_0(f_P|T_1)$ is a version of dP/dP_0. Since $\mathrm{E}_0(f_P|T_1)$ is T_1-measurable, we have by Theorem 2.6 that T_1 is sufficient. T_0 is minimal sufficient, and therefore

$$\sigma(T_0) \subseteq \sigma(T_1)[\mathcal{P}],$$

so that

$$\mathcal{C}_0 \subseteq \mathcal{C}[\mathcal{P}].$$

Hence, $\mathcal{C}_0 = \mathcal{C}[\mathcal{P}]$, as we wanted to show.

We will now assume that $\mathcal{C}_0 = \mathcal{C}[\mathcal{P}]$. By Lemma 2.2 we have that $\mathrm{E}_0(f_P|T_0)$ is a version of dP/dP_0, $\forall P \in \mathcal{P}$. Then T_0 is sufficient, by Theorem 2.6. We only have to show that

$$\mathcal{C}_0 \subseteq \sigma(T)[\mathcal{P}],$$

for any T sufficient. Now, since T is sufficient, we can choose dP/dP_0 T-measurable. Then $\mathcal{C} \subseteq \sigma(T)[\mathcal{P}]$, but since $\mathcal{C}_0 \subseteq \mathcal{C}[\mathcal{P}]$, we obtain that $\mathcal{C}_0 \subseteq \sigma(T)[\mathcal{P}]$. □

Corollary 2.8 *Under the previous assumptions, there exists a minimal sufficient statistic.*

Proof: Let $\mathcal{C} = \sigma(f_P : P \in \mathcal{P})$, where f_P is any version of dP/dP_0, $P \in \mathcal{P}$. By Lemma 2.3, there exists a separable σ-algebra $\mathcal{C}_1$ such that

$$\mathcal{C}_1 \subseteq \mathcal{C} \subseteq \mathcal{C}_1 \vee \mathcal{N}(\mathcal{P}).$$

Let T_1 be such that $\sigma(T_1) = \mathcal{C}_1$. Hence $\sigma(T_1) = \mathcal{C}[\mathcal{P}]$, and by Theorem 2.7, we have that T_1 is minimal sufficient. □

2.1.4 The general case

Let us consider now the case where we only assume that the measures of $\mathcal{P}$ are dominated by a σ-finite measure μ (*i.e.*, we do not suppose any more, as in Theorem 2.7 and Corollary 2.8, that $\mathcal{N}(P) = \mathcal{N}(P')$, $\forall P, P' \in \mathcal{P}$). Using Theorem 2.9 we will show some theorems that characterize minimal sufficient statistics in this more general context.

Theorem 2.9 *(Halmos-Savage's Theorem) Let $\mathcal{P}$ be a class of probability measures dominated by a σ-finite measure. Then $\mathcal{P}$ has a countable equivalent subclass.*

Proof: We will assume that $\mathcal{P} \ll \mu$. We will show that there exists a countable subclass $\{P_m\}_{m=1}^{\infty}$ of $\mathcal{P}$ such that for all $A \in \mathcal{A}$ we have

$$P_m(A) = 0 \quad \forall m \Rightarrow P(A) = 0, \quad \forall P \in \mathcal{P}.$$

We can assume without loss of generality that μ is a probability measure. Define

$$\begin{aligned} A_P &= \{dP/d\mu > 0\}, \\ \underline{B} &= \{B \in \mathcal{A} : \exists P \text{ such that } B \subseteq A_P \text{ and } \mu(B) > 0\} \text{ and} \\ \underline{C} &= \{C : C = \bigcup_{n=1}^{\infty} B_n, B_n \in \underline{B}\}. \end{aligned}$$

Let $s = \sup\{\mu(C) : C \in \underline{C}\}$. Since $s \leq 1$ (since μ is a probability) there exists a sequence $\{C_n\}_{n=1}^{\infty}$ such that $\mu(C_n) \nearrow s$. Writing $C = \cup_{n=1}^{\infty} C_n$, we have $\mu(C) = s$ and $C \in \underline{C}$.

Hence, there exists a sequence $\{B_m\}_{m=1}^{\infty}$, of elements of $\underline{B}$ such that $C = \bigcup_{m=1}^{\infty} B_m$. Since $B_m \in \underline{B}$, there exists P_m such that $B_m \subseteq A_{P_m}$ and $\mu(B_m) > 0$.

Now, let $A \in \mathcal{A}$ be such that $P_m(A) = 0$ $\forall m$, and suppose that there exists $P \in \mathcal{P}$ such that $P(A) > 0$. We will show that this will lead to a contradiction. We will assume that $A \subseteq A_P$, since otherwise A could be substituted by $A \cap A_P$. Since $P \ll \mu$, we have $\mu(A) > 0$. We also have that $A \in \underline{B}$, and therefore $C \cup A \in \underline{C}$. And since $0 = P_m(A \cap B_m) = \int_{A \cap B_m} (dP_m/d\mu)\, d\mu$, we have that $\mu(A \cap B_m) = 0$, with $dP_m/d\mu > 0$ in B_m. Therefore,

$$\mu(A \cap C) \leq \sum_{m=1}^{\infty} \mu(A \cap B_m) = 0.$$

Hence, $\mu(A \cap C^c) = \mu(A) > 0$. Now, using that $C \cup A \in \underline{C}$ we obtain

$$s = \mu(A \cup C) = \mu(C) + \mu(A \cap C^c) > \mu(C) = s,$$

which gives the contradiction. □

Let $\{P_m\}_{m=1}^{\infty}$ be the countable equivalent subset of $\mathcal{P}$, whose existence is given by Theorem 2.9. Let $P_0 = \sum_{m=1}^{\infty} c_m P_m$, where $\{c_m\}_{m=1}^{\infty}$ is a sequence of positive real numbers such that $\sum_{m=1}^{\infty} c_m = 1$. Hence we have that $\mathcal{N}(P_0) = \mathcal{N}(\mathcal{P})$, and therefore $\mathcal{P} \ll P_0$. Using this definition of P_0, we obtain the following generalization of Theorems 2.6, 2.7 and Corollary 2.8, for a situation where we only assume that $\mathcal{P}$ is dominated.

Theorem 2.10 *If $\mathcal{P} \ll \mu$ and P_0 is defined as above, then:*

(i) A statistic T is sufficient if and only if for each $P \in \mathcal{P}$ there exists a T-measurable version of dP/dP_0;

(ii) A statistic T_0 is minimal sufficient if and only if $\sigma(T_0) = \sigma(dP/dP_0 : P \in \mathcal{P})[\mathcal{P}]$;

(iii) A minimal sufficient statistic exists.

Proof: We will consider $\mathcal{P}' = \{P' : P' = P_0 \text{ or } P' = \frac{1}{2}(P_0 + P) \text{ where } P \in \mathcal{P}\}$. The measures in $\mathcal{P}'$ are equivalent, which is easy to see, using that $\mathcal{P} \ll P_0$.

Therefore, we can use Theorem 2.6, 2.7 and Corollary 2.8 on $\mathcal{P}'$. We then have that

$$\frac{d}{dP_0}\left[\frac{1}{2}(P_0 + P)\right] = \frac{1}{2}\left(1 + \frac{dP}{dP_0}\right).$$

Therefore, there exists a T-measurable version of dP/dP_0, $\forall P \in \mathcal{P}$ if and only if there exists a T-measurable version of dP'/dP_0, $\forall P' \in \mathcal{P}'$, and

$$\sigma\left(\frac{dP}{dP_0} : P \in \mathcal{P}\right) = \sigma\left(\frac{dP'}{dP_0} : P' \in \mathcal{P}'\right).$$

To prove the first part of the theorem, it is enough to show that T is sufficient for $\mathcal{P}$ if and only if T is sufficient for $\mathcal{P}'$. Hence, suppose first that T is sufficient for $\mathcal{P}$. Let $\pi(A|t)$ be a regular conditional probability of P given T, $\forall P \in \mathcal{P}$. Hence $\pi(A|t)$ is also a regular conditional probability for P_0 given T, since for $A \in \mathcal{A}$ and $B \in \mathcal{B}$ we have

$$\begin{aligned}
\int_B \pi(A|t)P_{0T}(dt) &= \int_B \pi(A|t)(\sum_{m=1}^{\infty} c_m P_{mT})(dt) \\
&= \sum_{m=1}^{\infty} c_m \int_B \pi(A|t)P_{mT}(dt) = \sum_{m=1}^{\infty} c_m P_m(A \cap t^{-1}(B)) \\
&= P_0(A \cap t^{-1}(B)).
\end{aligned}$$

In a similar way we can show that $\pi(A|t)$ is a regular conditional probability of P' given T $\forall P' \in \mathcal{P}'$. Therefore T is sufficient for $\mathcal{P}'$.

Conversely, if T is sufficient for $\mathcal{P}'$, then there exists a Markov kernel $\pi(A|t)$ which is a regular conditional probability of P' given T, $\forall P' \in \mathcal{P}'$. Hence we have that $\pi(A|t)$ is a regular conditional probability of P given T, $\forall P \in \mathcal{P}$, since if $P' = \frac{1}{2}(P_0 + P)$ then $P = 2P' - P_0$. For $A \in \mathcal{A}$ and $B \in \mathcal{B}$ we have that

$$\begin{aligned}
\int_B \pi(A|t)P_T(dt) &= \int_B \pi(A|t)(2P'_T - P_{0T})(dt) \\
&= 2\int_B \pi(A|t)P'_T(dt) - \int_B \pi(A|t)P_{0T}(dt) \\
&= 2P'(A \cap t^{-1}(B)) - P_0(A \cap t^{-1}(B)) \\
&= P(A \cap t^{-1}(B)).
\end{aligned}$$

Therefore, T is sufficient for $\mathcal{P}$. We conclude then that

$$\begin{aligned} T \text{ is sufficient for } \mathcal{P} &\iff T \text{ is sufficient for } \mathcal{P}' \\ &\iff \exists \text{ a } T\text{-measurable version of } \frac{dP'}{dP_0}, \quad \forall P' \in \mathcal{P}' \\ &\iff \exists \text{ a } T\text{-measurable version of } \frac{dP}{dP_0}, \quad \forall P \in \mathcal{P} \end{aligned}$$

where the second equivalence follows from Theorem 2.6. In this way, we have proved part (i) of the theorem.

We will show part (ii): Let T_0 be minimal sufficient for $\mathcal{P}$. Hence T_0 is sufficient for $\mathcal{P}$, and therefore for $\mathcal{P}'$. Let T be sufficient for $\mathcal{P}'$. Hence T is sufficient for $\mathcal{P}$ and, as T_0 is minimal sufficient for $\mathcal{P}$, we obtain that $\sigma(T_0) \subseteq \sigma(T)[\mathcal{P}]$. Since $\mathcal{N}(\mathcal{P}) = \mathcal{N}(\mathcal{P}')$, we have that $\sigma(T_0) \subseteq \sigma(T)[\mathcal{P}']$.

Therefore, we have proved that T_0 is minimal sufficient for $\mathcal{P}'$. In a similar way we can prove that T_0 is minimal sufficient for $\mathcal{P}$ if T_0 is minimal sufficient for $\mathcal{P}'$.

Item (ii) follows from the implications:

$$\begin{aligned} T_0 \text{ is minimal sufficient for } \mathcal{P} &\iff \\ T_0 \text{ is minimal sufficient for } \mathcal{P}' &\iff \\ \sigma(dP'/dP_0 : P' \in \mathcal{P}') = \sigma(T_0)[\mathcal{P}'] &\iff \\ \sigma(dP/dP_0 : P \in \mathcal{P}) = \sigma(T_0)[\mathcal{P}], & \end{aligned}$$

where the second implication follows from Theorem 2.7.

Item (iii) of the theorem can be proved exactly as in Corollary 2.8. □

The theorem that follows gives a simple method to verify from the density $dP/d\mu$, if a given statistic T is or is not sufficient. This result is known as Fisher-Neyman's factorization criterion.

Theorem 2.11 *(Fisher-Neyman Criterion) Let $\mathcal{P}$ be dominated by a σ-finite measure μ. Then $T = t(X)$ is sufficient if and only if there exists a version of $dP/d\mu$ of the form*

$$\frac{dP}{d\mu}(x) = h_P[t(x)]k(x).$$

Proof: Let P_0 be as in Theorem 2.10. We will assume that T is sufficient. Let us consider a T-measurable version of dP/dP_0, *i.e.*, $\frac{dP}{dP_0}(x) = h_P(t(x))$. Hence,

$$\frac{dP}{d\mu}(x) = \frac{dP}{dP_0}(x)\frac{dP_0}{d\mu}(x) = h_P(t(x))k(x).$$

Conversely, we will assume that $(dP/d\mu)(x)$ is in the form given by the theorem. Hence

$$\frac{dP_0}{d\mu}(x) = \sum_{m=1}^{\infty} c_m h_{P_m}(t(x))k(x)$$

and

$$\frac{dP}{dP_0}(x) = \frac{(dP/d\mu)(x)}{(dP_0/d\mu)(x)} = \frac{h_P(t(x))}{\sum_{m=1}^{\infty} c_m h_{P_m}(t(x))}, \quad [P_0]$$

since $\sum_{m=1}^{\infty} c_m h_{P_m}(t(x))k(x) > 0$ $[P_0]$, which shows that dP/dP_0 is T-measurable. □

2.1.5 Completeness

We now discuss the important concept of completeness.

Definition 2.12 *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model and let $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ be a statistic.*

(i) T is called complete with respect to $\mathcal{P}$ if for any measurable function $f : \mathcal{Y} \to \mathbb{R}$ satisfying

$$E_P(f(T)) = \int_Y f(t) P_T(dt) = 0, \quad \forall P \in \mathcal{P}, \tag{2.1}$$

we have $f \circ T = 0$ $[\mathcal{P}]$, i.e., $P(f \circ T = 0) = P_T(f = 0) = 1$ for any $P \in \mathcal{P}$;

(ii) T is called boundedly complete if for any bounded and measurable function $f : \mathcal{Y} \to \mathbb{R}$ satisfying (2.1), we have $f \circ T = 0$ $[\mathcal{P}]$.

Note that in the definition we implicitly assume that f is P_T-integrable, $\forall P \in \mathcal{P}$, *i.e.*,

$$\int f^+ dP_T < \infty \text{ and } \int f^- dP_T < \infty, \quad \forall P \in \mathcal{P}.$$

Minimal canonical statistics of exponential families are complete under weak regularity conditions, as the following lemma shows.

Lemma 2.13 *Let $\mathcal{P}$ be an exponential family and T a minimal canonical statistic of $\mathcal{P}$. If $\mathcal{P}$ has open kernel, then T is complete with respect to $\mathcal{P}$.*

Proof: Let

$$\frac{dP_\theta}{d\mu}(x) = a(\theta) b(x) e^{\theta \cdot t(x)}, \quad x \in \mathcal{X}, \quad \theta \in \Theta \subseteq \mathbb{R}^k$$

be a minimal representation. Since $\mathcal{P}$ has an open kernel, then $\text{int}\,\Theta \neq \emptyset$. Without loss of generality, we can assume that $0 \in \Theta$. Let f be such that

$$\int_{\mathcal{Y}} f(z) P_{\theta T}(dz) = 0, \quad \forall \theta \in \Theta.$$

We have,

$$\begin{aligned} \int_{\mathcal{Y}} f(z) P_{\theta T}(dz) &= \int_{\mathcal{X}} f(t(x)) a(\theta) b(x) e^{\theta \cdot t(x)} \mu(dx) \\ &= \frac{a(\theta)}{a(0)} \int_{\mathcal{X}} e^{\theta \cdot t(x)} f(t(x)) P_0(dx), \end{aligned}$$

and then

$$\int_{\mathcal{X}} e^{\theta \cdot t(x)} f^+(t(x)) P_0(dx) = \int_{\mathcal{X}} e^{\theta \cdot t(x)} f^-(t(x)) P_0(dx).$$

Define the measures v_- and v_+ by

$$v_\pm(B) = \int_{t^{-1}(B)} f^\pm(t(x)) P_0(dx), \quad \forall B \in \mathcal{B}(\mathbb{R}^k).$$

Hence, v_+ and v_- are finite measures on $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k))$ that satisfy

$$\int_{\mathbb{R}^k} e^{\theta \cdot z} v_+(dz) = \int_{\mathbb{R}^k} e^{\theta \cdot z} v_-(dz), \quad \forall \theta \in \Theta.$$

That is, the Laplace transforms of v_+ and v_- are identical on Θ, which contains a k-dimensional open ball. Hence, by the inversion theorem for the Laplace transform we conclude that $v_+ = v_-$. Since the measures in $\mathcal{P}$ are equivalent, we conclude that

$$f^+ \circ t = f^- \circ t \qquad [P_{\theta_0}], \quad \forall \theta \in \Theta,$$

or

$$f \circ t = 0 \quad [P_\theta], \quad \forall \theta \in \Theta,$$

as we wanted to show. □

The proof of the lemma above uses the theorem on the uniqueness of the Laplace transform. In other cases, when $\mathcal{P}$ is not an exponential family, it is sometimes possible to show that a given statistic T is complete with respect to $\mathcal{P}$, using a theorem on uniqueness of another transformation, say, the distribution function.

In the theorem that follows we will show that if a statistic is complete, then, under quite general conditions, sufficiency implies minimal sufficiency. Note that the concept of a complete statistic is widely used. Besides the use in connection with sufficiency, this concept is widely used together with ancillarity, in the theory of estimators with minimal variance, as well as in the theory of tests.

Theorem 2.14 *Let $\mathcal{P}$ be dominated by a σ-finite measure μ, and let T be sufficient and complete with respect to $\mathcal{P}$. Then T is minimal sufficient.*

Proof: Let P_0 be defined as in Theorem 2.10 and $\mathcal{C} = \sigma(dP/dP_0 : P \in \mathcal{P})$. By Theorem 2.10, it is enough to show that $\sigma(T) = \mathcal{C}[\mathcal{P}]$. Since T is sufficient we have $\mathcal{C} \subseteq \sigma(T)[\mathcal{P}]$, therefore we only have to show that $\sigma(T) \subseteq \mathcal{C}[\mathcal{P}]$. Since $\mathcal{N}(\mathcal{P}) = \mathcal{N}(P_0)$, it is sufficient to show that $\sigma(T) \subseteq \mathcal{C} \vee \mathcal{N}(P_0)$, and according to Lemma 2.1 we have that

$$\mathcal{C} \vee \mathcal{N}(P_0) = \{A \in \mathcal{A} : I_A = \mathrm{E}_0[I_A|\mathcal{C}], \ [P_0]\}.$$

Let $A \in \sigma(T)$ and $\phi_A = \mathrm{E}_0[I_A|\mathcal{C}]$. We will show that:

(a) $I_A = \mathrm{E}_0[\phi_A|T]$ $[P_0]$

(b) $\mathrm{E}_0[\phi_A|T] = \phi_A$ $[P_0]$.

The theorem now follows from (a) and (b).

To show (a), let $g(x) = I_A(x) - \mathrm{E}_0[\phi_A|T](x)$. The function g is T-measurable, *i.e.*, g is of the form $f \circ T$. Moreover we have that

$$\begin{aligned}
\mathrm{E}_P f(T) &= \int (I_A - \mathrm{E}_0[\phi_A|T])dP = P(A) - \int \mathrm{E}_0[\phi_A|T]dP \\
&= P(A) - \int \mathrm{E}_P[\phi_A|T]dP = P(A) - \int \phi_A dP \\
&= P(A) - \int \mathrm{E}_0[I_A|\mathcal{C}]dP \\
&= P(A) - \int \mathrm{E}_P[I_A|\mathcal{C}]dP = P(A) - \int I_A dP = 0,
\end{aligned}$$

where the second equality holds because T is sufficient and the last equality is easily obtained if we substitute I_A by any measurable function $X : \mathcal{X} \to I\!R$ (see Problem 2.6).

To show (b), note that,

$$\sigma(\phi_A) \subseteq \mathcal{C} \subseteq \sigma(T) \vee \mathcal{N}(\mathcal{P}) = \sigma(T) \vee \mathcal{N}(P_0).$$

Then (b) follows from Lemma 2.2. □

Theorem 2.15 *Let $\mathcal{P}$ be an exponential family of order k, and let*

$$\frac{dP}{d\mu}(x) = a(P)b(x)e^{\alpha(P)\cdot t(x)}$$

be a representation of $\mathcal{P}$. Then the canonical statistic T is sufficient, and if the representation is minimal then T is minimal sufficient.

Proof: The sufficiency is a consequence of Theorem 2.11. We will assume that the representation is minimal. Let $P_0 \in \mathcal{P}$ be arbitrary. Then

$$\frac{dP}{dP_0}(x) = \frac{a(P)}{a(P_0)} e^{\{\alpha(P)-\alpha(P_0)\}\cdot t(x)} = \tilde{a}(P)e^{\tilde{\alpha}(P)\cdot t(x)},$$

say. By minimality, there exist $P_1, \ldots, P_k$ such that $\tilde{\alpha}(P_1), \ldots, \tilde{\alpha}(P_k)$ are linearly independent vectors in $I\!R^k$. Since

$$\begin{pmatrix} \log \frac{dP_1}{dP_0}(x) \\ \vdots \\ \log \frac{dP_k}{dP_0}(x) \end{pmatrix} = \begin{pmatrix} \log \tilde{a}(P_1) \\ \vdots \\ \log \tilde{a}(P_k) \end{pmatrix} + \begin{pmatrix} \tilde{\alpha}(P_1)^\top \\ \vdots \\ \tilde{\alpha}(P_k)^\top \end{pmatrix} t(x),$$

and the matrix $(\tilde{a}(P_1)\dots\tilde{\alpha}(P_k))$ is non-singular. We have that T is a one-to-one function of $\{\log dP_j/dP_0 : j = 1,\dots,k\}$, and therefore of $\{\frac{dP_j}{dP_0} : j = 1,\dots,k\}$. Then

$$\sigma\left(\frac{dP_j}{dP_0} : j = 1,\dots,k\right) = \sigma(T),$$

and since $\sigma(\frac{dP}{dP_0} : P \in \mathcal{P}) \subseteq \sigma(T)[\mathcal{P}]$ we have

$$\sigma\left(\frac{dP}{dP_0} : P \in \mathcal{P}\right) = \sigma(T)[\mathcal{P}],$$

from which we can conclude that T is minimal sufficient by Theorem 2.7. □

2.1.6 A result on separable σ-algebras

We now show a technical result on separable σ-algebras, that has been used throughout the section. Note that the proof of this result is independent of the rest of the section.

Theorem 2.16 *Any separable σ-algebra is generated by a random variable,* i.e.*, a real measurable function $X : \mathcal{X} \to I\!R$.*

Proof: Let $\mathcal{A}$ be a separable σ-algebra generated by the sets $A_1, A_2, \dots$. Given $(\mathcal{X}, \mathcal{A})$ and $(I\!R, \mathcal{B}(I\!R))$ we define

$$X(w) = \sum_{n=1}^{\infty} N^{-n} I_{A_n}(w),$$

where $w \in \mathcal{X}$ and N is an even positive integer larger than 2. Evidently, $\sigma(X) \subseteq \mathcal{A}$. We have to prove that $\mathcal{A} \subset \sigma(X)$. To see this, it is sufficient to show that $\{A_n\}_{n=1}^{\infty} \subseteq \sigma(X)$, since this implies that $\sigma(\{A_n\}_{n=1}^{\infty}) \subseteq \sigma(X)$.

Let us consider the random variables

$$X_n(w) = \sum_{m=1}^{n} N^{-m} I_{A_m}(w) \text{ and } Y_n(w) = \sum_{m=n}^{\infty} N^{-m} I_{A_m}(w),$$

$n = 1, 2, \dots$. It is clear that $X(w) = X_{n-1}(w) + Y_n(w)$, for $n \in \mathbf{N}$, (letting $X_0(w) = 0$).

We have that

$$\begin{aligned} N^n X(w) &= \sum_{m=1}^{n-1} N^{-m+n} I_{A_m}(w) + I_{A_n}(w) + \sum_{m=n+1}^{\infty} N^{-m+n} I_{A_m}(w) \\ &= N^n X_{n-1}(w) + I_{A_n}(w) + N^n Y_{n+1}(w). \end{aligned}$$

Since we have that $N^n Y_{n+1}(w) \leq \sum_{k=0}^{\infty} N^{-k} = \frac{1}{N-1} \leq 1/3$, $N \geq 4$ and $N^n X_{n-1}(w) = 0$ or $N^n X_{n-1}(w) \geq N$, $(N \geq 4)$, and as $N^n X_{n-1}(w)$ is always even we have, with $f(x) = [x]$, that $w \in A_n \Leftrightarrow [N^n X(w)]$ is odd, that is, $A_n = g_n^{-1}(f^{-1}(U))$, where $g_n = N^n X$ and U is the set of odd numbers in $\mathbf{N}$. The function f is $\mathcal{B}/2^Z$ measurable, $f : I\!R \to Z$, g is $\mathcal{A}/\mathcal{B}$-measurable, $g : \mathcal{X} \to I\!R$, that is, $A_n \in \sigma(X)$ since $U \in 2^Z$. f is $\mathcal{B}/2^Z$-measurable because $f^{-1}(n) = \{x \in I\!R : [x] = n\} = [n, n+1) \in \mathcal{B}$. □

2.1.7 On the minimal sufficiency of the likelihood function

Consider the case where $\mathcal{P}$ is parametrized, $\mathcal{P} = \{\mathcal{P}_\omega : \omega \in \Omega\}$, let

$$q(x;\omega) = \frac{dP_\omega}{dP_0}(x)$$

and let r be the mapping which maps a point $x \in \mathcal{X}$ to the likelihood function

$$r(x) = q(x;\cdot).$$

Endowing the range space $\mathcal{R}^\mathcal{P}$ of r with the product σ-algebra $\mathcal{B}^\mathcal{P}$, where $\mathcal{B}$ is the Borel σ-algebra in $\mathcal{R}$, one obtains that r becomes measurable and that

$$\sigma\{r\} = \sigma\{q(\cdot;\omega) : \omega \in \Omega\}. \tag{2.2}$$

Here, and everywhere until the end of Example 2.22, equalities, inclusions, etc. are strict, *i.e.* not modulo null sets. This proposition represents one precise interpretation of the common phrase "the likelihood function is minimal sufficient". However, rather than this interpretation, the phrase reflects the useful fact that if T is a statistic generating the same partition of $\mathcal{X}$ as the mapping r , *i.e.*

$$T(x) = T(\tilde{x}) \Leftrightarrow q(x;\omega) = q(\tilde{x};\omega) \text{for every } \omega \in \Omega, \tag{2.3}$$

then, as a rule, T is minimal sufficient. That some regularity conditions are needed to ensure the minimal sufficiency of such a statistic T is illustrated by the next example.

Example 2.17 Let X_1 and X_2 be independent and normally distributed, $X_1 \sim N(0,1), X_2 \sim N(\omega,1)$ with $\omega \in \Omega = I\!R$. Then X_2 is minimal sufficient with respect to the family $\mathcal{P} = \{P_\omega : \omega \in \Omega\}$ of joint distributions of $X = (X_1, X_2)$. Taking $\mathcal{X} = I\!R^2$, X as the identity mapping on $\mathcal{X}$, a version of dP_ω/dP_0 is given by

$$q(x:\omega) = (1 - \delta(x_1 - \omega))e^{\omega x_2 - \frac{\omega^2}{2}}$$

where δ is the function of $I\!R$ which is 1 at the origin and 0 otherwise. Clearly,

$$q(x;\cdot) = q(x';\cdot) \Leftrightarrow x = x'$$

which means that X generates the same partition of $\mathcal{X}$ as does $r : x \to q(x;\cdot)$, although X is not minimal sufficient.

Theorem 2.18 *Suppose that T is a statistic which generates the same partition of $\mathcal{X}$ as the mapping $r : x \to q(x;\cdot)$. If either*

1) $\mathcal{P}$ is discrete

or

2) Ω is a subset of a Euclidean space and $q(x;\omega)$ is continuous in ω for each fixed x

then T is minimal sufficient.

Remark: The discrete case is straightforward to verify. If, in a given case, the conditions of the theorem are not fulfilled it may well be that minimal sufficiency can be established by the below method of proof which, as should be apparent, offers scope for considerable generalization.

Proof: For any mapping f on an arbitrary measure space $(E,\mathcal{C})$, let $\delta(f)$ denote the partition σ-algebra determined by f, *i.e.* the σ-algebra of those sets $C \in \mathcal{C}$ which are unions of elements of the partition of E generated by f. Clearly, $\delta(f) = \{C \in \mathcal{C} : C = f^{-1}(f(C))\}$ and, in case f is a measurable mapping,

$$\sigma(f) \subset \delta(f). \tag{2.4}$$

Under mild regularity conditions one has, in fact, that $\sigma(f) = \delta(f)$. This will be seen from the following proposition which is a special case of Theorem 3, p.145, in Hoffmann-Jørgensen (1990). □

Lemma 2.19 *Let E, F and G be Borel subsets of complete separable metric spaces, endowed with the Borel σ-algebras. Let f and g be measurable mappings from E into F and G, respectively, and suppose that the partition of E generated by f is finer than the partition generated by g. Then there exists a measurable mapping h from F into G such that $g = h \circ g$.*

Taking g to be the indicator function of an element of $\delta(f)$ one finds that this element belongs to $\sigma(f)$. Hence one has the following result.

Corollary 2.20 *Suppose f is a measurable mapping from a Borel subset of a complete, separable metric space into a complete, separable metric space. Then $\sigma(f) = \delta(f)$.*

It follows that, always,

$$\sigma(T) = \delta(T) = \delta(r) \tag{2.5}$$

and hence, in view of (2.2) and (2.4), that T is sufficient.

Now, suppose Ω is a subset of a Euclidean space and that $q(x:\cdot)$ is continuous on Ω for every $x \in \mathcal{X}$. Let Ω_0 be a dense subset of Ω and let r_0 be the mapping on $\mathcal{X}$ such that $r_0(x)$ is the restriction of $q(x;\cdot)$ to Ω_0. Then by continuity, $\sigma(r) = \sigma(r_0)$ and r_0 determines the same partition of $\mathcal{X}$ as r (and T). Therefore, on account of (2.5) and the Corollary,

$$\sigma(T) = \delta(r_0) = \sigma(r_0) = \sigma(r)$$

and since $\sigma(r)$ is minimal sufficient, so is T.

Corollary 2.21 *Suppose that T is a statistic which generates the same partition of $\mathcal{X}$ as the likelihood function $p(x:\cdot)$,* i.e.

$$T(x) = T(\tilde{x}) \Leftrightarrow cp(x:\omega) = \tilde{c}p(\tilde{x};\omega) \text{ for every } \omega \in \Omega$$

for some positive c and $\tilde{c}$ which do not depend on ω but may depend on x and $\tilde{x}$, respectively. If either

1) $\mathcal{P}$ is discrete

or

2) Ω is a subset of a Euclidean space and $p(x;\omega)$ is positive and continuous in ω for each fixed x then T is minimal sufficient.

Proof: Note that P_0 is of the form $\Sigma c_n P_{\omega_n}$ (by Halmos-Savage's Theorem). Taking the version of $dP_0/d\mu$ given by

$$p_0(x) = \sum c_n p(x;\omega_n)$$

and setting $q(x;\omega) = p(x;\omega)/p_0(x)$ one obtains that Theorem 2.18 applies. □

Example 2.22 The model function for a sample $x_1, \ldots, x_n$ from the Cauchy distribution with mode point ω is

$$p(x;\omega) = \pi^{-n} \prod_{j=1}^{n} \frac{1}{1+(x_j-\omega)^2}.$$

If $x = (x_1, \ldots, x_n)$ and $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_n)$ satisfy $cp(x;\cdot) = \tilde{c}p(\tilde{x};\cdot)$ then

$$c\prod_{j=1}^{n}(1+(\tilde{x}_j-\omega)^2) = \tilde{c}\prod_{j=1}^{n}(1+(x_j-\omega)^2).$$

Both sides of this equation are polynomials in ω and hence the equality holds for all $\omega \in \mathcal{R}$ precisely when these two polynomials have the same roots. Since the roots are $\tilde{x}_j \pm i, j = 1, \ldots, n$, respectively $x_j \pm i, j = 1, \ldots, n$ one sees that the order statistic $(x_{(1)}, \ldots, x_{(n)})$ is minimal sufficient.

2.1.8 Sufficiency and exponential families

There is a sense in which "only" exponential families permit a genuine reduction of data without loss of information. We will now make this claim precise.

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model, such that $\mathcal{X}$ is a region (an open and connected set) in $\mathcal{R}^r$, and $\mathcal{A}$ is the Borel σ-algebra of subsets of $\mathcal{X}$. Assume that $\mathcal{P}$ is dominated by the Lebesgue measure λ on $(\mathcal{X}, \mathcal{A})$. Let f_P denote the density of $P \in \mathcal{P}$ with respect to λ, and assume that f_P is positive on $\mathcal{X}$.

By Theorem 2.18, we find that a statistic T is sufficient for the model $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)})$ (corresponding to observing a random sample of size n from the model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$), if and only if for all $P \in \mathcal{P}$ there exists a function h_P such that

$$\frac{dP^{(n)}}{dP_0^{(n)}}(\underline{x}) = \frac{f_P(x_1) \dots f_P(x_n)}{f_{p_0}(x_1) \dots f_{p_0}(x_n)} = h_P(T(x)) \; [P_0^{(n)}].$$

If this condition holds for all $\underline{x} \in \mathcal{X}^{(n)}$, without any restrictions regarding null sets, we say that T is *ϕ-sufficient*. We then have the following result.

Theorem 2.23 *Assume that the densities f_P are continuous on $\mathcal{X}$. Let $k < n$ be positive integers and let T be a continuous k-dimensional statistic, such that T is ϕ-sufficient for the model $(\mathcal{X}^{(n)}, \mathcal{A}^{(n)}, \mathcal{P}^{(n)})$. Then*

(i) If $k = 1, \mathcal{P}$ is an exponential family of order 1.

(ii) If the densities f_P have continuous partial derivatives on $\mathcal{X}$, then $\mathcal{P}$ is an exponential family of order less than or equal to k.

Proof: See Barndorff-Nielsen and Pedersen (1968).

2.2 Ancillarity

Let us consider the statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. In some cases, we can find a statistic V, whose marginal distribution P_V does not depend on the choice of $P \in \mathcal{P}$. That is, $\mathcal{P}_V = \{P_V : P \in \mathcal{P}\}$ has a single element and, therefore the marginal distribution P_V does not contain any information on $P \in \mathcal{P}$. In this case, we say that the statistic V is *ancillary*.

The term ancillary—which means "auxiliary"—apparently does not have anything to do with the definition above. Its use will be justified with more details in the chapter on inferential separation. Roughly speaking, we can say that the observed value of the ancillary statistic works as "auxiliary", in the sense that it shows the contents of "information" of the sample.

The concept of ancillarity was introduced by Fisher in the mid-20s. At the end of the 50s, Basu (1955, 1958 and 1959) made a more detailed analysis of the concept. In the following, we will show Basu's most important results, as well as some examples of ancillary statistics.

2.2.1 Definitions

We will assume, as in the section on sufficiency, that the σ-algebra $\mathcal{A}$ is separable, and that the class of probability measures $\mathcal{P}$ is dominated by a σ-finite measure.

Definition 2.24 *A statistic U is ancillary if the class of marginal distributions of U, $\mathcal{P}_U = \{P_U : P \in \mathcal{P}\}$, consists of a single element.*

By analogy with the definition of minimal sufficiency, it would be obvious to call a statistic U_0 maximal ancillary, if $\sigma(U) \subseteq \sigma(U_0) \vee \mathcal{N}(P)$ for any ancillary statistic U. Note that there rarely exists a maximum ancillary statistic because in many situations it happens that U_1 and U_2 are ancillary but (U_1, U_2) is *not* ancillary. The following example illustrates this.

Example 2.25 Let (X_i, Y_i), $i = 1, \ldots, n$ be independent random vectors with a two-dimensional normal distribution, $E(X_i) = E(Y_i) = 0$, $Var(X_i) = Var(Y_i) = 1$ and $Cov(X_i, Y_i) = \rho$, with $\rho \in (-1, 1)$, for $i = 1, \ldots, n$. Note that we have a model that is parametrized only by ρ. The statistics $\mathbf{X} = (X_1, \ldots, X_n)^\top$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ are ancillary whereas the joint distribution of $(\mathbf{X}, \mathbf{Y})$ depends on ρ and therefore is not ancillary.

Definition 2.26 *A statistic U_0 is maximal ancillary if*

(i) U_0 is ancillary;

(ii) If U is ancillary and $\sigma(U_0) \subseteq \sigma(U) \vee \mathcal{N}(\mathcal{P})$, then $\sigma(U) \subseteq \sigma(U_0) \vee \mathcal{N}(\mathcal{P})$.

Note that Definition 2.26 does not exclude the possibility that there exist more than one maximum ancillary statistic for a given model, as can be seen in the example that follows.

Example 2.27 *(The Multinomial Distribution.)* Let us consider the 2×2 contingency table

$$\begin{bmatrix} X_{11} & X_{12} & X_{1+} \\ X_{21} & X_{22} & X_{2+} \\ X_{+1} & X_{+2} & n \end{bmatrix}$$

with the total n fixed, where $X_{i+} = X_{i1} + X_{i2}$ and $X_{+i} = X_{1i} + X_{2i}$, $i = 1, 2$. We assume that the table above is a realization of n independent multinomial trials with probabilities

$$\begin{array}{cc} (1+\theta)/6 & (2-\theta)/6 \\ (1-\theta)/6 & (2+\theta)/6 \end{array}$$

corresponding to the cells related with X_{11}, X_{12}, X_{21} and X_{22}, respectively. The parameter θ varies in the interval $(-1, 1)$. It is easy to see that $(X_{11}, X_{1+}, X_{+1})^\top$ is minimal sufficient and that X_{+1} and X_{1+} are both maximal ancillary (see Problem 2.19). But,

$$P_\theta(X_{1+} = n; X_{+1} = n) = P_\theta(X_{11} = n) = \{(1+\theta)/6\}^n,$$

and therefore $(X_{1+}, X_{+1})^\top$ is not ancillary.

2.2.2 Basu's Theorem

We now consider a theorem due to Basu which is useful for proving independence.

Theorem 2.28 *(Basu's theorem) Let $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ and $U : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Z}, \mathcal{C})$ be two statistics. We will assume that T is sufficient. Then:*

(i) If T and U are independent under any measure in $\mathcal{P}$, and if no pair of measures in $\mathcal{P}$ are mutually singular, then U is ancillary;

(ii) If T and U are independent under one measure in $\mathcal{P}$, and if the measures in $\mathcal{P}$ are equivalent, then U is ancillary;

(iii) If U is ancillary and if T is boundedly complete with respect to $\mathcal{P}$, then T and U are independent for any $P \in \mathcal{P}$.

Proof:

(i) Let $P_1, P_2 \in \mathcal{P}$. We need to show that $P_1(U \in C) = P_2(U \in C)$ for any $C \in \mathcal{C}$. Since T is sufficient, there exists a Markov kernel π, that is a regular conditional probability of P given T for any $P \in \mathcal{P}$. In particular, we have that

$$P_i(T \in B, U \in C) = \int_B \pi(U \in C|t) P_{iT}(dt), \quad \forall B \in \mathcal{B}, \quad \forall C \in \mathcal{C},$$

for $i = 1, 2$. Since T and U are independent we have that for $B \in \mathcal{B}$ and $C \in \mathcal{C}$ that,

$$P_i(T \in B, U \in C) = P_i(T \in B) P_i(U \in C) = \int_B P_i(U \in C) P_{iT}(dt),$$

for $i = 1, 2$. Comparing the two equations we obtain that there exists a P_{iT}-null set, N_i, such that

$$P_i(U \in C) = \pi(U \in C|t), \quad t \notin N_i, \quad i = 1, 2.$$

Since P_1 and P_2 are not mutually singular, $N_1 \cup N_2 \neq \mathcal{Y}$, and for $t \in \mathcal{Y} \setminus (N_1 \cup N_2)$ we obtain,

$$P_1(U \in C) = \pi(U \in C|t) = P_2(U \in C),$$

that is, $P_1(U \in C) = P_2(U \in C)$, which is valid for each $C \in \mathcal{C}$, as we wanted to show.

(ii) We will assume that T and U are independent on $P_0 \in \mathcal{P}$. Hence, for each $C \in \mathcal{C}$,

$$P_0(U \in C) = \pi(U \in C|t), \quad t \notin N_0,$$

where N_0 is a P_{0T}-null set. Since N_0 is a P_T-null set, for any $P \in \mathcal{P}$, we have,

$$\begin{aligned} P(U \in C) &= \int_{\mathcal{Y}} \pi(U \in C|t) P_T(dt) \\ &= \int_{\mathcal{Y}} P_0(U \in C) P_T(dt) = P_0(U \in C), \end{aligned}$$

which is valid for each $C \in \mathcal{C}$ and this shows that U is ancillary.

(iii) Let $B \in \mathcal{B}$ and $C \in \mathcal{C}$ be arbitrary. We need to show that

$$P(T \in B, U \in C) = P(T \in B)P(U \in C), \quad \forall P \in \mathcal{P}.$$

Let $f : \mathcal{Y} \to I\!R$ be defined by

$$f(t) = P(U \in C) - \pi(U \in C|t).$$

The function f does not depend on P (because U is ancillary) and f is measurable and bounded. Since,

$$\int_{\mathcal{Y}} f(t)P_T(dt) = 0, \quad \forall P \in \mathcal{P},$$

using that T is sufficient and boundedly complete, we obtain that $P_T(f = 0) = 1$, for any $P \in \mathcal{P}$. From this fact,

$$\begin{aligned} P(T \in B, U \in C) &= \int_B \pi(U \in C|t)P_T(dt) \\ &= \int_B P(U \in C)P_T(dt) \\ &= P(U \in C)P(T \in B), \end{aligned}$$

which is valid for any $C \in \mathcal{C}$ and $B \in \mathcal{B}$, implying that U and T are independent.

□

Item (iii) of Basu's theorem is often useful to show independence, as the following example illustrates.

Example 2.29 Let $X_1, \ldots, X_n$ be independent random variables with distribution $N(\mu, \sigma^2)$. For σ^2 fixed (known), $\overline{X}_+ = \frac{1}{n}\sum_{i=1}^n X_i$ is minimal sufficient and complete. Moreover, the distribution of $SSD = \sum_{i=1}^n (X_i - \overline{X}_+)^2$ does not depend on μ and, therefore, SSD is ancillary. By Basu's theorem, it follows that $\overline{X}_+$ and SSD are independent, for any $\mu \in I\!R$ and fixed σ^2 and, therefore, for any $(\mu, \sigma^2) \in I\!R \times I\!R_+$.

In a more general way, $g(X_1, \ldots, X_n)$ is independent of $\overline{X}_+$ if, and only if, for any $a \in I\!R$ we have that $g(a + X_1, \ldots, a + X_n)$ has the same distribution that $g(X_1, \ldots, X_n)$.

If μ and σ^2 are unknown then, the statistic $(\overline{X}_+, s^2)$ (where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_+)^2$) is minimal sufficient and complete. Let

$$U = \left(\frac{X_1 - \overline{X}_+}{s}, \ldots, \frac{X_n - \overline{X}_+}{s} \right)^\top.$$

The distribution of U does not depend on (μ, σ^2), therefore, U is ancillary. Moreover, U and $(\overline{X}, s^2)$ are independent.

Example 2.30 Let $(X_i, Y_i)^\top$, $i = 1, \ldots, n$ be independent random vectors with distribution $N_2(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2)^\top$ and

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}.$$

For $\rho = 0$, $(\overline{X}_+, \overline{Y}_+, \Sigma X_i^2, \Sigma Y_i^2)$ is minimal sufficient and complete, as well as

$$\tilde{T} = (\overline{X}_+, \overline{Y}_+, SSD_X, SSD_Y).$$

Moreover, the distribution of

$$V_{XY} = \frac{SSD_{XY}}{\sqrt{SSD_X SSD_Y}}$$

does not depend on $(\mu, \sigma_1^2, \sigma_2^2)$. By Basu's theorem, we have then that, for $\rho = 0$, V_{XY} is independent of $\tilde{T}$.

2.3 First-order ancillarity

The concepts of sufficiency and ancillarity have an immediate statistical interpretation, as we have seen before. The same does not happen with completeness, which is often called an essentially technical concept. In this section, we will try to obtain a statistical meaning of this notion. Therefore, we will extend the concepts of ancillarity and sufficiency, which will also be useful in Chapter 4 (for more details see Lehmann, 1981).

2.3.1 Examples

To develop the following ideas it will be enough to consider parametric families in $\mathbb{R}$. Let us first consider the following example.

Example 2.31 Let f be a given probability density. Let us consider the family of distributions $\{P_\theta : \theta \in \mathbb{R}\}$, where P_θ is the distribution corresponding to the density f_θ given by $f_\theta(x) = f(x - \theta), \forall x \in \mathbb{R}$, that is, we have a location model. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, with $X_1 \sim P_\theta$. Table 2.1 shows the minimal sufficient statistic, T, for θ, based on the sample, when f is a density of the indicated distributions. In the table $\overline{X}_+$ denotes the sample mean and $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of the sample $X_1, \ldots, X_n$.

It can be shown that a minimal sufficient statistic T is complete in the case of the normal and exponential distributions and that it is not complete in the other cases indicated in the table. We will compare T for the listed distributions and see if we can find some pattern that enables us to characterize the completeness.

First, the dimension of T is one in the cases where T is complete and is larger than one in the other cases. This indicates that T provides, in a certain way, the largest reduction in the data when T is complete. Meanwhile, this is not yet the crucial point.

Table 2.1: Minimal sufficient statistic for some distributions

Distribution given by f	Minimal Sufficient Statistic T
Normal	$\overline{X}_+$
Exponential	$X_{(1)}$
Uniform	$(X_{(1)}, X_{(n)})$
Logistic	$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$
Cauchy	$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$
Double Exponential	$(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$

Define $Y_i = X_{(n)} - X_{(i)}$, for $i = 1, \ldots, n-1$. It is easy to see that the statistic $\tilde{Y} = (Y_1, \ldots, Y_{n-1})^\top$ is ancillary, in any location model. In the case of the logistic, Cauchy and double exponential distributions, where T is not complete, each Y_i can be calculated as a function of T. Then T and $\tilde{Y}$ are correlated. Even in the case of the uniform distribution, Y_1 is a function of T. We conclude that a reduction by sufficiency was not capable of "eliminating" all the ancillarity contained in the data, in the sense of turning all the ancillary statistics independent of T. In the case of the normal and exponential distributions, each Y_i is not a function of T, being some of them independent of T (for example, in the case of the normal, $Cov(T, Y_i) = Cov(\overline{X}_+, X_{(i)} - X_{(n)}) = 0$). The discussion above suggests that completeness might be associated with the capability of T to eliminate ancillarity.

Let us consider the uniform distribution in the interval (θ_1, θ_2), where θ_1 and θ_2 are two parameters to be determined. Again, the statistic $T = (X_{(1)}, X_{(n)})$ is minimal sufficient. In this case, Y_1 is not ancillary any more, but $(Y_2, \ldots, Y_{n-1})$ are and in this case, T is more efficient in eliminating the ancillarity, than in the uniform in $(0, \theta)$ distribution case. It should be pointed that T is now complete!

Let us recall that Basu's theorem (Theorem 2.28 (iii)) says that if T is boundedly complete and sufficient then T is independent of any ancillary statistic. It is evident that, if we had a converse of Basu's theorem, then we would have the statistical interpretation of bounded completeness. Unfortunately, there is no hope that this is possible, since the ancillarity is globally related to the distribution of the statistic whereas completeness is only related to

the expectation. This suggests that a modification of the concept of ancillarity should be made so that it involves only the expectation.

Before continuing we will see an example where a minimal sufficient statistic is not independent of the ancillary statistic (then, by Basu's theorem, the minimal sufficient statistic cannot be complete).

Example 2.32 *Let X be a discrete random variable taking values on*

$$\{-5,-4,-3,-2,-1,1,2,3,4,5\}$$

with probabilities:

$$\begin{array}{rclrcl} P(X=-5) &=& \xi p^2 q & P(X=1) &=& \gamma pq \\ P(X=-4) &=& \xi pq^2 & P(X=2) &=& q^3/2 \\ P(X=-3) &=& p^3/2 & P(X=3) &=& p^3/2 \\ P(X=-2) &=& -1/2q^3 & P(X=4) &=& \alpha pq^2 \\ P(X=1) &=& wpq & P(X=5) &=& \alpha p^2 q \end{array}$$

where $\alpha, \gamma, \xi, w, p$ and q are positive constants such that

$$\alpha + \gamma = \xi + w = 3/2 \text{ and } p = 1 - q \in (0,1).$$

It can be shown that (see Problem 2.21):

(i) $T = |X|$ is minimal sufficient for p;

(ii) $P(X > 0) = 1/2$ and therefore $V = 1_{\{X>0\}}$ is ancillary;

(iii) If $\alpha \neq \xi$ then V is not independent of T.

2.3.2 Main results

Definition 2.33 *A statistic V is called first-order ancillary when $E_\theta(V)$ does not depend on θ and, a statistic T is called first-order sufficient if $E_\theta(X|T)$ does not depend on θ.*

Evidently, ancillarity (sufficiency) implies first-order ancillarity (sufficiency).

Theorem 2.34 *A sufficient statistic T is boundedly complete for θ if and only if any bounded real function of T is uncorrelated with any bounded first-order ancillary statistic for θ.*

Proof: ($\Rightarrow$) Let T be a boundedly complete sufficient statistic. Consider any bounded first-order ancillary statistic (for θ), V, and a given real bounded function f. We will show that $f(T)$ and V are uncorrelated. Without loss of generality, suppose that $\mathrm{E}_\theta(V) = 0$, $\forall \theta \in \Theta$. We have that,

$$\begin{aligned} \mathrm{Cov}_\theta(f(T), V) &= \mathrm{E}_\theta[f(T)V] - \mathrm{E}_\theta(V)\mathrm{E}_\theta(f(T)) = \mathrm{E}_\theta[f(T)V] \\ &= \mathrm{E}_\theta\{f(T)\mathrm{E}_\theta(V|T)\}, \quad \forall\theta \in \Theta. \end{aligned} \tag{2.6}$$

We claim that $\mathrm{E}_\theta(V|T) = 0$ $[P_\theta]$ $\forall\theta \in \Theta$. In fact, we have that

$$0 = \mathrm{E}_\theta(V) = \mathrm{E}_\theta[\mathrm{E}_\theta(V|T)], \quad \forall\theta \in \Theta.$$

Note that $\mathrm{E}_\theta(V|T)$ does not depend on θ because T is sufficient. Since T is boundedly complete, then, $\mathrm{E}_\theta(V|T) = 0$ $[P_\theta]$ $\forall\theta \in \Theta$. Then by (2.6), $f(T)$ and V are uncorrelated, which is valid for any real bounded function f and boundedly complete T.

($\Leftarrow$) If T is not boundedly complete, then there exists a bounded function f such that $\mathrm{E}_\theta(f(T)) = 0$, $\forall\theta \in \Theta$ and $f(T) \neq 0$, with strictly positive probability for some $\theta_0 \in \Theta$. Define $V(x) = f(T(x))$. The statistic V is bounded ancillary of first-order, since, $\mathrm{E}_\theta(V) = \mathrm{E}_\theta(f(T)) = 0$, $\forall\theta \in \Theta$. Moreover, $\mathrm{Cov}_{\theta_0}(V, f(T)) = \mathrm{E}_{\theta_0}[f^2(T)] > 0$ because otherwise $f(T)$ would have a degenerate distribution under P_{θ_0} which would contradict the definition of θ_0. □

Lehmann (1981) claims that "the analogous result (of Theorem 2.34) holds for completeness instead of bounded completeness if attention is restricted to statistics with finite variance". This statement is somewhat imprecise. In order to clarify it we give a definition and then prove a result of the kind Lehmann seems to have had in mind.

Definition 2.35 *Let $\mathcal{F}$ be a class of real valued, measurable functions. We say that the statistic T is $\mathcal{F}$-complete for some set of functions $\mathcal{F}$, if the condition*

$$\forall f \in \mathcal{F} : E_\theta[f(T)] = 0 \forall\theta$$

implies $f = 0$ $[P_{T\theta}]$, $\forall\theta$.

Theorem 2.36 *Let T be a sufficient statistic and let $\mathcal{F}_T$ be the class of functions given by*

$$\mathcal{F}_T = \left\{ f : f \text{ is real valued and } E[f^2(T)] < \infty \right\}.$$

Then T is $\mathcal{F}_T$-complete if and only if for all $f \in \mathcal{F}_T$ and first-order ancillary statistic V, such that Var $_\theta(V) < \infty$ $\forall\theta$ we have Cov $_\theta(V, f(T)) = 0$, $\forall\theta$.

Proof: ($\Rightarrow$) Let T be $\mathcal{F}_T$-complete for θ and $f \in \mathcal{F}_T$. Let V be first-order ancillary, such that Var $_\theta(V) < \infty$, $\forall\theta$. Also let $\phi(T) = E(V|T)$. Then

$$\text{Cov }_\theta(f(T), V) = \mathrm{E}_\theta(f(T)V) = \mathrm{E}_\theta(f(T)E(V|T)).$$

But

$$0 = \mathrm{E}_\theta(V) = \mathrm{E}_\theta[E(V|T)] = \mathrm{E}_\theta[\phi(T)]$$

and

$$E[\phi^2(T)] = E[E^2(V|T)] = \text{Var } [E(V|T)] \leq \text{Var } (V) < \infty.$$

Then $\phi \in \mathcal{F}_T$ and so $\phi = 0$ $[P_{T\theta}]$ implies that Cov $_\theta(f(T), V) = 0$, $\forall\theta$.

($\Leftarrow$) Let $f \in \mathcal{F}_T$ such that $\text{E}_\theta[f(T)] = 0$ $\forall\theta$ and assume $\exists\theta_0$ such that $P_{\theta_0}(f(T) \neq 0) > 0$. If we let $V = f(T)$ then V is first-order ancillary and has finite variance. Then Cov $_\theta(V, f(T)) = 0$, *i.e.* $\text{E}_\theta[f^2(T)] = 0$, $\forall\theta$. Hence, Var $_{\theta_0}[f(T)] = 0$, implying that $P_{\theta_0}(f(T) = 0) = 1$ which is a contradiction. Therefore, $P_\theta(f(T) \neq 0) = 0$, $\forall\theta$. □

Theorem 2.34 gives us an interesting characterization of the concept of bounded completeness in terms of first-order ancillarity. Note that when the concept of ancillarity is weakened by defining the first-order ancillarity, we obtain the analogue of Basu's theorem but *with converse*. The existence of this converse shows that the concept of ancillarity has been weakened in the exact measure for this purpose. Besides it provides an interpretation of bounded completeness. The concepts of first-order ancillarity and sufficiency will be useful to interpret the definitions of sufficiency and ancillarity that will be developed for inference functions in Chapter 4.

2.4 Problems

Sufficiency

Problem 2.1 Let $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ be a sufficient statistic. Then there exists a Markov kernel $\pi(\cdot|\cdot)$, which is a regular conditional probability of P given T for any $P \in \mathcal{P}$. Let $X : (\mathcal{X}, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable with finite expectation for all $P \in \mathcal{P}$. Show that

$$E_P[X|T](x) = \int_{\mathcal{X}} X(x')\pi(dx'|T(x)) \ [P].$$

Thus, a conditional P-mean of X given T does not depend on $P \in \mathcal{P}$.

Problem 2.2 Let $T : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{B})$ and $T_0 : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}_0, \mathcal{B}_0)$. If there exists a measurable function $f : (\mathcal{Y}, \mathcal{B}) \to (\mathcal{Y}_0, \mathcal{B}_0)$ such that $T_0 = f \circ T$ $[P]$ for any $P \in \mathcal{P}$, show that $\sigma(T_0) \subseteq \sigma(T)$ $[\mathcal{P}]$. Show the converse implication, in the particular case where $(\mathcal{Y}_0, B_0) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Problem 2.3 Assume that the measures of $\mathcal{P}$ are equivalent. Let $P_0 \in \mathcal{P}$ be arbitrary. Let f_P and g_P be two versions of $\frac{dP}{dP_0}$. We know that $f_P = g_P$ $[P]$ for any $P \in \mathcal{P}$. Show that

$$\sigma(f_P : P \in \mathcal{P}) = \sigma(g_P : P \in \mathcal{P}) \ [\mathcal{P}].$$

Problem 2.4 Let $\theta \in (0, \infty)$ and suppose that $X_1, \ldots, X_n$ are independent with distribution $U(0, \theta)$. Show, using Theorem 2.10, that $X_{(n)} = \max\{X_1, \ldots, X_n\}$ is a minimal sufficient statistic.

Problem 2.5 Let X and Y be independent with exponential distribution, with mean θ^{-1} and θ, respectively, where $\theta \in (0, \infty)$. Find a minimal sufficient statistic. Hint: Use Theorem 2.7.

Problem 2.6 Let $\mathcal{C}$ be as in Theorem 2.14, that is, $\mathcal{C} = \sigma(\frac{dP}{dP_0} : P \in \mathcal{P})$. Let $X : (\mathcal{X}, \mathcal{A}) \to (I\!R, \mathcal{B}(I\!R))$ be a random variable with finite expectation under any $P \in \mathcal{P}$. Show that

$$\int E_P[X|\mathcal{C}]dP = \int E_0[X|\mathcal{C}]dP \quad \forall P \in \mathcal{P}.$$

Problem 2.7 Under the conditions of Theorem 2.15, show that, if the representation is minimal, there exist $P_1, \ldots, P_k$, such that $\tilde{\alpha}(P_1), \ldots, \tilde{\alpha}(P_k)$ are linearly independent vectors in $I\!R^k$.

Problem 2.8 Let $X_1, \ldots, X_n$ be independent with distribution $U(\theta, \theta + 1)$, where $\theta \in I\!R$. Let $X_{(1)} = \min\{X_1, \ldots, X_n\}$, and $X_{(n)}$ be as in Problem 2.4. Show that $(X_{(1)}, X_{(n)})$ is minimal sufficient.

Problem 2.9 Let X_{ij}, $i = 1, \ldots, k$ and $j = 1, \ldots, c$ be independent random variables, whose joint distribution is multinomial with parameter θ_{ij}, $i = 1, \ldots, k$ and $j = 1, \ldots, c$. Let $\mathcal{P}$ be a family of multinomial distributions with $\theta_{ij} = \theta_{i+}\theta_{+j}$. Find the order of the exponential family $\mathcal{P}$ and find the minimal sufficient statistic.

Problem 2.10 Let X be a k-dimensional statistic and let $\mathcal{P}$ be the set of distributions of X. We will assume that $\mathcal{P}$ is an exponential family with canonical statistic X. Let $T = f(X)$ be a statistic such that $\sigma(T) \subseteq \sigma(X) \vee \mathcal{N}(\mathcal{P})$. (Therefore, T may not be equivalent to X, and a reduction of X by T implies a real reduction of information). Show that, if T is sufficient, then, $ord\mathcal{P} \leq k$.

Ancillarity

Problem 2.11 Let $\mathcal{X} = (-1, 0, 1, 2, \ldots)$, let $\mathcal{A}$ be the class of subsets of $\mathcal{X}$ and let $\mathcal{P} = \{P_\theta : \theta \in (0, 1)\}$ be the family of probabilities determined by

$$\begin{aligned} P_\theta(\{-1\}) &= \theta \\ P_\theta(\{x\}) &= (1-\theta)^2\theta^x \quad x = 0, 1, \ldots. \end{aligned}$$

Let $T(x) = x$, $x \in \mathcal{X}$. Show that T is boundedly complete with respect to $\mathcal{P}$, but that it is not complete.

Problem 2.12 Let $X_1, \ldots, X_n$ be independent and identically distributed as $N(\mu, \sigma^2)$, $\mu \in I\!R$ and $\sigma^2 \in I\!R_+$. Show that

$$T_1 = \frac{\sum_{i=1}^{n-1}(X_{i+1} - X_i)^2}{\sum_{i=1}^{n}(X_i - \overline{X}_+)^2}$$

and

$$T_2 = \frac{X_{(n)} - \overline{X}_+}{X_{(n)} - X_{(1)}}$$

are both independent of $(\overline{X}_+, s^2)$.

Problem 2.13 Let $X_1, \ldots, X_n$ be independent random variables with distribution $Ga(\beta, \lambda)$, $i = 1, \ldots, n$.

(i) Show that the maximum likelihood estimator $(\hat{\lambda}, \hat{\beta})$ of (λ, β) is a solution of the equations

$$\begin{aligned} \lambda\beta &= \overline{X}_+ \\ \psi(\lambda) - \log\lambda &= \log\frac{\tilde{X}}{\overline{X}_+}, \end{aligned}$$

where $\tilde{X} = (\prod_{i=1}^n X_i)^{1/n}$, and ψ is the digamma function.

(ii) Show that $h(X_1, \ldots, X_n)$ is independent of $\overline{X}_+$ if and only if $h(aX_1, \ldots, aX_n)$ has the same distribution as $h(X_1, \ldots, X_n)$ for any $a > 0$.

(iii) Show that $\hat{\lambda}\hat{\beta}$ and $\hat{\lambda}$ are independent.

Problem 2.14 Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with density

$$\frac{1}{\beta} e^{-(x-\alpha)\ \beta}, \quad x \geq \alpha,$$

where $\alpha \in I\!R$ and $\beta > 0$. Find the maximum likelihood estimator $(\hat{\alpha}, \hat{\beta})$ of (α, β), and show that $\hat{\alpha}$ and $\hat{\beta}$ are independent.

Problem 2.15 Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be independent random variables, with the X_i's having density given by

$$f(x) = \frac{1}{\theta_1} e^{-x/\theta_1}, \quad x > 0,$$

and the density of the Y_i's given by

$$f(y) = \frac{1}{\theta_2} e^{-y/\theta_2}, \quad y > 0,$$

where $\theta_1 > 0$ and $\theta_2 > 0$. We consider the hypothesis $\theta_1 = \theta_2$

(i) Find the likelihood ratio test for $\theta_1 = \theta_2$, and show that it only depends on $\overline{X}_+/\overline{Y}_+$.

(ii) Show that $\overline{X}_+/\overline{Y}_+$ has distribution $F(2n, 2n)$ if $\theta_1 = \theta_2$.

(iii) Show that the likelihood ratio test of $\theta_1 = \theta_2$ of level α has acceptance region

$$1/C \leq \overline{X}_+/\overline{Y}_+ \leq C,$$

where C is the $1 - \frac{\alpha}{2}$ quantile of the distribution $F(2n, 2n)$.

(iv) Find the power function of the likelihood ratio test for $\theta_1 = \theta_2$.

(v) Write the density of $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ in the exponential form with $(\overline{X}_+, \overline{X}_+ + \overline{Y}_+)$ as canonical statistic, and show that for $\theta_1 = \theta_2$, $\overline{X}_+ + \overline{Y}_+$ is sufficient and complete. Show that $\overline{X}_+/\overline{Y}_+$ and $\overline{X}_+ + \overline{Y}_+$ are independent if $\theta_1 = \theta_2$.

Problem 2.16 (Fisher-Behrens' problem). Let $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$ be independent with $X_i \sim N(\mu_1, \sigma_1^2)$, $i = 1, \ldots, n$ and $Y_i \sim N(\mu_2, \sigma_2^2)$, $i = 1, \ldots, m$. Let H_0 be the hypothesis $H_0 : \mu_1 = \mu_2$. Let $\mathcal{P}_0$ be the class of distributions corresponding to H_0. Show that $\mathcal{P}_0$ is an exponential family of order 4, and that $T = (\Sigma X_i, \Sigma Y_i, \Sigma X_i^2, \Sigma Y_i^2)$ is minimal sufficient. Show that T is not complete with respect to $\mathcal{P}_0$.

Problem 2.17 Let us suppose that $X_1, \ldots, X_n$ are independent, with distribution $U(0, \theta)$. Show using Theorem 2.6 (sufficiency) that $X_{(n)}$ is minimal sufficient. (This was already proved in Problem 2.4 of sufficiency). Furthermore, show that $X_{(n)}$ and $X_{(1)}/X_{(n)}$ are independent.

Problem 2.18 Let $X_1, \ldots, X_n$ be independent with distribution $U(\theta, \theta + 1)$, $\theta \in \mathbb{R}$. Show that the minimal sufficient statistic $(X_{(1)}, X_{(n)})$ is not boundedly complete with respect to the class of distributions of $(X_1, \ldots, X_n)$.

Problem 2.19 Show the claims made in Example 1.10.

Consider a set of independent and identically distributed random variables $X_1, \ldots, X_n$, normally distributed $N(\mu, \sigma^2)$.

1. Let $X_0 = X_n$, and assume that $\mu = 0$. Show that

$$\sum_{i=1}^{n} X_i^2 \text{ and } \frac{\sum_{i=1}^{n} X_{i-1} X_i}{\sum_{i=1}^{n} X_i^2}$$

are independent.

2. Now let μ be arbitrary, let

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and let

$$m_r = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^r, \; r = 2, 3, \ldots,$$

$$g_1 = \frac{m_3}{m_2^{3/2}}$$
$$g_2 = \frac{m_4}{m_2^2} - 3.$$

Show that $(g_1, g_2)^\top$ is independent of $(\bar{X}, m_2)^\top$.

3. From now on, assume that $\mu = 0$. Let $\beta \in (-1, 1)$ and define $Y_1, \ldots, Y_n$ by the system of equations

$$Y_i + \beta Y_{i-1} = X_i,\ i = 1, \ldots, n,$$

where $Y_0 = Y_n$. Show that $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ has a multivariate normal distribution, and show that its probability density function with respect to Lebesgue measure is

$$f(\mathbf{y}) = \frac{1 - (-\beta)^n}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left\{(1+\beta^2)\sum_{i=1}^n y_i^2 + 2\beta\sum_{i=1}^n y_{i-1}y_i\right\}\right].$$

4. Let

$$S = \sum_{i=1}^n Y_i^2$$
$$R = \frac{\sum_{i=1}^n Y_{i-1}Y_i}{S}.$$

Show that if $(\beta, \sigma^2)^\top$ varies in $(-1, 1) \times (0, \infty)$, then $(S, R)^\top$ is minimal sufficient for Y. Is $(S, R)^\top$ also minimal sufficient if σ^2 is known and β varies in $(-1, 1)$?

5. Let $g_0(r)$ denote the probability density function of R with respect to Lebesgue measure in the case $\beta = 0$ (you are not required to find g_0). Find, for arbitrary $(\beta, \sigma^2)^\top$ in $(-1, 1) \times (0, \infty)$, the density of $(S, R)^\top$ with respect to Lebesgue measure on $\mathbf{R}^2$. The final expression for this density must be indicated as explicitly as possible, except that $g_0(r)$ may enter in the expression. Hint: The result in Question 1 may be useful.

First-order ancillarity

Problem 2.20 Show the claims made in Example 1.13.

Problem 2.21 Show the claims made in Example 1.25.

Chapter 3

INFERENTIAL SEPARATION

In this chapter we will generalize the classical concepts of sufficiency and ancillarity studied in Chapter 2, for the case when there is a nuisance parameter. We will introduce five new concepts of sufficiency (S-, L-, G-, M- and I-sufficiency) and four new ones of ancillarity (S-, G-, M- and I-ancillarity), making reference to the classical concepts introduced in the preceding chapter as B-sufficiency and B-ancillarity (the B comes from Basu and Bahadur). In Section 3.1 we will present a motivation for the definition of these concepts, and in Sections 3.2–3.6 we will present the corresponding mathematical theory including some examples. See Basu (1978) for a review of some complimentary results on partial sufficiency.

3.1 Introduction

Statistical inference is usually done following the steps given below:

(i) Propose a statistical model of the form $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ for a certain experiment, where $\mathcal{X}$ is the sample space, $\mathcal{A}$ is a σ-algebra and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measures parametrized by $\theta \in \Theta$.

(ii) Find the likelihood function

$$L(\theta) = \frac{dP_\theta}{d\nu}(x), \tag{3.1}$$

where ν is a measure on $(\mathcal{X}, \mathcal{A})$ which dominates P_θ, for all $\theta \in \Theta$;

(iii) Calculate the maximum likelihood estimator $\hat{\theta}(x)$, *i.e.*, the value of θ that maximizes $L(\theta)$ in (3.1);

(iv) Conduct hypothesis tests, using the likelihood ratio.

In many situations this paradigm is not sufficient to solve all problems, particularly when θ is multidimensional. For example, let $\theta = (\psi, \phi)$ have two components, ψ and ϕ, both multidimensional. Suppose that we want to make inference only on one of the components. Such

a parameter is known as the *parameter of interest* and the other as the *nuisance parameter*. Obviously this case is not considered in the above paradigm. Inferential separation techniques deal with this situation, making inference on the parameter of interest and eliminating the nuisance parameter, as it will be illustrated in Examples 3.1 and 3.2.

Example 3.1 Let $X_1, \ldots, X_n$ be independent random variables with identical distribution $N(\mu, \sigma^2)$. The likelihood function is:

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-\frac{1}{2\sigma^2}(X_i - \mu)^2} \\ &= \exp\{-\frac{n}{2\sigma^2}(\bar{X}_+ - \mu)^2\}(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2} SSD\right\}, \end{aligned}$$

where $\bar{X}_+ = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $SSD = \sum_{i=1}^{n}(X_i - \bar{X}_+)^2$. We obtain the estimators

$$\hat{\mu} = \bar{X}_+ \quad \textit{and} \quad \hat{\sigma}^2 = \frac{1}{n} SSD.$$

Here we find the first problem. It is known from the theory of linear models that the usual estimator of σ^2 is

$$s^2 = \frac{1}{n-1} SSD.$$

An argument to justify the use of s^2 instead of $\hat{\sigma}^2$ is that $E(s^2) = \sigma^2$, while $E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$. Later we will present another justification for choosing s^2 in terms of the concept of L-sufficiency or in terms of G-sufficiency.

In Example 3.1, the difference between $\hat{\sigma}^2$ and s^2 is very small for n big. We will see in the following example that this difference may become significant.

Example 3.2 Let $X_{ij} \sim N(\mu_i, \sigma^2)$, $j = 1, 2$, $i = 1, \ldots, n$, be independent random variables. This corresponds to an experiment with n groups, each with two observations, where the mean changes from group to group. This is a typical example of paired observations.

Using the likelihood function, we obtain the following estimators for the parameters of the model:

$$\begin{aligned} \hat{\mu}_i &= \bar{X}_{i+} = \frac{1}{2}(X_{i1} + X_{i2}) \quad \text{and} \\ \hat{\sigma}^2 &= \frac{1}{2n}\sum_{i,j}(X_{ij} - \bar{X}_{i+})^2. \end{aligned}$$

Consequently,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_{i1} - X_{i2}}{2}\right)^2 \underset{n \to \infty}{\overset{a.s.}{\longrightarrow}} E\left[\frac{X_{11} - X_{12}}{2}\right]^2 = \frac{1}{2}\sigma^2.$$

Therefore, $\hat{\sigma}^2$ is not a reasonable estimator for σ^2 since it is not consistent. Using $X_{i1} - X_{i2} \sim N(0, 2\sigma^2)$ to estimate σ^2 we obtain the estimator

$$s^2 = \frac{1}{2n}\sum_{i=1}^{n}(X_{i1} - X_{i2})^2 = 2\hat{\sigma}^2 \underset{n \to \infty}{\overset{a.s.}{\longrightarrow}} \sigma^2,$$

which solves the problem of lack of consistency of $\hat{\sigma}^2$.

This type of examples have motivated some critics to the maximum likelihood method. For example, LeCam (1990) says: "Maximum likelihood estimators are considered the best estimators in all circumstances. However, there are many cases in which they plainly misbehave". However, we assert that the maximum likelihood method is not wrong in itself, but it is important to choose correctly the model to work with. Both examples discussed above show that it is not always convenient to deal with the whole likelihood function. Instead of working with the distribution of X under P_θ in order to make inference, which would lead us to the complete likelihood function, we propose to use the marginal distribution of a "sufficient" statistic, $U = u(X)$, or the conditional distribution of X given an "ancillary" statistic, $V = v(X)$. Let us consider now the situation where the parameter θ can be decomposed in the form $\theta = (\psi, \phi)$, where $\psi = \psi(\theta)$ is the parameter of interest. Let us also suppose that

$$L(\theta) = f(x; \theta) = h(u; \psi(\theta))g(x|u; \theta), \tag{3.2}$$

where $U = u(X)$ is a statistic, and h and g represent the marginal density of U and the conditional density of X given U, respectively. We are supposing in (3.2) that the marginal density of U depends on θ only through the parameter of interest $\psi = \psi(\theta)$ and that ψ parametrizes the family of distributions $\{h(u; \psi(\theta)); \theta \in \Theta\}$. If in some sense it is possible to say that the conditional distribution of X given $U = u$ does not contain information about ψ, then we say that U is *sufficient* (in the broad sense) for ψ in the presence of ϕ. In this case we claim that inference on ψ has to be done using the marginal distribution of U, *i.e.*, $h(u; \psi(\theta))$. This basic principle is called the *sufficiency principle*.

Another situation that may happen is that the likelihood can be written as

$$L(\theta) = f(x; \theta) = g(x|v; \psi(\theta))h(v; \theta), \tag{3.3}$$

where $V = v(X)$ is a statistic. We assume in (3.3) that the conditional distribution of X given $V = v$ depends on θ only through the parameter of interest $\psi = \psi(\theta)$ and that ψ parametrizes the family of distributions $\{g(x|v; \psi(\theta)) : \theta \in \Theta\}$. If, in some sense, the marginal distribution of V does not contain information about ψ, then V will be called *ancillary* (in the broad sense) for ψ in the presence of ϕ. In this case any inference on ψ should be done using the conditional distribution of X given $V = v$. This principle is called the "*ancillarity principle*". The term ancillary that means "auxiliary", is used because although the statistic V does not contains any information about ψ, the observed value v of V shows what is the conditional distribution that has to be used in the inference on ψ.

To formalize the definitions of sufficiency and ancillarity given above, we have to be precise with what is the exact meaning of a distribution that does not contain information about ψ. In the following sections we will give different definitions of "does not contain information", which will correspond to the concepts of sufficiency and ancillarity mentioned in the first paragraph. It is important to see that the concepts of B-sufficiency and B-ancillarity do not apply in this context except if ϕ is a constant and θ is the parameter of interest.

Example 3.3 Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with

$$P(X_i = 1) = 1 - P(X_i = 0) = \theta, \quad 0 < \theta < 1.$$

The likelihood function in this example is

$$L(\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{X_+}(1-\theta)^{n-X_+},$$

where $X_+ = X_1 + \ldots + X_n$. The statistic X_+ has a binomial distribution and the conditional distribution of $(X_1, \ldots, X_n)|X_+ = x_+$ is the uniform distribution on the set $C = \{(X_1, \ldots, X_n) \in \{0,1\}^n : X_1 + \ldots + X_n = x_+\}$. Then we have the factorization

$$L(\theta) = \begin{pmatrix} n \\ X_+ \end{pmatrix} \theta^{X_+}(1-\theta)^{n-X_+} \left[\begin{pmatrix} n \\ X_+ \end{pmatrix}^{-1} 1_C((X_1, \ldots, X_n)) \right].$$

In this case the conditional distribution of $(X_1, \ldots, X_n)$ given $X_+ = x_+$ does not depend on the parameter θ, therefore in this case X_+ is B-sufficient for θ.

Example 3.4 The first two examples presented about a B-ancillary statistic were given by Fisher (1934) in his discussion on the model of location and scale. One is the following, let $X_1, \ldots, X_n$ be independent random variable with distribution $X_i \sim P_{\mu,\sigma}$, where $P_{\mu,\sigma}$ is a distribution with density

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right),$$

where f is a given density. Evidently the statistic $V = v(X)$ given by

$$v(X) = \left(\frac{X_1 - X_3}{X_1 - X_2}, \frac{X_1 - X_4}{X_1 - X_2}, \ldots, \frac{X_1 - X_n}{X_1 - X_2}\right),$$

is B-ancillary. Fisher named $v(X)$ the "sample configuration" claiming that the inference on μ and σ should be done using the conditional distribution of $(\hat{\mu}, \hat{\sigma}^2)$ given $v(X) = v$.

We have following expression for the distribution of $P_{\mu,\sigma}$:

$$a + bP_{\mu,\sigma} = P_{a+b\mu,|b|\sigma},$$

whenever b is different from zero. This shows that the family $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in I\!R, \sigma > 0\}$ is generated by the group of affine transformations of $I\!R$.

The B-ancillarity property of a statistic is often a consequence of its invariance under a set of transformations, like for $v(X)$ in the example shown above. This idea will be used to define the concepts of G-sufficiency and G-ancillarity in Section 3.3.

3.1.1 S-ancillarity

The concepts of S-sufficiency and S-ancillarity are the most immediate generalizations of those of B-ancillarity and B-sufficiency. There are no special polemics about these concepts, which is not the case for the rest of the ancillarity and sufficiency concepts. In this section we will introduce some examples about S-ancillarity, leaving the formal detailed presentation to Section 3.2.

Example 3.5 Let us consider a continuous time Poisson process with intensity λ. When the i-th event occurs we observe the i-th random variable X_i, $i = 1, 2, \ldots$. Let us assume that these random variables $X_1, \ldots, X_n$ are independent given the process (for every $n \in \mathbb{N}$). This situation happens commonly in the study of an insurance portfolio, where the event is the occurrence of a claim and X_i is the value of the i-th claim.

Let us suppose that the process is observed in the interval $(0, 1)$. Let N be the number of claims in this interval. Let us assume that

$$X_i \sim N(\mu, 1), \quad i = 1, 2, \ldots \quad .$$

In this case the likelihood function is given by the Poisson distribution with mean λ and, given $N = n$, the random variables $X_1, \ldots, X_n$ are independent and identically distributed with the same distribution as in Example 3.1, with $\sigma^2 = 1$. The likelihood is given by

$$L(\mu, \lambda) = \left\{ \prod_{i=1}^{n} (2\pi)^{-1/2} e^{-1/2(x_i - \mu)^2} \right\} \frac{\lambda^n}{n!} e^{-\lambda}, \tag{3.4}$$

and the maximum likelihood estimators of λ and μ are, respectively,

$$\begin{aligned} \hat{\lambda} &= N \\ \hat{\mu} &= \begin{cases} 1/N \sum_{i=1}^{N} X_i, & \text{if } N \geq 1 \\ \text{undefined}, & \text{if } N = 0 \end{cases} . \end{aligned}$$

In this example it is intuitively clear that in order to talk about the goodness of the estimator $\hat{\mu}$, i.e., the "distance" between $\hat{\mu}$ and μ, we have to fix n and consider the variance

$$Var(1/n \sum_{i=1}^{n} X_i) = Var(\hat{\mu} | N = n) = 1/n.$$

Therefore, the random variable N acts like an index of the content of information of the sample about μ. That is, to evaluate $\hat{\mu}$ it makes no sense to compare different values of N since the variance of $\hat{\mu}$ depends on the observed value of the random variable N. Thus, if we observe $N = 1$ or $N = 1000$ we have very different situations.

If λ is known, the model is parametrized by $\theta = \mu$, and in this case the distribution of N does not depend on θ and therefore, N is B-ancillary. However, if $\lambda \in (0, \infty)$ is unknown, the model is parametrized by $\theta = (\mu, \lambda)$, where $\mu \in \mathbb{R}$ is the parameter of interest. We will say that N is S-ancillary for μ, according to the definitions given below, since the class of marginal distributions for N is the same for any value of μ.

Next we give the first definition of S-ancillarity. Let us consider the situation where the parameter $\theta = (\psi, \phi) \in \Theta$, (like in Section 3.1) with $\psi \in \Theta_1$ and $\phi \in \Theta_2$. We say that a statistic V is *S-ancillary* with respect to ψ when $\Theta = \Theta_1 \times \Theta_2$ and the likelihood function can be factorized in the following way

$$L(\theta) = g(x|v;\psi)h(v;\phi). \tag{3.5}$$

Observe that the conditional distribution of X given V does not depend on ϕ and that the marginal distribution of V does not depend on the parameter of interest ψ.

In the example given above we have the parameter $\theta = (\mu, \lambda) \in I\!R \times (0, \infty)$ where the marginal distributions of N are the same foe each value of μ. Hence, λ parametrizes the marginal distributions of N and clearly N is S-ancillary.

Example 3.6 *(Linear regression)* Let us consider a very frequent example where the ancillarity principle is used, although not always explicitly.

Let $Y = (Y_1, \ldots, Y_n)^\top$ be a vector of independent random variables with conditional distribution

$$Y_i|X_i = x_i \sim N(\alpha + \beta x_i, \sigma^2),$$

where $X = (X_1, \ldots, X_n)^\top$ is a random vector with $X \sim P \in \mathcal{P}$, and $\mathcal{P}$ is a family of distributions. Let us assume that $\psi = (\alpha, \beta, \sigma^2)$ is the parameter of interest and that $\mathcal{P} = \{P_\phi : \phi \in \Psi\}$ is parametrized by ϕ which varies independently of ψ.

The joint distribution of (X, Y) can be factorized in the following form:

$$f(x, y; \psi, \phi) = g(y|x;\psi)h(x;\phi).$$

Therefore we have that X is S-ancillary for ψ and that, according to the ancillarity principle, inference on $\psi = (\alpha, \beta, \sigma^2)$ has to be done with the conditional distribution of Y given $X = x$. This conditional procedure avoids the problem of having to specify the marginal distribution of X which implies that $x_1, \ldots, x_n$ have to be considered as constants in the analysis, being themselves realizations of random variables. However, it is important to notice that it is fundamental that ϕ does not have any relation with the parameters $(\alpha, \beta, \sigma^2)$ since otherwise the marginal distribution of X may contain additional information about these parameters.

The role of X, as auxiliary statistic, is easily illustrated as follows. Let

$$\hat{\beta} = \frac{SPD_{XY}}{SSD_X},$$

where $SPD_{XY} = \sum_{i=1}^n (X_i - \bar{X}_+)(Y_i - \bar{Y}_+)$ and $SSD_X = \sum_{i=1}^n (X_i - \bar{X}_+)^2$. Then the conditional variance of $\hat{\beta}$ given X is

$$Var(\hat{\beta}|X = x) = \frac{\sigma^2}{SSD_X} \quad .$$

Therefore the random variable SSD_X is an index of the conditional distribution of $\hat{\beta}$ given X. Evidently SSD_X contains important information about the distribution of $\hat{\beta}$, being X

random. If SSD_X is very large the slope of the line would be known very precisely and, otherwise, if SSD_X is small we would have little information about the slope of the line. That is, the marginal variance of $\hat{\beta}$ would be very pessimistic in the first case and very optimistic in the second one, and it would lead us to adopt an unacceptable procedure. The S-ancillarity principle would lead us, instead, to make the correct inference, conditioning on X.

3.1.2 The nonformation principle

Let us summarize the ideas already presented when formulating the nonformation principle. From this general principle we will derive the ancillarity and sufficiency principles.

Let X be a random variable defined in the measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric model in $(\mathcal{X}, \mathcal{A})$. Let $\psi = \psi(\theta)$ be the parameter of interest. Sometimes, it is useful to think of ψ as a component of the vector θ, writing $\theta = (\psi, \phi)$, where ϕ is a nuisance parameter. However, it is important to mention that the nuisance parameter ϕ will have just a secondary role in the discussion that follows. Therefore we simply can say that we are interested in making inference on a function ψ of the original parameter θ, *i.e.*, on $\psi(\theta) = \psi$. This has to be considered since the nuisance parameter can be defined in different ways for the same parameter of interest.

Recalling what was said in Subsection 3.1.1, in the case when there exists an statistic $U = u(X)$ with marginal distribution parametrized only by ψ, the conditional distribution of X given U does not contain any information about ψ, then U is said to be *sufficient* for ψ given ϕ. The *sufficiency principle* indicates that in this case inference has to be done using the marginal distribution of U. That is, the likelihood is factorized in the following way

$$L(\theta) = h(u; \psi(\theta))g(x|u; \theta). \tag{3.6}$$

We say that the factor $g(x|u; \theta)$, $\theta \in \Theta$, together with the given x, is *nonformative* with respect to the inference on ψ.

If the conditional distribution of X given a statistic V is parametrized by ψ and the marginal distribution of V does not contain any information about ψ, then V is called *ancillary* for ψ given ϕ. The *ancillarity principle* says that inference on ψ has to be done using the conditional distribution of X given $V = v$, where v is the observed value of V. In this case the likelihood can be factorized in the following way

$$L(\theta) = g(x|v; \psi(\theta))h(v; \theta), \tag{3.7}$$

and we say that the factor $h(v; \theta)$, $\theta \in \Theta$, together with the given v is *nonformative* with respect to the inference on ψ.

Let us formulate the general nonformation principle. Let S and T be two statistics and $g(s|t; \theta)$ be a density of the conditional distribution of S given $T = t$. In the case that this conditional distribution, together with the observed values s and t of S and T, do not contain information about the parameter of interest $\psi(\theta)$, we say that, in a wide sense, it

is *nonformative* with respect to the parameter ψ. The *nonformation principle* says that if the sub-model defined by the factor $g(s|t;\theta)$ is nonformative with respect to ψ, then this factor should be discarded, only using the complement of the likelihood to do the inference on ψ. This point will be explained in more detail later. Note that the nonformation principle is a generalization of the sufficiency principle, that corresponds to the particular case where $S = X$, and of the ancillarity principle, in the particular case when T is constant. In this way we say that the nonformation principle generalizes the sufficiency and ancillarity principles.

Apparently it is not possible to give a unique definition of nonformation. Up to this moment we have already discussed two definitions of nonformation:

B: The family of distributions contains a unique element, *i.e.*, the distribution does not depend on the value of θ;

S: The family of distributions corresponding to a fixed value of ψ is the same for all the values of ψ (the letter S comes from Sverdrup and Sandved).

In Sections 3.3–3.6 we will introduce four other concepts of absence of information. We use the letters L (from "likelihood"), G (from group), M (from maximum) and I (from information).

Sufficiency and ancillarity are the two extremes of the wide collection of possibilities of nonformation. It is important to recall that many times the nonformation principle is applied in stages. Thus, we eliminate successively the nonformative factors of the likelihood, arriving to the *exhaustive model*, which will be the model to be used for inference on the parameter of interest, ψ.

Let us explain this process in the case of a likelihood with three factors. Let U and V be statistics such that the conditional distribution of U given V depends on θ only through $\psi = \psi(\theta)$ and that ψ parametrizes this family of distributions, that is, the conditional distribution of U given $V = v$ has conditional density of the form

$$g(u|v;\psi(\theta)\,)\,, \tag{3.8}$$

and ψ parametrizes the family

$$\{g(u|v;\psi(\theta)\,) : \theta \in \Theta\}.$$

Hence we can write the likelihood of θ as:

$$L(\theta) = g(u|v;\psi(\theta)\,)\; h_1(v;\theta)\; h_2(x|u,v;\theta). \tag{3.9}$$

If the two factors h_1 and h_2 are both nonformative with respect to ψ, by the nonformation principle, we should ignore them. Thus, we should make inference on ψ using the exhaustive factor of the likelihood, $g(u|v;\psi)$.

Let us consider a typical case of this procedure. Let us assume that expression (3.9) has the form

$$L(\theta) = g(u|v;\psi)\; h_1(v;\phi)\; h_2(x|u,v)\,.$$

Hence, since the factor $h_2(x|u,v)$ does not depend on θ, we have that (U,V) is B-sufficient for θ. We ignore, thus, the factor h_2 of the likelihood, because this is nonformative with respect to θ, and therefore with respect to $\psi(\theta)$. The distribution of the B-sufficient statistic (U,V) is given by $g(u|v;\psi)h_1(v;\phi)$.

We suppose now that this factorization is such that V is S-ancillary for ψ in the marginal distribution of (U,V). Hence, the factor h_1 is nonformative with respect to ψ, and we ignore it from the likelihood. In this way, inference on ψ should be made using the exhaustive factor $g(u|v;\psi)$. The reasoning above shows that, in general, we can use different types of nonformation for each factor of the likelihood.

Example 3.7 Let $X_1,\ldots,X_n$ be independent random variables with inverse Gaussian distribution $X_i \sim N^-(\chi,\psi)$, given by the density

$$f(x;\chi,\psi) = \sqrt{\frac{\chi}{2\pi x^3}}\,\exp[\sqrt{\chi\psi} - \frac{1}{2}(\chi x^{-1} + \psi x)] \quad (x>0).$$

The likelihood function is

$$L(\chi,\psi) = \Big(\frac{\chi}{2\pi}\Big)^{n/2} \prod_{i=1}^{n} x_i^{-3/2}\,\exp\{n\sqrt{\chi\psi} - \frac{1}{2}(\chi X_- + \psi X_+)\} \tag{3.10}$$

where $X_- = X_1^{-1} + \ldots + X_n^{-1}$ and $X_+ = X_1 + \ldots + X_n$. Thus, this is an exponential family of order 2, with minimal canonical statistic $(X_-, X_+)^\top$.

We define the parameters $\omega = (\chi\psi)^{1/2}$ and $\mu = (\chi/\psi)^{1/2}$. We consider the case where we want to make inference on μ when $\omega = \omega_0$ is known. It is easy to see that for $c>0$ we have

$$cN^-(\chi,\psi) = N^-(c\chi, c^{-1}\psi).$$

Thus, multiplying $(X_1,\ldots,X_n)^\top$ by c, $\omega=\omega_0$ remains constant and μ is multiplied by c. We now define $S = (X_+/X_-)^{1/2}$ and $T = (X_-X_+)^{1/2}$. Hence, T is invariant under multiplication of $(X_1,\ldots,X_n)^\top$ by c, and S has been multiplied by c.

The likelihood of μ, with $\omega=\omega_0$ fixed is

$$\begin{aligned} L(\mu) &= \big(\tfrac{\omega_0\mu}{2\pi}\big)^{n/2} \textstyle\prod_{i=1}^n x_i^{-3/2}\exp\{n\omega_0 - 1/2\omega_0(\mu X_- + \mu^{-1}X_+)\} \\ &= \big(\tfrac{\omega_0\mu}{2\pi}\big)^{n/2} \textstyle\prod_{i=1}^n x_i^{-3/2}\exp\{n\omega_0 - 1/2\omega_0 T(\frac{\mu}{S} + \frac{S}{\mu})\}. \end{aligned}$$

It is evident that both (X_-,X_+) and (S,T) are B-sufficient for μ and that T is B-ancillary. Thus, inference on μ should be made using the conditional distribution of S given $T=t$. This distribution was found by Jørgensen (1982) and is given by

$$S|T=t \sim N_g^-(-n/2, t\omega_0\mu, t\omega_0\mu^{-1})$$

where $N_g^-(\lambda, \chi, \psi)$ denotes the generalized inverse Gaussian distribution with density

$$f(x; \lambda, \chi, \psi) = \frac{(\psi/\chi)^{\lambda/2}}{2K_\lambda(\sqrt{\chi\psi})} x^{\lambda-1} \exp\{-\frac{1}{2}(\chi x^{-1} + \psi x)\}, \quad x > 0$$

where

$$K_\lambda(w) = 1/2 \int_0^\infty x^{\lambda-1} e^{-1/2(x+x^{-1})} dx$$

is called the modified Bessel function of third type with index λ.

The concepts of nonformation introduced here should be considered as a help to "separate" the inference in pieces for a given model. This separation consists in general, in a factorization of the likelihood function.

Unfortunately it does not seem to be possible to give a unique general definition of nonformation. Thus, even if it seems to be not so elegant, we give five different definitions. We should emphasize, however, that each of these concepts accounts for quite different problems. This is the present state of the developed theory.

It is important to emphasize that the principles not always lead to a unique procedure. Besides, apparently, these principles are not derived from other more basic principles. Thus, the major justification of the theory that we will develop are the examples.

3.1.3 Discussion

The ideas of sufficiency and ancillarity, that started being developed by Fisher in the 20's and 30's, generated a polemic that lasted up to the present time, although now the discussion is less intense. We will not give the details of the arguments given by both parts. It is interesting to note that the sufficiency principle is almost unanimously accepted by the different schools of modern statistics; however, the same does not happen with the ancillarity principle. This seems strange at first glance, given that from the point of view of the exposition made above, the two principles are extremes cases of the same nonformation principle.

A plausible explanation for the easy acceptance of the sufficiency principle is that frequently the procedures produced by this principle coincide with those of widely used statistical procedures. An example of this is the maximum likelihood estimator, which is always a function of any sufficient statistic. Another example is the classic theorem of Lehmann-Scheffé, that says that the conditional expectation of an unbiased estimator of a parameter, given a complete sufficient statistic, is an unbiased estimator of minimal variance (UMVU). Thus, the use of the sufficiency principle sounds in general familiar to many statisticians.

The ancillarity principle implies a change of reference, which is perhaps what can explain the resistance to its acceptance. That is, if the original observations X are defined in a space $\mathcal{X}$, when we apply the ancillarity principle, using the value of an ancillary statistic, say $V = v_0$, we start to work with a subset of the sample space, $\{x \in \mathcal{X} : v(x) = v_0\}$. That is, by applying the ancillarity principle we are forced to change the sample space.

The main argument against the ancillarity principle is that the conditional procedure is less efficient in average in hypothetical repetitions, when compared with the non-conditional procedure. The main arguments in favour of the ancillarity principle are:

(i) The randomness contained in the ancillary statistic is a source of noise which is eliminated with the conditional procedure.

(ii) The conditional procedure uses the "true" content of information of the sample about the parameters.

(iii) Given the sample, hypothetical repetitions of the experiment are irrelevant for the interpretation of the sample.

A more detailed analysis of Example 3.6 about regression, allows us to illustrate the use of the conditional procedure with more detail and it clearly illustrates the last point mentioned above. We will compare two possibilities, $SSD_X = 1,000$ and $SSD_X = 0.001$. To simplify the discussion, we say that the expected value of SSD_X, using the distribution of X, is $SSD_X = 1$. In the first case, the conditional variance of $\hat{\beta}$ is $\sigma^2/1,000$ and in the second case it is $1,000\sigma^2$. The expected value of the variance of $\hat{\beta}$ is nearly σ^2. If we observe the value $SSD_X = 1,000$, then we estimate β with good precision and it would be scientifically incorrect to ignore this information and say that "the mean variance of $\hat{\beta}$ is σ^2". In a similar way, if the observed value of SSD_X is 0.001 it would be too optimistic to say that "this does not matter, because the mean variance of $\hat{\beta}$ is σ^2". Hence, from the principle that an estimator should be given the precision that the observed sample permits (and not the one it would have in average) we conclude that the conditional procedure gives the correct conclusion with respect to the precision of the estimator.

Now let us we give a famous example due to Cox, which will facilitate the presentation of a general argument in favour of the use of the ancillarity and sufficiency principles.

Example 3.8 *(Cox, 1958)* Suppose that we want to measure the length of a bar of iron and to do that we have two instruments. We will also assume that the measurements obtained with each instrument follow the normal distribution with common mean μ and variance 1 for one instrument and 100 for the other. The choice of the instrument to measure the bar is made by tossing a coin that has probability $p \in (0, 1)$ of showing heads and, by convention, if a head is obtained then the measurement is made using the instrument with the smallest variance, otherwise it is made using the other instrument.

Let V be the random variable that indicates the result of the coin tossing experiment and let U be the random variable that indicates the result of the measurement. Define $X = (U, V)$. We have that $V \sim Bi(1, p)$. The model has parameter $\theta = (p, \mu)$, and

$$U|V = v \sim N(\mu, v), \quad \textit{where } v = \begin{cases} 1, & \textit{if } v \textit{ is heads} \\ 100, & \textit{otherwise .} \end{cases}$$

Evidently, V is S-ancillary with respect to μ, because $N(\mu, 1)$ and $N(\mu, 100)$ are both parametrized by $\mu \in I\!R$. The maximum likelihood estimator of μ, $\hat{\mu}$, has a distribution given by the density

$$f_{\hat{\mu}}(y; \mu, p) = p\phi(y - \mu) + (1 - p)\frac{1}{10}\phi\left(\frac{y - \mu}{10}\right), \tag{3.11}$$

where ϕ is the density of the standard normal distribution. Meanwhile, after the choice of the instrument the distribution of $\hat{\mu}$ is $N(\mu, 1)$ or $N(\mu, 100)$. If we know which instrument was used, it is evidently more sensible to use the distribution $N(\mu, 1)$ or $N(\mu, 100)$ instead of (3.11) in order to judge the precision of $\hat{\mu}$. That is, it is better to use the conditional distribution of U given $V = v$. We emphasize that in order for the reasoning made above to be valid it is essential for V to be S-ancillary and for the value of V to be known, otherwise we would have a mixed model which cannot be solved by simple conditioning.

In a case like in the example above it is quite reasonable to work with the conditional distribution given an ancillary statistic. Until now, no general argument has been showed to favour this procedure. We could question then if this situation is general or if it is due only to a peculiarity of this example. We now give an argument presented by Cox (1958) which shows that it is reasonable to condition on ancillary statistics.

Let X be a random variable whose distribution, P, is a member of the family of distributions $\mathcal{P}$. Let $U = u(X)$ be a B-ancillary statistic with respect to $\mathcal{P}$. Assume that an experiment was conducted where we observed $X = x$. Let $u = u(x)$. The experiment can be interpreted as being composed of two sub-experiments:

(I) We observe $U = u$;

(II) We observe that $X = x$ in the conditional distribution given $U = u$.

Let $\mathcal{P}(\,\cdot\,|U = u)$ be the family of distributions corresponding to experiment II. Evidently the two sub-experiments give jointly the same information as the original experiment. But since U is B-ancillary, and hence its distribution does not depend on $P \in \mathcal{P}$, sub-experiment I does not provide any information about $\mathcal{P}$. On the other hand, sub-experiment II evidently contains information about $\mathcal{P}$, since, $\mathcal{P}(\,\cdot\,|U = u)$ contains more than one element. That is, the information contained in sub-experiment II does not depend on the distribution of U, but on the value of U that was observed. Hence, the distribution of U is irrelevant for the purpose of inference but not the value of U. Since the information about P is all contained in experiment II, it is reasonable that we make conditional inference.

As it was already mentioned in Examples 3.5 and 3.6, conditional inference is, in some sense, more "robust" (or stable) than inference made with the full likelihood. This is due to the fact that the marginal distribution of U is not used, and hence, the validity of the conditional procedure only depends on the validity of the conditional distribution of $X|U = u$. That is, this procedure continues to be valid even if the marginal distribution is incorrect.

Once again we stress that the name ancillary, *i.e.*, auxiliary, is justified since an ancillary statistic is useful to decide which conditional distribution we should use. An ancillary statistic carries information about the precision of our conclusions about $P \in \mathcal{P}$.

If U is B-sufficient for $\mathcal{P}$, we can use similar arguments to the ones given in the case of ancillarity, dividing the experiment into the same two sub-experiments. Such arguments show that sub-experiment II does not contain information about $\mathcal{P}$ and that experiment I contains all the information about P. Then inference should be made using the marginal distribution of U. It happens that the marginal distribution of U is known *before* conducting the experiment, which implies that several statistical procedures automatically obey the sufficiency principle and hence in these cases the statistic obeys this principle without noticing it. It is obvious that the arguments given above can be extended to justify the ancillarity and sufficiency principles in the broad sense, and for the general nonformation principle.

Although Cox's example seems to provide an irrefutable argument in favour of the ancillarity and sufficiency principles, there exist examples where the application of these principles is less obvious. The following example, presented by Fisher (1935), marked the start of the discussion about ancillarity and illustrates this fact very well.

Example 3.9 *(Fisher's exact test)* The data in the table below were collected to investigate if there exists a relation between the criminal tendency of an individual and its genetic constitution. Among 13 monozygotic twin brothers and sisters of sentenced criminals, 10 were also sentenced, whereas among 17 bizygotic twin brothers and sisters of sentenced criminals only 2 were also sentenced.

	sentenced	not sentenced	total
monozygotic	10	3	13
bizygotic	2	15	17
	12	18	30

Let p_1 be the probability of a monozygotic twin brother or sister of a sentenced criminal to be also a criminal and p_2 be the corresponding probability of a bizygotic twin brother or sister. The following table represents the model for the general situation represented in this example,

X_1	$n_1 - X_1$	n_1
X_2	$n_2 - X_2$	n_2
X_+	$n_+ - X_+$	n_+

where X_1 and X_2 are independent random variables and $X_i \sim Bi(n_i, p_i)$, $i = 1, 2$.

In his 1935 article, Fisher suggested that in order to test the hypothesis $p_1 = p_2$, the conditional distribution given $X_+ = x_+$ should be used. This was based on the argument that X_+ does not contain information about the hypothesis of interest. It is convenient to define

$$\psi = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}\,.$$

Note that $p_1 = p_2$ if and only if $\psi = 1$. The distribution of X_+ is given by

$$\begin{aligned} P(X_+ = x_+) &= (1-p_1)^{n_1} p_2^{x_+} (1-p_2)^{n_2 - x_+} \\ &\quad \sum_y \left\{ \begin{pmatrix} n_1 \\ y \end{pmatrix} \begin{pmatrix} n_2 \\ x_+ - y \end{pmatrix} \psi^y I_{[0,n_1]}(y) I_{[0,n_2]}(x_+ - y) \right\}, \end{aligned}$$

and the conditional distribution of X_1 given X_+ is

$$P(X_1 = x_1 | X_+ = x_+) = \frac{\begin{pmatrix} n_1 \\ x_1 \end{pmatrix} \begin{pmatrix} n_2 \\ x_+ - x_1 \end{pmatrix} \psi^{x_1} I_{[0,n_1]}(x_1) I_{[0,n_2]}(x_+ - x_1)}{\sum_y \begin{pmatrix} n_1 \\ y \end{pmatrix} \begin{pmatrix} n_2 \\ x_+ - y \end{pmatrix} \psi^y I_{[0,n_1]}(y) I_{[0,n_2]}(x_+ - y)} \tag{3.12}$$

which depends on (p_1, p_2) only through ψ. Evidently X_+ is not B-ancillary nor S-ancillary with respect to ψ. Nevertheless, Fisher claimed that X_+ does not contain information about ψ and hence concluded that inference on ψ should be made using the conditional distribution (3.12). The test based on (3.12) is called Fisher's exact test, because the p-value of this test can be exactly computed using (3.12) contrary to the traditional Pearson's χ^2-test which uses the asymptotic χ^2-distribution.

In the last example, it is natural to question the fact that X_+ is ancillary with respect to ψ and at the same time that its distribution depends on ψ. As we will see, it will be possible to define concepts of ancillarity, even in the case where the distribution of the statistic depends on the parameter of interest. We will show that X_+ is M-ancillary for ψ and hence does not contain information about it, in an objective sense, which will confirm Fisher's intuition.

In 1935, Fisher used the example above to start the discussion about conditional inference. In the discussion that followed, this example arouse recurrently and given the difficulty to interpret the statistic X_+ as ancillary, it is not surprising that the debate turned out to be not very constructive. Besides, Fisher's imprecise way of expressing himself contributed to the fact that the true meaning of some basic concepts was not clear at all. If to this we add that these concepts were inconsistently used, it is not surprising that incorrect and irrelevant answers were given in the debate that followed, where several articles can be viewed as attempts to interpret Fisher's paper.

Fisher's main idea about conditional inference can be summarized in the ancillarity principle: inference should be made conditionally on an ancillary statistic.

In order to understand and use this principle it is necessary to answer the following questions:

(i) What is an ancillary quantity?

(ii) Why condition?

According to Fisher, a variable is ancillary if it does not contain information about the problem of interest. This is essentially the definition of ancillarity and, in a more general way, of nonformation given above. Evidently, we will need a more rigorous definition, a mathematical one, in order to continue. Such definition will have to contemplate the largest possible number of examples where our intuition leads us to use the nonformation principle. In this way, we will give an answer to question (i) above.

Regarding question (ii), the examples given before illustrate why we should follow the principle of nonformation, mainly Cox's example. Besides, Cox's argument about the representation of an experiment in terms of sub-experiments justifies the use of conditional inference.

3.2 S-nonformation

In this section we will study in detail the concept of S-nonformation, which, as it will be seen, is a generalization of the concept of B-nonformation. As we have already seen in Section 3.1, there exist many situations where the B-nonformation principle is not suitable as, for example, when only a part of the parameter is of our interest. Fraser (1956) made the first formulations of this kind when he introduced the concept of S-sufficiency and after that Sverdrup (1965) introduced the notion of ancillarity, both specially adapted for the situation where there is a nuisance parameter. Sandved (1967 and 1972) modified Sverdrup's suggestion, obtaining a quite similar notion to the one we will next describe.

3.2.1 Definition

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model, where the family $\mathcal{P}$ is of the form $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subseteq I\!R^k$. As before, let us consider a decomposition of the parameter $\theta = (\theta_1, \theta_2)$ where θ_1 has dimension k_1 and θ_2 has dimension k_2 (k_1 and k_2 can be both greater than 1) and $k_1 + k_2 = k$. We will assume that θ_1 is the parameter of interest and that θ_2 is the nuisance parameter, *i.e.*, we are interested in doing inference only on θ_1. Define

$$\begin{aligned} \Theta_1 &= \{\theta_1 \in I\!R^{k_1} : \exists\, \theta_2 \in I\!R^{k_2} \text{ such that } (\theta_1, \theta_2) \in \Theta\} \qquad \text{and} \\ \Theta_2 &= \{\theta_2 \in I\!R^{k_2} : \exists\, \theta_1 \in I\!R^{k_1} \text{ such that } (\theta_1, \theta_2) \in \Theta\}. \end{aligned}$$

Definition 3.10 *Consider the statistic $U : (\mathcal{X}, \mathcal{A}) \to (\mathcal{Y}, \mathcal{D})$ such that*

(i) $\Theta = \Theta_1 \times \Theta_2$.

(ii) The family of marginal distributions of U, $\mathcal{P}_U$, is parametrized by θ_1;

(iii) The family of conditional distributions given U, $\mathcal{P}(\cdot\,|U)$ is parametrized by θ_2.

In this case, U is called S-sufficient with respect to θ_1 and S-ancillary with respect to θ_2.

Later we will introduce the concept of nonformation, that has as particular cases the concepts of S-sufficiency and S-ancillarity defined above. We will consider a more general situation than the one in Definition 3.10. Assume that θ is the parameter of the model and that we are interested in making inference on a function of θ, say $\psi(\theta)$. Evidently the case above is included in this one, taking the function ψ as a projection of θ on the first coordinate (*i.e.*, $\psi(\theta) = \theta_1$, where $\theta = (\theta_1, \theta_2)$). We assume that there exist statistics U and V such that for each value v, in the domain of V, the family of conditional distributions

$$\{P_{U\theta}(\cdot|V = v) : \psi(\theta) = \psi_0\} \tag{3.13}$$

is the same for each value ψ_0 of the function ψ. We say then that (3.13) is *S-nonformative* with respect to ψ.

Example 3.11 Let X_1 and X_2 be independent random variables with Poisson distributions with parameters λ_1 and λ_2 respectively, i.e., $X_1 \sim Po(\lambda_1)$ and $X_2 \sim Po(\lambda_2)$. We can parametrize the distribution of (X_1, X_2) by $\theta = (\lambda_1, \lambda_2) \in \Theta = I\!R_+ \times I\!R_+$. Then the statistic $U = X_1$ is S-ancillary with respect to λ_2 and is S-sufficient with respect to λ_1 since

$$\begin{aligned} \mathcal{P}_U &= \{Po(\lambda_1) : \lambda_1 \in I\!R_+\}, \\ \mathcal{P}(\cdot|U) &= \{Po(\lambda_2) : \lambda_2 \in I\!R_+\} \text{ and} \\ \Theta &= I\!R_+ \times I\!R_+. \end{aligned}$$

Evidently an analogous result is valid for $V = X_2$.

Let us consider the submodel

$$\Theta = \{(\lambda_1, \lambda_2) \in I\!R_+ \times I\!R_+ : \lambda_2 \leq \lambda_1\} \quad .$$

In this case, $U = X_1$ is not S-sufficient for λ_1 nor S-ancillary for λ_2, since $\Theta \neq \Theta_1 \times \Theta_2$. This type of restriction in the parameters arises for example in the following way. In an experiment of counting radioactive particles, let us suppose that we have a source and that the experiment is conducted in two stages: first we measure the number of particles that are detected by some instrument (in per-time units), and next we measure the number of particles that arrive to the receptacle if we put a plaque between the source and the receptacle (in the same period of time). If λ_1 and λ_2 are the parameters of the distributions of the first stage and of the second stage respectively, then $\lambda_1 \geq \lambda_2$.

Define the parameters $\theta_1 = \lambda_1 + \lambda_2$ and $\theta_2 = \lambda_2/\lambda_1$. Evidently $\theta = (\theta_1, \theta_2) = (\lambda_1 + \lambda_2, \lambda_2/\lambda_1) \in I\!R_+ \times (0, 1]$ parametrizes the sub-model considered (θ_1 is the "total charge" and θ_2 is the "decay rate"). The statistic $U = X_1 + X_2$ is S-sufficient for θ_1 and is S-ancillary with respect to θ_2 since $\mathcal{P}_U = \{Po(\lambda_1 + \lambda_2) = Po(\theta_1) : \theta_1 \in I\!R_+\}$ only depends on θ_1 and it is easy to see that

$$X_1|U = u \sim Bi(u, \lambda_1/(\lambda_1 + \lambda_2)) = Bi(u, \frac{1}{1 + \theta_2}).$$

We should mention that Fisher (1950) claimed that inference on λ_2/λ_1 should be made conditionally on $X_1 + X_2$.

Example 3.12 Let us consider the independent random variables $X_{ij} \sim Po(\alpha_i\beta_j)$, with $i = 1, \ldots, r$ and $j = 1, \ldots, s$. The joint density of these random variables is given by:

$$\left[\prod_{i,j} e^{-(\alpha_i\beta_j)}\right] \left[\prod_{i,j} \frac{1}{x_{ij}!}\right] \exp\left[\sum_i x_{i+} \log \alpha_i + \sum_j x_{+j} \log \beta_j\right] .$$

This is an regular exponential family of order $r+s-1$. Let $\tilde{\alpha}_i = \frac{\alpha_i}{\alpha_+}$, $\tilde{\beta}_j = \frac{\beta_j}{\beta_+}$ and $\mu = \alpha_+\beta_+$. The joint density can be expressed in terms of these new parameters in the following way:

$$\frac{[\prod_i x_{i+}!][\prod_j x_{+j}!]}{x_{++}!} \frac{1}{\prod_{i,j} x_{ij}!} \times \begin{pmatrix} x_{++} \\ x_{1+} \ldots x_{s+} \end{pmatrix} \tilde{\alpha}_1^{x_{1+}} \ldots \tilde{\alpha}_s^{x_{r+}}$$

$$\times \begin{pmatrix} x_{++} \\ x_{+1} \ldots x_{+s} \end{pmatrix} \tilde{\beta}^{x_{+1}} \ldots \tilde{\beta}^{x_{+s}} \times \frac{\mu^{x_{++}}}{x_{++}!} e^{-\mu}.$$

The four factors of the expression above correspond respectively to the conditional distribution of $\{X_{ij}\}$ given $(X_{1+}, \ldots, X_{r+}, X_{+1}, \ldots, X_{+s})$, the conditional distribution of $(X_{1+}, \ldots, X_{r+})$ given X_{++}, the conditional distribution of $(X_{+1}, \ldots, X_{+s})$ given X_{++} and the density of X_{++}. Note that (X_{i+}) and (X_{+j}) are conditionally independent given X_{++}. We conclude that (X_{i+}) is S-ancillary with respect to the inference on $\tilde{\beta}_j$, as well as (X_{+j}) is S-ancillary with respect to $\tilde{\alpha}_i$.

Next we will analyse some effects that S-sufficiency and S-ancillarity can have on statistical inference. Let $T = t(X)$ be an S-sufficient statistic for θ_1 and S-ancillary for θ_2. We assume that $\mathcal{P}$ is dominated by a σ-finite measure μ, and let

$$f(x;\theta) = \frac{dP_\theta}{d\mu}(x).$$

Then the marginal density of T depends only on θ_1, since T is S-sufficient with respect to θ_1, *i.e.* it has density $f_T(t;\theta_1)$. Similarly, the conditional density given T is

$$f(x;\theta)/f_T(t;\theta_1) = f(x|t;\theta_2)$$

since T is S-ancillary with respect to θ_2. Then we have that

$$f(x;\theta) = f_T(t(x);\theta_1) f(x|t(x);\theta_2). \tag{3.14}$$

Therefore, the likelihood function for θ is

$$L(\theta_1, \theta_2; x) = L_1(\theta_1; t) L_2(\theta_2; x|t), \tag{3.15}$$

where $L_1(\theta_1;t) = f_T(t,\theta_1)$ and $L_2(\theta;x|t) = f(x|t;\theta_2)$. We assume that $L(\theta_1,\theta_2;x)$ has a unique maximum, $(\hat{\theta}_1(x), \hat{\theta}_2(x))$. Then $L_1(\hat{\theta}_1(x);t) = \max L_1(\theta_1;t)$ where the maximum is taken over $\theta_1 \in \Theta_1$. Therefore, we obtain the same estimate of θ_1 using the original

likelihood L or the marginal L_1. Note that $\hat{\theta}_1(x)$ depends on x only through $T = t(x)$, say $\hat{\theta}_1(x) = h(t(x))$. Since θ_1 is the parameter of interest, we can "reduce" the data to $T = t(x)$ and estimate θ_1, as well as assess the distribution of $\hat{\theta}_1(x) = h(t(x))$, using the marginal distribution of T.

The situation for θ_2 (still under (3.14) and (3.15)) is different. We have that

$$L_2(\hat{\theta}_2(x); x|t) = \max L_2(\theta_2; x|t),$$

where $\theta_2 \in \Theta_2$, so that it is the same to estimate θ_2 using the original likelihood L or the conditional likelihood L_2. Meanwhile, it turns out that it makes a difference to use the conditional likelihood to assess the statistical properties of $\hat{\theta}_2(x)$, as it can be seen in Examples 3.5 and 3.6 of Section 3.1.

We will next give an interpretation of the notions of S-ancillarity and S-sufficiency in terms of the score function. We define the score function, $U : \Theta \longrightarrow I\!R^k$ as

$$U(\theta_1, \theta_2) = \begin{pmatrix} U_1(\theta_1, \theta_2) \\ U_2(\theta_1, \theta_2) \end{pmatrix} = \frac{\partial \log L(\theta_1, \theta_2; x)}{\partial(\theta_1, \theta_2)},$$

where $U(\theta_1, \theta_2)$ and $U_2(\theta_1, \theta_2)$ are the components of $U(\theta_1, \theta_2)$ of dimensions k_1 and k_2, respectively. Then (3.15) is equivalent to

$$U_1(\theta_1, \theta_2) = U_1(\theta_1) = \frac{\partial \log L_1(\theta_1; t)}{\partial \theta_1}$$

and

$$U_2(\theta_1, \theta_2) = U_2(\theta_2) = \frac{\partial \log L_2(\theta_2; x|t)}{\partial \theta_2}.$$

Hence, S-ancillarity implies that $U_1(\theta_1, \theta_2)$ depends on θ only through θ_1, and S-sufficiency implies that $U_2(\theta_1, \theta_2)$ depends on θ only through θ_2.

The next example is somewhat more complex than the others, where we will illustrate a more recent and realistic application. As it will be seen, we will have to extend the notion of S-nonformation, if we wish to take into account cases like this.

Example 3.13 *(Incubation time of the AIDS virus)* The incubation time of the AIDS virus is very long and, in general, it is not known with precision when the infection took place. Therefore, to determine the incubation time (i.e., the period of time elapsed between the infection and the time of appearance of the first symptoms) only data of people infected by blood transfusion was used, where it is known exactly the date of infection. Let us say that the study has been made in 1987. The observations are pairs (X_i, Y_i), where X_i is the time of infection and $Y_i \leq 1987$ is the time of appearance of the first symptoms.

Let us say that the incubation time has distribution function $F(\cdot; \psi)$ parametrized by ψ, i.e.,

$$F(z; \psi) = P(Y_i - X_i \leq z), \quad z \geq 0.$$

We assume that people are contaminated according to a non-homogeneous Poisson process of intensity $h(t)$, that is, the number of contaminated people by blood transfusion in the time interval (t_1, t_2) has Poisson distribution with mean $\int_{t_1}^{t_2} h(t)dt$. Therefore, the process of times of contamination, given by the X_is above, is also a Poisson process with intensity

$$\phi(t) = h(t)F(T-t;\psi), \tag{3.16}$$

where $T = 1987$ is the date of conclusion of the research. That is, the intensity of the process of times of infection is multiplied by $F(T-t;\psi)$, which gives us the probability that such an individual contaminated at time t manifests symptoms before the conclusion of the research and thus that such an individual will be included in the sample. We know then the number of observations $N = n$ and the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. The likelihood function can be found as the marginal distribution of $(N, X_1, \ldots, X_N)$ multiplied by the conditional density of $(Y_1, \ldots, Y_n)$. We will assume that given $(N, X_1, \ldots, X_N)$, $Y_1, \ldots, Y_n$ are independent and Y_i has density $f(y - X_i;\psi)/F(T - x_i;\psi)$, $f(z;\psi) = F'(z;\psi)$. Given $N = n$, $X_1, \ldots, X_n$ are independent, and X_i has density

$$\frac{\phi(x)}{\int_{T_0}^{\top} \phi(t)dt} \ ,$$

where T_0 is the initial time of the epidemic. Finally N has Poisson distribution with mean $\int_{T_0}^{\top} \phi(t)dt$. The likelihood function is

$$\begin{aligned} L(\psi; h(\cdot)) &= \left\{\prod_{i=1}^{n} \frac{f(y_i - x_i;\psi)}{F(T-x_i;\psi)}\right\} \left\{\frac{1}{n!}\prod_{i=1}^{n}\sum h(x_i)F(T-x_i;\psi)\right\} \\ &\quad \exp\left\{-\int_{T_0}^{\top} h(t)F(T-t;\psi)dt\right\} \\ &= \left\{\prod_{i=1}^{n} \frac{f(y_i - x_i;\psi)}{F(T-x;\psi)}\right\} \left\{\frac{1}{n!}\prod_{i=1}^{n}\phi(x_i)\exp\left[-\int_{T_0}^{\top}\phi(t)dt\right]\right\}. \end{aligned} \tag{3.17}$$

We are taking h, in (3.16), as a parameter. Note that $h(\cdot)$ is a function which is contained in an infinite-dimensional space, and hence the model we are using is non-parametric. Since part of this model is parametric (parametrized by ψ) and the other is non-parametric, we say that this is a semi-parametric model. Evidently we cannot use the definitions of S-ancillarity and S-sufficiency given in Definition 3.10. We will see that these concepts can be extended so as to include this case. We will see that if we allow $h(\cdot)$ to vary freely then $(N, X_1, \ldots, X_N)$ is, in some sense, S-ancillary for ψ. Roughly speaking, this is due to the fact that since $h(\cdot)$ varies freely and $\psi \in \Psi$, then, by (3.16), we see that $\phi(\cdot)$ also varies freely. Later we will need a much more precise meaning of $(N, X_1, \ldots, X_N)$ being S-ancillary for ψ. In this way, we make inference on ψ using only the first factor of (3.17). In the literature it has already been studied the use of the full likelihood given by (3.16), modelling $h(t)$ as $\exp(a + bt)$, where a and b are parameters. In the latter case $(N, X_1, \ldots, X_N)$ is not S-ancillary any more.

It is important to stress the difference between the case where no assumptions and some assumptions on $h(\cdot)$ are made. If we do not make any assumptions on h, inference on ψ should be based on the first factor of (3.17). In this case, if our model for the incubation time, i.e., for the form of $F(\,\cdot\,;\psi)$ is sensible, the model as presented would lead us to sensible conclusions about ψ. Meanwhile, if we make assumptions on $h(\cdot)$, the second factor of (3.17) turns out to have importance because of (3.16). Therefore, if we fail in specifying h, we can err when making inference on ψ. We see then that it is safer, in some sense, to use a more flexible model (letting h vary freely) and make inferential separation than to try to better specify the statistical model. This can be crucial in the process of modelling when there is still not enough information about the phenomenon which is being modelled, to justify assumptions on h.

Based on data from the U.S., using the conditional likelihood and assuming that the distribution of the incubation time is Weibull which has a distribution function given by

$$F(t,\psi) = 1 - \exp\{-(\alpha t)^{\beta}\}$$

where $\psi = (\alpha, \beta)$, it was estimated that $\alpha = 0.07$ and $\beta = 2.5$, with t measured in years. Such a distribution has median 12.3 and the probability of an incubation time of 7 or less years is only 0.155.

3.2.2 S-nonformation in exponential families

A detailed treatment of S-ancillarity in exponential families can be found in Barndorff-Nielsen and Blæsild (1975).

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a family parametrized by θ, with a decomposition of the parameter given by $\theta = (\theta_1, \theta_2)$. If there exists a factorization of the likelihood of the form

$$L(\theta) = L_1(\theta_1)L_2(\theta_2), \quad \theta = (\theta_1, \theta_2) \in \Theta,$$

then θ_1 and θ_2 are L-independent (where "L" comes from "Likelihood"). If L_1 and L_2 correspond respectively to the conditional and marginal likelihood for a statistic, say, U, then U is called a cut. In the case of a cut the factorization hence takes the form

$$L(\theta) = L(\theta_1; y|u)L(\theta_2; u).$$

In this case, θ_1 and θ_2 are called L-independent parameters corresponding to the cut. If $\Theta = \Theta_1 \times \Theta_2$, where $\theta_1 \in \Theta_1$ and $\theta_2 \in \Theta_2$ vary freely (and are L-independent) it is evident (by already given arguments) that the maximum likelihood estimator $(\hat{\theta}_1, \hat{\theta}_2)^\top$ can be found maximizing $L_1(\theta_1)$ and $L_2(\theta_2)$ separately.

In the case of exponential families, it is common to find examples where a mixed parametrization provides L-independent parameters and hence one component of the canonical statistic is a cut. We will present a theorem of Barndorff-Nielsen and Blæsild (1975) that will give us a necessary and sufficient condition for this to happen.

Let $\mathcal{P}$ be a regular exponential family of order k and minimal representation

$$\frac{dP_\theta}{d\nu}(x) = a(\theta)e^{\theta\cdot t(x)} \quad [\nu], \quad \theta \in \Theta\,.$$

Following the notation of Section 1.5, let A be the $m\times k$ matrix of rank m, B the $(k-m)\times k$ matrix of rank $k-m$, such that $AB^\top = 0$ and $V = v(X) = At(X)$, where X has distribution in $\mathcal{P}$. We define $\psi(\theta) = \mathrm{E}_\theta V = A\tau(\theta)$ and

$$\sigma(\theta) = B\theta,$$

as the components of the mixed parameter. In Theorem 1.24 we showed that the domain of (ψ,σ) is the Cartesian product $\psi(\Theta)\times\sigma(\Theta) = \Psi\times\Sigma$. A function $f:\Theta\to I\!R^k$ is called additive if

$$f(\psi,\sigma) + f(\psi_0,\sigma_0) = f(\psi,\sigma_0) + f(\psi_0,\sigma)$$

$\forall\psi,\psi_0\in\Psi$ and $\sigma,\sigma_0\in\mathcal{Z}$.

Theorem 3.14 *(Barndorff-Nielsen and Blæsild, 1975) A statistic V is a cut and is S-ancillary for σ if and only if, θ and $\log a$ are additive as functions of ψ and σ.*

Proof: By Theorem 1.24, ψ and σ are variationally independent. Evidently, the conditional distribution of X given V depends on θ only through $\sigma(\theta)$. Therefore, the ancillarity here only demands that the marginal distribution of V is independent of σ.

We first assume that this is true. Then by (3.14) we have that

$$f(x;\psi,\sigma) = g(v(x);\psi)h(x|v(x);\sigma), \tag{3.18}$$

where f, g and h are the densities of X, V and X given V, respectively. The densities are positive, thus

$$\frac{f(x;\psi,\sigma)f(x;\psi_0,\sigma_0)}{f(x;\psi_0,\sigma)f(x;\psi,\sigma_0)} = 1 \tag{3.19}$$

that is

$$\begin{aligned}&\{\theta(\psi,\sigma)+\theta(\psi_0,\sigma_0)-\theta(\psi,\sigma_0)-\theta(\psi_0,\sigma)\}^\top t(x)\\ &= -\log a(\psi,\sigma) - \log a(\psi_0,\sigma_0) + \log a(\psi,\sigma_0) + \log a(\psi_0,\sigma).\end{aligned}$$

The components of t are affinely independent, which shows that θ and $\log a$ are additive. On the other hand, if θ and $\log a$ are additive, then obviously, (3.19) is satisfied. This can be written in the following way

$$\frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}}(x) = \frac{dP_{\psi_0,\sigma}}{dP_{\psi_0,\sigma_0}}\frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}}.$$

The second factor is

$$\frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}}(x) = \frac{a(\psi,\sigma_0)}{a(\psi_0,\sigma_0)}e^{\{\theta(\psi,\sigma_0)-\theta(\psi_0,\sigma_0)\}^\top t(x)}.$$

Now, $B\{\theta(\psi,\sigma_0)-\theta(\psi_0,\sigma_0)\}=0$, hence $\theta(\psi,\sigma_0)-\theta(\psi_0,\sigma_0)=A^\top\eta$ for some η. Thus

$$\frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}}(x)=\frac{a(\psi,\sigma_0)}{a(\psi_0,\sigma_0)}e^{\eta^\top At(x)}. \tag{3.20}$$

The marginal density of $V=At(X)$ can be found through the expression:

$$\begin{aligned} g(v;\psi,\sigma) &= \int \tfrac{dP_{\psi,\sigma}}{dP_{\psi_0,\sigma_0}} dP_{V\psi_0,\sigma_0} \\ &= \int \tfrac{dP_{\psi_0,\sigma}}{dP_{V\psi_0,\sigma_0}} \tfrac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}} dP_{V\psi_0,\sigma_0} \end{aligned}$$

which by (3.20), is equivalent to

$$\begin{aligned} g(v;\psi,\sigma) &= \frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}} \int \frac{dP_{\psi_0,\sigma}}{dP_{\psi_0,\sigma_0}} dP_{V\psi_0,\sigma_0} \\ &= \frac{dP_{\psi,\sigma_0}}{dP_{\psi_0,\sigma_0}} f(v;\sigma). \end{aligned}$$

Since $\int g(At(x);\psi,\sigma)P_{\psi_0,\sigma_0}(dx)=1$ we have that

$$\int f(At(x),\sigma)P_{\psi,\sigma_0}(dx)=1,$$

for any ψ. We use now the fact that V is complete if σ is fixed, then

$$f(At(x),\sigma)=1.$$

From here it follows that $g(v;\psi,\sigma)$ does not depend on σ. □

It is evident that a mixed parametrization is not always associated with a cut. It is enough to see that in the normal distribution, $N(\mu,\sigma^2)$ provides a mixed parametrization, but $\bar{X}_+$ is not a cut.

We will now see an example of a cut in an exponential family.

Example 3.15 Let P be the regular exponential family of trinomial distributions, given by the density

$$\frac{dP_{p_1,p_2}}{d\nu}(x)=\frac{n!}{x_1!x_2!(n-x_1-x_2)!}p_1^{x_1}p_2^{x_2}(1-p_1-p_2)^{n-x_1-x_2},$$

with respect to the counting measure, ν, where $p=(p_1,p_2)^\top\in\{(p_1,p_2)^\top : p_1>0, p_2>0, p_1+p_2<1\}$. Let $V=X_1+X_2$. The corresponding mixed parametrization is given by

$$\psi=E_P(V)=n(p_1+p_2)$$

$$\sigma=\theta_1-\theta_2=\log\frac{p_1}{p_2},$$

where $A = (1, 1)$, $B = (1, -1)$ and $\theta_i = \log\{p_i/(1 - p_1 - p_2)\}$, $i = 1, 2$ are the canonical parameters. Then

$$\theta_1 = \log \frac{\psi}{n - \psi} + \log \frac{e^\sigma}{1 + e^\sigma}$$

$$\theta_2 = \log \frac{\psi}{n - \psi} + \log \frac{1}{1 + e^\sigma},$$

which is obviously an additive function. We also have that

$$\log a(\theta) = -n \log(1 - p_1 - p_2) = -n \log \frac{n - \psi}{n}$$

which is also additive. Hence, using Theorem 2.16, V is a cut and S-ancillary for σ. In this example it is easy to verify this directly, because

$$\frac{dP_{\psi,\sigma}}{d\nu}(x) = \frac{n!}{x_1!x_2!(n - x_1 - x_2)!}(\frac{\psi}{n})^v(1 - \frac{\psi}{n})^{n-v}e^{\sigma x_1}(1 + e^\sigma)^{-v}.$$

Thus, $V \sim Bi(n, \psi/n)$ and the conditional distribution of X_1, given V is $X_1|V = v \sim Bi(v, e^\sigma/(1 + e^\sigma))$. It follows that V is S-ancillary for the inference on σ and S-sufficient for the inference on ψ.

3.3 G-nonformation

The concept of G-nonformation is based on the notion of a transformation model. Together with the exponential families, the composite transformation models form the two most important types of statistical models. Many models are members of both classes, and form the class of transformation exponential models. The families of normal and gamma distributions are examples of models of this last type, since they are exponential families and they are closed under transformations of scale, which implies that, as we will see, that they are also transformation models.

In Section 3.3.1 we will consider more basic aspects of transformation models. A broader introduction can be found in Barndorff-Nielsen (1988), see also Barndorff-Nielsen, Blæsild, Jensen, and Jørgensen (1982) and Barndorff-Nielsen, Blæsild and Eriksen, (1989). Section 3.3.2 deals with the concept of nonformation especially adapted for transformation models, namely G-nonformation. This concept generalizes the concept of G-sufficiency, that was introduced by Barnard (1963). We will use the techniques developed here to treat the non-trivial example of Cox's proportional risks which will be left for Section 3.3.3.

3.3.1 Transformation models

Next we will give the basic concepts of the theory of transformation models. To do so, we will review some basic definitions of the theory of groups.

Let $\mathcal{Y}$ be a given set. The class of injective transformations from $\mathcal{Y}$ into $\mathcal{Y}$ with the operation of composition is a group, called the symmetric group of $\mathcal{Y}$ and will be denoted by $\mathrm{S}(\mathcal{Y})$. Given a group G, we say that G acts on $\mathcal{Y}$ if there exists a homomorphism $\gamma : G \to \mathrm{S}(\mathcal{Y})$. In this case, we say that the homomorphism γ is an action of the group G on $\mathcal{Y}$, and we use the notation

$$\gamma(g)(y) = gy = g(y) \ ,$$

for $g \in G$ and $y \in \mathcal{Y}$. This notation is justified by the fact that G and $\gamma(G) \subseteq \mathrm{S}(\mathcal{Y})$ are homomorphic and hence algebraically equivalent.

We say that the action γ is transitive if, for each y_1 and $y_2 \in \mathcal{Y}$, there exists a $g \in G$ such that $y_1 = gy_2$. We say that the action γ is free if for each g_1 and $g_2 \in G$ with $g_1 \neq g_2$, we have that $g_1 x \neq g_2 x$ for all $x \in \mathcal{Y}$.

Given $y_0 \in \mathcal{Y}$ an orbit of y_0 is the set $Gy_0 = \{gy_0 : g \in G\}$. Evidently the collection of the orbits of $\mathcal{Y}$ gives us a partition of $\mathcal{Y}$. Assume that on each orbit of $\mathcal{Y}$ we choose a fixed element which we will call a representative of the orbit. Then each point $y \in \mathcal{Y}$ can be determined by specifying the representative of the orbit to the which y belongs, say y_0, and an element of the group G, say g_0, that transforms y_0 into y, i.e., $y = g_0 y_0$. In this way, we express y by two components $y = (g_0, y_0)$ and we call this decomposition an orbital decomposition. Note that unless the action of the group is free, the first component of the orbital decomposition is not unique.

Let us consider the function $t : \mathcal{Y} \to \mathcal{Z}$. A function t is called invariant if $t(gy) = t(y)$, for all $y \in \mathcal{Y}$ and $g \in G$, that is, t is constant on the orbits of $\mathcal{Y}$. If a function t is invariant and also it is such that $t(y) = t(y')$ implies that $y = g(y')$, for some $g \in G$, then t distinguishes the orbits of $\mathcal{Y}$ and we say that t is maximum invariant. If $t(x) = t(x')$ implies that $t(gx) = t(gx')$, $\forall g \in G$ we say that t is equivariant. In this case, we can define $gt(x) = t(gx)$, which defines an action of G on $t(\mathcal{Y})$.

We will consider now a situation where we have a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ and a group G that acts on the sample space $\mathcal{X}$, with action γ. We will assume, from now on, that the action is such that the transformations $\gamma(g) : x \mapsto gx$ are $\mathcal{A}$-measurable for all $g \in G$. Then for each $g \in G$ and each measure ν on $(\mathcal{X}, \mathcal{A})$ we can define the g-transformed measure as

$$g(\nu)(B) = \nu(g^{-1}(B)), \ \ \forall B \in \mathcal{A}.$$

If a random variable X has distribution $P \in \mathcal{P}$ then the transformed random variable $g(X)$ has distribution $g(P)$. Hence, the action of G on $\mathcal{X}$ induces an action on the space of all as measures in $(\mathcal{X}, \mathcal{A})$ and in particular, on the space of all probability measures on $(\mathcal{X}, \mathcal{A})$. A measure ν such that $g(\nu) = \nu \ \forall g \in G$ is called an invariant measure.

If the family $\mathcal{P}$ is given by $\mathcal{P} = \{gP : g \in G\}$, for some $P \in \mathcal{P}$, we say that $\mathcal{P}$ is a transformation model generated by G. Note that the definition given above is independent of the choice of $P \in \mathcal{P}$. Let ν be an invariant measure. Then the densities of $\mathcal{P}$ with respect to a ν are given by

$$\frac{dgP}{d\nu}(x) = \frac{dgP}{dg\nu}(x) = \frac{dP}{d\nu}(g^{-1}x).$$

Thus, if f is the density with respect to an invariant measure ν, then the family of densities $\{f(g^{-1}\cdot) : g \in G\}$ corresponds to the transformation model.

The statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is called closed under the action of G if $P \in \mathcal{P} \Rightarrow g(P) \in \mathcal{P}$, $\forall g \in G$. Note that we can associate to the action γ of the group G on $\mathcal{X}$ the action $\bar{\gamma}$ of G on $\mathcal{P}$ given by $\bar{\gamma}(g)P = g(P)$, $\forall P \in \mathcal{P}$, $g \in G$. In this way, it makes sense to speak of the orbits of $\mathcal{P}$. Evidently, a transformation model generated by G is a model which is closed under the action of G and is composed of a single orbit, that is, the action of G on $\mathcal{P}$ is transitive. A model closed under the action of G with more than one orbit is called a composite transformation model. It is clear that every composite transformation model is a disjoint union of its orbits and hence of transformation models.

We assume now that $\mathcal{P}$ is parametrized by two parameters ψ and ϕ,

$$\mathcal{P} = \{P_{\psi,\phi} : \psi \in \Psi, \phi \in \Phi\},$$

such that for each $g \in G$ we have that $g(P_{\psi,\phi}) = P_{\psi,\phi'}$ for some $\phi' \in \Phi$. We will denote ϕ' by $g(\phi)$. In this case, the parameter ϕ is called the *group parameter* and ψ the *index parameter* or *invariant parameter*.

The definitions will now be illustrated with some examples.

Example 3.16 *(Location-scale model)* Let f be a density in $\mathbb{R}$. Let us consider the location and scale family

$$\mathcal{P} = \{P_{\mu,\sigma} : (\mu, \sigma) \in \Omega\}$$

where $\Omega = \mathbb{R} \times \mathbb{R}_+$ and for each $(\mu, \sigma) \in \Omega$ the distribution $P_{\mu,\sigma}$ has density

$$f(x; \mu, \sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right), \quad \forall x \in \mathbb{R}.$$

This model is generated by the group G given by:

$$G = \{(\mu, \sigma) : \sigma > 0, \mu \in \mathbb{R}\},$$

with the operation of composition given by

$$(\mu_1, \sigma_1) \circ (\mu_2, \sigma_2) = (\sigma_1\mu_2 + \mu_1, \sigma_1\sigma_2).$$

The action of G on $\mathbb{R}$ is given by

$$(\mu, \sigma)x = \mu + \sigma x, \quad \forall \quad (\mu, \sigma) \in G, \quad x \in \mathbb{R}.$$

It is easy to see that G generates the family $\mathcal{P}$ and that $\mathcal{P}$ is a transformation model. The action of G is transitive on $\mathbb{R}$ and on $\mathcal{P}$.

In the case where f is the density of the standard normal distribution, the action of G on $\mathcal{P} = \{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$ is free, and (μ, σ) parametrizes $\mathcal{P}$.

Example 3.17 *(The family of gamma distributions)* Let $X_1, \ldots, X_n$ be independent random variables with distribution

$$X_i \sim Ga(\beta, \lambda), \quad i = 1, \ldots, n.$$

Since,

$$cX_i \sim Ga(c\beta, \lambda),$$

for $c \in I\!R_+$, then, the model is closed under the action of the group of scale transformations. Hence we have a composite transformation model. The parameter λ is invariant and β is the group parameter. The joint density of $X_1, \ldots, X_n$ is

$$f(x; \beta, \lambda)dx = \frac{\beta^{n\lambda}}{\Gamma(\lambda)^n} \prod_{i=1}^{n} x_i^{\lambda} \exp\left(-\beta^{-1} \sum_{i=1}^{n} x_i\right) \nu(dx),$$

where $\nu(dx) = \prod_{i=1}^{n}(x_i^{-1}dx_i)$ is an invariant measure.

Example 3.18 *(The family of distributions of von Mises-Fisher)* Let us consider the distribution of von Mises-Fisher introduced in Example 1.11. Let $X \sim vM_k(\mu, \lambda)$, with the following density with respect to the surface measure of S^{k-1},

$$f(x; \mu, \lambda) = a(\lambda)e^{\lambda\mu \cdot x},$$

with $\mu, x \in S^{k-1}$ and $\lambda \geq 0$. Consider the group of orthogonal transformations of S^{k-1} given by $x \mapsto Ax$, where A is an orthogonal matrix ($A^TA = I$). The measure of surface of S^{k-1} is invariant under these transformations, and hence AX has density

$$f(A^{-1}x; \mu, \lambda) = a(\lambda)e^{\lambda\mu^\top A^{-1}x} = a(\lambda)e^{\lambda(A\mu)\cdot x} = f(x; A\mu, \lambda),$$

where we used that $A^\top = A^{-1}$, which shows that $AX \sim vM_k(A\mu, \lambda)$. Hence, the family is a composite transformation model with index parameter λ.

Note that the action of the group on S^{k-1} is not free, since multiplication by A gives us a rotation that in general leaves two fixed points.

There exist many other important examples of composite transformation models, many of which come from multivariate analysis. For example, the model of multivariate analysis of variance (see Barndorff-Nielsen, 1988, pp. 75-78) and Example 3.22. At this stage the reader should have perceived that we only gave examples of transformation models constructed with continuous distributions. In general, it is not very interesting to consider discrete parametric models generated by groups, since, if the sample space $\mathcal{X}$ is discrete, the symmetric group $S(\mathcal{X})$ is discrete and then the parameter space is also discrete.

3.3.2 Definition of G-nonformation

We will consider a family $\mathcal{P}$ of probability measures on $(\mathcal{X}, \mathcal{A})$, parametrized by Θ, that is $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let $\psi = \psi(\theta) \in \psi(\Theta)$ be the parameter of interest. For each $\psi_0 \in \psi(\Theta)$ define the family $\mathcal{P}(\psi_0) = \{P_\theta \in \mathcal{P} : \psi(\theta) = \psi_0\}$. In this context we have the following definition.

Definition 3.19 *If for each $\psi_0 \in \psi(\Theta)$ the family $\mathcal{P}(\psi_0)$ is a transformation model generated by a group of transformations with transitive action on $\mathcal{X}$, then $\mathcal{P}$ is called G-nonformative with respect to ψ.*

The concept of G-nonformation can be justified through the concept of perfect adjustment, that we will give next. This concept is very important for the study of nonformation and will be discussed from different points of view throughout this chapter.

To begin this discussion, let us consider the case where the family $\mathcal{P}(\psi_0)$ is determined by the family of densities

$$\mathcal{D}_{\psi_0} = \{f(g^{-1}\,\cdot) \,:\, g \in G\},$$

where f is a density of a given probability in $\mathcal{P}(\psi_0)$, with respect to an invariant measure ν. Note that under general regularity conditions, there exists such an invariant measure. We assume that the density f has mode point $x_0 \in \mathcal{X}$. Then $f(g^{-1}\,\cdot)$ has mode point gx_0. If the action of G on $\mathcal{X}$ is transitive, then any x in $\mathcal{X}$ is a mode point of some density $f(g^{-1}\,\cdot)$. Therefore, if we observe the value $x \in \mathcal{X}$, there exists $g \in G$ such that $f(g^{-1}\,\cdot) \in \mathcal{D}_{\psi_0}$ has mode point x, which is interpreted as a perfect adjustment of $f(g^{-1}\,\cdot)$, in the sense that the mode point of the density is the most plausible observation among the possible ones.

In the case of the definition of G-nonformation, a similar reasoning as in the previous paragraph is valid for each family $\mathcal{P}(\psi)$ with $\psi \in \psi(\Theta)$, that is, given $x \in \mathcal{X}$, each family $\mathcal{P}(\psi)$ has at least one element that adjusts perfectly to x. The following reasoning shows that the family $\mathcal{P}$ together with the observed value x do not contain information about ψ. Let ψ_0 and ψ_1 be two values of ψ. Given any observed value $x \in \mathcal{X}$, $\mathcal{P}(\psi_0)$ as well as $\mathcal{P}(\psi_1)$, give a perfect adjustment for x. Therefore, it is not reasonable that we prefer the value of ψ_0 over ψ_1 or vice-versa, based in the quality of the adjustment. This argument is valid for any pair of values of ψ. Note that if it is possible to obtain some information about ψ, through x and of the family $\mathcal{P}$, then there should be at least two values of ψ for which we could prefer one over the other. We conclude that the family $\mathcal{P}$ together with the observed value x do not contain information about ψ.

We will now consider the definitions of G-sufficiency and G-ancillarity. We assume that the likelihood of θ can be factorized in the following way:

$$L(\theta) = g(u; \psi(\theta))h(x|u; \theta),$$

where $U = u(X)$ is a statistic. If the parameter $\psi = \psi(\theta)$ parametrizes the family of marginal distributions of U and if for each u_0 in the domain of U the family $\{h(x|u_0; \theta) : \theta \in \Theta\}$ is

G-nonformative with respect to ψ then U is called G-sufficient with respect to the inference on ψ.

Let $V = v(X)$ be a statistic such that,

$$L(\theta) = h(x|v; \psi(\theta))g(v; \theta).$$

We assume that ψ parametrizes the family of the conditional distributions given V, and that the family $\{g(v; \theta) : \theta \in \Theta\}$ is G-nonformative with respect to ψ. In this case, we say that V is G-ancillary with respect to the inference on ψ.

Example 3.20 In many cases it is convenient to combine several definitions of nonformation, to obtain a reduced model to make inference. In the case of the normal distribution, $N(\mu, \sigma^2)$, we first use B-sufficiency to conclude that $(\bar{X}_+, s^2)$ is B-sufficient for the parameter (μ, σ^2) and hence we will use $(\bar{X}_+, s^2)$ to make inference on (μ, σ^2). We assume now that σ^2 is the parameter of interest. We will show that s^2 is G-sufficient for σ^2. First, the distribution of s^2 depends only on σ^2 and is parametrized by σ^2. Besides, $\bar{X}_+$ and s^2 are independent, and hence the conditional distribution of $\bar{X}_+$ given s^2 is the marginal distribution of $\bar{X}_+$, that is $N(\mu, \sigma^2/n)$. Let us consider now the group G of translations in $\mathbb{R}$. Evidently a translation $x \to x + c$ transforms the distribution $N(\mu, \sigma^2/n)$ into the distribution $N(\mu + c, \sigma^2/n)$, and hence, the action of G the family is transitive. Thus, the model

$$\{N(\mu, \sigma^2/n) : \mu \in \mathbb{R}\}$$

is generated by a group of translations and a transitive action of G on $\mathbb{R}$. Hence, s^2 is G-sufficient for σ^2. A similar argument shows that $\bar{X}_+$ is G-ancillary for σ^2. Therefore, inference on σ^2 should be made using the conditional distribution of s^2 given $\bar{X}_+$. Now, as $\bar{X}_+$ and s^2 are independent, this equivalent to using the marginal distribution of s^2. That is, G-sufficiency of s^2 as well as G-ancillarity of $\bar{X}_+$ lead us to the same exhaustive model, namely, the marginal distribution of s^2.

Let us consider again the parametric family $\mathcal{P} = \{P_{\psi,\phi} : \psi \in \Psi, \phi \in \Phi\}$ where ϕ is the group parameter and ψ is invariant, and let us suppose that if we want to make inference on ψ, ϕ is a nuisance parameter and that we have a statistic $U = u(X)$, such that $u(\cdot)$ is maximum invariant. Then

$$P_{\psi,g(\phi)}(u(X) \in A) = P_{\psi,\phi}(u(gX) \in A) = P_{\psi,\phi}(u(X) \in A),$$

$\forall A \in \mathcal{B}$. Hence, the distribution of U depends on (ψ, ϕ) only through ψ.

We will show that U is G-sufficient for ψ. Note that the conditional distribution of X given U is in an orbit Gx_0, since U is maximum invariant. Evidently, the action of G on Gx_0 is transitive. Besides, for each $A \in \mathcal{B}$ we have,

$$\begin{aligned} g(P_{\psi,\phi})(X \in A|U = u) &= P_{\psi,\phi}(g(X) \in A|U = u) \\ &= P_{\psi,g\phi}(X \in A|U = u), \end{aligned}$$

that is, the family of conditional distributions of X given U is generated by a transitive group. Hence, this model is G-nonformative with respect to ψ and then U is G-sufficient for ψ.

We now assume that ψ is known, that is, we are now considering a family $\mathcal{P}_\psi = \{P_{\psi,\phi} : \phi \in \Phi\}$. We then have a transformation model. Since the distribution of U depends only on ψ (which is now known) we conclude that U is B-ancillary. Hence we will see that we should use the conditional distribution of X given U to make inference on ϕ. Note that we need that ϕ parametrizes the family of conditional distributions.

Example 3.21 *(Continuation of Example 3.8)* The inverse Gaussian distribution follows the transformation law

$$cN^-(\chi, \psi) = N^-(c\chi, c^{-1}\psi).$$

Hence, it is closed under the action of the group of transformations of scale and is a composite transformation model. For n observations $(X_1, \ldots, X_n)$, the action is

$$(X_1, \ldots, X_n) \to (cX_1, \ldots, cX_n).$$

Using the results of Example 3.7, we have that in the space of the sufficient statistic the action is

$$(X_-, X_+) \to (c^{-1}X_-, cX_+).$$

The parameter of the group is $\mu = \sqrt{\chi/\psi}$ and $\omega = \sqrt{\chi\psi}$ is the invariant parameter. In the space of (X_-, X_+), $T = (X_-X_+)^{\frac{1}{2}}$ is maximum invariant. Hence, inference on ω should be made using the marginal distribution of T.

If ω is known then inference on μ should be made using the conditional distribution of $S = (X_+/X_-)^{\frac{1}{2}}$ given $T = t$, which is parametrized by μ.

Example 3.22 We will consider the inference on a correlation matrix of the multivariate normal distribution. Let $X_1, \ldots, X_n$ be independent random vectors with distribution $X_i \sim N_k(\mu, \Sigma)$, where $\mu \in I\!R^k$ and Σ is a $k \times k$ positive-definite matrix .

Let $\rho = \{\rho_{ij}\}_{i,j=1,\ldots,k}$, with

$$\rho_{ij} = \frac{\Sigma_{ij}}{(\Sigma_{ii}\Sigma_{jj})^{\frac{1}{2}}},$$

be the correlation matrix, which will be our parameter of interest. Let

$$X_+ = X_1 + \ldots + X_n$$

$$S = \sum_{i=1}^{n} (X_i - \bar{X}_+)(X_i - \bar{X}_+)^\top,$$

where $\bar{X}_+ = X_+/n$. Define the matrix of empirical correlations $R = \{R_{ij}\}_{i,j=1,\ldots,k}$, where

$$R_{ij} = \frac{S_{ij}}{(S_{ii}S_{jj})^{\frac{1}{2}}}.$$

The distribution of R depends only on ρ and is parametrized by ρ. We will now show that R is G-sufficient for ρ.

First, we observe that (X_+, S) is B-sufficient for (μ, Σ). Consider the group G of transformations of $\mathbb{R}^{nk}$, defined by

$$X_{ij} \mapsto a_j + b_j X_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, k,$$

where $b_j \neq 0 \ \forall j$ and $a_j \in \mathbb{R} \ \forall j$. Here $X \in \mathbb{R}^{nk}$ is represented as $\{X_{ij}\}_{i=1,\ldots,n, j=1,\ldots,k}$. The model is obviously closed under this action and the parameters (μ, Σ) are transformed in the following way

$$\begin{aligned} \mu_j &\mapsto a_j + b_j \mu_j \\ \Sigma_{jj} &\mapsto b_j^2 \Sigma_{jj} \\ \rho &\mapsto \rho, \end{aligned}$$

that is, ρ is the invariant parameter. In the same way, the statistic (X_+, S) is transformed by

$$\begin{aligned} X_{+j} &\mapsto a_j + b_j X_{+j} \\ S_{jj} &\mapsto b_j^2 S_{jj} \\ R &\mapsto R, \end{aligned}$$

which shows that R is maximum invariant. By the arguments above, it follows that R is G-sufficient for ρ.

3.3.3 Cox's proportional risks model

We will consider in this subsection the classical Cox model of proportional risks, which is widely used in survival analysis. This example will be useful to illustrate the construction of the likelihood function in a more complex situation since this is a semi-parametric model, i.e., part of the model will be parametrized by an infinite-dimensional parameter.

We assume that we observe the duration until a terminal event in a group of n individuals. This event can be the death or any event such that when it occurs we end the observation in this individual, as for example, the recovery of an illness, the occurrence of a defect in a machine, etc. We will always use the expression "the individual died" to indicate that the phenomenon happened.

For each individual we are given a set of independent variables (covariates) represented by $X_i \in \mathbb{R}^k$ for the i-th individual. In the context of a medical treatment these variables can represent, for example, the treatment given to the individual, its age, sex, among others. We want to make a model that explains the chances that an individual survives, as a function of the covariates, similarly to what is made in generalized linear models.

The distribution of the waiting time until the death of an individual can be described by the instantaneous risk function, which represents the probability to die in the next instant

of an individual, given that she or he survived until that moment. That is, the risk function is given by

$$P(\text{die in } (t, t+\delta)|\text{alive at time } t) = h(t)\delta + o(\delta), \tag{3.21}$$

where $o(\delta)/\delta \longrightarrow 0$ as $\delta \to 0$. Another way to express the idea above is to say that

$$P(\text{dying after time } (s+t) \,|\, \text{alive until time } s) = \exp\{-\int_s^{s+t} h(u)du\}. \tag{3.22}$$

It will be useful to compare the model above with a non-homogeneous Poisson process.

Let us consider m individuals with risks $h_1(t), \ldots, h_m(t)$. Let T be a random variable which describes the time of the first death among the m individuals and, I the individual that died at instant T. Using the interpretation of the instantaneous risk given in (3.21), we see that

$$P(I = i|T = t) = \frac{h_i(t)}{\sum_{j=1}^m h_j(t)}, \tag{3.23}$$

and so similarly to what was obtained in (3.22),

$$P(T > t) = \exp\left\{-\int_0^t \left[\sum_{j=1}^m h_j(u)\right] du\right\}.$$

Thus, the density of T is

$$\left\{\sum_{j=1}^m h_j(t)\right\} \ \exp\left\{-\int_0^t \left[\sum_{j=1}^m h_j(u)\right] du\right\}. \tag{3.24}$$

For the n individuals with covariates $X_1, \ldots, X_n$ we consider a model where the instantaneous risk function for the i-th individual is

$$h_i(t) = \lambda(t)e^{\beta^\top X_i}, \tag{3.25}$$

where $\lambda(t)$ is a factor of intensity, common to all the individuals and β is a parameter. This is the classical Cox model of proportional risks . The specification of (3.25) is fundamental for the inference that will be developed.

Let $D_1, \ldots, D_n$ be the times in which the individuals die. Let $T_1 \leq T_2 \leq \ldots \leq T_n$ be the ordered times D_i, and let $R_1, \ldots, R_n$ the ranks of D_i, i.e., R_i is the individual that dies at instant T_i. We can write the likelihood function as a product of conditional densities in the following way:

$$\begin{aligned} L(\beta, \lambda(\cdot)) &= \prod_{i=1}^n p(r_i, t_i|(r_1, t_1), \ldots, (r_{i-1}, t_{i-1})) \\ &= \prod_{i=1}^n \{p(r_i|t_i, (r_1, t_1), \ldots, (r_{i-1}, t_{i-1}))p(t_i|(r_1, t_1), \ldots, (r_{i-1}, t_{i-1}))\}. \end{aligned}$$

Let $I_i = \{R_i, R_{i+1}, \ldots, R_n\} = \{1, \ldots, n\} \setminus \{R_1, \ldots, R_{i-1}\}$ be the individuals that are alive at instant T_{i-1} ($T_0 = 0$). Given the variables $(r_1, t_1), \ldots, (r_{i-1}, t_{i-1})$, we have a situation as in (3.23) and (3.24), where the m ($m = i$) are the individuals given by I_i, and t_{i-1} is the origin of the time axis. From here we obtain,

$$\begin{aligned} L(\beta, \lambda(\cdot)) &= \prod_{i=1}^{n} \frac{\lambda(t_i) e^{\beta^\top X_{r_i}}}{\sum_{j \in I_i} \lambda(t_i) e^{\beta^\top X_{r_j}}} \left\{ \sum_{j \in I_i} \lambda(t_i) e^{\beta^\top X_{r_j}} \right\} \\ &\quad \exp\left\{ - \int_{t_{i-1}}^{t_i} \left[\sum_{j \in I_i} \lambda(u) e^{\beta^\top X_{r_j}} \right] du \right\} \\ &= \exp\left(\sum_{i=1}^{n} \beta^\top X_i \right) \prod_{i=1}^{n} \lambda(t_i) \exp\left\{ - \left[\sum_{j \in I_i} e^{\beta^\top X_{r_j}} \right] \int_{t_{i-1}}^{t_i} \lambda(u) du \right\} \\ &= \frac{\exp(\sum_{i=1}^{n} \beta^\top X_i)}{\prod_{i=1}^{n} \{\sum_{j \in I_i} \exp(\beta^\top X_{r_j})\}} \prod_{i=1}^{n} \left\{ \sum_{j \in I_i} \exp(\beta^\top X_{r_j}) \right\} \lambda(t_i) \\ &\quad \exp\left\{ - \left[\sum_{j \in I_i} e^{\beta^\top X_{r_j}} \right] \int_{t_{i-1}}^{t_i} \lambda(u) du \right\} \\ &= p(r_1, \ldots, r_n; \beta) p(t_1, \ldots, t_n | r_1, \ldots, r_n; \beta, \lambda(\,\cdot\,)). \end{aligned}$$

We want to give arguments to conclude that $(R_1, \ldots, R_n)$ is G-sufficient for β. That is, we want to show that the family of conditional distributions of $(T_1, \ldots, T_n)$ given $(R_1, \ldots, R_n)$, for β fixed, is the family generated by a group G, with transitive action. Let G be the group of all increasing transformations of $(0, \infty)$ into $(0, \infty)$, i.e., all the increasing transformations of the time axis. For $\phi \in G$, T_i is transformed into $\phi(T_i)$ and R_i is invariant. The transformation ϕ turns the model above into another model of the same type, with proportional risks, without altering the value of β, but transforming the common risk $\lambda(\cdot)$ as

$$\lambda(\phi^{-1}(t)) \phi'(\phi^{-1}(t)).$$

Our model is $\lambda(\cdot) \in \Lambda$, where Λ is the set of all the functions on $(0, \infty)$ with positive values, then for λ fixed we have

$$\Lambda = \{\lambda(\phi^{-1}(\cdot)) \phi'(\phi^{-1}(\cdot)) : \phi \in G\}.$$

Hence, we have that the conditional model given by $(R_1, \ldots, R_n)$ is generated by a group with a transitive action. This shows that $(R_1, \ldots, R_n)$ is G-sufficient for β. The corresponding marginal likelihood is

$$p(r_1, \ldots, r_n; \beta) = \frac{\exp(\sum_{i=1}^{n} \beta^\top X_i)}{\prod_{i=1}^{n} \{\sum_{j \in I_i} e^{\beta^\top X_{r_j}}\}}.$$

The model of proportional risks can be extended for the very important practical case of censored data (see, for example, Kalbfleisch and Prentice (1980)).

3.4 M-nonformation

The concept of transformation model can be applied essentially to the case of continuous distributions, as we argued in the previous section. Let us recall that the main justification for introducing the concept of G-nonformation is that the model generated by a group with a transitive action, has "perfect fit", which does not depend so much on the concept of group action. In this section, we will see the notion of M-nonformation, which will capture the idea of "perfect fit" and which will be possible to be applied in discrete models, extending in this way the concept of G-nonformation and complementing the other two previous concepts.

The notion of M-nonformation is associated to some concepts related to the mode point of a distribution (from where the letter "M" in M-nonformation comes), which we will develop in the following.

Let us consider the parametric family $\mathcal{Q} = \{Q_\omega : \omega \in \Omega\}$ of probability measures on a measurable space $(\mathcal{X}, \mathcal{A})$, where $\mathcal{X} \subseteq I\!R^k$. We say that a point, $x \in \mathcal{X}$, is realizable if $x \in \cup_{\omega\in\Omega} S_\omega$, where S_ω is the support of the measure Q_ω. We assume that $\mathcal{Q}$ is dominated by a σ-finite measure ν and that

$$q(x;\omega) = \frac{dQ_\omega}{d\nu}(x), \quad \forall x \in \mathcal{X},$$

is the density of $Q_\omega \in \mathcal{Q}$ with respect to the measure ν. If there exists $x_0 \in \mathcal{X}$ such that

$$q(x_0;\omega) = \sup_{x\in\mathcal{X}} q(x;\omega)$$

we say that x_0 is the mode point of the density $q(\cdot,\omega)$. The family of densities $\mathcal{D} = \{q(\cdot\,;\,\omega) : \omega \in \Omega\}$ is called universal if for each possible $x_0 \in \mathcal{X}$ there exists $\omega \in \Omega$, such that x_0 is a mode point of the density $q(\cdot\,;\,\omega)$, i.e., for all possible $x_0 \in \mathcal{X}$, x_0 there exists $\omega \in \Omega$, such that

$$q(x;\omega) \leq q(x_0,\omega), \quad \forall x \in \mathcal{X}. \tag{3.26}$$

We will use the following extension of the concept of universal family. A realizable point $x_0 \in \mathcal{X}$ is called a mode point of the family of densities $\mathcal{D}$ when

$$\forall \epsilon > 0, \exists \omega \in \Omega \,:\, (1+\epsilon) q(x_0;\omega) \geq \sup_{x\in\mathcal{X}} q(x;\omega). \tag{3.27}$$

If each realizable point $x \in \mathcal{X}$ is a mode point of the family $\mathcal{D}$, then $\mathcal{D}$ will be called universal. If each realizable $x \in \mathcal{X}$ is a mode point of some member of the family $\mathcal{D}$, then $\mathcal{D}$ will be called strictly universal. Note that in the case of strict universality each realizable point of $\mathcal{X}$ must be a mode point of some density of $\mathcal{D}$, whereas, in the case of the universality (non-strict) each realizable point of $\mathcal{X}$ must be a mode point of the family $\mathcal{D}$. Therefore, strict universality is a particular case of universality. If $\sup_{x\in\mathcal{X}} q(x;\omega)$ is independent of ω we say that $\mathcal{D}$ has constant mode.

The main interpretation of universality is that for any observation $x_0 \in \mathcal{X}$, it is possible to obtain a perfect fit (or almost perfect) of the model $\mathcal{D}$. In the case of strict universality,

there exists $\omega \in \Omega$ such that x_0 is a mode point of $q(\cdot;\omega)$, so that we obtain a model with perfect fit. If we only have universality (non-strict) there exists a $\omega \in \Omega$ for which $q(x_0;\omega)$ is as close as we want to $\sup_{x\in\mathcal{X}} q(x;\omega)$, so that we obtain a fit as good as we desire.

It is important to note that universality is characteristic of families of densities and not of a family of distributions, in the sense that universality depends on the chosen measure with respect to which the densities are defined. In many cases, there exists a natural measure, that is almost always used, as for example, the counting measure in the discrete case or the Lebesgue measure in the continuous case. An application of the concept of universality in the discrete case is simpler because for a fixed measure ν, the density is unique, which does not happen in the continuous case, where a density can be modified in a set of measure zero. However, it is usually possible to choose, uniquely, a continuous version of the density.

The following lemma will be useful to establish a relation between M-nonformation and G-nonformation, as well as to use the notion of universality in certain continuous cases.

Lemma 3.23 *Let G be a group of transformations in $\mathcal{X}$, with a transitive action. We assume that, for some $\omega_0 \in \Omega$,*

$$\mathcal{D} = \{q(g^{-1}(\cdot);\omega_0) : g \in G\}.$$

Then $\mathcal{D}$ is strictly universal with constant mode.

The lemma above says that under general regularity conditions G-nonformation of a family implies universality with respect to the invariant measure. Therefore, in the case of continuous distributions, where we frequently find models closed under a group action, the question of the choice of a dominating measure for the definition of universality can be answered taking the invariant measure. As we will see, in this case G-nonformation will be a special case of M-nonformation.

Theorem 3.24 *Let $U = u(X)$ be a statistic such that the density of Q_ω has the following factorization:*

$$q(x;\omega) = h(u;\omega)g(x|u;\omega)$$

in terms of the marginal density of U and of the conditional density given U. We assume that x_0 is a mode point of $\mathcal{D}$ and let $u_0 = u(x_0)$. Then x_0 is also a mode point for the family of conditional densities

$$\{g(\cdot|u_0;\omega) : \omega \in \Omega\}.$$

Corollary 3.25 *Under the hypothesis of the theorem above, if $\mathcal{D}$ is universal then the family of conditional densities $\{g(\ \cdot\ |u;\omega) : \omega \in \Omega\}$ is also universal.*

Definition 3.26 *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model. Assume that there exists a σ-finite measure ν such that $\mathcal{P} \ll \nu$ and that the family $\mathcal{P}$ has the parametrization $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. Let us consider the statistics $U = u(X)$ and $V = v(X)$ and the parametric function $\psi = \psi(\theta)$. The sub-model*

$$\{P_{\theta U}(\ \cdot\ |V = v) : \theta \in \Theta\}$$

is called M-nonformative with respect to ψ if for each value v of V and $\psi_0 \in \psi(\Theta)$ the family $\{P_{\theta U}(\cdot|V = v) : \psi(\theta) = \psi_0\}$ is universal. In this case, if $U = u(X) = X$ then V is called M-sufficient, and if V is constant then U is called M-ancillary.

The justification for the definition of M-nonformation is similar to the one given for G-nonformation. Since this point is crucial, we will repeat the argument. First, if for each value ψ_0 of the parameter ψ the family $\{P_{\theta U}(\cdot|V = v) : \psi(\theta) = \psi_0\}$ is universal then, by the previous arguments, each of these families will provide a perfect fit for the observed value of the statistic U. In this way, the observed value of U will not allow us to prefer any given value of ψ over others, in terms of the goodness of fit, because all the families fit perfectly. We are here using the so-called Orwell's principle, since, if all fits are perfect, we do not admit that some are better than others (see Orwell, 1945).

Example 3.27 In Fisher's example (Example 3.9) we can see that for each fixed value of ψ, the family of distributions of X_+ is universal. Hence, X_+ is M-ancillary with respect to ψ.

As we have already mentioned, Lemma 3.23 allows us to conclude that, under regularity conditions, G-nonformation implies M-nonformation. In this way, several examples of Section 3.3, such as Examples 3.21 and 3.22, can be justified using M-nonformation instead of G-nonformation. For example, it is evident that the family of distributions of $\bar{X}_+ \sim N(\mu, \sigma^2/n)$ is universal for σ^2 fixed.

Even exploring the same example of the last paragraph, where we wanted to make inference on σ^2, it is evident that using only $\bar{X}_+$ it is not possible to obtain information about σ^2, since this would be equivalent to estimating the variance of the normal distribution with a single observation, and μ unknown. Note that this last point is fundamental, since if μ is known, or even if we have information that μ is in a given interval, it is possible to extract from a single observation some information about σ^2. For example, if we know that $\mu \in (0, 1)$, the observation $\bar{X}_+ = 1$ indicates that it is less likely that the variance is very large, whereas $\bar{X}_+ = 1000$ implies that very likely the variance is large. This idea has already been expressed very clearly by Barnard (1963) who, when introducing the concept now known as G-sufficiency, called it: "sufficiency for ψ in the absence of knowledge of ϕ" (adapted to the notation used in Section 3.3), that is, "sufficiency for ψ, in the absence of information about ϕ".

In Problems 3.9, 3.10 and 3.11 some examples of M-nonformation can be found. The discussion about M-nonformation will be continued in the following sections.

3.5 I-nonformation

M-nonformation is a property that depends on the choice of the dominating measure of the family of distributions we are working with. We will show that if we choose adequately the dominating measure, under very general regularity conditions, the family of densities obtained is universal. In this way, saying that there exists a measure, with respect to which

a given family of distributions is M-nonformative, is a relatively weak claim. But, it is not always evident how to show the universality of a given family of densities.

We introduce now the concept of I-nonformation, which will take into account many situations already studied with the notions of M- and G-nonformation. It also arises naturally in the discussion of conditional inference in exponential families. The definition of I-nonformation will not depend on the dominating measure of the family of distributions and it will be relatively easy to use. In order to define I-nonformation we will introduce a version of the concept of saturated model which will have a quite clear interpretation, in the continuous as well as in the discrete case.

3.5.1 Definitions

Let us consider the parametric family $\mathcal{Q} = \{Q_\omega; \omega \in \Omega\}$ of probability measures dominated by the measure ν. This family is represented by the densities $q(x;\omega) = \partial Q_\omega / \partial \nu\,(x)$, with $\omega \in \Omega$. Let $\mathcal{S}$ be the set of realizable points of $\mathcal{Q}$. We assume that the maximum likelihood estimator, $\hat{\omega}(x)$, is defined for all $x \in \mathcal{S}$.

Definition 3.28 *Under the conditions above, the family $\mathcal{Q}$ is called saturated by the maximum likelihood estimator if the function $\hat{\omega} : \mathcal{S} \to \Omega$ is one-to-one.*

In some situations the maximum likelihood estimator is not defined in the whole set $\mathcal{S}$. For example, in discrete exponential families, the maximum likelihood estimator is not defined in the extreme points of the convex support of the canonical statistic. In these cases, we should interpret the definition above, not considering the points where the maximum likelihood estimator is not defined, or using an extended domain of the parameter. Note that the discrete case also shows that $\hat{\omega}$ is in general not subjective.

Observe that the concepts of G- and M-nonformation have apparently no immediate interpretation in terms of the score function. Meanwhile, this does not happen with the concept of saturated model by the maximum likelihood estimator. To see this, consider the score function, $U(\omega; x) = \partial \log q(x;\omega)/\partial\omega$ of the parameter ω. Then the family $\mathcal{Q}$ is saturated by the maximum likelihood estimator if and only if the equation $U(\omega; x) = 0$ defines a one-to-one relation between x and ω.

Definition 3.29 *Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a parametric statistical model with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and $\psi = \psi(\theta)$ be the parameter of interest. Consider the statistics U and V. If for each value $\psi_0 \in \psi(\Theta)$, the family $\{P_{\theta U}(\,\cdot\,|V = v) : \psi(\theta) = \psi_0\}$ is saturated by the maximum likelihood estimator, then, we say that the family $\{P_{\theta U}(\,\cdot\,|V = v) : \theta \in \Theta\}$ is I-nonformative with respect to ψ.*

The justification for this definition is similar to the one given for G- and M-nonformation. That is, for each value of ψ, the conditional distribution of $U|V = v$ is saturated by maximum likelihood estimator and hence it gives a perfect fit for the observation of U. This argument,

together with "Orwell's principle" presented in Section 3.4, implies that U does not provide information about ψ.

Evidently, we have associated with the notion of I-nonformation the concepts of I-sufficiency and I-ancillarity, defined in the usual way.

In order to justify the new definition of nonformation, we will interrupt for the moment the discussion of I-nonformation and we will turn our attention to the concept of M-nonformation.

We assume that the maximum likelihood estimator $\hat{\omega}(x)$ exists and that $q(x;\hat{\omega}(x))$ is finite for all $x \in \mathcal{S}$. Note that $q(x;\hat{\omega}(x))$ is well defined even if $\hat{\omega}(x)$ is not unique and if $\hat{\omega}(x)$ does not exist in extreme points of $\mathcal{S}$, it is sometimes possible to define $q(x;\hat{\omega}(x))$ by continuity.

We define the normed likelihood by

$$\bar{L}(\omega) = \bar{q}(x;\omega) = \frac{q(x;\omega)}{q(x;\hat{\omega}(x))}$$

and the canonical measure corresponding to $\bar{L}$ by

$$\bar{\nu}(dx) = q(x;\hat{\omega}(x))\nu(dx).$$

Then we have that,

$$q(x;\omega)\nu(dx) = \bar{q}(x;\omega)\bar{\nu}(dx),$$

so that $\bar{L}(\omega)$ is the version of the likelihood that corresponds to the canonical measure $\bar{\nu}$. The normed likelihood will play an important role in Section 3.6.

Proposition 3.30 *If the maximum likelihood estimator $\hat{\omega}(x)$ exists and $q(x;\hat{\omega}(x))$ is finite for each $x \in \mathcal{S}$, and $\hat{\omega}(\mathcal{S}) = \Omega$, then a family of densities $\bar{\mathcal{D}} = \{\bar{q}(\cdot;\omega) : \omega \in \Omega\}$ is strictly universal with respect to the canonical measure $\bar{\nu}$, with constant mode. If, furthermore, the maximum likelihood estimator $\hat{\omega}(x)$ is unique for each $x \in \mathcal{S}$, then the mode point $m(\omega)$ of $\bar{q}(\cdot;\omega)$ is unique for each $\omega \in \Omega$ if and only if the family of distributions $\mathcal{Q}$ is saturated by the maximum likelihood estimator.*

Proof: By the definition of $\hat{\omega}(x)$ we have

$$\bar{q}(x;\omega) \leq \bar{q}(x;\hat{\omega}(x)) = 1, \quad \forall x \in \mathcal{S}, \quad \omega \in \Omega,$$

and hence for a given $x_0 \in \mathcal{S}$

$$\bar{q}(x;\hat{\omega}(x_0)) \leq 1 = \bar{q}(x_0;\hat{\omega}(x_0)).$$

Thus, x_0 is a mode point for $\bar{q}(\cdot,\hat{\omega}(x_0))$, and as $\hat{\omega}(\mathcal{S}) = \Omega$, we conclude that $\bar{\mathcal{D}}$ is strictly universal.

We will show the direct implication of the second part of the proposition, leaving the converse implication as an exercise for the reader.

Assume that for each $\omega \in \Omega$ there exists a unique mode point $m(\omega)$ of $\bar{q}(\cdot;\omega)$. Let $x_1, x_2 \in \mathcal{S}$. Assume that $\hat{\omega}(x_1) = \hat{\omega}(x_2) = \omega_0 \in \Omega$. Then $\bar{q}(x_1;\omega_0) = \bar{q}(x_2;\omega_0) = 1$ and then x_1 and x_2 are modes of $\bar{q}(\cdot;\omega_0)$. Hence, by the hypothesis of uniqueness of the mode, $x_1 = x_2$. We conclude that the function $\hat{\omega}$ is one-to-one and then the family $\mathcal{Q}$ is saturated by the maximum likelihood estimator. □

According to the proposition above, it is almost always possible to show M-nonformation. It is enough to consider the family of densities with respect to the canonical measure. Meanwhile, a careful examination can show that many times the canonical measure depends on the parameter ψ. Let us see this in more detail.

Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of distributions, $\psi = \psi(\theta)$ the parameter of interest, U and V statistics. Consider the family $\mathcal{P}_{\psi_0,v} = \{P_{\theta U}(\,\cdot\,|V = v) : \psi(\theta) = \psi_0\}$. Let $\bar{\nu}_{\psi_0,v}$ be the canonical measure of the family $\mathcal{P}_{\psi_0,v}$. Define

$$\bar{\mathcal{D}}_{\psi_0,v} = \{\bar{p}_{\psi_0}(u|v;\theta) : \psi(\theta) = \psi_0\},$$

where $\bar{p}_{\psi_0}(\cdot|v;\theta)$ is the conditional density of $U|V = v$ with respect to $\bar{\nu}_{\psi_0,v}$. Proposition 3.30 shows that the family $\bar{\mathcal{D}}_{\psi_0,v}$ is universal with respect to the canonical measure $\bar{\nu}_{\psi_0,v}$ and hence is M-nonformative with respect to ψ. It is obvious that, in general, the canonical measure $\bar{\nu}_{\psi_0,v}$ depends on ψ_0. The fact that the dominating measure depends on the parameter is not foreseen in the definition of universality and nor in the one of M-nonformation, but, as we will see in Example 3.31, this will cause an undesirable behaviour of M-nonformation.

Note that the canonical measure is, in general, not proportional to the commonly used standardized measure, as for example, Lebesgue or counting measure. On the other hand, the canonical measure can have a clear statistical interpretation in certain contexts. For example, in the case of a model generated by a group action, the canonical measure is, under general assumptions, invariant by the group action.

The following example shows a case where M-nonformation, with respect to the canonical measure, has undesirable properties. We will see that the concept of I-nonformation solves this problem.

Example 3.31 Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, with $X_1 \sim N(\mu, \sigma^2)$ and $(\mu, \sigma^2) \in I\!R \times I\!R_+$. We know that $(\bar{X}_+, SSD)$ is B-sufficient and that in the marginal distribution of $(\bar{X}_+, SSD)$ the statistic SSD is G-sufficient for σ^2 (see Example 3.20). We have that $\bar{X}_+ \sim N(\mu, \sigma^2/n)$ and that, if σ^2 is fixed, the model is universal relative to Lebesgue measure, so $\bar{X}_+$ is M-ancillary for σ^2. Since, in this model, the maximum likelihood estimator of μ is $\bar{X}_+$, then we have a saturated model by the maximum likelihood estimator, and hence $\bar{X}_+$ is I-ancillary for σ^2. Therefore, $\bar{X}_+$ is ancillary for σ^2 and then we should use the conditional distribution given $\bar{X}_+$. Since $\bar{X}_+$ and SSD are independent, we should make inference on σ^2 using the marginal distribution of SSD.

Assume now that we know that $\mu > 0$. In this case, as we have already mentioned in Section 3.3, the statistic $\bar{X}_+$ gives us some information about σ^2. Note that the maximum likelihood estimator of μ now is given by

$$\hat{\mu}_0 = \begin{cases} \bar{X}_+, & \text{if } \bar{X}_+ > 0 \\ 0, & \text{otherwise .} \end{cases}$$

Therefore, the model $N(\mu, \sigma^2/n)$, with σ^2 fixed, is not saturated by the maximum likelihood estimator and nor is it universal with respect to Lebesgue measure. Hence, $\bar{X}_+$ is not any more G-, I-ancillary and M-ancillary with respect to Lebesgue measure, as it should be. That is, $\bar{X}_+$ is ancillary in the absence of information about the nuisance parameter μ and it is not any more ancillary if we provide information about μ.

Using Proposition 3.30 we easily see that, even in the presence of the information "$\mu > 0$", the statistic $\bar{X}_+$ is M-ancillary for σ^2 with respect to the canonical measure. In this way, the notion of M-nonformation relative to the canonical measure does not behave as it would be desirable.

The notion of I-nonformation manages to deal with many cases where usually the concept of M- or of G-nonformation is used. We will show in Subsection 3.5.2 that I-nonformation is useful in conditional inference in exponential families. Meanwhile, it is important to note that, in this way, G- and M-nonformation, and I-nonformation are not implied by S-nonformation. To see this, consider the L-independent pair of parameters (ψ, ϕ) with domain $\Psi \times \Phi$ and suppose that this parameter corresponds to a cut. If we reduce the domain of the parameter to $\Psi \times \Phi_0$, where $\Phi_0 \subseteq \Phi$ then we still have a corresponding cut. But, this reduction in general does not allow the application of I-nonformation, according to the discussion in Example 3.31.

3.5.2 Conditional inference in exponential families

Inference in exponential families will give us an important class of examples where to apply the concept of I-nonformation.

Let us consider a regular exponential family with minimal representation

$$\frac{dP_\theta}{d\nu}(x) = a(\theta)b(x)e^{\theta \cdot x},$$

with $\theta \in \Theta$ and $x \in \mathbb{R}^k$. We call families of this type natural exponential families. We define $\tau(\theta) = -a'(\theta)/a(\theta)$. We know that τ is a bijection between Θ and int$C$, the interior of the convex support of X. Therefore, the maximum likelihood estimator is given by

$$\hat{\theta}(x) = \tau^{-1}(x),$$

which is a one-to-one function on int$C$ and hence the family $\{P_\theta : \theta \in \Theta\}$ is saturated by the maximum likelihood estimator.

We will work now with the regular exponential family with minimal representation given by

$$f(x; \psi, \phi) = a(\psi, \phi)b(x)e^{\psi u(x) + \phi v(x)}, \quad x \in \mathcal{X}.$$

We assume that we want to make inference on the parameter ψ. Note that the minimal canonical statistic $(U, V) = (u(X), v(X))$ is B-sufficient for (ψ, ϕ). For ψ fixed, V is B-sufficient for ϕ and hence the conditional distribution $U|V = v$ depends on (ψ, ϕ) only

through ψ. Now, is V ancillary for ψ in some sense? Now, we know that for each fixed ψ, the marginal distribution of V is a natural exponential family and hence it is saturated by the maximum likelihood estimator. Thus, the statistic V is I-ancillary for ψ. Therefore, we should make inference on ψ using the conditional distribution of U given V.

Example 3.32 Let us recall Fisher's example of Section 3.1. We have that $X_i \sim Bi(n_i, p_i)$, $i = 1, 2$, where X_1 and X_2 are independent random variables. The parameter of interest is $\psi = \log[p_1(1-p_1)/\{p_2(1-p_2)\}]$. Writing the likelihood in terms of the canonical parameters $\theta_i = \log\{p_i/(1-p_i)\}$, $i = 1, 2$, we obtain

$$L(\theta_1, \theta_2) = \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) \left(\begin{array}{c} n_2 \\ x_2 \end{array} \right) e^{\theta_1 x_1 + \theta_2 x_2} (1 + e^{\theta_1})^{-n_1} (1 + e^{\theta_2})^{-n_2}.$$

Working in terms of x_1 and $x_+ = x_1 + x_2$ we have

$$L(\theta_1, \theta_2) = \left(\begin{array}{c} n_1 \\ x_1 \end{array} \right) \left(\begin{array}{c} n_2 \\ x_+ - x_1 \end{array} \right) e^{\psi x_1 + \theta_2 x_+} (1 + e^{\psi + \theta_2})^{-n_1} (1 + e^{\theta_2})^{-n_2}.$$

Hence, X_+ is I-ancillary with respect to ψ, and inference on ψ should be made in the conditional distribution of X_1 given X_+, which justifies Fisher's exact test.

A second example follows which involves the binomial distribution.

Example 3.33 Let us consider the independent binomial random variables,

$$X_i \sim Bi(n, p_i), \quad i = -k, -k+1, \ldots, k,$$

where $\log p_i/(1-p_i) = \alpha + i\beta$.

This model is known as the logistic model of dose-response with doses allocated symmetrically around zero. The likelihood is given by

$$\begin{aligned} L(\alpha, \beta) &= \prod_{i=-k}^{k} \left(\begin{array}{c} n \\ x_i \end{array} \right) (1 + e^{\alpha + \beta i})^{-n} e^{x_i(\alpha + \beta i)} \qquad (3.28) \\ &= e^{\alpha x_+ + \beta \tilde{x}_+} \prod_{i=-k}^{k} \left(\begin{array}{c} n \\ x_i \end{array} \right) (1 + e^{\alpha + \beta i})^{-n}, \end{aligned}$$

where $X_+ = X_{-k} + \cdots + X_k$ and $\tilde{X}_+ = -kX_{-k} + \cdots + kX_k$.

The representation (3.28) shows that we are dealing with an exponential family with minimal canonical statistic $(X_+, \tilde{X}_+)$. The distribution of $(X_+, \tilde{X}_+)$ is

$$p(x_+, \tilde{x}_+; \alpha, \beta) = a(\alpha, \beta) b(x_+, \tilde{x}_+) \exp\left\{\alpha x_+ + \beta \tilde{x}_+\right\},$$

where

$$a(\alpha, \beta) = 2^{n(2k+1)} \prod_{i=-k}^{k} (1 + e^{\alpha + \beta i})^{-n}$$

and

$$b(x_+, \tilde{x}_+) = 2^{-n(2k+1)} \Sigma^* \prod_{i=-k}^{k} \left(\begin{array}{c} n \\ a_i \end{array} \right),$$

with Σ^* being the sum under the restrictions $\Sigma a_i = x_+$ and $\Sigma i a_i = \tilde{x}_+$.

We easily conclude that X_+ is I-ancillary for β and that, hence, inference on β should be made in the conditional distribution of $\tilde{X}_+$ given X_+.

It is possible to show that X_+ is M-ancillary for β with respect to the counting measure. This is due to the fact that the distribution of the sum of independent binomial variables is unimodal (see Keilson and Gerber, 1971). Note that if $k \neq 1$, the statistic $\tilde{X}_+$ is not M-ancillary for α, but it is I-ancillary for α.

We will consider a situation, such as in Section 1.5, where we have an exponential family of order k with canonical parameter θ and minimal canonical statistic $t(X)$. Let (ψ, σ) be a mixed parametrization with respect to $V = At(X)$, where A is a $m \times k$ matrix. We recall that $\psi(\theta) = A\tau(\theta)$ and $\sigma(\theta) = B\theta$, where $AB^\top = 0$. Interchanging the roles of the parameters, we will consider inference on $\sigma(\theta)$, where ψ is now the nuisance parameter. Using Theorem 1.21, we conclude that for σ fixed, the distribution of V is a natural exponential family and hence we have a model saturated by the maximum likelihood estimator. We conclude that V is I-ancillary for σ and in this way, we should make inference on σ using the conditional distribution of X given $V = v$, which depends on θ only through $\sigma(\theta)$. We will see several applications of this result.

The theory here developed will not be applicable to the following example, but nevertheless we will give a satisfactory solution to the presented problem.

Example 3.34 It is curious that one of the most popular statistical methods, the t-test, cannot be easily explained by the methods used until now. Let us see this in more detail: recall that if $X_1, \ldots, X_n$ are independent and identically distributed random variables, with $X_1 \sim N(\mu, \sigma^2)$ then we have the likelihood

$$\begin{aligned} L(\mu, \sigma^2) &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 + \frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} x_i \right\}. \end{aligned} \tag{3.29}$$

From (3.29), we obtain that, for μ fixed, the statistic $\sum_{i=1}^{n} (x_i - \mu)^2$ is sufficient. Note that this statistic depends on μ, which makes our methods non-applicable. Nevertheless, we deal with this problem in the following way.

Using (3.29) we see that $(\Sigma X_i, \Sigma X_i^2)$ is a B-sufficient statistic and that, for $\mu = 0$, ΣX_i^2 is sufficient. Therefore, to test the hypothesis "$\mu = 0$" we use the conditional distribution $\Sigma X_i | \Sigma X_i^2$, which is equivalent to using the conditional distribution

$$T = \frac{\sqrt{n} \frac{1}{n} \Sigma X_i}{\sqrt{\frac{1}{n-1} (\Sigma X_i^2 - 1/n (\Sigma X_i)^2}} \, \Big| \Sigma X_i^2 \, . \tag{3.30}$$

For $\mu = 0$, the normal distribution $N(0, \sigma^2)$ is generated by a group of transformations of scale, and T in (3.30) is invariant by this group action. Thus T is ancillary and ΣX_i^2 is B-minimal sufficient and complete, which shows, by Basu's Theorem, that T and ΣX_i^2 are independent. If one wishes to test $\mu = \mu_0$, it is sufficient to work with $X_i - \mu_0$, obtaining in this way the usual t-test.

The example above illustrates the fact that the inference on one component of the canonical statistic is easy, inference on the mean of a component of the canonical statistic can be difficult and it might even happen that we do not find any solution based on the methods studied here.

3.5.3 The relation between S- and I-nonformation

As we have already mentioned S-nonformation, in general, is not compatible with I-nonformation, in the sense that S-nonformation does not imply I-nonformation and vice-versa. A more refined analysis of this point reveals some fundamental aspects of the theory of inferential separation which we will present next. Note that it is enough to study the case of ancillarity.

Let us fix the notation to be used. We will consider a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where $\mathcal{P}$ is a parametric family of probability measures, with $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. We assume that the parameter θ is decomposed into $\theta = (\psi, \phi)$ and that the component ψ is the parameter of interest. We will work with the statistic $U = u(X)$ which will be ancillary for ψ.

Let us consider first the case of S-ancillarity, where the likelihood can be factorized in the following way

$$L(\theta) = g(x|u; \psi)h(u; \phi), \tag{3.31}$$

where $(\psi, \phi) \in \Theta = \Psi \times \Phi$. Note that we are assuming that ψ and ϕ are variationally independent (*i.e.*, the parameter (ψ, ϕ) varies in the Cartesian product $\Psi \times \Phi$, where $\psi \in \Psi$ and $\phi \in \Phi$). The crucial point to characterize S-ancillarity, is that the knowledge of the marginal distribution of U, or equivalently of ϕ, does not provide information about ψ. This is due to the fact that the marginal distribution of U does not depend on ψ and even knowing the exact value of ϕ, say ϕ_0, we can only say that $(\psi, \phi) \in \Psi \times \{\phi_0\}$, which does not reduce the uncertainty about ψ.

The cases of G-, M- and I-ancillarity are different situations, where we have the following decomposition of the likelihood

$$L(\theta) = g(x|u; \psi)h(u; \psi, \phi). \tag{3.32}$$

It is important to observe that here the marginal distribution of U also depends on ψ. Therefore, knowing this marginal distribution we also know the value of the parameter ψ, which is essentially different from the case of S-ancillarity.

Note that in the cases of G-, M- and I-ancillarity, if we had a sample of values of U, we could obtain some information about ψ using the marginal distribution of U. That is, we

can, for example, estimate ψ using a sample of values of U and the marginal distribution of U, even though this procedure is highly inefficient. This, apparently, contrasts the claim that U does not contain information about ψ, but a careful examination of the arguments here presented eliminates the paradox.

We recall that the common argument for the cases of M-, G- and I-nonformation, to justify that the value of U, together with its marginal distribution, does not contain information about ψ, is based on the idea of perfect fits. For example, in the case of M-ancillarity, the family of marginal distributions of U, with ψ fixed, always gives the distribution whose mode point is the observed value of U, *i.e.*, with perfect fit of the given value of U. It is important to emphasize that this argument says that *a single* observation of U, together with the marginal distribution de U, does not provide any information about ψ. Nothing is said with respect to the information contained in a samples of values of U.

Another point that distinguishes the concept of S-ancillarity from the rest, is that in the first case the parameters ψ and ϕ are variationally independent, and in the other cases this does not necessarily happen. In this way, in the cases of G-, M- or I-ancillarity, we should say that U is ancillary for ψ *in the absence of information about* ϕ. In the case of S-ancillarity, it is indifferent to have or not to have information about ϕ.

Example 3.35 Let us consider the regular exponential family with minimal representation given by

$$a(\psi,\phi)b(x)e^{u(x)\psi+v(x)\phi}, \quad x\in\mathcal{X} \tag{3.33}$$

such as in Subsection 3.5.2. We assume that we want to make inference about ψ and that ϕ is the nuisance parameter. As we have already said, if we had a single observation, the statistic V would be I-ancillary for ψ.

We assume now that we have the sample $X_1,\ldots,X_n$, of this family. The joint distribution of this sample is also an exponential family with representation

$$a(\psi,\phi)^n\prod_{i=1}^n b(x_i)e^{\psi\sum_{i=1}^n u(x_i)+\phi\sum_{i=1}^n v(x_i)}. \tag{3.34}$$

We define the statistics V_n and U_n by $v_n(x_1,\ldots,x_n)=\sum_{i=1}^n v(x_i)$ and $u_n(x_1,\ldots,x_n)=\sum_{i=1}^n u(x_i)$. We know that (U_n,V_n) is B-sufficient. The same arguments as before show that V_n is I-ancillary for ψ. This allows us to claim that inference on ψ should be made using the conditional distribution of U_n given V_n.

Note that we show that the statistic V_n is I-ancillary for ψ, but not that the statistic $v_n'(X_1,\ldots,X_n)=(v(X_1),\ldots,v(X_n))$ is I-ancillary for ψ. This is because the family of marginal distributions of V_n' is not saturated by the maximum likelihood estimator, since any $\tilde{x}=(x_1,\ldots,x_n)\in\mathcal{X}^n$ such that $\sum_{i=1}^n v(x_i)=x_0\in\mathcal{X}$ gives the same value of the maximum likelihood estimator of ϕ (for ψ fixed).

We conclude that V_n does not contain information about ψ, since it is I-ancillary. On the other hand, V_n' can give some information about ψ, which is in no way a contradiction since V_n' is not I-ancillary.

Summarizing the discussion above, we have that S-nonformation is essentially different from the other nonformation concepts presented until now.

3.6 L-nonformation

Until now we have studied five different nonformation concepts. In this way, it is natural to ask ourselves if there exists some concept of nonformation that combines all the previous concepts, obtaining in this way, a general definition. We will see that the concept of L-nonformation will provide a partial answer to this question, in the sense that it will include some of the already given definitions (not all though). On the other hand, the following discussion will lead us to analyse more carefully the existing relation between the concepts of sufficiency and ancillarity.

L-sufficiency, which is a particular case of L-nonformation, will turn out to be the most important definition of sufficiency. Until now, there does not exist any similar concept to L-ancillarity. L-sufficiency, as it will be seen, will have the advantage to be a relatively simple definition, and it will be given directly in terms of the likelihood.

In Subsection 3.6.1 we will define the concept of L-sufficiency.

3.6.1 L-sufficiency

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a parametric family of probability measures dominated by the σ-finite measure ν.

Let U be a statistic and $\psi = \psi(\theta)$ be a parametric function, that will be treated as the parameter of interest. The *profile likelihood* for ψ is defined by

$$\tilde{L}(\psi) = \sup_{\theta|\psi} L(\theta),$$

where $L(\theta) = f(x;\theta)$ is the likelihood with respect to the σ-finite measure ν and the symbol $\sup_{\theta|\psi_0}$ denotes the supremum over the set $\theta \in \Theta$, such that $\psi(\theta) = \psi_0$. Let $\hat{\theta}$ be the maximum likelihood estimator of θ without restrictions and let $\hat{\theta}(\psi)$ be the maximum likelihood estimator under the restriction $\psi(\theta) = \psi$, so that, $\tilde{L}(\psi) = L(\hat{\theta}(\psi))$.

Definition 3.36 *A statistic U is called L-sufficient for ψ if:*

(i) The marginal distribution of U depends on θ only through $\psi(\theta)$, and ψ parametrizes the family marginal distributions of U;

(ii) The σ-algebra generated by the normed profile likelihood $\tilde{L}(\psi)/L(\hat{\theta})$ is contained in the σ-algebra generated by U.

If these two σ-algebras are identical, U will be called minimal L-sufficient.

In the original definition, due to Rémon (1984), item (i) was not included. Its inclusion is due to Barndorff-Nielsen (1988). Note that this condition is required in all our previous definitions of nonformation.

Example 3.37 Let $X_1, \ldots, X_n$ be independent random variables with distribution $N(\mu, \sigma^2)$. Assume that we want to make inference on σ^2. The likelihood is given by

$$L(\mu, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

Hence, the profile likelihood for σ^2 is

$$\tilde{L}(\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}SSD}$$

using the previous notation. Also,

$$L(\hat{\mu}, \hat{\sigma}^2) = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}$$

where $\hat{\sigma}^2 = \frac{1}{n}SSD$. The normed profile likelihood is then

$$(\hat{\sigma}^2/\sigma^2)^{\frac{n}{2}} e^{\frac{n}{2}-\frac{1}{2\sigma^2}SSD}.$$

Therefore, SSD or, equivalently, $\hat{\sigma}^2$, is L-sufficient for σ^2. This leads us to the same conclusion obtained with G-sufficiency, that is, the estimator based on the marginal distribution of SSD is $\frac{1}{n-1}SSD$.

The definition of L-sufficiency can be interpreted as an extension of Fisher and Neyman's Factorization Theorem for B-sufficiency. We define $a(x) = L(\hat{\theta})$, so that we can write condition (ii) as

$$\tilde{L}(\psi) = a(x)b(u; \psi), \tag{3.35}$$

where, $b(u; \psi) = \tilde{L}(\psi)/L(\hat{\theta})$ is a function that depends on x only through u, according to point (ii) of the definition. On the other hand if there exist functions $a(x)$ and $b(u; \psi)$ such that (3.35) holds, then $\hat{\psi}$ is a function of x only through u, and

$$\frac{\tilde{L}(\psi)}{L(\hat{\theta})} = \frac{b(u; \psi)}{b(u; \hat{\psi})},$$

which shows that condition (ii) is satisfied. Hence, the factorization (3.35) characterizes L-sufficiency, given that the distribution of U depends on θ only through $\psi(\theta)$.

Note that $a(x) = L(\hat{\theta})$ is a density of the canonical measure with respect to ν, where nu is such that $a(x)\nu(dx) = \bar{\nu}(dx)$, and that the factor $b(u; \psi)$ em (3.35) is interpreted as the normed profile likelihood which will be denoted by $\bar{L}(\psi)$.

We will consider the case where $\psi = \theta$, so that item (i) of Definition 3.36 is always satisfied. In this case, the factorization (3.35) is equivalent to the factorization criterion of

Fisher-Neyman, according to which U is B-sufficient for θ if and only if $L(\theta)$ has a factorization of the form $L(\theta) = a(x)b(u;\theta)$. Hence, B-sufficiency is a particular case of L-sufficiency. Turning back to the case of a general $\psi(\theta)$, the factorization (3.35) shows that L-sufficiency is equivalent to the application of Fisher-Neyman's factorization criterion to the profile likelihood $\tilde{L}(\psi)$.

The concept of L-sufficiency has a simple interpretation in terms of the score function, as we will next show. To simplify the arguments, we will suppose that $\theta = (\psi, \phi)$, where ϕ is a nuisance parameter, and we will consider the following decomposition of the score function:

$$U(\psi,\phi) = \begin{pmatrix} U_1(\psi,\phi) \\ U_2(\psi,\phi) \end{pmatrix} = \frac{\partial \log L}{\partial(\psi,\phi)}(\psi,\phi).$$

By condition (i) of Definition 3.36, we have the factorization

$$L(\psi,\phi) = h(u;\psi)g(x|u;\psi,\phi),$$

and then

$$U_1(\psi,\phi) = H(u;\psi) + G_1(x|u;\psi,\phi)$$
$$U_2(\psi,\phi) = G_2(x|u;\psi,\phi),$$

where

$$H(u;\psi) = \frac{\partial \log h(u;\psi)}{\partial\psi},$$
$$B(u;\psi) = \frac{\partial \log b(u;\psi)}{\partial\psi},$$
$$G_1(x|u;\psi,\phi) = \frac{\partial \log g(x|u;\psi,\phi)}{\partial\psi}$$

and

$$G_2(x|u;\psi,\phi) = \frac{\partial \log g(x|u;\psi,\phi)}{\partial\phi}.$$

Note that the estimator of ϕ for ψ fixed, denoted by $\hat{\phi}(\psi)$, is defined by

$$G_2(x|u;\psi,\hat{\phi}(\psi)) = 0. \tag{3.36}$$

With this notation, we have that

$$\frac{\partial \log \tilde{L}(\psi)}{\partial\psi} = H(u;\psi) + G_1(x|u;\psi,\hat{\phi}(\psi)),$$

where we use the chain rule to differentiate $\log \tilde{L}(\psi) = \log L(\psi, \hat{\phi}(\psi))$, and equation (3.36). Using (3.35), we have that condition (ii) of Definition 3.36 is equivalent to

$$H(u;\psi) + G_1(x|u;\psi,\hat{\phi}(\psi)) = B(u;\psi),$$

which is equivalent to supposing that $G_1(x|u;\psi,\hat{\phi}(\psi))$ is a function of x only through u. In this way we have an interpretation of L-sufficiency using the score function.

We will next study the relation between L-sufficiency and the other concepts of sufficiency.

Theorem 3.38 *The concepts of B-, S- and G-sufficiency are particular cases of L-sufficiency.*

Proof: Let U be S-sufficient for ψ, let $\theta = (\psi, \phi)$ be the corresponding pair of L-independent parameters and let

$$L(\psi, \phi) = L_1(\psi)L_2(\phi) = h(u; \psi)g(x|u; \phi),$$

be the corresponding factorization. Then $\tilde{L}(\psi) = L_1(\psi)L_2(\hat{\phi})$ and $L(\hat{\psi}, \hat{\phi}) = L_1(\hat{\psi})L_2(\hat{\phi})$, where $\hat{\psi}$ maximizes $L_1(\cdot)$ and $\hat{\phi}$ maximizes $L_2(\cdot)$. Hence,

$$\frac{\tilde{L}(\psi)}{L(\hat{\psi}, \hat{\phi})} = \frac{L_1(\psi)}{L_1(\hat{\psi})}.$$

Since $L_1(\psi) = h(u; \psi)$, then evidently $\hat{\psi}$ is a function of x only through u. Hence, U is L-sufficient for ψ. This proof is also valid in the case of B-sufficiency, which is a special case of S-sufficiency.

Now, let us consider the case where U is G-sufficient for ψ. Let $\theta = (\psi, \phi)$, following the notation of Section 3.3. For each $g \in G$ we have

$$\sup_{\phi} f(gx; \phi, \psi) = \sup_{\phi} f(gx; g\phi, \psi) = \sup_{\phi} f(x; \phi, \psi) = \tilde{L}(\psi).$$

Hence, $\tilde{L}(\psi)$ is invariant by the action of G. Since U is maximum invariant, $\tilde{L}(\psi)$ is a function of U, which shows that U is L-sufficient for ψ. □

We will give now a general justification for L-sufficiency, which shows that it is related with a certain type of M-sufficiency, according to Theorem 3.39 below. By condition (i) of Definition 3.36 we have the factorization

$$L(\theta) = h(u; \psi)g(x|u; \theta),$$

in terms of the marginal and conditional densities of U. By the factorization (3.35) we have then the relation

$$h(u; \psi) \sup_{\theta|\psi} g(x|u; \theta) = a(x)b(u; \psi),$$

that is,

$$\sup_{\theta|\psi} \frac{g(x|u; \theta)}{a(x)} = \frac{b(u; \psi)}{h(u; \psi)}. \tag{3.37}$$

Since $a(x)\nu(dx)$ represents a canonical measure, the function

$$\bar{g}(x|u; \theta) = \frac{g(x|u; \theta)}{a(x)}$$

represents the conditional density of X given U with respect to a measure derived from the canonical measure. Hence, (3.37) implies that the supremum of $\bar{g}(x|u; \theta)$ for ψ fixed depends on x only through u. By the same arguments used in the proof of Proposition 3.30, this implies that the family $\{\bar{g}(\cdot|u; \theta) : \psi(\theta) = \psi_0\}$ is strictly universal with constant mode for each ψ_0 fixed.

Theorem 3.39 *A statistic U is L-sufficient for ψ if and only if U is M-sufficient and there exists a measure such that the family of densities $\{g(x|u;\theta) : \psi(\theta) = \psi_0\}$ with respect to this measure is strictly universal with constant mode for each ψ_0 fixed.*

Proof: The considerations preceding the theorem show that L-sufficiency implies the existence of a measure such that the family of conditional densities is universal for each ψ fixed, and hence U is M-sufficient. To show the converse implication, let U be M-sufficient for ψ, and suppose that the family of conditional densities $g(x|u;\theta)$, besides being universal for $\phi(\theta)$ fixed, also has constant mode, so that

$$\sup_{\theta|\psi} g(x|u;\theta) = m(u;\psi),$$

where $m(u;\psi)$ is a function that depends on x only through u. Thus,

$$\tilde{L}(\psi) = h(u;\psi)m(u;\psi).$$

By the factorization criterion (3.35), this shows that U is L-sufficient for ψ. □

By Theorem 3.39, we have then, that L-sufficiency represents a particular type of M-sufficiency, and by Theorem 3.38 we have still that B-, S- and G-sufficiency are particular cases of this type of M-sufficiency. In this way, a general argument to justify B-, S-, G- and L-sufficiency is based on the concept of strict universality with constant mode.

It is easy to verify that a large part of the examples of L-sufficiency can also be justified using I-sufficiency. But, apparently, there does not exist any formal relation between L- and I-sufficiency. It is evident that these two concepts are not equivalent, and that I-sufficiency does not include, in general, S-sufficiency, as we showed in Section 3.4.

Example 3.40 Let $X_1, \ldots, X_n$ be independent with density

$$f(x;\mu,\lambda) = a(\lambda)b(x)e^{\lambda t(y;\mu)} \tag{3.38}$$

for x and μ in $\Omega \subseteq \mathbb{R}^k$. We assume that $t(x;\mu) \leq t(x;x) = 0$, $\forall x, \mu \in \Omega$. Then the likelihood is

$$L(\mu,\lambda) = a(\lambda)^n \prod_{i=1}^{n} b(x_i) \exp\left\{-\frac{\lambda}{2} D(x;\mu)\right\},$$

where $D(x;\mu) = -2\{t(x_1;\mu) + \ldots + t(x_n;\mu)\}$ is called the deviance. We will make inference on λ. The profile likelihood for λ is

$$\tilde{L}(\lambda) = a(\lambda)^n \prod_{i=1}^{n} b(x_i) \exp\left\{-\frac{\lambda}{2} D(x;\hat{\mu})\right\},$$

where $\hat{\mu}$ minimizes the deviance $D(x;\cdot)$. Hence, the statistic $U = D(X;\hat{\mu})$ satisfies condition (ii) of Definition 3.36, but to be L-sufficient, the distribution of U should depend on (μ,λ) only through λ. Several distributions of the form (3.38) satisfy this condition. An example

is the normal distribution $N(\mu, 1/\lambda)$, which gives the same inference on $\sigma^2 = 1/\lambda$ already derived from G- and M-sufficiency. Another example is the gamma distribution, written in the following way

$$f(x;\mu,\lambda) = \frac{\lambda^\lambda}{\Gamma(\lambda)} x^{-1} \exp\left\{-\lambda\{\log(\mu/x) + x/\mu - 1\}\right\}.$$

A third example is the inverse Gaussian distribution, with density

$$f(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda}{2}\frac{(x-\mu)^2}{x\mu^2}\right\}.$$

We mention also that any density of the form

$$f(x;,\mu,\lambda) = a(\lambda)e^{\lambda t(x-\mu)}, \quad x \in \Omega \tag{3.39}$$

is of the form (3.38). If $\Omega = I\!R$, any function $t(\cdot)$ such that $\exp t$ is integrable can be used in (3.39). The case $\Omega = [0, 2\pi)$ and $t(x) = \cos x$ will give the distribution of von Mises-Fisher. The case of L-nonformation and L-ancillarity is discussed in Jørgensen (1993).

3.7 Models with many nuisance parameters

In this last section we will show a technique to eliminate a large number of nuisance parameters, that is, we will deal with cases where the number of nuisance parameters increases with the sample size. We will then have an opportunity to apply some of the techniques developed in the previous sections in a non trivial way.

We will first consider the general problem of the elimination of nuisance parameters by means of conditioning on a B-sufficient statistic. Next, we will consider the case where the number of nuisance parameters increases with the sample size.

Let X be a random variable with distribution P_θ where $\theta = (\psi, \phi)$, ψ is the parameter of interest. We assume that there exists a statistic $T = t(X)$ such that T is B-sufficient for ϕ if ψ is fixed. This implies that the likelihood is of the form

$$L(\psi, \phi) = g(x|t;\psi)h(t;\psi,\phi). \tag{3.40}$$

If the statistic T is ancillary, in some sense, we can use the conditional distribution of X given $T = t$ to make inference on ψ.

Meanwhile, without imposing more conditions, the factor $h(t;\psi,\phi)$ is not, in general, nonformative with respect to ψ, and can, in fact, contain information about ψ. This happens usually in the case where T is not boundedly complete for ψ fixed, which frequently happens when $T = (U, V)$, where V is B-ancillary for ψ fixed. In this case, (3.40) changes to

$$L(\psi, \phi) = g(x|v,u;\psi)h_1(v;\psi)h_2(u|v;\psi,\phi), \tag{3.41}$$

where $h_1(v;\psi)$ in general will depend on ψ. If this is the case, this factor will obviously contain information about ψ.

If T is boundedly complete for ψ fixed, the situation depicted above given by (3.41) is impossible, and in this case the model $X|T=t$ is, in a sense that will be explained in the next theorem, the only possibility to obtain a model that involves only ψ. The theorem comes from the theory of hypothesis testing.

Theorem 3.41 *Let H_0 be a hypothesis on the model $\mathcal{P}$, and U be minimal sufficient under H_0. We assume that A is a critical region for H_0 which is similar,* i.e., *$P(A)=\alpha$ for any $P\in\mathcal{P}$. If U is boundedly complete under H_0, then*

$$P(A|U=u)=\alpha.$$

Proof: The condition $P(A)=\alpha$ can be written as

$$\mathrm{E}_P(1_A)=\alpha \qquad \forall P\in H_0$$

or

$$\mathrm{E}_P\{E(1_A-\alpha|U)\}=0, \quad \forall P\in H_0,$$

where the conditional expectation does not depend on P. Since U is boundedly complete and $\mathrm{E}_P(1_A-\alpha|U)$ is bounded, we conclude that

$$E(1_A-\alpha|U)=0 \quad [\mathcal{P}] \text{ or}$$

$$P(A|U=u)=\alpha.$$

If T, in the discussion above, is boundedly complete, this theorem leads us to work with the conditional distribution given T, to make inference on ψ. Nevertheless, T is not necessarily nonformative with respect to ψ. The discussion in Section 3.5.2 shows that this can be related with the domain of ϕ, which should be as large as possible.

We consider now the situation where the dimension of ϕ is large, more precisely, its size is the sample size. Let $X_1,\ldots,X_n$ be independent random variables with distribution

$$X_i\sim P_{\psi,\phi_i}, \quad i=1,\ldots,n,$$

where ψ is the parameter of interest and $\phi_i=(\phi_1,\ldots,\phi_n)$ is the nuisance parameter. We assume that $V_i=v(X_i)$ is a B-minimal sufficient statistic for ϕ_i, with ψ fixed, in such a way that the likelihood is

$$L(\psi,\phi_i)=\prod_{i=1}^{n} g(x_i|v_i;\psi)\prod_{i=1}^{n} h(v_i;\psi,\phi_i).$$

It is not necessary to require that V_i is boundedly complete, although the discussion above shows that this would be desirable.

This situation was considered by Andersen (1970, 1973), who considered estimation based in the conditional likelihood

$$L_c(\psi) = \prod_{i=1}^{n} g(x_i|v_i;\psi),$$

and showed that, under regularity conditions, the conditional estimator $\hat{\psi}_c$, that maximizes $L_c(\psi)$, is consistent and asymptotically normal. He also showed (Andersen, 1971) that the likelihood ratio test based on L_c converges to the χ^2 distribution.

The problem with the maximum likelihood estimator when there is an infinite number of nuisance parameters was first studied by Neyman and Scott (1948), and was considered in Example 3.2 of Section 3.1.

Example 3.42 This example, besides from illustrating the ideas of exponential families and inferential separation, provides a useful statistical model which is used in pedagogical and psychological experiments, where n persons are exposed to m tests (called items).

Let X_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$ be random variables with Bernoulli distribution

$$p_{ij} = P(X_{ij} = 1) = 1 - P(X_{ij} = 0).$$

The model used is given by

$$p_{ij} = \frac{e^{\alpha_i + \beta_j}}{1 + e^{\alpha_i + \beta_j}},$$

that is, the generalized linear model with logistic link and with two factors with no interaction. The parameter $(\alpha_1, \ldots, \alpha_n, \beta_1, \ldots, \beta_m)^\top$ varies freely in $\mathbb{R}^{n+m}$.

The density of the Bernoulli distribution is $p^x(1-p)^{1-x}$, and hence, the joint density of all the X_{ij}s is

$$\prod_{i,j}(1 - p_{ij}) \left(\frac{p_{ij}}{1 - p_{ij}}\right)^{x_{ij}} =$$

$$\left\{\prod_{i,j}[1 + e^{\alpha_i + \beta_j}]^{-1}\right\} \exp\left(\sum_{i=1}^{n} \alpha_i X_{i+} + \sum_{j=1}^{m} \beta_j X_{+j}\right),$$

where the symbol "+" as an index, indicates the sum. Using the identity $X_{+m} = X_{++} - X_{+1} - \ldots - X_{+m-1} = \sum_i X_{i+} - X_{+1} - \ldots - X_{+m-1}$, we can verify that the model can be parametrized by $\tilde{\alpha}_i = \alpha_i + \beta_m$, $i = 1, \ldots, n$ and $\tilde{\beta}_j = \beta_j - \beta_m$, $j = 1, \ldots, m-1$. The parameter $(\tilde{\alpha}_1, \ldots, \tilde{\alpha}_n, \tilde{\beta}_1, \ldots, \tilde{\beta}_{m-1})^\top$ varies freely on $\mathbb{R}^{n+m-1}$. This is a regular exponential family of order $n + m - 1$.

This model was introduced by G. Rasch under the name item analysis. This and other models are explained in the books by Rasch (1960) and Cox and Snell (1989). The model was created to be used in the analysis of intelligence tests used for admission to the army.

The data can be represented in the following way:

$$\begin{array}{cccccc} & & & question & & \\ & i \setminus j & 1 & \cdots & j & \cdots & m \\ & \vdots & \vdots & & \vdots & & \vdots \\ person & i & X_{i1} & \cdots & X_{ij} & \cdots & X_{im} \\ & \vdots & \vdots & & \vdots & & \vdots \\ & n & X_{n1} & \cdots & X_{nj} & \cdots & X_{nm} \end{array}$$

The i, j-th cell of the table contains 1 if person i answered question j correctly, and 0 if the answer was wrong.

We can assume that

$$X_{ij} \sim Bi(1, p_{ij}).$$

Let δ_i be a number that measures the ability of the i-th person to answer the type of questions used. Large values of δ_i mean that the person has a good ability to answer the questions. Let ϵ_j be a parameter that indicates the degree of difficulty of the question numbered j, where large ϵ_j means that the question is difficult. Then it is reasonable to infer that the probability of the i-th person to answer correctly to the j-th question, p_{ij}, depends on δ_i and ϵ_j, or that

$$p_{ij} = \pi(\delta_i, \epsilon_j).$$

We assume that the ability and the difficulty are measured in a scale from 0 to ∞, in such a way the we can compensate the fact that if the difficulty is doubled the ability is also doubled. We conclude that π depends on δ and ϵ only through δ/ϵ, i.e.,

$$\pi(\delta, \epsilon) = \pi\left(\frac{\delta}{\epsilon}\right).$$

It is reasonable to assume that

$$\pi(v) \to \begin{cases} 1 & \text{if } v \to \infty \\ 0 & \text{if } v \to 0. \end{cases}$$

A function that satisfies this criterion is

$$\pi(v) = \frac{v}{1+v}, \qquad v \in I\!R_+.$$

Rasch then chose the model given by

$$\begin{aligned} p_{ij} &= \pi(\frac{\delta_i}{\epsilon_j}) \\ &= \frac{\frac{\delta_i}{\epsilon_j}}{1 + \frac{\delta_i}{\epsilon_j}} \\ &= \frac{\exp\{\log \delta_i - \log \epsilon_j\}}{1 + \exp\{\log \delta_i - \log \epsilon_j\}} \quad \delta_i > 0\,, \quad \epsilon_j > 0. \end{aligned}$$

From the point of view of generalized linear models, it is important to verify the link, but the canonical link that Rasch used has the advantage of giving a regular exponential family, where it is possible to apply the results developed in this chapter.

Let us consider now the case $m = 2$. Define $\tilde{\beta} = \beta_1 - \beta_2$. Then

$$P(X_{i1} = 1) = \frac{e^{\tilde{\alpha}_i + \tilde{\beta}}}{1 + e^{\tilde{\alpha}_i + \tilde{\beta}}} \text{ and } P(X_{i2} = 1) = \frac{e^{\tilde{\alpha}_i}}{1 + e^{\tilde{\alpha}_i}}.$$

Obviously, the parameter $\tilde{\beta}$ characterizes the difference between the two questions. The likelihood equations are

$$X_{i+} = (1 + e^{-\tilde{\alpha}_1 - \tilde{\beta}})^{-1} + (1 + e^{-\tilde{\alpha}_i})^{-1} \qquad i = 1, \ldots, n$$

and

$$X_{+1} = \sum_{i=1}^{n} (1 + e^{-\tilde{\alpha}_i - \tilde{\beta}})^{-1}.$$

According to the possible values of X_{i+}, which are $0, 1$ and 2, we have

$$\tilde{\alpha} = \begin{cases} -\infty & \text{if } X_{i+} = 0 \\ -\frac{1}{2}\tilde{\beta} & \text{if } X_{i+} = 1 \\ \infty & \text{if } X_{i+} = 2. \end{cases}$$

In this way, the following equation is obtained

$$X_{+1} = N_0 0 + N_1 (1 + e^{-\frac{1}{2}\tilde{\beta}})^{-1} + N_2 1,$$

where N_j is the number of X_{i+} equal to j, $j = 0, 1, 2$. The last equation is hence

$$-\frac{1}{2}\tilde{\beta} = \log \frac{N_1 - X_{+1} + N_2}{X_{+1} - N_2}.$$

Since $X_{+1} - N_2$ is the number of pairs (X_{i1}, X_{i2}) with value $(1, 0)$, and $N_1 - X_{+1} + N_2$ is the number of pairs with value $(1, 0)$, by the law of large numbers we have that

$$\frac{X_{+1} - N_2}{N} \to \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\tilde{\alpha}_i}}{(1 + e^{-\tilde{\alpha}_i - \tilde{\beta}})(1 + e^{-\tilde{\alpha}_i})}$$

and

$$\frac{N_1 - X_{+1} + N_2}{X_{+1} - N_2} \to \frac{1}{n} \sum_{i=1}^{n} \frac{e^{-\tilde{\alpha}_i - \tilde{\beta}}}{(1 + e^{-\tilde{\alpha}_1 - \tilde{\beta}})(1 + e^{-\tilde{\alpha}_1})}.$$

Hence,

$$-\frac{1}{2}\tilde{\beta} \to \log(e^{-\tilde{\beta}}) = -\tilde{\beta}.$$

This shows that $\tilde{\beta}$ converges to the wrong value, as $\hat{\sigma}^2$ in Example 3.2 of Section 3.1.

If we consider the conditional distribution of X_{i1} given $X_{i+} = x_{i+}$, we have the conditional density

$$P(X_{i+}|X_{i+} = x_{i+}) = \begin{cases} 1_{\{x_{i1}=0\}} & \text{if } x_{i+} = 0 \\ \frac{e^{\tilde{\beta}x_{i1}}}{1+e^{\tilde{\beta}}} & \text{if } x_{i+} = 1 \\ 1_{\{x_{i1}=1\}} & \text{if } x_{i+} = 2. \end{cases}$$

In order to justify this model, we note that X_{i+} is M-ancillary for $\tilde{\beta}$. The likelihood function for these conditional densities is

$$L_c(\tilde{\beta}) = e^{\tilde{\beta}N_{10}}(1 + e^{\tilde{\beta}})^{-N_{01}-N_{10}},$$

where N_{ij} is the number of pairs (X_{i1}, X_{i2}) so (i, j). We obtain then the conditional estimator

$$\tilde{\beta}_c = \log \frac{N_{10}}{N_{01}} \to \log(\frac{1}{e^{-\tilde{\beta}}}) = \tilde{\beta},$$

using the previous argument. This estimator converges to the true value.

This example shows once again the importance of using the marginal or conditional distribution for making inference, if there are nuisance parameters. The general case of Example 3.42 can be solved with the theory presented by Andersen (1973).

3.8 Problems

S-nonformation

Problem 3.1 Let X and Y independent random variables with

$$P(X = 0) = q, \quad P(X = 1) = p, \quad p + q = 1$$

and

$$P(Y = -1) = a, \quad P(Y = 0) = q \quad P(Y = 1) = p - a,$$

where $(a, p)^\top$ has domain

$$\Theta = \{(a, p)^\top : 0 \leq a \leq p, \quad \frac{1}{2} \leq p \leq \frac{2}{3}\}.$$

The parameter of interest is p. Show that $X + Y^2$ is S-sufficient with respect to p.

Problem 3.2 Let f be the function defined on $\mathbb{R}^2$ by

$$f(u_1, u_2) = \frac{u_1^{\lambda_1-1}u_2^{-\lambda_2-\lambda_1-1}}{\Gamma(\lambda_1)\Gamma(\lambda_2)\beta_1^{\lambda_1}\beta_2^{\lambda_2}} e^{-\frac{1}{\beta_2 u_2}(\frac{\beta_2}{\beta_1}u_1+1)},$$

where $u_1 > 0$ and $u_2 > 0$, and $(\lambda_1, \lambda_2, \beta_1, \beta_2) \in \mathbb{R}^4_+$.

(i) Show that f is a density function and that the family of distributions with density f and $(\lambda_1, \lambda_2, \beta_1, \beta_2) \in \mathbb{R}^4_+$ is an exponential family. Give the order of this family.

(ii) Show that u_1/u_2 is S-sufficient with respect to (λ_1, β_1) and that u_2 is S-sufficient for (λ_2, β_2).

Problem 3.3 Let $X_1, \ldots, X_n$ be independent, and $X_i \sim N_p(\mu, \Sigma)$, $i = 1, \ldots, n$, where $\mu \in \mathbb{R}^p$ and Σ is $p \times p$ is positive definite.

(i) Show, using Basu's Theorem, that the maximum likelihood estimator $\bar{X}_+$ for μ and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_+)(X_i - \bar{X}_+)^\top$ for Σ are independent.

(ii) Show that $\bar{X}_+$ is not S-ancillary with respect to Σ.

(iii) Let $p = 2$ and $n = 1$. Show that X_{11} is a cut, and find the corresponding pair of L-independent parameters.

Problem 3.4 Let $(U, V)^\top$ be a random vector with density

$$f(u, v) = \frac{v^{\lambda-1} e^{-v/u^2}}{u^{2\lambda}\Gamma(\lambda)(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(u-\alpha)^2\right\}, \quad u \in \mathbb{R}, \quad v \in \mathbb{R}_+$$

where $\alpha \in \mathbb{R}$ and $\lambda, \sigma^2 \in \mathbb{R}_+$.

(i) Show that $U \sim N(\alpha, \sigma^2)$, and that the conditional distribution of V given $U = u$ is the distribution $Ga(u^2, \lambda)$.

(ii) Show that the distribution of (U, V) is an exponential family of order 3.

(iii) Let $(U_1, V_1)^\top, \ldots, (U_n, V_n)^\top$ be independent and identically distributed random vectors with density f. Show that $(U_1, \ldots, U_n)$ is S-sufficient for (α, σ^2) and S-ancillary for λ.

(iv) Show that

$$\left(\sum_{i=1}^n U_i, \sum_{i=1}^n U_i^2\right) \quad \textit{and} \ \sum_{i=1}^n \log \frac{V_i}{U_i^2}$$

are independent.

(v) Show that the maximum likelihood estimators $\hat{\alpha}, \hat{\sigma}^2$ and $\hat{\lambda}$ are independent.

Problem 3.5 Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric family of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$. Let ψ be a parametric function and $x \to (u(x), v(x))$, $x \in \mathcal{X}$, a measurable bijection. We assume that $U = u(X)$ and $V = v(X)$ are cuts and that the set $\Psi = \{\psi(\theta) : \theta \in \Theta\}$ parametrizes the marginal distributions of V as well as the conditional distributions given U. Hence, U is S-ancillary with respect to ψ and V is S-sufficient with respect to ψ. In this case, the principle of S-ancillarity states that inference on ψ should be made using the marginal distribution of V, whereas the principle of S-ancillarity says that inference on ψ should be made using the conditional distributions given U. Show that, if P is boundedly complete, then the two principles suggest the same class of distributions for inference on ψ.

G-nonformation

Problem 3.6 Let $Q = \{f_\psi(\cdot) : \psi \in \Psi\}$ be a family of densities in $\mathbb{R}$, such that no $f_\psi(\cdot)$ can be obtained from another member of Q by means of a transformation of location and scale. Let $P_{\psi,\mu,\sigma}$ be defined by the density

$$x \to \frac{1}{\sigma} f_\psi \left(\frac{x-\mu}{\sigma} \right).$$

(i) Show that for $(a, b) \in \mathbb{R} \times \mathbb{R}_+$,

$$a + bP_{\psi,\mu,\sigma} = P_{\psi,a+b\mu,b\sigma}$$

and that this corresponds to a group of transformations of the sample space.

(ii) Find a G-sufficient statistic for ψ, based on a random sample $X_1, \ldots, X_n$ from $P_{\psi,\mu,\sigma}$.

(iii) If ψ is known, find a B-ancillary statistic for (μ, σ).

Problem 3.7 Consider the generalized inverse Gaussian distribution in Example 3.7 of Section 3.1. Let $X_1, \ldots, X_n$ be random variables with distribution $N^-(\lambda, \chi, \psi)$, where λ is known. Study the inference on the parameters $\mu = \sqrt{\chi/\psi}$ and $\omega = \sqrt{\chi\psi}$ from a G-sufficiency point of view. Also, study the inference on μ when ω is known.

Problem 3.8 Define the function $a(\lambda)$ by

$$a(\lambda)^{-1} = \int_0^{2\pi} e^{\lambda \cos x} dx, \quad \lambda \geq 0.$$

Consider the distribution $P_{\mu,\lambda}$ defined by the density

$$f(x; \mu, \lambda) = a(\lambda) e^{\lambda \cos(x-\mu)} \quad x \in [0, 2\pi)$$

where $\lambda \geq 0$ and $\mu \in [0, 2\pi)$. This is called the von Mises' distribution. The distribution $P_{\mu,0}$ is the uniform distribution on $[0, 2\pi)$, whereas for $\lambda > 0$, $f(\cdot; \mu, \lambda)$ has mode point μ and is more and more concentrated around μ when λ increases.

(i) Let G be the group of rotations, that is, translations modulo 2π, defined by

$$x \to (g + x) \ mod\ 2\pi, \quad x \in [0, 2\pi).$$

Let the elements of G be denominated by $g \in [0, 2\pi)$. Show that

$$g(P_{\mu,\lambda}) = P_{g(\mu),\lambda}.$$

(ii) Let $X_1, \ldots, X_n$ be independent random variables with distribution $P_{\mu,\lambda}$. Show that the statistic

$$(X_{1+}, X_{2+})^\top = \left(\sum_{i=1}^n \cos X_i, \sum_{i=1}^n \sin X_i \right)^\top$$

is B-sufficient for (μ, λ) and that the family $\{P_{\mu,\lambda} : \mu \in [0, 2\pi), \lambda \geq 0\}$ is an exponential family of order 2. Show that $(X_{1+}, X_{2+})^\top$ is a minimal canonical statistic. The vector $(X_{1+}, X_{2+})^\top$ is called the resultant vector. Hint: Use trigonometric relations.

(iii) Let R be the length of the resultant vector defined by

$$R = (X_{1+}^2 + X_{2+}^2)^{1/2},$$

and $D = \frac{1}{R}(X_{1+}, X_{2+})^\top$, called the direction. Show that, with respect to the space of $(X_{1+}, X_{2+})^\top$, R is maximum invariant and D is equivariant.

(iv) Show that R is G-sufficient for λ.

(v) Show that, if μ is known, R is B-ancillary with respect to μ, hence, inference on μ should be made using the conditional distribution of D given $R = r$.

M-nonformation

Problem 3.9 Show that the family $\{N(\mu, \sigma^2) : \mu \in I\!R\}$ is strictly universal with constant mode point with respect to Lebesgue measure, for any given $\sigma^2 > 0$. Show the use of this result for inference on σ^2 in the distribution $N(\mu, \sigma^2)$.

Problem 3.10 Show that the family of exponential densities $\lambda e^{-\lambda x}$ $(x > 0)$ for $\lambda > 0$ is universal, but not strictly universal.

Problem 3.11 Show that the binomial family $\{Bi(n, p) : p \in (0, 1)\}$ is universal. Is it strictly universal? What would be the use of this result?

I-nonformation

Problem 3.12 Show that the negative binomial distribution $Nb(\lambda, p)$ is saturated by the maximum likelihood estimator for any fixed value of λ. Show that the binomial distribution $Bi(n, p)$ is saturated by the maximum likelihood estimator for n fixed.

Problem 3.13 Let $X_i \sim Po(\lambda_i)$, $i = 1, 2$ be independent, and define $\psi = \log(\lambda_1/\lambda_2)$. Show that $X_1 + X_2$ is I-ancillary for ψ.

Problem 3.14 Show that the family of arcsine distributions

$$f(x; \theta) = \frac{\sin(\pi\theta)}{\pi} x^{\theta-1}(1-x)^{-\theta}, \quad 0 < x < 1$$

for $\theta \in (0, 1)$, is saturated by the maximum likelihood estimator.

Problem 3.15 Let $X_i \sim N^-(\chi, \psi)$, $i = 1, \ldots, n$, be independent. Show that $X_- = X_1^{-1} + \cdots + X_n^{-1}$ is I-ancillary for ψ, and suggest a conditional test for the hypothesis $H_0 : \psi = \psi_0$. (Here we use the notation of Example 3.7).

L-nonformation

Problem 3.16 Let $X_i \sim N^-(\chi, \psi)$ be independent, $i = 1, \ldots, n$, and let $\omega = \sqrt{\chi\psi}$ be the parameter of interest. Show that the statistic $T = (X_- X_+)^{\frac{1}{2}}$ is L-sufficient for ω. (Here we use the notation of Example 3.7).

Problem 3.17 Let $X_1, \ldots, X_n$ be independent with negative binomial distribution, $X_i \sim Nb(\lambda, p)$. Show that X_+ is I-ancillary for the parameter λ.

Problem 3.18 Let $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ be independent and

$$\begin{aligned} X_i &\sim Bi\left(r, \phi_i\psi/(1+\phi_i\psi)\right) \\ Y_i &\sim Bi\left(s, \phi_i/(1+\phi_i)\right). \end{aligned}$$

Discuss inference on ψ.

Problem 3.19 Let $Y \sim N_k(X\beta, \sigma^2 I)$ be a linear model. Show that $SSD = ||Y - \hat{\mu}||^2$, where $\hat{\mu} = X\hat{\beta}$, is L-sufficient for σ^2, and use this to derive the usual estimator of σ^2, SSD/f, where f is the number of degrees of freedom.

Chapter 4

INFERENCE FUNCTIONS

4.1 Introduction

The traditional theory of point estimation deals with properties of estimators which are functions only of the observations, *i.e.* statistics. Constraints, such as invariance, unbiasedness or asymptotic normality, are usually imposed in the class of estimators to be considered. A criterion of ordering (or partial ordering) in the restricted class of estimators, such as the variance or the asymptotic variance of the estimator, is then given and used to define an optimal estimate. We refer to two classical examples of these paradigmatic developments: the traditional unbiased minimum variance estimation and the maximum likelihood theory, where efforts are put into finding estimators that are asymptotically normally distributed with minimum variance. In spite of the beauty of these theories, some criticisms can be made. For instance, the method of uniform unbiased estimation may produce estimators which are absurd (see Kendall and Stuart, 1979 p. 36). Moreover, the minimum variance unbiased method is not invariant under smooth reparametrizations, *i.e.* the estimator obtained depends on the arbitrarily chosen parametrization. On the other hand, the method of maximum likelihood estimation is invariant under smooth reparametrizations, but these estimators may be (asymptotically) inefficient or even inconsistent. In Chapter 3 we gave some alternatives to the pure maximum likelihood method that satisfactorily solve some of these situations. In particular, the so-called Neyman-Scott paradox has been addressed via the notion of nonformation. Here we give another alternative to the maximum likelihood theory by developing a theory of what we call inference functions or estimating equations.

In the approach of estimating equations we consider estimators which can be expressed as solutions of an equation such as

$$\Psi(x;\theta) \,=\, 0 \ . \tag{4.1}$$

Here Ψ is a function of the given data, say x, and the parameter, say θ, of a certain statistical model. We call Ψ an inference function, also known as an estimating function (the precise definition will be given later). The equation (4.1) is often called an estimating equation. Following the same procedure as in the classical theories, one introduces some constraints in

the class of inference functions to be considered, and a criterion for ordering the estimators obtained from the estimating equations in the restricted class. Equivalently one might introduce a criterion for ordering the inference functions, and choose the uniformly best (if it exists). It is clear that in most of the classical "well-behaved" cases, the maximum likelihood estimator is given by the solution of an estimating equation. Moreover, the criterion for ordering inference functions is closely related to the asymptotic variance of the associated estimators. In that way, the approach of estimating equations can be viewed as a generalization of the maximum likelihood theory. Due to the optimal behaviour of the maximum likelihood in "regular" cases, it is not surprising that the optimal inference function will give us exactly the maximum likelihood estimator. However, there are some situations in which the maximum likelihood theory fails and the estimating equation theory works well. For instance, we will see that the theory of estimating equations provides an alternative justification for conditional inference.

The earliest mention of the idea of estimating equations is probably due to Fisher (1935) (he used the term "equation of estimation"). A remarkable example of an early non-trivial use of inference functions can be found in Kimball (1946), where estimating equations were used to give confidence regions for the parameters of the family of Gumbel distributions (or extreme value distributions). There, the idea of "stable" estimating equations, *i.e.* inference functions whose expectations are independent of the parameter, was introduced, anticipating the theory of sufficiency and ancillarity for inference functions proposed by McLeish and Small (1987) and Small and McLeish (1988a,b).

The theory of optimality of inference functions appears in the pioneering paper of Godambe (1960). In the same year Durbin (1960) introduced the notion of unbiased linear inference function and proved some optimality theorems particularly suited to applications in time series analysis.

Since that time, the theory of inference functions has been developed a great deal, both by Godambe (c.f. Godambe, 1976, 1980, 1984; Godambe and Thompson, 1974, 1976), and by others in different contexts and with different names and approaches. We mention, for instance, the so-called theory of M-estimators developed in the seventies in order to obtain robust estimators, and the quasi-likelihood methods used in generalized linear models. As one can see, the theory of inference functions was not only inspired by an alternative optimality theory for point estimation. One could say that there is now a firm and well established theory of inference functions, with many branches, some of them based on very deep mathematical foundations. A notorious example is the sophisticated theory of weak convergence due to Hoffman-Jørgensen(1990), used for proving the consistency of estimators associated with estimating equations in a very general context (not even measurability is assumed, see also Van der Vaart and Wellner, 1996). In the present text, we concentrate on the use of inference functions for point estimation. We do not intend a complete coverage of the theory, but merely to give an overview of some important aspects.

The chapter is organized as follows. Section 4.2 presents some basic notions of the theory of inference functions, in particular the concept of unbiased inference function and the implications of the unbiasedness in terms of the consistency of the associated estimates, under

regularity conditions. The theory of optimality of inference functions is discussed in Section 4.3. This section begins by presenting the basic notions of the theory of inference functions for the case where the parameter is one-dimensional, including also the notion of regular asymptotic linear estimate and some connections with the classic theory of point estimation. The section closes with a generalization of the theory to the case of a multidimensional parameter. Section 4.4 studies the theory of inference functions in the presence of nuisance parameters. We first give a general formulation of the optimality theory for inference functions in this new context. Next, we specialize to models with a one-dimensional parameter of interest and arbitrary nuisance parameters, for which we can easily obtain some results that allow us to find explicitly optimal inference functions. At this point we give an alternative justification for conditional inference when one has a likelihood factorization of a particular form. A criterion for the existence of optimal inference functions is then given in a very general context (of semiparametric models) and it is proved that the only possibility for obtaining an optimal inference function is the so-called efficient score function (*i.e.* the orthogonal projection of the partial score function onto the orthogonal complement of the nuisance tangent space). This generalizes the results obtained in the previous section.

4.2 Preliminaries

Consider a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$. Here $\mathcal{P}$ is a family of probability measures defined on a common measurable space $(\mathcal{X}, \mathcal{A})$. We study first the case where $\mathcal{P}$ is a parametric family, say

$$\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq I\!R^k\},$$

for some fixed $k \in I\!N$. Later on we extend the theory to a more general context where one can also have a nuisance parameter, not necessarily finite dimensional.

Let us consider a function $\Psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$ such that for each $\theta \in \Theta$ the component function $\Psi(\,\cdot\,;\theta) : \mathcal{X} \longrightarrow I\!R^k$, obtained by fixing the value of the parameter, is measurable. Such a function is called an *inference function*. We stress that Ψ is an $I\!R^k$-valued function, where k is the dimension of the parameter space Θ.

Given an inference function, say Ψ, one can define for each possible value of the observation, say $x \in \mathcal{X}$, an estimate $\hat{\theta} = \hat{\theta}(x)$ as the solution to the following equation

$$\Psi(x;\theta) = 0$$

with respect to θ. Accordingly, one associates an estimator to the inference function Ψ. Obviously, the inference function Ψ must satisfy some regularity conditions in order for the estimate $\hat{\theta}$ to be well defined and well behaved, but for now we postpone this discussion.

It is interesting to observe that given an inference function, say $\Psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$, one can easily construct other inference functions with the same estimator associated. To see this, suppose that one has an inference function, say $\Phi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$, such that for all $x \in \mathcal{X}$ and all $\theta \in \Theta$

$$\Phi(\,x\,;\,\theta\,) \;=\; C(\theta)\Psi(\,x\,;\,\theta\,),$$

where $C(\theta)$ is a $k \times k$ matrix of full rank. The two inference functions Ψ and Φ are then said to be *equivalent*. We denote the equivalence of Ψ and Φ by "$\Psi \sim \Phi$". It is easy to see that " $\sim$" really defines an equivalence relation (see Problem 4.1) and that two equivalent inference functions have the same estimator associated. Hence, we must identify the inference functions contained in the same equivalence class. We will see that this is done implicitly in the theory developed in this chapter.

In the following, we work with a sample, say $X_1, \ldots, X_n$, of size n from an unknown (but fixed) distribution $P_\theta \in \mathcal{P}$ based on which one would like to estimate P_θ, or equivalently θ. Of course, the theory of inference functions can be applied in a more general context, but, due to the introductory nature of this chapter, it is better to concentrate on the treatment of this specific case.

Clearly, one can redefine the problem above in terms of the "product statistical model", say $(\mathcal{X}^n, \mathcal{A}^n, \mathcal{P}^n)$, and define an inference function in this context as a function from $\mathcal{X}^n \times \Theta$ into $\mathbb{R}^k$ in a completely general form. However, in the classical theory of inference functions one usually restricts the attention to the particular case of inference functions defined as follows. Let $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ be an inference function defined as before. Consider the inference function for the extended model $\Psi_n : \mathcal{X}^n \times \Theta \longrightarrow \mathbb{R}^k$ given by

$$\Psi_n(\,\mathbf{x}\,;\,\theta\,) \;=\; \sum_{i=1}^{n} \Psi(x_i;\theta)$$

where $\mathbf{x} = (x_1, \ldots, x_n)^\top$. We then define the estimator $\hat{\theta} \;=\; \hat{\theta}_n \;=\; \hat{\theta}_n(\mathbf{x})$ as the solution of the equation

$$\sum_{i=1}^{n} \Psi(x_i;\hat{\theta}_n) \;=\; \Psi_n(\mathbf{x};\hat{\theta}_n) \;=\; 0. \tag{4.2}$$

In other words, we are considering inference functions in the extended model which are additive over the sample. This approach has the advantage that it makes it easy to tie up with the classical asymptotic theory under a repeated sampling scheme, and with the theory of empirical processes (see Van der Vaart and Wellner, 1996).

An inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ is said to be *unbiased* if for each $\theta \in \Theta$

$$\int \Psi(x;\theta) P_\theta(dx) \;=\; 0\,. \tag{4.3}$$

Let us introduce the notation $\Psi(\theta)$ to denote the random vector $\Psi(\cdot;\theta)$ and $\mathrm{E}_\theta(X)$ the expectation of the random vector X under P_θ. Then for each $\theta \in \Theta$, the expectation (4.3) becomes

$$\mathrm{E}_\theta\{\Psi(\theta)\} \;=\; 0\,.$$

Clearly, if an inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ is unbiased, then the inference function $\Psi_n : \mathcal{X}^n \times \Theta \longrightarrow \mathbb{R}^k$ based on a sample of size n (*i.e.* $\Psi_n(\mathbf{x};\theta) = \sum_{i=1}^n \Psi(x_i;\theta)$) is also unbiased (and vice-versa).

In the rest of this section we try to elucidate the meaning of the assumption of unbiasedness in the context of a one-dimensional parameter (*i.e.* $k = 1$). We will not obtain full generality of the arguments presented here, but try to be illustrative.

We start by fixing a value of the parameter, say $\theta_0 \in \Theta \subseteq I\!R$, which is regarded as the "true value of the parameter θ". Define the function $\lambda_{\theta_0} : \Theta \longrightarrow I\!R$ by

$$\lambda_{\theta_0}(\theta) = \lambda(\theta) = \mathrm{E}_{\theta_0}\{\Psi(\theta)\}. \tag{4.4}$$

Clearly, Ψ is unbiased if and only if

$$\lambda_{\theta_0}(\theta_0) = 0 \ , \ \forall\, \theta_0 \in \Theta.$$

Now if Ψ is not unbiased then there exists at least one member of Θ, say θ_1, such that $\lambda_{\theta_1}(\theta_1) \neq 0$. If, additionally, $\lambda_{\theta_1} : \Theta \longrightarrow I\!R$ is continuous in a neighbourhood of θ_1, then there exists an $\epsilon > 0$ such that for all $\theta \in (\theta_1 - \epsilon, \theta_1 + \epsilon)$

$$\lambda_{\theta_1}(\theta) \ \neq \ 0\,.$$

Hence, for n large enough, there would be no roots of Ψ_n in the interval $(\theta_1 - \epsilon, \theta_1 + \epsilon)$, P_{θ_1}-almost surely. Then the sequence of estimators defined by the roots of Ψ_n would not be (strongly) consistent! The assumption of continuity of λ_{θ_1} holds in most of the cases encountered in practice and it is somehow related with the stability of the inference function in play.

Clearly, the problem presented above does not occur with unbiased inference functions. However, one should be careful to post the reciprocal of the argument presented. One possibility is given in the next theorem.

Theorem 4.1 *Suppose that $\Psi_n(\cdot)$ is continuous $[P_{\theta_0}]$, and that there exists a $\delta_0 > 0$ such that for all $\theta \in (\theta_0 - \delta_0, \theta_0)$*

$$\lambda(\theta) > 0$$

and for all $\theta \in (\theta_0, \theta_0 + \delta_0)$

$$\lambda(\theta) < 0.$$

Then there exists a sequence of roots of Ψ_n, say $\{\hat{\theta}_n\}$, such that

$$\hat{\theta}_n \xrightarrow{P_{\theta_0}} \theta_0.$$

Proof: Take $\delta \in (0, \delta_0)$. From the strong law of large numbers

$$\Psi_n(\theta_0 - \delta) \longrightarrow \lambda(\theta_0 - \delta) > 0 \quad [P_{\theta_0}],$$

and

$$\Psi_n(\theta_0 + \delta) \longrightarrow \lambda(\theta_0 + \delta) < 0 \quad [P_{\theta_0}],$$

as $n \to \infty$. Hence, there exists an $n_0 = n_0(\delta)$ such that for all $n \geq n_0$ one has, almost surely with respect to P_{θ_0}, that

$$\Psi_n(\theta_0 - \delta) > 0$$

and

$$\Psi_n(\theta_0 + \delta) < 0.$$

Then from the continuity of Ψ_n, there exists, almost surely with respect to P_{θ_0}, a root $\hat{\theta}_n(\delta)$ of Ψ_n in the interval $(\theta_0 - \delta, \theta_0 + \delta)$ such that

$$P_{\theta_0}(|\hat{\theta}_n(\delta) - \theta_0| < \delta) \longrightarrow 1,$$

as $n \to \infty$. Now, instead of $\hat{\theta}_n(\delta)$, we take the root $\hat{\theta}_n$ which is closest to θ_0. Clearly, this root does not depend on δ and also satisfies

$$P_{\theta_0}(|\hat{\theta}_n - \theta_0| < \delta) \longrightarrow 1\,, \tag{4.5}$$

as $n \longrightarrow \infty$. Note that the existence of $\hat{\theta}_n(\delta)$ ensures the existence of at least one root of Ψ_n that satisfies (4.5). From the arbitrariness of the choice of δ in $(0, \delta_0)$, one concludes from (4.5) that $\hat{\theta}_n$ converges in probability to θ_0 under P_{θ_0}. □

4.3 Optimality of inference functions

In this section, we study a criterion for selecting inference functions which is reminiscent of the classical theory of point estimation. The idea is to restrict attention to a class of inference functions in such a way that it will be easy to tie up with the classical theory of point estimation. It is then desirable to have a set of conditions that ensure at least that the associated estimators are consistent and asymptotically normally distributed. Then we choose the inference function with the smallest asymptotic variance. This "optimal" inference function is equivalent to the score function in most of the well behaved cases, providing an alternative justification for the maximum likelihood estimator. However, the original formulation of the optimality theory, due to Godambe, is noteworthy in that it does not involve asymptotic arguments. Actually, Godambe's conditions used for defining regular inference functions are not sufficient to ensure asymptotic normality. This makes the coincidence of the "optimal" inference function and the score function even more interesting, provided one agrees with the intuitive justification of the definition of information.

4.3.1 The one-dimensional case

Let $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ be a statistical model, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. In order to postpone technicalities, we first study the case where the parameter space Θ is one-dimensional; *i.e.* Θ is a subset of the real line. Moreover, we assume that each distribution in $\mathcal{P}$ is dominated by

a common σ-finite measure μ. For each $P_\theta \in \mathcal{P}$, we choose a version of the Radon-Nikodym derivative with respect to μ, which we denote by

$$p(\cdot;\theta) = \frac{dP_\theta}{d\mu}(\cdot).$$

An inference function $\psi : \mathcal{X} \times \Theta \to I\!R$ is said to be a *regular inference function* when the following conditions are satisfied for all $\theta \in \Theta$:

i) $\mathrm{E}_\theta\{\psi(\theta)\} = 0$ (*i.e.* ψ is unbiased);

ii) The partial derivative of $\psi(x;\theta)$ with respect to θ exists for μ-almost every $x \in \mathcal{X}$.

ii) The order of integration and differentiation may be interchanged as follows:

$$\frac{d}{d\theta}\int_{\mathcal{X}} \psi(x;\theta)p(x;\theta)d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left[\psi(x;\theta)p(x;\theta)\right] d\mu(x);$$

iv) $0 < \mathrm{E}_\theta\{\psi^2(\theta)\} < \infty$;

v) $0 < \mathrm{E}^2_\theta\left|\partial\psi(\theta)/\partial\theta\right| < \infty$.

We let $\mathcal{G}$ denote the class of regular inference functions.

If the score function $U(x;\theta) = \partial \log p(x;\theta)/\partial\theta$ is a regular inference function and $\Theta \subseteq I\!R$ is an open interval, then the statistical model is said to be *regular*. We will consider only regular statistical models. However, the reader should be aware that there are simple examples of non-regular models. The next example illustrates this point.

Example 4.2 *(A non-regular model)* Consider the uniform distribution on the interval $[0,\theta]$, where $\theta \in \Theta = (0,\infty)$. Then we may take $p(x;\theta) = \theta^{-1}I_{[0,\theta)}(x)$. Consequently,

$$\log p(x;\theta) = \begin{cases} \log\theta^{-1} & \text{if } x \in [0,\theta) \\ -\infty & \text{otherwise.} \end{cases}$$

Note that $\partial \log p(x;\theta)/\partial\theta$ is not defined for $x \geq \theta$, but since the set $[\theta,\infty)$ has P_θ-measure zero, we may define the score function arbitrarily there. Accordingly, we let

$$U(x;\theta) = \begin{cases} -\theta^{-1} & \text{if } x \in [0,\theta) \\ 0 & \text{otherwise.} \end{cases}$$

Clearly U is not unbiased, so the model is not regular.

Before developing the optimality theory for inference functions, we analyse the assumption of regularity in more detail. We first define the function $i : \Theta \to (0,\infty)$ by

$$i(\theta) = \mathrm{E}_\theta\{U^2(\theta)\},$$

which is called the *Fisher information*. Using condition *iii*), it can be easily shown that

$$i(\theta) = -\mathrm{E}_\theta \left[\frac{\partial^2 \log p(x;\theta)}{\partial \theta^2} \right].$$

provided that the second partial derivative $\partial^2 \log p(x;\theta)/\partial\theta^2$ exists for μ-almost every x. Henceforth we assume the existence of this derivative. Condition *iv*) ensures that $i(\theta) > 0$ for each $\theta \in \Theta$.

It is convenient to denote the partial derivatives $\partial\psi/\partial\theta$ and $\partial^2\psi/\partial\theta^2$ and by $\psi'(\theta)$ and $\psi''(\theta)$, respectively.

We now define the Godambe information and provide an intuitive justification. We define two real functions associated with a regular inference function $\psi \in \mathcal{G}$. The *variability* $V_\psi(\cdot)$ of ψ is defined as

$$V_\psi : \theta \longmapsto V_\psi(\theta) = \mathrm{E}_\theta \psi^2(\theta) = V_\theta\{\psi(\theta)\},$$

while the *sensitivity* $S_\psi(\cdot)$ of ψ is defined as

$$S_\psi : \theta \longmapsto S_\psi(\theta) = \mathrm{E}_\theta\{\psi'(\theta)\}.$$

Clearly, one seeks inference functions with low variability, meaning that the inference function assumes a value close to its mean (zero) at the true value of the parameter. One also seeks high absolute value of the sensitivity, meaning that small parameter changes to the inference function in a neighbourhood of the true value provoke large changes in the inference function.

Different inference functions cannot be compared on the basis of variability alone. Equivalent inference functions (*i.e.* those yielding identical estimators) can have different variability (for example, multiply a given inference function by 2). One way to deal with this problem, which also takes the sensitivity into account, is to use the *Godambe information* (in reference to the pioneering work of V.P. Godambe), which can be expressed as

$$J_\psi(\theta) = \frac{S_\psi^2(\theta)}{V_\psi(\theta)}.$$

Inference functions with high absolute sensitivity and low variability then correspond to functions having high Godambe information.

An alternative interpretation of the Godambe information is obtained through the notion of *standardized inference function*. Recall the equivalence relation "$\sim$" between regular inference functions. From each equivalence class, say $[\psi]$, choose the (unique) function in that class having sensitivity equal to one (see Problem 4.8 item i)). For each equivalence class $[\psi]$ (associated with a regular inference function ψ), this function is the *standardized version* of ψ defined by

$$\tilde{\psi}(x;\theta) = \frac{\psi(x;\theta)}{S_\psi(\theta)},$$

$\theta \in \Theta$, $x \in \mathcal{X}$. Having chosen this representative from each equivalence class, we make comparisons between different classes by comparing the variability of their representatives.

Now the variability $V_{\tilde{\psi}}$ of the standardized version of ψ can be obtained from the Godambe information of ψ; *i.e.* for each $\theta \in \Theta$

$$V_{\tilde{\psi}}(\theta) = J_{\psi}^{-1}(\theta).$$

Hence, we need only use the Godambe information to compare inference functions.

Note that the above discussion makes no direct reference to asymptotic arguments. This is interpreted in the literature (in particular the many papers by Godambe) as meaning that there is a finite sample justification (maximizing the Godambe information) for the optimality theory for inference functions. However there are some criticisms of this viewpoint (see Pfanzagl, 1990, p. 36) since the arguments make no reference to properties of the sequence of estimators $\{\hat{\theta}_n\}$, which is in fact the primary object of interest.

We now study the asymptotic distribution of the roots of a regular inference function ψ. Suppose $x_1, x_2, x_3, \ldots$ are a sequence of independent observations drawn from an unknown distribution $P_\theta \in \mathcal{P}$. In the following theorem, we consider a sequence of estimators $\{\hat{\theta}_n\}_{n\geq 1} = \{\hat{\theta}_n(x_1, \cdots, x_n)\}_{n\geq 1}$ satisfying the equations

$$\sum_{i=1}^{n} \psi(x_i; \hat{\theta}_n) = 0, \tag{4.6}$$

for each $n \geq 1$. The following result shows that, under regularity conditions, J_ψ is the inverse of the asymptotic variance of $\{\hat{\theta}_n\}_{n\geq 1}$, providing a further (asymptotic) argument for preferring inference functions with large Godambe information.

Theorem 4.3 *Let $\psi : \mathcal{X} \times \Theta \to I\!R$ be a regular inference function and let $\{\hat{\theta}_n\}_{n\geq 1}$ be a sequence of estimators satisfying (4.6) for each $n \geq 1$. Assume that the model is regular. Suppose that there is a $\theta \in \Theta$ such that,*

$$\hat{\theta}_n \xrightarrow{P} \theta \quad \text{as } n \to \infty\,,$$

under P_θ and that $\psi(x;\cdot)$ is twice continuously differentiable, and that there exists a constant c and a measurable function $M : \mathcal{X} \longrightarrow I\!R$ such that for all $x \in \mathcal{X}$ and all $\theta_* \in (\theta - c, \theta + c)$,

$$|\psi''(x;\theta_*)| < M(x)\,, \tag{4.7}$$

and $\int_{\mathcal{X}} M(x)p(x;\theta)\mu(dx) < \infty$. We then have that for all $\theta \in \Theta$,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, J_{\psi}^{-1}(\theta))$$

under P_θ.

Here "$\xrightarrow{\mathcal{D}}$" denotes convergence in distribution (relative to P_θ) and "$\xrightarrow{P}$" convergence in probability.

The assumptions of Theorem 4.3 can be relaxed somewhat. Actually it is not necessary to assume the kind of boundedness given by (4.7), but this assumption simplifies very much the proof. In Problem 4.9 the proof of Theorem 4.3 without assumption (4.7) (but assuming instead that $\int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x) \in I\!R$) is sketched for the interested reader.

Proof: For each $n \in N$ and each sample $\mathbf{x} = (x_1, ..., x_n)$ define the functions $\Psi_n(\mathbf{x}; \cdot) : \Theta \to I\!R$ by, for $\theta \in \Theta$,

$$\Psi_n(\mathbf{x};\theta) = \sum_{i=1}^{n} \psi(x_i;\theta),$$

and let $\Psi'_n(\mathbf{x}; \cdot)$ and $\Psi''_n(\mathbf{x}; \cdot)$ denote the derivatives of Ψ_n with respect to θ. A Taylor expansion with Lagrange reminder term of $\Psi_n(\hat{\theta}_n)$ around θ yields

$$0 = \Psi_n(\mathbf{x};\hat{\theta}_n) = \Psi_n(\mathbf{x};\theta) + (\hat{\theta}_n - \theta)\Psi'_n(\mathbf{x};\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)^2\Psi''_n(\mathbf{x};\theta^*_n). \tag{4.8}$$

where θ^*_n lies between θ and $\hat{\theta}_n$. Solving for $\theta - \hat{\theta}_n$ and multiplying by $\sqrt{n}$ gives,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{-\frac{1}{\sqrt{n}}\Psi_n(\mathbf{x};\theta)}{\frac{1}{n}\Psi'_n(\mathbf{x};\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)\frac{1}{n}\Psi''_n(\mathbf{x};\theta^*_n)}\,. \tag{4.9}$$

The Central Limit Theorem gives

$$\frac{1}{\sqrt{n}}\Psi_n(\theta) \xrightarrow{\mathcal{D}} N\left(0, V_\psi(\theta)\right)\,.$$

We show that the denominator of (4.9) converges in probability to $S_\psi(\theta)$, which in light of Slutsky's Theorem and he last remark implies the theorem.

Note that, from the law of large numbers, $\frac{1}{n}\Psi'_n(\theta) \xrightarrow{P} S_\psi(\theta)$. Hence we just have to prove that $\frac{1}{n}\Psi''_n(\theta^*_n)$ remains bounded in probability as $n \to \infty$, because $\hat{\theta}_n - \theta \xrightarrow{P} 0$. Using (4.7) and the law of large numbers,

$$\left|\frac{1}{n}\Psi''_n(\theta^*_n)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\psi''(x_i;\theta^*_n)\right| \leq \frac{1}{n}\sum_{i=1}^{n} M(x_i) \xrightarrow{P} \int_{\mathcal{X}} M(x)p(x;\theta)\mu(dx) < \infty\,. \tag{4.10}$$

We conclude that $\frac{1}{n}\Psi''_n(\theta^*_n)$ remains bounded in probability, which proves the theorem. □

4.3.2 Regular asymptotic linear estimators

In the last section we gave an interpretation of the Godambe information based on the asymptotic normality of the sequence of estimates associated with a sufficiently regular inference function. In the proof of Theorem 4.3, on the asymptotic normality of the roots of inference functions, the crucial step was a Taylor expansion of a regular inference function. In this section we use the same idea to propose a rich class of estimators which are consistent and

asymptotically normally distributed. That is, we consider the class of estimates for which we have the expansion we need, which of course will contain the class of estimates considered in the Theorem 4.3.

Proceeding in this way, we stress what from our point of view is the mathematical kernel of the theory of optimality. This differs from the original formulation in which the motivation was more in the direction of reproducing, in the context of inference functions, the classical theory of optimality of point estimation, a kind of analogy which is explored in the next section. Finally we note that the class of estimators we consider here has often been used to obtain optimality results in more general contexts than the one considered here, see for instance Bickel *et al.* (1993).

We say that a sequence of estimates $\{\hat{\theta}_n\}_{n\in N} = \{\hat{\theta}_n(x_1,\ldots,x_n)\}_{n\in N}$, based on a sample $x_1,\ldots,x_n$, is *regular asymptotic linear* if there exists a measurable function $\psi : \mathcal{X}\times\Theta \longrightarrow I\!R$ and a random sequence $\{a_n\}$ such that for each $\theta\in\Theta$

$$n/a_n \xrightarrow{P} 1, \text{ under } P_\theta; \tag{4.11}$$

$$\psi(\theta) \in L^2(P_\theta)\ ; \tag{4.12}$$

$$\int \psi(x;\theta) dP_\theta(x) \ = \ 0\ ; \tag{4.13}$$

$$\hat{\theta}_n \ = \ \theta \ + \ \frac{1}{a_n}\sum_{i=1}^{n}\psi(x_i;\theta) \ + \ o_{P_\theta}(n^{-1/2})\ . \tag{4.14}$$

The function ψ is said to be the *influence function* associated with $\{\hat{\theta}_n\}$. This definition is slightly more general than the definition by Bickel *et al.* (1993), whose definition corresponds to taking $a_n = n$.

Theorem 4.4 *If $\{\hat{\theta}_n\}_{n\in N}$ is regular asymptotic linear, then for each $\theta\in\Theta$,*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, V_\psi(\theta)),$$

where $V_\psi(\theta) = \int \psi^2(x;\theta)\, dP_\theta(x)$ denotes the variability of ψ.

Proof: From (4.11)–(4.14) we have that

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta) &= \frac{\sqrt{n}}{a_n}\sum_{i=1}^{n}\psi(x_i;\theta) + \sqrt{n}\, o_{P_\theta}(n^{-1/2}) \\ &= \frac{n}{a_n\sqrt{n}}\sum_{i=1}^{n}\psi(x_i;\theta) + o_{P_\theta}(1).\end{aligned}$$

On the other hand, by the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x_i;\theta) \xrightarrow{\mathcal{D}} N(0, V_\psi(\theta)),$$

and

$$n/a_n \xrightarrow{P} 1\,,$$

as $n \to \infty$. By Slutsky's Theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{n}{a_n}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(x_i;\theta) \xrightarrow{\mathcal{D}} N(0, V_\psi(\theta)).$$

□

Proposition 4.5 *Under the assumptions of Theorem 4.3 (or Problem 4.9), the sequence of roots, $\{\hat{\theta}_n\}$, of a regular inference function $\psi : \chi \times \Theta \to I\!R$ is regular asymptotic linear.*

Proof: Without loss of generality, assume that ψ is standardized, *i.e.*

$$S_\psi(\theta) = \mathrm{E}_\theta\{\psi'(\,\theta)\} = 1.$$

Take $a_n = \Psi'_n(\mathbf{x};\theta) + \frac{1}{2}(\hat{\theta}_n - \theta)\Psi''_n(\mathbf{x};\theta^*_n)$. From the law of large numbers one has

$$\frac{1}{n}\Psi'_n(\theta) \xrightarrow{P} \mathrm{E}_\theta\{\psi'(\theta)\} = 1.$$

From the argument given in the proof of Theorem 4.3 the second term of a_n divided by n converges in probability to zero. Using Slutsky's Theorem it follows that $n/a_n \xrightarrow{P} 1$. The theorem follows by solving Equation 4.8 for $\hat{\theta}_n$ and defining the term $o_{P_\theta}(n^{-1/2}) = 0$. □

The previous theorem shows that regular asymptotic linearity is sufficient to have consistency and asymptotic normality. We show now that, in a certain sense, these are the minimal natural conditions, in the sense that in order to drop one of the assumptions (4.11)–(4.14) and keep the consistency and asymptotic normality, one should introduce rather complicated and artificial conditions. First of all, we have already given arguments showing that the unbiasedness, *i.e.* (4.13), is a kind of minimal condition for having consistency provided one also has asymptotic normality. Moreover, (4.12) and (4.13) ensure asymptotic normality via the Central Limit Theorem in a natural way when an independent identically distributed scheme is assumed. Of course, in view of Feller's reciprocal of the Central Limit Theorem, Lindeberg's condition would be the most general condition to have asymptotic normality. However, we should agree that sequences of estimators based on inference functions that satisfy Lindeberg's condition but not (4.14) are rare.

4.3.3 Generalizations of classical estimation theory

In this section, we explore the mathematical structure associated with the class of regular inference functions to obtain some parallels with the classical theory of point estimation. This approach is in the spirit of the original formulation of the optimality theory of inference functions due to Godambe. The assumption of regularity of the inference function is crucial here.

It can be shown that the asymptotic variance of the sequence of estimators $\{\hat{\theta}_n\}_{n\geq 1}$ associated with a regular inference function ψ cannot be smaller than the inverse of the Fisher information $i(\theta)$, provided that the sequence of estimators is consistent and asymptotically normally distributed. On the other hand, $i(\theta)$ is precisely the Godambe information of the score function. This is closely connected with the following theorem.

Theorem 4.6 *Under the previous assumptions, if ψ is a regular inference function, then for every $\theta \in \Theta$,*

$$J_\psi(\theta) \leq i(\theta). \tag{4.15}$$

Equality is attained in (4.15) if and only if ψ is equivalent to the score function.

Proof: First note that since ψ is unbiased, we have

$$\int_{\mathcal{X}} \psi(x;\theta)p(x;\theta)d\mu(x) = 0. \tag{4.16}$$

Differentiating this expression with respect to θ and interchanging the order of integration and differentiation, we obtain

$$\begin{aligned} 0 &= \int_{\mathcal{X}} \psi'(x;\theta)p(x;\theta)\, d\mu(x) + \int_{\mathcal{X}} \psi(x;\theta)\frac{\partial p(x;\theta)}{\partial\theta}\, d\mu(x) \\ &= \mathrm{E}_\theta\psi'(\theta) + \mathrm{E}_\theta[\psi(\theta)U(\theta)]. \end{aligned} \tag{4.17}$$

It follows that

$$\mathrm{E}^2_\theta\{\psi'(\theta)\} = \mathrm{E}^2_\theta[\psi(\theta)U(\theta)]. \tag{4.18}$$

From (4.16) (and the unbiasedness of the score function), the right-hand side of this equation is the squared correlation between $\psi(\theta)$ and $U(\theta)$. By the Cauchy-Schwartz inequality,

$$\mathrm{E}^2_\theta[\psi(\theta)U(\theta)] \leq \mathrm{E}_\theta\{\psi^2(\theta)\}\mathrm{E}_\theta\{U^2(\theta)\}.$$

Applying this result to (4.18) and dividing through by $\mathrm{E}_\theta\psi^2(\theta)$, we obtain the desired inequality $J_\psi(\theta) \leq i(\theta)$. Equality is obtained in this expression if and only if there exist real parameters $a(\theta)$ and $k(\theta)$ such that $\psi(\theta) = a(\theta) + k(\theta)U(\theta)$. But since $\psi(\theta)$ is unbiased, we must have $a(\theta) = 0$. Hence the Godambe information attains the upper bound $i(\theta)$ if and only if the inference function is equivalent to the score function $U(\theta)$. □

We stress that the bound obtained in (4.15) is a generalization of the classical information inequality. Thus, consider a statistic $T = t(X)$ with finite variance, *i.e.* $V_\theta\{t(X)\} < \infty$ for all $\theta \in \Theta$. Define the unbiased inference function ψ given by

$$\psi(x;\theta) = t(x) - \mathrm{E}_\theta t(X)$$

for all $x \in \mathcal{X}$ and $\theta \in \Theta$. We assume that ψ is regular. (The reader is invited to check that this is equivalent to the classical assumptions of estimation theory). Noting that $\mathrm{E}_\theta\psi'(\theta) = -\partial\{\mathrm{E}_\theta t(X)\}/\partial\theta$ and $\mathrm{E}_\theta\{\psi^2(\theta)\} = V_\theta\{t(X)\}$, (4.15) implies

$$\frac{\left[\frac{\partial}{\partial\theta}\mathrm{E}_\theta t(X)\right]^2}{V_\theta\{t(X)\}} = \frac{\mathrm{E}^2_\theta\psi'(\theta)}{\mathrm{E}_\theta\{\psi^2(\theta)\}} \leq i(\theta), \tag{4.19}$$

which is the classical information inequality. Moreover, if T is unbiased, *i.e.* $\mathrm{E}_\theta t(X) = \theta$ for all $\theta \in \Theta$, inequality (4.19) becomes

$$\frac{1}{V_\theta\{t(X)\}} \leq i(\theta), \tag{4.20}$$

which is the Cramér-Rao inequality.

We note that equality holds in (4.20) if and only if the statistical model is an exponential family. To see this, suppose that $\psi(x;\theta) = t(x) - \theta$ is an optimal inference function (*i.e.* the bound (4.20) holds with equality). Then $\psi(x;\theta) = k(\theta)U(x;\theta)$ for some $k : \Theta \longrightarrow I\!R \setminus \{0\}$. Hence, for some $\theta_0 \in \Theta$ and $a : \mathcal{X} \longrightarrow I\!R$, the log-likelihood function has the form

$$\begin{aligned} l(x;\theta) &= a(x) + \int_{\theta_0}^{\theta} U(z)dz \\ &= a(x) + t(x)\int_{\theta_0}^{\theta} \frac{1}{k(z)}dz - \int_{\theta_0}^{\theta} \frac{z}{k(z)}dz \\ &= a(x) + t(x)\alpha(\theta) - \beta(\theta), \end{aligned}$$

where α and β are the integrals appearing in the second equality. Hence we have

$$p(x;\theta) = e^{a(x)}e^{t(x)\alpha(\theta)-\beta(\theta)},$$

which specifies an exponential family. In particular, the Cramér-Rao bound is attained with equality if and only if θ is the mean parameter of an exponential family.

The connection between classical estimation theory and the theory of optimal inference functions is not restricted to the information inequality. Many other analogies exist. For example, the following theorem generalizes the classical Rao-Blackwell theorem of estimation theory.

Theorem 4.7 *Let $t(X)$ be a B-sufficient statistic (sufficient in the sense defined in Chapter 2) for the family $\{P_\theta : \theta \in \Theta \subseteq I\!R\}$, and let Φ be a regular inference function. Define*

$$\psi(t_0;\theta) = E_\theta\{\Phi(X;\theta)|T = t_0\}.$$

Then under the previous regularity conditions, for all $\theta \in \Theta$,

$$J_\Phi(\theta) \leq J_\psi(\theta), \tag{4.21}$$

with equality if and only if, for all $\theta \in \Theta$,

$$\Phi(x;\theta) = \psi(t(x);\theta) \quad [P_\theta].$$

Before presenting the proof, we show that this theorem is indeed a generalization of the Rao-Blackwell theorem. Define, for each $\theta \in \Theta$ and $x \in \mathcal{X}$,

$$\Phi(x;\theta) = f(x) - \theta,$$

where $f(x)$ is an unbiased estimator of θ. In the notation of the theorem, one has

$$\psi(t_0;\theta) = f^*(t_0) - \theta,$$

where $f^*(t_0)$ is the conditional expectation of $f(X)$ given $T = t_0$. The inequality (4.21) then becomes

$$V_\theta\{f^*(t(X))\} \leq V_\theta\{f(X)\},$$

with equality if and only if $f(x) = f^*(t(x))$ $[P_\theta]$. This is exactly the Rao-Blackwell theorem.

It is also possible to obtain an analogue of the Lehmann-Scheffé Theorem in the context of inference functions. However, there is no need to do so, because we have already found an optimal inference function, namely the score function.

The theorem basically says that if we want to maximize the Godambe information, then we need only consider inference functions that depend on the data through a sufficient statistic. This idea is consistent with the principle of sufficiency.

Proof: Note that for all $\theta \in \Theta$,

$$\mathrm{E}_\theta\psi(\theta) = \mathrm{E}_\theta\Phi(\theta) = 0.$$

Moreover,

$$\mathrm{E}_\theta\Phi^2(\theta) = \mathrm{E}_\theta\left[\mathrm{Var}_\theta\{\Phi(\theta)|T\}\right] + \mathrm{E}_\theta\psi^2(\theta).$$

It then follows that

$$\begin{aligned} 0 &= \frac{d}{d\theta}\mathrm{E}_\theta\psi(\theta) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left[\psi(x;\theta)p(x;\theta)\right]dx \\ &= \int \psi'(x;\theta)p(x;\theta)dx + \int \psi(x;\theta)\frac{\partial}{\partial\theta}p(x;\theta)dx \\ &= \mathrm{E}_\theta\psi'(\theta) + \mathrm{E}_\theta[\psi(\theta)U(\theta)], \end{aligned}$$

and therefore

$$-\mathrm{E}_\theta\psi'(\theta) = \mathrm{E}_\theta[\psi(\theta)U(\theta)].$$

Here we have invoked the regularity conditions to interchange the order of integration and differentiation. Since T is sufficient,

$$\begin{aligned} -\mathrm{E}_\theta\psi'(\theta) = \mathrm{E}_\theta[\psi(\theta)U(\theta)] &= \mathrm{E}_\theta\left\{\mathrm{E}_\theta[\Phi(\theta)|T]U(\theta)\right\} \\ &= \mathrm{E}_\theta[\Phi(\theta)U(\theta)] \\ &= -\mathrm{E}_\theta(\Phi'(\theta)). \end{aligned}$$

Hence we obtain, for all $\theta \in \Theta$,

$$J_\Phi(\theta) = \frac{\mathrm{E}_\theta^2[\psi'(\theta)]}{\mathrm{E}_\theta\psi^2(\theta) + \mathrm{E}_\theta[\mathrm{Var}_\theta\{\Phi(\theta)|T\}]} \leq J_\psi(\theta).$$

□

4.3.4 The multidimensional case

In this section, we extend the concepts of the previous section to the multidimensional context. The parameter space Θ is then assumed to be an open subset of $\mathbb{R}^k$, where $k \in \mathbb{N}$. As before, it is assumed that we have a parametric family of distributions $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$ and a σ-finite measure μ defined on a given measurable space $(\mathcal{X}, \mathcal{A})$. For each $P_\theta \in \mathcal{P}$, we choose a version of the Radon-Nikodym derivative (with respect to μ), denoted by

$$p(\,\cdot\,;\theta) = \frac{dP_\theta}{d\mu}(\cdot) \ .$$

We now extend the notion of regular inference function to this multidimensional context. A function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ is said to be a *regular inference function* when the following conditions are satisfied for all $\theta = (\theta_1, \cdots, \theta_k) \in \Theta$ and for $i, j = 1, \ldots, k$.

(i) $\mathrm{E}_\theta\{\Psi(\theta)\} = 0$;

(ii) The partial derivative $\partial\Psi(x;\theta)/\partial\theta_i$ exists for μ-almost every $x \in \mathcal{X}$;

(iii) The order of integration and differentiation may be interchanged as follows:

$$\frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} \Psi(x;\theta)p(x;\theta)d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left[\Psi(x;\theta)p(x;\theta)\right]d\mu(x) \ ;$$

(iv) $\mathrm{E}_\theta\{\psi_i(\theta)\psi_j(\theta)\} \in \mathbb{R}$ and the $k \times k$ matrix

$$V_\psi(\theta) = \mathrm{E}_\theta\{\Psi(\theta)\Psi^\top(\theta)\}$$

is positive-definite;

(v) $\mathrm{E}_\theta\left\{\frac{\partial\psi_i}{\partial\theta_r}(\theta)\frac{\partial\psi_j}{\partial\theta_s}(\theta)\right\} \in \mathbb{R}$ and the $k \times k$ matrix

$$S_\psi(\theta) = \mathrm{E}_\theta\{\nabla_\theta\Psi(\theta)\}$$

is nonsingular.

Here ψ_i denotes the ith component of the vector function

$$\Psi(\cdot) = (\psi_1(\cdot), \ldots, \psi_k(\cdot))^\top ,$$

and ∇_θ denotes the gradient operator relative to the vector θ, defined by

$$\nabla_\theta f(\theta) = \frac{\partial f}{\partial\theta^\top}(\theta).$$

It is easy to see that the preceding conditions generalize those given earlier for the one-dimensional case.

As before, if the score function

$$U(x;\theta) = \nabla_\theta^\top \log p(x;\theta)$$

is a regular inference function and Θ is an open region of $\mathbb{R}^k$, then the model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$ is said to be *regular*. Henceforth, we consider only regular models. In the multidimensional context, the Fisher information is the matrix-valued function $i : \Theta \longrightarrow \mathbb{R}^{k\times k}$ defined by

$$i(\theta) = \mathrm{E}_\theta[U(\theta)U^\top(\theta)].$$

Given a regular inference function Ψ, the *Godambe information* is the matrix-valued function $J_\Psi : \Theta \longrightarrow \mathbb{R}^{k\times k}$ defined by

$$\begin{aligned} J_\Psi(\theta) &= \left[\mathrm{E}_\theta^{-1}\{\nabla_\theta\Psi(\theta)\}\mathrm{E}_\theta\{\Psi(\theta)\Psi^\top(\theta)\}\mathrm{E}_\theta^{-\top}\{\nabla_\theta\Psi(\theta)\}\right]^{-1} \\ &= S_\Psi^\top(\theta)V_\Psi^{-1}(\theta)S_\Psi(\theta). \end{aligned}$$

Here $B^{-\top} = (B^{-1})^\top$. The following theorem gives an asymptotic interpretation of the Godambe information. The reader is invited to provide a finite sample justification similar to the one given in the preceding section.

Theorem 4.8 *Take $\theta \in \Theta$ fixed. Under the previous assumptions, if the sequence of estimators $\{\hat\theta_n\}_{n=1}^\infty$ associated with a regular inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ converges in probability to θ, then this sequence is asymptotically normal,*

$$\sqrt{n}(\hat\theta_n - \theta) \xrightarrow{\mathcal{D}} N_k\left(0, J_\Psi^{-1}(\theta)\right)$$

under P_θ.

The proof of this theorem can be obtained by modifying that of the one-dimensional version given previously (see Problem 4.10).

In the one-dimensional case, we used the Godambe information to compare two regular inference functions. We prefer the function having the larger Godambe information. In the multidimensional case, the Godambe information is no longer a number but a $k \times k$ matrix. Accordingly, we must be more careful when comparisons are made between inference functions. Several equivalent approaches for partially ordering the class of regular inference functions are presented in the literature (see Bhapkar (1972), Kale and Chandrasekar (1983), Chandrasekhar and Kale (1984), Chandrasekhar (1988), Joseph and Durairajan (1991) or Durairajan (1989)).

We shall here use the Löwner partial ordering of matrices. For $k \times k$ matrices A and B, we write "$A \geq B$" if and only if $A - B$ is positive-semi-difinite. Then given two inference functions Ψ and Φ, we prefer Ψ if, for all $\theta \in \Theta$, $J_\Psi(\theta) \geq J_\Phi(\theta)$. The regular inference function Ψ is said to be *optimal* when, for all $\theta \in \Theta$ and all regular inference functions Φ, one has

$$J_\Psi(\theta) \geq J_\Phi(\theta).$$

The natural interpretation of the criterion given above is the following. Due to the asymptotic normality of the sequence of estimators $\{\hat{\theta}_n\}_{n=1}^{\infty}$ associated with an inference function, one can construct an asymptotic confidence region for the parameter; a k-dimensional ellipsoid. Now, the volume of this ellipsoid is given by the determinant of the inverse Godambe information (*i.e.* the determinant of the asymptotic covariance matrix). Since $A \geq B$ implies that $\det(A - B) \geq 0$, we conclude that the optimal inference function minimizes the volume of the asymptotic confidence region.

The next result is again a multidimensional analogue of a previous theorem.

Theorem 4.9 *Under the previous assumptions, given a regular inference function Ψ, one has, for all $\theta \in \Theta$,*

$$J_{\Psi}(\theta) \leq i(\theta).$$

Equality holds if and only if Ψ is equivalent to the score function.

To show the theorem, we will first define some special subspaces of L^2 in which the components of the regular inference function and the components of the score function reside. These spaces are closed and hence are Hilbert spaces when considered with the natural inner product of L^2. With this convenient mathematical structure we will obtain a decomposition of each regular inference function into two orthogonal components, one of them equivalent to the score function. We will prove then that the information of the component of the score function that is equivalent to the score function is larger than the Godambe information of the original inference function.

We note that the technique we introduce here is rather general and, except for a more sophisticated decomposition of the regular inference function, the procedure is essentially the same as we will develop for the case with an arbitrary nuisance parameter.

For each $\theta \in \Theta$ define

$$L_0^2(P_\theta) = \left\{ f \in L^2(P_\theta) \, : \, \int f(x) \, dP_\theta(x) = 0 \right\}$$

and

$$\mathcal{U}_\theta = \text{span}\{U_i(\cdot;\theta) : i = 1, \ldots, k\}$$

where $U_i(\cdot;\theta)$ is the ith component of the score function, U, evaluated at θ. Note that if $\Psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$ is a regular inference function with components $\psi_1, \ldots, \psi_k$, then $\psi_i : \mathcal{X} \times \Theta \longrightarrow I\!R$ and $\psi_i(\cdot;\theta) \in L_0^2(P_\theta)$ for $i = 1, \ldots, k$. Furthermore, if the inference function Ψ is equivalent to the score function (in the sense of the equivalence relation "$\sim$" defined before) then, for $i = 1, \ldots, k$, $\psi(\cdot;\theta) \in \mathcal{U}_\theta$. The next lemma gives a convenient mathematical structure for L_0^2 and $\mathcal{U}_\theta$.

Lemma 4.10 *We have the inclusions $\mathcal{U}_\theta \subseteq L_0^2(P_\theta) \subset L^2(P_\theta)$, $\forall \theta \in \Theta$. Moreover, $\mathcal{U}_\theta$ and $L_0^2(P_\theta)$ are real, closed and convex vector subspaces of $L^2(P_\theta)$ (when endowed with the inner product of $L^2(P_\theta)$).*

Proof: Let $\theta \in \Theta$ be fixed but otherwise arbitrary. The inclusion $\mathcal{U}_\theta \subseteq L_0^2(P_\theta)$ follows from the fact that the score function is unbiased and belongs to $L^2(P_\theta)$. The other inclusion is obvious.

The convexity and the vector space structure of $\mathcal{U}_\theta$ and $L_0^2(P_\theta)$ are straightforward. Furthermore, since $\mathcal{U}_\theta$ is a finite-dimensional vector space, it is closed. For the closedness of $L_0^2(P_\theta)$ we argue as follows. Take an arbitrary generalized sequence $\{f_\alpha\}_{\alpha\in A}$ contained in $L_0^2(P_\theta)$ such that for some $f \in L^2(P_\theta)$, $f_\alpha \overset{L^2(P_\theta)}{\longrightarrow} f \quad$ as $\alpha \nearrow$. Here A is a given ascendent directed set. By the continuity of the inner product

$$0 = \int f_\alpha(x)\, dP_\theta(x) =< f_\alpha, \mathbf{1} >_\theta \longrightarrow < f, \mathbf{1} >_\theta = \int f(x)\, dP_\theta(x)$$

where $< \cdot, \cdot >$ is the inner product of $L^2(P_\theta)$ and $\mathbf{1}$ is the constantly equal to 1 function. Then $\int f(x)\, dP_\theta(x) = 0$ which implies that $f \in L_0^2(P_\theta)$. Since the sequence $\{f_\alpha\}$ is arbitrary we conclude that $L_0^2(P_\theta)$ is closed. □

We now define $\mathcal{A}_\theta$ to be the orthogonal complement of $\mathcal{U}_\theta$ with respect to $L_0^2(P_\theta)$, $\theta \in \Theta$.

Lemma 4.11 *Given a regular inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ with components $\psi_1, \ldots, \psi_k$, the following decomposition holds for each $x \in \mathcal{X}$ and $\theta \in \Theta$*

$$\Psi(x;\theta) = \Psi_a(x;\theta) + \Psi_U(x;\theta)$$

where

$$\Psi_a(x;\theta) = (\psi_{a1}(x;\theta), \ldots, \psi_{ak}(x;\theta))^\top$$
$$\Psi_U(x;\theta) = (\psi_{U1}(x;\theta), \ldots, \psi_{Uk}(x;\theta))^\top$$

with

$$\psi_{a1}(\cdot;\theta), \ldots, \psi_{ak}(\cdot;\theta) \in \mathcal{A}_\theta$$
$$\psi_{U1}(\cdot;\theta), \ldots, \psi_{Uk}(\cdot;\theta) \in \mathcal{U}_\theta$$

Proof: Let $i \in \{1, \ldots, k\}$ be arbitrary. From Lemma 4.10, $\psi_i(\cdot;\theta) \in L_0^2(P_\theta)$. By the orthogonal projection theorem, ψ_i can be decomposed as $\psi_i(\cdot;\theta) = \psi_{ai}(\cdot;\theta) + \psi_{Ui}(\cdot;\theta)$ and $\psi_{ai}(\cdot;\theta) \in \mathcal{A}_\theta$ and $\psi_{Ui}(\cdot;\theta) \in \mathcal{U}_\theta$, as was required. □

Lemma 4.12 *Let $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ be a regular inference function (except from the fact that here we allow the sensitivity to vanish) such that $\psi_i(\cdot;\theta) \in \mathcal{A}_\theta$ for $\theta \in \Theta$ and $i = 1, \ldots, k$. Then*

$$\int \left\{ \frac{\partial}{\partial \theta^i} \Psi(x;\theta) \right\} dP_\theta(x) = 0, \quad i = 1, \ldots, k$$

i.e. *the sensitivity of Ψ, S_Ψ, vanishes at θ.*

Proof: Take $\theta \in \Theta$ and $i \in \{1, \ldots, k\}$. Note that

$$\int \Psi(x;\theta) \frac{\partial}{\partial \theta^i} p(x;\theta)\, d\mu(x) = \int \Psi(x;\theta) U_i(x;\theta) p(x;\theta)\, d\mu(x) = 0$$

because $\psi_1(\cdot;x), \ldots, \psi(\cdot;\theta) \in \mathcal{A}_\theta$, *i.e.* they are orthogonal to the score function. Hence,

$$\begin{aligned}
\int \frac{\partial}{\partial \theta_i} \{\Psi(x;\theta) p(x;\theta)\}\, d\mu(x) &= \int \left\{ \frac{\partial}{\partial \theta_i} \Psi(x;\theta) \right\} p(x;\theta)\, d\mu(x) \\
&\quad + \int \Psi(x;\theta) \frac{\partial}{\partial \theta_i} p(x;\theta)\, d\mu(x) \\
&= \int \frac{\partial}{\partial \theta_i} \{\Psi(x;\theta) p(x;\theta)\}\, d\mu(x) \\
&= 0.
\end{aligned}$$

□

We now have the necessary elements to prove the optimality theorem.

Proof: Let $\theta \in \Theta$ be arbitrary. From the linearity of the gradient operator and the expectation we obtain

$$\begin{aligned}
S_\Psi(\theta) &= \mathrm{E}_\theta \{\nabla_\theta \Psi(\theta)\} = \mathrm{E}_\theta \left[\nabla_\theta \{\Psi_a(\theta) + \psi_U(\theta)\}\right] \\
&= \mathrm{E}_\theta \{\nabla_\theta \Psi_a(\theta)\} + \mathrm{E}_\theta \{\nabla_\theta \Psi_U(\theta)\} \\
&= S_{\Psi_a}(\theta) + S_{\Psi_U}(\theta) = S_{\Psi_U}(\theta).
\end{aligned}$$

The last equality follows from Lemma 4.12. On the other hand,

$$\begin{aligned}
V_\Psi(\theta) &= \mathrm{E}_\theta \{\Psi(\theta) \Psi(\theta)^\top\} \\
&= \mathrm{E}_\theta [\{\Psi_a(\theta) + \Psi_U(\theta)\} \{\Psi_a(\theta) + \Psi_U(\theta)\}^\top] \\
&= \mathrm{E}_\theta \{\Psi_a(\theta) \Psi_a^\top(\theta)\} + E\{\Psi_U(\theta) \Psi_U^\top(\theta)\} \\
&= V_{\Psi_a}(\theta) + V_{\Psi_U}(\theta).
\end{aligned}$$

Here the cross-products vanish because of orthogonality.

Hence, we have

$$\begin{aligned}
J_\Psi^{-1}(\theta) &= S_\Psi^{-1}(\theta) V_\Psi(\theta) S_\Psi^{-\top}(\theta) \\
&= S_{\Psi_U}^{-1}(\theta) V_\Psi(\theta) S_{\Psi_U}^{-\top}(\theta) \\
&= S_{\Psi_U}^{-1}(\theta) \left\{V_{\Psi_U}(\theta) + V_{\Psi_a}(\theta)\right\} S_{\Psi_U}^{-\top}(\theta) \\
&= J_{\Psi_U}^{-1}(\theta) + S_{\Psi_U}^{-1}(\theta) V_{\Psi_a} S_{\Psi_U}^{-\top}(\theta).
\end{aligned}$$

The second term in the last expression is positive-definite, so $J_\Psi^{-1}(\theta) \geq J_{\Psi_U}^{-1}(\theta)$. It follows easily that $J_\Psi(\theta) \leq J_{\Psi_U}(\theta)$. Note that $J_{\Psi_U}(\theta) = J_U(\theta)$, because $\psi_U \sim U$. Also, $J_U(\theta) = i(\theta)$. We have hence shown that $J_\Psi(\theta) \leq i(\theta)$ and that this upper bound is attained by the score function. □

4.4 Inference functions with nuisance parameters

We now develop the theory of inference functions for models with nuisance parameters. The idea of *nuisance parameters* seems to have more than one interpretation in the statistical literature. We use this term here in the sense defined in the following.

Consider a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, where the family of probability measures $\mathcal{P}$ is parametrized as

$$\mathcal{P} = \{P_{\theta\xi} : \theta \in \Theta \subseteq I\!R^k, \xi \in \mathcal{Z}\} \tag{4.22}$$

Here Θ is an open region of $I\!R^k$ and $\mathcal{Z}$ is an arbitrary set (typically infinite-dimensional). It is assumed that the parametrization of $\mathcal{P}$ given in (4.22) is identifiable, *i.e.*, the mapping $(\theta, \xi) \longmapsto P_{\theta\xi}$, from $\Theta \times \mathcal{Z}$ into $\mathcal{P}$, is one-to-one.

The parameter θ is considered as the parameter of interest, for which we would like to make inference, and the parameter ξ is understood as a nuisance parameter, which we have no interest in estimating. This kind of model is the prototype of a semiparametric model and this is the most general context we consider in this chapter. The terminology "semiparametric" is meant to reflect the fact that only the parameter θ is assumed to be finite-dimensional, whereas ξ is not necessarily finite-dimensional.

An *inference function*, in this new context, is a function, $\Psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$, of the observations and the parameter of interest taking values in $I\!R^k$ (recall that $\Theta \subseteq I\!R^k$), such that for each $\theta \in \Theta$, the function, $\Psi(\,\cdot\,;\theta) : \mathcal{X} \longrightarrow I\!R^k$, obtained by fixing the value of the parameter of interest, is measurable ($\mathcal{A}$, Borel($I\!R^k$) measurable). As before, we use the convenient notation $\Psi(\theta)$ to mean the random vector $\Psi(\,\cdot\,;\theta)$; moreover, $\psi_1, \ldots, \psi_k$ denote the k components of the function Ψ, ∇_θ is the gradient operator (with respect to θ), and so on.

We note that an inference function must *not* depend on the nuisance parameter. However, the nuisance parameter plays an important role in the theory that we develop here, and one should take some care in this respect. For instance, see the following definition of unbiased inference function.

An inference function Ψ is said to be *unbiased* when for each $\theta \in \Theta$ and each $\xi \in \mathcal{Z}$ one has

$$\mathrm{E}_{\theta\xi}\{\Psi(\theta)\} = 0.$$

Here $\mathrm{E}_{\theta\xi}(X)$ denotes, of course, the expectation of X under $P_{\theta\xi}$.

We study in the following an extension of the optimality theory to the context of models with nuisance parameters. We give also a brief introduction to a kind of inferential separation theory in this context. We close the section by developing some differential-geometric based arguments, showing that the so-called efficient score function is the only possible optimal inference function one can have, and give a criterion for the existence of optimal inference functions. No previous knowledge in differential geometry is assumed.

4.4.1 Optimality theory

We begin by assuming that the family $\mathcal{P}$ is dominated by a σ-finite measure μ and that for each $P_{\theta\xi} \in \mathcal{P}$ a version of the Radon-Nikodym derivation was chosen, and is denoted by

$$p(\,\cdot\,;\theta,\xi) = \frac{dP_{\theta\xi}}{d\mu}(\,\cdot\,).$$

An inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$ is said to be *regular* if for all $\theta \in \Theta$, all $\xi \in \mathcal{Z}$ and for $i, j = 1, \ldots, k$, the following conditions hold:

1. $\mathrm{E}_{\theta\xi}\{\Psi(\theta)\} = 0$;
2. The partial derivative $\partial\Psi(x;\theta)/\partial\theta_i$ exists for μ-almost every $x \in \mathcal{X}$;
3. The order of integration and differentiation may be interchanged as follows:

$$\frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} \Psi(x;\theta)p(x;\theta,\xi)d\mu(x) = \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left[\Psi(x;\theta)p(x;\theta,\xi)\right]d\mu(x) \ ;$$

4. $\mathrm{E}_{\theta\xi}\{\psi_i(\theta)\psi_j(\theta)\} \in I\!R$, and the $k \times k$ matrix

$$V_\Psi(\theta,\xi) = \mathrm{E}_{\theta\xi}\{\Psi(\theta)\Psi^\top(\theta)\}$$

   is positive definite;

5. $\mathrm{E}_{\theta\xi}\left\{\frac{\partial\psi_i}{\partial\theta_m}(\theta)\frac{\partial\psi_j}{\partial\theta_n}(\theta)\right\} \in I\!R$ and the $k \times k$ matrix

$$S_\Psi(\theta,\xi) = \mathrm{E}_{\theta\xi}\{\nabla_\theta\Psi(\theta)\}$$

   is nonsingular.

Conditions 1)–5) above are similar to the conditions given in the definition of regular inference function in the multivariate case, the only difference being the presence of the nuisance parameter. The class of regular inference functions is denoted $\mathcal{G}$.

It is convenient to introduce the notion of *regular quasi-inference function* (or in short *quasi-inference function*) which is a function, say $\Phi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow I\!R^k$, of the parameters and the observations, such that for each $\theta \in \Theta$ and each $\xi \in \mathcal{Z}$ the conditions (i)–(v) hold with $\Psi(\theta)$ replaced by $\Phi(\theta,\xi)$. We assume that the model is regular if the partial score function $U : \mathcal{X} \times \Theta \longrightarrow I\!R^k$ which for each $x \in \mathcal{X}$, $\theta \in \Theta$ and $\xi \in \mathcal{Z}$ is given by

$$U(x;\theta,\xi) \ = \ \nabla_\theta^\top[\log\{p(x;\theta,\xi)\}]$$

is a regular quasi-inference function and Θ is an open region of $I\!R^k$. Note that we here allow the partial score function to depend on the nuisance parameter.

As before, we define the *Godambe information* for a quasi-inference function, say $\Phi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow \mathbb{R}^k$, as the matrix-valued function $J_\Phi : \Theta \times \mathcal{Z} \longrightarrow \mathbb{R}^{k\times k}$, which for each $\theta \in \Theta$ and each $\xi \in \mathcal{Z}$ is given by

$$\begin{aligned} J_\Phi(\theta,\xi) &= \mathrm{E}^\top_{\theta\xi}[\nabla_\theta\{\Phi(\theta,\xi\}]\mathrm{E}^{-1}_{\theta\xi}\{\Phi(\theta,\xi)^\top\Phi(\theta,\xi)\}\mathrm{E}_{\theta\xi}[\nabla_\theta\{\Phi(\theta,\xi\}] \\ &= S^\top_\Psi(\theta,\xi)V^{-1}_\Psi(\theta,\xi)S_\Psi(\theta,\xi)\ . \end{aligned} \tag{4.23}$$

We note that, in the literature, the Godambe information is usually defined only for regular inference functions, but if a regular quasi-inference function does not depend on the nuisance parameter (*i.e.* it is a regular inference function) then the traditional notion of Godambe information coincides with our definition. In the following development, we deal several times with quasi-inference functions—such as the partial score function and the efficient score function—and then it is convenient to work with the extended setup given here.

We make the same use of the Godambe information as we did before, *i.e.*, we want to maximize the Godambe information. Thus, a regular inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ is said to be *optimal* when, for all $\theta \in \Theta$, all $\xi \in \mathcal{Z}$ and all regular inference functions $\Phi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$

$$J_\Psi(\theta,\xi) \geq J_\Phi(\theta,\xi). \tag{4.24}$$

Note that the symbol "$\geq$" in (4.24) means that the matrix $J_\Psi(\theta,\xi)$ is "larger than or equal to" the matrix $J_\Phi(\theta,\xi)$, in the sense of the partial ordering of matrices (as in Section 4.3). Moreover, (4.24) should hold for *all* $\theta \in \Theta$ and *all* $\xi \in \mathcal{Z}$. It is natural, then, to ask whether an optimal inference function exists. We anticipate the answer: there are simple situations where there exist no optimal regular inference functions. We will illustrate this situation with an example and give a theorem characterizing the situation at the end of this section.

4.4.2 One-dimensional interest parameter

Before developing a theory characterizing the existence of optimal inference functions for models with nuisance parameters, we present some classical results, due to Godambe, which allow us to compute optimal inference functions in many practical situations. For the sake of simplicity we study, in this section, models with a one-dimensional parameter of interest, but in the next section we work in full generality. It must be said that the development given here is slightly more general than Godambe's original formulation, because we consider nuisance parameters of arbitrary nature and avoid assumptions about differentiability with respect to the nuisance parameter.

The following technical (and trivial) lemma will be the kernel of the proofs that follow. But first it is convenient to introduce the following notation. Given a regular inference function $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$, we define

$$\tilde{\Psi}(x;\theta) = \frac{\Psi(x;\theta,\xi)}{\mathrm{E}_{\theta\xi}\{\Psi'(\theta)\}},$$

which is called the standardized version of Ψ. Here $\Psi'(\theta) = \nabla_\theta\Psi(\theta)$.

Lemma 4.13 *Under the previous regularity conditions, for each regular inference function* Ψ *and* $\Phi : \mathcal{X} \times \Theta \longrightarrow I\!R$ *and each* $\theta \in \Theta$ *and* $\xi \in \mathcal{Z}$*, the following assertions hold:*

(i)

$$\frac{E_{\theta\xi}\{\Psi(\theta)U(\theta;\xi)\}}{E_{\theta\xi}\{\Psi'(\theta)\}} \;=\; -1,$$

where $U(\theta;\xi)$ *is the partial score function at* $(\theta;\xi)$*;*

(ii)

$$E_{\theta\xi}\left\{\tilde{\Phi}^2(\theta)\right\} \;=\; E_{\theta\xi}\left[\{\tilde{\Phi}(\theta) - \tilde{\Psi}(\theta)\}^2\right] \;+\; 2E_{\theta\xi}\left\{\tilde{\Phi}(\theta)\tilde{\Psi}(\theta)\right\} \;-\; E_{\theta\xi}\left\{\tilde{\Psi}^2(\theta)\right\} \;.$$

Proof: Since Ψ is unbiased, one has

$$\int \Psi(x;\xi)p(x;\theta,\xi)d\mu(x) \;=\; 0.$$

Differentiating the expectation above with respect to θ and interchanging the order of differentiation and integration, we obtain

$$\mathrm{E}_{\theta\xi}\left\{\frac{\partial}{\partial\theta}\Psi(\theta)\right\} \;+\; \mathrm{E}_{\theta\xi}\{\Psi(\theta)U(\theta;\xi)\} \;=\; 0$$

which is equivalent to the first part of the lemma. The second part is straightforward. □

The following theorem gives a useful tool for computing optimal inference functions.

Theorem 4.14 *Assume the previous regularity conditions. Consider two functions* $A : \Theta \longrightarrow I\!R\backslash\{0\}$ *and* $R : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow I\!R$*. Suppose that, for each regular inference function* Φ*, one has, for each* $\theta \in \Theta$ *and* $\xi \in \mathcal{Z}$*,*

$$\int R(x;\theta,\xi)\Phi(x;\theta)p(x;\theta,\xi)d\mu(x) = 0 \;.$$

If a regular inference function Ψ *can be written in the form, for all* $\theta \in \Theta$*,*

$$\Psi(x;\theta) \;=\; A(\theta)U(x;\theta,\xi) \;+\; R(x;\theta,\xi), \tag{4.25}$$

for x $[P_{\theta\xi}]$*, (*Ψ *does not depend on* ξ *even though* U *and* R *do), then* Ψ *is optimal. Furthermore, a regular inference function* Φ *is optimal if and only if*

$$\tilde{\Phi}(\theta) \;=\; \tilde{\Psi}(\theta) \;\; [P_{\theta\xi}] \;\forall\theta \in \Theta, \forall\xi \in \mathcal{Z},$$

provided that there exists a decomposition as (4.25) above.

Proof: Take an arbitrary $(\theta, \xi) \in \Theta \times \mathcal{Z}$. Given $\Phi \in \mathcal{G}$ one has

$$\begin{aligned} \mathrm{E}_{\theta\xi}\{\tilde{\Phi}(\theta)\tilde{\Psi}(\theta)\} &= \mathrm{E}_{\theta\xi}\left[\frac{\Phi(\theta)A(\theta)U(\theta,\xi) + \Phi(\theta)R(\cdot;\theta,\xi)}{\mathrm{E}_{\theta\xi}\{\Phi'(\theta)\}\mathrm{E}_{\theta\xi}\{\Psi'(\theta)\}}\right] \\ &= \frac{A(\theta)}{\mathrm{E}_{\theta\xi}\{\Psi'(\theta)\}}\frac{\mathrm{E}_{\theta\xi}\{\Phi(\theta)U(\theta,\xi)\}}{\mathrm{E}_{\theta\xi}\{\Phi'(\theta)\}} \\ &= -\frac{A(\theta)}{\mathrm{E}_{\theta\xi}\{\Psi'(\theta)\}}. \end{aligned} \tag{4.26}$$

Hence the value of $\mathrm{E}_{\theta\xi}\{\tilde{\Phi}(\theta)\tilde{\Psi}(\theta)\}$ does not depend on Φ, in particular,

$$\mathrm{E}_{\theta\xi}\{\tilde{\Phi}(\theta)\tilde{\Psi}(\theta)\} = \mathrm{E}_{\theta\xi}\{\tilde{\Psi}^2(\theta)\} > 0.$$

On the other hand, from (ii) of Lemma 4.13, one has

$$\begin{aligned} \mathrm{E}_{\theta\xi}\{\tilde{\Phi}^2(\theta)\} &= \mathrm{E}_{\theta}[\{\tilde{\Phi}(\theta) - \tilde{\Psi}(\theta)\}^2] + 2\mathrm{E}_{\theta\xi}\{\tilde{\Phi}(\theta)\tilde{\Psi}(\theta)\} - \mathrm{E}_{\theta\xi}\{\tilde{\Psi}^2(\theta)\} \\ &= \mathrm{E}_{\theta}[\{\tilde{\Phi}(\theta) - \tilde{\Psi}(\theta)\}^2] + \mathrm{E}_{\theta\xi}\{\tilde{\Psi}^2(\theta)\} \\ &\geq \mathrm{E}_{\theta\xi}\{\tilde{\Psi}^2(\theta)\}, \end{aligned} \tag{4.27}$$

for each $\Phi \in \mathcal{G}$. Thus, $\forall\theta \in \Theta, \forall\xi \in \mathcal{Z}, \forall\Phi \in \mathcal{G}$,

$$J_\Phi(\theta,\xi) = \frac{1}{\mathrm{E}_{\theta\xi}\{\tilde{\Phi}^2(\theta)\}} \leq \frac{1}{\mathrm{E}_{\theta\xi}\{\tilde{\Psi}^2(\theta)\}} = J_\Psi(\theta,\xi). \tag{4.28}$$

We conclude that Ψ is optimal. For the second part of the theorem, note that one has equality in (4.27), and hence in (4.28), if and only if $\forall\theta \in \Theta, \forall\xi \in \mathcal{Z}$, $\mathrm{E}_{\theta\xi}[\{\tilde{\Phi}(\theta) - \tilde{\Psi}(\theta)\}^2] = 0$. That is, if a regular inference function Φ is optimal then $\tilde{\Phi}(\cdot;\theta) = \tilde{\Psi}(\cdot;\theta)$ $[P_{\theta\xi}]$, $\forall\theta \in \Theta, \forall\xi \in \mathcal{Z}$. □

We now study the situation where we have a likelihood factorization of the following form. Suppose that there exists a statistic $T = t(X)$ such that, for all $\theta \in \Theta$ all $\xi \in \mathcal{Z}$ and all $x \in \mathcal{X}$,

$$p(x;\theta,\xi) = f_t(x;\theta)h\{t(x);\theta,\xi\}. \tag{4.29}$$

Theorem 4.15 *Assume that the previous regularity conditions hold and that there exists a statistic T such that one has the decomposition (4.29). Moreover, suppose that the class $\{P^t_{\theta\xi} : \xi \in \mathcal{Z}\}$, where $P^t_{\theta\xi}$ is the distribution of $T(x)$ under $P_{\theta\xi}$ (*i.e. *$X \sim P_{\theta\xi}$), is complete. Then the regular inference function given by*

$$\Psi(x;\theta) = \frac{\partial}{\partial\theta}\log f_t(x;\theta),\ \forall x \in \mathcal{X}, \forall\theta \in \Theta \tag{4.30}$$

is optimal. Moreover, if Φ is also an optimal inference function then Φ is equivalent to Ψ.

The theorem above gives an alternative justification for the use of conditional inference. As an immediate consequence, we also justify the use of the correct denominator for the estimation of the variance in the normal model and solve the paradox of Neyman-Scott.

Proof: Take $\theta \in \Theta$ and $\xi \in \mathcal{Z}$ fixed. From (4.29),

$$U(x;\theta,\xi) = \frac{\partial}{\partial\theta}\log p(x;\theta,\xi) = \frac{\partial}{\partial\theta}\log f_t(x;\theta) \,+\, \frac{\partial}{\partial\theta}\log h(x;\theta,\xi)\,. \tag{4.31}$$

We apply Theorem 4.14 to prove that ψ is a ("unique") optimal inference function. More precisely, defining $A(\theta) = 1$ and $R(x;\theta,\xi) = -\partial \log h\{t(x);\theta,\xi\}/\partial\theta$, and using (4.31) we can write Ψ in the form

$$\Psi(x;\theta) = \frac{\partial}{\partial\theta}\log f_t(;\theta) = A(\theta)U(x;\theta,\xi) + R(x;\theta,\xi)\,.$$

According to Theorem 4.14, if R is orthogonal to every regular inference function, then Ψ is optimal, moreover Ψ is the unique optimal inference function, apart from equivalent inference functions.

Take an arbitrary regular inference function ϕ. We show that ϕ and R are orthogonal. Note that for each $\xi \in \mathcal{Z}$,

$$0 = \int \phi(x;\theta)p(x;\theta,\xi)d\mu(x) = \int \phi(x;\theta)f_t(x;\theta)h\{t(x);\theta,\xi)d\mu(x)\,.$$

On the other hand $\mathrm{E}_{\theta\xi}(\phi|T) = \int \phi(x;\theta)f_t(x;\theta)d\mu(x)$, which is independent of ξ. We write $\mathrm{E}_\theta(\phi|T)$ for $\mathrm{E}_{\theta\xi}(\phi|T)$, and we have $\mathrm{E}_{\theta\xi}\{E_\theta(\phi|T)\} = 0$. Since T is complete, $E_\theta(\phi|T) = 0$, $P_{\theta\xi}$ almost surely. We have then,

$$\mathrm{E}_{\theta\xi}\{\phi(\theta)R(\theta,\xi)\} = \mathrm{E}_{\theta\xi}\{R(\theta,\xi)E_\theta(\phi|T)\} = 0\,.$$

□

4.4.3 Optimality theory for the general case

We treat next the optimality theory in a context where the parameter of interest is k dimensional and inference is to be done in the presence of an arbitrary nuisance parameter.

We introduce next several submodels of $\mathcal{P}$ required for the development. Given any value $\theta \in \Theta \subseteq I\!R^k$ of the parameter of interest, define the submodel

$$\mathcal{P}_\theta \;=\; \{P_{\theta\xi} : \xi \in \mathcal{Z}\}\,.$$

We interpret $\mathcal{P}_\theta$ as the "nonparametric component" of $\mathcal{P}$ at θ. It is convenient to introduce also the following notation for the class of densities of $\mathcal{P}_\theta$,

$$\mathcal{P}_\theta^* \;=\; \left\{ \frac{dP_{\theta\xi}}{d\mu}(\,\cdot\,) := p(\cdot;\theta,\xi) : \xi \in \mathcal{Z} \right\}.$$

We shall use often $\mathcal{P}_\theta^*$ to characterize $\mathcal{P}_\theta$.

Next we consider "well behaved" submodels of $\mathcal{P}_\theta^*$ which are parametrized by a one-dimensional parameter. Formally, given $(\theta, \xi) \in \Theta \times \mathcal{Z}$, a *differentiable path* (in the direction of the nuisance parameter) at (θ, ξ) is an application from a left neighbourhood of zero, $V \in [0, \infty)$ on $\mathcal{P}_\theta^*$, represented by the generalized sequence $\{p_t\}_{t \in V} \subseteq \mathcal{P}_\theta^*$, such that there exists a function $\nu \in L_0^2(P_{\theta\xi})$ and for each $t \in V$ there is $r_t \in L_0^2(P_{\theta\xi})$ with

$$p_t(\,\cdot\,) = p(\,\cdot\,;\theta,\xi) + p(\,\cdot\,;\theta,\xi)\nu(\,\cdot\,) + p(\,\cdot\,;\theta,\xi)r_t(\,\cdot\,) \tag{4.32}$$

and

$$r_t \overset{L^2(P_{\theta\xi})}{\longrightarrow} 0\,, \quad \text{as } t \downarrow 0\,. \tag{4.33}$$

The $L_0^2(P_{\theta\xi})$ function ν is called the *tangent* of the path $\{p_t\}$. Solving (4.32) for ν one obtains, for each $t \in V$,

$$\nu(\,\cdot\,) = \frac{p_t(\,\cdot\,) - p(\,\cdot\,;\theta,\xi)}{t p(\,\cdot\,;\theta,\xi)} - r_t(\,\cdot\,)\,. \tag{4.34}$$

The expression above leads to the interpretation of the tangent ν as (an L^2 approximation to) the score function in a submodel of $\mathcal{P}_\theta$ given by $\{p_t : t \in V\}$ at $t = 0$. The idea of differentiable paths traces back to the pioneer article of Stein (1956) where the "worst regular" one-dimensional submodel was used for reducing a nonparametric problem to estimation in a regular one-dimensional model. Here we make a different use of such notion.

The class of all tangents of differentiable paths at $(\theta, \xi) \in \Theta \times \mathcal{Z}$ is termed the *nuisance tangent set* and the $L_0^2(P_{\theta\xi})$ closure of its span is referred as the *nuisance tangent space*. More precisely, the nuisance tangent set at (θ, ξ) is given by

$$T_N^0(\theta,\xi) := \left\{ \begin{array}{ll} \nu \in L_0^2(P_{\theta\xi}): & \exists\{p_t\} \subseteq \mathcal{P}_\theta^*,\, \exists\{r_t\} \subseteq L_0^2(P_{\theta\xi}) \\ & \\ & \text{such that (4.32) and (4.33) hold} \end{array} \right\}\,.$$

The nuisance tangent space at (θ, ξ) is

$$T_N(\theta,\xi) := cl_{L^2(P_{\theta\xi})}\left[span\left\{T_n^0(\theta,\xi)\right\}\right]\,.$$

In the literature (see Pfanzagl, 1990; Bickel *et al.* , 1993 and references therein) the notions of path differentiability and tangent space introduced above are known as strong or L^2 path differentiability and tangent spaces. Other useful notions of path differentiability can be obtained by using alternative definitions for the convergence of the sequence $\{r_t\}$ given in (4.33) (for example convergence in the supremum norm or in the L^1 sense).

The following proposition establishes the first connection between the notions introduced above and the theory of (quasi-) inference functions.

Proposition 4.16 *Let $\Psi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow \mathbb{R}^k$ be a regular quasi-inference function with components $\psi_1, \ldots, \psi_k$. For all $\theta \in \Theta$, $\xi \in \mathcal{Z}$ and $i \in \{1, \ldots, k\}$,*

$$\psi_i(\,\cdot\,;\theta,\xi) \in T_N^{\perp}(\theta,\xi)\,.$$

Here and in the rest of this section $T_N^{\perp}(\theta,\xi)$ is the orthogonal complement of $T_N(\theta,\xi)$ in $L_0^2(P_{\theta\xi})$. We denote the inner product and the norm of $L_0^2(P_{\theta\xi})$ by $\langle\,\cdot\,,\,\cdot\,\rangle_{\theta\xi}$ and $\|\,\cdot\,\|_{\theta\xi}$ respectively. **Proof:** Take $\theta \in \Theta$, $\xi \in \mathcal{Z}$ and $i \in \{1, \ldots, k\}$ fixed and an arbitrary $\nu \in T_N^0(\theta,\xi)$. We prove that ν and $\psi_i(\,\cdot\,;\theta,\xi)$ are orthogonal in the sense of $L^2(P_{\theta\xi})$. This implies the proposition, because of the continuity of the inner product. Using (4.34), for each $t \in V$,

$$\begin{aligned}
\langle \nu(\,\cdot\,), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} &= \langle [\{p_t(\,\cdot\,) - p(\,\cdot\,;\theta,\xi)\}/p(\,\cdot\,;\theta,\xi)] - r_t(\,\cdot\,), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\
&= \int_{\mathcal{X}} \psi_i(x;\theta,\xi) p_t(x) d\mu(x) - \int_{\mathcal{X}} \psi_i(x;\theta,\xi) p(x;\theta,\xi) d\mu(x) \\
&\quad -\langle r_t(\,\cdot\,), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\
&= -\langle r_t(\,\cdot\,), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\,.
\end{aligned}$$

Since $r_t \overset{L^2(P_{\theta\xi})}{\longrightarrow} 0$, from the continuity of the inner product, we conclude that

$$\langle \nu(\,\cdot\,), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} = 0\,.$$

□

If the quasi-inference function does not depend on the nuisance parameter (i.e. it corresponds to a genuine inference function), then we can obtain a sharper result.

Proposition 4.17 *Let $\Psi : \mathcal{X} \times \Theta \longrightarrow \mathbb{R}^k$ be a regular inference function with components $\psi_1, \ldots, \psi_k$. For all $\theta \in \Theta$ and $i \in \{1, \ldots, k\}$,*

$$\psi_i(\,\cdot\,;\theta) \in \bigcap_{\xi\in\mathcal{Z}} T_N^{\perp}(\theta,\xi)\ .$$

In fact the proposition above holds for a certain class of quasi-inference function "containing" the regular inference functions (see exercise 4.11) **Proof:** Take $\theta \in \Theta$ and $i \in \{1, \ldots, k\}$ fixed and arbitrary $\xi \in \mathcal{Z}$ and $\nu \in T_N^0(\theta,\xi)$. We prove that ν and $\psi_i(\,\cdot\,;\theta)$ are orthogonal in the sense of $L^2(P_{\theta\xi})$.

Using (4.34), for each $t \in V$,

$$\langle \nu(\,\cdot\,), \psi_i(\,\cdot\,;\theta)\rangle_{\theta\xi} = -\langle r_t(\,\cdot\,), \psi_i(\,\cdot\,;\theta)\rangle_{\theta\xi}$$

Since $r_t \overset{L^2(P_{\theta\xi})}{\longrightarrow} 0$, we conclude that $\langle \nu(\,\cdot\,), \psi_i(\,\cdot\,;\theta)\rangle_{\theta\xi} = 0$. □

We define the *efficient score function*, $U^E : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow I\!R^k$, by the orthogonal projection of the partial score function, U onto $T_N^{\perp}(\theta, \xi)$. More precisely, for each $\theta \in \Theta$ and $\xi \in \mathcal{Z}$, the efficient score function at (θ, ξ) is given by

$$U^E(\,\cdot\,;\theta,\xi) \;=\; \Pi\{U(\,\cdot\,;\theta,\xi)|T_N^{\perp}(\theta,\xi)\},$$

where $\Pi(\cdot|A)$ is the orthogonal projection operator onto A (with respect to the inner product $\langle f\,,\,g\rangle_{\theta\xi}$).

The space spanned by the components $U_1^E, \ldots, U_k^E$ of the efficient score function at $(\theta, \xi) \in \Theta \times \mathcal{Z}$ is denoted by $E(\theta, \xi)$, i.e.

$$E(\theta,\xi) = span\{U_i^E(\,\cdot\,;\theta,\xi) : i = 1,\ldots,k\}\,.$$

Note that $E(\theta, \xi)$ is a closed (since it is finite-dimensional vector space) subspace of $L_0^2(P_{\theta\xi})$. Hence given any regular quasi-inference function $\Psi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow I\!R^k$ with components $\psi_1, \ldots, \psi_k$ we have, for all $\theta \in \Theta$, $\xi \in \mathcal{Z}$ and $i \in \{1, \ldots, k\}$ the decomposition

$$\psi_i(\,\cdot\,;\theta,\xi) = \psi_i^A(\,\cdot\,;\theta,\xi) \;+\; \psi_i^E(\,\cdot\,;\theta,\xi)\;, \tag{4.35}$$

where $\psi_i^E(\,\cdot\,;\theta,\xi) \in E(\theta,\xi)$ and $\psi_i^A(\,\cdot\,;\theta,\xi) \in A(\theta,\xi) := E^{\perp}(\theta,\xi)$. Here $A(\theta, \xi)$ is the orthogonal complement of $E(\theta, \xi)$ in $L_0^2(P_{\theta\xi})$. The decomposition above induces the following decomposition of each regular quasi-inference function

$$\Psi(\,\cdot\,;\theta,\xi) = \Psi^A(\,\cdot\,;\theta,\xi) \;+\; \Psi^E(\,\cdot\,;\theta,\xi)\;, \tag{4.36}$$

where the components $\psi_i^A(\,\cdot\,;\theta,\xi), \ldots, \psi_i^A(\,\cdot\,;\theta,\xi)$ of Ψ^A at (θ, ξ) are in $A(\theta, \xi)$ and the components $\psi_i^E(\,\cdot\,;\theta,\xi), \ldots, \psi_i^E(\,\cdot\,;\theta,\xi)$ of Ψ^E at (θ, ξ) are in $E(\theta, \xi)$.

We show next that taking the "component" Ψ^E of a regular (quasi-) inference function improves the Godambe information. However, at this stage a technical difficulty appears, the function Ψ^E is not necessarily a regular quasi-inference function, and hence does not necessarily possesses a well-defined Godambe information. For this reason we introduce next an extension of the notion of sensitivity, and consequently of Godambe information, which will make us able to speak of Godambe information of some non-regular inference functions. To motivate our extended notion of sensitivity, consider a regular inference function $\Psi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow I\!R^k$. We characterize the sensitivity of Ψ in an alternative form that will suggest the extension one should define. For each $(\theta, \xi) \in \Theta \times \mathcal{Z}$ and each $i, j \in \{1, \ldots, k\}$ we have

$$\begin{aligned}
0 \;&=\; \frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} \psi_j(x;\theta,\xi)p(x;\theta,\xi)d\mu(x) \qquad (4.37)\\
&\quad\;\text{(differentiating under the integral sign)}\\
&=\; \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left\{\psi_j(x;\theta,\xi)p(x;\theta,\xi)\right\}d\mu(x)\\
&=\; \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left\{\psi_j(x;\theta,\xi)\right\}p(x;\theta,\xi)d\mu(x) + \int_{\mathcal{X}} \psi_j(x;\theta,\xi)\frac{\partial}{\partial\theta_i}\left\{p(x;\theta,\xi)\right\}d\mu(x)\,.
\end{aligned}$$

Hence

$$
\begin{aligned}
&\int_{\mathcal{X}} \frac{\partial}{\partial \theta_i} \{\psi_j(x;\theta,\xi) p(x;\theta,\xi)\}\, d\mu(x) \\
&= -\int_{\mathcal{X}} \psi_j(x;\theta,\xi) \frac{\partial}{\partial \theta_i} \{p(x;\theta,\xi)\}\, d\mu(x) \\
&= \int_{\mathcal{X}} \psi_j(x;\theta,\xi) U_i(x;\theta,\xi) p(x;\theta,\xi) d\mu(x) = -\langle \psi_j(\,\cdot\,;\theta,\xi), U_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\
&\text{(decomposing } U_i = U_i^A + U_i^E \text{ with } U_i^E \in T_n^\perp \text{ and } U_i^A \in T_N) \\
&= -\langle \psi_j(\,\cdot\,;\theta,\xi), U_i^E(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} - \langle \psi_j(\,\cdot\,;\theta,\xi), U_i^A(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\
&\text{(Since } \psi_j \in T_n^\perp \text{ and } U_i^A \in T_N) \\
&= -\langle \psi_j(\,\cdot\,;\theta,\xi), U_i^E(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\
&\text{(decomposing } \psi_j = \psi_i^A + \psi_i^E \text{ and using the orthogonality of } U_i^E \text{ and } \psi_i^A) \\
&= -\langle \psi_j^E(\,\cdot\,;\theta,\xi), U_i^E(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\,.
\end{aligned}
$$

We conclude that the sensitivity of Ψ at (θ,ξ) is given by

$$
S_\Psi(\theta,\xi) = \left[-\langle \psi_j^E(\,\cdot\,;\theta,\xi), U_i^E(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\right]_{i=1,\ldots,k}^{j=1,\ldots,k}\,. \tag{4.38}
$$

Here $[a_{ij}]_{i=1,\ldots,k}^{j=1,\ldots,k}$ denotes the matrix formed by a_{ij}'s with i indexing the columns and j indexing the lines.

We define the *extended sensitivity* (or simply the *sensitivity*) of Ψ by the matrix in the right-hand side of (4.38). The (extended) Godambe information is defined in the same way we did before but using the extended sensitivity instead of the sensitivity. Note that both, the standard and the extended, versions of the sensitivity (and the Godambe information) coincide in the case where Ψ is regular. Moreover, the extended sensitivity is defined for each quasi-inference function whose components are in L_0^2, not only for regular inference functions. According to the new definition both Ψ and Ψ^E posses the same sensitivity.

Proposition 4.18 *Given a regular inference function Ψ, for all $\theta \in \Theta$ and all $\xi \in \mathcal{Z}$,*

$$
J_\Psi(\theta,\xi) \leq J_{\Psi^E}(\theta,\xi)\,.
$$

Proof: For each $\theta \in \Theta$ and $\xi \in \mathcal{Z}$,

$$
\begin{aligned}
J_\Psi^{-1}(\theta,\xi) &= S_\Psi^{-1}(\theta,\xi) V_\Psi(\theta,\xi) S_\Psi^{-T}(\theta,\xi) \\
&= S_{\Psi^E}^{-1}(\theta,\xi)\{V_{\Psi^E}(\theta,\xi) + V_{\Psi^A}(\theta,\xi)\} S_{\Psi^E}^{-T}(\theta,\xi) \\
&= S_{\Psi^E}^{-1}(\theta,\xi) V_{\Psi^E}(\theta,\xi) S_{\Psi^E}^{-T}(\theta,\xi) + S_{\Psi^E}^{-1}(\theta,\xi) V_{\Psi^A}(\theta,\xi) S_{\Psi^E}^{-T}(\theta,\xi) \\
&\geq S_{\Psi^E}^{-1}(\theta,\xi) V_{\Psi^E}(\theta,\xi) S_{\Psi^E}^{-T}(\theta,\xi) = J_{\Psi^E}(\theta,\xi)\,.
\end{aligned}
$$

□

The following proposition gives further properties of regular inference functions, which will allow us to establish an upper bound for the Godambe information.

Proposition 4.19 *Given a regular inference function Ψ, for all $\theta \in \Theta$ and all $\xi \in \mathcal{Z}$, we have:*

(i) $\Psi^E \sim U^E$;

(ii) $span\{\Psi_i^E(\,\cdot\,;\theta,\xi) : i = 1,\ldots,k\} = E(\theta,\xi)$;

(iii) $J_{\Psi^E}(\theta,\xi) = J_{U^E}(\theta,\xi)$.

Proof: Take $\theta \in \Theta$ and $\xi \in \mathcal{Z}$ fixed.
(i) Assume without loss of generality that the components of the efficient score function $U_1^E(\,\cdot\,;\theta,\xi),\ldots,U_k^E(\,\cdot\,;\theta,\xi)$ are orthonormal in $L_0^2(P_{\theta\xi})$. For each $i \in \{1,\ldots,k\}$, expanding $\psi_i(\,\cdot\,;\theta,\xi)$ in a Fourier series with respect to a basis whose first k elements are $U_1^E(\,\cdot\,;\theta,\xi),\ldots,U_k^E(\,\cdot\,;\theta,\xi)$ one obtains

$$\begin{aligned}\psi_i(\,\cdot\,;\theta,\xi) &= \langle U_1^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\, U_1^E(\,\cdot\,;\theta,\xi) \\ &\quad + \cdots + \langle U_k^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\, U_k^E(\,\cdot\,;\theta,\xi) + \psi_i^A(\,\cdot\,;\theta,\xi)\,.\end{aligned}$$

That is,

$$\begin{aligned}\psi_i^E(\,\cdot\,;\theta,\xi) &= \langle U_1^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\, U_1^E(\,\cdot\,;\theta,\xi) \\ &\quad + \cdots + \langle U_k^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi}\, U_k^E(\,\cdot\,;\theta,\xi)\,.\end{aligned} \tag{4.39}$$

Moreover, for $j = 1,\ldots,k$

$$\begin{aligned}\langle U_j^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} &= \langle U_j^E(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} + \langle U_j^A(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\ &= \langle U_j(\,\cdot\,;\theta,\xi), \psi_i(\,\cdot\,;\theta,\xi)\rangle_{\theta\xi} \\ &= \int_{\mathcal{X}} \psi_i(x;\theta,\xi) U_j(x;\theta,\xi) p(x;\theta,\xi) d\mu(x) \\ &= -\int_{\mathcal{X}} \left\{\frac{\partial}{\partial\theta_j}\psi_i(x;\theta,\xi)\right\} p(x;\theta,\xi) d\mu(x)\,.\end{aligned} \tag{4.40}$$

The last equality above comes from the following

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta_i}\int_{\mathcal{X}} \psi_j(x;\theta,\xi) p(x;\theta,\xi) d\mu(x) \\ &\quad \text{(differentiating under the integral sign)} \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left\{\psi_j(x;\theta,\xi) p(x;\theta,\xi)\right\} d\mu(x) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left\{\psi_j(x;\theta,\xi)\right\} p(x;\theta,\xi) d\mu(x) + \int_{\mathcal{X}} \psi_j(x;\theta,\xi)\frac{\partial}{\partial\theta_i}\left\{p(x;\theta,\xi)\right\} d\mu(x) \\ &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta_i}\left\{\psi_j(x;\theta,\xi)\right\} p(x;\theta,\xi) d\mu(x) + \int_{\mathcal{X}} \psi_j(x;\theta,\xi) U_i(x;\theta,\xi) p(x;\theta,\xi) d\mu(x)\,. \end{aligned}$$

We conclude from (4.39) and (4.40) that $\Psi^E(\,\cdot\,;\theta,\xi) = -S_\Psi(\theta,\xi)U^E(\,\cdot\,;\theta,\xi)$, which means that Ψ^E and U^E are equivalent.
(ii) From the previous discussion $span\{\Psi_i^E(\,\cdot\,;\theta,\xi) : i = 1,\ldots,k\}$ is the space spanned by $-S_\Psi(\theta,\xi)U^E(\,\cdot\,;\theta,\xi)$ which is the span of $\{U_i^E(\,\cdot\,;\theta,\xi) : i = 1,\ldots,k\}$, since the sensitivity by assumption is of full rank.
(iii) Straightforward.

□

A consequence of the two last proposition is that J_{U^E} is an upper bound for the Godambe information of regular quasi inference functions. This upper bound is attained by any (if any exists) extended regular inference functions with components in E. In particular if U^E is a regular (quasi-) inference function, then it is an optimal (quasi-) inference function. We consider next some examples.

Example 4.20 *(Semiparametric location model)* Consider the following semiparametric extension of the location model.

$$\mathcal{P} = \left\{ P_{\theta\xi} : \frac{P_{\theta\xi}}{d\mu}(\,\cdot\,) = \xi(\,\cdot\, - \theta),\ \theta \in \Theta = I\!R\ ,\ \xi \in \mathcal{Z} \right\}\ . \tag{4.41}$$

Here $\mathcal{Z}$ is the class of probability densities $\xi : I\!R \longrightarrow I\!R$ such that (4.42)-(4.47) given below hold.

$$\forall x \in I\!R,\ \xi(x) > 0; \tag{4.42}$$

$$\int_{I\!R} \xi(x)d\mu(x) = 1; \tag{4.43}$$

$$\int_{I\!R} x\xi(x)d\mu(x) = 0; \tag{4.44}$$

$$\int_{I\!R} x^2\xi(x)d\mu(x) < \infty; \tag{4.45}$$

$$\xi \text{ is differentiable almost everywhere;} \tag{4.46}$$

$$\int_{I\!R} \frac{\{\xi'(x)\}^2}{\xi(x)}d\mu(x) < \infty\,. \tag{4.47}$$

Conditions (4.42) and (4.43) ensure that ξ is a density of a probability measure with support equal to the whole real line. From condition (4.44) the parametrization $(\theta,\xi) \mapsto P_{\theta\xi}$ is identifiable. Conditions (4.46) and (4.47) ensure that the score function

$$U(\,\cdot\,;\theta) = \frac{\xi'(\,\cdot\, - \theta)}{\xi(\,\cdot\, - \theta)}$$

is well-defined and in $L^2(P_{\theta\xi})$.

We calculate next the nuisance tangent space at $\theta = 0$ and $\xi \in \mathcal{Z}$. It is easy to see that the nuisance tangent space at an arbitrary point of the parameter space is

$$T_N(\theta, \xi) = \{\nu(\cdot - \theta) : \nu \in T_N(0, \xi)\}\,. \tag{4.48}$$

We prove next that $T_N(0,\xi) = [span\{\nu(\cdot) = (\cdot)\}]^\perp$. Consider the path, given for $t \in V$ by

$$p_t(\cdot) = \xi(\cdot) + t(\cdot)\nu(\cdot)\,.$$

Here $\nu \in C_k^\infty \cap [span\{(\cdot)\}]^\perp$ and C_k^∞ is the space of smooth (C^∞) functions from $\mathbb{R}$ to $\mathbb{R}$ that are compact supported. It is easy to verify that for t small enough $p_t \in \mathcal{P}_0^*$, hence $\{p_t\}$ is a differentiable path (at $(0,\xi)$) with tangent ν. That is, $\nu \in T_N^0(0,\xi)$. We conclude that $C_k^\infty \cap [span\{(\cdot)\}]^\perp \subseteq T_N^0(0,\xi)$. Furthermore, it is well-known that C_k^∞ is dense in $L^2(P_{\theta\xi})$, hence $C_k^\infty \cap [span\{(\cdot)\}]^\perp$ is dense in $[span\{(\cdot)\}]^\perp \subset L^2(P_{\theta\xi})$. We conclude that $[span\{(\cdot)\}]^\perp \subseteq T_N(0,\xi)$.

Let us prove now that $T_N(0,\xi) \subseteq [span\{(\cdot)\}]^\perp$. Take $\nu \in T_N^0(0,\xi)$, we prove that $(\cdot)$ and ν are orthogonal, which implies what we want to prove here.

$$\begin{aligned}
\langle \nu(\cdot\,,(\cdot)\rangle_{0\xi} &= \langle \frac{p_t(\cdot) - \xi(\cdot)}{\xi(\cdot)} - r_t(\cdot), \nu(\cdot)\rangle_{0\xi} \\
&= \frac{1}{t}\left\{\int_{\mathbb{R}} \nu(x)p_t(x)d\mu(x) - \int_{\mathbb{R}} \nu(x)\xi(x)d\mu(x)\right\}\langle -r_t(\cdot), \nu(\cdot)\rangle_{0\xi} \\
&= = \langle -r_t(\cdot), \nu(\cdot)\rangle_{0\xi}\ .
\end{aligned}$$

Since r_t converges to zero in $L^2(P_{0\xi})$, we conclude that $\langle \nu(\cdot\,,(\cdot)\rangle_{0\xi} = 0$. We have then proved that the nuisance tangent space is given by

$$T_N(\theta,\xi) = [span\{\nu(\cdot) = (\cdot - \theta)\}]^\perp$$

and hence

$$T_N^\perp(\theta,\xi) = span\{\nu(\cdot) = (\cdot - \theta)\}\,.$$

We calculate next the efficient score function. To project $U(\cdot\,;\theta,\xi)$ onto $T_N^\perp(\theta,\xi)$ we take an orthonormal basis $\{e_i : i = 1, 2, \ldots\}$ in $L_0^2(P_{\theta\xi})$ with the first element being the function $e_1 = (\cdot - \theta)$. Expanding $U(\cdot\,;\theta,\xi)$ in terms of that basis we obtain

$$U(\cdot\,;\theta,\xi) = \langle U(\cdot\,;\theta,\xi), (\cdot - \theta)\rangle_{\theta\xi}\ (\cdot - \theta) + \sum_{i=2}^{\infty} \langle U(\cdot\,;\theta,\xi), e_i(\cdot - \theta)\rangle_{\theta\xi}\ e_i(\cdot - \theta)\,.$$

Hence the efficient score function is given by

$$U^E(\cdot\,;\theta,\xi) = \langle U(\cdot\,;\theta,\xi), (\cdot - \theta)\rangle_{\theta\xi}\ (\cdot - \theta)\,.$$

The Fourier coefficient $\langle U(\,\cdot\,;\theta,\xi),(\,\cdot\,-\theta)\rangle_{\theta\xi}$ can be calculated explicitly if we insert the assumption that the density ξ is continuous and $\lim_{x\to\pm\infty} x\xi(x) = 0$ (integrate by parts). In that case the Fourier coefficient is -1. It can also be shown that inserting this assumption in the model, the nuisance tangent space remains unchanged. Clearly, the efficient score function does not depend on the nuisance parameter and hence, from the previous discussion it is optimal. It is easy to see that under a repeated sample schema the root of the efficient is the sample mean.

Example 4.21 *(Models without nuisance parameter)* It is easy to see that in a model without nuisance parameter the "nuisance tangent space" is the space $\{0\}$ and hence its orthogonal complement in L_0^2 is L_0^2 itself. If the score function is in L_0^2 then it is the efficient score function. Moreover, the efficient score function does not depend on the nuisance parameter and hence is optimal. That is the score function is optimal, which is in accordance with the previous results.

4.5 Problems

Basic setup

Problem 4.1 Show that the relation "$\sim$ " is an equivalence relation, i.e., given any inference function ψ, ϕ and ζ, then

(i) $\psi \sim \psi$;

(ii) $\psi \sim \phi \Rightarrow \phi \sim \psi$;

(iii) $\psi \sim \phi$ and $\phi \sim \zeta \Rightarrow \psi \sim \zeta$.

Problem 4.2 Consider a regular inference function ψ for a one-dimensional parameter θ.

(i) Let $C(\theta)$ be a constant function, and define

$$\phi(\theta) = C(\theta)\psi(\theta).$$

Show that $C(\theta)$ may be chosen in such a way that for all θ

$$V_\phi(\theta) = -S_\phi(\theta).$$

(ii) Show that $C(\theta) = 1$ if ψ is the score function.

(iii) Show that, with C chosen as in (i),

$$J_\psi(\theta) = -S_\phi(\theta) = V_\phi(\theta),$$

Express this formula in terms of the derivatives of the likelihood function in the case where ψ is the score function.

(iv) Define a quasi-likelihood function by

$$L(\theta) = \int_{\theta_0}^{\theta} \phi(\theta)\, d\theta.$$

Show that the stationary points of L are the same as the solutions to $\psi(\theta) = 0$. Why might one prefer the stationary points that correspond to local maxima of L? Show that L is a version of the log likelihood function in the case where ψ is the score function.

(v) Comment on the possibilities of generalizing these results to dimension $k > 1$, in particular the difficulty of generalizing the definition of L.

Problem 4.3 Let $\psi : \mathcal{X} \times \Theta \longrightarrow I\!R^k$ be an inference function based on one observation. For each $n \in I\!N$ define ψ_n using the expression (4.2).

(i) Note that ψ and ψ_n are defined in different sample spaces. Write the definition of the function ψ_n given the domain and image and state precisely the measurability basic condition to ψ_n.

(ii) Show that ψ is unbiased if and only if ψ_n is unbiased.

Problem 4.4 (Location model) Consider the following parametric family:

$$\mathcal{P} = \{P_\mu \ll \nu : \frac{dP_\mu}{d\nu}(\cdot) = f(\cdot - \mu), \mu \in I\!R\},$$

where ν is the Lebesgue measure on the real line, and $f : I\!R \longrightarrow I\!R$ is such that

(a) $f \in C^2(I\!R)$;

(b) $f(x) > 0, \forall x \in I\!R$;

(c) $\int f(x) d\nu(x) = 1$;

(d) $\int x f(x) d\nu(x) = 0$;

(e) f' is integrable.

Clearly f is a probability density of a distribution with support equal to the whole real line.

(i) Compute the score function for μ and show that $\int f'(x) d\nu(x) = 0$ is a necessary and sufficient condition to the unbiasedness of the score function;

(ii) Show that if $f(x) \to 0$ as $x \to \pm\infty$ then the score function is unbiased (Hint: use the fundamental theorem of calculus.)

(iii) Compute the function ν_μ defined by (4.4) and show that ν_μ satisfies the conditions of Theorem 4.1;

(iv) Show that the same conclusion holds for f satisfying (b)–(e) and the following condition:

$$\text{(a')}\ f(x) \geq 0 \text{ for } x, [\nu];$$

(v) Show that if f is the density of a centred exponential distribution (i.e. shifted to have mean zero mean as required by (d)) then the score function is not an unbiased inference function;

(vi) Show that condition (d) ensures the identifiability of the model.

Problem 4.5 (The Huber estimator of location) Consider the location model defined in Problem 4.4. Let K be a fixed positive real number. Define the inference function $\phi_K : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ given by

$$\phi_K(x;\mu) = \begin{cases} x-\mu & \text{if } |x-\mu| < K \\ \{sgn(x-\mu)\}K & \text{if } |x-\mu| \geq K \end{cases}$$

The root of this inference function is called the Huber estimator of location.

(i) Show that, under a sequence of independent and identically distributed observations, for K large enough the sample mean is a root of ϕ_K, except for a set of Lebesgue measure zero;

(ii) Show that, under a sequence of independent and identically distributed observations, for K sufficiently close to zero, the sample median is a root of ϕ_K, ν-almost everywhere;

(iii) Compute the function λ_μ for the inference function ϕ_K and show that λ_μ is a continuous function;

(iv) Show that if f is a symmetric function (i.e. $f(x) = f(-x), \forall x \in \mathbb{R}$) then for each $\mu \in \mathbb{R}$, λ_μ satisfies the conditions of Theorem 4.1, and hence has consistent roots;

(v) Consider the Gumbel distribution with density given, for all $x \in \mathbb{R}$, by

$$f(x) = e^{-x}e^{-e^{-x}}$$

Show that λ_0 has a root different from 0 and hence ϕ_K is not unbiased. Show that the Huber estimator is not consistent in this case.

Problem 4.6 (Scale model) Consider the parametric family

$$\mathcal{P} = \left\{ P_\sigma \ll \lambda : \frac{dP_\sigma}{d\lambda}(\cdot) = \sigma^{-1} f(\cdot/\sigma), \sigma \in I\!R_+ \right\},$$

where λ is the Lebesgue measure and f is as in Problem 4.4 with condition (d) replaced by

$$(d') \quad \int x^2 f(x) d\lambda(x) = 1.$$

(i) Compute the score function for σ and find conditions under which this score function is unbiased;

(ii) Study the function λ_σ;

(iii) Show that if f is the density of the standardized exponential distribution, then the score function is unbiased;

(iv) What can one conclude about the maximum likelihood estimator in this case?

Problem 4.7 (Ratio of means) Suppose that $Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots, Y_{2n}$ are independent random variables such that for $i = 1, \ldots, n$ one has

$$E(Y_{1i}) = \theta E(Y_{2i}).$$

In other words, one has independent pairs of observations such that the rate of the means in each pair is constant. We assume, of course, that the expectations are all finite and that $\theta \neq 0$. If we do not know anything else about the random variables above, we may use the inference function

$$\psi_n(\tilde{y}, \theta) = \sum_{i=1}^{n} (y_{1i} - \theta y_{2i}),$$

where $\tilde{y} = (y_{11}, \ldots, y_{1n}, y_{21}, \ldots, y_{2n})^\top$ are the observations. Show that

$$\hat{\theta} = \hat{\theta}(\tilde{y}) = \sum y_{1i} / \sum y_{2i}$$

is a root of ψ_n. More generally, we may consider the inference function

$$\psi_n^w(\tilde{y}, \theta) = \sum_{i=1}^{n} w_i (y_{1i} - \theta y_{2i}),$$

where $w_1, \ldots, w_n$ is a set of weights. Give conditions under which ψ_n^w is unbiased and the functions λ_θ satisfy the conditions of Theorem 4.1. Discuss the consistency of the root of ψ_n^w.

Optimality theory

Problem 4.8 Consider the one-parameter context (i.e. $k = 1$).

(i) Show that the sensitivity function is linear, i.e., for each $K \neq 0$ and each regular inference function ψ, $S_{K\psi}(\theta) = KS_{\psi}(\theta)$;

(ii) Show that in each equivalence class defined by the relation $\sim$, there is only one inference function with sensitivity equal to 1;

(iii) Show that the Godambe information defines a partial ordering in the class of regular inference functions;

(iv) What can you conclude from (iii) regarding the existence of an optimal inference function (Hint: Consider Zorn's Lemma);

(v) Extend items (i)–(iv) to the k-dimensional ($k \geq 1$) case.

Problem 4.9 This exercise aims to improve Theorem 4.3 by replacing the assumption of boundedness of ψ'' given in (4.7) by the assumption that $\int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x) \in I\!R$. Consider then the set-up and notation introduced in Theorem 4.3 and in its proof. The proof sketched will be the same as the proof of Theorem 4.3 except from the argument given in (4.10), where the assumption (4.7) is used.

(i) We start with the following lemma of convergence in probability. Prove that a sequence of random variables (defined in a common probability space) $\{X_n\}$ converges in probability to the random variable X as $n \to \infty$ if and only if each subsequence of $\{X_n\}$ possesses a subsubsequence that converges almost surely to X. Hint: $\Rightarrow$ is a well-known result (see any reasonable book of probability). For proving $\Leftarrow$ use contradiction ($A \to B$ if and only if not $B \Rightarrow$ not A);

(ii) Show that $\theta_n^* \xrightarrow{P} \theta$. Hint: Use the facts that $\hat{\theta}_n \xrightarrow{P} \theta$ and $|\theta_n^* - \theta| \leq |\hat{\theta}_n - \theta|$. Show that $\frac{1}{n}\Psi_n''(\mathbf{x};\theta) \xrightarrow{P} \int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)\, d\mu(x)$;

(iv) Given a subsequence of $\{\hat{\theta}_n\}$ show that there exist a subsubsequence of that subsequence converging almost surely to θ. Analogously, given a subsequence of $\{\frac{1}{n}\Psi_n(\mathbf{x};\theta)\}$ show that there exist a subsubsequence of that subsequence converging almost surely to $\int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)$;

(v) Given a subsequence $\{n(m) : m \in N\} \subseteq \{n : n \in N\}$ show that there is a subsubsequence, say $\{n(m(i)) : i \in N\} \subseteq \{n(m) : m \in N\} \subseteq \{n : n \in N\}$ such that $\frac{1}{n}(m(i))\Psi_{n(m(i))}(\mathbf{x};\theta)$ converges almost surely to $\int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)$ and $\hat{\theta}_{n(m(i))}$ converges almost surely to θ;

(vi) Given a subsequence of $\{\frac{1}{n}\Psi_n(\mathbf{x};\theta_n^*)\}$ show that there exist a subsubsequence of that sequence converging to $\int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)$. Hint: This is the delicate step of the proof. Given a subsequence $\{n(m) : m \in N\} \subseteq \{n : n \in N\}$ take subsubsequence $\{n(m(i)) : i \in N\} \subseteq \{n(m) : m \in N\} \subseteq \{n : n \in N\}$ defined in the previous item. Consider the set $A = \{\omega : \theta^*_{n(m(i))}(\omega) \to \theta$ and $\frac{1}{n(m(i))}\Psi_{n(m(i))}(\mathbf{x};\theta)(\omega) \to \int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)\}$ (here ω is an element of the sample space on which the random variables are defined and θ_n^* and $\Psi_n(\mathbf{x};\theta)$ are viewed as random variables on that space). Define $A^* = \{\omega : \frac{1}{n(m(i))}\Psi_{n(m(i))}(\mathbf{x};\theta^*_{n(m(i))}(\omega))(\omega) \to \int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)\}$. Note that by construction $P_\theta(A) = 1$. Show that $A \subseteq A^*$. To prove the last statement take an arbitrary element $\omega \in A$ and use the continuity of $\Psi(\mathbf{x}(\omega);\,\cdot\,)$;

(vii) Using the first item and the previous item conclude that

$$\frac{1}{n}\Psi_n(\theta_n^*) \xrightarrow{P} \int_{\mathcal{X}} \psi''(x;\theta)p(x;\theta)d\mu(x)\,;$$

(viii) Conclude the proof of the theorem.

Problem 4.10 The goal of this exercise is to obtain a multidimensional versions of the theorem on asymptotic normality of consistent sequences of roots of regular inference functions.

(i) Write a multidimensional version of Problem 4.9;

(ii) Write a multidimensional version of Theorem 4.3;

(iii) Using Theorem 4.3 and Problem 4.9 prove the theorems given in the previous items. Hint: Consider the Cramér-Wold Theorem to reduce the problem to an arbitrary linear combination;

Problem 4.11 *(i)* A regular quasi-inference function $\Psi : \mathcal{X} \times \Theta \times \mathcal{Z} \longrightarrow \mathbb{R}^k$ is said to be nuisance ancillary if for each $\theta \in \Theta$ and $\xi, \xi_* \in \mathcal{Z}$, $\mathrm{E}_{\theta\xi}\{\Psi(X;\theta,\xi_*) = \mathbf{0}$. Write and prove a proposition analogue to proposition 4.17 using nuisance ancillary quasi-inference functions instead of inference functions.

(ii) A quasi-inference function $\Psi : \mathcal{X} \times \Theta \times \mathcal{Z} \to \mathbb{R}$ is said to be E-ancillary with respect to the nuisance parameter if there is a generalized sequence $\{\Psi_\alpha : \mathcal{X} \times \Theta \times \mathcal{Z} \to \mathbb{R}\}_{\alpha\in A}$ such that for all $\theta \in \Theta$ and $\xi_1, \xi_2 \in \mathcal{Z}$,

$$\Psi_\alpha(\cdot;\theta,\xi_1) \overset{L^2(P_{\theta\xi})}{\longrightarrow} \Psi(\cdot;\theta,\xi_1)$$

and

$$\mathrm{E}_{\theta\xi_1}\{\Psi_\alpha(\cdot;\theta,\xi_2)\} = 0$$

(see Small and McLeish, 1988a). Write and prove a proposition analogue to proposition 4.17 using E-ancillary quasi-inference functions instead of inference functions.

Appendix A

MOMENT GENERATING FUNCTIONS

The moment generating function is an indispensable tool for handling exponential families. The present appendix gives a brief but self-contained introduction to the requisite theory for moment generating functions, characteristic functions and Fourier-Laplace transforms for multivariate distributions. Although standard, this material is not easily available in the literature. Both Lukacs (1970), who treats characteristic functions, and Kawata (1972), who treats Fourier transforms, concentrate on the univariate case.

A.1 Definition and properties

Let $\mathcal{M}_k$ denote the set of all probability distributions on $\mathbb{R}^k$. For $P \in \mathcal{M}_k$, we define the *moment generating function* of P by

$$M_P(s) = \int e^{s \cdot x} P(dx), \qquad s \in \mathbb{R}^k,$$

and we define the *cumulant generating function* of P by

$$\mathcal{K}_P(s) = \log M_P(s), \qquad s \in \mathbb{R}^k.$$

The *effective domain* of M_P, respectively $\mathcal{K}_P$, is defined by

$$\Theta_P = \{s \in \mathbb{R}^k : \ M_P(s) < \infty\}.$$

If X is a random vector with distribution P we write M_X instead of M_P, and similarly $\mathcal{K}_X$ and Θ_X instead of $\mathcal{K}_P$ and Θ_P, respectively. When no confusion arises we sometimes omit the subscript P, respectively X, etc.

Theorem A.1 *Assume that $X \sim P$, where $P \in \mathcal{M}_k$. Then*

(i) $0 < M_P(s) \leq \infty$ *for* $s \in \mathbb{R}^k$.

(ii) $M_P(0) = 1$ *and* $\mathcal{K}_P(0) = 0$.

(iii) If B *is an* $\ell \times k$ *matrix and* c *an* $\ell \times 1$ *vector then*

$$\begin{aligned} \Theta_{BX+c} &= \{s \in \mathbb{R}^\ell \colon B^T s \in \Theta_X\} \\ M_{BX+c}(s) &= M_X(B^T s) e^{s \cdot c}, \quad s \in \mathbb{R}^\ell \ , \end{aligned}$$

and

$$\mathcal{K}_{BX+c}(s) = \mathcal{K}_X(B^T s) + s \cdot c, \quad s \in \mathbb{R}^\ell.$$

(iv) Let $X = (X_1^T, X_2^T)^T$, *where* X_1 *is* d*-dimensional and* X_2 *is* $(k-d)$*-dimensional, and let* $s = (s_1^T, s_2^T)^T$ *be the corresponding partition of* s. *If* X_1 *and* X_2 *are independent, then*

$$M_X(s) = M_{X_1}(s_1) M_{X_2}(s_2), \quad s \in \mathbb{R}^k$$

and

$$\mathcal{K}_X(s) = \mathcal{K}_{X_1}(s_1) + \mathcal{K}_{X_2}(s_2), \quad s \in \mathbb{R}^k$$

Proof: See Problem A.1. □

Example A.2 Let us derive the moment generating function of the multivariate normal distribution. Consider first the case where $X \sim N_k(0, I_k)$. Then for $s \in \mathbb{R}^k$,

$$\begin{aligned} M_X(s) &= (2\pi)^{-k/2} \int \exp(-\frac{1}{2} x \cdot x + s \cdot x) dx \\ &= (2\pi)^{-k/2} \exp(\frac{1}{2} s \cdot s) \int \exp\{-\frac{1}{2}(x \cdot x - 2s \cdot x + s \cdot s)\} dx \\ &= \exp(\frac{1}{2} s \cdot s) \int (2\pi)^{-k/2} \exp\{-\frac{1}{2}(x - s) \cdot (x - s)\} dx \\ &= \exp(\frac{1}{2} s \cdot s). \end{aligned}$$

If B is an $\ell \times k$ matrix an μ an $\ell \times 1$ vector then $BX + \mu \sim N_\ell(\mu, \sum)$, where $\sum = BB^T$. By Theorem A.1 (iii) we have

$$\begin{aligned} M_{BX+\mu}(s) &= \exp\left\{\frac{1}{2}(B^T s) \cdot B^T s + s \cdot \mu\right\} \\ &= \exp\left\{\frac{1}{2} s^T \Sigma s + s \cdot \mu\right\}, \quad s \in \mathbb{R}^\ell, \end{aligned}$$

which is hence the moment generating function of the ℓ-variate normal distribution $N_\ell(\mu, \sum)$. Since the matrix B was arbitrary, the result holds even if $\sum$ is singular.

The next theorem concerns convexity properties of M_P and $\mathcal{K}_P$. But first we introduce *Hölder's inequality*.

Proposition A.3 *Let $P \in \mathcal{M}_k$. Then for any $s_1, s_2 \in I\!R^k$ and $0 \le \alpha \le 1$ we have*

$$\int \exp\{\alpha s_1 \cdot x \quad + \quad (1-\alpha) s_2 \cdot x\} P(dx) \le \left\{ \int e^{s_1 \cdot x} P(dx) \right\}^{\alpha} \left\{ \int e^{s_2 \cdot x} P(dx) \right\}^{1-\alpha} . \quad \text{(A.1)}$$

If $0 < \alpha < 1$ and $s_1 \neq s_2$ the inequality is strict if and only if P is not concentrated on an affine subspace of $I\!R^k$.

Proof: The logarithm is a concave function, and hence for any $a_1 > 0$, $a_2 > 0$ and $0 \le \alpha \le 1$ we have

$$\alpha \log a_1 + (1-\alpha) \log a_2 \le \log(\alpha a_1 + (1-\alpha) a_2). \quad \text{(A.2)}$$

Let $c_i = \int e^{s_i \cdot x} P(dx)$, $i = 1, 2$. If either $c_1 = \infty$ or $c_2 = \infty$ the inequality (A.1) is trivial. If both c_1 and c_2 are finite let

$$a_i = e^{s_1 \cdot x} / c_i, \qquad i = 1, 2. \quad \text{(A.3)}$$

Inserting this in (A.2) and taking the exponential function on both sides we get

$$\exp\{\alpha s_1 \cdot x + (1-\alpha) s_2 \cdot x\} / (c_1^{\alpha} c_2^{1-\alpha}) \quad \le \quad \alpha e^{s_1 \cdot x} / c_1 + (1-\alpha) e^{s_2 \cdot x} / c_2. \quad \text{(A.4)}$$

By integrating both sides of (A.4) with respect to x we obtain (A.1).

Since log is a strictly concave function we have strict inequality in (A.2) if $0 < \alpha < 1$ and $a_1 \neq a_2$. Hence, if $0 < \alpha < 1$ and $s_1 \neq s_2$, then equality in (A.1) is obtained if and only if in (A.3) $a_1 = a_2$ with probability 1 with respect to P, which is equivalent to

$$(s_1 - s_2) \cdot x = \log(c_1 / c_2) \quad \text{(A.5)}$$

with probability 1 with respect to P. Since $s_1 - s_2 \neq 0$, the set of xs that satisfy (A.5) is an affine subspace of $I\!R^k$, and hence the condition for strict inequality in (A.1) follows. □

Using the result of Proposition A.3, we may now obtain the convexity properties of M_P and $\mathcal{K}_P$.

Theorem A.4 *Let $P \in \mathcal{M}_k$. Then*

(i) The set Θ_P is convex.

(ii) M_P is a convex function on Θ_P.

(iii) $\mathcal{K}_P$ is a convex function on Θ_P, and strictly convex if and only if P is not concentrated on an affine subspace of $\mathbf{R}^k$.

Proof: Assume that $s_1, s_2 \in \Theta_P$ and $0 \le \alpha \le 1$. Then by Hölder's inequality

$$\begin{aligned} M_P(\alpha s_1 + (1-\alpha)s_2) &= \int \exp\{\alpha s_1 \cdot x + (1-\alpha)s_2 \cdot x\} P(dx) \\ &\le M_P(s_1)^\alpha M_P(s_2)^{1-\alpha}. \end{aligned} \tag{A.6}$$

By the definition of Θ_P we have $M_P(s_i) < \infty$ for $i = 1, 2,$ and hence by (A.6) $M_P(\alpha s_1 + (1-\alpha)s_2) < \infty$. This implies that $\alpha s_1 + (1-\alpha)s_2 \in \Theta_P$, and hence Θ_P is convex.

The convexity of $\mathcal{K}_P$ follows from (A.1) by taking logs on both sides. The condition for strict convexity follows from the condition for strict inequality in Hölder's inequality. Finally, the convexity of M_P follows because M_P is the composition of exp and $\mathcal{K}_P$, where exp is convex and increasing and $\mathcal{K}_P$ is convex, see Problem A.11. □

Example A.5 Consider a multinomial random vector $X = (X_1, \ldots, X_k)^T$ with probability function

$$f(x_1, \ldots, x_k) = \binom{n}{x_1 \ldots x_k} p_1^{x_1} \ldots p_k^{x_k},$$

where $p_i \ge 0$, $x_i \ge 0$, $i = 1, \cdots, k$, $p_1 + \cdots + p_k = 1$ and $x_1 + \cdots + x_k = n$. The moment generating function of X is, for $s = (s_1, \ldots, s_k)^T \in \mathbb{R}^k$,

$$\begin{aligned} M_X(s) &= \sum_{x_1+\cdots+x_k=n\,;\, x_i \ge 0} \binom{n}{x_1 \ldots x_k} p_1^{x_1} \cdot \ \ldots \ \cdot p_k^{x_k} \exp(s_1 x_1 + \cdots + s_k x_k) \\ &= \sum_{x_1+\cdots+x_k=n} \binom{n}{x_1 \ldots x_k} \prod_{j=1}^k \left(\frac{p_j e^{s_i}}{\sum_{i=1}^k p_i e^{s_i}} \right)^{x_j} \left(\sum_{i=1}^k p_i e^{s_i} \right)^{x_1+\cdots+x_k} \\ &= \left(\sum_{i=1}^k p_i e^{s_i} \right)^n . \end{aligned}$$

Since $X_1 + \cdots + X_k = n$ the distribution is concentrated on an affine subspace of $\mathbb{R}^k$, and it is not difficult to verify directly (Problem A.2) that the cumulant generating function of X is not strictly convex, in agreement with Theorem A.4 (iii).

Let us consider the linear transformation $Y = BX = (X_1, \ldots, X_{k-1})^T$, where

$$B = \begin{pmatrix} 1 & 0 & \cdots & & 0 \\ 0 & 1 & & & \\ \vdots & & \ddots & & \\ 0 & & \cdots & 1 & 0. \end{pmatrix}$$

By Theorem A.1 (iii) the moment generating function of Y is, for $s = (s_1, \ldots, s_{k-1}) \in \mathbb{R}^{k-1}$,

$$M_Y(s) = \left(\sum_{i=1}^{k-1} p_i e^{s_i} + p_k \right) .$$

A.2 The characteristic function and the Fourier-Laplace transform

A very important property of the moment generating function is that it is analytic, making the powerful tools of complex function theory available to us.

For this purpose we need to extend the definition of the moment generating function to complex arguments. For $P \in \mathcal{M}_k$ we define the *Fourier-Laplace* transform of P by

$$\tilde{M}_P(z) = \int e^{z \cdot x} \tilde{M}_P(z) = \int e^{z \cdot x} P(dx), \qquad z \in \mathbb{C}^k. \tag{A.7}$$

More explicitly, let $z = (z_1, \ldots, z_k)^T$ and write $z_j = s_j + it_j$ where i is the imaginary unit. Then

$$\tilde{M}(s_1 + it_1, \ldots, s_k + it_k) \;=\; \int e^{s \cdot x} \cos(t \cdot x) P(dx) + \int e^{s \cdot x} \sin(t \cdot x) P(dx),$$

where $s = (s_1, \ldots, s_k)^T$ and $t = (t_1, \ldots, t_k)^T$.

A special case of the Fourier-Laplace transform is the *characteristic function*, defined by $\varphi_P(t) = \tilde{M}_P(it)$ for $t \in \mathbb{R}^k$. Hence

$$\begin{aligned} \varphi_P(t) &= \int e^{it \cdot x} P(dx) \\ &= \int \cos(t \cdot x) P(dx) + i \int \sin(t \cdot x) P(dx), \qquad t \in \mathbb{R}^k. \end{aligned}$$

We use the same conventions regarding random variables as in the case of moment generating functions, and write $\tilde{M}_X$ instead of $\tilde{M}_P$ if $X \sim P$ etc.

The next theorem summarizes the elementary properties of $\tilde{M}_P$ and φ_P. By $\Re z$ we mean the real part of the complex number z.

Theorem A.6 *Let $P \in \mathcal{M}_k$. Then*

(i) The integral (A.7) is absolutely convergent if and only if $z \in \tilde{\Theta}_P$, where

$$\tilde{\Theta}_P = \Theta_P + i\mathbb{R}^k = \{s + it\colon\ s \in \Theta_P, \quad t \in \mathbb{R}^k\}.$$

(ii) $\tilde{M}_P(s) = M_P(s)$ *for* $s \in \Theta_P$.

(iii) $\tilde{M}_P(it) = \varphi_P(t)$ *for* $t \in \mathbb{R}^k$.

(iv) $\tilde{M}_P(0) = \varphi_P(0) = 1$.

(v) $\left|\tilde{M}_P(z)\right| \le M_P(\Re\, z)$ *for* $z \in \tilde{\Theta}_P$, *where* $\Re\, z = (\Re\, z_1, \ldots, \Re\, z_k)^T$.

(vi) $|\varphi_P(t)| \leq 1 \quad$ *for* $\quad t \in \mathbf{R}^k$.

(vii) *Let B be an $\ell \times k$ matrix and c an $\ell \times 1$ vector, and let X have distribution P. Then*

$$\tilde{M}_{BX+c}(z) = \tilde{M}_X(B^T z)e^{z\cdot c},\ B^T Z \in \tilde{\Theta}_P$$

and

$$\varphi_{BX+c}(t) = \varphi_X(B^T t)e^{it\cdot c}, \qquad t \in I\!R^k$$

(viii) *Let $X = (X_1^T, X_2^T)^T$ be a partition of X with components of dimension d and $k-d$, respectively, and let $z = (z_1^T, z_2^T)^T$ and $t = (t_1^T, t_2^T)^T$ be similar partitions of $z \in \mathbb{C}^k$ and $t \in \mathbf{R}^k$. If X_1 and X_2 are independent then*

$$\tilde{M}_X(z) = \tilde{M}_{X_1}(z_1)\tilde{M}_{X_2}(z), \qquad z \in \tilde{\Theta}_P$$

and

$$\varphi_X(t) = \varphi_{X_1}(t_1)\varphi_{X_2}(t_2), \qquad t \in I\!R^k.$$

Proof: For $z = s + it \in \mathbb{C}^k$ we have

$$|e^{z\cdot x}| = \left|e^{s\cdot x}e^{it\cdot x}\right| = e^{s\cdot x}. \tag{A.8}$$

Hence, by majorization we have that the integral (A.7) is convergent if and only if $s = \Re\ z \in \Theta_P$, which shows (i). By (A.8)

$$\left|\tilde{M}_P(z)\right| \leq \int |e^{z\cdot x}|\, P(dx) = M_P(s),$$

which shows (v) and, for $s = 0$, (vi). The remaining parts of the theorem are trivial, and the proofs are left for the reader. □

A.3 Analytic properties of univariate moment generating functions

To facilitate the discussion of the analytic properties we begin with the univariate case, which allows us to rely on the familiar theory for analytic functions in one variable. The results in the multivariate case, treated in Section A.5, may be derived from the corresponding univariate result.

Let $P \in \mathcal{M}_1$ and let $\tilde{M}_P$ be the Fourier-Laplace transform of P and M_P the moment generating function of P. The effective domain for M_P is then an interval on the real line, and $\tilde{\Theta}_P$ is the corresponding vertical strip in the complex plane. The following theorem shows that $\tilde{M}_P$ is an analytic function.

Theorem A.7 *Let $P \in \mathcal{M}_1$ and assume that $0 \in$ int Θ_P. Then the Fourier-Laplace transform $\tilde{M}_P$ is analytic on int $\tilde{\Theta}_P$. The Taylor expansion of $\tilde{M}_P$ around 0 is*

$$\tilde{M}_P(z) = \sum_{j=0}^{\infty} \frac{\mu_j(P)}{j!} z^j, \tag{A.9}$$

where

$$\mu_j(P) = \int x^j P(dx)$$

is the jth moment for P, which exists for any $j = 0, 1, 2, \ldots$. The cumulant generating function $\mathcal{K}_P$ is analytic with Taylor expansion

$$\mathcal{K}_P(s) = \sum_{j=0}^{\infty} \frac{\mathcal{K}_j(P)}{j!} s^j \tag{A.10}$$

around 0, where $\mathcal{K}_j(P)$ is the jth cumulant of P, which exists for any $j = 0, 1, 2, \ldots$. In particular, P has mean and variance (writing $X \sim P$)

$$E(X) = \mathcal{K}_P'(0) \qquad Var(X) = \mathcal{K}_P''(0),$$

respectively. If P is not degenerate, then $Var(X) > 0$.

Proof: Let $z_0 = s_0 + it_0 \in$ int $\tilde{\Theta}_P$ and write $\tilde{M}_P$ as follows for $|z - z_0| < \varepsilon$

$$\begin{aligned} \tilde{M}_P(z) &= \int \exp\{(z - z_0)x + z_0 x\} P(dx) \\ &= \int \sum_{j=0}^{\infty} \frac{(z - z_0)^j x^j}{j!} \, e^{z_0 x} P(dx), \end{aligned} \tag{A.11}$$

where $\varepsilon > 0$ is such that $|z - z_0| < \varepsilon$ implies $z \in$ int $\tilde{\Theta}_P$. For any $n \geq 0$ we have

$$\begin{aligned} \left| \sum_{j=0}^{n} \frac{(z - z_0)^j x^j}{j!} \, e^{z_0 x} \right| &\leq \sum_{j=0}^{\infty} \frac{|z - z_0|^j \, |x|^j}{j!} \, |e^{z_0 x}| \\ &= \exp(|z - z_0| \, |x| + s_0 x) \\ &\leq (e^{\varepsilon x} + e^{-\varepsilon x}) e^{s_0 x}. \end{aligned} \tag{A.12}$$

We have $s_0 \pm \varepsilon \in \Theta_P$, so the integral of (A.12) with respect to P is finite. By Lebesgue's dominated convergence theorem we may hence interchange integration and summation in (A.11). Thus, for $|z - z_0| < \varepsilon$

$$\tilde{M}_P(z) = \sum_{j=0}^{\infty} \frac{(z - z_0)^j}{j!} \int x^j e^{z_0 x} P(dx),$$

which shows that $\tilde{M}_P$ is analytic on int $\tilde{\Theta}_P$. For $z_0 = 0$ we obtain (A.9), and that $\mu_j(P)$ exists for any $j \geq 0$.

Let Log denote the principal branch of the complex logarithm. Since $M_P(s) > 0$ for $s \in \Theta_P$, $\text{Log}\tilde{M}_P(z)$ exists in the domain

$$R = \{z \in \text{int } \tilde{\Theta}_P \colon \tilde{M}_P(z) \notin (-\infty, 0]\} \tag{A.13}$$

and is analytic in R. Hence $\mathcal{K}_P(s) = \log M_P(s)$ is analytic with Taylor series (A.10). Finally, if P is not degenerate, then $\mathcal{K}_P''(0) = \text{Var } (X) > 0$, cf. Problem A.20. □

Example A.8 *Let us find the characteristic function of the normal distribution. From Example A.2 we know that the moment generating function of the univariate normal distribution* $N(0,1)$ *is* $\exp\left(\frac{1}{2}s^2\right)$. *By analytic continuation, the corresponding characteristic function is* $\exp\left\{\frac{1}{2}(it)^2\right\} = \exp\left(-\frac{1}{2}t^2\right)$. *By Theorem A.6 (viii) the characteristic function of* $N_k(0, I_k)$ *is hence*

$$\exp(-\frac{1}{2}t_1^2) \cdot \cdots \cdot \exp(-\frac{1}{2}t_k^2) = \exp(-\frac{1}{2}t \cdot t),$$

where $t = (t_1, \ldots, t_k)$. *Finally, by Theorem A.6 (vii) we conclude that the characteristic function of the normal distribution* $N_k(\mu, \Sigma)$ *is*

$$\exp(-\frac{1}{2}t^T \Sigma t + it \cdot \mu),$$

where we have used the transformation $X \to BX + \mu, \quad \Sigma = BB^T$, *as in Example A.2.*

By Theorem A.7, $M_P(s)$ is continuous and differentiable on int Θ_P, and in the next theorem we show that M_P is continuous at the boundary of Θ_P in the one-dimensional case. This is not generally so in the multivariate case, cf. Barndorff-Nielsen (1978, p. 105).

Theorem A.9 *Let* $P \in \mathcal{M}_1$, *and assume that* Θ_P *has a finite endpoint* θ_0. *Let* $\lim_{\theta \to \theta_0}$ *denote either* $\lim_{\theta \downarrow \theta_0}$ *or* $\lim_{\theta \uparrow \theta_0}$, *depending on whether* θ_0 *is the upper or lower endpoint of* Θ_0. *Then the following two statements are equivalent*

(i) $\theta_0 \in \Theta_P$

(ii) $\lim_{\theta \to \theta_0} M_P(\theta)$ *exists and is finite.*

If (i) or (ii) holds, then $M_P(\theta_0) = \lim_{\theta \to \theta_0} M_P(\theta)$.

Proof: We assume that θ_0 is the lower endpoint of Θ_P, the proof in the opposite case being similar. If $\theta_0 \in \Theta_P$, then $M_P(\theta_0) < \infty$. For $\theta_0 \leq \theta \leq \theta_0 + \varepsilon$, we have

$$e^{\theta x} \leq e^{\theta_0 x} + e^{(\theta_0 + \varepsilon)x}. \tag{A.14}$$

For $\theta_0 + \varepsilon \in \Theta_P$ we have $M_P(\theta_0 + \varepsilon) < \infty$, which together with (A.14) and Lebesgue's Dominated Convergence Theorem implies that $\lim_{\theta \downarrow \theta_0} M_P(\theta)$ exists and is equal to $M_P(\theta_0)$. This shows the implication $(i) \Rightarrow (ii)$. Now assume that (ii) holds. By Fatou's Lemma, applied to the sequence of positive functions $e^{\theta x}$ for a sequence of θs, we find that $M_P(\theta_0) \leq \lim_{\theta \to \theta_0} M_P(\theta)$. Hence $\theta_0 \in \Theta_P$, concluding the proof. □

A.4 The uniqueness theorem for characteristic functions

We now show that a distribution is characterized by its characteristic function. The Fourier-Laplace transform provides a link between the moment generating function M_P and the characteristic function φ_P. This allows us to use the fact that φ_P characterizes P to show that M_P also characterizes P.

Theorem A.10 *Let $P_1, P_2 \in \mathcal{M}_k$ be two distributions such that*

$$\varphi_{P_1}(t) = \varphi_{P_2}(t) \qquad for \qquad t \in I\!R^k.$$

Then $P_1 = P_2$.

Proof: We show how a distribution $P \in \mathcal{M}_K$ may be recovered from its characteristic function φ. The starting point of the proof is the fact that

$$e^{-it\cdot s}\varphi(t) = \int e^{it\cdot(x-s)}P(dx).$$

By integrating both sides of this equation with respect to the density function of the normal distribution $N_k(0, a^{-1}I_k)$, we get

$$\begin{aligned}
& \left(\frac{2\pi}{a}\right)^{-k/2} \int \varphi(t) \exp(-it\cdot s - \frac{1}{2}at\cdot t)dt \\
= & \iint e^{it\cdot(x-s)}P(dx)\left(\frac{2\pi}{a}\right)^{-k/2} e^{-1/2at\cdot t}dt \\
= & (\frac{2\pi}{a})^{-k/2} \iint \exp\left\{it\cdot(x-s) - \frac{1}{2}at\cdot t\right\} dtP(dx) \\
= & \int \exp\{-\frac{1}{2a}(s-x)\cdot(s-x)\}P(dx),
\end{aligned}$$

where we have use the result of Example A.8. Hence we have the relation

$$\left(\frac{2\pi}{a}\right)^{-\frac{k}{2}} \int \exp(-it\cdot s - \frac{1}{2}at\cdot t)\varphi(t)dt = \int f_a(s-x)P(dx), \tag{A.15}$$

where f_a is the density function of the normal distribution $N_k(0, a\ I_k)$. The right-hand side of equation (A.15) is the density function of the convolution of P with the normal distribution $N_k(0, a\ I_k)$, whereas the left-hand side depends on P only through φ. Since the convolution on the right-hand side converges in distribution to P as a tends to 0, we may hence recover P from φ, which proves the theorem. □

Corollary A.11 *(Cramér-Wold) Let X have distribution $P \in \mathcal{M}_k$. Then P is uniquely determined by the set of marginal distributions of $\theta \cdot X$ for $\theta \in \mathbb{R}^k$.*

Proof: The characteristic function of $\theta \cdot X$ is for $s \in \mathbb{R}$

$$\varphi_{\theta \cdot X}(s) = \int \exp(is\theta \cdot x) P(dx) = \varphi_X(s\theta). \tag{A.16}$$

If the distribution of $\theta \cdot X$ is known for any $\theta \in \mathbb{R}^k$, then by (A.16) the characteristic function $\varphi_X(u)$ is known for any $u = s\theta \in \mathbb{R}^k$. Hence, by the uniqueness theorem, the distribution of X is known. □

Using Theorem A.10 we may now show that a univariate analytic moment generating function characterizes its distribution. The corresponding result for the multivariate case is shown in the next section.

Theorem A.12 *Let P_1 and P_2 belong to $\mathcal{M}_1$. If there exists an open set $S \subseteq \Theta_{P_1} \cap \Theta_{P_2}$ such that*

$$M_{P_1}(s) = M_{P_2}(s) \quad \text{for} \quad s \in S, \tag{A.17}$$

then $P_1 = P_2$.

Proof: Let $\theta_0 \in S$ and define

$$Q_i(dx) = \{e^{\theta_0 x} / M_{P_i}(\theta_0)\} P_i(dx), \quad i = 1, 2. \tag{A.18}$$

Then Q_1 and Q_2 are distributions in $\mathcal{M}_1$, and actually (A.18) is an example of a linear exponential family. The Fourier-Laplace transform of Q_i is

$$\tilde{M}_{Q_i}(z) = \tilde{M}_{P_i}(z + \theta_0) / M_{P_i}(\theta_0), \Re\, z \in \Theta_{P_i} - \theta_0.$$

Defining $R = \operatorname{int} \tilde{\Theta}_{P_1} \cap \operatorname{int} \tilde{\Theta}_{P_2} - \theta_0$, we have $S - \theta_0 \subseteq R$, and by (A.17) $\tilde{M}_{Q_1}$ and $\tilde{M}_{Q_2}$ are identical on $S - \theta_0$. By analytic continuation, $\tilde{M}_{Q_1}$ and $\tilde{M}_{Q_2}$ are hence identical on R, and since R includes the imaginary axis we conclude that Q_1 and Q_2 have the same characteristic function. Hence $Q_1 = Q_2$, and by (A.18) this implies $P_1 = P_2$. □

A.5 Analytic properties of multivariate moment generating functions

We now generalize the results of Section A.3 to the multivariate case. We first show that an analytic moment generating function characterizes its distribution.

Theorem A.13 *Let $P \in \mathcal{M}_k$ and let M_P be the moment generating function of P with effective domain Θ_P. If int $\Theta_P \neq \emptyset$ then M_P characterizes P.*

Proof: Let $\theta_0 \in \text{int } \Theta_P$, and define the distribution Q by

$$Q(dx) = \{e^{\theta_0 \cdot x}/M_P(\theta_0)\}P(dx). \tag{A.19}$$

with moment generating function

$$M_Q(s) = M_P(\theta_0 + s)/M_P(\theta_0), \quad s \in \Theta_P - \theta_0.$$

If $X \sim Q$ and $\theta \in I\!R^k$, then $\theta \cdot X$ has moment generating function

$$M_{\theta \cdot X}(s) = M_P(\theta_0 + s\theta)/M_P(\theta_0),$$

and since int $\Theta_P \neq \emptyset$ we have int $\Theta_{\theta \cdot X} \neq \emptyset$ for any $\theta \in I\!R^k$. By Theorem A.12 the distribution of $\theta \cdot X$ may be recovered from M_Q, which in turn is defined in terms of M_P. Hence, by Corollary A.11, the distribution of Q may be recovered from M_P. By (A.19) P is uniquely determined by Q, and hence the conclusion of the theorem follows. □

Corollary A.14 *Let $P \in \mathcal{M}_k$ and assume that int $\Theta_P \neq \emptyset$. Then the function $s \to M_P(\theta_0 + s\theta)$ is analytic for any $\theta_0 \in$ int Θ_P and $\theta \in I\!R^k$. In particular, M_P is analytic separately in each coordinate.*

Proof: Follows immediately from the proof of Theorem A.13. □

As a prologue to the multivariate version of Theorem A.7 we look at the Taylor expansion of the exponential function. The reason is that in the proof of Theorem A.7 the Taylor expansion of the Fourier-Laplace transform was obtained by integrating the Taylor expansion of the exponential function term by term. In the multivariate case we need the expansion of $\exp(z_1 + \cdots + z_k)$, which may be obtained from the Taylor expansion of exp. Thus

$$\begin{aligned}
\exp(z_1 + \cdots + z_k) &= \sum_{i=0}^{\infty} \frac{(z_1 + \cdots + z_k)^i}{i!} \\
&= \sum_{i=0}^{\infty} \frac{1}{i!} \sum_{\substack{i_1, \ldots, i_k \geq 0 \\ i_1 + \cdots + i_k = i}} \binom{i}{i_1 \ldots i_k} z_1^{i_1} \cdots z_k^{i_k} \\
&= \sum_{i_1 = \cdots = i_k = 0}^{\infty} \frac{z_1^{i_1} \cdots z_k^{i_k}}{i_1! \cdots i_k!}.
\end{aligned}$$

This shows that $\exp(z_1 + \cdots + z_k)$ is analytic as a function of the complex arguments $z_1, \ldots, z_k$, by displaying it as the sum of its Taylor series. In general, a function of several complex variables, f, is said to be analytic in a region S if for every $z_0 \in S$, f is given by its Taylor expansion around z_0 in some neighbourhood of z_0,

$$f(z) = \sum_{i_1 = \cdots = i_k = 0}^{\infty} \frac{\partial^{i_1 + \cdots + i_k} f(z_0)}{\partial z_1^{i_1} \ldots \partial z_k^{i_k}} \cdot \frac{(z_1 - z_{01})^{i_1} \cdot \cdots \cdot (z_k - z_{0k})^{i_k}}{i_1! \cdots i_k!}.$$

The multivariate version of Theorem A.7 is

Theorem A.15 *Let $P \in \mathcal{M}_k$ and assume that $0 \in$ int Θ_P. Then the Fourier-Laplace transform $\tilde{M}_P$ is analytic on int $\tilde{\Theta}_P$. The Taylor expansion of $\tilde{M}_P$ is*

$$\tilde{M}_P(z) = \sum_{i_1=\ldots=i_k=0} \frac{\mu_{i_1\ldots i_k}(P)}{i_1! \cdot \ \ldots \ \cdot i_k!} z_1^{i_1} \cdot \ \ldots \ \cdot z_k^{i_k} \tag{A.20}$$

where

$$\mu_{i_1\ldots i_k}(P) = \int x_1^{i_1} \ldots x_k^{i_k} P(dx)$$

is the $(i_1, \ldots, i_k)th$ moment of P and $z = (z_1, \ldots, z_k)^T$, $x = (x_1, \ldots, x_k)^T$. The cumulant generating function $\mathcal{K}_P$ is analytic with Taylor expansion around 0.

$$\mathcal{K}_p(s) = \sum_{i_1=\ldots=i_k=0}^{\infty} \frac{\mathcal{K}_{i_1\ldots i_k}(P)}{i_1! \ldots i_k!} \ s_1^{i_1} \cdot \ \ldots \ \cdot s_k^{i_k}, \tag{A.21}$$

where $\mathcal{K}_{i_1\ldots i_k}(P)$ is the $(i_1, \ldots, i_k)th$ cumulant of P and $s = (s_1, \ldots, s_k)^T$. In particular, P has moments and cumulants of arbitrary order.

Proof: Let $z_0 = s_0 + it_0 \in$ int $\tilde{\Theta}_P$ be given, and let $\varepsilon > 0$ be such that $|z - z_0| < \varepsilon$ implies $z \in$ int $\tilde{\Theta}_P$. Here $|z|$ denotes the Euclidean norm on $\mathbb{C}^k$, obtained by identifying $\mathbb{C}^k$ with the Euclidean space $\mathbb{R}^{2k}$. Using the univariate Taylor expansion of the exponential function we obtain

$$\begin{aligned} \tilde{M}_P(z) &= \int \exp\{(z - z_0) \cdot x + z_0 \cdot x\} P(dx) \\ &= \int \sum_{i=0}^{\infty} \frac{\{(z - z_0) \cdot x\}^i}{i!} \ e^{z_0 \cdot x} P(x). \end{aligned}$$

The following inequality justifies the use of Lebesgue's theorem of dominated convergence, writing $s_0 = (s_{01}, \ldots, s_{0k})^T$ etc.,

$$\begin{aligned} \left| \sum_{i=0}^{n} \frac{\{(z - z_0) \cdot x\}^i}{i!} \ e^{z_0 \cdot x} \right| &\leq \sum_{i=0}^{\infty} \frac{|(z - z_0) \cdot x|^i}{i!} \ e^{s_0 \cdot x} \\ &= \exp\{|(z - z_0) \cdot x| + s_0 \cdot x\} \\ &\leq \exp\{|(z_1 - z_{01})| \, x_1 + \cdots + |(z_k - z_{0k}) x_k| + s_0 \cdot x\} \quad \text{(A.22)} \\ &\leq \prod_{i=1}^{k} \{(e^{\varepsilon x_i} + e^{-\varepsilon x_i}) e^{s_{oi} x_i}\}. \quad \text{(A.23)} \end{aligned}$$

We may choose ε such that the rectangle with vertices $s_{0i} \pm \varepsilon$ is contained in int Θ_P. Then the function (A.23) is integrable with respect to P, and interchanging summation and integration

in (A.22) we obtain

$$\begin{aligned}\tilde{M}_P(z) &= \sum_{i=0}^{\infty} \int \frac{\{(z-z_0)\cdot x\}^i}{i!}\, e^{z_0\cdot x} P(dx) \\ &= \sum_{i=0}^{\infty} \int \frac{1}{i} \sum_{\substack{i_1,\ldots,i_k \geq 0 \\ i_1+\cdots+i_k=i}} \binom{i}{i_1 \ldots i_k} (z_1-z_{01})^{i_1} x_1^{i_1} \cdot \ \ldots \ \cdot (z_k-z_{0k})^{i_k} x_k^{i_k} e^{z_0\cdot x} P(dx) \\ &= \sum_{i_1=\ldots=i_k=0}^{\infty} \frac{(z_1-z_{01})^{i_1}\ldots(z_k-z_{0k})^{i_k}}{i_1!\cdot\ldots\cdot i_k!} \int x_1^{i_1}\cdot\ldots\cdot x_k^{i_k} e^{z_0\cdot x} P(dx),\end{aligned}$$

which shows that $\tilde{M}_P$ is analytic in int $\tilde{\Theta}_P$. For $z_0 = 0$ we obtain (A.20). Since $M_P(s) > 0$ for $s \in \Theta_P$, $\mathrm{Log}\tilde{M}_P$ is defined and analytic in a region of $\mathbb{C}^k$ containing $\tilde{\Theta}_P$, and hence $\mathcal{K}_P = \log M_P$ is analytic, with Taylor expansion (A.21). □

Corollary A.16 *Let $P \in \mathcal{M}_k$, let $X \sim P$, and assume that $0 \in$ int Θ_P. Then X has mean vector $\mu = (\mu_1, \ldots, \mu_k)^T$ with*

$$\mu_i = \frac{\partial \mathcal{K}_P(s_1,\ldots,s_k)}{\partial s_i} \mid_{s=0}$$

and variance matrix $V = \{v_{ij}, \quad i,j = 1,\ldots,k\}$ with

$$v_{ij} = \frac{\partial^2 \mathcal{K}_P(s_1,\ldots,s_k)}{\partial s_i \partial s_j} \mid_{s=0} .$$

The variance matrix V is positive-definite if and only if P is not concentrated on an affine subspace of $\mathbb{R}^k$.

Proof: The expressions for μ_i and v_{ij} follow from the results of Theorem A.15. The condition for positive-definiteness of V follows from general properties of variance matrixes, cf. Problem A.20. □

A.6 Problems

Problem A.1 Prove Theorem A.1.

Problem A.2 Show that the cumulant generating function of the multinomial distribution is not strictly convex.

Problem A.3 Show, without using Theorem A.4, that the cumulant generating function of the multivariate normal distribution $N_k(\mu, \Sigma)$ is strictly convex if and only if Σ is positive-definite.

Problem A.4 Use the convexity of the exponential function to show, without using Theorem A.4, that the moment generating function M_P is convex.

Problem A.5 Find the moment and cumulant generating functions of the following distributions: degenerate, uniform, exponential, gamma, normal, inverse Gaussian, Poisson, binomial and negative binomial.

Problem A.6 Let P be the measure with probability density $\frac{1}{2}\, e^{-|x|}$. Find M_P and Θ_P for this measure.

Problem A.7 Let P be the Cauchy distribution with density function

$$f(x) = 1/\{\pi(1 + x^2)\}.$$

Show that $\Theta_P = \{0\}$.

Problem A.8 Show that the moment generating function of a univariate distribution P is strictly convex if and only if P is not concentrated in zero.

Problem A.9 Show that the only univariate distributions whose cumulant generating functions are not strictly convex are the degenerate distributions.

Problem A.10 Let $U = \{a + bt\colon\ t \in \mathbb{R}\}$ with $a, b \in \mathbb{R},\ b \neq 0$, denote an affine subspace of $\mathbb{R}^2$. Give an example of a distribution P concentrated on U, but not degenerate, and show that $\mathcal{K}_P$ is not strictly convex.

Problem A.11 Show that if $f\colon A \to B$, and $g\colon B \to \mathbb{R}$, where $A \subseteq \mathbb{R}^k$ is convex and $B \subseteq \mathbb{R}$ is an interval, and f and g are convex functions and g increasing, then $g \circ f$ is convex. Give conditions under which $g \circ f$ is strictly convex. Show that $\exp\{f(x)\}$ is a convex function for f convex.

Problem A.12 Show that if $P \in \mathcal{M}_k$ has bounded support, then $\Theta_P = \mathbb{R}^k$.

Problem A.13 Consider the moment generating function M_X in Example A.5 for $k = 2$. Derive the relation between M_X and the moment generating function of the binomial distribution.

Problem A.14 Consider the moment generating function M_Y in Example A.5. Give necessary and sufficient conditions for $\mathcal{K}_Y = \log M_Y$ to be strictly convex.

Problem A.15 Make a plot of $\tilde{\Theta}_P$ for the distributions mentioned in Problem A.5.

Problem A.16 Verify the details of the proof of Theorem A.10 in the case $k = 1$.

Problem A.17 Show that the cumulants of the standard normal distribution $N(0,1)$ are $\mathcal{K}_1 = 0$, $\mathcal{K}_2 = 1$ and $\mathcal{K}_j = 0$ for $j \geq 2$.

Problem A.18 Write the Taylor expansions (A.9) and (A.10) for each of the distributions of Problem A.5. In particular, find mean and variance in each case.

Problem A.19 Let P and Q be distributions in $\mathcal{M}_k$ such that there exists an open set S with $0 \in S \subseteq \Theta_P \cap \Theta_Q$ with $M_P(s) = M_Q(s)$ for $s \in S$. Show that $P = Q$. Can you suggest any improvements of this result?

Problem A.20 Show that $Var(X) \geq 0$ for any random variable X, and that $Var(X) > 0$, unless X is degenerate. Use this result to show that the variance-covariance matrix V for a random vector X is non-negative definite, and that V is positive definite, unless X is concentrated on an affine subspace of $\mathbb{R}^k$.

Problem A.21 Prove that $X + aY \xrightarrow{d} X$ as $a \to 0$ for any two random vectors X and Y. This is important for the proof of one of the theorems in this appendix. Which?

Bibliography

[1] Andersen, E.B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *J. Roy. Statist. Soc. Ser.* **B 32**, 283–301.

[2] Andersen, E.B. (1971). Asymptotic properties of conditional likelihood ratio tests. *J. Amer. Statist. Assoc.* **66**, 630–633.

[3] Andersen, E.B. (1973). *Conditional Inference and Models for Measuring.* Mentalhygiejnisk Forlag, Copenhagen.

[4] Andersen, H.C. (1866). *Eventyr Fortalte for Børn.*

[5] Bahadur, R.R. (1955a). A characterization of sufficiency. *Ann. Math. Statist.* **26**, 286–293.

[6] Bahadur, R.R. and Lehmann, E.L. (1955). Two comments on sufficiency and statistical decision functions. *Ann. Math. Statist.* **26**, 139–141.

[7] Bahadur, R.R. (1954). Sufficiency and statistical decision functions. *Ann. Math. Statist.* **25**, 423–462.

[8] Bahadur, R.R. (1955). Statistics and subfields. *Ann. Math. Statist.* **26**, 490–497.

[9] Barnard, G.A. (1963). Some logical aspects of the fiducial argument. *J. Roy. Statist. Soc. Ser.* **B 25** 111–114.

[10] Barnard, G.A. (1973). Maximum likelihood and nuisance parameter. *Sankhyā* **A 35**, 133–135.

[11] Barndorff-Nielsen, O.E. (1973). On M-ancillarity. *Biometrika* **60**, 447–455.

[12] Barndorff-Nielsen, O.E. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

[13] Barndorff-Nielsen, O.E. (1988). *Parametric Statistical Models and Likelihood*. Lecture Notes in Statistics, Springer-Verlag, New York.

[14] Barndorff-Nielsen, O.E. and Blæsild, P. (1975). *S*-ancillarity in exponential families. *Sankhyā* **A 37**, 334–385.

[15] Barndorff-Nielsen, O.E., Blæsild, P., Jensen, J.L. and Jørgensen, B. (1982). Exponential transformation models. *Proc. Roy. Soc. London Ser.* **A 379**, 41–65.

[16] Barndorff-Nielsen, O.E., Blæsild, P. and Eriksen, P.S. (1989). *Decomposition and Invariance of Measures, and Statistical Transformation Models*. Lectures Notes in Statistics, Springer-Verlag, New York.

[17] Barndorff-Nielsen, O.E. and Pedersen, K. (1968). Sufficient data reduction and exponential families. *Math. Scand.* **22**, 197–202.

[18] Basu, D. (1955). On statistics independent of a complete sufficient statistic. *Sankhyā* **15**, 377–380.

[19] Basu, D. (1958). On statistics independent of a sufficient statistic. *Sankhyā* **20**, 223–229.

[20] Basu, D. (1959). The family of ancillary statistics. *Sankhyā* **20**, 247–256.

[21] Basu, D. (1978). On partial sufficiency: a review. *J. Statist. Plann. Inference* **2**, 1–13.

[22] Basu, D. and Ghosh, J.E. (1967). Sufficient statistics in sampling from a finite universe. *Proc.* 36 *th Session Internat. Statist. Inst.* (in ISI Bulletin), 850–858.

[23] Bhapkar, V.P. (1972). On a measure of efficiency of an estimating equation. *Sankhyā* **A 34**, 467–472.

[24] Berk, R.H. (1972). Consistency and asymptotic normality of MLE's for exponential models. *Ann. Math. Statist.* **43**, 193–200.

[25] Bickel,P.J. , Klaasen, C.S.J. , Ritov, R. and Wellner, J.A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

[26] Chandrasekar, B. (1988). An optimality criterion for vector unbiased statistical estimation functions. *J. Statist. Plann. Inference.* **18**, 115–117.

[27] Chandrasekar, B. and Kale, B.K. (1984). Unbiased statistical estimation function for the parameters in the presence of nuisance parameters. *J. Statist. Plann. Inference* **9**, 45–54.

[28] Chow, Y.S. and Teicher, H. (1988). *Probability Theory: Independence, Interchangeability, and Martingales*, second ed. New York: Springer-Verlag.

[29] Cox, D.R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29**, 357–372.

[30] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. Chapman and Hall, London.

[31] Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data* (2nd Ed.). London: Chapman & Hall.

[32] Durairajan, T.M. (1989). Characterization and uniqueness of optimal estimating function. *J. Statist. Plann. Inference* **18**, 115–117.

[33] Durbin, J. (1960). Estimation of parameters in time-series regression models. *J. Roy. Statist. Soc. Ser.* **B 22**, 139–153.

[34] Fisher, R.A. (1921). The mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser.* **A 222**, 309–368.

[35] Fisher, R.A. (1934). Two new properties of mathematical likelihood. *Proc. Royal Soc. London Ser.* **A 144**, 285–307.

[36] Fisher, R.A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc.* **98**, 39–54.

[37] Fisher, R.A. (1950). The significance of deviation from expectation in a Poisson series. *Biometrika* **6**, 17–24.

[38] Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh. (Third ed. 1973, Collier Macmillan, London.)

[39] Fraser, D.A.S. (1956). Sufficient statistics with nuisance parameters. *Ann. Math. Statist.* **27**, 838–842.

[40] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **81**, 1208–1212.

[41] Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63**, 277–284.

[42] Godambe, V.P. (1980). On sufficiency and ancillarity in the presence of a nuisance parameter. *Biometrika* **67**, 269–276.

[43] Godambe, V.P. (1984). On ancillarity and Fisher information in the presence of a nuisance parameter. *Biometrika* **71**, 626–629.

[44] Godambe, V.P. and Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Statist.* **2**, 568–571.

[45] Godambe, V.P. and Thompson, M.E. (1976). Some aspects of the theory of estimating equations. *J. Statist. Plann. Inference* **2**, 95–104.

[46] Haberman, S.J. (1974). *The analysis of frequency data.* University of Chicago Press.

[47] Haberman, S.J. (1977). Maximum likelihood estimate in experimental response models. *Ann. Statist.* **5**, 815–841.

[48] Halmos, P.R. and Savage, L.J. (1949). Application of the Radon-Nikodym theorem to the theory of sufficient statistics. *Ann. Math. Statist.* **20**, 225–241.

[49] Hoffmann-Jørgensen (1990) Asymptotic likelihood theory, *in:* Butkovié, D. , Kraljevic, H. , Kurepa, S. and Hoffmann-Jørgensen, J. *Functional Analysis III*: Proceedings of the Postgraduate School and Conference held at Inter-University Center. Mathematical Institute, Aarhus University.

[50] Johansen, S. (1979). *Introduction to the Theory of Regular Exponential Families*. Lecture Notes 3, Institute of Mathematical Statistics, University of Copenhagen.

[51] Jørgensen, B. (1982). *Statistical Properties of the Generalized Inverse Gaussian Distribution*. Lecture Notes in Statistics 9, Springer-Verlag, New York.

[52] Jørgensen, B. (1993). The rules of conditional inference: Is there a universal definition of nonformation? Technical report #130, Department of Statistics, University of British Columbia.

[53] Jørgensen, B. and Labouriau, R.S. (1992). *Famílias exponenciais e Inferência Teórica.* Monografias de Matemática no 52, IMPA, Rio de Janeiro.

[54] Joseph, B. and Durairajan, T.M. (1991). Equivalence of various optimality criteria for estimating functions. *J. Statist. Plann. Inference* **27**, 355–360.

[55] Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. Wiley, New York.

[56] Kale, B.K. and Chandrasekar, B. (1983). Equivalence of optimality criteria for vector unbiased statistics. *J. Indian Statist. Assoc.* **21**, 49–58.

[57] Kawata, T. (1972). *Fourier Analysis in Probability Theory*. New York: Academic Press.

[58] Keilson, J. and Gerber, H. (1971). Some results for discrete unimodality. *J. Amer. Statist. Ass.* **66**, 386–389.

[59] Kendall, B.K. and Stuart, A. (1979). *The Advanced Theory of Statistics*, 4th ed., Vol. 2. Charles Griffin, London.

[60] Kimball, B.K. (1946). Sufficient statistical estimation functions for the parameters of the distribution of maximum values. *Ann. Math. Statist.* **17**, 299–309.

[61] LeCam, L. (1990). Maximum likelihood: An introduction. *Internat. Statist. Rev.* **58**, 153–171.

[62] Lehmann, E.L. (1959). *Testing Statistical Hypothesis*. Wiley, New York.

[63] Lehmann, E.L. (1981). An interpretation of completeness and Basu's theorem. *J. Amer. Statist. Assoc.* **76**, 335–340.

[64] Lukacs, E.E. (1970). *Characteristic Functions*. Second Edition. London, Griffin.

[65] McLeish, D.L. and Small, C.G. (1987). *The Theory and Applications of Statistical inference Functions.* Lecture Notes in Statistics 44, Springer-Verlag, New York.

[66] Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrika* **16**, 1–32.

[67] Orwell, G. (1945). *Animal Farm: A Fairy Story*. Secker and Warburg (1987 edition), London.

[68] Pfanzagl, J. (1990). *Estimation in Semiparametric Models: Some Recent Developments.* Lecture Notes in Statistics, **63**. Springer-Verlag.

[69] Plausonio, A. (1968). *De Re Ætiopia*. XXII edition. Editora Rodeziana. Barcelona-São Paulo.

[70] Rasch, G. (1960). *Probabilistic Models for some Intelligence and Attainment Tests*. Studies in Mathematical Psychology I. Danish Institute for Educational Research, Copenhagen.

[71] Rémon, M. (1984). On a concept of partial sufficiency: L-sufficiency. *Internat. Statist. Rev.* **52**, 127–135.

[72] Rockafellar, R.T. (1970). *Convex Analysis.* Princeton University Press.

[73] Sandved, E. (1967). A principle for conditioning on an ancillary statistic. *Skand. Aktuar. J.* **50**, 39–47.

[74] Sandved, E. (1972). Ancillary statistics in models without and with nuisance parameters. *Scand. Actuar. J.* **55**, 81–91.

[75] Small, C.G. and McLeish, D.L. (1988a). Generalizations of ancillarity, completeness and sufficiency in an inference function space. *Ann. Statist.* **16**, 534–551.

[76] Small, C.G. and McLeish, D.L. (1988b). Projection as method for increasing sensitivity and eliminating nuisance parameters. *Technical Report Series* STAT-88-11. University of Waterloo.

[77] Stein, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1**, 187–195. California Press, Berkeley.

[78] Sverdrup, E. (1965). Estimates and test procedures in connection with stochastic models for deaths, recoveries, and transfers between different states of health. *Scand. Actuar. J.* **48**, 184–211.

[79] Van der Vaart, A. and Wellner, J. (1996) *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer-Verlag, New York. ISBN 0-387-94640-3.