ABSTRACT

Title of dissertation: STATISTICAL INFERENCE BASED ON ESTIMATING FUNCTIONS IN EXACT AND MISSPECIFIED MODELS

Ryan Janicki, Doctor of Philosophy, 2009

Dissertation directed by: Professor Abram Kagan
Department of Mathematics, Statistics Program

Estimating functions, introduced by Godambe, are a useful tool for constructing estimators. The classical maximum likelihood estimator and the method of moments estimator are special cases of estimators generated as the solution to certain estimating equations. The main advantage of this method is that it does not require knowledge of the full model, but rather of some functionals, such as a number of moments.

We define an estimating function $\mathbf{\Psi}$ to be a Fisher estimating function if it satisfies $E_{\boldsymbol{\theta}}\left(\mathbf{\Psi}\mathbf{\Psi}^T\right) = -E_{\boldsymbol{\theta}}\left(\partial\mathbf{\Psi}/\partial\boldsymbol{\theta}\right)$. The motivation for considering this class of estimating functions is that a Fisher estimating function behaves much like the Fisher score, and the estimators generated as solutions to these estimating equations behave much like maximum likelihood estimators. The estimating functions in this class share some of the same optimality properties as the Fisher score function and they have applications for estimation in submodels, elimination of nuisance parameters, and combinations of independent samples. We give some applications of estimating

functions to estimation of a location parameter in the presence of a nuisance scale parameter. We also consider the behavior of estimators generated as solutions to estimating equations under model misspecification when the misspecification is small and can be parameterized. A problem related to model misspecification is attempting to distinguish between a finite number of competing parametric families. We construct an estimator that is consistent and efficient, regardless of which family contains the true distribution.

Statistical Inference Based on Estimating Functions in Exact and Misspecified Models

by

Ryan Janicki

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Abram Kagan, Chair/Advisor
Professor Eric Slud
Professor Benjamin Kedem
Professor Paul Smith
Professor Prakash Narayan

Acknowledgments

I would like to express my gratitude to the many people who have helped me during my graduate career and have made this dissertation possible.

I would like to thank my advisor, Professor Abram Kagan, for giving me the opportunity to work with him over the past four years. Professor Kagan has taught me more than I thought possible. He has given me many challenging problems to work on and has been patient with me as I tried to solve them. It has been an honor to learn from such a great mathematician.

I want to thank all the people I've gotten to know and the friends I've made during my six years here. Thanks to Professors Eric Slud, Benjamin Kedem, Paul Smith, and Prakash Narayan for serving on my dissertation committee and thanks to Professors Eric Slud and Paul Smith for their valuable comments and suggestions which helped improve the dissertation. To my family, thanks for all your support. Thanks to all the professors in the department who have taught me so much. Thanks to my officemates who have made me laugh so many times. Thanks to Hyejin Kim for being a great friend and helping me through my early years as a graduate student.

For many, many reasons, thanks to my beautiful wife Avanti and my beautiful daughter Leela.

Table of Contents

Chapter 1

Introduction

1.1 Overview

The topic of this dissertation is statistical inference for samples $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ drawn from a population $P_{\boldsymbol{\theta}}$ parameterized by a scalar or vector valued parameter $\boldsymbol{\theta}$. Estimating functions $\boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta})$ for a statistical model $(\mathcal{X}, \mathcal{A}, \mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\})$ are a convenient tool since they can be based on partial information on $\mathcal{P}$ and yet preserve basic properties of the classical methods, such as the method of moments and the method of maximum likelihood. Estimating functions are also widely used in generalized linear models.

In Chapter 2, "Estimating functions," we present some general results on estimating functions and the estimators generated from estimating equations. A novelty is the concept of Fisher estimating functions and properties of the estimators they generate.

In Chapter 3, "Estimating functions for location parameter families," we describe the behavior of estimators of a location parameter and estimators of a location parameter in the presence of a nuisance scale parameter generated from estimating equations. Using the special characteristics of the parameters, we construct a simple modification of the Pitman estimator in the presence of a nuisance scale parameter.

Chapter 4, "Estimators by estimating equations in misspecified models," contains new results on the large sample (i.e., when $n \longrightarrow \infty$) behavior of estimators defined by estimating functions under misspecified models when the misspecification is of different order of smallness compared to $1/\sqrt{n}$. The models of misspecification under study look rather realistic for applications and, at the same time, convenient for a rigorous mathematical analysis.

In Chapter 5, "Analogues of classical tests based on estimating functions," we construct analogues of Rao's and Wald's classical tests and study their asymptotic properties when the parameter estimators for the extended and/or restricted model are obtained from estimating equations. The results can be used in reducing the dimension of parametric models.

Some results that are, in a sense, by-products of research in the above main topics are presented in Chapter 6, "Combining estimators." They deal with combining estimators obtained from independent estimating equations. Also, a characterization of multivariate distributions depending on a multivariate location parameter by linearity of the Pitman estimator of a linear function of the parameter is obtained. It is worth noting that the class also contains non-Gaussian distributions, in contrast to the univariate case.

1.2 Outline of results obtained in the dissertation

1. Theorem (2.3.1): We show the superadditivity of the efficient information on a structural parameter $\boldsymbol{\theta}_1$ in the presence of a nuisance parameter $\boldsymbol{\theta}_2$ based on independent estimating functions $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$.

2. Theorem (2.3.2): If $\boldsymbol{\theta}_2^*$ is a $\sqrt{n}-$consistent estimator of the nuisance parameter $\boldsymbol{\theta}_2$ and $\hat{\boldsymbol{\theta}}_1$ is a consistent solution of the estimating equation

$$\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1 \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) = \mathbf{0}, \tag{1.1}$$

where $\hat{\boldsymbol{\Psi}}_1$ is the efficient estimating function for $\boldsymbol{\theta}_1$ in the presence of $\boldsymbol{\theta}_2$ based on $\boldsymbol{\Psi}$, then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\right) \xrightarrow{d} N_r\left(\mathbf{0}, \mathbf{B}^{11}(\boldsymbol{\theta})\right) \tag{1.2}$$

as $n \longrightarrow \infty$.

3. Theorem (3.2.1): A simple modification of the Pitman estimator of a location parameter μ in the presence of a nuisance scale parameter σ is given by

$$\tilde{t}_n = \overline{X} - \frac{S}{n\hat{I}} \sum_{i=1}^{n} \hat{J}_1\left(\frac{X_i - \overline{X}}{S}\right). \tag{1.3}$$

We show that

$$\sqrt{n}\left(\tilde{t}_n - \mu\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\hat{I}}\right), \tag{1.4}$$

where $\sigma^2/\hat{I}$ is the efficient information quantity.

4. Theorem (3.2.3): We construct a modified version of the polynomial Pitman estimator for a location parameter μ in the presence of a nuisance scale parameter σ, given by

$$\tilde{\mu}_n = \overline{X} - \frac{S}{n\hat{I}^{(k)}} \sum_{i=1}^{n} \hat{J}_1^{(k)}\left(\frac{X_i - \overline{X}}{S}\right) \tag{1.5}$$

and show that

$$\sqrt{n}\left(\tilde{\mu}_n - \mu\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\hat{I}^{(k)}}\right) \tag{1.6}$$

as $n \longrightarrow \infty$.

5. Theorem (4.2.3): When the model $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}$ is misspecified and the misspecification can be parameterized through $\boldsymbol{\eta}_n = \mathbf{c}/\sqrt{n}$, the behavior of $\hat{\boldsymbol{\theta}}_n$, the solution to a general estimating equation for the assumed model $\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}$ is

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{B}(\boldsymbol{\theta}), \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right). \tag{1.7}$$

If $\|\boldsymbol{\eta}_n\| = o(1/\sqrt{n})$, then the asymptotic behavior of $\hat{\boldsymbol{\theta}}_n$ is not affected by model misspecification.

6. Theorem (4.3.1): If $\hat{\theta}_n$ is an unbiased estimator of θ when the true distribution is $F(x;\theta)$ or the true distribution is $G(x;\theta)$, then a lower bound for the variance of $\hat{\theta}_n$ when the true distribution is F is given by

$$\mathrm{Var}_1\left(\hat{\theta}_n\right) \geq \frac{1}{nI_1(\theta) - \frac{n^2(E_2 J_1(x))^2}{(1+\Delta_1)^n}} \tag{1.8}$$

for some $\Delta_1 \geq 0$.

7. Theorem (4.3.3): Let $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$ where $\mathcal{P}_1 = \{f_1(x;\theta_1) : \theta_1 \in \Theta_1 \subseteq \mathbb{R}\}$ and $\mathcal{P}_2 = \{f_2(x;\theta_2) : \theta_2 \in \Theta_2 \subseteq \mathbb{R}\}$, and let $\hat{\theta}_1$ be the MLE for θ_1 and $\hat{\theta}_2$ be the QMLE for θ_2. We show

$$P_1\left(\sup_{\theta\in\Theta_1} \prod_{i=1}^{n} f_1(X_i;\theta) > \sup_{\theta\in\Theta_2} \prod_{i=1}^{n} f_2(X_i;\theta)\right) \longrightarrow 1 \tag{1.9}$$

as $n \longrightarrow \infty$ and

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) \xrightarrow{d,P_1} N\left(0, \frac{1}{I_1(\theta)}\right) \tag{1.10}$$

as $n \longrightarrow \infty$, where

$$\hat{\theta}_n = \hat{\theta}_1 I \left\{ \prod_{i=1}^n f_1(X_i; \hat{\theta}_1) > \prod_{i=1}^n f_2(X_i; \hat{\theta}_2) \right\} \\ + \hat{\theta}_2 I \left\{ \prod_{i=1}^n f_1(X_i; \hat{\theta}_1) \leq \prod_{i=1}^n f_2(X_i; \hat{\theta}_2) \right\}. \tag{1.11}$$

8. Theorem (5.2.1), Theorem (5.2.2): We give a version of the Wald test statistic

$$\mathcal{W}_n = n\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n) \left[\frac{\partial \mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\hat{\boldsymbol{\theta}}_n) \frac{\partial \mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} \right]^{-1} \mathbf{R}(\hat{\boldsymbol{\theta}}_n) \tag{1.12}$$

based on an estimating function $\boldsymbol{\Psi}$ for testing the hypothesis $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ and show that the sequence of tests based on $\mathcal{W}_n$ is consistent.

9. Theorem (5.3.3), Theorem (5.3.4), Theorem (5.3.5): We give a version of the Rao score test statistic

$$\mathcal{R}_n = \frac{1}{n} \left(\sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) \right)^T \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n) \left(\sum_{i=1}^n \boldsymbol{\Psi}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) \right) \tag{1.13}$$

based on a Fisher estimating function $\boldsymbol{\Psi}$ for testing the hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$, show that the sequence of tests $\mathcal{R}_n$ is consistent, and prove $\mathcal{R}_n - \mathcal{W}_n = o_p(1)$.

10. Theorem (5.3.7): The limiting distribution of $\mathcal{R}_n$ and $\mathcal{W}_n$ under a linear hypothesis and the form of misspecification described in Chapter 4 is shown to be a non-central χ^2 distribution.

11. Theorem (6.1.1), Theorem (6.1.3): Let $\mathbf{X}$ and $\mathbf{Y}$ be independent random samples whose distributions depend on a common parameter $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_1$ be the solution to the estimating equation $\boldsymbol{\Psi}_1(\mathbf{X}; \boldsymbol{\theta}) = \mathbf{0}$ and $\hat{\boldsymbol{\theta}}_2$ be the solution to the estimating equation $\boldsymbol{\Psi}_2(\mathbf{Y}; \boldsymbol{\theta}) = \mathbf{0}$. We show the best linear combination

of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ is asymptotically as good as the estimator generated from the optimal linear combination of the estimating functions $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$.

12. Theorem (6.2.3), Theorem (6.2.5): We give the form of the Pitman estimator of a linear function of a bivariate location parameter and show that the Pitman estimator is linear if and only if the characteristic function is of the form

$$\phi(t,s) = \exp\{Q(t,s) + h(c_2 t - c_1 s)\} \tag{1.14}$$

for some quadratic form Q and some differentiable function h.

Chapter 2

Estimating functions

2.1 Introduction

Let $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ be a family of distributions of a random element $\mathbf{X}$ taking values in a measurable space $(\mathcal{X}, \mathcal{A})$ depending on a parameter $\boldsymbol{\theta} \in \Theta$. In other words, $P(\mathbf{X} \in A; \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(A),\ A \in \mathcal{A},\ \boldsymbol{\theta} \in \Theta$.

Our goal is to estimate $\boldsymbol{\theta}$ based on our observation $\mathbf{X}$. Estimating equations, introduced by Godambe ([11]), are a useful tool for constructing estimators. The classical maximum likelihood estimator (MLE) and method of moments estimator are special cases of estimators generated by estimating equations. The advantage of the method of moments estimator is that no assumptions about the probability measure need to be made except for the structure of the moments, while the advantage of using maximum likelihood is that the estimator will be optimal. However, to construct the MLE it is necessary to have full distributional specification.

A vector function $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta}) : \mathcal{X} \times \Theta \mapsto \mathbb{R}^s$ is called an *estimating function* for $\mathcal{P}$ if for all $\boldsymbol{\theta} \in \Theta$,

1. $E_{\boldsymbol{\theta}} \boldsymbol{\Psi} = \mathbf{0}$,

2. $E_{\boldsymbol{\theta}} |\Psi_j|^2 < \infty,\ j = 1, \ldots, s$, and

3. the matrix

$$\mathbf{C}_{\mathbf{\Psi}}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left(\frac{\partial \mathbf{\Psi}}{\partial \boldsymbol{\theta}}\right) \tag{2.1}$$

is nonsingular.

In addition, we will assume that the covariance matrix

$$\mathbf{B}_{\mathbf{\Psi}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\mathbf{\Psi}\mathbf{\Psi}^T\right) \tag{2.2}$$

is positive definite for all $\boldsymbol{\theta} \in \Theta$.

Let $\mathcal{G}$ be the set of all estimating functions $g : \mathcal{X} \times \Theta \mapsto \mathbb{R}^s$ for $\mathcal{P}$. The choice of the estimating function $\mathbf{\Psi}$ and the properties of the estimator it generates are part of the theory of estimating functions.

Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a sample from a population $P_{\boldsymbol{\theta}} \in \mathcal{P}$. The justification for the term "estimating function" comes from the fact that under general regularity conditions (e.g. [42], p. 46), there exists a measurable function, that is a statistic, $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, which is a solution to the *estimating equation*

$$\sum_{i=1}^{n} \mathbf{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{2.3}$$

such that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) = \frac{1}{\sqrt{n}}\mathbf{C}_{\mathbf{\Psi}}^{-1}(\boldsymbol{\theta})\sum_{i=1}^{n} \mathbf{\Psi}\left(\mathbf{X}_i; \boldsymbol{\theta}\right) + \mathbf{o}_p(1), \tag{2.4}$$

where $\mathbf{o}_p(1)$ represents a random vector $\mathbf{R}(\mathbf{X}_1, \ldots, \mathbf{X}_n; \boldsymbol{\theta})$ with the property that as $n \longrightarrow \infty$, $\|\mathbf{R}(\mathbf{X}_1, \ldots, \mathbf{X}_n; \boldsymbol{\theta})\| \longrightarrow 0$ in $P_{\boldsymbol{\theta}}$-probability. Hence,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{0}, \mathbf{C}_{\mathbf{\Psi}}^{-1}(\boldsymbol{\theta})\mathbf{B}_{\mathbf{\Psi}}(\boldsymbol{\theta})\mathbf{C}_{\mathbf{\Psi}}^{-1}(\boldsymbol{\theta})^T\right) \tag{2.5}$$

as $n \longrightarrow \infty$. One simple condition for (2.4) to hold is that there exists a function $\mathring{\Psi}(\mathbf{x})$ with $E_{\boldsymbol{\theta}}\mathring{\Psi}^2(\mathbf{x}) < \infty$ such that

$$\|\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}_1) - \boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}_2)\| \leq \mathring{\Psi}(\mathbf{x})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \tag{2.6}$$

for every $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ in a neighborhood of the true value $\boldsymbol{\theta}$.

The matrix

$$\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{C}^T_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \tag{2.7}$$

is called the *information* on $\boldsymbol{\theta}$ associated with the estimating function $\boldsymbol{\Psi}$. This definition was introduced by Bhapkar ([5]) in 1972 based on the work of Godambe ([11]), and can be viewed as a generalization of the Fisher information

$$\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{x};\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}}\log f(\mathbf{x};\boldsymbol{\theta})^T\right). \tag{2.8}$$

Suppose $P_{\boldsymbol{\theta}}$ is absolutely continuous with respect to some sigma-finite measure μ with $dP_{\boldsymbol{\theta}}/d\mu = f(\mathbf{x};\boldsymbol{\theta})$, where $f(\mathbf{x};\boldsymbol{\theta})$ is twice differentiable in $\boldsymbol{\theta}$ and satisfies

$$\frac{\partial}{\partial\boldsymbol{\theta}}\int f(\mathbf{x};\boldsymbol{\theta})\mu(d\mathbf{x}) = \int\frac{\partial}{\partial\boldsymbol{\theta}}f(\mathbf{x};\boldsymbol{\theta})\mu(d\mathbf{x}). \tag{2.9}$$

If the matrix $\mathbf{I}(\boldsymbol{\theta})$ is positive definite for all $\boldsymbol{\theta} \in \Theta$, the Fisher score $\mathbf{J} = \mathbf{J}(\mathbf{x};\boldsymbol{\theta}) = \partial\log f(\mathbf{x};\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is an estimating function which satisfies $\mathbf{B}_{\mathbf{J}}(\boldsymbol{\theta}) = \mathbf{C}_{\mathbf{J}}(\boldsymbol{\theta})$ (see [34], p. 134). In this case $\mathbf{I}_{\mathbf{J}}(\boldsymbol{\theta})$ reduces to $\mathbf{I}(\boldsymbol{\theta})$, so that this definition of the information is consistent with Fisher's definition.

Some of the properties of the information on $\boldsymbol{\theta}$ contained in the estimating function $\boldsymbol{\Psi}$ are as follows:

Lemma 2.1.1. *Let* $\mathbf{S} = \mathbf{S}(\mathbf{X})$ *be sufficient for the family* $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$. *The function*

$$\boldsymbol{\Psi}^*(\mathbf{S}; \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \left(\boldsymbol{\Psi}(\mathbf{X}; \boldsymbol{\theta}) \mid \mathbf{S} \right) \tag{2.10}$$

defines an estimating function for $\mathcal{P}$ *and*

$$\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \leq \mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\theta}). \tag{2.11}$$

Proof. Assume $\theta \in \mathbb{R}$. Clearly $E_\theta\left(\Psi^*\right) = 0$, and $E_\theta\left(\Psi^2\right) = E_\theta\left(\Psi^{*2}\right) + E_\theta \mathrm{Var}_\theta\left(\Psi \mid S\right)$ so that $0 < E_\theta\left(\Psi^{*2}\right) = \mathrm{Var}_\theta\left(\Psi^*\right) < \infty$. Since S is sufficient for θ, the conditional distribution of X given S is independent of θ, so assuming we can interchange the operations of integration and differentiation, we have

$$\begin{aligned} E_\theta\left(\frac{\partial}{\partial\theta}\Psi^*\right) &= E_\theta\left(\frac{\partial}{\partial\theta}E\left(\Psi \mid S\right)\right) = E_\theta\left(E\left(\frac{\partial}{\partial\theta}\Psi \mid S\right)\right) \\ &= E_\theta\left(\frac{\partial}{\partial\theta}\Psi\right) \neq 0 \end{aligned} \tag{2.12}$$

so that Ψ^* is an estimating function for $\mathcal{P}$.

Let $f(x;\theta)$ denote the density function. We can differentiate the identity $0 = E_\theta \Psi(x;\theta)$ with respect to θ (again assuming we can interchange the operations of integration and differentiation) to get

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} E_\theta \Psi(x;\theta) = \frac{\partial}{\partial\theta}\int \Psi(x;\theta) f(x;\theta)\mu(dx) \\ &= \int \frac{\partial}{\partial\theta}\Psi(x;\theta) f(x;\theta)\mu(dx) + \int \Psi(x;\theta)\frac{\partial}{\partial\theta} f(x;\theta)\mu(dx) \\ &= E_\theta\left(\frac{\partial}{\partial\theta}\Psi(x;\theta)\right) + E_\theta\left(\Psi(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)\right) \end{aligned} \tag{2.13}$$

so that

$$-E_\theta\left(\frac{\partial}{\partial\theta}\Psi(x;\theta)\right) = E_\theta\left(\Psi(x;\theta) J(x;\theta)\right) \tag{2.14}$$

where $J(x;\theta)$ is the Fisher score. Similarly, if S has density $q(s;\theta)$,

$$-E_\theta\left(\frac{\partial}{\partial\theta}\Psi^*(s;\theta)\right) = E_\theta\left(\Psi^*(s;\theta)J_q(s;\theta)\right), \tag{2.15}$$

where J_q is the score function corresponding to the density $q(s;\theta)$. Since S is sufficient for θ, the density can be factorized as $f(x;\theta) = g(s(x),\theta)h(x)$, hence

$$J(x;\theta) = \frac{\partial}{\partial\theta}\log f(x;\theta) = \frac{\partial}{\partial\theta}\log g(s(x),\theta). \tag{2.16}$$

We have

$$\begin{aligned} E_\theta\left(\Psi(x;\theta)J(x;\theta)\mid s\right) &= E_\theta\left(\Psi(x;\theta)\frac{\partial}{\partial\theta}\log g(s;\theta)\,\middle|\, s\right) \\ &= \frac{\partial}{\partial\theta}\log g(s;\theta)E_\theta\left(\Psi(x;\theta)\mid s\right) \\ &= E_\theta\left(\frac{\partial}{\partial\theta}\log f(x;\theta)\,\middle|\, s\right)\Psi^*(s;\theta) \\ &= \frac{\partial}{\partial\theta}\log q(s;\theta)\Psi^*(s;\theta). \end{aligned} \tag{2.17}$$

The last equality follows from a well-known property of the Fisher score, which does not rely on the sufficiency of S, but does require mild regularity conditions on the densities. The proof can be found in [32], p. 330. Equations (2.13) - (2.15) give us

$$\begin{aligned} I_\Psi(\theta) = \frac{\left(E_\theta\frac{\partial}{\partial\theta}\Psi(x;\theta)\right)^2}{E_\theta\Psi^2(x;\theta)} &= \frac{\left[E_\theta\left(\Psi(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)\right)\right]^2}{E_\theta\Psi^{*2}(s;\theta) + E_\theta\mathrm{Var}\left(\Psi(x;\theta)\mid s\right)} \\ &\leq \frac{\left[E_\theta E_\theta\left(\Psi(x;\theta)\frac{\partial}{\partial\theta}\log f(x;\theta)\mid s\right)\right]^2}{E_\theta\Psi^{*2}(s;\theta)} \\ &= \frac{\left[E_\theta\left(\Psi^*(s;\theta)\frac{\partial}{\partial\theta}\log q(s;\theta)\right)\right]^2}{E_\theta\Psi^{*2}(s;\theta)} = \frac{\left(E_\theta\frac{\partial}{\partial\theta}\Psi^*(s;\theta)\right)^2}{E_\theta\Psi^{*2}(s;\theta)} \\ &= I_{\Psi^*}(\theta). \end{aligned} \tag{2.18}$$

$\square$

For the case of a multivariate parameter, see [30], p. 37.

A more general monotonicity property holds for the Fisher score $\mathbf{J}(\mathbf{x};\boldsymbol{\theta})$. For any statistic $\mathbf{T} = \mathbf{T}(\mathbf{X})$, the information associated with the estimating function $\mathbf{J}^*(\mathbf{t};\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\mathbf{J}(\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{t}\right)$, $\mathbf{I}_{\mathbf{T}}(\boldsymbol{\theta})$, is less than or equal to the Fisher Information, $\mathbf{I}(\boldsymbol{\theta})$, with equality if and only if $\mathbf{T}$ is sufficient for $\boldsymbol{\theta}$.

For a general estimating function we do not have this property of monotonicity. That is if $\mathbf{T} = \mathbf{T}(\mathbf{X})$ is a statistic and the matrix $E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}} E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi} \mid \mathbf{T}\right)\right)$ is nonsingular then $\boldsymbol{\Psi}^*(\mathbf{t};\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{t}\right)$ will be an estimating function for $\mathcal{P}$, since clearly $E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}^*\right) = \mathbf{0}$, and for each component Ψ_j^*, $E_{\boldsymbol{\theta}}\left(\Psi_j^2\right) = E_{\boldsymbol{\theta}}\left(\Psi_j^{*2}\right) + E_{\boldsymbol{\theta}}\mathrm{Var}_{\boldsymbol{\theta}}\left(\Psi_j \mid T\right)$ so that $E_{\boldsymbol{\theta}}\left(\Psi_j^{*2}\right) \leq E_{\boldsymbol{\theta}}\left(\Psi_j^2\right) < \infty$. We cannot, however, say that $\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \leq \mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\theta})$ or $\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \geq \mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\theta})$.

Lemma 2.1.2. *Let* $\mathbf{U} = \mathbf{U}(\mathbf{X})$ *be an ancillary statistic. That is, the distribution of* $\mathbf{U}$ *does not depend on the parameter* $\boldsymbol{\theta}$*. If* $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ *is an estimating function for* $\boldsymbol{\theta}$*, then so too is*

$$\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^*(\mathbf{x};\boldsymbol{\theta}) = \boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}) - E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}) \mid \mathbf{u}\right) \tag{2.19}$$

and

$$\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \leq \mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\theta}). \tag{2.20}$$

Proof. See [6]. □

There is a close relationship between Lemmas 2.1.1 and 2.1.2. Suppose there exists a complete sufficient statistic $S = S(X)$ for θ. By Basu's Theorem S is independent of every ancillary statistic $U = U(X)$, so that for square integrable functions g and h, $\mathrm{Cov}_\theta\left(g(S), h(U)\right) = 0$. The space of functions of U is a linear subspace of the vector space of all square integrable functions of X. Small and McLeish ([36], [37]) showed that the subspace of functions of S is the orthogonal complement of this linear vector space. Motivated by this observation, they showed that an estimating function can be improved upon by projecting it onto the orthogonal complement of the space of ancillary statistics, even if a complete sufficient statistic does not exist.

Lemma 2.1.3. *Let $\mathbf{X}$ be distributed according to the probability measure $P_{\boldsymbol{\theta}}$ and $\mathbf{Y}$ be distributed according to the probability measure $Q_{\boldsymbol{\theta}}$, where $\mathbf{X}$ and $\mathbf{Y}$ are independent random vectors and $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s$. If $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_1(\mathbf{x};\boldsymbol{\theta})$ is an estimating function for $\boldsymbol{\theta}$ and $\boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_2(\mathbf{y};\boldsymbol{\theta})$ is an estimating function for $\boldsymbol{\theta}$ and the matrix*

$$\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta}) + \mathbf{C}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta}) \tag{2.21}$$

is nonsingular, then

$$\boldsymbol{\Psi}(\mathbf{x},\mathbf{y};\boldsymbol{\theta}) = \boldsymbol{\Psi}_1(\mathbf{x};\boldsymbol{\theta}) + \boldsymbol{\Psi}_2(\mathbf{y};\boldsymbol{\theta}) \tag{2.22}$$

is also an estimating function for $\boldsymbol{\theta}$, and we have

$$\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \leq \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \mathbf{I}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta}). \tag{2.23}$$

If $\mathbf{B}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})$ and $\mathbf{B}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})$ then we have equality in equation (2.23).

Proof. If $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ are estimating functions, then $\boldsymbol{\Psi} = \boldsymbol{\Psi}_1 + \boldsymbol{\Psi}_2$ has zero expectation and components which are square integrable. The covariance matrix

$$\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right) = \mathbf{B}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta}) + \mathbf{B}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta}) \tag{2.24}$$

is positive definite, so that (using assumption (2.21)) $\boldsymbol{\Psi}$ is an estimating function for $\boldsymbol{\theta}$. The information on $\boldsymbol{\theta}$ associated with $\boldsymbol{\Psi}$ is $\mathbf{C}_{\boldsymbol{\Psi}}{}^T(\boldsymbol{\theta})\mathbf{B}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$, so to prove

(2.23) we need to show that

$$\begin{aligned}
&\left(\mathbf{C}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{C}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})\right)^T\left(\mathbf{B}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{B}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})\right)^{-1}\left(\mathbf{C}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{C}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})\right) \\
&\qquad\qquad\qquad \leq \mathbf{C}^T_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\mathbf{C}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{C}^T_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})\mathbf{C}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta}).
\end{aligned} \tag{2.25}$$

This follows from the zero expectation of an estimating function, the independence of $\mathbf{X}$ and $\mathbf{Y}$, and the fact that any covariance matrix is nonnegative definite:

$$\begin{aligned}
\mathbf{0} &\leq E_{\boldsymbol{\theta}}\left(\mathbf{C}^T_{\boldsymbol{\Psi_1}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_1}}\boldsymbol{\Psi_1}+\mathbf{C}^T_{\boldsymbol{\Psi_2}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}\boldsymbol{\Psi_2}-\mathbf{C}^T_{\boldsymbol{\Psi}}\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\boldsymbol{\Psi}\right) \\
&\qquad\qquad \left(\mathbf{C}^T_{\boldsymbol{\Psi_1}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_1}}\boldsymbol{\Psi_1}+\mathbf{C}^T_{\boldsymbol{\Psi_2}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}\boldsymbol{\Psi_2}-\mathbf{C}^T_{\boldsymbol{\Psi}}\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\boldsymbol{\Psi}\right)^T \\
&= \mathbf{C}^T_{\boldsymbol{\Psi_1}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_1}}\mathbf{C}_{\boldsymbol{\Psi_1}}+\mathbf{C}^T_{\boldsymbol{\Psi_2}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}\mathbf{C}_{\boldsymbol{\Psi_2}}+\mathbf{C}^T_{\boldsymbol{\Psi}}\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\mathbf{C}_{\boldsymbol{\Psi}} \\
&\qquad\qquad - E_{\boldsymbol{\theta}}\left(\mathbf{C}^T_{\boldsymbol{\Psi_1}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_1}}\boldsymbol{\Psi_1}+\mathbf{C}^T_{\boldsymbol{\Psi_2}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}\boldsymbol{\Psi_2}\right)\left(\mathbf{C}^T_{\boldsymbol{\Psi}}\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\left(\boldsymbol{\Psi_1}+\boldsymbol{\Psi_2}\right)\right)^T \\
&\qquad\qquad - E_{\boldsymbol{\theta}}\left(\mathbf{C}^T_{\boldsymbol{\Psi}}\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\left(\boldsymbol{\Psi_1}+\boldsymbol{\Psi_2}\right)\right)\left(\mathbf{C}^T_{\boldsymbol{\Psi_1}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_1}}\boldsymbol{\Psi_1}+\mathbf{C}^T_{\boldsymbol{\Psi_2}}\mathbf{B}^{-1}_{\boldsymbol{\Psi_2}}\boldsymbol{\Psi_2}\right)^T \\
&= \mathbf{I}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{I}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})+\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})-2\left(\mathbf{C}_{\boldsymbol{\Psi_1}}+\mathbf{C}_{\boldsymbol{\Psi_2}}\right)^T\mathbf{B}^{-1}_{\boldsymbol{\Psi}}\mathbf{C}_{\boldsymbol{\Psi}} \\
&= \mathbf{I}_{\boldsymbol{\Psi_1}}(\boldsymbol{\theta})+\mathbf{I}_{\boldsymbol{\Psi_2}}(\boldsymbol{\theta})-\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}).
\end{aligned} \tag{2.26}$$

If $\mathbf{B}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})$ and $\mathbf{B}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})$ then also $\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$, and the above inequality becomes an equality. $\square$

The information on $\boldsymbol{\theta}$ associated with the estimating function $\boldsymbol{\Psi}$ can be used as a tool for comparing different estimating functions for the same family $\mathcal{P}$. The estimating function $\boldsymbol{\Psi}_1$ is said to be *more informative* or *better* than the estimating function $\boldsymbol{\Psi}_2$ if

$$\mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) \geq \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta}) \tag{2.27}$$

in the sense that $\mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) - \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})$ is a non-negative definite matrix. Similarly, we say that an estimating function $\boldsymbol{\Psi}$ is *optimal* in a class $\mathcal{C}$ of estimating functions for $\mathcal{P}$ if it is more informative than any function $\tilde{\boldsymbol{\Psi}} \in \mathcal{C}$.

It is well-known and is easily proved that under mild regularity conditions $\mathbf{I}(\boldsymbol{\theta}) \geq \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ for any estimating function $\boldsymbol{\Psi} \in \mathcal{G}$ ([11], [28]), so that the optimal estimating function in the class $\mathcal{G}$ of all estimating functions is the Fisher score. As a second example, let $X \in \mathbb{R}$ with $E(x) = \alpha(\theta)$ and $\mathrm{Var}(x) = \sigma^2(\theta)$ where $\alpha(\theta)$ and $\sigma^2(\theta)$ are known differentiable functions of a scalar-valued parameter θ. The optimal estimating function in the class of linear estimating functions $\Psi(x;\theta) = a(\theta) + b(\theta)x$ is given by

$$\Psi(x;\theta) = \frac{\left(\frac{d}{d\theta}\alpha(\theta)\right)(x - \alpha(\theta))}{\sigma^2(\theta)}. \tag{2.28}$$

This can be proved using (2.30) below.

It is easy to show that the above definition of optimality is equivalent to the following definition: An estimating function $\boldsymbol{\Psi} \in \mathcal{C}$ is optimal if

$$\begin{aligned} &E_{\boldsymbol{\theta}}\left(\mathbf{J} - \mathbf{C}_{\boldsymbol{\Psi}}^T \mathbf{B}_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Psi}\right)\left(\mathbf{J} - \mathbf{C}_{\boldsymbol{\Psi}}^T \mathbf{B}_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Psi}\right)^T \\ &\qquad\qquad \leq E_{\boldsymbol{\theta}}\left(\mathbf{J} - \mathbf{C}_{\tilde{\boldsymbol{\Psi}}}^T \mathbf{B}_{\tilde{\boldsymbol{\Psi}}}^{-1} \tilde{\boldsymbol{\Psi}}\right)\left(\mathbf{J} - \mathbf{C}_{\tilde{\boldsymbol{\Psi}}}^T \mathbf{B}_{\tilde{\boldsymbol{\Psi}}}^{-1} \tilde{\boldsymbol{\Psi}}\right)^T \end{aligned} \tag{2.29}$$

for any $\tilde{\boldsymbol{\Psi}} \in \mathcal{C}$, where $\mathbf{J}$ is the Fisher score. This definition allows for a geometric interpretation of optimality. Suppose $\mathcal{C}$ is a closed subspace of the Hilbert space $L^2(P_{\boldsymbol{\theta}})$ (with inner product $(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)_{\boldsymbol{\theta}} = E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_1^T \boldsymbol{\Psi}_2\right)$ and norm $\|\boldsymbol{\Psi}\|_{\boldsymbol{\theta}} = (\boldsymbol{\Psi}, \boldsymbol{\Psi})_{\boldsymbol{\theta}}^{1/2}$). As noted above, the optimal estimating function in the class of all estimating functions $\mathcal{G}$ is the Fisher score $\mathbf{J}$. The optimal estimating function in the closed linear span of a subset of $\mathcal{G}$ is the estimating function $\boldsymbol{\Psi}$ for which $\mathbf{C}_{\boldsymbol{\Psi}}^T \mathbf{B}_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Psi}$ is closest to the Fisher score. That is, the optimal estimating function in $\mathcal{C}$ is

$$\boldsymbol{\Psi} = \hat{E}_{\boldsymbol{\theta}}\left(\mathbf{J} \mid \mathcal{C}\right), \tag{2.30}$$

the orthogonal projection of the Fisher score into the space $\mathcal{C}$ (see [15] for a full proof).

2.2 Fisher estimating functions

Let $\mathbf{A} = \mathbf{A}(\boldsymbol{\theta})$ be an $s \times s$ nonsingular matrix whose elements $a_{ij}(\boldsymbol{\theta})$ are differentiable. If $\boldsymbol{\Psi}$ is an estimating function for $\mathcal{P}$, then so is $\boldsymbol{\Phi} = \mathbf{A}\boldsymbol{\Psi}$. The system

$$\sum_{i=1}^{n} \boldsymbol{\Phi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0} \tag{2.31}$$

is equivalent to

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{2.32}$$

Since equation (2.31) and equation (2.32) will have the same solution, we say that $\boldsymbol{\Phi}$ is *equivalent* to $\boldsymbol{\Psi}$. Thus, to choose $\boldsymbol{\Phi}$ as an estimating function over $\boldsymbol{\Psi}$ is simply a matter of preference.

In what follows we will show there exists a special matrix $\mathbf{A}$ corresponding to the Fisher form of the estimating function equivalent to $\boldsymbol{\Psi}$. We call $\boldsymbol{\Psi}$ a *Fisher estimating function* for $\mathcal{P}$ if

$$\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}). \tag{2.33}$$

The motivation for this definition is that Fisher estimating functions and the estimators they generate share a crucial property of the Fisher score and the maximum likelihood estimator.

If $\boldsymbol{\Psi}$ is an estimating function for $\mathcal{P}$ and the elements of $\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ and $\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ are differentiable, then $\hat{\boldsymbol{\Psi}} = \mathbf{C}_{\boldsymbol{\Psi}}^T \mathbf{B}_{\boldsymbol{\Psi}}^{-1} \boldsymbol{\Psi}$ is a Fisher estimating function equivalent to

$\boldsymbol{\Psi}$, since

$$\mathbf{B}_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\Psi}}\hat{\boldsymbol{\Psi}}^T\right) = \mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right)\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}} = \mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}} \tag{2.34}$$

and

$$\mathbf{C}_{\hat{\boldsymbol{\Psi}}}(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}\right) = -\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\Psi}\right) = \mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}}. \tag{2.35}$$

We call $\hat{\boldsymbol{\Psi}}$ the *Fisher form* of $\boldsymbol{\Psi}$. This definition coincides with Heyde's definition of the standardized form of the estimating function $\boldsymbol{\Psi}$ used in equation (2.30).

Theorem 2.2.1. *If the matrices $\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ and $\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ are nonsingular with entries that are differentiable in $\boldsymbol{\theta}$, there exists a unique Fisher form $\hat{\boldsymbol{\Psi}}$ of the estimating function $\boldsymbol{\Psi}$.*

Proof. Suppose there exist two Fisher estimating functions, $\hat{\boldsymbol{\Psi}}_1 = \mathbf{A}_1\boldsymbol{\Psi}$ and $\hat{\boldsymbol{\Psi}}_2 = \mathbf{A}_2\boldsymbol{\Psi}$, which are both equivalent to $\boldsymbol{\Psi}$. Then $\hat{\boldsymbol{\Psi}}_2 = \mathbf{A}_2\mathbf{A}_1^{-1}\hat{\boldsymbol{\Psi}}_1$. Since $\hat{\boldsymbol{\Psi}}_1$ is a Fisher estimating function,

$$\mathbf{C}_{\hat{\boldsymbol{\Psi}}_2}(\boldsymbol{\theta}) = \mathbf{A}_2(\boldsymbol{\theta})\mathbf{A}_1^{-1}(\boldsymbol{\theta})\mathbf{C}_{\hat{\boldsymbol{\Psi}}_1}(\boldsymbol{\theta}) = \mathbf{A}_2(\boldsymbol{\theta})\mathbf{A}_1^{-1}(\boldsymbol{\theta})\mathbf{B}_{\hat{\boldsymbol{\Psi}}_1}(\boldsymbol{\theta}). \tag{2.36}$$

Since $\hat{\boldsymbol{\Psi}}_2$ is also assumed to be a Fisher estimating function,

$$\mathbf{C}_{\hat{\boldsymbol{\Psi}}_2}(\boldsymbol{\theta}) = \mathbf{B}_{\hat{\boldsymbol{\Psi}}_2}(\boldsymbol{\theta}) = \mathbf{A}_2(\boldsymbol{\theta})\mathbf{A}_1^{-1}\mathbf{B}_{\hat{\boldsymbol{\Psi}}_1}(\boldsymbol{\theta})\mathbf{A}_1^{-1}(\boldsymbol{\theta})^T\mathbf{A}_2(\boldsymbol{\theta})^T. \tag{2.37}$$

Therefore, $\mathbf{A}_1^{-1}(\boldsymbol{\theta})^T\mathbf{A}_2(\boldsymbol{\theta})^T = \mathbf{I}_{s\times s}$, or $\mathbf{A}_1(\boldsymbol{\theta}) = \mathbf{A}_2(\boldsymbol{\theta})$. □

2.3 Structural and nuisance parameters

Let $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$ where $\boldsymbol{\theta}_1 \in \Theta_1 \subseteq \mathbb{R}^r$ is the parameter of interest, $\boldsymbol{\theta}_2 \in \Theta_2 \subseteq \mathbb{R}^q$ is a nuisance parameter, and $r + q = s$. Consider the set $\mathcal{G}_1$ of all estimating functions $g : \mathcal{X} \times \Theta_1 \mapsto \mathbb{R}^r$ for $\boldsymbol{\theta}_1$. Chandresekar and Kale ([10]) showed that for any $\boldsymbol{\Psi}$ in $\mathcal{G}_1$,

$$\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \geq \mathbf{I}^{11}(\boldsymbol{\theta}) \tag{2.38}$$

where $\mathbf{I}^{11}(\boldsymbol{\theta})$ is the $r \times r$ submatrix of the inverse of the Fisher information matrix $\mathbf{I}^{-1}(\boldsymbol{\theta})$ corresponding to $\boldsymbol{\theta}_1$ when $\mathbf{I}^{-1}(\boldsymbol{\theta})$ is partitioned according to the partition of $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$.

It was shown by Godambe and Thompson that under certain regularity conditions, there exists an optimal estimating function in $\mathcal{G}_1$ ([12], [14]). Suppose there exists a statistic $(\mathbf{S}, \mathbf{U})$ which is sufficient for $\boldsymbol{\theta}$. If the sufficient statistic has the property that the conditional distribution of $\mathbf{S}$ given $\mathbf{U}$, $h(\mathbf{s}; \boldsymbol{\theta}_1|\mathbf{u})$, depends on $\boldsymbol{\theta}$ only through $\boldsymbol{\theta}_1$, and the family of probability distributions $\{P_{\boldsymbol{\theta}}^{(\mathbf{U})}\}$ of $\mathbf{U}$ is complete for every fixed $\boldsymbol{\theta}_1$, then the conditional Fisher score function

$$\mathbf{l}\,(\mathbf{x}; \boldsymbol{\theta}_1) = \frac{\partial}{\partial \boldsymbol{\theta}_1} \log h(\mathbf{s}, \boldsymbol{\theta}_1|\mathbf{u}) \tag{2.39}$$

will be the optimal estimating function in $\mathcal{G}_1$ for $\boldsymbol{\theta}_1$.

We use the method of projection ([27], [37]) to minimize the effects of the nuisance parameter $\boldsymbol{\theta}_2$ when estimating the structural parameter $\boldsymbol{\theta}_1$ using a general estimating function. Let $\boldsymbol{\Psi}$ be a Fisher estimating function for $\mathcal{P} = \{P_{\boldsymbol{\theta}}\}$. We partition $\boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta})$ as,

$$\boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\Psi}_1(\mathbf{x}; \boldsymbol{\theta}) \\ \boldsymbol{\Psi}_2(\mathbf{x}; \boldsymbol{\theta}) \end{bmatrix}, \tag{2.40}$$

$\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ as,

$$\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{B}_{11}(\boldsymbol{\theta}) & \mathbf{B}_{12}(\boldsymbol{\theta}) \\ \mathbf{B}_{21}(\boldsymbol{\theta}) & \mathbf{B}_{22}(\boldsymbol{\theta}) \end{bmatrix}, \tag{2.41}$$

and $\mathbf{B}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})$ as

$$\mathbf{B}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) = \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{B}^{11}(\boldsymbol{\theta}) & \mathbf{B}^{12}(\boldsymbol{\theta}) \\ \\ \mathbf{B}^{21}(\boldsymbol{\theta}) & \mathbf{B}^{22}(\boldsymbol{\theta}) \end{bmatrix} \tag{2.42}$$

with dimensions corresponding to the partition of $\boldsymbol{\theta}$.

Let $\hat{\boldsymbol{\Psi}}_1 = \boldsymbol{\Psi}_1 - \hat{E}_{\boldsymbol{\theta}}(\boldsymbol{\Psi}_1 \,\big|\, \boldsymbol{\Psi}_2)$, where $\hat{E}_{\boldsymbol{\theta}}(\,\bullet \,\big|\, \boldsymbol{\Psi}_2)$ is the projection operator onto the space spanned by the components of $\boldsymbol{\Psi}_2$. An element in the space spanned by the components of $\boldsymbol{\Psi}_2$ can be written as $\mathbf{C}(\boldsymbol{\theta})\boldsymbol{\Psi}_2$, so to find $\hat{E}_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_1 \mid \boldsymbol{\Psi}_2\right)$, we need to find the matrix $\mathbf{C}(\boldsymbol{\theta})$ such that

$$E_{\boldsymbol{\theta}}\left((\boldsymbol{\Psi}_1 - \mathbf{C}(\boldsymbol{\theta})\boldsymbol{\Psi}_2)^T \boldsymbol{\Psi}_2\right) = 0. \tag{2.43}$$

We can set $\mathbf{C}(\boldsymbol{\theta}) = \mathbf{B}_{12}(\boldsymbol{\theta})\mathbf{B}_{22}^{-1}(\boldsymbol{\theta})$ to attain equality (2.43).

The function $\hat{\boldsymbol{\Psi}}_1 = \boldsymbol{\Psi}_1 - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\boldsymbol{\Psi}_2$ is sometimes called the *efficient estimating function* for $\boldsymbol{\theta}_1$ in the presence of $\boldsymbol{\theta}_2$ based on the estimating function $\boldsymbol{\Psi}$. This is due to the fact that

$$\begin{aligned} \hat{\mathbf{I}}_{1,\boldsymbol{\Psi}}(\boldsymbol{\theta}) &= E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\Psi}}_1 \hat{\boldsymbol{\Psi}}_1^T\right) = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \boldsymbol{\theta}_1}\hat{\boldsymbol{\Psi}}_1\right) \\ &= \mathbf{B}_{11}(\boldsymbol{\theta}) - \mathbf{B}_{12}(\boldsymbol{\theta})\mathbf{B}_{22}^{-1}(\boldsymbol{\theta})\mathbf{B}_{21}(\boldsymbol{\theta}) = \left(\mathbf{B}^{11}(\boldsymbol{\theta})\right)^{-1}, \end{aligned} \tag{2.44}$$

the inverse of the first block of matrix (2.42). We call the matrix $\hat{\mathbf{I}}_{1,\boldsymbol{\Psi}}(\boldsymbol{\theta})$ the *efficient information* on $\boldsymbol{\theta}_1$ contained in the estimating function $\boldsymbol{\Psi}$.

The next Theorem generalizes a result by Kagan and Rao ([27]) concerning the efficient score function.

Theorem 2.3.1. *(Superadditivity) Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be independent random vectors and* $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_1(\mathbf{X};\boldsymbol{\theta})$ *and* $\boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_2(\mathbf{Y};\boldsymbol{\theta})$ *be Fisher estimating functions for the* *parameter* $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$. *If* $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) = \boldsymbol{\Psi}_1(\mathbf{X}; \boldsymbol{\theta}) + \boldsymbol{\Psi}_2(\mathbf{Y}; \boldsymbol{\theta})$,

$$\hat{\mathbf{I}}_{1,\boldsymbol{\Phi}}(\boldsymbol{\theta}) \geq \hat{\mathbf{I}}_{1,\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \hat{\mathbf{I}}_{1,\boldsymbol{\Psi}_2}(\boldsymbol{\theta}). \tag{2.45}$$

If $\mathbf{X}$ *and* $\mathbf{Y}$ *are identically distributed and* $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2$ *then* (2.45) *becomes an equality.*

Proof. We partition the estimating functions

$$\boldsymbol{\Psi}_1 = \begin{bmatrix} \boldsymbol{\Psi}_{11} \\ \\ \boldsymbol{\Psi}_{12} \end{bmatrix}, \quad \boldsymbol{\Psi}_2 = \begin{bmatrix} \boldsymbol{\Psi}_{21} \\ \\ \boldsymbol{\Psi}_{22} \end{bmatrix} \tag{2.46}$$

with dimensions corresponding to the dimension of the partition of $\boldsymbol{\theta}$. The efficient information on $\boldsymbol{\theta}_1$ contained in $\boldsymbol{\Phi}$ is given by

$$\begin{aligned}
\hat{\mathbf{I}}_{1,\boldsymbol{\Phi}}(\theta) &= E_{\boldsymbol{\theta}}\left(\boldsymbol{\Phi}_1 - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Phi}_2\right)\left(\boldsymbol{\Phi}_1 - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Phi}_2\right)^T \\
&= E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} + \boldsymbol{\Psi}_{21} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\left(\boldsymbol{\Psi}_{12} + \boldsymbol{\Psi}_{22}\right)\right) \\
&\qquad \left(\boldsymbol{\Psi}_{11} + \boldsymbol{\Psi}_{21} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\left(\boldsymbol{\Psi}_{12} + \boldsymbol{\Psi}_{22}\right)\right)^T \\
&= E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Psi}_{12}\right)\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Psi}_{12}\right)^T \\
&\qquad + E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{21} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Psi}_{22}\right)\left(\boldsymbol{\Psi}_{21} - \mathbf{I}_{12,\boldsymbol{\Phi}}\mathbf{I}_{22,\boldsymbol{\Phi}}^{-1}\boldsymbol{\Psi}_{22}\right)^T \\
&\geq E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Psi}_1}\mathbf{I}_{22,\boldsymbol{\Psi}_1}^{-1}\boldsymbol{\Psi}_{12}\right)\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Psi}_1}\mathbf{I}_{22,\boldsymbol{\Psi}_1}^{-1}\boldsymbol{\Psi}_{12}\right)^T \\
&\qquad + E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{21} - \mathbf{I}_{12,\boldsymbol{\Psi}_2}\mathbf{I}_{22,\boldsymbol{\Psi}_2}^{-1}\boldsymbol{\Psi}_{22}\right)\left(\boldsymbol{\Psi}_{21} - \mathbf{I}_{22,\boldsymbol{\Psi}_2}\mathbf{I}_{22,\boldsymbol{\Psi}_2}^{-1}\boldsymbol{\Psi}_{22}\right)^T \\
&= E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\Psi}}_1\hat{\boldsymbol{\Psi}}_1^T\right) + E_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\Psi}}_2\hat{\boldsymbol{\Psi}}_2^T\right) = \hat{\mathbf{I}}_{1,\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \hat{\mathbf{I}}_{1,\boldsymbol{\Psi}_2}(\boldsymbol{\theta}).
\end{aligned} \tag{2.47}$$

The inequality follows from the fact that $\hat{E}_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} \mid \boldsymbol{\Psi}_{12}\right) = \mathbf{I}_{12,\boldsymbol{\Psi}_1}(\theta)\mathbf{I}_{22,\boldsymbol{\Psi}_1}^{-1}(\theta)\boldsymbol{\Psi}_{12}$ which implies that for any $r \times q$ matrix $\mathbf{A}(\boldsymbol{\theta})$,

$$\begin{aligned}
&E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} - \mathbf{A}(\boldsymbol{\theta})\boldsymbol{\Psi}_{12}\right)\left(\boldsymbol{\Psi}_{11} - \mathbf{A}(\boldsymbol{\theta})\boldsymbol{\Psi}_{12}\right)^T \\
&\quad \geq E_{\boldsymbol{\theta}}\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\mathbf{I}_{22,\boldsymbol{\Psi}_1}^{-1}(\boldsymbol{\theta})\boldsymbol{\Psi}_{12}\right)\left(\boldsymbol{\Psi}_{11} - \mathbf{I}_{12,\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\mathbf{I}_{22,\boldsymbol{\Psi}_1}^{-1}(\boldsymbol{\theta})\boldsymbol{\Psi}_{12}\right)^T.
\end{aligned} \tag{2.48}$$

If $\mathbf{\Psi}_1 = \mathbf{\Psi}_2$ and $\mathbf{X}$ and $\mathbf{Y}$ are identically distributed, then

$$\begin{aligned}
\mathbf{I}_{12,\mathbf{\Phi}}(\boldsymbol{\theta})\mathbf{I}_{22,\mathbf{\Phi}}^{-1}(\boldsymbol{\theta}) &= \left(\mathbf{I}_{12,\mathbf{\Psi}_1}(\boldsymbol{\theta}) + \mathbf{I}_{12,\mathbf{\Psi}_2}(\boldsymbol{\theta})\right)\left(\mathbf{I}_{22,\mathbf{\Psi}_1}(\boldsymbol{\theta}) + \mathbf{I}_{22,\mathbf{\Psi}_2}(\boldsymbol{\theta})\right)^{-1} \\
&= \mathbf{I}_{12,\mathbf{\Psi}_1}(\boldsymbol{\theta})\mathbf{I}_{22,\mathbf{\Psi}_1}^{-1}(\boldsymbol{\theta}) \qquad (2.49) \\
&= \mathbf{I}_{12,\mathbf{\Psi}_2}(\boldsymbol{\theta})\mathbf{I}_{22,\mathbf{\Psi}_2}^{-1}(\boldsymbol{\theta}).
\end{aligned}$$

Therefore we have equality in equation (2.45). □

In general, $\hat{\mathbf{\Psi}}_1 = \hat{\mathbf{\Psi}}_1(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ depends on both the structural and the nuisance parameter. For fixed $\boldsymbol{\theta}_2$, $\hat{\mathbf{\Psi}}_1$ is a Fisher estimating function for $\boldsymbol{\theta}_1$, and the equation

$$\sum_{i=1}^{n} \hat{\mathbf{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2\right) = \mathbf{0} \qquad (2.50)$$

will have a solution $\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\theta}}_1(\mathbf{X}_1, \ldots, \mathbf{X}_n, \boldsymbol{\theta}_2)$. The advantage of this method is that when we have an initial $\sqrt{n}-$consistent estimate $\boldsymbol{\theta}_2^* = \boldsymbol{\theta}_2^*(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ of $\boldsymbol{\theta}_2$, $\hat{\boldsymbol{\theta}}_1$ will be efficient.

Theorem 2.3.2. *Suppose for each j,*

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\mathbf{\Psi}}_1}{\partial \boldsymbol{\theta}_j}(\mathbf{X}_i; \boldsymbol{\theta}) - E\left(\frac{\partial \hat{\mathbf{\Psi}}_1}{\partial \boldsymbol{\theta}_j}(\mathbf{x}; \boldsymbol{\theta})\right) \right\| = o_p(1) \qquad (2.51)$$

as $n \longrightarrow \infty$, and $\mathbf{m}_j(\boldsymbol{\theta}) = E\left(\partial \hat{\mathbf{\Psi}}_1 / \partial \boldsymbol{\theta}_j\right)$ is continuous in $\boldsymbol{\theta}$. If $\boldsymbol{\theta}_2^*$ is an initial estimator of $\boldsymbol{\theta}_2$ such that $\sqrt{n}(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2) = \mathbf{O}_p(1)$, for all large n, there exists a consistent solution $\hat{\boldsymbol{\theta}}_1$ of the equation

$$\sum_{i=1}^{n} \hat{\mathbf{\Psi}}_1(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*) = \mathbf{0} \qquad (2.52)$$

such that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \xrightarrow{d} N_r\left(\mathbf{0}, \mathbf{B}^{11}(\boldsymbol{\theta})\right) \qquad (2.53)$$

as $n \longrightarrow \infty$.

Proof. Let $\boldsymbol{\theta}_0^T = \left(\boldsymbol{\theta}_{10}^T, \boldsymbol{\theta}_{20}^T\right)$ denote the true value of the parameter. Since $\boldsymbol{\theta}_2^*$ is a consistent estimator of $\boldsymbol{\theta}_{20}$ we have for some $\tilde{\boldsymbol{\theta}}_2$ between $\boldsymbol{\theta}_2^*$ and $\boldsymbol{\theta}_{20}$,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}\right) + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2}\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}_2\right)\left(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2\right)$$
$$\xrightarrow{p} E_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\Psi}}_1\left(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}\right)\right) \tag{2.54}$$

as $n \longrightarrow \infty$. In particular,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}_{10}, \boldsymbol{\theta}_2^*\right) \xrightarrow{p} \mathbf{0} \tag{2.55}$$

as $n \longrightarrow \infty$. By the continuity of $\mathbf{m}_1(\boldsymbol{\theta})$ and (2.51),

$$\begin{aligned}\left\|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_1}\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) - \mathbf{m}_1(\boldsymbol{\theta}_0)\right\| &\leq \left\|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_1}\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) - \mathbf{m}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right)\right\| \\ &+ \left\|\mathbf{m}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) - \mathbf{m}_1(\boldsymbol{\theta}_0)\right\| \\ &\leq \sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}\right) - \mathbf{m}_1(\boldsymbol{\theta})\right\| + \left\|\mathbf{m}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right) - \mathbf{m}_1(\boldsymbol{\theta}_0)\right\| \\ &\xrightarrow{p} \left\|\mathbf{m}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}\right) - \mathbf{m}_1\left(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}\right)\right\|\end{aligned} \tag{2.56}$$

as $n \longrightarrow \infty$. Fix $r > 0$. Let $\lambda = 1/2\left\|\mathbf{m}_1^{-1}(\boldsymbol{\theta}_0)\right\|^{-1}$. For some $s < r$,

$$\left\|\mathbf{m}_1\left(\boldsymbol{\theta}_1, \boldsymbol{\theta}_{20}\right) - \mathbf{m}_1\left(\boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{20}\right)\right\| < \lambda \tag{2.57}$$

for all $\boldsymbol{\theta}_1 \in \mathbf{B}_s(\boldsymbol{\theta}_{10})$. By the Inverse Function Theorem (see [33], p. 193 and [20], Lemma 2), $\mathbf{B}_{\lambda s}\left(\mathbf{M}_n(\boldsymbol{\theta}_{10})\right) \subseteq \mathbf{M}_n\left(\mathbf{B}_s(\boldsymbol{\theta}_{10})\right)$, where

$$\mathbf{M}_n(\boldsymbol{\theta}_1) = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2^*\right). \tag{2.58}$$

Since $\mathbf{M}_n(\boldsymbol{\theta}_{10}) \xrightarrow{p} \mathbf{0}$, for sufficiently large n, $\mathbf{0} \in \mathbf{B}_{\lambda s}(\mathbf{M}_n(\boldsymbol{\theta}_{10})) \subseteq \mathbf{M}_n(\mathbf{B}_s(\boldsymbol{\theta}_{10}))$. That is, for sufficiently large n, there is a $\hat{\boldsymbol{\theta}}_n \in \mathbf{B}_s(\boldsymbol{\theta}_{10}) \subset \mathbf{B}_r(\boldsymbol{\theta}_{10})$ such that $\mathbf{M}_n\left(\hat{\boldsymbol{\theta}}_n\right) = \mathbf{0}$. Since $r > 0$ is arbitrary, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_{10}$ as $n \longrightarrow \infty$.

To show (2.53) we can expand (2.52) in a Taylor series around the point $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ (dropping the extra subscript) to get

$$\begin{aligned}
\mathbf{0} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2^*\right) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) + \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_1}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) \\
&\qquad + \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2)
\end{aligned} \tag{2.59}$$

for some $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ between $(\hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2^*)$ and $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. By (2.51),

$$\begin{aligned}
\left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_j}\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}\right) - \mathbf{m}_j(\boldsymbol{\theta}) \right\| &\leq \sup_{\boldsymbol{\theta}} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_j}(\mathbf{X}_i; \boldsymbol{\theta}) - \mathbf{m}_j(\boldsymbol{\theta}) \right\| \\
&\qquad + \left\| \mathbf{m}_j\left(\tilde{\boldsymbol{\theta}}\right) - \mathbf{m}_j(\boldsymbol{\theta}) \right\| = o_p(1)
\end{aligned} \tag{2.60}$$

since $(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2)$ converges in probability to $\boldsymbol{\theta}$ and $\mathbf{m}_j(\boldsymbol{\theta})$ is assumed continuous. Therefore,

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\theta}}_2) &\xrightarrow{p} E_{\boldsymbol{\theta}}\left(\frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2}(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right) \\
&= E_{\boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\Psi}_1}{\partial \boldsymbol{\theta}_2}(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right) - \mathbf{B}_{12}\mathbf{B}_{22}^{-1} E_{\boldsymbol{\theta}}\left(\frac{\partial \boldsymbol{\Psi}_2}{\partial \boldsymbol{\theta}_2}(\mathbf{x}; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right) \\
&= -\mathbf{C}_{12} + \mathbf{B}_{12}\mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{C}_{22} = -\mathbf{B}_{12} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{22} \\
&= \mathbf{0}.
\end{aligned} \tag{2.61}$$

Then, since $\sqrt{n}\,(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2) = \mathbf{O}_p(1)$, the last term in (2.59) is $\mathbf{o}_p(1)$. Similarly,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{\boldsymbol{\Psi}}_1}{\partial\boldsymbol{\theta}_1}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_1,\tilde{\boldsymbol{\theta}}_2) &\xrightarrow{p} E_{\boldsymbol{\theta}}\left(\frac{\partial\hat{\boldsymbol{\Psi}}_1}{\partial\boldsymbol{\theta}_1}(\mathbf{X}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right) \\
&= E_{\boldsymbol{\theta}}\left(\frac{\partial\boldsymbol{\Psi}_1}{\partial\boldsymbol{\theta}_1}(\mathbf{x};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right) - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}E_{\boldsymbol{\theta}}\left(\frac{\partial\boldsymbol{\Psi}_2}{\partial\boldsymbol{\theta}_1}(\mathbf{x};\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\right) \\
&= -\mathbf{C}_{11} + \mathbf{B}_{12}\mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{C}_{21} = -\mathbf{B}_{11} + \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21} \\
&= -(\mathbf{B}^{11})^{-1}.
\end{aligned} \tag{2.62}$$

Rearranging (2.59) gives

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) &= \frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{\boldsymbol{\Psi}}_1}{\partial\boldsymbol{\theta}_1}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_1,\tilde{\boldsymbol{\theta}}_2)\right)^{-1}\sum_{i=1}^{n}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) + \mathbf{o}_p(1) \\
&= -\mathbf{B}^{11}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) + \mathbf{o}_p(1).
\end{aligned} \tag{2.63}$$

proving (2.53). □

Theorem 2.3.2 requires that our initial estimator $\boldsymbol{\theta}_2^*$ of $\boldsymbol{\theta}_2$ is $\sqrt{n}$−consistent. The assumption can be relaxed under further moment assumptions on $\partial\hat{\boldsymbol{\Psi}}_1/\partial\boldsymbol{\theta}_2$.

Theorem 2.3.3. *Suppose*

$$E_{\boldsymbol{\theta}}\left\|\frac{\partial\hat{\boldsymbol{\Psi}}_1}{\partial\boldsymbol{\theta}_2}\right\|^{\alpha} < \infty \tag{2.64}$$

for some $\alpha \in [1,2)$ and $\boldsymbol{\theta}_2^*$ is an initial estimator of $\boldsymbol{\theta}_2$ such that

$$(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2) = \mathbf{O}_p\left(\frac{1}{n^{1/\alpha-1/2}}\right). \tag{2.65}$$

If $\hat{\boldsymbol{\theta}}_1$ is a solution of the equation

$$\sum_{i=1}^{n}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}_1,\boldsymbol{\theta}_2^*) = \mathbf{0} \tag{2.66}$$

such that

$$\mathbf{0} = \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1 \left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_1, \boldsymbol{\theta}_2^* \right) = \sum_{i=1}^{n} \hat{\boldsymbol{\Psi}}_1 \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) + \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2} \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) \left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1 \right)$$
$$+ \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}}{\partial \boldsymbol{\theta}_2} \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) \left(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2 \right) + \mathbf{o}_p(1) \tag{2.67}$$

then

$$\sqrt{n} \left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta} \right) \xrightarrow{d} N_r \left(\mathbf{0}, \mathbf{B}^{11} \left(\boldsymbol{\theta} \right) \right) \tag{2.68}$$

as $n \longrightarrow \infty$.

Proof. Since $E_{\boldsymbol{\theta}} \left(\partial \hat{\boldsymbol{\Psi}}_1 / \partial \boldsymbol{\theta}_2 \right) = \mathbf{0}$ and $E_{\boldsymbol{\theta}} \left\| \partial \hat{\boldsymbol{\Psi}}_1 / \partial \boldsymbol{\theta}_2 \right\|^{\alpha} < \infty$, by the Marcinkiewz-Zygmund law of large numbers (see [4], p. 256),

$$\frac{1}{n^{1/\alpha}} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2} \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) \xrightarrow{p} \mathbf{0} \tag{2.69}$$

as $n \longrightarrow \infty$. Since $\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2 = \mathbf{O}_p(n^{-1/\alpha+1/2})$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2} \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) \left(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2 \right)$$
$$= \frac{1}{n^{1/\alpha}} \sum_{i=1}^{n} \frac{\partial \hat{\boldsymbol{\Psi}}_1}{\partial \boldsymbol{\theta}_2} \left(\mathbf{X}_i; \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \right) n^{1/\alpha - 1/2} \left(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2 \right) = \mathbf{o}_p(1) \mathbf{O}_p(1) \tag{2.70}$$
$$= \mathbf{o}_p(1)$$

as $n \longrightarrow \infty$. The remainder of the proof is identical to the proof of Theorem 2.3.2. □

The next example, due to Kagan and Rao ([27]), shows that the efficient estimating function can be a useful tool in calculating a statistic for the full parameter $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$. Consider the standard linear regression model

$$\mathbf{X} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2.71}$$

where $\mathbf{X}$ is an observable $n \times 1$ random vector, $\mathbf{A}$ is a known design matrix of order $n \times s$ of full rank $s \leq n$, and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of errors with $E_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Var}_{\boldsymbol{\theta}}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_{n \times n}$, σ^2 unknown.

The least-squares estimator of $\boldsymbol{\theta}$,

$$\hat{\boldsymbol{\theta}}_n = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{X} \tag{2.72}$$

is the solution of the estimating equation

$$\boldsymbol{\Psi}(\mathbf{X}; \boldsymbol{\theta}) = \mathbf{A}^T (\mathbf{X} - \mathbf{A}\boldsymbol{\theta}) = \mathbf{0} \tag{2.73}$$

and is the best linear unbiased estimator (BLUE) of $\boldsymbol{\theta}$ in the sense that for any linear unbiased estimator $\tilde{\boldsymbol{\theta}}_n = \mathbf{B}\mathbf{X}$,

$$\mathrm{Var}_{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n\right) = \sigma^2 \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \leq \mathrm{Var}_{\boldsymbol{\theta}}\left(\tilde{\boldsymbol{\theta}}_n\right). \tag{2.74}$$

The best linear estimating function for $\boldsymbol{\theta}$ is

$$\mathbf{J}_{\text{lin}} = \hat{E}_{\boldsymbol{\theta}}(\mathbf{J} \mid \mathbf{X}) = \frac{1}{\sigma^2} \mathbf{A}^T (\mathbf{X} - \mathbf{A}\boldsymbol{\theta}), \tag{2.75}$$

which is the Fisher form of the estimating function in equation (2.73). We partition $\mathbf{A} = (\mathbf{A}_1 \ \mathbf{A}_2)$ and $\mathbf{J}_{\text{lin}}^T = \left(\mathbf{J}_{1,\text{lin}}^T, \mathbf{J}_{2,\text{lin}}^T\right)$ according to the partition of $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T)$. The information on $\boldsymbol{\theta}$ contained in the estimating function $\mathbf{J}_{\text{lin}}$ is

$$\mathbf{I}_{\text{lin}} = E_{\boldsymbol{\theta}}\left(\mathbf{J}_{\text{lin}} \mathbf{J}_{\text{lin}}^T\right) = \frac{1}{\sigma^2} \mathbf{A}^T \mathbf{A} = \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{A}_1^T \mathbf{A}_1 & \mathbf{A}_1^T \mathbf{A}_2 \\ \mathbf{A}_2^T \mathbf{A}_1 & \mathbf{A}_2^T \mathbf{A}_2 \end{pmatrix} = \frac{1}{\sigma^2} \begin{pmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{pmatrix} \tag{2.76}$$

which is independent of $\boldsymbol{\theta}$.

The efficient estimating function for $\boldsymbol{\theta}_1$ in the presence of $\boldsymbol{\theta}_2$ based on the estimating function $\mathbf{J}_{\text{lin}}$ is

$$\begin{aligned}
\hat{\mathbf{J}}_{1,\text{lin}} &= \mathbf{J}_{1,\text{lin}} - \hat{E}_{\boldsymbol{\theta}}\left(\mathbf{J}_{1,\text{lin}} \mid \mathbf{J}_{2,\text{lin}}\right) \\
&= \frac{1}{\sigma^2}\mathbf{A}_1^T\left(\mathbf{X}-\mathbf{A}\boldsymbol{\theta}\right) - \frac{1}{\sigma^2}\left(\mathbf{A}_1^T\mathbf{A}_2\right)\left(\mathbf{A}_2^T\mathbf{A}_2\right)^{-1}\mathbf{A}_2^T\left(\mathbf{X}-\mathbf{A}\boldsymbol{\theta}\right) \\
&= \frac{1}{\sigma^2}\left(\mathbf{A}_1^T - \left(\mathbf{A}_1^T\mathbf{A}_2\right)\left(\mathbf{A}_2^T\mathbf{A}_2\right)^{-1}\mathbf{A}_2\right)\mathbf{X} \\
&\qquad - \frac{1}{\sigma^2}\left(\mathbf{A}_1^T\mathbf{A}_1 - \left(\mathbf{A}_1^T\mathbf{A}_2\right)\left(\mathbf{A}_2^T\mathbf{A}_2\right)^{-1}\mathbf{A}_2^T\mathbf{A}_1\right)\boldsymbol{\theta}_1 \\
&= \frac{1}{\sigma^2}\left(\mathbf{A}_1^T - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{A}_2^T\right)\mathbf{X} - \frac{1}{\sigma^2}\left(\mathbf{I}^{11}\right)^{-1}\boldsymbol{\theta}_1
\end{aligned} \tag{2.77}$$

which is independent of $\boldsymbol{\theta}_2$. The estimating equation

$$\hat{\mathbf{J}}_{1,\text{lin}} = \mathbf{0} \tag{2.78}$$

has solution

$$\hat{\boldsymbol{\theta}}_1 = \mathbf{I}^{11}\left(\mathbf{A}_1^T - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{A}_1^T\right)\mathbf{X}. \tag{2.79}$$

Similarly, $\hat{\mathbf{J}}_{2,\text{lin}}$ is independent of $\boldsymbol{\theta}_1$, and the estimating equation

$$\hat{\mathbf{J}}_{2,\text{lin}} = \mathbf{0} \tag{2.80}$$

has solution

$$\hat{\boldsymbol{\theta}}_2 = \mathbf{I}^{22}\left(\mathbf{A}_2^T - \mathbf{I}_{12}\mathbf{I}_{22}^{-1}\mathbf{A}_1^T\right)\mathbf{X}. \tag{2.81}$$

Simple calculations show that the variance of $\left(\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T\right)^T$ is $\sigma^2\left(\mathbf{A}^T\mathbf{A}\right)^{-1}$ and that $\hat{\boldsymbol{\theta}}_n^T = \left(\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T\right)$. However, calculating $\hat{\boldsymbol{\theta}}_n$ from (2.73) requires inverting one $s \times s$ matrix, while calculating $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ separately from (2.78) and (2.80) requires inverting an $r \times r$ matrix and a $p \times p$ matrix. The computational complexity of inverting an $s \times s$ matrix is $Cs^{2+\delta}$ for some constants $C > 0$ and $0 < \delta < 1$. Thus, if $r = q = s/2$ the computational complexity of generating $\hat{\boldsymbol{\theta}}_n$ from (2.73) is 2^δ times higher than first partitioning the parameter and then estimating the subvectors separately using (2.78) and (2.80).

Chapter 3

Estimating functions for location parameter families

3.1 Equivariant estimators

In Chapter 2 we studied estimators generated by estimating equations. In a sense, the estimators were not given in an explicit form. Here, using the special characteristics of a parameter, we present efficient estimators in an explicit form.

The *loss* incurred by estimating $g(\theta)$ by $\tilde{g}$ is measured by a *loss function* $L(\tilde{g}; g)$. Typically $L(\tilde{g}; g) \geq 0$ and $L(\tilde{g}; g)$ attains its minimum value in g at $g = \tilde{g}$. Two common examples of loss functions are

1.
$$L(\tilde{g}; g) = \mid \tilde{g}(\mathbf{x}) - g(\theta) \mid^p, \ p > 0$$

2.
$$L(\tilde{g}; g) = \begin{cases} 1 & \text{if } \mid \tilde{g}(\mathbf{x}) - g(\theta)| \geq \Delta \\ 0 & \text{if } \mid \tilde{g}(\mathbf{x}) - g(\theta)| < \Delta \end{cases} \quad \text{for } \Delta > 0.$$

The first loss function is quadratic loss when $p = 2$ and absolute value loss when $p = 1$. The second loss function is known as "0-1" loss; it is related to estimation by confidence intervals.

If $\tilde{g}(\mathbf{X})$ is an estimator for $g(\theta)$, its performance is measured by the expected loss, called the *risk*:

$$R(\tilde{g}; g) = E_\theta \left(L(\tilde{g}; g) \right) = \int_{\mathcal{X}} L(\tilde{g}(\mathbf{x}); g(\theta)) dP_\theta(\mathbf{x}). \tag{3.1}$$

The risk corresponding to the first loss function when $p = 2$ is the mean squared error, $E_\theta \left(\tilde{g}(\mathbf{X}) - g(\theta) \right)^2$, and the risk corresponding to the second loss function is $P_\theta \left(| \tilde{g}(\mathbf{X}) - g(\theta)| \geq \Delta \right)$.

We would like to find an estimator for $g(\theta)$ which minimizes the risk function for all values of $\theta \in \Theta$. Unfortunately, such a statistic does not exist in general. For example, let $\tilde{g}(\mathbf{X}) = g(\theta_0)$ for some fixed number θ_0. The risk of this estimator will be close to 0 for values of θ that are close to θ_0 but will likely be very large for other values of the parameter. For this reason we need to restrict our attention to a smaller class of estimators which possess certain characteristics.

In what follows, we will consider the class of *equivariant* estimators. The motivation for the use of an equivariant estimator can found in [9]. Equivariant estimators are appropriate for problems in which certain symmetries exist; they also have the advantage that in many cases there exists a best equivariant estimator which minimizes the risk uniformly for all values of θ.

First, consider the case where the parameter μ is a *location parameter*. If $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$ is a probability space and X is a random variable in $\mathbb{R}$ distributed according to the probability measure P, then for any $\mu \in \mathbb{R}$, the random variable $X + \mu$ will be distributed according to the probability measure P_μ where $P_\mu(B) = P(B - \mu)$ for any $B \in \mathcal{B}(\mathbb{R})$. The family $\mathcal{P} = \{P_\mu : \mu \in \mathbb{R}\}$ is called a location family. If $E|X| < \infty$ we can assume without loss of generality that $E_\mu \left(X \right) = \mu$.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from $P_\mu \in \mathcal{P}$ and let $T(\mathbf{X})$ be an estimator of μ. Suppose that instead of μ, we wish to estimate $\mu + a$ for some $a \in \mathbb{R}$. Two reasonable approaches to this estimation problem are to estimate $\mu + a$ by $T(\mathbf{X}) + a$, or to first transform the data with $\tilde{X}_i = X_i + a$ and use $T(\mathbf{X} + a)$ as our estimator. We call a statistic *location equivariant* if

$$T(X_1 + a, \ldots, X_n + a) = T(X_1, \ldots, X_n) + a \tag{3.2}$$

for any $a \in \mathbb{R}$. If (3.2) is violated, then the estimation problem depends on the choice of the origin in $\mathbb{R}$ (see also [18], ch. 2).

When the loss function is squared error loss, there exists an equivariant estimator of μ which minimizes the risk uniformly for any $\mu \in \mathbb{R}$. The estimator is called the Pitman estimator, and it is given by

$$t_n = \overline{X} - E_0 \left(\overline{X} \mid X_1 - \overline{X}, \ldots, X_n - \overline{X} \right). \tag{3.3}$$

If the density function $f(x - \theta)$ exists, then (3.3) can be written in integral form as

$$t_n = \frac{\int u f(X_1 - u) \cdots f(X_n - u) du}{\int f(X_1 - u) \cdots f(X_n - u) du}. \tag{3.4}$$

The Pitman estimator of a univariate location parameter is unbiased and efficient ([31], [40]). That is,

$$\sqrt{n}\,(t_n - \mu) \xrightarrow{d} N\left(0, \frac{1}{I(f)}\right) \tag{3.5}$$

as $n \longrightarrow \infty$, where

$$I(f) = \int \left(\frac{f'(x)}{f(x)} \right)^2 f(x) dx \tag{3.6}$$

is the Fisher information of a location parameter.

If, in addition, there exists an equivariant estimator $\tilde{t}_n$ of μ such that

$$E_\mu \mid \tilde{t}_n - \mu \mid^3 < \infty, \tag{3.7}$$

the Pitman estimator of μ is also admissible under quadratic loss in the entire class of estimators of μ ([39]). That is, if $\hat{\mu}_n = \hat{\mu}_n(X_1, \ldots, X_n)$ is any estimator of μ such that

$$E_\mu \left(\hat{\mu}_n - \mu\right)^2 \leq E_\mu \left(t_n - \mu\right)^2 \tag{3.8}$$

for all $\mu \in \mathbb{R}$, then

$$E_\mu \left(\hat{\mu}_n - \mu\right)^2 = E_\mu \left(t_n - \mu\right)^2 \tag{3.9}$$

for almost all μ.

In the class of all distributions F with fixed finite variance σ^2, the risk of the Pitman estimator is maximized when $t_n = \overline{X}$. If F is Gaussian, then $\overline{X}$ is independent of the vector of residuals $(X_1 - \overline{X}, \ldots, X_n - \overline{X})$, so

$$E_0 \left(\overline{X} \mid X_1 - \overline{X}, \ldots, X_n - \overline{X}\right) = 0. \tag{3.10}$$

Hence the normal distribution is a least favorable distribution. It was shown by Kagan, et al. ([23], [24]) that if $n \geq 3$, the normal distribution is the unique distribution for which $t_n = \overline{X}$; the result is known as the KLR-theorem.

3.2 Location-scale families

Let $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P)$ be a probability space and let X be a random variable in $\mathbb{R}$ distributed according to the probability measure P. For any $\mu \in \mathbb{R}$ and any $\sigma \in \mathbb{R}^+$, the random variable $\sigma X + \mu$ will be distributed according to the probability measure

$P_{\mu,\sigma}$, where $P_{\mu,\sigma}(B) = P(1/\sigma(B-\mu))$ for any $B \in \mathcal{B}(\mathbb{R})$. We call such a family $\mathcal{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R},\ \sigma \in \mathbb{R}^+\}$ a location-scale family. For the purpose of this section, the location parameter μ is the parameter of interest and the scale parameter σ is a nuisance parameter. As before, we can assume without loss of generality that $E_{\mu,\sigma}(X) = \mu$ for all $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$.

3.2.1 Modified Pitman estimator

Let $\boldsymbol{\theta} = (\mu, \sigma)^T$. For a location-scale family with known f, the Fisher information matrix is of the form

$$I(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{pmatrix} \int \frac{f'(x)^2}{f(x)}dx & \int x\frac{f'(x)^2}{f(x)}dx \\ \int x\frac{f'(x)^2}{f(x)}dx & \int \frac{(xf'(x)+f(x))^2}{f(x)}dx \end{pmatrix} = \frac{1}{\sigma^2}\begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}, \tag{3.11}$$

which is independent of μ and depends on σ only through the coefficient $1/\sigma^2$. The inverse of this matrix has the form

$$I^{-1}(\boldsymbol{\theta}) = \sigma^2 \begin{pmatrix} I^{11} & I^{12} \\ I^{21} & I^{22} \end{pmatrix} \tag{3.12}$$

where $I^{11} = (I_{11} - I_{12}I_{22}^{-1}I_{21})^{-1}$. This gives us the Cramér-Rao lower bound for the variance of any unbiased estimator of μ in the presence of the nuisance scale parameter σ of $\sigma^2 I^{11}$.

Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random sample from $P_{\mu,\sigma}$. Denote an estimator for μ by $\hat{\mu}_n = \hat{\mu}_n(X_1, \ldots, X_n)$. As in the previous section, we will restrict our attention to equivariant estimators of μ. For the location-scale family, an estimator $\hat{\mu}_n$ is equivariant if for any $a \in \mathbb{R}$ and any $b \in \mathbb{R}^+$ we have

$$\hat{\mu}_n(bX_1 + a, \ldots, bX_n + a) = b\hat{\mu}_n(X_1, \ldots, X_n) + a. \tag{3.13}$$

We will compare estimators using the quadratic loss function

$$L(\hat{\mu};\theta) = \left(\frac{\hat{\mu}-\mu}{\sigma}\right)^2. \tag{3.14}$$

Notice that for any equivariant estimator $\hat{\mu}_n$, the risk function is independent of the parameter, since

$$\begin{aligned} E_\theta L(\hat{\mu};\theta) &= \int_{\mathbb{R}^n} \left(\frac{\hat{\mu}_n(x_1,\ldots,x_n)-\mu}{\sigma}\right)^2 \prod_{i=1}^n \frac{1}{\sigma} f\left(\frac{x_i-\mu}{\sigma}\right) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left(\frac{\hat{\mu}_n(\sigma x_1+\mu,\ldots,\sigma x_n+\mu)-\mu}{\sigma}\right)^2 \prod_{i=1}^n f(x_i) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \left(\frac{\sigma\hat{\mu}_n(x_1,\ldots,x_n)+\mu-\mu}{\sigma}\right)^2 \prod_{i=1}^n f(x_i) d\mathbf{x} \\ &= \int_{\mathbb{R}^n} \hat{\mu}_n^2(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}. \end{aligned} \tag{3.15}$$

Finding a best equivariant estimator with respect to the quadratic loss function amounts to minimizing $E_{0,1}\hat{\mu}^2(\mathbf{X})$ over the class of equivariant estimators. It can be shown (see [9]) that the Pitman estimator of μ in the presence of the nuisance parameter σ has the form

$$\hat{t}_n(X_1,\ldots,X_n) = \overline{X} - S\frac{E_{0,1}\left(\overline{X}S \,\middle|\, \frac{X_1-\overline{X}}{S},\ldots,\frac{X_n-\overline{X}}{S}\right)}{E_{0,1}\left(S^2 \,\middle|\, \frac{X_1-\overline{X}}{S},\ldots,\frac{X_n-\overline{X}}{S}\right)}, \tag{3.16}$$

where S is the sample standard deviation.

The Pitman estimator of μ has the property that

$$\sqrt{n}\left(\hat{t}_n-\mu\right) \xrightarrow{d} N\left(0,\frac{\sigma^2}{\hat{I}_1}\right), \tag{3.17}$$

where $\hat{I}_1 = I_{11} - I_{12}I_{22}^{-1}I_{21}$. Since the asymptotic distribution is Gaussian, with mean 0 and variance which achieves the Cramér-Rao lower bound, the Pitman estimator of μ is an efficient estimator.

The goal of the remainder of this section is to construct an equivariant estimator $\tilde{t}_n$ of μ, which we will call the modified Pitman estimator, that is asymptotically as good as the Pitman estimator $\hat{t}_n$. Since the Pitman estimator is the best equivariant estimator, we will necessarily have that $R(\hat{t}_n, \mu) \leq R(\tilde{t}_n, \mu)$. However, the modified estimator has the advantage that it has an explicit form that is easier to work with.

In the location model, a version of the Pitman estimator was constructed by modifying the score equation using an iterative approach based on the Newton-Raphson algorithm (see [19] and [22]). In the location model, the maximum likelihood estimator $\hat{\mu}_n$ is the solution of the likelihood equation

$$\sum_{i=1}^{n} -\frac{f'(X_i-\mu)}{f(X_i-\mu)} = \sum_{i=1}^{n} -J(X_i-\mu) = 0. \tag{3.18}$$

Applying the iterative procedure to (3.18), and replacing μ with the $\sqrt{n}-$consistent initial estimate of μ, $\overline{X}$, we obtain the statistic

$$\overline{X} + \frac{1}{\sum_{i=1}^{n} J'(X_i-\overline{X})} \sum_{i=1}^{n} J(X_i-\overline{X}). \tag{3.19}$$

Finally, we can replace J' with its expectation $-I$, the Fisher information, to obtain the modified Pitman estimator

$$\tilde{\mu}_n = \overline{X} - \frac{1}{nI} \sum_{i=1}^{n} J(X_i-\overline{X}). \tag{3.20}$$

The modified Pitman estimator is an equivariant, asymptotically efficient estimator of μ.

For the location-scale model, the maximum likelihood estimator is the solution to the pair of equations

$$\begin{aligned}\sum_{i=1}^{n} -\frac{1}{\sigma} J\left(\frac{X_i-\mu}{\sigma}\right) &= \sum_{i=1}^{n} J_1(X_i;\mu,\sigma) = 0\\ \sum_{i=1}^{n}\left[-\frac{1}{\sigma} - \frac{(X_i-\mu)}{\sigma^2} J\left(\frac{X_i-\mu}{\sigma}\right)\right] &= \sum_{i=1}^{n} J_2(X_i;\mu,\sigma) = 0\end{aligned} \tag{3.21}$$

where as before, $J(x) = f'(x)/f(x)$.

Since we are interested only in estimation of μ, and regard σ as a nuisance parameter, we need to eliminate σ in our estimating function. To generate an estimating function for μ, we can use the method of projection discussed in Section 2.3. Let $\hat{J}_1 = J_1 - \hat{E}_{\boldsymbol{\theta}}(J_1|J_2)$ where the operator $\hat{E}$ is the projection operator onto the space spanned by J_2. Since $E_{\boldsymbol{\theta}}\left((J_1 - I_{12}I_{22}^{-1}J_2)J_2\right) = 0$ it follows that $\hat{J}_1 = J_1 - I_{12}I_{22}^{-1}J_2$, where I_{11} and I_{22} are the elements of the Fisher information matrix (3.11).

We can apply the same iterative procedure to $\hat{J}_1$ to get the expression

$$\mu + \frac{1}{\sum_{i=1}^{n} \frac{\partial}{\partial\mu}\hat{J}_1(X_i;\mu,\sigma)} \sum_{i=1}^{n} \hat{J}_1\left(X_i;\mu,\sigma\right). \tag{3.22}$$

Replacing $\hat{J}_1'$ with its expectation $-1/\sigma^2\hat{I}_1 = -1/\sigma^2(I_{11} - I_{12}I_{22}^{-1}I_{21})$, μ with $\overline{X}$, and σ with S, we get the modified Pitman estimator

$$\begin{aligned}&\tilde{t}_n(X_1,\ldots,X_n)\\ &\quad= \overline{X} - \frac{S}{n\hat{I}}\sum_{i=1}^{n}\left[J\left(\frac{X_i-\overline{X}}{S}\right) - \frac{I_{12}}{I_{22}}\left(1 + \frac{X_i-\overline{X}}{S} J\left(\frac{X_i-\overline{X}}{S}\right)\right)\right]\\ &\quad= \overline{X} - \frac{S}{n\hat{I}}\sum_{i=1}^{n}\varphi\left(\frac{X_i-\overline{X}}{S}\right).\end{aligned} \tag{3.23}$$

Theorem 3.2.1. *Suppose $E_{\boldsymbol{\theta}}X^4 < \infty$ and φ is twice differentiable with*

$$\left|\varphi\left(\frac{x-\mu}{\sigma}\right)\right| \leq h(x), \tag{3.24}$$

for all μ and σ, where $E_{\boldsymbol{\theta}} h^2(x) < \infty$. Then

$$\sqrt{n}\left(\tilde{t}_n - \mu\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\hat{I}_1}\right) \text{ as } n \longrightarrow \infty. \tag{3.25}$$

Proof. Since $\mathbf{J} = (J_1, J_2)^T$ is a Fisher estimating function, the following equalities hold:

$$E_{\boldsymbol{\theta}} J_1^2 = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \mu} J_1\right) = \frac{1}{\sigma^2} I_{11}, \tag{3.26}$$

$$E_{\boldsymbol{\theta}}\left(J_2^2\right) = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \sigma} J_2\right) = \frac{1}{\sigma^2} I_{22}, \tag{3.27}$$

$$E_{\boldsymbol{\theta}}\left(J_1 J_2\right) = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \mu} J_2\right) = -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial \sigma} J_1\right) = \frac{1}{\sigma^2} I_{12} = \frac{1}{\sigma^2} I_{21}, \tag{3.28}$$

and

$$E_{\boldsymbol{\theta}}\left(\hat{J}_1^2\right) = \frac{1}{\sigma^2} E_{\boldsymbol{\theta}}\left(\varphi^2\right) = \frac{1}{\sigma^2} \hat{I}_1. \tag{3.29}$$

We can expand φ in a Taylor series around the point $\left(X_i - \mu\right)/\sigma$ to get

$$\begin{aligned}
\sqrt{n}(\tilde{t}_n - \mu) &= \sqrt{n}(\overline{X} - \mu) - \frac{\sqrt{n} S}{n \hat{I}_1} \sum_{i=1}^{n} \varphi\left(\frac{X_i - \overline{X}}{S}\right) \\
&= \sqrt{n}(\overline{X} - \mu) - \frac{S}{\sqrt{n} \hat{I}_1} \sum_{i=1}^{n} \varphi\left(\frac{X_i - \mu}{\sigma}\right) \\
&\quad - \frac{\sqrt{n} S}{n \hat{I}_1} \sum_{i=1}^{n} \varphi'\left(\frac{X_i - \mu}{\sigma}\right)\left(\frac{X_i - \overline{X}}{S} - \frac{X_i - \mu}{\sigma}\right) \\
&\quad - \frac{\sqrt{n} S}{n \hat{I}_1} \sum_{i=1}^{n} \varphi''\left(\xi_i\right)\left(\frac{X_i - \overline{X}}{S} - \frac{X_i - \mu}{\sigma}\right)^2
\end{aligned} \tag{3.30}$$

for some ξ_i between $(X_i - \overline{X})/S$ and $(X_i - \mu)/\sigma$. Straight forward calculation show

$$\varphi'\left(\frac{x - \mu}{\sigma}\right) = \sigma^2\left(\frac{\partial}{\partial \mu} J_1\left(x; \mu, \sigma\right) - \frac{I_{12}}{I_{22}} \frac{\partial}{\partial \sigma} J_1(x; \mu, \sigma)\right) \tag{3.31}$$

so that

$$\frac{1}{n} \sum_{i=1}^{n} \varphi'\left(\frac{X_i - \mu}{\sigma}\right) \xrightarrow{p} -\hat{I}_1. \tag{3.32}$$

Similarly,

$$\left(\frac{x-\mu}{\sigma}\right)\varphi'\left(\frac{x-\mu}{\sigma}\right)=\sigma^2\frac{\partial}{\partial\sigma}J_1\left(x;\mu,\sigma\right)+\sigma\hat{J}_1\left(x;\mu,\sigma\right) \\ -\frac{I_{12}}{I_{22}}\sigma^2\frac{\partial}{\partial\sigma}J_2\left(x;\mu,\sigma\right) \tag{3.33}$$

so that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)\varphi'\left(\frac{X_i-\mu}{\sigma}\right)\xrightarrow{p}\sigma^2 I_{21}+0-\frac{I_{12}}{I_{22}}\sigma^2 I_{22}=0. \tag{3.34}$$

The first and third terms combined in equation (3.30) are $o_p(1)$ since

$$\begin{aligned}
&\sqrt{n}\left(\overline{X}-\mu\right)-\frac{\sqrt{n}}{n}\frac{S}{\hat{I}_1}\sum_{i=1}^{n}\varphi'\left(\frac{X_i-\overline{X}}{S}\right)\left(\frac{X_i-\overline{X}}{S}-\frac{X_i-\mu}{\sigma}\right)\\
&\quad=\sqrt{n}\left(\overline{X}-\mu\right)-\frac{\sqrt{n}}{n}\frac{S}{\hat{I}_1}\sum_{i=1}^{n}\left\{\varphi'\left(\frac{X_i-\mu}{\sigma}\right)\right.\\
&\qquad\qquad\left.\left[\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{\sigma}{S}-1\right)-\left(\frac{\overline{X}-\mu}{\sigma}\right)\right]\right\}\\
&\quad=\sqrt{n}\left(\overline{X}-\mu\right)+\sqrt{n}\left(S-\sigma\right)\frac{1}{\hat{I}_1}\frac{1}{n}\sum_{i=1}^{n}\varphi'\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{X_i-\mu}{\sigma}\right)\\
&\qquad\qquad+\sqrt{n}\left(\overline{X}-\mu\right)\frac{1}{\hat{I}_1}\frac{1}{n}\sum_{i=1}^{n}\varphi'\left(\frac{X_i-\mu}{\sigma}\right)\\
&\quad=\sqrt{n}\left(\overline{X}-\mu\right)\left[1+\frac{1}{\hat{I}_1}\frac{1}{n}\sum_{i=1}^{n}\varphi'\left(\frac{X_i-\mu}{\sigma}\right)\right]+O_p(1)o_p(1)\\
&\quad=O_p(1)\left[1-\frac{\hat{I}_1}{\hat{I}_1}+o_p(1)\right]+o_p(1)\\
&\quad=o_p(1).
\end{aligned} \tag{3.35}$$

The last term in equation (3.30) is also $o_p(1)$ since

$$
\begin{aligned}
&\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi''\left(\xi_i\right)\left(\frac{X_i-\overline{X}}{S}-\frac{X_i-\mu}{\sigma}\right)^2\right| \\
&\qquad\le \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left|\varphi''\left(\xi_i\right)\right|\left[\left(\frac{X_i-\mu}{\sigma}\right)\left(\frac{\sigma}{S}-1\right)-\frac{\overline{X}-\mu}{\sigma}\right]^2 \\
&\qquad\le 2\frac{\sqrt{n}}{n}\sum_{i=1}^{n}h\left(\xi_i\right)\left[\left(\frac{X_i-\mu}{\sigma}\right)^2\left(\frac{\sigma}{S}-1\right)^2+\left(\frac{\overline{X}-\mu}{\sigma}\right)^2\right] \\
&\qquad\le \frac{2}{S^2}\sqrt{n}\left(S-\sigma\right)^2\left(\frac{1}{n}\sum_{i=1}^{n}h^2(\xi_i)\frac{1}{n}\sum_{j=1}^{n}\left(\frac{X_j-\mu}{\sigma}\right)^2\right)^{1/2} \\
&\qquad\qquad\qquad + \frac{2}{\sigma^2}\sqrt{n}\left(\overline{X}-\mu\right)^2\frac{1}{n}\sum_{i=1}^{n}h(\xi_i) \\
&\qquad= o_p(1)\left(Eh^2(x)+o_p(1)\right)^{1/2}+o_p(1)\left(Eh(x)+o_p(1)\right) \\
&\qquad= o_p(1).
\end{aligned}
\tag{3.36}
$$

Therefore

$$
\begin{aligned}
\sqrt{n}\left(\tilde{t}_n-\mu\right) &= \frac{S}{\hat{I}_1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varphi\left(\frac{X_i-\mu}{\sigma}\right)+o_p(1) \\
&\xrightarrow{d} N\left(0,\frac{\sigma^2}{\hat{I}_1}\right)
\end{aligned}
\tag{3.37}
$$

as $n \longrightarrow \infty$ by the central limit theorem and Slutsky's theorem. $\square$

The proof of Theorem 3.2.1 depends mainly on the fact that the Fisher score is a Fisher estimating function. For this reason, Theorem 3.2.1 can be generalized to the case when our estimator is generated by an arbitrary Fisher estimating function of the form

$$
\mathbf{\Phi}(X;\mu,\sigma) = -\frac{1}{\sigma}\begin{bmatrix}\Phi_1\left(\frac{X-\mu}{\sigma}\right) \\ \Phi_2\left(\frac{X-\mu}{\sigma}\right)\end{bmatrix}. \tag{3.38}
$$

The covariance matrix of (3.38) is of the form

$$
\mathbf{B}_{\mathbf{\Phi}}(\boldsymbol{\theta}) = \frac{1}{\sigma^2}\begin{bmatrix}B_{11} & B_{12} \\ B_{21} & B_{22}\end{bmatrix} \tag{3.39}
$$

which is also independent of μ and depends on σ only through the coefficient $1/\sigma^2$. Let $\hat{\Phi}_1 = \Phi_1 - \hat{E}_{\boldsymbol{\theta}}(\Phi_1 \mid \Phi_2) = \Phi_1 - B_{12}B_{22}^{-1}\Phi_2$.

Theorem 3.2.2. *Suppose $E_{\boldsymbol{\theta}}X^4 < \infty$ and $\hat{\Phi}$ is twice differentiable with*

$$\left|\hat{\Phi}\left(\frac{x-\mu}{\sigma}\right)\right| \leq h(x) \tag{3.40}$$

for all μ and σ, where $E_{\boldsymbol{\theta}}h^2(x) < \infty$. Then the statistic

$$\tilde{w}_n = \overline{X} - \frac{S}{n\hat{B}_1}\sum_{i=1}^{n}\hat{\Phi}_1\left(\frac{X_i - \overline{X}}{S}\right) \tag{3.41}$$

is an equivariant estimator of μ and has the property that

$$\sqrt{n}\left(\tilde{w}_n - \mu\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\hat{B}_1}\right) \tag{3.42}$$

where $\hat{B}_1 = B_{11} - B_{12}B_{22}^{-1}B_{21}$.

Proof. The proof is identical to the proof of Theorem 3.2.1. □

3.2.2 Polynomial Pitman estimator

Let

$$\mathbf{Z} = \left(\frac{X_1 - \overline{X}}{S}, \cdots, \frac{X_n - \overline{X}}{S}\right) \tag{3.43}$$

be the standardized residuals and

$$\mathcal{C}_k = \{\tilde{\mu}_n = \overline{X} + SQ(\mathbf{Z}) \mid Q \text{ is a polynomial of degree } \leq k\}. \tag{3.44}$$

For simplicity, we call the class $\mathcal{C}_k$ the class of equivariant polynomial estimators of degree $\leq k$ of a location parameter μ in the presence of a nuisance scale parameter σ. For any $\tilde{\mu}_n \in \mathcal{C}_k$, under quadratic loss,

$$E_{\mu,\sigma}\left(\frac{\tilde{\mu}_n - \mu}{\sigma}\right)^2 = E_{0,1}\left(\tilde{\mu}_n\right)^2, \tag{3.45}$$

so that the risk is constant, and we should expect that there is an estimator that minimizes this risk.

Let Λ_k be the closed linear span of all functions of the form $SQ(\mathbf{Z})$ where Q is any polynomial of degree at most k. For any $\tilde{\mu}_n = \overline{X} + SQ(\mathbf{Z}) \in \mathcal{C}_k$ we have

$$\begin{aligned}
E_{0,1}(\overline{X} + SQ(\mathbf{Z}))^2 &= E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right) + \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right) + SQ(\mathbf{Z})\right)^2 \\
&= E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)\right)^2 + E_{0,1}\left(\hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right) + SQ(\mathbf{Z})\right)^2 \\
&\qquad + 2E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)\right)\left(\hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right) + SQ(\mathbf{Z})\right) \\
&\geq E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)\right)^2 \\
&\qquad + 2E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)\right)\left(\hat{E}_{0,1}\left(\overline{X} + SQ(\mathbf{Z}) \mid \Lambda_k\right)\right) \\
&= E_{0,1}\left(\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)\right)^2
\end{aligned} \tag{3.46}$$

where the last line follows from the fact that $\overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right)$ is orthogonal to every element in Λ_k. Since

$$\hat{\mu}_n^{(k)} = \overline{X} - \hat{E}_{0,1}\left(\overline{X} \mid \Lambda_k\right) \tag{3.47}$$

uniformly minimizes the risk in the class $\mathcal{C}_k$, we call $\hat{\mu}_n^{(k)}$ the *polynomial Pitman estimator* of a location parameter μ in the presence of a scale parameter σ. The polynomial Pitman estimator is asymptotically Gaussian:

$$\sqrt{n}\left(\hat{\mu}_n^{(k)} - \mu\right) \xrightarrow{d} N\left(0, \frac{\sigma^2}{\hat{I}_1^{(k)}}\right). \tag{3.48}$$

It will be explained below that $\hat{I}_1^{(k)}/\sigma^2$ is the information on μ in the presence of σ contained in the space of polynomials of degree at most k. This means that $\hat{\mu}_n^{(k)}$ is not only optimal in the class $\mathcal{C}_k$, but also efficient as an estimator of μ.

The polynomial Pitman estimator may have a complicated structure and it may be difficult to calculate the coefficients explicitly. In this section we consider equivariant polynomial estimators of a location parameter in the presence of a nuisance scale parameter with simpler structure that are asymptotically efficient in the space of polynomials of degree $\leq k$.

One such estimator was constructed by Kagan et al. (see [26]). This estimator can be written as

$$\tilde{\mu}_n^{(k)} = \overline{X} + S \sum_{j=2}^{k} \hat{A}_j g_j \tag{3.49}$$

where $g_j = 1/n \sum_{i=1}^{n} [(X_i - \overline{X})/S]^j$ and the constants $\hat{A}_j$ depend only on the quantities $\gamma_i = \int x^i dF(x/\sigma) / \left(\int x^2 dF(x/\sigma)\right)^{i/2}$ for $i = 1, \ldots, 2k$. $\tilde{\mu}_n^{(k)}$ is asymptotically equivalent to the polynomial Pitman estimator in the sense that

$$E_{\boldsymbol{\theta}} \left(\sqrt{n} \left(\tilde{\mu}_n^{(k)} - \hat{\mu}_n^{(k)}\right)\right)^2 = o(1) \tag{3.50}$$

as $n \longrightarrow \infty$.

We will construct an alternative estimator using the methods of Section 3.2.1. To do this, we will first need to review some results concerning estimators and information on a general finite-dimensional linear space $\mathcal{H}$ and apply these ideas to the case when $\mathcal{H} = \overline{\text{Span}\{1, x, \ldots, x^k\}}$.

In [21], Kagan considers estimation of a general finite-dimensional parameter $\boldsymbol{\theta}$ when the estimator is an element of a finite-dimensional Hilbert space $\mathcal{H}$. Let $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^2\}$ be a family of probability measures indexed by a bivariate parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. Let $\mathcal{H} = \overline{\text{Span}\{1, x, \ldots, x^k\}}$ be the closed linear span of the elements $1, \ldots, x^k$. On $\mathcal{H}$ we specify the family of scalar products

$$\left\{ (\varphi_1, \varphi_2)_{\boldsymbol{\theta}} = \int_{\mathcal{X}} \varphi_1 \varphi_2 dP_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^2,\ \varphi_1, \varphi_2 \in \mathcal{H} \right\}. \tag{3.51}$$

Let $\pi_{ij}(\boldsymbol{\theta}) = (x^i, x^j)_{\boldsymbol{\theta}},\ i, j = 1, \ldots, k,$ and $\pi_i(\boldsymbol{\theta}) = (x^i, 1)_{\boldsymbol{\theta}}$ for $i = 1, \ldots, k$. We assume that the functions $\pi_{ij}(\boldsymbol{\theta})$ are differentiable in $\boldsymbol{\theta}$, $\pi_i(\boldsymbol{\theta})$ are twice differentiable in $\boldsymbol{\theta}$, and that the Gram matrix

$$\begin{bmatrix} \pi_{11}(\boldsymbol{\theta}) & \cdots & \pi_{1k}(\boldsymbol{\theta}) \\ \cdots & \cdots & \cdots \\ \pi_{k1}(\boldsymbol{\theta}) & \cdots & \pi_{kk}(\boldsymbol{\theta}) \end{bmatrix} \tag{3.52}$$

is nonsingular for all $\boldsymbol{\theta} \in \Theta$.

Let

$$\Lambda^{(r,q)}(\boldsymbol{\theta}) = \begin{bmatrix} 0 & \frac{\partial}{\partial \theta_r} \pi_1(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_r} \pi_k(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_q} \pi_1(\boldsymbol{\theta}) & \pi_{11}(\boldsymbol{\theta}) & \cdots & \pi_{1k}(\boldsymbol{\theta}) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial}{\partial \theta_q} \pi_k(\boldsymbol{\theta}) & \pi_{k1}(\boldsymbol{\theta}) & \cdots & \pi_{kk}(\boldsymbol{\theta}) \end{bmatrix} \tag{3.53}$$

for $r, q = 1, 2$ and let $\Lambda^{(r,q)}_{i,j}(\boldsymbol{\theta}),\ i, j = 1, \ldots, k+1$ be the cofactors of the matrix $\Lambda^{(r,q)}(\boldsymbol{\theta})$. The functions

$$J_r = J_r(\mathcal{H}; \boldsymbol{\theta}) = -\sum_{i=1}^{k} \frac{\Lambda^{(r,r)}_{1,i+1}(\boldsymbol{\theta})}{\Lambda^{(r,r)}_{11}(\boldsymbol{\theta})} x^i,\ r = 1, 2, \tag{3.54}$$

are the unique functions in $\mathcal{H}$ which satisfy

$$(J_r, \varphi)_{\boldsymbol{\theta}} = \frac{\partial}{\partial \theta_r} (\varphi, 1)_{\boldsymbol{\theta}},\ r = 1, 2, \tag{3.55}$$

for all $\varphi \in \mathcal{H}$ and all $\boldsymbol{\theta}$. It can be shown that the functions $J_r(\mathcal{H}; \boldsymbol{\theta})$ are the projections of the Fisher score onto the space $\mathcal{H}$. That is, $J_1(\boldsymbol{\theta}; \mathcal{H}) = \hat{E}_{\boldsymbol{\theta}}(J_1 \mid \mathcal{H})$

and $J_2(\boldsymbol{\theta};\mathcal{H}) = \hat{E}_{\boldsymbol{\theta}}\left(J_2 \mid \mathcal{H}\right)$. The vector $\mathbf{J}(\mathcal{H};\boldsymbol{\theta}) = (J_1(\mathcal{H};\boldsymbol{\theta}), J_2(\mathcal{H};\boldsymbol{\theta}))^T \in \mathcal{H}\times\mathcal{H}$ is the called *score vector* of the space $\mathcal{H}$.

The functions $J_1(\boldsymbol{\theta};\mathcal{H})$ and $J_2(\boldsymbol{\theta};\mathcal{H})$ have 0 expectation. It follows from equation (3.55) that

$$(J_r, J_q)_{\boldsymbol{\theta}} = -\left(\frac{\partial}{\partial\theta_q}J_r, 1\right)_{\boldsymbol{\theta}}. \tag{3.56}$$

Therefore, the the score vector of the space $\mathcal{H}$ satisfies all the properties of a Fisher estimating function. The corresponding polynomial estimating equation

$$\sum_{i=1}^{n}\mathbf{J}\left(\boldsymbol{\theta};\mathcal{H}_i\right) = \mathbf{0} \tag{3.57}$$

will have a solution $\tilde{\boldsymbol{\theta}}_n$, which is not necessarily a polynomial, such that

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_2\left(\mathbf{0}, \mathbf{I}^{-1}\left(\boldsymbol{\theta};\mathcal{H}\right)\right) \tag{3.58}$$

as $n \longrightarrow \infty$.

The 2×2 matrix $\mathbf{I}(\boldsymbol{\theta};\mathcal{H}) = [I_{rq}(\boldsymbol{\theta};\mathcal{H})]$, where

$$I_{rq}(\boldsymbol{\theta};\mathcal{H}) = \left(J_r(\boldsymbol{\theta};\mathcal{H}), J_q(\boldsymbol{\theta};\mathcal{H})\right)_{\boldsymbol{\theta}},\ r,q = 1,2, \tag{3.59}$$

is called the information on $\boldsymbol{\theta}$ contained in the space $\mathcal{H}$. An explicit formula for the elements of the information matrix is given by

$$I_{rq}(\boldsymbol{\theta};\mathcal{H}) = -\frac{\mid \Lambda^{(r,q)}(\boldsymbol{\theta}) \mid}{\Lambda_{11}(\boldsymbol{\theta})},\ r,q = 1,2. \tag{3.60}$$

The information on $\boldsymbol{\theta}$ contained in the space $\mathcal{H}$ has some of the same properties as the Fisher information matrix such as additivity and monotonicity.

A version of the Cramér-Rao inequality exists for unbiased estimators $\boldsymbol{\varphi} = (\varphi_1, \varphi_2)^T \in \mathcal{H}\times\mathcal{H}$ of $\boldsymbol{\theta}$. Let $\mathbf{B}_{\boldsymbol{\varphi}}(\boldsymbol{\theta};\mathcal{H}) = E_{\boldsymbol{\theta}}\left[\left(\boldsymbol{\varphi}-\boldsymbol{\theta}\right)\left(\boldsymbol{\varphi}-\boldsymbol{\theta}\right)^T\right]$. For any $\boldsymbol{\varphi} \in \mathcal{H}\times\mathcal{H}$

such that $(\varphi_i, 1)_{\boldsymbol{\theta}} = \theta_i,\ i = 1, 2,$

$$\mathbf{B}_{\boldsymbol{\varphi}}(\boldsymbol{\theta}; \mathcal{H}) \geq \mathbf{I}^{-1}(\boldsymbol{\theta}, \mathcal{H}) = \begin{bmatrix} I^{11}(\boldsymbol{\theta}; \mathcal{H}) & I^{12}(\boldsymbol{\theta}; \mathcal{H}) \\ I^{21}(\boldsymbol{\theta}; \mathcal{H}) & I^{22}(\boldsymbol{\theta}; \mathcal{H}) \end{bmatrix}. \tag{3.61}$$

This can be proved by using fact that any covariance matrix is non-negative definite:

$$\begin{aligned}
\mathbf{0} &\leq E_{\boldsymbol{\theta}}\left(\boldsymbol{\varphi} - \boldsymbol{\theta} - \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H})\, \mathbf{J}(\boldsymbol{\theta}; \mathcal{H})\right)\left(\boldsymbol{\varphi} - \boldsymbol{\theta} - \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H})\mathbf{J}(\boldsymbol{\theta}; \mathcal{H})\right)^T \\
&= \mathbf{B}_{\boldsymbol{\varphi}}(\boldsymbol{\theta}; \mathcal{H}) - \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) E_{\boldsymbol{\theta}}\left(\mathbf{J}(\boldsymbol{\theta}; \mathcal{H}), \boldsymbol{\varphi}\right) - E_{\boldsymbol{\theta}}\left(\boldsymbol{\varphi}, \mathbf{J}(\boldsymbol{\theta}; \mathcal{H})\right) \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) \\
&\qquad + \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) E_{\boldsymbol{\theta}}\left(\mathbf{J}(\boldsymbol{\theta}; \mathcal{H})\mathbf{J}(\boldsymbol{\theta}; \mathcal{H})^T\right) \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) \qquad (3.62) \\
&= \mathbf{B}(\boldsymbol{\theta}; \mathcal{H}) - \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} \boldsymbol{\varphi} - \frac{\partial}{\partial \boldsymbol{\theta}} E_{\boldsymbol{\theta}} \boldsymbol{\varphi} \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) + \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}) \\
&= \mathbf{B}(\boldsymbol{\theta}; \mathcal{H}) - \mathbf{I}^{-1}(\boldsymbol{\theta}; \mathcal{H}).
\end{aligned}$$

It follows that for any unbiased estimator $\varphi_1 \in \mathcal{H}$ of θ_1,

$$(\varphi_1, 1)_{\boldsymbol{\theta}} \geq I^{11}(\boldsymbol{\theta}; \mathcal{H}). \tag{3.63}$$

See [21] for further discussion and detailed proofs of the above results.

Suppose $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} = (\mu, \sigma)^T \in \mathbb{R} \times \mathbb{R}_+\}$ is a location-scale family and that the probability measures $P_{\boldsymbol{\theta}}$ are absolutely continuous with respect to some sigma-finite measure ν so that the densities $(1/\sigma)\, f((x-\mu)/\sigma)$ exist. Let $J_1(x) = f'(x)/f(x)$ and $J_2(x) = (1 + x f'(x)/f(x))$. We assume that both J_1 and J_2 belong to $L^2(f)$. The functions

$$-\frac{1}{\sigma} J_1\left(\frac{x-\mu}{\sigma}\right) \tag{3.64}$$

and

$$-\frac{1}{\sigma} J_2\left(\frac{x-\mu}{\sigma}\right) \tag{3.65}$$

are the Fisher score functions for μ and σ respectively.

Let

$$J_1^{(k)} = \hat{E}_{\boldsymbol{\theta}} \left(-\frac{1}{\sigma} J_1 \left(\frac{x-\mu}{\sigma} \right) \middle| \mathcal{H} \right) \tag{3.66}$$

and

$$J_2^{(k)} = \hat{E}_{\boldsymbol{\theta}} \left(-\frac{1}{\sigma} J_2 \left(\frac{x-\mu}{\sigma} \right) \middle| \mathcal{H} \right) \tag{3.67}$$

be the polynomial versions of the Fisher score and $\hat{J}_1^{(k)} = J_1^{(k)} - \hat{E}_{\boldsymbol{\theta}} \left(J_1^{(k)} \middle| J_2^{(k)} \right)$ be the polynomial version of the efficient Fisher score for the location parameter μ in the presence of the nuisance scale parameter σ.

Let $\alpha_j = \int x^j f(x)\nu(dx) < \infty$ for $j = 1, \ldots, 2k$. The function $J_1^{(k)}$ can be written in the form

$$c_0(\boldsymbol{\theta}) + c_1(\boldsymbol{\theta}) \left(\frac{x-\mu}{\sigma} \right) + \cdots + c_k(\boldsymbol{\theta}) \left(\frac{x-\mu}{\sigma} \right)^k . \tag{3.68}$$

Since $-1/\sigma J_1 - \hat{J}_1^{(k)}$ is orthogonal to every element in $\mathcal{H}$, the coefficients $c_0(\boldsymbol{\theta}), \ldots, c_k(\boldsymbol{\theta})$ are determined by the system of equations

$$\begin{aligned} &E_{\boldsymbol{\theta}} \left(\left(-\frac{1}{\sigma} J_1 - \hat{J}_1^{(k)} \right) \left(\frac{x-\mu}{\sigma} \right)^j \right) \\ &\quad = E_{\boldsymbol{\theta}} \left(-\frac{1}{\sigma} J_1 \left(\frac{x-\mu}{\sigma} \right)^j \right) - c_0(\boldsymbol{\theta}) E_{\boldsymbol{\theta}} \left(\frac{x-\mu}{\sigma} \right)^j - \cdots - c_k(\theta) E_{\boldsymbol{\theta}} \left(\frac{x-\mu}{\sigma} \right)^{j+k} \\ &\quad = E_{\boldsymbol{\theta}} \left(-\frac{1}{\sigma} J_1 \left(\frac{x-\mu}{\sigma} \right)^j \right) - c_0(\boldsymbol{\theta})\alpha_j - \cdots - c_k(\boldsymbol{\theta})\alpha_{j+k} = 0. \end{aligned} \tag{3.69}$$

Using integration by parts gives

$$
\begin{aligned}
&E_{\boldsymbol{\theta}}\left(-\frac{1}{\sigma}J_1\left(\frac{x-\mu}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right)^j\right) \\
&\qquad = -\int \frac{1}{\sigma}J_1\left(\frac{x-\mu}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right)^j \frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)\nu(dx) \\
&\qquad = -\frac{1}{\sigma}\int x^j f'(x)\nu(dx) = \frac{j}{\sigma}\int x^{j-1}f(x)\nu(dx) \\
&\qquad = \frac{j}{\sigma}\alpha_{j-1}.
\end{aligned}
\tag{3.70}
$$

We then have the following system of equations

$$
\begin{aligned}
0 &= 0 - c_0(\boldsymbol{\theta}) - c_1(\boldsymbol{\theta})\alpha_1 - \cdots - c_k(\boldsymbol{\theta})\alpha_{k+1} \\
0 &= \frac{1}{\sigma} - c_0(\boldsymbol{\theta})\alpha_1 - c_1(\boldsymbol{\theta})\alpha_2 - \cdots - c_k(\boldsymbol{\theta})\alpha_{k+2} \\
\cdots & \qquad\qquad \cdots \\
0 &= \frac{k\alpha_{k-1}}{\sigma} - c_0(\boldsymbol{\theta})\alpha_k - c_1(\boldsymbol{\theta})\alpha_{k+1} - \cdots - c_k(\boldsymbol{\theta})\alpha_{2k}
\end{aligned}
\tag{3.71}
$$

which has solution

$$
\mathbf{c}(\boldsymbol{\theta}) = \begin{bmatrix} c_0(\boldsymbol{\theta}) \\ c_1(\boldsymbol{\theta}) \\ \cdots \\ c_k(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_k \\ \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_{k+1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_k & \alpha_{k+1} & \alpha_{k+2} & \cdots & \alpha_{2k} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 1/\sigma \\ 2\alpha_1/\sigma \\ \cdots \\ k\alpha_{k-1}/\sigma \end{bmatrix}
\tag{3.72}
$$

so that the coefficients depend on the parameter only through the factor $1/\sigma$.

Similarly, since

$$
\begin{aligned}
&E_{\boldsymbol{\theta}}\left(-\frac{1}{\sigma}J_2\left(\frac{x-\mu}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right)^j\right)\\
&\quad= -\int \frac{1}{\sigma}\left(1+\left(\frac{x-\mu}{\sigma}\right)J\left(\frac{x-\mu}{\sigma}\right)\right)\left(\frac{x-\mu}{\sigma}\right)^j\frac{1}{\sigma}f\left(\frac{x-\mu}{\sigma}\right)\nu(dx)\\
&\quad= -\frac{1}{\sigma}\int\left(1+x\frac{f'(x)}{f(x)}\right)x^j f(x)\nu(dx) = -\frac{\alpha_j}{\sigma}-\frac{1}{\sigma}\int x^{j+1}f'(x)\nu(dx) \qquad (3.73)\\
&\quad= -\frac{\alpha_j}{\sigma}+\frac{j+1}{\sigma}\int x^j f(x)\mu(dx) = -\frac{\alpha_j}{\sigma}+\frac{(j+1)\alpha_j}{\sigma}\\
&\quad= \frac{j\alpha_j}{\sigma}
\end{aligned}
$$

and

$$
E_{\boldsymbol{\theta}}\left(\left(-\frac{1}{\sigma}J_2-J_2^{(k)}\right)\left(\frac{x-\mu}{\sigma}\right)^j\right) = \frac{j\alpha_j}{\sigma} - b_0(\boldsymbol{\theta})\alpha_j - \cdots - b_k(\boldsymbol{\theta})\alpha_{j+k} = 0 \qquad (3.74)
$$

for $j = 1, \ldots, k$, the coefficients $b_0(\boldsymbol{\theta}), \ldots, b_k(\boldsymbol{\theta})$ also depend on the parameter only through a factor $1/\sigma$. Therefore, the information on $\boldsymbol{\theta}$ contained in $\mathcal{H}$ is given by

$$
\mathbf{I}(\boldsymbol{\theta};\mathcal{H}) = \frac{1}{\sigma^2}\begin{bmatrix} I_{11}^{(k)} & I_{12}^{(k)} \\ I_{21}^{(k)} & I_{22}^{(k)} \end{bmatrix} \qquad (3.75)
$$

where $I_{ij}^{(k)}$ are constants independent of $\boldsymbol{\theta}$.

The efficient polynomial score function can be written as

$$
\hat{J}_1^{(k)} = J_1^{(k)} - \frac{I_{12}^{(k)}}{I_{22}^{(k)}}J_2^{(k)} = -\frac{1}{\sigma}\left\{c_0 + c_1\left(\frac{x-\mu}{\sigma}\right) + \cdots + c_k\left(\frac{x-\mu}{\sigma}\right)^k\right\}. \qquad (3.76)
$$

Using the iterative procedure on $\hat{J}_1^{(k)}$ discussed in Section 3.2.1 we can obtain the modified polynomial Pitman estimator,

$$
\tilde{\mu}_n^{(k)} = \overline{X} - \frac{S}{\hat{I}_1^{(k)}}\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}c_j\left(\frac{X_i-\overline{X}}{S}\right)^j \qquad (3.77)
$$

where $\hat{I}_1^{(k)} = \hat{I}_{11}^{(k)} - \hat{I}_{12}^{(k)}\left(\hat{I}_{22}^{(k)}\right)^{-1}\hat{I}_{21}^{(k)}$.

Theorem 3.2.3. *Assume that for some $k \geq 3$, $\int x^{2k} f(x)\nu(dx) < \infty$. The modified polynomial Pitman estimator*

$$\tilde{\mu}_n^{(k)} = \overline{X} - \frac{S}{\hat{I}^{(k)}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{k} c_j \left(\frac{X_i - \overline{X}}{S} \right)^j \tag{3.78}$$

is an efficient estimator of μ. That is

$$\sqrt{n} \left(\tilde{\mu}_n^{(k)} - \theta \right) \xrightarrow{d} N \left(0, \frac{\sigma^2}{\hat{I}_1^{(k)}} \right) \tag{3.79}$$

as $n \longrightarrow \infty$.

Proof. Since the function $\mathbf{J}(\mathcal{H}; \boldsymbol{\theta}) = (J_1(\mathcal{H}; \boldsymbol{\theta}), J_2(\mathcal{H}; \boldsymbol{\theta}))^T = \left(J_1^{(k)}, J_2^{(k)} \right)^T$ is a Fisher estimating function, so too is $\hat{J}_1^{(k)}$ and,

$$\begin{aligned} -E_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \mu} \hat{J}_1^{(k)} \right) &= E_{\boldsymbol{\theta}} \left(-\frac{c_1}{\sigma^2} - 2\frac{c_2}{\sigma^2} \left(\frac{x-\mu}{\sigma} \right) - \cdots - k\frac{c_k}{\sigma^2} \left(\frac{x-\mu}{\sigma^2} \right)^{k-1} \right) \\ &= E_{\boldsymbol{\theta}} \left(\hat{J}_1^{(k)} \right)^2 = \frac{1}{\sigma^2} \left(I_{11}^{(k)} - \frac{\left(I_{12}^{(k)} \right)^2}{I_{22}^{(k)}} \right) = \frac{\hat{I}_1^{(k)}}{\sigma^2} \end{aligned} \tag{3.80}$$

while

$$\begin{aligned} -E_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \sigma} \hat{J}_1^{(k)} \right) &= E_{\boldsymbol{\theta}} \left(-\frac{c_0}{\sigma^2} - 2\frac{c_1}{\sigma^2} \left(\frac{X_i-\mu}{\sigma} \right) - \cdots - (k+1)\frac{c_k}{\sigma^2} \left(\frac{X_i-\mu}{\sigma} \right)^{k} \right) \\ &= -E_{\boldsymbol{\theta}} \left(\frac{\partial}{\partial \sigma} J_1^{(k)} - I_{12}^{(k)} \left(I_{22}^{(k)} \right)^{-1} \frac{\partial}{\partial \sigma} J_2^{(k)} \right) \\ &= \frac{1}{\sigma^2} \left(I_{11}^{(k)} - I_{12}^{(k)} \left(I_{22}^{(k)} \right)^{-1} I_{22}^{(k)} \right) = 0. \end{aligned} \tag{3.81}$$

We have

$$
\begin{aligned}
\sqrt{n}\left(\hat{\mu}^{(k)}-\mu\right) &= \sqrt{n}\left(\overline{X}-\mu\right) - \frac{S}{\hat{I}_1^{(k)}}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} c_j\left(\frac{X_i-\overline{X}}{S}\right)^j \\
&= \sqrt{n}\left(\overline{X}-\mu\right) - \frac{S}{\hat{I}_1^{(k)}}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} c_j\left(\frac{\sigma}{S}\right)^j\left(\frac{X_i-\mu}{\sigma}+\frac{\mu-\overline{X}}{\sigma}\right)^j \\
&= \sqrt{n}\left(\overline{X}-\mu\right) - \frac{S}{\hat{I}_1^{(k)}}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\sum_{l=0}^{j} c_j\left(\frac{\sigma}{S}\right)^j\binom{j}{l}\left(\frac{X_i-\mu}{\sigma}\right)^{j-l}\left(\frac{\mu-\overline{X}}{\sigma}\right)^l
\end{aligned}
\tag{3.82}
$$

In the above summand, the terms corresponding to $l=0$ are

$$
\begin{aligned}
&\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} c_j\left(\frac{\sigma}{S}\right)^j\left(\frac{X_i-\mu}{\sigma}\right)^j \\
&\quad= \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{c_j}{S^j}\left(\sigma^j-S^j\right)\left(\frac{X_i-\mu}{\sigma}\right)^j + \frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=0}^{k} c_j\left(\frac{X_i-\mu}{\sigma}\right)^j \\
&\quad= -\sqrt{n}\left(S-\sigma\right)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{c_j}{S^j}\left(\sum_{l=0}^{j} S^{j-l}\sigma^l\right)\left(\frac{X_i-\mu}{\sigma}\right)^j - \frac{\sigma}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right) \\
&\quad= -\sqrt{n}\left(S-\sigma\right)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\left\{(j+1)\frac{c_j}{\sigma}\left(\frac{X_i-\mu}{\sigma}\right)^j - \frac{c_j}{\sigma}\left(\frac{X_i-\mu}{\sigma}\right)^j\right\} \\
&\qquad\quad - \frac{\sigma}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right) + o_p(1) \\
&\quad= -\sqrt{n}\left(S-\sigma\right)\frac{\sigma}{n}\sum_{i=1}^{n}\left\{\frac{\partial}{\partial\sigma}\hat{J}_1^{(k)}(X_i;\mu,\sigma) - \hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right)\right\} \\
&\qquad\quad - \frac{\sigma}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right) + o_p(1) \\
&\quad= -\frac{\sigma}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}(X_i;\mu,\sigma) + o_p(1).
\end{aligned}
\tag{3.83}
$$

The terms corresponding to $l = 1$ are

$$\begin{aligned}
&\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} jc_j\left(\frac{\sigma}{S}\right)^j\left(\frac{X_i-\mu}{\sigma}\right)^{j-1}\left(\frac{\mu-\overline{X}}{\sigma}\right)\\
&\qquad= -\sqrt{n}\left(\overline{X}-\mu\right)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} j\frac{c_j}{\sigma}\left(\frac{\sigma}{S}\right)^j\left(\frac{X_i-\mu}{\sigma}\right)^{j-1}\\
&\qquad= -\sqrt{n}\left(\overline{X}-\mu\right)\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} j\frac{c_j}{\sigma}\left(\frac{X_i-\mu}{\sigma}\right)^{j-1} + o_p(1)\\
&\qquad= -\sqrt{n}\left(\overline{X}-\mu\right)\frac{\sigma}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\mu}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right) + o_p(1)\\
&\qquad= \frac{\hat{I}_1^{(k)}}{\sigma}\sqrt{n}\left(\overline{X}-\mu\right) + o_p(1)
\end{aligned} \tag{3.84}$$

The remaining terms involve $\sqrt{n}\left(\overline{X}-\mu\right)^l$, $l \geq 2$ so are $o_p(1)$. Therefore

$$\begin{aligned}
\sqrt{n}\left(\hat{\mu}_n^{(k)}-\mu\right) &= \sqrt{n}\left(\overline{X}-\mu\right) - \sqrt{n}\left(\overline{X}-\mu\right)\frac{S}{\hat{I}_1^{(k)}}\frac{\hat{I}_1^{(k)}}{\sigma} + \frac{S\sigma}{\hat{I}_1^{(k)}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right)\\
&\qquad - \frac{S}{\hat{I}_1^{(k)}}\frac{\sqrt{n}}{n}\sum_{i=1}^{n}\sum_{j=2}^{k}\sum_{l=2}^{j} c_j\left(\frac{\sigma}{S}\right)^j\binom{j}{l}\left(\frac{X_i-\mu}{\sigma}\right)^{j-l}\left(\frac{\mu-\overline{X}}{\sigma}\right)^j + o_p(1)\\
&= \frac{\sigma^2}{\hat{I}_1^{(k)}}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{J}_1^{(k)}\left(X_i;\mu,\sigma\right) + o_p(1)\\
&\xrightarrow{d} N\left(0,\frac{\sigma^2}{\hat{I}_1^{(k)}}\right)
\end{aligned} \tag{3.85}$$

as $n \longrightarrow \infty$. □

It is of interest to consider when a polynomial estimator of the form $\overline{X}+SQ(Z)$ is linear. That is, when the polynomial estimator is equal to $\overline{X}$. This question was answered by Kagan, et al. ([26]). They found that if the distribution function $F(x)$ has more than k growth points, all its moments $\alpha_1, \alpha_2, \ldots$ are finite, and $\alpha_1 = 0$, then the sample mean $\overline{X}$ is optimal as an estimator of μ in the presence of σ in the class $\overline{X} + SQ(Z)$ if and only if either $\alpha_2, \ldots, \alpha_{k+1}$ coincide with the corresponding moments of the normal distribution or $\alpha_2, \ldots, \alpha_{k+1}$ coincide with the corresponding moments of some centralized gamma distribution, or $\alpha_2, -\alpha_3, \ldots, (-1)^{k+1}\alpha_{k+1}$ coincide with the corresponding moments of some centralized gamma distribution, where a centralized gamma distribution is a distribution with characteristic function

$$\frac{e^{-i\gamma pt}}{(1-i\gamma t)^p}, \quad \gamma > 0,\ p > 0. \tag{3.86}$$

Chapter 4

Estimators by estimating equations in misspecified models

4.1 Misspecified models and quasi-maximum likelihood

In this introductory section we explain what is meant by a *misspecified model* and describe the behavior of the maximum likelihood estimator and estimators generated as solutions to estimating equations under model misspecification. The results summarized below are due mainly to Huber ([16]) and White ([44]).

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent random vectors having a common distribution function G which is absolutely continuous with respect to some sigma-finite measure μ, with $dG/d\mu = g$. We assume the model to be $\mathcal{F} = \{F(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ where Θ is a compact subset of $\mathbb{R}^s$. The densities $f(\mathbf{x}; \boldsymbol{\theta}) = dF(\mathbf{x}; \boldsymbol{\theta})/d\mu(\mathbf{x})$ are assumed to exist. The true distribution G may or may not belong to the working family of distributions $\mathcal{F}$. If G is not an element of $\mathcal{F}$, the model is said to be misspecified.

White defined the *quasi-log-likelihood* of the sample to be

$$L_n(\boldsymbol{\theta}) = L_n(\mathbf{X}_1, \ldots, \mathbf{X}_n, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \log f(\mathbf{X}_i; \boldsymbol{\theta}) \tag{4.1}$$

(not to be confused with the quasi-log-likelihood used in the theory of generalized linear models). It can be shown that under general regularity conditions there exists a measurable function $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ which maximizes the quasi-log-likelihood. That is, there exists a statistic $\hat{\boldsymbol{\theta}}_n$ such that

$$L_n(\hat{\boldsymbol{\theta}}_n) \geq L_n(\boldsymbol{\theta}) \tag{4.2}$$

for any $\boldsymbol{\theta} \in \Theta$. We call $\hat{\boldsymbol{\theta}}_n$ the *quasi-maximum likelihood estimator (QMLE)* of $\boldsymbol{\theta}$ based on the sample $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

If the model is correctly specified so that $G(\mathbf{x}) = F(\mathbf{x}; \boldsymbol{\theta}_0) \in \mathcal{F}$ for some $\boldsymbol{\theta}_0 \in \Theta$, the QMLE is simply the MLE and $\hat{\boldsymbol{\theta}}_n$ is consistent for $\boldsymbol{\theta}_0$ and asymptotically normal. However, if $G \notin \mathcal{F}$, it is not obvious that the QMLE should converge at all. It was shown by White that if there is a distribution $F(\mathbf{x}; \boldsymbol{\theta}^*) \in \mathcal{F}$ that is "closest" to $G(\mathbf{x})$, then the QMLE will be consistent for $\boldsymbol{\theta}^*$ and will also be asymptotically normal.

The *Kullback-Leibler Information Criterion* (KLIC) is defined as

$$I(g : f, \boldsymbol{\theta}) = E_g \left(\log \left[\frac{g(x)}{f(x; \boldsymbol{\theta})} \right] \right) = \int_{\mathcal{X}} \log \left[\frac{g(x)}{f(x; \boldsymbol{\theta})} \right] g(x) \mu(dx). \tag{4.3}$$

A basic result concerning the KLIC is stated in the following Lemma.

Lemma 4.1.1. *(Kullback-Leibler) Let f and g be probability densities with respect to a sigma-finite measure μ and let S be the region in which $f > 0$. If $\int_S (f(x) - g(x)) \mu(dx) \geq 0$ then*

$$\int_S \log \left(\frac{f(x)}{g(x)} \right) f(x) \mu(dx) \geq 0 \tag{4.4}$$

with equality if and only if $f(x) = g(x)$ a.e. $[\mu]$.

Proof. See [32], p. 59. $\square$

The KLIC does not define a metric on the space of density functions as it is not symmetric in its arguments and it does not satisfy the triangle inequality. However it does give us a tool for measuring the "closeness" of the density $f(x;\boldsymbol{\theta})$ to the density $g(x)$.

Theorem 4.1.2. *Suppose $E_g(\log g(\mathbf{x}))$ exists and $|\log f(\mathbf{x};\boldsymbol{\theta})| \le h(x)$ for all $\boldsymbol{\theta} \in \Theta$ where h is integrable with respect to G. If $I(g : f, \boldsymbol{\theta})$ has a unique minimum at $\boldsymbol{\theta}^* \in \Theta$,*

$$\hat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}^* \ a.s.\ [G] \tag{4.5}$$

as $n \longrightarrow \infty$.

Proof. See [44], p. 4. □

Hence the QMLE is a consistent estimator of the parameter which minimizes the KLIC.

Under certain additional general regularity conditions, the QMLE is also asymptotically normal. Let

$$\mathbf{A}(\boldsymbol{\theta}) = \begin{bmatrix} E_g\left(\frac{\partial^2 \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_1}\right) & \cdots & E_g\left(\frac{\partial^2 \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1^T \partial \boldsymbol{\theta}_s}\right) \\ \cdots & \cdots & \cdots \\ E_g\left(\frac{\partial^2 \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s^T \partial \boldsymbol{\theta}_1}\right) & \cdots & E_g\left(\frac{\partial^2 \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s^T \boldsymbol{\theta}_s}\right) \end{bmatrix} \tag{4.6}$$

be the matrix of second order partial derivatives and

$$\mathbf{B}(\boldsymbol{\theta}) = \begin{bmatrix} E_g\left(\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}^T\right) & \cdots & E_g\left(\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s}^T\right) \\ \cdots & \cdots & \cdots \\ E_g\left(\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s}\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s}^T\right) & \cdots & E_g\left(\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s}\frac{\partial \log f(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_s}^T\right) \end{bmatrix} \tag{4.7}$$

be the covariance matrix of the gradient of $\log f(\mathbf{x};\boldsymbol{\theta})$.

Theorem 4.1.3.

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right) \xrightarrow{d,G} N_s\left(\mathbf{0}, \mathbf{A}^{-1}(\boldsymbol{\theta}^*)\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}^{-1}(\boldsymbol{\theta}^*)\right) \tag{4.8}$$

Proof. See [44], p. 6. □

If the model is correctly specified, so that $g(\mathbf{x}) \equiv f(\mathbf{x}; \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}_0 \in \Theta$, then $\mathbf{B}(\boldsymbol{\theta}_0) - \mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{0}$. However, under misspecification, this is not necessarily the case, and the covariance matrix in Theorem 4.1.3 does not necessarily collapse.

Similar results hold if our estimator $\hat{\boldsymbol{\theta}}_n$ is the solution of an estimating equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{4.9}$$

In this case, $\hat{\boldsymbol{\theta}}_n$ will also be consistent and asymptotically normal. However, there does not seem to be an analogous version of the KLIC for relating the distance between the true distribution and the working distribution in this setup. The following Theorem is due to Huber ([16]).

Theorem 4.1.4. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random vectors with distribution function G and let $\hat{\boldsymbol{\theta}}_n$ be a solution of the estimating equation*

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}. \tag{4.10}$$

If

1. *$\boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ for each fixed $\mathbf{x}$,*

2. *$\lambda(\boldsymbol{\theta}) = E_G\left(\boldsymbol{\Psi}(\mathbf{x}; \boldsymbol{\theta})\right)$ exists for all $\boldsymbol{\theta} \in \Theta$ and has a unique zero at $\boldsymbol{\theta}^* \in \Theta$, and*

3. there exists a continuous function which is bounded away from zero, $b(\boldsymbol{\theta}) \geq b > 0$, such that

(a) $\sup_{\boldsymbol{\theta}} \|\boldsymbol{\Psi}\|/b(\boldsymbol{\theta})$ *is integrable*

(b) $\liminf_{\|\boldsymbol{\theta}\| \longrightarrow \infty} \|\lambda(\boldsymbol{\theta})\|/b(\boldsymbol{\theta}) \geq 1$

(c) $E_G\left(\limsup_{\|\boldsymbol{\theta}\| \longrightarrow \infty} \|\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}) - \lambda(\boldsymbol{\theta})\|/b(\boldsymbol{\theta})\right) < 1,$

then

$$\hat{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}^* \tag{4.11}$$

in $G-$probability as $n \longrightarrow \infty$.

Under further regularity conditions, the estimator $\hat{\boldsymbol{\theta}}_n$ can be shown to be asymptotically normal:

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right) \xrightarrow{d,G} N_s\left(\mathbf{0}, \Lambda^{-1}(\boldsymbol{\theta}^*)\mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}^*)\left(\Lambda^T(\boldsymbol{\theta}^*)\right)^{-1}\right), \tag{4.12}$$

where $\Lambda(\boldsymbol{\theta}^*) = \partial/\partial\boldsymbol{\theta}\lambda(\boldsymbol{\theta}) \mid_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}$.

4.2 Behavior of estimators under small model misspecification

In this section we consider the behavior of an estimator generated as the solution of an estimating equation under model misspecification when the degree of misspecification is small and can be smoothly parameterized. Suppose $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample with probability density function $\tilde{p}(\mathbf{x};\boldsymbol{\xi}_n)$, where $\boldsymbol{\xi}_n^T = (\boldsymbol{\theta}^T, \boldsymbol{\eta}_n^T)$ for $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s$ and $\boldsymbol{\eta}_n \in \Xi \subseteq \mathbb{R}^m$, and $\|\boldsymbol{\eta}_n\| \longrightarrow 0$ as $n \longrightarrow \infty$. However, we believe the random sample to have probability density function $p(\mathbf{x};\boldsymbol{\theta})$ where $p(\mathbf{x};\boldsymbol{\theta}) = \tilde{p}(\mathbf{x};\boldsymbol{\theta},\mathbf{0})$.

For example, suppose we believe the random variable X to have the form $X = \theta + \epsilon$, where ϵ is a mean zero random variable with density f, while the true form of the random variable is $X = \theta + \epsilon + \eta_n Y$, where Y is a mean zero random variable, independent of ϵ, with density g. Then the assumed density is $f(x-\theta)$ while the true density is

$$h(x;\theta,\eta_n) = \int f(x-\theta-\eta_n u)g(u)du. \tag{4.13}$$

Clearly $f(x-\theta) = h(x;\theta,0)$.

A second example is Huber's contamination model ([17]). In this model, we assume the distribution is $F(\mathbf{x};\theta)$, while the true distribution is

$$H(\mathbf{x};\theta,\eta_n) = (1-\eta_n)F(\mathbf{x};\theta) + \eta_n G(\mathbf{x};\theta). \tag{4.14}$$

This model has the interpretation that with high probability $(1-\eta_n)$ an observation $\mathbf{X}$ will be distributed according to $F(\mathbf{x};\theta)$, while with small probability the observation will be distributed according to $G(\mathbf{x};\theta)$. In the contamination model we have $F(\mathbf{x};\theta) = H(\mathbf{x};\theta,0)$.

In what follows, we will assume that the square root of the density $\tilde{p}(\mathbf{x};\boldsymbol{\xi})$ is *differentiable in quadratic mean* at the point $\boldsymbol{\xi}^T = (\boldsymbol{\theta}^T,\mathbf{0}^T)$. That is, we assume there exists a vector $\mathring{\mathbf{l}}^T(\mathbf{x}) = \left(\mathring{\mathbf{l}}_1^T(\mathbf{x}), \mathring{\mathbf{l}}_2^T(\mathbf{x})\right)$ such that

$$\begin{aligned}
&\int \left[\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi}+\mathbf{h})} - \sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi})} - \frac{1}{2}\mathbf{h}^T\mathring{\mathbf{l}}(\mathbf{x})\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi})}\right]^2 \mu(d\mathbf{x}) \\
&\qquad = \int \left[\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi}+\mathbf{h})} - \sqrt{p(\mathbf{x};\boldsymbol{\theta})} - \frac{1}{2}\mathbf{h}^T\mathring{\mathbf{l}}(\mathbf{x})\sqrt{p(\mathbf{x};\boldsymbol{\theta})}\right]^2 \mu(d\mathbf{x}) \qquad (4.15) \\
&\qquad = o(\|\mathbf{h}\|^2) \text{ as } \mathbf{h} \longrightarrow 0.
\end{aligned}$$

Typically $\mathring{\mathbf{l}}$ is the Fisher score at the point $\boldsymbol{\xi}$, which in our case is

$$\mathring{\mathbf{l}} = \begin{pmatrix} \frac{\partial \log \tilde{p}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\eta})}{\partial \boldsymbol{\theta}} \big|_{\boldsymbol{\eta}=\mathbf{0}} \\ \frac{\partial \log \tilde{p}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \big|_{\boldsymbol{\eta}=\mathbf{0}} \end{pmatrix} = \begin{pmatrix} \frac{\partial \log p(\mathbf{x};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\ \frac{\partial \log \tilde{p}(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} \big|_{\boldsymbol{\eta}=\mathbf{0}} \end{pmatrix}. \qquad (4.16)$$

Differentiability in quadratic mean is usually a weaker condition than pointwise differentiability. For example, a sufficient condition for differentiability in quadratic mean of $\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi})}$ is that $\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi})}$ is continuously differentiable in $\boldsymbol{\xi}$ for every $\mathbf{x}$ and that the Fisher information matrix

$$\tilde{\mathbf{I}}(\boldsymbol{\xi}) = E_{\tilde{p}}\left(\frac{\partial}{\partial \boldsymbol{\xi}} \log \tilde{p}(\mathbf{x};\boldsymbol{\xi}) \frac{\partial}{\partial \boldsymbol{\xi}} \log \tilde{p}(\mathbf{x};\boldsymbol{\xi})^T\right) \qquad (4.17)$$

is well defined and continuous in $\boldsymbol{\xi}$ ([42], p. 95).

The assumption of differentiability in quadratic mean of the square root of the density allows us to take an expansion of the log likelihood.

Theorem 4.2.1. *Suppose that* $\Theta \times \Xi$ *is an open subset of* $\mathbb{R}^{s+m}$ *and that* $\sqrt{\tilde{p}(\mathbf{x};\boldsymbol{\xi}_n)}$ *is differentiable in quadratic mean at* $\boldsymbol{\xi}^T = (\boldsymbol{\theta}^T, \mathbf{0}^T)$*. If* $\mathbf{h}_n = \mathbf{c}/\sqrt{n}$ *for some fixed* $\mathbf{c}^T = (\mathbf{c}_1^T, \mathbf{c}_2^T) \in \mathbb{R}^{s+m}$*, then*

$$\begin{aligned}
\log \prod_{i=1}^n \frac{\tilde{p}(\mathbf{x};\boldsymbol{\xi}+\mathbf{h}_n)}{\tilde{p}(\mathbf{x};\boldsymbol{\xi})} &= \log \prod_{i=1}^n \frac{\tilde{p}(\mathbf{x};\boldsymbol{\xi}+\mathbf{h}_n)}{p(\mathbf{x};\boldsymbol{\theta})} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{c}^T\mathring{\mathbf{l}}(\mathbf{X}_i) - \frac{1}{2}\mathbf{c}^T\tilde{\mathbf{I}}(\boldsymbol{\xi})\mathbf{c} + o_p(1)
\end{aligned} \qquad (4.18)$$

as $n \longrightarrow \infty$. *If* $\|\mathbf{h}_n\| = o(1/\sqrt{n})$,

$$\log \prod_{i=1}^{n} \frac{\tilde{p}(\mathbf{x}; \boldsymbol{\xi} + \mathbf{h}_n)}{\tilde{p}(\mathbf{x}; \boldsymbol{\xi})} = \log \prod_{i=1}^{n} \frac{\tilde{p}(\mathbf{x}; \boldsymbol{\xi} + \mathbf{h}_n)}{p(\mathbf{x}; \boldsymbol{\theta})} = o_p(1) \tag{4.19}$$

as $n \longrightarrow \infty$.

Proof. See [42], p. 94. □

When the model is correctly specified, so that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ is a random sample with density function $p(\mathbf{x}; \boldsymbol{\theta})$, we can choose an estimating function $\boldsymbol{\Psi}(\mathbf{X}; \boldsymbol{\theta})$ from which we can obtain a statistic $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ as a solution to the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}, \tag{4.20}$$

and easily find the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$ as in Chapter 2. We now describe the behavior of the estimator $\hat{\boldsymbol{\theta}}_n$, constructed as if the true family is $\mathcal{P} = \{p(\mathbf{x}; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$, using the estimating function $\boldsymbol{\Psi}$ for $\mathcal{P}$, when the true density is $\tilde{p}(\mathbf{x}; \boldsymbol{\xi}_n) = \tilde{p}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\eta}_n)$ and $p(\mathbf{x}; \boldsymbol{\theta}) = \tilde{p}(\mathbf{x}; \boldsymbol{\theta}, \mathbf{0})$. Roughly speaking, if we know the behavior of a statistic $\mathbf{T}_n$ under the probability measure P_n and we know that the probability measure Q_n is sufficiently close to P_n, then we should be able to derive the behavior of $\mathbf{T}_n$ under Q_n.

This notion of "closeness" is made precise when we introduce the concept of *contiguity*. If P_n and Q_n are sequences of probability measures defined on the measurable spaces $(\Omega_n, \mathcal{F}_n)$, the measures Q_n are said to be *contiguous* with respect to the measures P_n, written $Q_n \lhd P_n$, if for any sequence $A_n \in \mathcal{F}_n$ such that $P_n(A_n) \longrightarrow 0$ as $n \longrightarrow 0$ then $Q_n(A_n) \longrightarrow 0$ as well. The measures P_n and Q_n are said to be *mutually contiguous*, written $P_n \lhd \rhd Q_n$, if both $P_n \lhd Q_n$ and $Q_n \lhd P_n$. Contiguity can be thought of as asymptotic absolute continuity of probability measures.

Theorem 4.2.2. *(LeCam's Third Lemma): Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $\mathbf{T}_n : \Omega_n \mapsto \mathbb{R}^k$ be a sequence of random vectors. If*

$$\begin{pmatrix} \mathbf{T}_n \\ \log \frac{dQ_n}{dP_n} \end{pmatrix} \xrightarrow{d,P_n} N_{k+1}\left(\begin{pmatrix} \boldsymbol{\mu} \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\tau} \\ \boldsymbol{\tau}^T & \sigma^2 \end{pmatrix} \right) \tag{4.21}$$

then

$$\mathbf{T}_n \xrightarrow{d,Q_n} N_k\left(\boldsymbol{\mu} + \boldsymbol{\tau}, \boldsymbol{\Sigma}\right). \tag{4.22}$$

Proof. See [42], p. 90. $\square$

We mention that when the conditions of Theorem 4.2.2 hold,

$$\log \frac{dQ_n}{dP_n} \xrightarrow{d,P_n} N\left(-\frac{1}{2}\sigma^2, \sigma^2\right). \tag{4.23}$$

This is a sufficient condition for the probability measures P_n and Q_n to be mutually contiguous ([42], p. 89). The mutual contiguity of P_n and Q_n is an important part of the proof of LeCam's Third Lemma in deriving the asymptotic behavior of $\mathbf{T}_n$ under Q_n.

Let $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{X}; \boldsymbol{\theta})$ be an estimating function for $\mathcal{P} = \{p(\mathbf{x}; \boldsymbol{\theta})\}$, and assume

$$\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \begin{bmatrix} \tilde{\mathbf{I}}^1_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) & \tilde{\mathbf{I}}^2_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} E_p\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}_1^T\right) & E_p\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}_2^T\right) \end{bmatrix} = E_p\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}^T\right). \tag{4.24}$$

exists and is finite.

Theorem 4.2.3. *Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. random variables distributed according to the density $\tilde{p}(\mathbf{x}; \boldsymbol{\xi}_n)$ for $\boldsymbol{\xi}_n^T = (\boldsymbol{\theta}^T, \boldsymbol{\eta}_n^T)$, where $\tilde{p}(\mathbf{x}; \boldsymbol{\theta}, \mathbf{0}) = p(\mathbf{x}; \boldsymbol{\theta})$. Suppose $\sqrt{\tilde{p}(\mathbf{x}; \boldsymbol{\xi})}$ is differentiable in quadratic mean at $\boldsymbol{\xi}^T = (\boldsymbol{\theta}^T, \mathbf{0}^T)$. If $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ is a solution to the estimating equation*

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0} \tag{4.25}$$

such that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) = \frac{1}{\sqrt{n}} \mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) + \mathbf{o}_p(1) \tag{4.26}$$

then

1. $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d,\tilde{p}} N_s\left(\mathbf{0}, \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right)$ *if* $\|\boldsymbol{\eta}_n\| = o\left(\frac{1}{\sqrt{n}}\right)$

2. $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d,\tilde{p}} N_s\left(\mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^{2}(\boldsymbol{\theta})\mathbf{c}, \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right)$ *if* $\boldsymbol{\eta}_n = \frac{1}{\sqrt{n}}\mathbf{c}$

for some fixed $\mathbf{c} \in \mathbb{R}^m$*.*

Proof. If $\|\boldsymbol{\eta}_n\| = o(1/\sqrt{n})$ we can use Theorem 4.2.1 to expand the likelihood ratio to get

$$\log \prod_{i=1}^{n} \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\eta}_n)}{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}, \mathbf{0})} = \sum_{i=1}^{n} \log \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\eta}_n)}{p(\mathbf{X}_i; \boldsymbol{\theta})} = o_p(1). \tag{4.27}$$

Then we have under $p = p\left(\mathbf{x}; \boldsymbol{\theta}\right)$,

$$\begin{aligned} \begin{pmatrix} \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \\ \log \prod_{i=1}^{n} \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\eta}_n)}{p(\mathbf{X}_i; \boldsymbol{\theta})} \end{pmatrix} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) \\ \mathbf{0} \end{pmatrix} + \mathbf{o}_p(1) \\ &\xrightarrow{d,p} N_{s+1}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \right) \end{aligned} \tag{4.28}$$

by the multivariate central limit theorem. By LeCam's Third Lemma,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d,\tilde{p}} N_s\left(\mathbf{0}, \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right). \tag{4.29}$$

If $\boldsymbol{\eta}_n = \mathbf{c}/\sqrt{n}$ for some fixed $\mathbf{c} \in \mathbb{R}^m$, we can again use the expansion given in Theorem 4.2.1, with $\tilde{\mathbf{c}}^T = (\mathbf{0}^T, \mathbf{c}^T) \in \mathbb{R}^{s+m}$ to get

$$\begin{aligned}\log \prod_{i=1}^n \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\xi} + \tilde{\mathbf{c}}/\sqrt{n})}{\tilde{p}(\mathbf{X}_i; \boldsymbol{\xi})} &= \sum_{i=1}^n \log \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}; \mathbf{c}/\sqrt{n})}{p(\mathbf{X}_i; \boldsymbol{\theta})} \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{c}^T \mathring{\mathbf{l}}_2(\mathbf{X}_i) - \frac{1}{2}\mathbf{c}^T \tilde{\mathbf{I}}_{22}(\boldsymbol{\xi})\mathbf{c} + \mathbf{o}_p(1).\end{aligned} \tag{4.30}$$

Then we have

$$\begin{aligned}\begin{pmatrix} \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \\ \log \prod_{i=1}^n \frac{\tilde{p}(\mathbf{X}_i; \boldsymbol{\theta}, \mathbf{c}/\sqrt{n})}{p(\mathbf{X}_i; \boldsymbol{\theta})} \end{pmatrix} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) \\ \mathbf{c}^T \mathring{\mathbf{l}}_2(\mathbf{X}_i) \end{pmatrix} + \begin{pmatrix} \mathbf{0} \\ -\frac{1}{2}\mathbf{c}^T \tilde{\mathbf{I}}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix} + \mathbf{o}_p(1) \\ &\xrightarrow{d,p} N_{s+1}\left(\begin{pmatrix} \mathbf{0} \\ -\frac{1}{2}\mathbf{c}^T \tilde{\mathbf{I}}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix}, \begin{pmatrix} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) & \mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^{2}(\boldsymbol{\theta})\mathbf{c} \\ \mathbf{c}^T \tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^{2T}(\boldsymbol{\theta}) \mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) & \mathbf{c}^T \tilde{\mathbf{I}}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix}\right).\end{aligned} \tag{4.31}$$

By LeCam's Third Lemma,

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d,\tilde{p}} N\left(\mathbf{C}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^{2}(\boldsymbol{\theta})\mathbf{c}, \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right). \tag{4.32}$$

□

As an example, consider Huber's contamination model with $p(x;\theta) = f(x-\theta)$ and

$$\tilde{p}(x; \theta, \eta_n) = (1-\eta_n) f(x-\theta) + \eta_n g(x-\theta) \tag{4.33}$$

for $\theta \in \mathbb{R}$. Let

$$t_n = \overline{X} - E_0\left(\overline{X} \mid X_1 - \overline{X}, \ldots, X_n - \overline{X}\right) \tag{4.34}$$

be the Pitman estimator of θ,

$$\tilde{t}_n = \overline{X} - \frac{1}{nI(f)} \sum_{i=1}^{n} J(X_i - \overline{X}) \tag{4.35}$$

be the modified Pitman estimator discussed in Section 3.2.1, and $\hat{\theta}_n$ be the MLE, all constructed as if the sample $X_1, \ldots, X_n$ is from the density $f(x - \theta)$. Suppose the MLE $\hat{\theta}_n$ can be written in the form

$$\sqrt{n}\left(\hat{\theta}_n - \theta\right) = \frac{1}{I(f)\sqrt{n}} \sum_{i=1}^{n} J(X_i - \theta) + o_p(1). \tag{4.36}$$

It was shown by Janssen, et al. (see [19]) that under mild regularity conditions,

$$\hat{\theta}_n - t_n = O_p\left(\frac{1}{n}\right) \tag{4.37}$$

and

$$\tilde{t}_n - t_n = O_p\left(\frac{1}{n}\right). \tag{4.38}$$

Then both $\sqrt{n}\,(t_n - \theta)$ and $\sqrt{n}\left(\tilde{t}_n - \theta\right)$ have the representation

$$\frac{1}{I(f)\sqrt{n}} \sum_{i=1}^{n} J(X_i - \theta) + o_p(1). \tag{4.39}$$

The score function can be written as

$$\frac{\partial}{\partial \eta} \log \tilde{p}(x; \theta, \eta_n) = \frac{g(x-\theta) - f(x-\theta)}{f(x-\theta) + \eta_n\left(g(x-\theta) - f(x-\theta)\right)} \tag{4.40}$$

so that

$$\mathring{l}_2(x) = \frac{g(x-\theta) - f(x-\theta)}{f(x-\theta)} \tag{4.41}$$

and

$$\tilde{I}_J^2(\theta) = \int J(x-\theta)\mathring{l}_2(x) f(x-\theta)dx = \int \frac{f'(x)}{f(x)} g(x)dx. \tag{4.42}$$

Using Theorem 4.2.3, we have that if $\eta_n = o(1/\sqrt{n})$, both $\sqrt{n}\,(t_n - \theta)$ and $\sqrt{n}\left(\tilde{t}_n - \theta\right)$ converge in distribution to $N(0, I^{-1}(f))$ as $n \longrightarrow \infty$ under $\tilde{p}(x;\theta,\eta_n)$. If $\eta_n = c/\sqrt{n}$ for some constant c, then both $\sqrt{n}\,(t_n - \theta)$ and $\sqrt{n}\left(\tilde{t}_n - \theta\right)$ converge in distribution to

$$N\left(\frac{c\int \frac{f'(x)}{f(x)} g(x)dx}{I(f)}, \frac{1}{I(f)}\right) \tag{4.43}$$

as $n \longrightarrow \infty$ under $\tilde{p}(x;\theta,\eta_n)$.

4.3 Finite unions of parametric families

4.3.1 A version of the Cramér-Rao inequality

Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample with probability density function $f(x;\theta)$ where $\theta \in \mathbb{R}$. The classic Cramér-Rao inequality states that if $T = T(\mathbf{X})$ is an unbiased estimator of θ, then

$$\mathrm{Var}_\theta(T) \geq \frac{1}{I(\theta)} \tag{4.44}$$

where $I(\theta)$ is the Fisher information contained in the random sample $\mathbf{X}$.

Andrews, et al. ([3]) considered the *dual-criterion* problem for estimation of a location parameter. The statistic $T = T(X_1, \ldots, X_n)$ is assumed to be an unbiased, equivariant estimator of the univariate parameter θ when the random sample $X_1, \ldots, X_n$ is distributed according to either of the two families $dF_\theta(x) = dF(x-\theta) = f(x-\theta)$ or $dG_\theta(x) = dG(x-\theta) = g(x-\theta)$. They sought to find a Cramér-Rao lower bound for the variance of such an estimator when the true distribution is

$$H_\phi(x-\theta) = \cos^2(\phi)F(x-\theta) + \sin^2(\phi)G(x-\theta) \tag{4.45}$$

for all fixed ϕ between 0 and $\pi/2$, which includes the case where the random sample is distributed according to F and the case when the random sample is distributed according to G. The found that the asymptotic lower bound for the variance of an equivariant, unbiased estimator T is

$$\frac{\sin^2(\phi)}{I(F)} + \frac{\cos^2(\phi)}{I(G)} \tag{4.46}$$

and they constructed an equivariant, unbiased estimator T^* for which

$$\mathrm{Var}_{H_\phi}(T^*) \sim \frac{\sin^2(\phi)}{I(F)} + \frac{\cos^2(\phi)}{I(G)} \tag{4.47}$$

as $n \longrightarrow \infty$.

We consider a similar problem for a general univariate parameter θ. Let $\mathcal{P}_1 = \{P_\theta^1 : \theta \in \Theta\}$ and $\mathcal{P}_2 = \{P_\theta^2 : \theta \in \Theta\}$ where Θ is an open subset of $\mathbb{R}$. Let $\mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. We assume there exists a sigma-finite measure μ such that $P_\theta \ll \mu$ for all $P_\theta \in \mathcal{P}$ so that the densities $f_1(x;\theta)$ and $f_2(x;\theta)$ exist. Suppose $\mathbf{X} = (X_1, \ldots, X_n)$ is a random sample with each component independently distributed according to $P_\theta \in \mathcal{P}$. Our goal is to find a lower bound for the variance of an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ which is unbiased for θ for any $P_\theta \in \mathcal{P}$. That is, an estimator $\hat{\theta}_n$ for which

$$\begin{aligned}\int \cdots \int (\hat{\theta}_n - \theta) f_1(x_1;\theta) \cdots f_1(x_n;\theta)\mu(dx_1)\cdots\mu(dx_n)& \\ = \int \cdots \int (\hat{\theta}_n - \theta) f_2(x_1;\theta) \cdots f_2(x_n;\theta)\mu(dx_1)\cdots\mu(dx_n)& \\ = 0.&\end{aligned} \tag{4.48}$$

Let E_1 and Var_1 denote expectation and variance with respect to $f_1(x;\theta)$ and E_2 and Var_2 denote expectation and variance with respect to $f_2(x;\theta)$. We need the following assumptions on f_1 and f_2:

1. the support of $f_1(x;\theta)$ is the same as the support of $f_2(x;\theta)$,

2. for each fixed x, $f_1(x;\theta)$ and $f_2(x;\theta)$ are differentiable with respect to θ,

3. $J_1(x) = f_1'(x;\theta)/f_1(x;\theta) \in L^1(P_\theta^2)$ and $J_2(x) = f_2'(x;\theta)/f_2(x;\theta) \in L^1(P_\theta^1)$.

4. $0 < I_1(\theta) = E_1\left(J_1^2(x)\right) < \infty$ and $0 < I_2(\theta) = E_2\left(J_2^2(x)\right) < \infty$,

5. $f_2(x;\theta)/f_1(x;\theta) \in L^2(P_\theta^1)$ and $f_1(x;\theta)/f_2(x;\theta) \in L^2(P_\theta^2)$,

6. $\partial/\partial\theta \int f_1(x;\theta)\mu(dx) = \int \partial/\partial\theta f_1(x;\theta)\mu(dx)$ and
   $\partial/\partial\theta \int f_2(x;\theta)\mu(dx) = \int \partial/\partial\theta f_2(x;\theta)\mu(dx)$.

Theorem 4.3.1. *Under the above assumptions, if $\hat{\theta}_n = \hat{\theta}_n\left(X_1,\ldots,X_n\right)$ is an estimator of θ such that $E_1\left(\hat{\theta}_n\right) = \theta = E_2\left(\hat{\theta}_n\right)$, then for each $\theta \in \Theta$,*

$$Var_1(\hat{\theta}_n) \geq \frac{1}{nI_1(\theta) - \frac{n^2(E_2J_1(x))^2}{(1+\Delta_1)^n}} \tag{4.49}$$

and

$$Var_2(\hat{\theta}_n) \geq \frac{1}{nI_2(\theta) - \frac{n^2(E_1J_2(x))^2}{(1+\Delta_2)^n}} \tag{4.50}$$

for some non-negative constants $\Delta_1 = \Delta_1(\theta)$ and $\Delta_2 = \Delta_2(\theta)$.

Proof. Differentiating

$$E_1(\hat{\theta}_n) = \theta = \int\cdots\int \hat{\theta}_n \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) \tag{4.51}$$

with respect to θ gives

$$\begin{aligned}1 &= \int \cdots \int \hat{\theta}_n \sum_{i=1}^{n} f_1'(x;\theta) \prod_{j\neq i} f_1(x_j;\theta)\mu(dx_1)\cdots\mu(dx_n) \\ &= \sum_{i=1}^{n} \int \cdots \int \hat{\theta}_n J_1(x_i) \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) \\ &= \int \cdots \int (\hat{\theta}_n - \theta) \sum_{i=1}^{n} J_1(x_i) \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n).\end{aligned} \tag{4.52}$$

We also have $\int \cdots \int \hat{\theta}_n \prod_{i=1}^{n} f_2(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) = \theta$, so for any real c,

$$\begin{aligned}&c\int \cdots \int (\hat{\theta}_n - \theta) \prod_{i=1}^{n} f_2(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) \\ &\quad= \int \cdots \int (\hat{\theta}_n - \theta) c \prod_{i=1}^{n} \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)} \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) \\ &\quad= 0.\end{aligned} \tag{4.53}$$

Subtracting equation (4.53) from equation (4.52) gives

$$1 = \int \cdots \int (\hat{\theta}_n - \theta) \left(\sum_{i=1}^{n} J_1(x_i) - c \prod_{i=1}^{n} \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)} \right) \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n). \tag{4.54}$$

It follows from the Cauchy-Schwarz inequality that

$$\mathrm{Var}_1(\hat{\theta}_n) \geq \frac{1}{E_1\left(\sum_{i=1}^{n} J_1(x_i) - c\prod_{i=1}^{n} \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)}\right)^2}. \tag{4.55}$$

We can rewrite the term $E_1\left(\prod_{i=1}^{n} f_2(x_i;\theta)/f_1(x_i;\theta)\right)^2$:

$$\begin{aligned}&E_1\left(\prod_{i=1}^{n} \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)}\right)^2 \\ &\quad= \int \cdots \int \left(\prod_{i=1}^{n} \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)}\right)^2 \prod_{i=1}^{n} f_1(x_i;\theta)\mu(dx_1)\cdots\mu(dx_n) \\ &\quad= \left(\int \left(\frac{f_2(x;\theta)}{f_1(x;\theta)}\right)^2 f_1(x;\theta)\mu(dx)\right)^n \\ &\quad= (1+\Delta_1)^n\end{aligned} \tag{4.56}$$

for some $\Delta_1 = \Delta_1(\theta) \geq 0$, since $\int (f/g)^2 g\mu(dx) \geq 1$ with equality if and only if $f = g$ a.e. $[\mu]$. The denominator of (4.55) then becomes

$$E_1\left(\sum_{i=1}^n J_1(x_i) - c\prod_{i=1}^n \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)}\right)^2 \\ = nI_1(\theta) - 2cnE_2(J_1(x)) + c^2(1+\Delta_1)^n. \tag{4.57}$$

Minimizing over c gives

$$c^* = n\frac{E_2(J_1(x))}{(1+\Delta_1)^n}. \tag{4.58}$$

Let

$$\hat{J}_1(\mathbf{X}) = J_1(\mathbf{X}) - \hat{E}_1\left(J_1(\mathbf{X}) \middle| \frac{f_2}{f_1}(\mathbf{X})\right) \\ = \sum_{i=1}^n J_1(x_i) - n\frac{E_2(J_1(x))}{(1+\Delta_1)^n}\prod_{i=1}^n \frac{f_2(x_i;\theta)}{f_1(x_i;\theta)} \tag{4.59}$$

where $\hat{E}$ is the projection operator, and

$$\hat{I}_1(\theta) = E_1(\hat{J}_1^2(\mathbf{x})) = nI_1(\theta) - \frac{n^2\left(E_2J_1(x)\right)^2}{(1+\Delta_1)^n}. \tag{4.60}$$

The Cramer-Rao lower bound for the variance of an unbiased estimator $\hat{\theta}_n$ of θ becomes

$$\mathrm{Var}_1(\hat{\theta}_n) \geq \frac{1}{\hat{I}_1(\theta)} = \frac{1}{nI_1(\theta) - \frac{n^2(E_2J_1(x))^2}{(1+\Delta_1)^n}}, \tag{4.61}$$

proving (4.49). The proof of (4.50) is identical. $\square$

If $f \not\equiv g$, then $\Delta_1 > 0$ and for each fixed θ,

$$\frac{n^2\left(E_2J_1(x)\right)^2}{(1+\Delta_1)^n} \longrightarrow 0 \tag{4.62}$$

as $n \longrightarrow \infty$ so that the asymptotic lower bound is $1/\left(nI_1(\theta)\right)$. Therefore Theorem 4.3.1 is consistent with (4.46) in the case when θ is a location parameter.

Let us now return to the setup of Section 4.2. Suppose our two families are $\mathcal{P}_1 = \{f_1(x;\theta) : \theta \in \Theta\}$ and $\mathcal{P}_{2,n} = \{f_2(x;\theta,\alpha_n) : \theta \in \Theta,\ \alpha_n \in \mathbb{R}\}$ where the densities f_1 and f_2 are related by the equation $f_1(x;\theta) = f_2(x;\theta,0)$. Suppose we can expand the density f_2 in a Taylor series expansion about the point $\alpha = 0$:

$$\begin{aligned} f_2(x;\theta,\alpha_n) &= f_2(x;\theta,0) + \alpha_n \frac{\partial}{\partial\alpha} f_2(x;\theta,0) + o(\alpha_n) \\ &= f_1(x;\theta) + \alpha_n \frac{\partial}{\partial\alpha} f_2(x;\theta,0) + o(\alpha_n) \end{aligned} \tag{4.63}$$

as $\alpha_n \longrightarrow 0$. Let $\hat{\theta}_n$ be an unbiased estimator for θ for any $P_\theta \in \mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_{2,n}$. For this example, the information bound can be computed asymptotically.

$$\begin{aligned} 1 + \Delta_1 &= \int \left(\frac{f_2(x;\theta,\alpha_n)}{f_1(x;\theta)} \right)^2 f_1(x;\theta)\mu(dx) \\ &= \int \frac{(f_1(x;\theta) + \alpha_n \partial/\partial\alpha f_2(x;\theta,0) + o(\alpha_n))^2}{f_1(x;\theta)} \mu(dx) \\ &= \int f_1(x;\theta)\mu(dx) + 2\alpha_n \int \partial/\partial\alpha f_2(x;\theta,0)\mu(dx) \\ &\qquad + (\alpha_n)^2 \int \frac{(\partial/\partial\alpha f_2(x;\theta,0))^2}{f_1(x;\theta)} \mu(dx) + o(\alpha_n^2) \\ &= 1 + c\alpha_n^2 + o(\alpha_n^2), \end{aligned} \tag{4.64}$$

under the assumption that $\int \partial/\partial\alpha f_2(x;\theta,0)\mu(dx) = \partial/\partial\alpha \int f_2(x;\theta,0)\mu(dx) = 0$. Also,

$$\begin{aligned} E_2 J_1(x) &= \int \frac{\partial/\partial\theta f_1(x;\theta)}{f_1(x;\theta)} f_2(x;\theta,\alpha_n)\mu(dx) \\ &= \int \left(\frac{\partial/\partial\theta f_1(x;\theta)}{f_1(x;\theta)} \right) (f_1(x;\theta) + \alpha_n \partial/\partial\alpha f_2(x;\theta,0) + o(\alpha_n))\mu(dx) \\ &= \int \partial/\partial\theta f_1(x;\theta)\mu(dx) \\ &\qquad + \alpha_n \int \left(\frac{\partial/\partial\theta f_1(x;\theta)}{f_1(x;\theta)} \right) \partial/\partial\alpha f_2(x;\theta,0)\mu(dx) + o(\alpha_n) \\ &= c_1\alpha_n + o(\alpha_n) \text{ as } \alpha_n \longrightarrow 0. \end{aligned} \tag{4.65}$$

The information inequality of Theorem 4.3.1 becomes

$$\operatorname{Var}_1(\sqrt{n}\hat{\theta}_n) \geq \frac{1}{I_1(\theta) - \frac{n(E_2 J_1(x))^2}{(1+\Delta_1)^n}} = \frac{1}{I_1(\theta) - \frac{n(c_1^2\alpha_n^2 + o(\alpha_n^2))}{(1+c\alpha_n^2+o(\alpha_n^2))^n}} \tag{4.66}$$

We can use Theorem 4.3.1 to compare the information bound when the two families $\mathcal{P}_1$ and $\mathcal{P}_{2,n}$ become closer as n increases by observing what happens as $\alpha_n \longrightarrow 0$ at different rates. If $\alpha_n = o(1/\sqrt{n})$ as $n \longrightarrow \infty$,

$$\begin{aligned} \frac{n(c_1\alpha_n^2 + o(\alpha_n^2))}{(1+\alpha_n^2 c + o(\alpha_n^2))^n} &\sim c_1 n(\alpha_n^2 + o(\alpha_n^2))\,(e^c)^{-\alpha_n^2 n} \\ &= c_1 n * o\left(\frac{1}{n}\right) e^{o(1/n)} = o(1) \end{aligned} \tag{4.67}$$

as $n \longrightarrow \infty$, so that asymptotically, there is no loss of information. If $\alpha_n = 1/\sqrt{n}$,

$$\begin{aligned} \frac{n(c_1\alpha_n^2 + o(\alpha_n^2))}{(1+c\alpha_n^2 + o(\alpha_n^2))^n} &\sim c_1 n\left(\alpha_n^2 + o(\alpha_n^2)\right)(e^c)^{-\alpha_n^2 n} \\ &= c_1 n\left(\frac{1}{n} + o\left(\frac{1}{n}\right)\right) e^{-c} = c_1 e^{-c}(1+o(1)) \end{aligned} \tag{4.68}$$

as $n \longrightarrow \infty$, so that asymptotically, there may be some loss of information. Note that since the denominator of (4.49) must be non-negative, by (4.68), the constants $c = c(\theta)$ and $c_1 = c_1(\theta)$ must satisfy the inequality $c_1 e^c \leq I_1(\theta)$. Finally, if $\alpha_n = 1/n^\delta$ for $0 < \delta < 1/2$,

$$\begin{aligned} \frac{n(c_1\alpha_n^2 + o(\alpha_n^2))}{(1+\alpha_n^2 c + o(\alpha_n^2))^n} &\sim c_1 n\left(\alpha_n^2 + o(\alpha_n^2)\right)(e^c)^{-\alpha_n^2 n} \\ &= \left(c_1 n^{1-2\delta} + o(n^{1-2\delta})\right)(e^c)^{-n^{1-2\delta}} = o(1) \end{aligned} \tag{4.69}$$

as $n \longrightarrow \infty$, so that again, asymptotically, there is no loss of information.

These results can be interpreted as follows: If the two possible distributions are sufficiently close, then there is no need to distinguish between them, and there will be no loss of information. If the two distributions are sufficiently far apart, then asymptotically, we should be able to distinguish between the two possible families perfectly and there will be no loss of information. However, when $\alpha_n = O\left(1/\sqrt{n}\right)$, the possible distributions are close enough to affect the information bound.

4.3.2 Efficient estimators

Let $\mathcal{P}_1 = \{f_1(x;\theta_1) : \theta_1 \in \Theta_1\}$ and $\mathcal{P}_2 = \{f_2(x;\theta_2) : \theta_2 \in \Theta_2\}$ be two parametric families with both Θ_1 and Θ_2 open subsets of $\mathbb{R}$. The Cramér-Rao lower bound guarantees that if $X_1, \ldots, X_n$ is a random sample distributed according to $P_1 \in \mathcal{P}_1$ that any unbiased estimator $T_n = T_n(X_1, \ldots, X_n)$ of θ_1 will have variance greater than or equal to $1/nI_1(\theta)$, where $I_1(\theta)$ is the Fisher information on θ corresponding to $f_1(x;\theta)$. Asymptotically, the maximum likelihood estimator $\hat{\theta}_1$ satisfies

$$\sqrt{n}\left(\hat{\theta}_1 - \theta\right) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta)}\right) \tag{4.70}$$

as $n \longrightarrow \infty$. We say that $\hat{\theta}_1$ is an efficient estimator since the limiting distribution of $\sqrt{n}\left(\hat{\theta}_1 - \theta\right)$ is Gaussian with variance equal to the Cramer-Rao lower bound. Similarly, if $X_1, \ldots, X_n$ is distributed according to $f_2(x;\theta_2)$ any unbiased estimator $\hat{\theta}_2$ of θ_2 will have variance greater than or equal to $1/I_2(\theta_2)$, and asymptotically, the MLE of θ_2 satisfies

$$\sqrt{n}\left(\hat{\theta}_2 - \theta_2\right) \xrightarrow{d} N\left(0, \frac{1}{I_2(\theta_2)}\right). \tag{4.71}$$

Now suppose our random sample $X_1, \ldots, X_n$ comes from $P \in \mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Our goal is to construct a statistic which behaves like $\hat{\theta}_1$ when $P \in \mathcal{P}_1$ and like $\hat{\theta}_2$ when $P \in \mathcal{P}_2$.

In what follows we assume that the functions

$$\frac{1}{n}\sum_{i=1}^{n}\log f_j(X_i;\theta_j) \tag{4.72}$$

follow the uniform law of large numbers for $j = 1, 2$. That is

$$\sup_{\theta\in\Theta_j}\left|\frac{1}{n}\sum_{i=1}^{n}\log f_j\left(X_i;\theta\right) - E\left(\log f_j\left(x;\theta\right)\right)\right| = o_p(1) \text{ as } n \longrightarrow \infty. \tag{4.73}$$

One well-known set of conditions for which the uniform law of large numbers holds is given in the following Lemma.

Lemma 4.3.2. *If Θ is a subset of $\mathbb{R}^k$ with compact closure, $m_{\boldsymbol{\theta}}(\mathbf{x})$ is continuous in $\boldsymbol{\theta} \in \overline{\Theta}$ for each $\mathbf{x}$, and $|m_{\boldsymbol{\theta}}(\mathbf{x})| \leq h(\mathbf{x})$ for all $\boldsymbol{\theta} \in \overline{\Theta}$, where $E|h(x)| < \infty$, then*

$$\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n}m_{\boldsymbol{\theta}}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x})\right| \xrightarrow{p} 0 \textit{ as } n \longrightarrow \infty. \tag{4.74}$$

Proof. Fix $\epsilon > 0$. For each $\boldsymbol{\theta} \in \overline{\Theta}$, let $B_\epsilon(\boldsymbol{\theta}) = \{\boldsymbol{\theta}' \in \overline{\Theta} : \; \|\boldsymbol{\theta}' - \boldsymbol{\theta}\| < \epsilon\}$. By the continuity of $m_{\boldsymbol{\theta}}(\mathbf{x})$ in $\boldsymbol{\theta}$, for $\epsilon_n \searrow 0$ as $n \longrightarrow \infty$,

$$\left(\sup_{\boldsymbol{\theta}'\in B_{\epsilon_n}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\mathbf{x}) - \inf_{\boldsymbol{\theta}'\in B_{\epsilon_n}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\mathbf{x})\right) \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{4.75}$$

Since the functions $m_{\boldsymbol{\theta}}(\mathbf{x})$ are assumed bounded by $h(\mathbf{x})$, it follows from the Lebesgue Dominated Convergence Theorem that

$$\lim_{n\longrightarrow\infty} E\left(\sup_{\boldsymbol{\theta}'\in B_{\epsilon_n}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\mathbf{x}) - \inf_{\boldsymbol{\theta}'\in B_{\epsilon_n}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\mathbf{x})\right) = 0. \tag{4.76}$$

Therefore, for each $\boldsymbol{\theta} \in \Theta$, there exists $\delta(\boldsymbol{\theta}) > 0$ such that

$$E\left(\sup_{\boldsymbol{\theta}'\in B_{\delta(\boldsymbol{\theta})}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\boldsymbol{x}) - \inf_{\boldsymbol{\theta}'\in B_{\delta(\boldsymbol{\theta})}(\boldsymbol{\theta})} m_{\boldsymbol{\theta}'}(\mathbf{x})\right) < \epsilon. \tag{4.77}$$

The set $\{B_{\delta(\boldsymbol{\theta})}(\boldsymbol{\theta})\}_{\boldsymbol{\theta}\in\Theta}$ forms an open cover of Θ, so by the compactness of Θ there exists a finite subcover, say $\{B_1, \ldots, B_k\}$ of Θ. If $\boldsymbol{\theta} \in B_j$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\inf_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x}) &\leq \frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x}) \\ &\leq \frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x}).\end{aligned} \tag{4.78}$$

By the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\sup_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(X_i) - Em_{\boldsymbol{\theta}}(\mathbf{x}) \xrightarrow{p} E\left(\sup_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(\mathbf{x}) - m_{\boldsymbol{\theta}}(\mathbf{x})\right) < \epsilon \tag{4.79}$$

and

$$-\left(Em_{\boldsymbol{\theta}}(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^{n}\inf_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(\mathbf{X}_i)\right) \xrightarrow{p} -E\left(m_{\boldsymbol{\theta}}(\mathbf{x}) - \inf_{\boldsymbol{\theta}'\in B_j} m_{\boldsymbol{\theta}'}(\mathbf{x})\right) > -\epsilon. \tag{4.80}$$

Therefore,

$$0 \leq \limsup_{n\longrightarrow\infty}\left(\sup_{\boldsymbol{\theta}\in B_j}\left|\frac{1}{n}\sum_{i=1}^{n} m_{\theta}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x})\right|\right) < \epsilon \tag{4.81}$$

in probability for all $\epsilon > 0$, or

$$\left(\sup_{\boldsymbol{\theta}\in B_j}\left|\frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x})\right|\right) = o_p(1). \tag{4.82}$$

Finally,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\boldsymbol{x})\right| &= \max_{1\leq j\leq k}\left(\sup_{\boldsymbol{\theta}\in B_j}\left|\frac{1}{n}\sum_{i=1}^{n} m_{\boldsymbol{\theta}}(\mathbf{X}_i) - Em_{\boldsymbol{\theta}}(\mathbf{x})\right|\right) \\ &= o_p(1).\end{aligned} \tag{4.83}$$

□

Lemma 4.3.3. *Let* $\mathcal{P}_1 = \{f_1(x;\theta_1) : \theta_1 \in \Theta_1 \subseteq \mathbb{R}\}$ *and* $\mathcal{P}_2 = \{f_2(x;\theta_2) : \theta_2 \in \Theta_2 \subseteq \mathbb{R}\}$ *be two parametric families with identifiable parameters and common support* *which is independent of the parameter. Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a random sample from* $P \in \mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$*. Suppose the true value of the parameter,* θ_0*, is an interior point of* Θ_1*, and there exists a point* θ^* *which is an interior point of* Θ_2 *and uniquely minimizes the KLIC,* $\int \log\left(f_1(x;\theta_0)/f_2(x;\theta^*)\right) f_1(x;\theta_0)\mu(dx) > 0$*. Let* E_1 *denote expectation with respect to the density* $p_1 = f_1(x;\theta_0)$*. If*

$$\sup_{\theta\in\Theta_j}\left|\frac{1}{n}\sum_{i=1}^{n}\log f_j\left(X_j;\theta\right) - E_1\left(\log f_j(x;\theta)\right)\right| = o_{p_1}(1) \text{ as } n \longrightarrow \infty, \tag{4.84}$$

for $j = 1, 2$*, then*

$$P_1\left(\sup_{\theta\in\Theta}\prod_{i=1}^{n} f_1(X_i;\theta) > \sup_{\theta\in\Theta}\prod_{i=1}^{n} f_2(X_i;\theta)\right) \longrightarrow 1 \text{ as } n \longrightarrow \infty. \tag{4.85}$$

Proof. Let $\hat{\theta}_1 = \hat{\theta}_{1,n}(\mathbf{X})$ be a sequence of maximum likelihood estimators which is consistent for θ_0 based on the random sample $\mathbf{X}$ when the true density is $f_1(x;\theta_0)$, which exist with probability 1 for all sufficiently large n under the stated conditions (see [9], p. 463), so that $\prod_{i=1}^{n} f_1(X_i;\hat{\theta}_1) \geq \prod_{i=1}^{n} f_1(X_i;\theta_1)$ for any $\theta_1 \in \Theta_1$. Let $\hat{\theta}_2 = \hat{\theta}_{2,n}(\mathbf{X})$ be a sequence of QMLEs which is consistent for θ^* based on the random sample $\mathbf{X}$ and the density $f_2(x;\theta_2)$, which also exist with probability 1 for all sufficiently large n under the stated conditions (see [44]), so that $\prod_{i=1}^{n} f_2(X_i;\hat{\theta}_2) \geq \prod_{i=1}^{n} f_2(X_i;\theta_2)$ for any $\theta_2 \in \Theta_2$. Then

$$\begin{aligned}
P_1&\left(\sup_{\theta\in\Theta}\prod_{i=1}^{n} f_1(X_i;\theta) > \sup_{\theta\in\Theta}\prod_{i=1}^{n} f_2(X_i;\theta)\right)\\
&= P_1\left(\prod_{i=1}^{n} f_1(X_i;\hat{\theta}_1) > \prod_{i=1}^{n} f_2(X_i;\hat{\theta}_2)\right)\\
&= P_1\left(\frac{1}{n}\sum_{i=1}^{n}\log f_1(X_i;\hat{\theta}_1) > \frac{1}{n}\sum_{i=1}^{n}\log f_2(X_i;\hat{\theta}_2)\right).
\end{aligned} \tag{4.86}$$

By assumption,

$$\begin{aligned}
&\left| \frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i;\hat{\theta}_1) - E_1\left(\log f_1(x;\theta)\right)\Big|_{\theta=\hat{\theta}_1} \right| \\
&\qquad \le \sup_{\theta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i;\theta) - E_1\left(\log f_1(x;\theta)\right) \right| = o_{p_1}(1),
\end{aligned} \tag{4.87}$$

and by Lemma 4.1.1,

$$E_1\left(\log \frac{f_1(x;\theta_0)}{f_1(x;\theta)}\right) = E_1\left(\log f_1(x;\theta_0)\right) - E_1\left(\log f_1(x;\theta)\right) > 0 \tag{4.88}$$

for any $\theta_0 \neq \theta$. Let $M(\theta) = E_1\left(\log f_1(x;\theta)\right)$ and $M_n(\theta) = 1/n \sum_{i=1}^{n} \log f_1(x;\theta)$. By (4.88), $M(\theta_0) \ge M(\theta)$ for any $\theta \in \Theta_1$, and $M_n(\theta_0) \xrightarrow{p} M(\theta_0)$ as $n \longrightarrow \infty$, hence $M_n(\hat{\theta}_1) \ge M_n(\theta_0) = M_n(\theta_0) - M(\theta_0) + M(\theta_0) = M(\theta_0) + o_{p_1}(1)$. Therefore,

$$\begin{aligned}
o_{p_1}(1) &\le M_n(\hat{\theta}_1) - M(\theta_0) \le M_n(\hat{\theta}_1) - M(\hat{\theta}_1) \\
&\le \sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| = o_{p_1}(1),
\end{aligned} \tag{4.89}$$

so that

$$\frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i;\hat{\theta}_1) - E_1\left(\log f_1(x;\theta_0)\right) = o_{p_1}(1). \tag{4.90}$$

By Theorem 4.1.2, the QMLE $\hat{\theta}_2$ is a consistent estimator of θ^*, and $\log \prod_{i=1}^{n} f_2(X_i;\hat{\theta}_2) \ge \log \prod_{i=1}^{n} f_2(X_i;\theta)$ for any $\theta \in \Theta_2$. Since we have assumed that θ^* uniquely minimizes the KLIC, we have $\mathrm{KLIC}(f_1; f_2, \theta) > \mathrm{KLIC}(f_1; f_2, \theta^*)$, or $E_1 \log f_2(x;\theta) \le E_1 \log f_2(x;\theta^*)$ for any $\theta \in \Theta_2$. We can therefore repeat the above arguments with f_1 replaced by f_2 and θ_0 replaced by θ^* to get

$$\frac{1}{n}\sum_{i=1}^{n} \log f_2(X_i;\hat{\theta}_2) - E_1\left(\log f_2(x;\theta^*)\right) = o_{p_1}(1). \tag{4.91}$$

Using equations (4.90), (4.91), we have

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i;\hat{\theta}_1) - \frac{1}{n}\sum_{i=1}^{n} \log f_2(X_i;\hat{\theta}_2)& \\
&= \frac{1}{n}\sum_{i=1}^{n} \log f_1(X_i;\hat{\theta}_1) - E_1\left(\log f_1(x;\theta_0)\right) + E_1\left(\log f_1(x;\theta_0)\right) \\
&\quad - \frac{1}{n}\sum_{i=1}^{n} \log f_2(X_i;\hat{\theta}_2) + E_1\left(\log f_2(x;\theta^*)\right) - E_1\left(\log f_2(x;\theta^*)\right) \qquad (4.92)\\
&= E_1\left(\log f_1(x;\theta_0)\right) - E_1\left(\log f_2(x;\theta^*)\right) + o_{p_1}(1) \\
&= E_1\left(\log \frac{f_1(x;\theta_0)}{f_2(x;\theta^*)}\right) + o_{p_1}(1) \\
&\xrightarrow{p_1} E_1\left(\log \frac{f_1(x;\theta_0)}{f_2(x;\theta^*)}\right) > 0
\end{aligned}
$$

□

We can use Lemma 4.3.3 to find a statistic $\hat{\theta}_n$ that behaves like $\hat{\theta}_1$ when $P \in \mathcal{P}_1$ and like $\hat{\theta}_2$ when $P \in \mathcal{P}_2$.

Theorem 4.3.4. *Let $(X_1, \ldots, X_n)$ be i.i.d. random variables distributed according to $P \in \mathcal{P} = \mathcal{P}_1 \cup \mathcal{P}_2$. Suppose $P \in \mathcal{P}_1$, so that $dP(x)/d\mu = f(x) = f_1(x;\theta_1)$ for some $\theta_1 \in \Theta_1$, and let $\hat{\theta}_1$ be the MLE for θ_1 and $\hat{\theta}_2$ be the QMLE for θ_2. If the conditions of Lemma 4.3.3 hold, the statistic*

$$
\begin{aligned}
\hat{\theta}_n = \hat{\theta}_1 I&\left\{\prod_{i=1}^{n} f_1(X_i;\hat{\theta}_1) > \prod_{i=1}^{n} f_2(X_i;\hat{\theta}_2)\right\} \\
&+ \hat{\theta}_2 I\left\{\prod_{i=1}^{n} f_1(X_i;\hat{\theta}_1) \leq \prod_{i=1}^{n} f_2(X_i;\hat{\theta}_2)\right\}
\end{aligned} \qquad (4.93)
$$

satisfies

1. *$\hat{\theta}_n \xrightarrow{p} \theta_1$ and*

2. *$\sqrt{n}\left(\hat{\theta}_n - \theta_1\right) \xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_1)}\right)$ as $n \longrightarrow \infty$.*

Proof. Let $A_n = \left\{ \prod_{i=1}^{n} f_1(X_i; \hat{\theta}_1) \leq \prod_{i=1}^{n} f_2(X_i; \hat{\theta}_2) \right\}$. Since $P_1(A_n) \longrightarrow 0$ as $n \longrightarrow \infty$, it follows that $I(A_n) \longrightarrow 0$ in P_1-probability, since for any $\epsilon > 0$,

$$\begin{aligned} P_1\left(\omega :\mid I(A_n) \mid > \epsilon\right) &= P_1\left(\omega : I(A_n) = 1\right) = P_1\left(\omega : \omega \in A_n\right) \\ &= P_1(A_n) \longrightarrow 0 \text{ as } n \longrightarrow \infty. \end{aligned} \tag{4.94}$$

Likewise, $I(A_n^c) \longrightarrow 1$ in P_1-probability as $n \longrightarrow \infty$. Since $\hat{\theta}_1$ is a consistent estimator of θ_1 and $\hat{\theta}_2$ is a consistent estimator of θ^*, it follows that $\hat{\theta}_n$ is a consistent estimator of θ_1.

To show asymptotic normality, we can see that

$$\begin{aligned} \sqrt{n}\left(\hat{\theta}_n - \theta_1\right) &= \sqrt{n}\left(\hat{\theta}_1 I(A_n^c) + \hat{\theta}_2 I(A_n) - \theta_1 I(A_n^c) - \theta_1 I(A_n)\right) \\ &= \sqrt{n}\left(\hat{\theta}_1 - \theta_1\right) I(A_n^c) + \sqrt{n}\left(\hat{\theta}_2 - \theta_1\right) I(A_n) \\ &\xrightarrow{d} N\left(0, \frac{1}{I_1(\theta_1)}\right) \end{aligned} \tag{4.95}$$

as $n \longrightarrow \infty$ since the QMLE $\hat{\theta}_2$ is a consistent estimator of θ^* and $\sqrt{n} I(A_n) \longrightarrow 0$ in P_1-probability by a calculation equivalent to (4.94). □

Corollary 4.3.4 can be easily extended to the case when $\mathcal{P} = \mathcal{P}_1 \cup \cdots \cup \mathcal{P}_k$, a finite union of parametric families $\mathcal{P}_j = \{f_j(x; \theta_j) : \theta_j \in \Theta_j\}$. As in Theorem 4.3.3, under certain regularity conditions,

$$\lim_{n \longrightarrow \infty} P_1\left(\prod_{i=1}^{n} f_1(X_i; \hat{\theta}_1) > \max_{j \neq 1} \prod_{i=1}^{n} f_j(X_i; \hat{\theta}_j)\right) = 1 \tag{4.96}$$

and the appropriate modification of (4.93) is

$$\begin{aligned} \hat{\theta}_n = \hat{\theta}_1 I &\left\{ \prod_{i=1}^{n} f_1(X_i; \hat{\theta}_1) > \max_{j \neq 1} \prod_{i=1}^{n} f_j(X_i; \hat{\theta}_j) \right\} + \cdots \\ &+ \hat{\theta}_k I \left\{ \prod_{i=1}^{n} f_k(X_i; \hat{\theta}_k) > \max_{j \neq k} \prod_{i=1}^{n} f_j(X_i; \hat{\theta}_k) \right\}. \end{aligned} \tag{4.97}$$

In some situations the statistic $\hat{\theta}_n$ in equation (4.93) trivially converges to the true parameter θ because the quantity that minimizes the information criterion is the true parameter itself. For example, if $f_1(x)$ is a symmetric density, and $f_2(x;\theta) = f_2(x-\theta)$ where $f_2(x)$ is also a symmetric density, then $\theta^* = 0$ minimizes the KLIC. The KLIC between f_1 and f_2 is given by

$$I(f_1 : f_2, \theta) = E_{f_1}\left(\log\left[\frac{f_1(x)}{f_2(x-\theta)}\right]\right), \tag{4.98}$$

and can be minimized by maximizing the quantity

$$\int \log\left(f_2(x-\theta)\right) f_1(x)\mu(dx). \tag{4.99}$$

Assuming we can interchange the operations of integration and differentiation, the minimum is attained at the value of θ where

$$\begin{aligned} \frac{\partial}{\partial\theta}\int \log\left(f_2(x-\theta)\right) f_1(x)\mu(dx) &= \int \frac{\partial}{\partial\theta}\log\left(f_2(x-\theta)\right) f_1(x)\mu(dx) \\ &= -\int \frac{f_1'(x-\theta)}{f_1(x-\theta)} f_2(x)\mu(dx) = 0. \end{aligned} \tag{4.100}$$

Since f_1 and f_2 are even and f_1' is odd, the above integral will be equal to 0 when $\theta = 0$.

As a non-trivial example, suppose $X_1, \ldots, X_n$ are i.i.d. lognormal random variables with density function

$$g(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{x}\exp\left\{-\frac{(\log x)^2}{2}\right\} I_{(0,\infty)}(x), \tag{4.101}$$

and our two families of densities are

$$\mathcal{P}_1 = \left\{ g(x;\theta) = \frac{1}{\sqrt{2\pi}\theta}\frac{1}{x}\exp\left\{-\frac{(\log x)^2}{2\theta^2}\right\} I_{(0,\infty)}(x) \,\middle|\, \theta > 0 \right\} \tag{4.102}$$

and

$$\mathcal{P}_2 = \left\{ f(x;\theta) = \frac{1}{\theta} \exp\left\{-\frac{x}{\theta}\right\} I_{(0,\infty)}(x) \,\middle|\, \theta > 0 \right\}. \tag{4.103}$$

The parameter which minimizes $I(g; f, \theta)$ is $\theta^* = \sqrt{e}$. The MLE

$$\hat{\theta}_g = \left(\frac{1}{n} \sum_{i=1}^{n} \log(X_i) \right)^{1/2} \tag{4.104}$$

converges in probability to 1 and the QMLE $\hat{\theta}_f = \overline{X}$ converges in probability to $\theta^* = \sqrt{e}$.

A simulation of 500 standard lognormal random variables (using R) shows that, in this example, this procedure chooses the MLE every time even though the QMLE converges rather slowly.

Table 4.1: Simulation of 500 standard lognormal random variables

i	X_i	$\hat{\theta}_g$	$\hat{\theta}_f$	$\log \prod_{i=1}^{n} g(X_i;\hat{\theta}_g)$	$\log \prod_{i=1}^{n} f(X_i;\hat{\theta}_f)$	$\hat{\theta}_n$
1	0.129	2.048	0.129	-0.088	-2.065	2.048
2	0.256	0.1740	-0.193	-0.534	-3.370	1.740
3	0.133	0.1837	0.173	-0.652	-5.360	1.837
4	2.940	1.680	0.865	-3.400	-3.572	1.680
5	0.265	1.616	0.745	-3.815	-4.247	1.616
...	...	...	...	...	...	...
20	0.417	1.197	1.335	-25.172	-29.850	1.197
...	...	...	...	...	...	...
30	0.541	1.164	1.721	-44.765	-72.528	1.164
...	...	...	...	...	...	...
40	5.799	1.071	1.613	-56.115	-84.971	1.071
...	...	...	...	...	...	...
50	0.690	1.107	1.698	-71.937	-117.611	1.107
...	...	...	...	...	...	...
100	0.251	1.030	1.802	-140.922	-265.736	1.030
...	...	...	...	...	...	...
500	0.755	0.998	1.676	-698.837	-1146.099	0.998

Chapter 5

Analogues of classical tests based on estimating functions

5.1 Estimation in a submodel

We call $\mathcal{P}^*$ a *submodel* of $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ if there exists an m-dimensional ($m < s$) parameter $\boldsymbol{\eta}$ such that

$$\mathcal{P}^* = \{P^*_{\boldsymbol{\eta}}\},\ P^*_{\boldsymbol{\eta}} = P_{\boldsymbol{\theta}(\boldsymbol{\eta})},\ \boldsymbol{\eta} \in \mathcal{H} \subseteq \mathbb{R}^m \tag{5.1}$$

where $\boldsymbol{\theta}(\boldsymbol{\eta})$ is a differentiable function of $\boldsymbol{\eta}$. Let

$$\mathbf{D}(\boldsymbol{\eta}) = \begin{bmatrix} \partial\theta_1(\boldsymbol{\eta})/\partial\eta_1 & \cdots & \partial\theta_1(\boldsymbol{\eta})/\partial\eta_m \\ \ldots & \ldots & \ldots \\ \partial\theta_s(\boldsymbol{\eta})/\partial\eta_1 & \cdots & \partial\theta_s(\boldsymbol{\eta})/\partial\eta_m \end{bmatrix}. \tag{5.2}$$

We assume $\mathbf{D}$ is of full rank.

Let $\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta})$ be an estimating function for the full model $\mathcal{P}$. There are many transformations of $\boldsymbol{\Psi}$ that will produce an estimating function for $\mathcal{P}^*$, but the optimal linear transformation involves the Fisher form of the estimating function $\boldsymbol{\Psi}$. To show this we will first need the following matrix inequality.

Lemma 5.1.1. *Let $\mathbf{B}$ be an $s \times s$ symmetric positive definite matrix and let $\mathbf{A}$ be an $m \times s$ matrix of full rank $m < s$. Then*

$$\mathbf{A}^T \left(\mathbf{A}\mathbf{B}\mathbf{A}^T\right)^{-1} \mathbf{A} \leq \mathbf{B}^{-1}. \tag{5.3}$$

Proof. Let $\boldsymbol{\Sigma}$ be a positive definite matrix such that $\boldsymbol{\Sigma}^T\boldsymbol{\Sigma} = \mathbf{B}$. For any $\mathbf{Y} \in \mathbb{R}^s$ and any $\boldsymbol{\beta} \in \mathbb{R}^m$, the quadratic form

$$\left(\boldsymbol{\Sigma}\mathbf{Y} - \boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\beta}\right)^T \left(\boldsymbol{\Sigma}\mathbf{Y} - \boldsymbol{\Sigma}\mathbf{A}^T\boldsymbol{\beta}\right) = \left(\mathbf{Y} - \mathbf{A}\boldsymbol{\beta}\right)^T \mathbf{B}\left(\mathbf{Y} - \mathbf{A}\boldsymbol{\beta}\right) \tag{5.4}$$

is nonnegative. Using the theory of least squares, the above quadratic form is minimized over $\boldsymbol{\beta}$ for fixed $\mathbf{Y}$ when $\hat{\boldsymbol{\beta}} = (\mathbf{ABA}^T)^{-1}\mathbf{ABY}$. Substituting $\hat{\boldsymbol{\beta}}$ into (5.4) and expanding the expression gives

$$\mathbf{Y}^T\mathbf{BY} \geq \mathbf{Y}^T\mathbf{BA}^T\left(\mathbf{ABA}^T\right)^{-1}\mathbf{ABY}. \tag{5.5}$$

Since this holds for any $\mathbf{Y}$ it follows that

$$\mathbf{B} \geq \mathbf{BA}^T\left(\mathbf{ABA}^T\right)^{-1}\mathbf{AB}. \tag{5.6}$$

Multiplying on the left and right by $\mathbf{B}^{-1}$ gives the result. □

Theorem 5.1.2. *The optimal estimating function for $\mathcal{P}^*$ in the class of linear transformations $\mathcal{C} = \{\mathbf{A}(\boldsymbol{\theta})\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta}) : \mathbf{A}(\boldsymbol{\theta})$ is an $m \times s$ matrix of full rank $\}$ is*

$$\boldsymbol{\Psi}^*(\mathbf{X};\boldsymbol{\eta}) = \mathbf{D}^T(\boldsymbol{\eta})\mathbf{C}_{\boldsymbol{\Psi}}^T(\boldsymbol{\theta}(\boldsymbol{\eta}))\mathbf{B}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\Psi}(\mathbf{X};\boldsymbol{\theta}(\boldsymbol{\eta})). \tag{5.7}$$

Proof. We fist verify that $\boldsymbol{\Psi}^* = \mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\boldsymbol{\Psi}$ is a Fisher estimating function. Clearly $\boldsymbol{\Psi}^*$ has zero expectation and components which are square integrable. Also,

$$\begin{aligned}\mathbf{B}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) &= E_{\boldsymbol{\eta}}\left(\boldsymbol{\Psi}^*(\boldsymbol{\Psi}^*)^T\right) = \mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}E_{\boldsymbol{\eta}}\left(\boldsymbol{\Psi}\boldsymbol{\Psi}^T\right)\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}}\mathbf{D} \\ &= \mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}}\mathbf{D} > \mathbf{0}\end{aligned} \tag{5.8}$$

while

$$\begin{aligned}\mathbf{C}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) &= -E_{\boldsymbol{\eta}}\left(\frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\Psi}^*\right) = -\mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}E_{\boldsymbol{\eta}}\left(\frac{\partial}{\partial\boldsymbol{\eta}}\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta}(\boldsymbol{\eta}))\right) \\ &= -\mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}E_{\boldsymbol{\eta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})\frac{\partial\boldsymbol{\theta}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}}\right) \\ &= \mathbf{D}^T\mathbf{C}_{\boldsymbol{\Psi}}^T\mathbf{B}_{\boldsymbol{\Psi}}^{-1}\mathbf{C}_{\boldsymbol{\Psi}}\mathbf{D} = \mathbf{B}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) > \mathbf{0}.\end{aligned} \tag{5.9}$$

Hence

$$\mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) = \mathbf{B}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) = \mathbf{C}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}). \tag{5.10}$$

Let $\tilde{\boldsymbol{\Psi}} = \mathbf{A}\boldsymbol{\Psi}$ be any estimating function for $\mathcal{P}^*$ in $\mathcal{C}$. We need to show $\mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) \geq \mathbf{I}_{\tilde{\boldsymbol{\Psi}}}(\boldsymbol{\eta})$. Since

$$\mathbf{I}_{\tilde{\boldsymbol{\Psi}}}(\boldsymbol{\eta}) = \mathbf{C}_{\tilde{\boldsymbol{\Psi}}}^T \mathbf{B}_{\tilde{\boldsymbol{\Psi}}}^{-1} \mathbf{C}_{\tilde{\boldsymbol{\Psi}}} = \mathbf{D}^T \mathbf{C}_{\boldsymbol{\Psi}}^T \mathbf{A}^T (\mathbf{A}\mathbf{B}_{\boldsymbol{\Psi}}\mathbf{A}^T)^{-1} \mathbf{A}\mathbf{C}_{\boldsymbol{\Psi}}\mathbf{D}, \tag{5.11}$$

it suffices to show $\mathbf{B}_{\boldsymbol{\Psi}}^{-1} \geq \mathbf{A}^T(\mathbf{A}\mathbf{B}_{\boldsymbol{\Psi}}\mathbf{A}^T)^{-1}\mathbf{A}$ which holds by Lemma 5.1.1. □

Under the same regularity conditions as mentioned in Chapter 2, the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}^*(\mathbf{X}_i; \boldsymbol{\eta}) = \mathbf{0} \tag{5.12}$$

will have a solution $\hat{\boldsymbol{\eta}}_n = \hat{\boldsymbol{\eta}}_n(\mathbf{X}_1, \ldots, \mathbf{X}_n)$ that is a consistent estimator for $\boldsymbol{\eta}$ when the model can be correctly parameterized by the $m-$dimensional parameter $\boldsymbol{\eta}$ and

$$\sqrt{n}\left(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}\right) \xrightarrow{d} N_m\left(\mathbf{0}, \mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\right) \tag{5.13}$$

as $n \longrightarrow \infty$. It follows from the delta method that

$$\sqrt{n}\left(\boldsymbol{\theta}\left(\hat{\boldsymbol{\eta}}_n\right) - \boldsymbol{\theta}\left(\boldsymbol{\eta}\right)\right) \xrightarrow{d} N_s\left(\mathbf{0}, \mathbf{D}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\,\mathbf{D}^T\right) \tag{5.14}$$

as $n \longrightarrow \infty$. Let $\hat{\boldsymbol{\theta}}_n$ be the solution to the estimating equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{X}_i; \boldsymbol{\theta}\right) = \mathbf{0}. \tag{5.15}$$

Then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{0}, \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\right) \tag{5.16}$$

as $n \longrightarrow \infty$. It follows from Lemma 5.1.1 that

$$\mathbf{D}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}\mathbf{D}^T = \mathbf{D}\left(\mathbf{D}^T\mathbf{I}_{\boldsymbol{\Psi}}\mathbf{D}\right)^{-1}\mathbf{D}^T \leq \mathbf{I}_{\boldsymbol{\Psi}}^{-1}. \tag{5.17}$$

This expresses the fact that in estimating a parameter by estimating functions, it is always better to parameterize the model from the very beginning with as few parameters as possible.

We now consider the problem of constructing a test statistic for the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$ for some known function $\boldsymbol{\theta}$.

5.2 Wald's test

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be a random sample from a distribution $P_{\boldsymbol{\theta}}$ in $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$. We are interested in testing the hypothesis that the model is over-parameterized, i.e. that the true model $\mathcal{P}^* = \{P^*_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H} \subseteq \mathbb{R}^m\}$ is a submodel of $\mathcal{P}$. Suppose there exists a function $\mathbf{R} : \mathbb{R}^s \longrightarrow \mathbb{R}^k$, where $k = m - s$, which links the parameters in the full model to those in the submodel, in that for every $P^*_{\boldsymbol{\eta}} \in \mathcal{P}^*$ where $P^*_{\boldsymbol{\eta}} = P_{\boldsymbol{\theta}(\boldsymbol{\eta})}$ we have $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$.

In 1943, Wald proposed a statistic based on the unrestricted maximum likelihood estimator for testing the hypothesis that $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ for some known function $\mathbf{R} : \mathbb{R}^s \longrightarrow \mathbb{R}^k$ (see [43]). If $\hat{\boldsymbol{\theta}}_n$ is the solution to the equation

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0} \tag{5.18}$$

then, under the hypothesis $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$, the statistic

$$n\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)\left(\frac{\partial \mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_n)\frac{\partial \mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right)^{-1}\mathbf{R}(\hat{\boldsymbol{\theta}}_n) \tag{5.19}$$

converges in distribution to a χ_k^2 random variable, where $k = s - m$ and $\mathbf{I}^{-1}(\boldsymbol{\theta})$ is the inverse of the Fisher information matrix.

A similar test statistic based on $\hat{\boldsymbol{\theta}}_n$, the solution to an arbitrary estimating equation

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{0} \tag{5.20}$$

can be used in place of the MLE.

Theorem 5.2.1. *Suppose $\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ is a positive definite matrix, the Jacobian*

$$\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{\partial R_1(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial R_1(\boldsymbol{\theta})}{\partial \theta_s} \\ \cdots & \cdots & \cdots \\ \frac{\partial R_k(\boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial R_k(\boldsymbol{\theta})}{\partial \theta_s} \end{bmatrix} \tag{5.21}$$

exists, is continuous in $\boldsymbol{\theta}$, and is of full rank m, and the matrix

$$\left[\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]^{-1} \tag{5.22}$$

is continuous in $\boldsymbol{\theta}$. Then, under $H_0 : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$,

1.

$$\sqrt{n}\mathbf{R}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} N_k \left(\mathbf{0}, \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \tag{5.23}$$

and

2.

$$\mathcal{W}_n = n\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n) \left[\frac{\partial \mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\hat{\boldsymbol{\theta}}_n) \frac{\partial \mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} \right]^{-1} \mathbf{R}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} \chi_k^2 \tag{5.24}$$

as $n \longrightarrow \infty$.

Proof. Let $\mathbf{\Psi}$ be an estimating function for the full model $\mathcal{P}$ and let $\hat{\boldsymbol{\theta}}_n$ be the solution to the estimating equation $\sum_{i=1}^{n} \mathbf{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}$. We can expand $\mathbf{R}(\hat{\boldsymbol{\theta}}_n)$ in a Taylor series around the point $\boldsymbol{\theta}$. For some $\boldsymbol{\theta}_n^*$ between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}$ we have

$$
\begin{aligned}
\sqrt{n}(\mathbf{R}(\hat{\boldsymbol{\theta}}_n) - \mathbf{R}(\boldsymbol{\theta})) &= \sqrt{n}\left(\mathbf{R}(\boldsymbol{\theta}) + \frac{\partial \mathbf{R}(\boldsymbol{\theta}_n^*)}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) - \mathbf{R}(\boldsymbol{\theta})\right) \\
&= \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) + \mathbf{o}_p(1) \qquad (5.25) \\
&\xrightarrow{d} N_k\left(\mathbf{0}, \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}(\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}}\right)
\end{aligned}
$$

as $n \longrightarrow \infty$. Under H_0, $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$, which proves 1. Part 2 follows from part 1, since

$$
\left(\frac{\partial \mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}(\hat{\boldsymbol{\theta}}_n)\frac{\partial \mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right)^{-1} \xrightarrow{p} \left(\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1} \qquad (5.26)
$$

as $n \longrightarrow \infty$. Then, by Slutsky's theorem,

$$
\mathcal{W}_n \xrightarrow{d} \chi_k^2. \qquad (5.27)
$$

□

For fixed $\alpha \in (0, 1)$, let $\chi_{\alpha,k}^2$ denote the critical point of the χ_k^2 distribution. The statistic $\mathcal{W}_n$ can be used as an asymptotic level α test statistic with critical region $K = (\chi_{\alpha,k}^2, \infty)$ for testing the null hypothesis $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$.

Let $\Theta_0 = \{\boldsymbol{\theta} \in \Theta : \mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}\}$ and let $\boldsymbol{\theta}^*$ be any point in Θ_0^c. The *power* of the test is defined to be

$$
P_{\boldsymbol{\theta}^*}\left(\mathcal{W}_n \in K\right). \qquad (5.28)
$$

The test is said to be *consistent* if it is asymptotically of level α and

$$
\lim_n P_{\boldsymbol{\theta}^*}\left(\mathcal{W}_n \in K\right) = 1 \qquad (5.29)
$$

for any fixed $\boldsymbol{\theta}^* \in \Theta_0^c$.

Theorem 5.2.2. *Under the same conditions as Theorem 5.2.1, the sequence of tests based on $\mathcal{W}_n$ is consistent at level α against any alternative $\boldsymbol{\theta}^* \in \Theta_0^c$.*

Proof. Fix $\boldsymbol{\theta}^* \in \Theta_0^c$. Then $\mathbf{R}(\boldsymbol{\theta}^*)$ is not equal to the zero vector. Since $\hat{\boldsymbol{\theta}}_n$ is a consistent estimator for $\boldsymbol{\theta}^*$ and

$$h(\boldsymbol{\theta}) = \mathbf{R}^T(\boldsymbol{\theta}) \left(\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) \frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{R}(\boldsymbol{\theta}) \tag{5.30}$$

is assumed continuous in $\boldsymbol{\theta}$, given $\epsilon > 0$ we have for all sufficiently large n,

$$P_{\boldsymbol{\theta}^*} \left(\left| h(\hat{\boldsymbol{\theta}}_n) - h(\boldsymbol{\theta}^*) \right| \geq \epsilon \right) < \epsilon. \tag{5.31}$$

Since

$$\left(\frac{\partial \mathbf{R}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}^*) \frac{\partial \mathbf{R}^T(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \right)^{-1} \tag{5.32}$$

is a positive definite matrix, and $\mathbf{R}(\boldsymbol{\theta}^*)$ is not the zero vector, the quadratic form

$$h(\boldsymbol{\theta}^*) = \mathbf{R}^T(\boldsymbol{\theta}^*) \left(\frac{\partial \mathbf{R}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}^*) \frac{\partial \mathbf{R}^T(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}} \right)^{-1} \mathbf{R}(\boldsymbol{\theta}^*) \tag{5.33}$$

is a positive number, say c. We have for any ϵ with $0 < \epsilon < c$,

$$0 < c - \epsilon = h(\boldsymbol{\theta}^*) - \epsilon < h(\hat{\boldsymbol{\theta}}_n) < h(\boldsymbol{\theta}^*) + \epsilon = c + \epsilon \tag{5.34}$$

with probability greater than $1 - \epsilon$ for all large n. Hence $\mathcal{W}_n = nh(\hat{\boldsymbol{\theta}}_n)$ tends to infinity in probability as $n \longrightarrow \infty$. Therefore,

$$\lim_n P_{\boldsymbol{\theta}^*} \left(\mathcal{W}_n \in K \right) = 1 \tag{5.35}$$

for any $\boldsymbol{\theta}^* \in \Theta_0^c$. $\square$

5.3 Rao's test

In 1947 Rao proposed an alternative to the Wald test statistic based only on the restricted MLE (see [32], p. 417). Suppose the parameter $\boldsymbol{\theta}$ is a function of an $m-$dimensional parameter $\boldsymbol{\eta}$. The restricted maximum likelihood estimator $\hat{\boldsymbol{\eta}}_n$ is the solution to the equation

$$\sum_{i=1}^{n} \mathbf{D}^T(\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\theta}} \log f(\mathbf{X}_i; \boldsymbol{\theta}(\boldsymbol{\eta})) = \mathbf{0}. \tag{5.36}$$

Rao showed that the efficient score statistic, given by

$$\frac{1}{n}\left(\sum_{i=1}^{n} \log f(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n)\right)^T \mathbf{I}^{-1}(\tilde{\boldsymbol{\theta}}_n) \left(\sum_{i=1}^{n} \log f(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n)\right), \tag{5.37}$$

where $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}}_n)$ and $\mathbf{I}(\boldsymbol{\theta})$ is the Fisher information on $\boldsymbol{\theta}$, also converges to a χ^2_k distribution.

A similar test statistic based on the estimator $\hat{\boldsymbol{\eta}}_n$ derived from the best estimating function $\boldsymbol{\Psi}^*$ for the submodel $\mathcal{P}^*$ can be constructed. Let $\boldsymbol{\Psi}$ be a Fisher estimating function for the full model and let $\hat{\boldsymbol{\theta}}_n$ be a solution to the estimating equation $\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) = \mathbf{0}$. In Section 5.1 we found the best estimating function $\boldsymbol{\Psi}^*$ for $\mathcal{P}^*$ based on a linear transformation of $\boldsymbol{\Psi}$ to be $\boldsymbol{\Psi}^* = \mathbf{D}^T\boldsymbol{\Psi}$. Let $\hat{\boldsymbol{\eta}}_n$ be the solution to the estimating equation $\sum_{i=1}^{n} \boldsymbol{\Psi}^*(\mathbf{X}_i; \boldsymbol{\eta}) = \sum_{i=1}^{n} \mathbf{D}^T(\boldsymbol{\eta})\boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}(\boldsymbol{\eta})) = \mathbf{0}$ and let $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}}_n)$.

Lemma 5.3.1. *If $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$ is continuously differentiable in $\boldsymbol{\eta}$ and the matrix $\mathbf{D}(\boldsymbol{\eta}) = \partial\boldsymbol{\theta}(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$ is of full rank, then under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) \xrightarrow{d} N_s(\mathbf{0}, \boldsymbol{\sigma}^2(\boldsymbol{\eta})) \text{ as } n \longrightarrow \infty, \tag{5.38}$$

where

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta}) \tag{5.39}$$

is of rank $k = s - m$.

Proof. Using the representations given in Chapter 2, we can write

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) = \frac{1}{\sqrt{n}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) + \mathbf{o}_p(1) \tag{5.40}$$

and

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \frac{1}{\sqrt{n}}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\sum_{i=1}^{n}\boldsymbol{\Psi}^*(\mathbf{X}_i;\boldsymbol{\eta}) + \mathbf{o}_p(1). \tag{5.41}$$

Since $\mathbf{D}(\boldsymbol{\eta})$ is assumed continuous in $\boldsymbol{\eta}$, we can take a Taylor series expansion of $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}(\hat{\boldsymbol{\eta}}_n)$ around the point $\boldsymbol{\eta}$ to get

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) &= \sqrt{n}\left(\boldsymbol{\theta}(\boldsymbol{\eta}) + \frac{\partial\boldsymbol{\theta}(\boldsymbol{\eta}_n^*)}{\partial\boldsymbol{\eta}}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) - \boldsymbol{\theta}(\boldsymbol{\eta})\right) \\ &= \mathbf{D}(\boldsymbol{\eta})\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) + \mathbf{o}_p(1)\end{aligned} \tag{5.42}$$

for some $\boldsymbol{\eta}_n^*$ between $\hat{\boldsymbol{\eta}}_n$. Under H_0 we have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n) &= \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) - \sqrt{n}(\boldsymbol{\theta}(\hat{\boldsymbol{\eta}}_n) - \boldsymbol{\theta}(\boldsymbol{\eta})) \\ &= \frac{1}{\sqrt{n}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) - \mathbf{D}(\boldsymbol{\eta})\frac{1}{\sqrt{n}}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\sum_{i=1}^{n}\boldsymbol{\Psi}^*(\mathbf{X}_i;\boldsymbol{\eta}) + \mathbf{o}_p(1) \\ &= \frac{1}{\sqrt{n}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) + \mathbf{o}_p(1) \\ &= \left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\right)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) + \mathbf{o}_p(1) \\ &\xrightarrow{d} N_s(\mathbf{0}, \boldsymbol{\sigma}^2(\boldsymbol{\eta}))\end{aligned} \tag{5.43}$$

as $n \longrightarrow \infty$, where

$$\begin{aligned}\boldsymbol{\sigma}^2(\boldsymbol{\eta}) &= \left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\right) \\ &\qquad \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\right)^T \\ &= \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta}),\end{aligned} \tag{5.44}$$

due to the fact that $\mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta}) = \mathbf{D}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\mathbf{D}(\boldsymbol{\eta})$. Let $\mathbf{P}(\boldsymbol{\theta})$ be the symmetric square root of the positive definite matrix $\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ so that $\mathbf{P}^2(\boldsymbol{\theta}) = \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$. Simple calculations show that the matrix $\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))$ is symmetric and idempotent, hence its rank is equal to its trace. Since multiplication by a nonsingular matrix does not change the rank, the rank of $\boldsymbol{\sigma}^2(\boldsymbol{\eta})$ is the same as the rank of $\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))$ and we have

$$\begin{aligned}\text{Rank}\left(\boldsymbol{\sigma}^2(\boldsymbol{\eta})\right) &= \text{Rank}\left(\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) = \text{trace}\left(\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{P}(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) \\ &= \text{trace}\left(\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{P}^2(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) = \text{trace}\left(\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) \\ &= \text{trace}\left(\mathbf{I}_{s\times s} - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) \\ &= s - \text{trace}\left(\mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\right) \\ &= s - \text{trace}\left(\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\mathbf{D}(\boldsymbol{\eta})\right) \\ &= s - \text{trace}\left(\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta})\right) = s - \text{trace}\left(\mathbf{I}_{m\times m}\right) \\ &= s - m = k.\end{aligned} \tag{5.45}$$

□

The next Lemma gives us the distribution of a quadratic form $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ when $\mathbf{Y}$ is a Gaussian random vector with a possibly singular covariance matrix.

Lemma 5.3.2. *(Ogasawara and Takahashi, 1951) Let* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. A necessary and sufficient condition that* $(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{A} (\mathbf{Y} - \boldsymbol{\mu})$ *have* χ^2_r *distribution is*

$$\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma} \tag{5.46}$$

and

$$r = Rank(\mathbf{A}\boldsymbol{\Sigma}) \tag{5.47}$$

Proof. See [32], p. 188. □

Theorem 5.3.3. *Under the conditions of Lemma 5.3.1, if* $\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})$ *is continuous in* $\boldsymbol{\theta}$ *and*

$$\sup_{\boldsymbol{\theta}\in\Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) - E\left(\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}) \right) \right\| = o_p(1) \tag{5.48}$$

then under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$*,*

$$\mathcal{R}_n = \frac{1}{n} \left(\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) \right)^T \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n) \left(\sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) \right) \xrightarrow{d} \chi^2_k \tag{5.49}$$

as $n \longrightarrow \infty$*.*

Proof. We can take a Taylor series expansion of $\boldsymbol{\Psi}(\mathbf{X}_i, \tilde{\boldsymbol{\theta}}_n)$ around the point $\hat{\boldsymbol{\theta}}_n$. Since $\sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{X}_i; \hat{\boldsymbol{\theta}}_n\right) = \mathbf{0}$, we have for some $\boldsymbol{\theta}_n^*$ between $\tilde{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\theta}}_n$,

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}(X_i; \hat{\boldsymbol{\theta}}_n) + \frac{\sqrt{n}}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{\Psi}(\mathbf{X}_i; \boldsymbol{\theta}_n^*)(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n) \\ &= -\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\sqrt{n}(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n) + \mathbf{o}_p(1) \\ &\xrightarrow{d} N_s(\mathbf{0}, \boldsymbol{\sigma}^2(\boldsymbol{\theta})) \text{ as } n \longrightarrow \infty, \end{aligned} \tag{5.50}$$

by Lemma 5.3.1 and (5.48), where

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta})). \tag{5.51}$$

Direct calculations show that

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = \boldsymbol{\sigma}^2(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta}) \tag{5.52}$$

and $\text{Rank}\big(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))\boldsymbol{\sigma}^2(\boldsymbol{\eta})\big) = k$. Since $\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})$ is assumed continuous, by Lemma (5.3.2) and Slutsky's theorem,

$$\frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n)\left(\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n)\right) \xrightarrow{d} \chi_k^2 \text{ as } n \longrightarrow \infty. \tag{5.53}$$

□

It is well known that despite their very different constructions, the Rao score test statistic (using the restricted MLE) is asymptotically equivalent to the Wald test statistic (based on the unrestricted MLE) when the hypothesis $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ is equivalent to the hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$, in the sense that the difference of the two statistics converges in probability to zero (see [38]). The choice of which statistic to use depends on the structure of the null hypothesis. We will now show that the analogues of the Rao and Wald test statistics given in Theorems 5.2.1 and 5.3.3 are also asymptotically equivalent.

Theorem 5.3.4. *Suppose the hypothesis* $\mathbf{R}(\boldsymbol{\theta}) = \mathbf{0}$ *is equivalent to the hypothesis that* $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$. *Then*

$$\mathcal{R}_n - \mathcal{W}_n = o_p(1) \tag{5.54}$$

as $n \longrightarrow \infty$.

Proof. Since $\mathbf{R}(\boldsymbol{\theta}(\boldsymbol{\eta})) = \mathbf{0}$ for any $\boldsymbol{\eta} \in \mathcal{H}$ it follows that $\mathbf{R}(\tilde{\boldsymbol{\theta}}_n) = \mathbf{R}(\boldsymbol{\theta}(\hat{\boldsymbol{\eta}}_n)) = \mathbf{0}$. Taking a Taylor series expansion around the point $\hat{\boldsymbol{\theta}}_n$ gives for some $\boldsymbol{\theta}_n^*$ between $\hat{\boldsymbol{\theta}}_n$

and $\tilde{\boldsymbol{\theta}}_n$

$$\begin{aligned}\mathbf{0} = \mathbf{R}(\tilde{\boldsymbol{\theta}}_n) &= \mathbf{R}(\hat{\boldsymbol{\theta}}_n) + \frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}_n^*)(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n) \\ &= \mathbf{R}(\hat{\boldsymbol{\theta}}_n) - \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n) + \mathbf{o}_p(1)\end{aligned} \tag{5.55}$$

since $\tilde{\boldsymbol{\theta}}_n - \hat{\boldsymbol{\theta}}_n = \mathbf{o}_p(1)$ under H_0. Then the statistic $\mathcal{W}_n$ can be written as

$$\begin{aligned}\mathcal{W}_n &= n\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)\left(\frac{\partial \mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\hat{\boldsymbol{\theta}}_n)\frac{\partial \mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right)^{-1}\mathbf{R}(\hat{\boldsymbol{\theta}}_n) \\ &= n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)^T \frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\frac{\partial \mathbf{R}^T(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) + o_p(1)\end{aligned} \tag{5.56}$$

while the statistic $\mathcal{R}_n$ can be written as

$$\begin{aligned}\mathcal{R}_n &= \frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n)\left(\sum_{i=1}^{n}\boldsymbol{\Psi}(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n)\right) \\ &= n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)^T \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) + o_p(1).\end{aligned} \tag{5.57}$$

From Lemma 5.3.1, $\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) \xrightarrow{d} \mathbf{Z} \sim N_s\left(\mathbf{0}, \boldsymbol{\sigma}^2(\boldsymbol{\eta})\right)$ where $\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})$ is of rank k. We can write $\mathbf{Z} = \mathbf{B}(\boldsymbol{\eta})\mathbf{Y}$ where $\mathbf{Y} \sim N_k\left(\mathbf{0}, \mathbf{I}_{k\times k}\right)$ and $\mathbf{B}(\boldsymbol{\eta})$ is an $s\times k$ matrix of rank k such that $\mathbf{B}(\boldsymbol{\eta})\mathbf{B}^T(\boldsymbol{\eta}) = \boldsymbol{\sigma}^2(\boldsymbol{\eta})$. Then $\mathcal{R}_n - \mathcal{W}_n$ can be written as (dropping the argument $\boldsymbol{\theta}(\boldsymbol{\eta})$),

$$\begin{aligned}\mathcal{R}_n - \mathcal{W}_n &= n\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right)^T\left[\mathbf{I}_{\boldsymbol{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) + o_p(1) \\ &= \mathbf{Y}^T\mathbf{B}^T\left[\mathbf{I}_{\boldsymbol{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\boldsymbol{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\mathbf{B}\mathbf{Y} + o_p(1) \\ &= \mathbf{Y}^T\mathbf{A}\mathbf{Y} + o_p(1).\end{aligned} \tag{5.58}$$

The matrix $\mathbf{A}$ is clearly symmetric. Since $\mathbf{R}(\boldsymbol{\theta}(\boldsymbol{\eta})) = \mathbf{0}$ for all $\boldsymbol{\eta} \in \mathcal{H}$ it follows that

$$\frac{\partial \mathbf{R}(\boldsymbol{\theta}(\boldsymbol{\eta}))}{\partial \boldsymbol{\eta}} = \frac{\partial \mathbf{R}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\frac{\partial \boldsymbol{\theta}\,(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \mathbf{0}. \tag{5.59}$$

Therefore,

$$\begin{aligned}\mathbf{A}^2 = \mathbf{B}^T &\left[\mathbf{I}_{\mathbf{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\left(\mathbf{I}_{\mathbf{\Psi}}^{-1} - \mathbf{D}\mathbf{I}_{\mathbf{\Psi}^*}^{-1}\mathbf{D}^T\right) \\ &\left[\mathbf{I}_{\mathbf{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\mathbf{B} \qquad (5.60) \\ = \mathbf{B}^T &\left(\mathbf{I}_{\mathbf{\Psi}} - \mathbf{I}_{\mathbf{\Psi}}\mathbf{D}\mathbf{I}_{\mathbf{\Psi}^*}^{-1}\mathbf{D}^T\mathbf{I}_{\mathbf{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right)\mathbf{B}.\end{aligned}$$

The term $\mathbf{C} = \mathbf{B}^T\mathbf{I}_{\mathbf{\Psi}}\mathbf{D}\mathbf{I}_{\mathbf{\Psi}^*}^{-1}\mathbf{D}^T\mathbf{I}_{\mathbf{\Psi}}\mathbf{B}$ is symmetric, and direct calculations show that $\mathbf{C}^2 = \mathbf{0}$. It follows that $\mathbf{C} = \mathbf{0}$. Therefore, $\mathbf{A}$ is a symmetric, idempotent matrix and rank($\mathbf{A}$) =trace($\mathbf{A}$). Using equation (5.45) gives

$$\begin{aligned}\text{Rank}(\mathbf{A}) = \text{trace}(\mathbf{A}) &= \text{trace}\left(\mathbf{B}^T\left[\mathbf{I}_{\mathbf{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\mathbf{B}\right) \\ &= \text{trace}\left(\left[\mathbf{I}_{\mathbf{\Psi}} - \frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\right]\boldsymbol{\sigma}^2\right) \\ &= k - \text{trace}\left(\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\left[\mathbf{I}_{\mathbf{\Psi}} - \mathbf{D}\mathbf{I}_{\mathbf{\Psi}^*}^{-1}\mathbf{D}^T\right]\right) \qquad (5.61) \\ &= k - \text{trace}\left(\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)^{-1}\left(\frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}}\mathbf{I}_{\mathbf{\Psi}}^{-1}\frac{\partial \mathbf{R}^T}{\partial \boldsymbol{\theta}}\right)\right) \\ &= k - k = 0.\end{aligned}$$

Therefore, $\mathbf{A} = \mathbf{0}$ and $\mathcal{R}_n - \mathcal{W}_n = o_p(1)$. □

As with the version of the Wald statistic, the version of the Rao statistic $\mathcal{R}_n$ can be used as an asymptotic level α test statistic with critical region $K = (\chi^2_{\alpha,k}, \infty)$. We now show that the sequence of tests based on $\mathcal{R}_n$ is consistent.

Theorem 5.3.5. *Let $\Theta_{\boldsymbol{\eta}} = \{\boldsymbol{\theta} \in \Theta : \boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta}),\ \boldsymbol{\eta} \in \mathcal{H}\}$. Suppose the conditions of Theorem 4.1.4 hold so that for each $\boldsymbol{\theta}^* \in \left(\overline{\Theta_{\boldsymbol{\eta}}}\right)^c$, there exists a unique point $\boldsymbol{\eta}^* \in \mathcal{H}$ such that $E_{\boldsymbol{\theta}^*}\hat{\mathbf{\Psi}}(\mathbf{x}; \boldsymbol{\eta}^*) = \mathbf{0}$. If the conditions of Theorem 5.3.3 hold, the sequence of tests based on $\mathcal{R}_n$ is consistent at level α against any alternative $\boldsymbol{\theta}^* \in \left(\overline{\Theta_{\boldsymbol{\eta}}}\right)^c$.*

Proof. Fix $\boldsymbol{\theta}^* \in \left(\overline{\Theta_{\boldsymbol{\eta}}}\right)^c$. By Theorem 4.1.4, $\hat{\boldsymbol{\eta}}_n$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\eta}^*$ and $\tilde{\boldsymbol{\theta}}_n = \boldsymbol{\theta}\left(\hat{\boldsymbol{\eta}}_n\right)$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)$. Then

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) &= \sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\right) - \sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)\right) + \sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)\right) \\ &= \sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)\right) + \mathbf{O}_p(1) \\ &= \sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p(1/\sqrt{n})\right)\end{aligned} \tag{5.62}$$

Then

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\Psi}\left(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n\right) &= -\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \tilde{\boldsymbol{\theta}}_n\right) + \mathbf{o}_p(1) \\ &= -\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\sqrt{n}\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p\left(1/\sqrt{n}\right)\right) + \mathbf{o}_p(1),\end{aligned} \tag{5.63}$$

and the quadratic form $\mathcal{R}_n$ becomes

$$\mathcal{R}_n = n\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p\left(1/\sqrt{n}\right)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p\left(1/\sqrt{n}\right)\right) + \mathbf{o}_p(1). \tag{5.64}$$

The vector $\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)$ is not the zero vector, so since $\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})$ is assumed positive definite,

$$\begin{aligned}&\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p\left(1/\sqrt{n}\right)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right) + \mathbf{O}_p\left(1/\sqrt{n}\right)\right) \\ &\qquad \xrightarrow{p} \left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta})\left(\boldsymbol{\theta}^* - \boldsymbol{\theta}\left(\boldsymbol{\eta}^*\right)\right) > 0\end{aligned} \tag{5.65}$$

and it follows that

$$\lim_n P_{\boldsymbol{\theta}^*}\left(\mathcal{R}_n \in K\right) = 1 \tag{5.66}$$

for any fixed $\boldsymbol{\theta}^* \in \left(\overline{\Theta_{\boldsymbol{\eta}}}\right)^c$. □

As an example, consider the family $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ and the submodel $\mathcal{P}' = \{P'_{\boldsymbol{\eta}} : \boldsymbol{\eta} \in \mathcal{H} \subseteq \mathbb{R}^m\}$ of $\mathcal{P}$, where the submodel is related to the full model through the function

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \left(\eta_1, \cdots, \eta_m, 0, \cdots, 0\right)^T = \left(\boldsymbol{\eta}^T, \mathbf{0}^T\right)^T. \tag{5.67}$$

Suppose the density functions $f(\mathbf{x};\boldsymbol{\theta})$ exist (with respect to the measure μ) and let $\boldsymbol{\Psi}(\mathbf{x};\boldsymbol{\theta})$ be a Fisher estimating function for the full model $\mathcal{P}$. Using the methods of Section 5.1, we can see that the best estimating function $\boldsymbol{\Psi}^*(\mathbf{x};\boldsymbol{\eta})$ for $\mathcal{P}'$ is given by

$$\boldsymbol{\Psi}^*\left(\mathbf{x};\boldsymbol{\eta}\right) = \mathbf{D}^T(\boldsymbol{\eta})\boldsymbol{\Psi}\left(\mathbf{x};\boldsymbol{\theta}(\boldsymbol{\eta})\right) = \begin{bmatrix} \Psi_1\left(\mathbf{x};\boldsymbol{\eta},\mathbf{0}\right) \\ \cdots \\ \Psi_m\left(\mathbf{x};\boldsymbol{\eta},\mathbf{0}\right) \end{bmatrix}, \tag{5.68}$$

since

$$\mathbf{D}(\boldsymbol{\eta}) = \frac{\partial\boldsymbol{\theta}(\boldsymbol{\eta})}{\partial\boldsymbol{\eta}} = \begin{bmatrix} \mathbf{I}_{m\times m} & \mathbf{0}_{m\times k} \\ \mathbf{0}_{k\times m} & \mathbf{0}_{k\times k} \end{bmatrix} \tag{5.69}$$

where $k = s - m$. Let $\hat{\boldsymbol{\theta}}_n$, $\hat{\boldsymbol{\eta}}_n$ and $\tilde{\boldsymbol{\theta}}_n$ be defined as in previous sections. Then by Theorem 5.3.3, under the hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}(\boldsymbol{\eta})$,

$$\frac{1}{n}\left(\sum_{i=1}^{n}\boldsymbol{\Psi}\left(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n\right)\right)^T \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n)\left(\sum_{i=1}^{n}\boldsymbol{\Psi}\left(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n\right)\right) \xrightarrow{d} \chi^2_k \tag{5.70}$$

as $n \longrightarrow \infty$.

Notice that under the null hypothesis, the density can be written as

$$f(\mathbf{x};\theta_1,\cdots,\theta_m,0,\ldots,0), \tag{5.71}$$

returning us to the setup of Chapter 4. We can use the results of Chapter 4 to find the behavior of the statistics $\mathcal{W}_n$ and $\mathcal{R}_n$ under the sequence of alternative hypotheses $H_n : \mathbf{X}_1,\ldots,\mathbf{X}_n$ are i.i.d. with density

$$f(\mathbf{x};\theta_1,\ldots,\theta_m,c_{m+1}/\sqrt{n},\ldots,c_s/\sqrt{n}). \tag{5.72}$$

To find the limiting distribution of $\mathcal{W}_n$ and $\mathcal{R}_n$ under the sequence H_n, we will make use of the following Lemma:

Lemma 5.3.6. *Let* $\mathbf{Y} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. *A set of necessary and sufficient conditions for* $\mathbf{Y}^T\mathbf{A}\mathbf{Y}$ *to follow a non-central* χ^2 *distribution with* r *degrees of freedom and non-centrality parameter* δ *is*

$$r = Rank(\mathbf{A}\boldsymbol{\Sigma}), \tag{5.73}$$

$$\boldsymbol{\mu}^T(\mathbf{A}\boldsymbol{\Sigma})^2 = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\Sigma}, \tag{5.74}$$

and

$$\delta = \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}. \tag{5.75}$$

Proof. See [41]. □

Let

$$\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}(\boldsymbol{\eta}) = \left[\tilde{\mathbf{I}}^1_{\boldsymbol{\Psi}}(\boldsymbol{\eta}) \quad \tilde{\mathbf{I}}^2_{\boldsymbol{\Psi}}(\boldsymbol{\eta})\right] = \left[E\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}_1^T\right) \quad E\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}_2^T\right)\right] = \left[E\left(\boldsymbol{\Psi}\mathring{\mathbf{l}}^T\right)\right] \tag{5.76}$$

as in Chapter 4.

Theorem 5.3.7. *Suppose the conditions of Theorem 5.2.1 and Theorem 5.3.3 hold. If the square root of the density* $f(\mathbf{x};\boldsymbol{\theta})$ *is differentiable in quadratic mean at the point* $\boldsymbol{\xi}^T = (\boldsymbol{\eta}^T, \mathbf{0}^T)$ *then under* $H_n : \boldsymbol{\theta}_n^T = (\boldsymbol{\eta}^T, \boldsymbol{\delta}_n^T)$,

$$\mathcal{W}_n = n\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)\left(\frac{\partial\mathbf{R}(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\mathbf{I}^{-1}_{\boldsymbol{\Psi}}(\hat{\boldsymbol{\theta}}_n)\frac{\partial\mathbf{R}^T(\hat{\boldsymbol{\theta}}_n)}{\partial\boldsymbol{\theta}}\right)^{-1}\mathbf{R}(\hat{\boldsymbol{\theta}}_n) \xrightarrow{d} \chi^2_{k,\gamma} \tag{5.77}$$

and

$$\mathcal{R}_n = \frac{1}{n}\left(\sum_{i=1}^n \boldsymbol{\Psi}\left(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n\right)\right)^T \mathbf{I}^{-1}_{\boldsymbol{\Psi}}(\tilde{\boldsymbol{\theta}}_n)\left(\sum_{i=1}^n \boldsymbol{\Psi}\left(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n\right)\right) \xrightarrow{d} \chi^2_{k,\gamma}, \tag{5.78}$$

where

$$\gamma = \mathbf{c}^T\tilde{\mathbf{I}}^{2T}_{\boldsymbol{\Psi}}\mathbf{I}_{\boldsymbol{\Psi}}\left(\mathbf{I}^{-1}_{\boldsymbol{\Psi}} - \mathbf{D}\mathbf{I}^{-1}_{\boldsymbol{\Psi}^*}\mathbf{D}^T\right)\mathbf{I}_{\boldsymbol{\Psi}}\tilde{\mathbf{I}}^2_{\boldsymbol{\Psi}}\mathbf{c}. \tag{5.79}$$

Furthermore,

$$\mathcal{W}_n - \mathcal{R}_n = o_p(1). \tag{5.80}$$

Proof. If the square root of the density $f(\mathbf{x};\boldsymbol{\theta})$ is differentiable in quadratic mean at the point $\boldsymbol{\xi}^T = (\boldsymbol{\eta}^T, \mathbf{0}^T)$, it follows from the comment following Theorem 4.2.2 that the measures $P_n = P_{\boldsymbol{\theta}_n}$ and $Q_n \equiv Q = P_{\boldsymbol{\eta},\mathbf{0}}$ are mutually contiguous. Then, since $\tilde{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}(\boldsymbol{\eta})$ in $Q-$probability, it follows that $\tilde{\boldsymbol{\theta}}_n \longrightarrow \boldsymbol{\theta}(\boldsymbol{\eta})$ in P_n-probability. To see this, fix $\epsilon > 0$, and let $A_n = \{\|\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}(\boldsymbol{\eta})\| > \epsilon\}$. Then by the definition of contiguity,

$$\lim_n Q(A_n) = 0 \Leftrightarrow \lim_n P_n(A_n) = 0. \tag{5.81}$$

Since $\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})$ is assumed continuous, $\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\tilde{\boldsymbol{\theta}}_n) \longrightarrow \mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta}))$ as $n \longrightarrow \infty$ in P_n-probability as well. Also, by Theorem 5.3.4, $\mathcal{W}_n - \mathcal{R}_n = o_p(1)$ when the true distribution is Q. Therefore, $\mathcal{W}_n - \mathcal{R}_n = o_p(1)$ under H_n as well, proving (5.80), and (5.77) will follow from (5.78).

Assume that the square root of the density $f(\mathbf{x};\boldsymbol{\theta})$ is differentiable in quadratic mean at the point $\boldsymbol{\xi} = (\eta_1, \ldots, \eta_m, 0, \ldots, 0)^T$. In the proof of Theorem 5.3.3 it was shown that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n\right) = \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) \left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}) - \mathbf{D}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}(\boldsymbol{\eta})\mathbf{D}^T\right) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) + \mathbf{o}_p(1). \tag{5.82}$$

By Theorem 4.2.1 we have

$$\prod_{i=1}^{n} \log \frac{f(\mathbf{X}_i; \boldsymbol{\theta}, \boldsymbol{\delta}_n)}{f(\mathbf{X}_i; \boldsymbol{\theta}, \mathbf{0})} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{c}^T \mathring{\mathbf{l}}_2(\mathbf{X}_i) - \frac{1}{2}\mathbf{c}^T \mathbf{I}_{22}(\boldsymbol{\xi})\mathbf{c} + \mathbf{o}_p(1). \tag{5.83}$$

where $\boldsymbol{\delta}_n = (c_{m+1}/\sqrt{n}, \ldots, c_s/\sqrt{n})^T$ and $\mathbf{c} = (c_{m+1}, \ldots, c_s)^T$. Under Q, by the multivariate central limit theorem,

$$\begin{pmatrix} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n\right) \\ \prod_{i=1}^{n} \log \frac{f(\mathbf{X}_i;\boldsymbol{\theta},\boldsymbol{\delta}_n)}{f(\mathbf{X}_i;\boldsymbol{\theta},\mathbf{0})} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \mathbf{I}_{\boldsymbol{\Psi}}\left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1} - \mathbf{D}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}\mathbf{D}^T\right)\boldsymbol{\Psi}(\mathbf{X}_i;\boldsymbol{\theta}) \\ \mathbf{c}^T \mathring{\mathbf{l}}_2(\mathbf{X_i}) \end{pmatrix}$$
$$+ \begin{pmatrix} \mathbf{0} \\ -\frac{1}{2}\mathbf{c}^T\mathbf{I}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix} + \mathbf{o}_p(1)$$
$$\xrightarrow{d} N_{s+1}\left(\begin{pmatrix} \mathbf{0} \\ -\frac{1}{2}\mathbf{c}^T\mathbf{I}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\sigma}^2(\boldsymbol{\eta}) & \boldsymbol{\Sigma}(\boldsymbol{\eta}) \\ \boldsymbol{\Sigma}^T(\boldsymbol{\eta}) & \mathbf{c}^T\mathbf{I}_{22}(\boldsymbol{\xi})\mathbf{c} \end{pmatrix} \right) \tag{5.84}$$

where

$$\boldsymbol{\sigma}^2(\boldsymbol{\eta}) = \mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\hat{\boldsymbol{\Psi}}}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\right)\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta})), \tag{5.85}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{\eta}) = \left(\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta})) - \mathbf{D}(\boldsymbol{\eta})\mathbf{I}_{\hat{\boldsymbol{\Psi}}}^{-1}(\boldsymbol{\eta})\mathbf{D}^T(\boldsymbol{\eta})\right)\mathbf{I}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}(\boldsymbol{\eta}))\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^2(\boldsymbol{\theta}(\boldsymbol{\eta}))\mathbf{c}. \tag{5.86}$$

Hence, by LeCam's Third Lemma, under P_n,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\Psi}\left(\mathbf{X}_i; \tilde{\boldsymbol{\theta}}_n\right) \xrightarrow{d} N_s\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\sigma}^2(\boldsymbol{\theta})\right). \tag{5.87}$$

Under H_n,

$$\mathcal{R}_n = \mathbf{Y}^T\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\theta})\mathbf{Y} + o_p(1) \tag{5.88}$$

where $\mathbf{Y} \sim N_s\left(\boldsymbol{\Sigma}(\boldsymbol{\theta}), \boldsymbol{\sigma}^2(\boldsymbol{\theta})\right)$. Straight forward calculations, similar to those used in the proofs of Lemma 5.3.1 and Theorem 5.3.4 show that $\boldsymbol{\Sigma}^T(\boldsymbol{\eta})\left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\eta})\boldsymbol{\sigma}^2(\boldsymbol{\eta})\right)^2 = \boldsymbol{\Sigma}^T(\boldsymbol{\eta})\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\eta})\boldsymbol{\sigma}^2(\boldsymbol{\eta})$, Rank$(\mathbf{I}_{\boldsymbol{\Psi}}^{-1}(\boldsymbol{\eta})\boldsymbol{\sigma}^2(\boldsymbol{\eta})) = k$, and $\gamma = \mathbf{c}^T\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^{2T}\mathbf{I}_{\boldsymbol{\Psi}}\left(\mathbf{I}_{\boldsymbol{\Psi}}^{-1} - \mathbf{D}\mathbf{I}_{\boldsymbol{\Psi}^*}^{-1}\mathbf{D}^T\right)\mathbf{I}_{\boldsymbol{\Psi}}\tilde{\mathbf{I}}_{\boldsymbol{\Psi}}^2\mathbf{c}$. Equation (5.78) then follows from Lemma 5.3.6. $\square$

Chapter 6

Combining estimators

6.1 Combining estimators vs. combining estimating functions

Let $\mathbf{X} = (X_1, \ldots, X_{n_1})^T$ be a random sample from a distribution $F_1(x;\theta)$ and let $\mathbf{Y} = (Y_1, \ldots, Y_{n_2})^T$ be a second random sample, independent of the first, from a distribution $F_2(y;\theta)$, where both F_1 and F_2 depend on a common univariate parameter θ. The sample sizes are of the same order in that $n_1 = \lambda_1 n$ and $n_2 = \lambda_2 n$, where λ_1 and λ_2 are positive real numbers such that $\lambda_1 + \lambda_2 = 1$.

Let $\Psi_1 = \Psi_1(x;\theta)$ be an estimating function for $\mathcal{P}_1 = \{F_1(x;\theta) : \theta \in \Theta\}$ and let $\hat{\theta}_1 = \hat{\theta}_1(X_1, \ldots, X_{n_1})$ be a solution to the estimating equation

$$\sum_{i=1}^{n_1} \Psi_1(X_i;\theta) = 0 \tag{6.1}$$

such that

$$\sqrt{n_1}\left(\hat{\theta}_1 - \theta\right) = \frac{1}{n_1} C_{\Psi_1}^{-1}(\theta) \sum_{i=1}^{n_1} \Psi_1(X_i;\theta) + o_p(1). \tag{6.2}$$

Similarly, let $\Psi_2 = \Psi_2(y;\theta)$ be an estimating function for $\mathcal{P}_2 = \{F_2(y;\theta) : \theta \in \Theta\}$ and let $\hat{\theta}_2 = \hat{\theta}_2(Y_1, \ldots, Y_{n_2})$ be a solution to the estimating equation

$$\sum_{i=1}^{n_2} \Psi_2(Y_i;\theta) = 0 \tag{6.3}$$

such that

$$\sqrt{n_2}\left(\hat{\theta}_2 - \theta\right) = \frac{1}{n_2} C_{\Psi_2}^{-1}(\theta) \sum_{i=1}^{n_2} \Psi_2(Y_i;\theta) + o_p(1). \tag{6.4}$$

The question we will address is how to combine the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ in the optimal way, and how that optimal estimator compares to the estimator based on the combined data from the two independent samples.

For any non-negative functions $w_1(\theta)$ and $w_2(\theta)$ with $w_1(\theta) + w_2(\theta) = 1$, the linear combination $w_1(\theta)\hat{\theta}_1 + w_2(\theta)\hat{\theta}_2$ will be asymptotically Gaussian:

$$\begin{aligned}\sqrt{n}\left(w_1(\theta)\hat{\theta}_1 + w_2(\theta)\hat{\theta}_2 - \theta\right) &= \frac{w_1(\theta)}{\sqrt{\lambda_1}}\sqrt{n_1}\left(\hat{\theta}_1 - \theta\right) + \frac{w_2(\theta)}{\sqrt{\lambda_2}}\sqrt{n_2}\left(\hat{\theta}_2 - \theta\right) \\ &\xrightarrow{d} N\left(0, \sigma^2(\theta)\right)\end{aligned} \tag{6.5}$$

as $n \longrightarrow \infty$ for

$$\sigma^2(\theta) = \frac{w_1^2(\theta)}{\lambda_1 I_{\Psi_1}(\theta)} + \frac{w_2^2(\theta)}{\lambda_2 I_{\Psi_2}(\theta)}. \tag{6.6}$$

The variance $\sigma^2(\theta)$ is minimized when

$$w_1(\theta) = \frac{\lambda_1 I_{\Psi_1}(\theta)}{\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta)} \tag{6.7}$$

and $w_2(\theta) = 1 - w_1(\theta)$, and the minimal variance is

$$\frac{1}{\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_{2}(\theta)}}, \tag{6.8}$$

the inverse of the information in the combined sample. Clearly,

$$\frac{1}{\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta)} \leq \min\left(\frac{1}{\lambda_1 I_{\Psi_1}(\theta)}, \frac{1}{\lambda_2 I_{\Psi_2}(\theta)}\right) \tag{6.9}$$

so that optimal linear combination of $\hat{\theta}_1$ and $\hat{\theta}_2$ is better than either $\hat{\theta}_1$ or $\hat{\theta}_2$.

Let θ_n^* be an initial consistent estimate of θ. For example, we can choose $\theta_n^* = \hat{\theta}_1/2 + \hat{\theta}_2/2$. We will estimate the weight functions with $w_1(\theta_n^*)$ and $w_2(\theta_n^*)$.

Theorem 6.1.1. *If the weight function $w_1(\theta)$ given by* (6.7) *is continuously differentiable, then*

$$\sqrt{n}\left(w_1(\theta_n^*)\hat{\theta}_1 + w_2(\theta_n^*)\hat{\theta}_2 - \theta\right) \xrightarrow{d} N\left(0, \frac{1}{\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta)}\right) \tag{6.10}$$

as $n \longrightarrow \infty$.

Proof. Expand the function $w_1(\theta_n^*)$ in a Taylor series around the point θ to get

$$w_1(\theta_n^*) = w_1(\theta) + \frac{dw_1(\tilde{\theta}_1)}{d\theta}(\theta_n^* - \theta) \tag{6.11}$$

for some $\tilde{\theta}_1$ between θ and θ_n^*. Since $w_1(\theta) + w_2(\theta) = 1$ for all θ, it follows that $dw_1(\theta)/d\theta = -dw_2(\theta)/d\theta$. We can then rewrite $\sqrt{n}\left(w_1(\theta_n^*)\hat{\theta}_1 + w_2(\theta_n^*)\hat{\theta}_2 - \theta\right)$ as

$$\begin{aligned}
\sqrt{n}\bigg(w_1(\theta_n^*)\hat{\theta}_1 &+ (1 - w_1(\theta_n^*))\hat{\theta}_2 - \theta\bigg) \\
&= \sqrt{n}\Bigg[\left(w_1(\theta) + \frac{dw_1(\tilde{\theta}_1)}{d\theta}(\theta_n^* - \theta)\right)\hat{\theta}_1 - w_1(\theta)\theta \\
&\quad + \left(1 - w_1(\theta) - \frac{dw_1(\tilde{\theta}_1)}{d\theta}(\theta_n^* - \theta)\right)\hat{\theta}_2 - w_2(\theta)\theta\Bigg] \\
&= w_1(\theta)\sqrt{n}\left(\hat{\theta}_1 - \theta\right) + w_2(\theta)\sqrt{n}\left(\hat{\theta}_2 - \theta\right) \\
&\quad + \frac{dw_1(\tilde{\theta}_1)}{d\theta}\sqrt{n}(\theta_n^* - \theta)(\hat{\theta}_1 - \hat{\theta}_2) \\
&= \frac{w_1(\theta)}{\lambda_1}\sqrt{n_1}\left(\hat{\theta}_1 - \theta\right) + \frac{w_2(\theta)}{\lambda_2}\sqrt{n_2}\left(\hat{\theta}_2 - \theta\right) + o_p(1)
\end{aligned} \tag{6.12}$$

since $(\theta_n^* - \theta) = o_p(1)$ and $dw_1(\tilde{\theta}_1)/d\theta\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_2) = O_p(1)$. Also,

$$\frac{w_1(\theta)}{\lambda_1}\sqrt{n_1}\left(\hat{\theta}_1 - \theta\right) \xrightarrow{d} N\left(0, \frac{w_1^2(\theta)}{\lambda_1^2}\lambda_1 I_{\Psi_1}^{-1}(\theta)\right) \tag{6.13}$$

while

$$\frac{w_2(\theta)}{\lambda_2}\sqrt{n_2}\left(\hat{\theta}_2 - \theta\right) \xrightarrow{d} N\left(0, \frac{w_2^2(\theta)}{\lambda_2^2}\lambda_2 I_{\Psi_2}^{-1}(\theta)\right). \tag{6.14}$$

Since $\hat{\theta}_1$ and $\hat{\theta}_2$ are independent, we have

$$\sqrt{n}\left(w_1(\theta_n^*)\hat{\theta}_1 + w_2(\theta_n^*)\hat{\theta}_2 - \theta\right) \xrightarrow{d} N\left(0, \sigma^2(\theta)\right) \tag{6.15}$$

where

$$\sigma^2(\theta) = \frac{w_1^2(\theta)}{\lambda_1 I_{\Psi_1}(\theta)} + \frac{w_2^2(\theta)}{\lambda_2 I_{\Psi_2}(\theta)}$$
$$= \frac{\lambda_1 I_{\Psi_1}(\theta)}{(\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta))^2} + \frac{\lambda_2 I_{\Psi_2}(\theta)}{(\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta))^2} \tag{6.16}$$
$$= \frac{1}{\lambda_1 I_{\Psi_1}(\theta) + \lambda_2 I_{\Psi_2}(\theta)}.$$

□

Now, suppose the common parameter $\boldsymbol{\theta}$ is $s-$dimensional. To combine the estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ in the optimal way, we should use the $(s \times s)$ matrices

$$\mathbf{w}_1(\boldsymbol{\theta}) = \lambda_1 \left(\lambda_1 \mathbf{I}_{\Psi_1}(\boldsymbol{\theta}) + \lambda_2 \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1} \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) \tag{6.17}$$

and $\mathbf{w}_2(\boldsymbol{\theta}) = \mathbf{I}_{s\times s} - \mathbf{w}_1(\boldsymbol{\theta})$ as the weight functions, so that

$$\mathbf{w}_1(\boldsymbol{\theta}) \left(\lambda_1 \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\right)^{-1} \mathbf{w}_1^T(\boldsymbol{\theta}) + \mathbf{w}_2(\boldsymbol{\theta}) \left(\lambda_2 \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1} \mathbf{w}_2^T(\boldsymbol{\theta}) = \left(\lambda_1 \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \lambda_2 \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1}. \tag{6.18}$$

It follows by an identical argument as was used in the proof of (6.1.1) that if $\mathbf{w}_1(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_n^*$ is an initial consistent estimator of $\boldsymbol{\theta}$, that

$$\sqrt{n}\left(\mathbf{w}_1(\boldsymbol{\theta}_n^*)\hat{\boldsymbol{\theta}}_1 + \mathbf{w}_2(\boldsymbol{\theta}_n^*)\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{0}, \left(\lambda_1 \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \lambda_2 \mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1}\right) \tag{6.19}$$

as $n \longrightarrow \infty$.

To compare the best linear combination of the estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ to an estimator $\hat{\boldsymbol{\theta}}_n$ using the entire combined sample $\mathbf{Z}^T = (\mathbf{X}^T, \mathbf{Y}^T)$ we need to first investigate how to combine the estimating functions $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$.

Theorem 6.1.2. *Let* $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_{n_1})$ *be a random sample with distribution function* $F_1(\mathbf{x};\boldsymbol{\theta})$*, with* $\boldsymbol{\theta} \in \mathbb{R}^s$*, and let* $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_1(\mathbf{X};\boldsymbol{\theta}) = \sum_{i=1}^{n_1} \boldsymbol{\Psi}_1(\mathbf{X}_i;\boldsymbol{\theta})$ *be an* *estimating function for* $\boldsymbol{\theta}$ *based on the sample* $\mathbf{X}$*. Let* $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2})$ *be a second random sample, independent of* $\mathbf{X}$*, with distribution function* $F_2(\mathbf{y};\theta)$*, and let* $\boldsymbol{\Psi}_2 = \boldsymbol{\Psi}_2(\mathbf{Y};\boldsymbol{\theta}) = \sum_{i=1}^{n_2} \boldsymbol{\Psi}_2(\mathbf{Y}_i;\boldsymbol{\theta})$ *be an estimating function for* $\boldsymbol{\theta}$ *based on the random sample* $\mathbf{Y}$*. Let* $\mathbf{Z}^T = (\mathbf{X}^T, \mathbf{Y}^T)$ *represent the combined samples. The best estimating function* $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{Z};\boldsymbol{\theta})$ *for* $\boldsymbol{\theta}$ *based on the combined samples* $\mathbf{Z}$ *in the class of estimating functions* $\boldsymbol{\Psi}^*(\mathbf{Z};\boldsymbol{\theta}) = \mathbf{A}_1(\boldsymbol{\theta})\boldsymbol{\Psi}_1 + \mathbf{A}_2(\boldsymbol{\theta})\boldsymbol{\Psi}_2$ *is*

$$\hat{\boldsymbol{\Psi}} = \hat{\boldsymbol{\Psi}}\,(\mathbf{Z};\boldsymbol{\theta}) = \mathbf{C}^T_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\boldsymbol{\Psi}_1(\mathbf{X};\boldsymbol{\theta}) + \mathbf{C}^T_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\boldsymbol{\Psi}_2(\mathbf{Y};\boldsymbol{\theta}). \qquad (6.20)$$

Proof. First notice that

$$\mathbf{C}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}) = \mathbf{C}^T_{\boldsymbol{\Psi}_1}\mathbf{B}^{-1}_{\boldsymbol{\Psi}_1}\mathbf{C}_{\boldsymbol{\Psi}_1} + \mathbf{C}^T_{\boldsymbol{\Psi}_2}\mathbf{B}^{-1}_{\boldsymbol{\Psi}_2}\mathbf{C}_{\boldsymbol{\Psi}_2} = \mathbf{I}_{\boldsymbol{\Psi}_1} + \mathbf{I}_{\boldsymbol{\Psi}_2} = \mathbf{B}_{\boldsymbol{\Psi}}(\boldsymbol{\theta}). \qquad (6.21)$$

Therefore $\boldsymbol{\Psi}$ defines an estimating function (since $\mathbf{C}_{\boldsymbol{\Psi}}$ must be nonsingular, due to the positive definiteness of $\mathbf{I}_{\boldsymbol{\Psi}_1}$ and $\mathbf{I}_{\boldsymbol{\Psi}_2}$) that is in Fisher form.

Let $\boldsymbol{\Psi}^* = \boldsymbol{\Psi}^*(\mathbf{X},\mathbf{Y};\boldsymbol{\theta}) = \mathbf{A}_1(\boldsymbol{\theta})\boldsymbol{\Psi}_1 + \mathbf{A}_2(\boldsymbol{\theta})\boldsymbol{\Psi}_2$ be any linear combination of the estimating functions $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$. To prove the optimality of the estimating function $\boldsymbol{\Psi}$ we need to show that

$$\begin{aligned}
\mathbf{I}_{\boldsymbol{\Psi}} &= \mathbf{I}_{\boldsymbol{\Psi}_1} + \mathbf{I}_{\boldsymbol{\Psi}_2} \\
&\geq \left(\mathbf{A}_1\mathbf{C}_{\boldsymbol{\Psi}_1} + \mathbf{A}_2\mathbf{C}_{\boldsymbol{\Psi}_2}\right)^T \left(\mathbf{A}_1\mathbf{B}_{\boldsymbol{\Psi}_1}\mathbf{A}_1^T + \mathbf{A}_2\mathbf{B}_{\boldsymbol{\Psi}_2}\mathbf{A}_2^T\right)^{-1} \left(\mathbf{A}_1\mathbf{C}_{\boldsymbol{\Psi}_1} + \mathbf{A}_2\mathbf{C}_{\boldsymbol{\Psi}_2}\right) \qquad (6.22) \\
&= \mathbf{I}_{\boldsymbol{\Psi}^*}
\end{aligned}$$

This follows from the non-negative definiteness of the covariance matrix for the vector $\mathbf{\Psi} - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{\Psi}^*$.

$$
\begin{aligned}
\mathbf{0} &\leq E_{\boldsymbol{\theta}} \left(\mathbf{\Psi} - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{\Psi}^* \right) \left(\mathbf{\Psi} - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{\Psi}^* \right)^T \\
&= E_{\boldsymbol{\theta}} \left(\mathbf{\Psi} \mathbf{\Psi}^T \right) - E_{\boldsymbol{\theta}} \left(\mathbf{\Psi} \mathbf{\Psi}^{*T} \right) \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{C}_{\Psi^*} - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} E_\theta \left(\mathbf{\Psi}^* \mathbf{\Psi}^T \right) \\
&\qquad + \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} E_{\boldsymbol{\theta}} \left(\mathbf{\Psi}^* \mathbf{\Psi}^{*T} \right) \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{C}_{\mathbf{\Psi}^*} \\
&= \mathbf{I}_{\mathbf{\Psi}} - E_{\boldsymbol{\theta}} \left(\mathbf{C}_{\mathbf{\Psi}_1}^T \mathbf{B}_{\mathbf{\Psi}_1}^{-1} \mathbf{\Psi}_1 + \mathbf{C}_{\mathbf{\Psi}_2}^T \mathbf{B}_{\mathbf{\Psi}_2}^{-1} \mathbf{\Psi}_2 \right) \left(\mathbf{A}_1 \mathbf{\Psi}_1 + \mathbf{A}_2 \mathbf{\Psi}_2 \right)^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{C}_{\mathbf{\Psi}^*} \\
&\qquad - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} E_{\boldsymbol{\theta}} \left(\mathbf{A}_1 \mathbf{\Psi}_1 + \mathbf{A}_2 \mathbf{\Psi}_2 \right) \left(\mathbf{C}_{\mathbf{\Psi}_1}^T \mathbf{B}_{\mathbf{\Psi}_1}^{-1} \mathbf{\Psi}_1 + \mathbf{C}_{\mathbf{\Psi}_2}^T \mathbf{B}_{\mathbf{\Psi}_2}^{-1} \mathbf{\Psi}_2 \right)^T + \mathbf{I}_{\mathbf{\Psi}^*} \\
&= \mathbf{I}_{\mathbf{\Psi}} - \left(\mathbf{A}_1 \mathbf{C}_{\mathbf{\Psi}_1} + \mathbf{A}_2 \mathbf{C}_{\mathbf{\Psi}_2} \right)^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{C}_{\mathbf{\Psi}^*} \\
&\qquad - \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \left(\mathbf{A}_1 \mathbf{C}_{\mathbf{\Psi}_1} + \mathbf{A}_2 \mathbf{C}_{\mathbf{\Psi}_2} \right) + \mathbf{I}_{\mathbf{\Psi}^*} \\
&= \mathbf{I}_{\mathbf{\Psi}} - 2 \mathbf{C}_{\mathbf{\Psi}^*}^T \mathbf{B}_{\mathbf{\Psi}^*}^{-1} \mathbf{C}_{\mathbf{\Psi}^*} + \mathbf{I}_{\mathbf{\Psi}^*} = \mathbf{I}_{\mathbf{\Psi}} - \mathbf{I}_{\mathbf{\Psi}^*}
\end{aligned}
\tag{6.23}
$$

□

There is also a geometric interpretation of this result. In Chapter 2 we showed that if $\mathcal{C} = \{\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_k\}$ is a finite collection of estimating functions and $\mathcal{H} = \overline{\mathrm{Span}\{\mathbf{\Psi}_1, \ldots, \mathbf{\Psi}_k\}}$ is the closed linear span of the elements in $\mathcal{C}$, then the optimal estimating function $\mathbf{\Psi}$ in $\mathcal{H}$ is the projection of the Fisher score function onto $\mathcal{H}$. That is, if $\mathbf{\Psi} = \hat{E}_{\boldsymbol{\theta}} (\mathbf{J} \mid \mathcal{H})$, then $\mathbf{I}_{\mathbf{\Psi}}(\boldsymbol{\theta}) \geq \mathbf{I}_{\mathbf{\Psi}^*}(\boldsymbol{\theta})$ for any estimating function $\mathbf{\Psi}^* \in \mathcal{H}$. It is easy to show that

$$
\mathbf{\Psi} = \hat{E}_{\boldsymbol{\theta}} \left(\mathbf{J} \mid \mathbf{\Psi}_1,\ \mathbf{\Psi}_2 \right) = \mathbf{C}_{\mathbf{\Psi}_1}^T \mathbf{B}_{\mathbf{\Psi}_1}^{-1} \mathbf{\Psi}_1 + \mathbf{C}_{\mathbf{\Psi}_2}^T \mathbf{B}_{\mathbf{\Psi}_2}^{-1} \mathbf{\Psi}_2, \tag{6.24}
$$

giving an alternate proof of Theorem 6.1.2.

We now describe the behavior of the estimator that is the solution to the best linear combination of the estimating functions $\mathbf{\Psi}_1$ and $\mathbf{\Psi}_2$ given in Theorem 6.1.2.

Theorem 6.1.3. *Suppose $\mathcal{P}_1 = \{P_{1,\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ and $\mathcal{P}_2 = \{P_{2,\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^s\}$ are two parametric families depending on a common parameter $\boldsymbol{\theta}$. Let $\boldsymbol{\Psi}_1$ be an estimating function for $\mathcal{P}_1$ and $\boldsymbol{\Psi}_2$ be an estimating function for $\mathcal{P}_2$. Let $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{Z})$ be the solution of the estimating equation*

$$\begin{aligned}\boldsymbol{\Psi}(\mathbf{Z};\boldsymbol{\theta}) &= \mathbf{C}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\sum_{i=1}^{n_1}\boldsymbol{\Psi}_1(\mathbf{X}_i;\boldsymbol{\theta}) + \mathbf{C}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\mathbf{B}^{-1}_{\boldsymbol{\Psi}_2}\sum_{i=1}^{n_2}\boldsymbol{\Psi}_2(\mathbf{Y}_i;\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n_1}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}) + \sum_{i=1}^{n_2}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y}_i;\boldsymbol{\theta}) = \mathbf{0}\end{aligned} \tag{6.25}$$

where $\hat{\boldsymbol{\Psi}}_1$ and $\hat{\boldsymbol{\Psi}}_2$ are the Fisher forms of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$ respectively. If

$$\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_j\left(\mathbf{X}_i;\boldsymbol{\theta}\right) - E\left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_j(\mathbf{x};\boldsymbol{\theta})\right)\right\| = o_p(1) \tag{6.26}$$

for $j = 1,2$ as $n \longrightarrow \infty$ then

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{0}, \left(\lambda_1\mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \lambda_2\mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1}\right) \tag{6.27}$$

as $n \longrightarrow \infty$.

Proof. Expand $\boldsymbol{\Psi}(\mathbf{Z};\hat{\boldsymbol{\theta}}_n)$ in a Taylor series around the point $\boldsymbol{\theta}$ to get

$$\begin{aligned}\mathbf{0} = \boldsymbol{\Psi}(\mathbf{Z};\hat{\boldsymbol{\theta}}_n) &= \sum_{i=1}^{n_1}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\hat{\boldsymbol{\theta}}_n) + \sum_{i=1}^{n_2}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y};\hat{\boldsymbol{\theta}}_n) \\ &= \sum_{i=1}^{n_1}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}) + \sum_{i=1}^{n_2}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y}_i;\boldsymbol{\theta}) \\ &\quad + \left(\sum_{i=1}^{n_1}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n) + \sum_{i=1}^{n_2}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y}_i;\tilde{\boldsymbol{\theta}}_n)\right)\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)\end{aligned} \tag{6.28}$$

for some vector $\tilde{\boldsymbol{\theta}}_n$ between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}$. Rearranging gives

$$\begin{aligned}\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) &= \left(-\frac{\lambda_1}{n_1}\sum_{i=1}^{n_1}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\tilde{\boldsymbol{\theta}}_n) - \frac{\lambda_2}{n_1}\sum_{i=1}^{n_2}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y}_i;\tilde{\boldsymbol{\theta}}_n)\right)^{-1} \\ &\qquad \times \left(\frac{\sqrt{\lambda_1}}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}) + \frac{\sqrt{\lambda_2}}{\sqrt{n_2}}\sum_{i=1}^{n_2}\hat{\boldsymbol{\Psi}}_2(\mathbf{Y}_i;\boldsymbol{\theta})\right)\end{aligned} \tag{6.29}$$

Here we use the fact that for the Fisher estimating function $\hat{\boldsymbol{\Psi}}_1(\mathbf{x};\boldsymbol{\theta})$ (and similarly for $\hat{\boldsymbol{\Psi}}_2$),

$$-\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}) \xrightarrow{p} -E_{\boldsymbol{\theta}}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\boldsymbol{\Psi}}_1(\mathbf{x};\boldsymbol{\theta})\right) = \mathbf{C}_{\hat{\boldsymbol{\Psi}}_1}(\boldsymbol{\theta}) = \mathbf{I}_{\hat{\boldsymbol{\Psi}}_1}(\boldsymbol{\theta}) = \mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) \tag{6.30}$$

as $n_1 \longrightarrow \infty$. Since the first and second sample are assumed independent, and

$$\frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{\boldsymbol{\Psi}}_1(\mathbf{X}_i;\boldsymbol{\theta}) \xrightarrow{d} N_s\left(\mathbf{0}, \mathbf{I}^{-1}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta})\right) \tag{6.31}$$

as $n_1 \longrightarrow \infty$ we have

$$\sqrt{n}\left(\tilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \xrightarrow{d} N_s\left(\mathbf{0}, \left(\lambda_1\mathbf{I}_{\boldsymbol{\Psi}_1}(\boldsymbol{\theta}) + \lambda_2\mathbf{I}_{\boldsymbol{\Psi}_2}(\boldsymbol{\theta})\right)^{-1}\right) \tag{6.32}$$

as $n \longrightarrow \infty$. $\square$

This result is somewhat counterintuitive, since the limiting distribution of the estimator $\tilde{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}_n(\mathbf{X}_1,\ldots,\mathbf{X}_{n_1},\mathbf{Y}_1,\ldots,\mathbf{Y}_{n_2})$, which is the solution of the estimating equation based on the best linear combination of the estimating functions $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, is the same as the limiting distribution of the best linear combination of the estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. In general, the estimator $\tilde{\boldsymbol{\theta}}_n$ is not a function of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$. However, asymptotically, we can do as well using only the estimators $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ as we can calculating an estimator $\tilde{\boldsymbol{\theta}}_n$ based on the entire combined sample.

6.2 Estimation of a bivariate location parameter

6.2.1 The Pitman estimator of a location parameter

If $X_1,\ldots,X_n$ is a random sample from the distribution $F(x-\mu)$, where μ is a univariate location parameter, we saw in Section 3.1 that the Pitman estimator of

μ under quadratic loss is

$$t_n = \overline{X} - E_0(\overline{X} \mid X_1 - \overline{X}, \ldots, X_n - \overline{X}) \tag{6.33}$$

and that (for $n \geq 3$) $t_n = \overline{X}$ if and only if F is Gaussian.

In this section we consider samples from a distribution depending on a bivariate location parameter. Let $\boldsymbol{\theta} = (\theta_1, \theta_2)^T \in \Theta = \mathbb{R}^2$ and $(\mathcal{X}, \mathcal{A}) = (\mathbb{R}^{2n}, \mathcal{B}_{2n})$, where $\mathcal{B}_{2n}$ is the $\sigma-$algebra of Borel sets in $\mathbb{R}^{2n}$, and

$$P_{\boldsymbol{\theta}}(A) = \int_A dF(\mathbf{x} - \boldsymbol{\theta}) = \int_A dF(\mathbf{x}_1 - \boldsymbol{\theta}) \cdots dF(\mathbf{x}_n - \boldsymbol{\theta}). \tag{6.34}$$

We will be interested in estimating linear functions of the parameter $\boldsymbol{\theta}$.

Let $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ be a random sample with distribution function $F(x - \theta_1, y - \theta_2)$. Our goal is to estimate $\Delta = c_1\theta_1 + c_2\theta_2$ where c_1 and c_2 are known constants. For reasons discussed in Section 3.1, it makes sense to restrict our attention to estimators that are equivariant. For the case of the bivariate location family, an estimator $\hat{\theta} = \hat{\theta}(\mathbf{X}, \mathbf{Y})$ of $c_1\theta_1 + c_2\theta_2$ is *equivariant* if for any $a_1 \in \mathbb{R},\ a_2 \in \mathbb{R}$,

$$\hat{\theta}(\mathbf{X} + a_1, \mathbf{Y} + a_2) = \hat{\theta}(\mathbf{X}, \mathbf{Y}) + c_1 a_1 + c_2 a_2. \tag{6.35}$$

One such estimator is $c_1\overline{X} + c_2\overline{Y}$.

When using an equivariant estimator, the estimation of $c_1\theta_1 + c_2\theta_2$ by $\hat{\theta}(\mathbf{X}, \mathbf{Y})$ should be identical to the estimation of $c_1(\theta_1 + a_1) + c_2(\theta_2 + a_2)$ by $\hat{\theta}(\mathbf{X} + a_1, \mathbf{Y} + a_2)$, and this should be reflected in the loss function. We say that the loss function $L(\delta; \theta_1, \theta_2)$ is *invariant* if

$$L(\delta(\mathbf{X} + a_1, \mathbf{Y} + a_2); \theta_1 + a_1, \theta_2 + a_2) = L(\delta(\mathbf{X}, \mathbf{Y}); \theta_1, \theta_2). \tag{6.36}$$

When calculating the loss of an equivariant estimator with an invariant loss function, both the bias and the risk of the estimator are independent of the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$. To see why, let $\delta(\mathbf{X}, \mathbf{Y})$ be any equivariant estimator. Then

$$\begin{aligned}
E_{\boldsymbol{\theta}}(\delta(\mathbf{X}, \mathbf{Y})) &= \int \delta(\mathbf{x}, \mathbf{y}) f(\mathbf{x} - \theta_1, \mathbf{y} - \theta_2) d\mathbf{x} d\mathbf{y} \\
&= \int \delta(\mathbf{x} + \theta_1, \mathbf{y} + \theta_2) f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \\
&= \int (\delta(\mathbf{x}, \mathbf{y}) + c_1\theta_1 + c_2\theta_2) f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \qquad (6.37) \\
&= \int (\delta(\mathbf{x}, \mathbf{y}) f(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} + c_1\theta_1 + c_2\theta_2 \\
&= E_{\mathbf{0}}(\delta(\mathbf{X}, \mathbf{Y})) + c_1\theta_1 + c_2\theta_2
\end{aligned}$$

so that the bias, $E_{\boldsymbol{\theta}}(\delta(\mathbf{X}, \mathbf{Y}) - (c_1\theta_1 + c_2\theta_2)) = E_{\mathbf{0}}\delta(\mathbf{X}, \mathbf{Y})$, is independent of $\boldsymbol{\theta}$. A similar calculation shows that if $L(\delta(\mathbf{X}, \mathbf{Y}))$ is an invariant loss function, then $R(\delta(\mathbf{X}, \mathbf{Y})) = E_{\boldsymbol{\theta}} L(\delta(\mathbf{X}, \mathbf{Y}))$ is independent of $\boldsymbol{\theta}$. This is an important fact, because it allows us to compare equivariant estimators. That is, if δ_1 and δ_2 are equivariant, then either $R(\delta_1, \boldsymbol{\theta}) > R(\delta_2, \boldsymbol{\theta})$, $R(\delta_1, \boldsymbol{\theta}) < R(\delta_2, \boldsymbol{\theta})$, or $R(\delta_1, \boldsymbol{\theta}) = R(\delta_2, \boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$.

Following the method used in Casella and Lehmann ([9]), we now describe the class of equivariant estimators of Δ.

Lemma 6.2.1. *A function $u(\mathbf{X}, \mathbf{Y})$ satisfies*

$$u(\mathbf{X}, \mathbf{Y}) = u(\mathbf{X} + a_1, \mathbf{Y} + a_2) \qquad (6.38)$$

for any $\mathbf{a} = (a_1, a_2)^T \in \mathbb{R}^2$ if and only if

$$u(\mathbf{X}, \mathbf{Y}) = u(X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}). \qquad (6.39)$$

Proof. If u satisfies (6.39) then clearly it satisfies (6.38). If u satisfies (6.38), then we can set $a_1 = -\overline{X}$ and $a_2 = -\overline{Y}$ to get (6.39). □

Lemma 6.2.2. *If $\delta(\mathbf{X}, \mathbf{Y})$ is equivariant, then an estimator $\delta'(\mathbf{X}, \mathbf{Y})$ is equivariant if and only if*

$$\delta'(\mathbf{X}, \mathbf{Y}) = \delta(\mathbf{X}, \mathbf{Y}) + u(\mathbf{X}, \mathbf{Y}), \tag{6.40}$$

where u is of the form (6.39).

Proof. If $\delta' = \delta + u$ then

$$\begin{aligned} \delta'(\mathbf{X} + a_1, \mathbf{Y} + a_2) &= \delta(\mathbf{X} + a_1, \mathbf{Y} + a_2) + u(\mathbf{X} + a_1, \mathbf{Y} + a_2) \\ &= \delta(\mathbf{X}, \mathbf{Y}) + u(\mathbf{X}, \mathbf{Y}) + c_1 a_1 + c_2 a_2 \qquad (6.41) \\ &= \delta(\mathbf{X}, \mathbf{Y}) + c_1 a_1 + c_2 a_2. \end{aligned}$$

Conversely, if δ' is equivariant then we can set $u = \delta' - \delta$. □

One equivariant estimator is $\delta_0(\mathbf{X}, \mathbf{Y}) = c_1\overline{X} + c_2\overline{Y}$. By Lemma 6.2.2, any equivariant estimator of $\Delta = c_1\theta_1 + c_2\theta_2$ can be written as $\delta(\mathbf{X}, \mathbf{Y}) = c_1\overline{X} + c_2\overline{Y} + u(\mathbf{X}, \mathbf{Y})$. Finding the minimum risk equivariant estimator under an invariant loss function then amounts to finding a function u which minimizes $R(\delta_0 + u; \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} L(\delta_0 + u; \boldsymbol{\theta})$. But since the risk is independent of the parameter $\boldsymbol{\theta}$, it suffices to minimize $R(\delta_0 + u; \mathbf{0}) = E_{\mathbf{0}} L(\delta_0 + u; \mathbf{0})$.

We define the Pitman estimator of Δ to be the minimum risk equivariant estimator under quadratic loss, $L(\tilde{g}, g) = (\tilde{g}(\mathbf{X}) - g(\boldsymbol{\theta}))^2$. This is an invariant loss function, so to find the form of the Pitman estimator of Δ we need to find the function u which minimizes $E_{\mathbf{0}} \left(\delta_0(\mathbf{X}, \mathbf{Y}) + u(\mathbf{Z})\right)^2$, where

$$\mathbf{Z} = (X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}) \tag{6.42}$$

is the vector of residuals. Let $v^* = v^*(\mathbf{Z}) = E_{\mathbf{0}}(\delta_0(\mathbf{X}, \mathbf{Y}) \mid \mathbf{Z})$. Then

$$\begin{aligned} E_{\mathbf{0}}((\delta_0 - v)^2 \mid \mathbf{Z}) &= E_{\mathbf{0}}((\delta_0 - v^* + v^* - v)^2 \mid \mathbf{Z}) \\ &= E_{\mathbf{0}}((\delta_0 - v^*)^2 \mid \mathbf{Z}) + (v^* - v)^2 + 2E_{\mathbf{0}}((\delta_0 - v^*)(v^* - v) \mid \mathbf{Z}) \\ &= E_{\mathbf{0}}((\delta_0 - v^*)^2 \mid \mathbf{Z}) + (v^* - v)^2 + 2(v^* - v)(E_{\mathbf{0}}(\delta_0 \mid \mathbf{Z}) - v^*) \\ &= E_{\mathbf{0}}((\delta_0 - v^*)^2 \mid \mathbf{Z}) + (v^* - v)^2 \\ &\geq E_{\mathbf{0}}((\delta_0 - v^*)^2 \mid \mathbf{Z}). \end{aligned} \tag{6.43}$$

Taking expectations of both sides gives $R(\delta_0 + v^*; \mathbf{0}) \leq R(\delta; \mathbf{0})$ for any equivariant estimator δ. This gives us the form of the minimum risk equivariant estimator of $c_1\theta_1 + c_2\theta_2$ under quadratic loss.

Theorem 6.2.3. *Let $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ be a sample from a location family $F(X - \theta_1, Y - \theta_2)$. Under quadratic loss, the Pitman estimator of $\Delta = c_1\theta_1 + c_2\theta_2$ is*

$$t_n(\mathbf{X}, \mathbf{Y}) = c_1\overline{X} + c_2\overline{Y} - E_{\mathbf{0}}\left(c_1\overline{X} + c_2\overline{Y} \,\middle|\, \mathbf{Z}\right) \tag{6.44}$$

where $\mathbf{Z} = (X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$.

The Pitman estimator of θ_1 based on the observations $X_1, \ldots, X_n$ is

$$t_x = \overline{X} - E_0\left(\overline{X} \mid X_1 - \overline{X}, \ldots, X_n - \overline{X}\right) \tag{6.45}$$

and the Pitman estimator of θ_2 based on the observations $Y_1, \ldots, Y_n$ is

$$t_y = \overline{Y} - E_0\left(\overline{Y} \mid Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}\right). \tag{6.46}$$

It is of interest to consider when $t_n = c_1t_x + c_2t_y$. This is the case when the samples $\mathbf{X}$ and $\mathbf{Y}$ are independent. The proof follows directly from the following Lemma:

Lemma 6.2.4. *Let ξ, η_1, and η_2 be random variables with $E \mid X \mid < \infty$. Suppose that (ξ, η_1) and η_2 are independent. Then*

$$E\left(\xi \mid \eta_1, \eta_2\right) = E\left(\xi \mid \eta_1\right) \;\; a.s. \tag{6.47}$$

Proof. See [34], p. 35. □

If, in addition, we assume that the distribution function F has density f with respect to Lebesgue measure, the Pitman estimator (6.44) can be written in integral form

$$t(\mathbf{X},\mathbf{Y}) = \frac{\iint (c_1 s + c_2 t) \prod_{i=1}^{n} f(X_i - s, Y_i - t) ds dt}{\iint \prod_{i=1}^{n} f(X_i - s, Y_i - t) ds dt}. \tag{6.48}$$

The above form of the Pitman estimator of $c_1\theta_1 + c_2\theta_2$ follows from (6.44) by calculating the conditional density of $(\overline{X}, \overline{Y})$ given the vector of residuals $\mathbf{Z}$, and calculating the conditional expectation $E_{\mathbf{0}}\left(c_1\overline{X} + c_2\overline{Y} \,\middle|\, \mathbf{Z}\right)$.

6.2.2 Linearity of the Pitman estimator

In the class $\mathcal{F}$ of bivariate distributions F with fixed, finite covariance

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}, \tag{6.49}$$

the risk of the Pitman estimator, $E_{\boldsymbol{\theta}}(t_n - \Delta)^2$, is maximized when it is of the form $t_n(\mathbf{X}, \mathbf{Y}) = \delta_0 = c_1\overline{X} + c_2\overline{Y}$ since

$$
\begin{aligned}
\frac{c_1^2}{n}\sigma_X^2 + \frac{c_2^2}{n}\sigma_Y^2 + 2\frac{c_1c_2}{n}\sigma_{XY} &= E_{\boldsymbol{\theta}}\left(\delta_0 - \Delta\right)^2 = E_{\mathbf{0}}\left(\delta_0\right)^2 \\
&= E_{\mathbf{0}}\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right) + E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2 \\
&= E_{\mathbf{0}}\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2 + E_{\mathbf{0}}\left(E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2 \\
&\qquad + 2E_{\mathbf{0}}\left(\left[\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right] E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right) \qquad (6.50) \\
&\geq E_{\mathbf{0}}\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2 + 2E_{\mathbf{0}}\left\{E_{\mathbf{0}}\left[\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right) E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right) \mid \mathbf{Z}\right]\right\} \\
&= E_{\mathbf{0}}\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2 + 2E_{\mathbf{0}}\left\{E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right) E_{\mathbf{0}}\left[\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right) \mid \mathbf{Z}\right]\right\} \\
&= E_{\mathbf{0}}\left(\delta_0 - E_{\mathbf{0}}\left(\delta_0 \mid \mathbf{Z}\right)\right)^2
\end{aligned}
$$

If $(X, Y)^T$ is from a bivariate Gaussian distribution with mean vector $(\theta_1, \theta_2)^T$ and covariance matrix $\mathbf{\Sigma}$, then $c_1\overline{X} + c_2\overline{Y}$ is independent of the vector of residuals and the Pitman estimator will then be $c_1\overline{X} + c_2\overline{Y}$. The next Theorem characterizes the family of all bivariate distributions for which the Pitman estimator of Δ is linear, namely, the class of distribution functions F for which $t_n(\mathbf{X}, \mathbf{Y}) = c_1\overline{X} + c_2\overline{Y}$. This is equivalent to describing the class of bivariate distribution functions for which

$$
E_{\mathbf{0}}(c_1\overline{X} + c_2\overline{Y} | X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y}) = 0. \qquad (6.51)
$$

In contrast to the Kagan-Linnik-Rao theorem concerning a univariate location parameter, the family of Gaussian distributions is not the unique family of distributions for which (6.51) holds.

Theorem 6.2.5. *Let $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$, $(n \geq 3)$ be a random sample from a location family $F(x - \theta_1, y - \theta_2)$ with $Ex^2 < \infty$ and $Ey^2 < \infty$. If the Pitman* *estimator of* $\Delta = c_1\theta_1 + c_2\theta_2$ *is linear, then* F *has a characteristic function of the form*

$$\phi(t,s) = e^{Q(t,s)+h(c_2t-c_1s)} \tag{6.52}$$

in a neighborhood of $\mathbf{0}$*, where* Q *is a quadratic form in* s *and* t*, and* h *is a differentiable function with* $h(0) = 0$*.*

Proof. Let $\mathbf{Z} = (X_1 - \overline{X}, \ldots, X_n - \overline{X}, Y_1 - \overline{Y}, \ldots, Y_n - \overline{Y})$ be the vector of residuals. Suppose $E_{\mathbf{0}}(c_1\overline{X} + c_2\overline{Y}|\mathbf{Z}) = 0$. Multiplying both sides of the equation by

$$\exp\left(i\sum_{j=1}^{n}\big(t_j(X_j - \overline{X}) + s_j(Y_j - \overline{Y})\big)\right) \tag{6.53}$$

for constants t_j and s_j and taking the expectation gives

$$E_{\mathbf{0}}\left((c_1\overline{X} + c_2\overline{Y})e^{i\sum_{j=1}^{n}\left(t_j(X_j-\overline{X})+s_j(Y_j-\overline{Y})\right)}\right) = 0. \tag{6.54}$$

Since $t_1(X_1 - \overline{X}) + \cdots + t_n(X_n - \overline{X}) = X_1(t_1 - \overline{t}) + \cdots + X_n(t_n - \overline{t})$, after multiplying through by n, we can rewrite equation (6.54) as

$$\begin{aligned}
&\sum_{k=1}^{n} E_{\mathbf{0}}\left((c_1X_k + c_2Y_k)e^{i\sum_{j=1}^{n}(X_j(t_j-\overline{t})+Y_j(s_j-\overline{s}))}\right) \\
&\quad = \sum_{k=1}^{n} E_{\mathbf{0}}\left((c_1X_k + c_2Y_k)\prod_{j=1}^{n} e^{i(X_j(t_j-\overline{t})+Y_j(s_j-\overline{s}))}\right) \\
&\quad = \sum_{k=1}^{n}\left(E_{\mathbf{0}}\left((c_1X_k + c_2Y_k)e^{i(X_k(t_k-\overline{t})+Y_k(s_k-\overline{s}))}\right)\prod_{j\neq k} E_{\mathbf{0}}\left(e^{i(X_j(t_j-\overline{t})+Y_j(s_j-\overline{s}))}\right)\right) \\
&\quad = \sum_{k=1}^{n}\left(E_{\mathbf{0}}\left((c_1X_k + c_2Y_k)e^{i(X_k(t_k-\overline{t})+Y_k(s_k-\overline{s}))}\right)\prod_{j\neq i}\phi(t_j-\overline{t}, s_j-\overline{s})\right) \\
&\quad = \sum_{k=1}^{n}\left(\left(\frac{c_1}{i}\frac{\partial}{\partial t}\phi(t_k-\overline{t}, s_k-\overline{s}) + \frac{c_2}{i}\frac{\partial}{\partial s}\phi(t_k-\overline{t}, s_k-\overline{s})\right)\prod_{j\neq i}^{n}\phi(t_j-\overline{t}, s_j-\overline{s})\right) \\
&\quad = 0
\end{aligned}$$

where $\phi(t,s) = E_{\mathbf{0}} e^{itx+isy}$ is the characteristic function for the pair $(X,Y)^T$ when $\theta_1 = \theta_2 = 0$. Since the characteristic function ϕ is uniformly continuous and $\phi(0,0) = 1$, there exists a $\delta > 0$ such that for any $(t,s) \in B_\delta(\mathbf{0}) = \{(t,s) : \|(t,s)\| < \delta\}$, we have $\phi(t,s) \neq 0$. Let all points $(t_k - \bar{t}, s_k - \bar{s})$ lie in the ball $B_\delta(\mathbf{0})$ and divide through by $\prod_{k=1}^n \phi(t_k - \bar{t}, s_k - \bar{s})$ to get

$$\begin{aligned} \sum_{k=1}^{n} & \left(\frac{c_1 \frac{\partial}{\partial t} \phi(t_k - \bar{t}, s_k - \bar{s}) + c_2 \frac{\partial}{\partial s} \phi(t_k - \bar{t}, s_k - \bar{s})}{\phi(t_k - \bar{t}, s_k - \bar{s})} \right) \\ &= \sum_{k=1}^{n} \varphi(t_k - \bar{t}, s_k - \bar{s}) = 0. \end{aligned} \tag{6.55}$$

Fix (t_1, s_1) and (t_2, s_2) in $B_{\delta/2}(\mathbf{0})$ and set $t_3 = -t_1 - t_2$, $s_3 = -s_1 - s_2$, and $t_4 = \cdots = t_n = s_4 = \cdots = s_n = 0$ (so that $\bar{t} = \bar{s} = 0$ as well). Equation (6.55) reduces to

$$\varphi(t_1, s_1) + \varphi(t_2, s_2) + \varphi(-(t_1 + t_2), -(s_1 + s_2)) = 0 \tag{6.56}$$

or

$$\varphi(t_1, s_1) + \varphi(t_2, s_2) = -\varphi(-(t_1 + t_2), -(s_1 + s_2)). \tag{6.57}$$

If we let $t_1 = -t_2$ and $s_1 = -s_2$, we see that

$$\varphi(t, s) = -\varphi(-t, -s). \tag{6.58}$$

Therefore, (6.57) can be written as

$$\varphi(t_1, s_1) + \varphi(t_2, s_2) = \varphi(t_1 + t_2, s_1 + s_2). \tag{6.59}$$

We see that ϕ is linear and must have the form

$$\varphi(t, s) = c_1 \frac{\partial}{\partial t} \log \phi(t, s) + c_2 \frac{\partial}{\partial s} \log \phi(t, s) = At + Bs \tag{6.60}$$

for some constants A and B (see [1], p. 215). This is a partial differential equation of the form $c_1 u_t + c_2 u_s = At + Bs$ which has general solution (see [8])

$$u(t,s) = Q(t,s) + h(c_2 t - c_1 s) \tag{6.61}$$

where Q is quadratic in s and t and h is a differentiable function. Therefore,

$$\phi(t,s) = \exp\{Q(t,s) + h(c_2 t - c_1 s)\} \tag{6.62}$$

for $\|(t,s)\| < \delta$. $\square$

Let $(X_1, \ldots, X_n)$ be a random sample with distribution function $F(x)$. It is well known that the sample mean $\overline{X}$ is independent of the vector of residuals $\left(X_1 - \overline{X}, \ldots, X_n - \overline{X}\right)$ if and only if F is Gaussian (see [23]). The next Theorem shows that for a bivariate random sample $\left((X_1, Y_1)^T, \ldots, (X_n, Y_n)^T\right)$, the independence of $c_1\overline{X} + c_2\overline{Y}$ and the vector of residuals $\mathbf{Z}$ characterizes the same family of bivariate distributions with characteristic function of the form (6.52).

Theorem 6.2.6. *Let* $\left((X_1, Y_1)^T, \ldots, (X_n, Y_n)^T\right)$ *($n \geq 3$) be a random sample from a distribution $F(x,y)$ with $Ex^2 < \infty$ and $Ey^2 < \infty$. The linear combination $c_1\overline{X} + c_2\overline{Y}$ is independent of the vector of residuals $\mathbf{Z}$ if and only if the characteristic function of F is of the form* (6.52).

Proof. If $c_1\overline{X} + c_2\overline{Y}$ is independent of $\mathbf{Z}$, then $E\left(c_1\overline{X} + c_2\overline{Y} \mid \mathbf{Z}\right) = 0$. By Theorem 6.2.5, the characteristic function of F must be of the form (6.52).

Suppose F has a characteristic function of the form (6.52). A necessary and sufficient condition for a vector of random variables to have independent components is that its characteristic function is the product of the characteristic functions of its components (see [35], p. 286). The characteristic function of $c_1\overline{X} + c_2\overline{Y}$ is

$$\begin{aligned} E\left(\exp\left\{iw\left(c_1\overline{X}+c_2\overline{Y}\right)\right\}\right) &= E\left(\exp\left\{i\sum_{k=1}^{n}\left(wc_1\frac{X_k}{n}+wc_2\frac{Y_k}{n}\right)\right\}\right) \\ &= \prod_{k=1}^{n} E\left(\exp\left\{i\left(wc_1\frac{X_k}{n}+wc_2\frac{Y_k}{n}\right)\right\}\right) \\ &= \prod_{k=1}^{n} \exp\{Q\left(wc_1/n, wc_2/n\right) + h\left(c_2wc_1/n - c_1wc_2/n\right)\} \\ &= \prod_{k=1}^{n} \exp\left\{Q\left(wc_1/n, wc_2/n\right)\right\} \end{aligned} \tag{6.63}$$

while the characteristic function of $\mathbf{Z}$ is

$$\begin{aligned} &E\left(\exp\left\{i\sum_{k=1}^{n}\left[t_k\left(X_k-\overline{X}\right)+s_k\left(Y_k-\overline{Y}\right)\right]\right\}\right) \\ &\qquad = \prod_{k=1}^{n} E\left(\exp\left\{i\left(X_k\left(t_k-\overline{t}\right)+Y_k\left(s_k-\overline{s}\right)\right)\right\}\right) \\ &\qquad = \prod_{k=1}^{n} \exp\left\{Q\left(t_k-\overline{t}, s_k-\overline{s}\right)+h\left(c_2\left(t_k-\overline{t}\right)-c_1\left(s_k-\overline{s}\right)\right)\right\}. \end{aligned} \tag{6.64}$$

The joint characteristic function of $c_1\overline{X} + c_2\overline{Y}$ and $\mathbf{Z}$ is

$$\begin{aligned} &E\left(\exp\left\{iw\left(c_1\overline{X}+c_2\overline{Y}\right)+i\sum_{k=1}^{n}\left[t_k\left(X_k-\overline{X}\right)+s_k\left(Y_k-\overline{Y}\right)\right]\right\}\right) \\ &= \prod_{k=1}^{n} E\left(\exp\left\{iX_k\left(w\frac{c_1}{n}+t_k-\overline{t}\right)+iY_k\left(w\frac{c_2}{n}+s_k-\overline{s}\right)\right\}\right) \\ &= \prod_{k=1}^{n} \exp\left\{Q\left(wc_1/n+t_k-\overline{t}, wc_2/n+s_k-\overline{s}\right)+h\left(c_2\left(t_k-\overline{t}\right)-c_1\left(s_k-\overline{s}\right)\right)\right\}. \end{aligned} \tag{6.65}$$

Since $\sum\limits_{k=1}^{n}\left(t_k-\overline{t}\right) = \sum\limits_{k=1}^{n}\left(s_k-\overline{s}\right) = 0$,

$$\sum_{k=1}^{n} Q\left(w\frac{c_1}{n}+t_k-\overline{t}, w\frac{c_2}{n}+s_k-\overline{s}\right) = \sum_{k=1}^{n}\left(Q\left(w\frac{c_1}{n}, w\frac{c_2}{n}\right)+Q\left(t_k-\overline{t}, s_k-\overline{s}\right)\right). \tag{6.66}$$

Therefore, the characteristic function in equation (6.65) is the product of the characteristic functions in equations (6.63) and (6.64) so that $c_1\overline{X} + c_2\overline{Y}$ is independent of $\mathbf{Z}$. □

Theorem 6.2.5 is a similar result to one proved by Yu ([45]). He considered the class of bivariate distributions depending on a univariate parameter θ of the form

$$F(x-\theta, y-\theta). \tag{6.67}$$

The Pitman estimator for θ based on a sample $(X_1, Y_1)^T, \ldots, (X_n, Y_n)^T$ was found to be

$$\begin{aligned} t_n(\mathbf{X}, \mathbf{Y}) &= w_1\overline{X} + w_2\overline{Y} \\ &\quad - E_0\left(w_1\overline{X} + w_2\overline{Y} \mid X_1 - \overline{Y}, \ldots, X_n - \overline{Y}, Y_1 - \overline{X}, \ldots, Y_n - \overline{X}\right) \end{aligned} \tag{6.68}$$

for appropriate non-negative constants w_1 and w_2 such that $w_1 + w_2 = 1$. The Pitman estimator was found to be linear in this setup if and only if the distribution has characteristic function of the form given in Theorem 6.2.5.

The family of distributions with characteristic functions given by (6.52) includes, but is not limited to, the family of bivariate Gaussian distributions. In addition, a Gaussian distribution convolved with a distribution with mass concentrated on the line $c_2Y - c_1X = 0$ will have a characteristic function of the form (6.52). But there are characteristic functions of the form given in (6.52) that cannot be represented as the product of the characteristic function of a Gaussian distribution and another characteristic function. The following example is due to Gennady Feldman at the Ukrainian Academy of Sciences.

There exists a characteristic function of the form

$$\phi(t,s) = e^{-at^2-bs^2}V(t), \;\; a>0, b>0 \tag{6.69}$$

that cannot be represented as $e^{-Q(t,s)}W(t)$ where Q is a nonnegative definite quadratic form and $W(t)$ is a characteristic function.

Note that a characteristic function of the form (6.52) can be reduced to the form (6.69) by a suitable change of variables. Let $V(t) = -1/4 + 5/4\cos(t)$; $V(t)$ is not a characteristic function. This can be verified using Cramer's Criterion, ([29], p. 65) which states that a bounded continuous function $f(t)$ is a characteristic function if and only if i) $f(0) = 1$ and ii) $\psi(x,A) = \int_0^A \int_0^A f(t-u)\exp\{ix(t-u)\}dtdu$ is real and non-negative for all real x and for all $A > 0$. Setting $x = 0$ and $A = 3\pi/2$ gives $\int_0^{3\pi/2}\int_0^{3\pi/2} V(t-u)dtdu = 5/2 - 9/16\pi^2 < 0$, so that $V(t)$ cannot be a characteristic function.

For any $\sigma \geq \sigma_0 = (4\log(5))^{-1}$, $\varphi(t) = e^{-\sigma t^2}V(t)$ is a characteristic function while for $0 \leq \sigma < \sigma_0$, $\varphi(t)$ is not a characteristic function. This is true because

$$\begin{aligned} p(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt \\ &= \frac{1}{16\sqrt{\sigma\pi}}\left[5e^{x/\sigma} - 2e^{(2x+1)/(4\sigma)} + 5\right]e^{-(x+1)^2/(4\sigma)} \end{aligned} \tag{6.70}$$

is nonnegative and integrates to 1 for any $\sigma \geq \sigma_0$, while $p(x) < 0$ for some x if $\sigma < \sigma_0$. In particular $e^{-\sigma_0 t^2 - bs^2}V(t)$ is a characteristic function since it is the product of two characteristic functions.

Suppose $e^{-\sigma_0 t^2 - bs^2}V(t) = e^{-Q(t,s)}W(t)$. For $s = 0$, $e^{-\sigma_0 t^2}V(t) = e^{-kt^2}W(t)$ for some $k > 0$. But this is impossible, since if $k \geq \sigma_0$, $e^{-(k-\sigma_0)t^2}W(t) = V(t)$. The left hand side is the product of two characteristic functions, while the right hand side is not a characteristic function. If $k < \sigma_0$, $e^{-(\sigma_0 - k)t^2} V(t) = W(t)$. The left hand side cannot be a characteristic function since $0 < \sigma_0 - k < \sigma_0$.

Theorem 6.2.5 can easily be generalized to the case of location families where the dimension of the parameter is greater than 2.

Theorem 6.2.7. *Let* $\mathbf{X}_1 = (X_1^1, \ldots, X_1^k), \ldots, \mathbf{X}_n = (X_n^1, \ldots, X_n^k)$ *be a sample of size* $n \geq 3$ *from a location family* $F(\mathbf{X} - \boldsymbol{\theta})$ *for* $\boldsymbol{\theta} \in \mathbb{R}^k$. *If the Pitman estimator of* $\Delta = c_1\theta_1 + \cdots + c_k\theta_k$ *has the form* $\mathbf{C}^T\overline{\mathbf{X}}$, *where* $\mathbf{C}^T = [c_1, \ldots, c_k]$, *then* F *has characteristic function of the form*

$$\begin{aligned} &\phi(t_1, \ldots, t_k) = \\ &\qquad \exp\{Q(t_1, \ldots, t_k) + h(c_2t_1 - c_1t_2, c_3t_2 - c_2t_3, \ldots, c_kt_{k-1} - c_{k-1}t_k)\}. \end{aligned} \tag{6.71}$$

where h *is a differentiable function with* $h(\mathbf{0}) = 0$ *and* Q *is a quadratic function in* $t_1, \ldots, t_k$.

Bibliography

[1] Aczel, J. (1966). *Lectures on Functional Equations and Their Applications.* New York: AP.

[2] Altham, P.M.E. (1984). Improving the precision of estimation by fitting a model. *J. Roy. Stat. Soc. B. Met.*, **46**, 118-119.

[3] Andrews, D.F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H., and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton: Princeton University Press.

[4] Athreya, K.B, and Lahiri, S.N. (2006) *Measure Theory and Probability Theory.* New York: Springer.

[5] Bhapkar, V.P. (1972). On a measure of efficiency of an estimating equation. *Sankhyā, Ser. A*, **34**, 467-472.

[6] Bhapkar, V.P. (1989). Conditioning on ancillary statistics and loss of information in the presence of nuisance parameters. *J. Stat. Plann. Inference*, **21**, 139-160.

[7] Bickel, P., Klassen, C., Ya'acov, R., and Wellner, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* New York: Springer.

[8] Evans, L.C. (1998). *Partial Differential Equations.* Providence: AMS.

[9] Casella, G. and Lehmann, E.L. (1998). *Theory of Point Estimation, Second Edition.* New York: Springer.

[10] Chandrasekar, B. and Kale, B.K. (1983). Unbiased statistical estimation functions in presence of nuisance parameter. *J. Stat. Plann. Inference*, **9**, 45-54.

[11] Godambe, V.P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Stat.*, **63**, 277-284.

[12] Godambe, V.P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, **63**, 277-84.

[13] Godambe, V.P. (1991). *Estimating functions.* Oxford: Clarendon Press.

[14] Godambe, V.P., and Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Ann. Stat.*, **2**, 568-71.

[15] Heyde, C. C., (1997). *Quasi-Likelihood And Its Applications*. New York: Springer.

[16] Huber, P. J., (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, **1**, 221-233.

[17] Huber, P. J., (1981). *Robust Statistics*. New York: Wiley.

[18] Ibragimov, I.A. and Has'minskii, R.Z., (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer-Verlag.

[19] Jansen, P., Jureckova, J., and Veraverbeke, N. (1985). Rate of convergence of one- and two-step M-estimators with applications to maximum likelihood and Pitman estimators. *Ann. Stat.*, **13**, 1222-1229.

[20] Jennrich, R. I. and Yuan, K.H. (1998). Asymptotics of estimating equations under natural conditions. *J. Multivariate Anal.*, **65**, 245-260.

[21] Kagan, A.M. (1976). Fisher information contained in a finite-dimensional linear space and a properly formulated version of the method of moments. *Problems Inform. Transmission*, **12**, 25-42.

[22] Kagan, A.M. (1986). A simple modification of Pitman estimates for a location parameter. *Theory Probab. Appl.*, **30**, 598-603.

[23] Kagan, A.M., Linnik, Yu. V., and Rao, C.R. (1973). *Characterization Problems in Mathematical Statistics*. New York: Wiley.

[24] Kagan, A.M., Linnik, Yu. V., and Rao, C.R. (1965). On a characterization of the normal law based on a property of the sample average. *Sankhyā, Ser. A*, **27**, 37-40.

[25] Kagan, A.M., Klebanov, L.B., and Fintushal, S.M. (1978). Asymptotic behavior of polynomial Pitman estimators. *J. Math. Sci.*, **9**, 862-870.

[26] Kagan, A.M., Melamed, I.A., and Zinger, A.A. (1982). A class of estimators of a location parameter in presence of a nuisance scale parameter. *Statistics and Probability: Essays in Honor of C.R. Rao*, 359-368.

[27] Kagan, A.M., and Rao, C.R. (2003). Some properties and applications of the efficient Fisher score. *J. Stat. Plann. Inference*, **117**, 343-352.

[28] Kale, B. K. (1962). An extension of the Cramér-Rao inequality for statistical estimation functions. *Skand. Aktuar.*, **45**, 60-89.

[29] Lukacs, E. (1960). *Characteristic Functions*. London: Griffin.

[30] Mukhopadhyay, P. (2004). *An Introduction to Estimating Functions*. Harrow U.K.: Alpha Science.

[31] Port, Sidney, C., and Stone, C. J. (1974). Fisher information and the Pitman estimator of a location parameter. *Ann. Stat.*, **2**, 225-247.

[32] Rao, C.R. (1973). *Linear Statistical Inference and its Applications, Second Edition*. New York: Wiley.

[33] Rudin, W. (1964). *Principles of Mathematical Analysis, Second Edition*. New York: McGraw-Hill.

[34] Shao, J. (1999). *Mathematical Statistics*. New York: Springer-Verlag.

[35] Shiryaev, A.N. (1996). *Probability, second edition*. New York: Springer-Verlag.

[36] Small, C.G., and McLeish, D.L. (1988). Generalizations of ancillarity, completeness and sufficiency in an inference function space. *Ann. Stat.*, **16**, 534-551.

[37] Small, C.G., and McLeish, D.L. (1989). Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, **73**, 534-551.

[38] Silvey, S.D. (1959). The Lagrangian multiplier test. *Ann. Math. Stat.*, **30**, 389-407.

[39] Stein, C., (1959). The admissibility of Pitman's estimator of a single location parameter. *Ann. Math. Stat.*, **30**, 970-979.

[40] Stone, C., (1974). Asymptotic properties of estimators of a location parameter. *Ann. Stat.*, **2**, 1127-1137.

[41] Styan, G.P.H., (1970). Notes on the distribution of quadratic forms in singular normal variables. *Biometrika*, **57**, 567-572.

[42] Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

[43] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.*, **54**, 426-482.

[44] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1-26.

[45] Yu, T. (2008). Estimation theory of a location parameter in small samples. Ph.D. Thesis, University of Maryland, College Park, Maryland.