Semiparametric Statistics

Lecture Notes

Klaassen, C.A.J.; van Es, A.J.

Semiparametric Statistics

Lecture Notes

Chris A.J. Klaassen
Bert van Es
Korteweg-de Vries Institute for Mathematics
University of Amsterdam

September 10, 2013

Contents

Preface

Classical mathematical statistics deals with models that are parametrized by a Euclidean, i.e. finite dimensional, parameter. Quite often such models have been and still are chosen in practical situations for their mathematical simplicity and tractability. However, these models are typically inappropriate since the implied distributional assumptions cannot be supported by hard evidence. It is natural then to relax these assumptions. This leads to the class of semiparametric models.

An example is the classical linear regression model with normal error distribution. If the normality assumption is replaced by the often less questionable assumption that the errors have a density with mean zero, we have a semiparametric linear regression model with the regression parameters as Euclidean parameter and the unknown error density as so called Banach parameter.

Semiparametrics has been initiated by Stein (1956), who presented a result on Fisher information matrices and who discussed its consequences for the asymptotic theory of estimation and of testing hypotheses. Apparently, Hájek (1962) has been unaware of this paper when he presented the first semiparametrically efficient test. However, the main development in semiparametrics this far has been in estimation theory. This development has started with Van Eeden (1970), Beran (1974), and Stone (1975) and got a strong impetus by Pfanzagl and Wefelmeyer (1982), Bickel (1982), and Begun, Hall, Huang, and Wellner (1983). A comprehensive account is given in Bickel, Klaassen, Ritov, and Wellner (1993, 1998).

This course will treat some highlights from the theory of semiparametric estimation as it has developed during the last quarter of the past century. More recent results will be discussed as well. Some topics are crucial for a proper understanding of the issues of semiparametrics and they will be treated. Others are less essential. From this last group a few have been chosen according to our own, very personal biases.

Semiparametrics has been studied in a local asymptotic setting, in which the Convolution Theorem yields bounds on the performance of regular estimators. Alternatively, local asymptotics can be based on the Local Asymptotic Minimax Theorem and on the Local Asymptotic Spread Theorem, both valid for any sequence of estimators. This Local Asymptotic Spread Theorem is a straightforward consequence of a finite sample Spread Inequality, which has some intrinsic value for estimation theory in general. We will discuss both the finite sample and Local Asymptotic Spread Theorem, as well as the Convolution Theorem.

These notes will *not* constitute a self-contained text. Often reference will be made to the original literature for relevant technical details. However, the main line of the argument will be understandable without consulting the literature referred to.

Preliminary notes by Edwin van den Heuvel of this course have been transformed into an

intermediate version by Bert van Es. I would like to thank them for this great help. Furthermore, I would like to thank many people for fruitful and inspiring discussions. Most of all, my thesis advisor Willem van Zwet, who has put me on the track towards the spread inequality many years ago, and Peter Bickel, Yanki Ritov, and Jon Wellner, with whom it has been very pleasant to collaborate over the many years that have passed since the start of the writing of our book in 1983.

Chapter 1

Introduction

The random quantity X takes values in the measurable space $(\mathcal{X}, \mathcal{A})$. It has *unknown* distribution P. This distribution belongs to the *known* class $\mathcal{P}$. The map

$$\nu : \mathcal{P} \to \mathbb{R}^m$$

defines a Euclidean parameter. We will study estimation of the unknown value of $\nu(P)$. Any measurable map

$$t : \mathcal{X} \to \mathbb{R}^m$$

defines an estimator $T = t(X)$ of $\nu(P)$. We will focus on optimality and efficiency of estimators T of $\nu(P)$.

To prove that an estimator is optimal, one has to show that no estimator is better. Typically, this is done by proving the validity of a bound on the performance of estimators and by showing subsequently that the proposed estimator attains this bound. This is exactly what we will do here. We will discuss such bounds in Chapter 2 through 7, and in Chapter 8 we will study construction of estimators attaining these bounds.

We will start by a discussion of the so-called spread inequality in Chapter 2. This inequality yields a bound on the distribution of an estimator averaged over the class $\mathcal{P}$ of underlying distributions P. It is formulated in terms of quantile functions, and it is valid without any restriction on the estimator.

In most (but not all) of the models we will consider, the observations $x_1, \dots, x_n$ may be viewed as realizations of i.i.d. random variables $X_1, \dots, X_n$. In the notation used above, this means that X is an n-vector $(X_1, \dots, X_n)$. Then X has distribution P^n for some unknown distribution P from some known class $\mathcal{P}$. Note that this is an abuse of notation since P^n, and $\mathcal{P}^n$ have been denoted by P and $\mathcal{P}$ above. In fact, we will abuse notation even to a larger extent. However, it will always be clear when we are in the general situation with

$$X,\ (\mathcal{X}, \mathcal{A}),\ P \in \mathcal{P},\ \nu : \mathcal{P} \to \mathbb{R}^m,\ T = t(X), t : \mathcal{X} \to \mathbb{R}^m, \tag{1.1}$$

and when we are in the i.i.d. situation with

$$\begin{gathered} X, X_1, \dots, X_n \text{ independent and identically distributed on } (\mathcal{X}, \mathcal{A}), \\ X \sim P \in \mathcal{P},\ \nu : \mathcal{P} \to \mathbb{R}^m,\ T_n = t_n(X_1, \dots, X_n),\ t_n : \mathcal{X}^n \to \mathbb{R}^m. \end{gathered} \tag{1.2}$$

The spread inequality from Chapter 2 is a finite sample inequality in the sense that, in the i.i.d. situation, it is valid for n finite. In the remainder of the course we will focus on asymptotics with $n \to \infty$. First, we will study regular parametric models. In these models $\mathcal{P}$ is parametrized by $\theta \in \Theta$, an open subset of $\mathbb{R}^k$; so $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$. In Chapter 3 it will be shown that regular parametric models are Locally Asymptotically Normal (LAN).

Under Local Asymptotic Normality the spread inequality yields an asymptotic bound on the performance of any sequence $\{T_n\}_{n\in\mathbb{N}}$ of estimators. The resulting Local Asymptotic Spread Theorem will be discussed in Chapter 4. Some geometric interpretations of it for estimation in the presence of nuisance parameters will be presented in Chapter 5.

In a Semiparametric model the class $\mathcal{P}$ of distributions cannot be parametrized as a regular parametric model, but it contains regular parametric submodels in a natural way. The Local Asymptotic Spread Theorem yields lower bounds for the asymptotic performance of an estimator sequence within each of these regular parametric submodels. The supremum of these lower bounds thus is an obvious lower bound to the asymptotic performance of estimators in the semiparametric model. This approach is the key idea in semiparametrics, and it will be discussed in Chapter 6, together with some of its consequences. Traditionally, the theory of efficient estimation in semiparametric models is developed via the Convolution Theorem, which is studied in Chapter 7.

Construction of estimators attaining the semiparametric asymptotic bound obtained in Chapters 6 and 7, will be studied in Chapter 8. These estimators are called (asymptotically) efficient. In Chapter 9 we will extend this theory for i.i.d. models to a large class of time series models. In Chapter 10 we will discuss estimation of a Banach parameter $\nu(P)$ where ν maps $\mathcal{P}$ into a Banach space $\mathcal{B}$. Finally, in Chapter 11 we will study cross sectional sampling, a time-saving method in survival analysis.

We close this chapter by presenting three examples of semiparametric models for i.i.d. situations with $X,\ X_1, \ldots, X_n$ i.i.d. $X \sim P \in \mathcal{P}$, and one example of a non-i.i.d. model.

Example 1.1 (Symmetric Location) Let $\mathcal{G}$ be the class of distributions on $\mathbb{R}$ symmetric about 0 and let ϵ be a random variable with unknown distribution $G \in \mathcal{G}$. Consider

$$X = \nu + \epsilon \tag{1.3}$$

with ν the unknown real location parameter of interest and G the unknown nuisance parameter. Then X has distribution $P_{\nu,G}$, which is symmetric about ν, and we could parametrize $\mathcal{P}$ as $\mathcal{P} = \{P_{\nu,G} : \nu \in \mathbb{R}, G \in \mathcal{G}\}$. Stein (1956) claimed that under this semiparametric model $\mathcal{P}$ it should be possible to estimate the location parameter ν asymptotically as well at any G_0 as under the parametric model $\{P_{\nu,G_0} : \nu \in \mathbb{R}\}$. Beran (1974) and Stone (1975) were the first to construct such estimators, which were called adaptive, because they adapt to the underlying distribution G. Van Eeden (1970) constructed an adaptive estimator earlier, but under the additional assumption that G is strongly unimodal, that is, G has a log-concave density. Therefore, it is fair to say that the symmetric location model triggered the development of semiparametrics. □

Example 1.2 (Linear Regression) Here we take $\mathcal{G}$ to be the class of all distributions on $\mathbb{R}$ with mean 0. We extend (1.1) to

$$Y = \nu^T Z + \epsilon, \tag{1.4}$$

where ϵ and Z are independent, ϵ has unknown distribution $G \in \mathcal{G}$ and $Z \in \mathbb{R}^m$ has unknown distribution such that the covariance matrix of Z,

$$\Sigma_Z = \mathrm{E}\,(Z - \mathrm{E}\,Z)(Z - \mathrm{E}\,Z)^T, \tag{1.5}$$

is nonsingular. If one observes i.i.d. copies of $X = (Y, Z)$, then it is possible to estimate $\nu \in \mathbb{R}^m$ adaptively, as we will see later. □

Example 1.3 (Cox's Proportional Hazards Model) Again, one observes i.i.d. copies of $X = (Y, Z)$ with Z an m-vector of covariates. Given $Z = z$ the one-dimensional survival time Y has hazard function

$$\lambda(y \mid z, \nu) = \exp(\nu^T z)\lambda(y), \quad y > 0. \tag{1.6}$$

Here $\lambda(y) = g(y)/(1 - G(y))$ is the unknown baseline hazard function with G a distribution function with density g. We will be interested in estimation of ν in the presence of the nuisance parameter λ. □

Example 1.4 (ARMA) Consider the ARMA(p, q) process where we observe $X_t,\ t = 1, \dots, n$, generated by

$$X_t = \rho_1 X_{t-1} + \dots + \rho_p X_{t-p} + \varphi_1 \epsilon_{t-1} + \dots + \varphi_q \epsilon_{t-q} + \epsilon_t, \tag{1.7}$$

and the square integrable starting values $X_{1-p}, \dots, X_0,\ \epsilon_{1-q}, \dots, \epsilon_0$. The innovations $\epsilon_{1-q}, \dots, \epsilon_n$ are i.i.d. with unknown density g, mean 0, and finite variance. The $\mathbb{R}^{p+q}$-valued parameter $\nu = (\rho_1, \dots, \rho_p,\ \varphi_1, \dots, \varphi_q)$ is restricted in such a way that the zeroes of $1 - \rho_1 B \cdots - \rho_p B^p$ and of $1 + \varphi_1 B \cdots + \varphi_q B^q$ all lie outside the unit circle in $\mathbb{C}$. This restriction guarantees stationarity and invertibility of the process (1.5); cf. Box, Jenkins, and Reinsel (1994). □

1.1 Exercises Chapter 1

Exercise 1.1 (Identifiability) Determine the map $\nu : \mathcal{P} \to \mathbb{R}^m$ that identifies the Euclidean parameter $\nu \in \mathbb{R}^m$ in Example 1.1 ($m = 1$) and Example 1.2. When such a map ν exists, the corresponding parameter ν is called identifiable. □

Exercise 1.2 (Identifiability of Cox's Model) Show that the Cox Proportional Hazards Model from Example 1.3 is identifiable if the covariance matrix of Z is nonsingular, by noting

$$\log\left(\frac{\lambda(y \,|\, z, \nu)}{\lambda(y \,|\, EZ, \nu)}\right) = \nu^T(z - EZ). \tag{1.8}$$

□

Chapter 2

Spread Inequality

Consider a class $\mathcal{P}$ of probability measures dominated by a σ-finite measure μ on $(\mathcal{X}, \mathcal{A})$. The map $\nu : \mathcal{P} \to \mathbb{R}$ defines our one-dimensional parameter of interest. One observes a realization x of the random variable X with unknown distribution $P \in \mathcal{P}$. As in the general case of Chapter 1, $T = t(X)$ is an estimator of $\nu(P)$.

Example 2.1 (Degenerate Estimators) Fix $P_0 \in \mathcal{P}$. Note that the degenerate estimator $T = \nu(P_0)$ cannot be improved in estimating $\nu(P)$ *at* $P = P_0$, but it is extremely bad, of course, at all P with $\nu(P) \neq \nu(P_0)$. □

This little example illustrates that the performance of an estimator can be judged only by considering its behavior at several (or all) $P \in \mathcal{P}$ simultaneously. In (most versions of) the Cramér-Rao inequality this is forced by the condition of unbiasedness on the estimator. In Bayesian statistics one chooses a prior. The Hájek-Le Cam convolution theorem restricts attention to regular estimators, as we will see in Chapter 7. The so-called Local Asymptotic Minimax theorem considers suprema over subclasses of $\mathcal{P}$.

Here we will choose a weight function $\tilde{W}$ on $\mathcal{P}$, which is a probability measure that describes the relative stress we want to put on the performance of the estimator T at the respective $P \in \mathcal{P}$. Thus we will study the average distribution of $T - \nu(P)$ defined by

$$G(y) = \int_{\mathcal{P}} P(T - \nu(P) \leq y) d\tilde{W}(P), \quad y \in \mathbb{R}. \tag{2.1}$$

Roughly speaking, the spread inequality says that this distribution G cannot be concentrated arbitrarily much. In fact, it states that, whatever the estimator T, the quantiles of G are at least as far apart as the corresponding quantiles of a particular distribution function K, so

$$\left| G^{-1}(v) - G^{-1}(u) \right| \geq \left| K^{-1}(v) - K^{-1}(u) \right|, \quad u, v \in (0,1). \tag{2.2}$$

Here K is a distribution defined in terms of the model and *not* the estimator.

Define the random vector $(X, \vartheta) = (X, \nu(P))$, where P is random with distribution $\tilde{W}$. Note that our estimation problem is completely described by the joint distribution of (X, ϑ), which generates a parametric model. Furthermore, note that (2.1) may be rewritten as

$$G(y) = P(T - \vartheta \leq y), \quad y \in \mathbb{R}. \tag{2.3}$$

We will denote the conditional density of X at x given $\vartheta = \theta$ by $p(x \mid \theta)$ with respect to μ, the density of ϑ at θ by $w(\theta)$ with respect to Lebesgue measure, and the joint density of (X, ϑ) by $f(x,\theta) = p(x \mid \theta)w(\theta)$ with respect to the product measure of μ and Lebesgue measure.

Differentiating G we see that under regularity conditions, G has density

$$\begin{aligned} g(y) &= \lim_{\epsilon\to 0} \epsilon^{-1}(1 - G(y-\epsilon) - (1 - G(y))) \\ &= \lim_{\epsilon\to 0} \mathrm{E}\left(\mathbf{1}_{[T-\vartheta>y]}\epsilon^{-1}\left\{\frac{f(X,\vartheta+\epsilon)}{f(X,\vartheta)} - 1\right\}\right) \\ &= \mathrm{E}\,\mathbf{1}_{[T-\vartheta>y]}S, \quad y \in \mathbb{R}, \end{aligned} \tag{2.4}$$

where the **Score Statistic** S is defined as

$$S = \frac{\dot f}{f}(X,\vartheta) = \frac{\dot p}{p}(X \mid \vartheta) + \frac{\dot w}{w}(\vartheta) \tag{2.5}$$

with $\dot{}$ denoting differentiation with respect to θ. Now, let H be the distribution function of S,

$$H(z) = P(S \le z), \quad z \in \mathbb{R}. \tag{2.6}$$

Then (2.4) yields (see also Exercise 2.1)

$$\begin{aligned} g(G^{-1}(s)) &= \mathrm{E}\,(\mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]} \mid S)S) \\ &= \int_0^1 \mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]} \mid S = H^{-1}(t))H^{-1}(t)dt \\ &= \int_0^1 \varphi(t)H^{-1}(t)dt. \end{aligned} \tag{2.7}$$

We note

$$1 - s = \int_0^1 \mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]} \mid S = H^{-1}(t))dt = \int_0^1 \varphi(t)dt \tag{2.8}$$

and $0 \le \varphi(t) \le 1$. Maximizing the right-hand side of (2.7) over all critical functions φ satisfying (2.8) we get

$$g(G^{-1}(s)) \le \int_s^1 H^{-1}(t)dt \tag{2.9}$$

in view of the monotonicity of H^{-1}. Integration yields

$$\begin{aligned} G^{-1}(v) - G^{-1}(u) &\ge \int_u^v \frac{1}{g(G^{-1}(s))}\,ds \ge \int_u^v \frac{1}{\int_s^1 H^{-1}(t)dt}ds \\ &= K^{-1}(v) - K^{-1}(u), \quad 0 < u < v < 1, \end{aligned} \tag{2.10}$$

with

$$K^{-1}(u) = \int_{\frac{1}{2}}^u \frac{1}{\int_s^1 H^{-1}(t)dt}ds. \tag{2.11}$$

All in all, we have

Theorem 2.1 (One-dimensional Spread Inequality) *The distribution function G defined in* (2.1) *and* (2.3) *is at least as spread out as the distribution function K defined by its quantile function K^{-1} in* (2.11)*; in formula (cf.* (2.2)*)*

$$G^{-1}(v) - G^{-1}(u) \geq K^{-1}(v) - K^{-1}(u), \quad 0 < u < v < 1. \tag{2.12}$$

Regularity conditions, details of the proof and many implications may be found in Van den Heuvel and Klaassen (1997) and Klaassen (1989a,b).

An alternative way to see (2.9) is to write

$$\begin{aligned}
&\int_s^1 H^{-1}(t)dt - g(G^{-1}(s)) \\
&\quad = \int_0^1 H^{-1}(t)\{\mathbf{1}_{[t>s]} - \mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]}|S = H^{-1}(t))\}dt \\
&\quad = \int_0^1 \{H^{-1}(t) - H^{-1}(s)\}\{\mathbf{1}_{[t>s]} - \mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]}|S = H^{-1}(t))\}dt \\
&\quad = \int_0^1 \{H^{-1}(t) - H^{-1}(s)\}\{\mathbf{1}_{[H^{-1}(t)>H^{-1}(s)]} - \mathrm{E}\,(\mathbf{1}_{[T-\vartheta>G^{-1}(s)]}|S = H^{-1}(t))\}dt \\
&\quad = \mathrm{E}\,\{S - H^{-1}(s)\}\{\mathbf{1}_{[S>H^{-1}(s)]} - \mathbf{1}_{[T-\vartheta>G^{-1}(s)]}\} \\
&\quad = \mathrm{E}\,|S - H^{-1}(s)||\mathbf{1}_{[S>H^{-1}(s)]} - \mathbf{1}_{[T-\vartheta>G^{-1}(s)]}| \geq 0.
\end{aligned} \tag{2.13}$$

Assume that S has no pointmasses and note that $T - \vartheta$ has neither since it has a density g with respect to Lebesgue measure. Then, equality can hold in (2.13) only if

$$\mathrm{E}\,|\mathbf{1}_{[H(S)>s]} - \mathbf{1}_{[G(T-\vartheta)>s]}| = 0. \tag{2.14}$$

Consequently, the spread inequality from (2.12) is an equality only if

$$0 = \mathrm{E} \int_0^1 |\mathbf{1}_{[H(S)>s]} - \mathbf{1}_{[G(T-\vartheta)>s]}|ds = \mathrm{E}\,|H(S) - G(T-\vartheta)|. \tag{2.15}$$

We have proved the "only if" part of (See note 1 in Appendix B.1 for the "if" part)

Theorem 2.2 (Spread Equality) *Let regularity conditions hold such that* Theorem 2.1 *is valid. S has no pointmasses and equality holds in* (2.12) *iff*

$$T - \vartheta = G^{-1}(H(S)), \;\; a.s. \tag{2.16}$$

Often a model, i.e. the class $\mathcal{P}$, contains natural parametric submodels that are of interest, or it is parametric itself. To study this situation we consider the parametric model

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}, \quad \Theta \subset \mathbb{R}^k, \;\Theta \text{ open}, \tag{2.17}$$

and the map $\nu : \mathcal{P} \to \mathbb{R}^m$ that defines our m-dimensional parameter of interest. Together with the parametrization $\theta \mapsto P_\theta$ this map ν defines the map $q : \Theta \to \mathbb{R}^m$ by $q(\theta) = \nu(P_\theta),\ \theta \in \Theta$. With $T = t(X)$ an estimator of $\nu(P)$ we are interested in the behavior of

$$T - q(\theta), \;\theta \in \Theta. \tag{2.18}$$

The weight function $\tilde{W}$ on $\mathcal{P}$ can be represented now by a density w with respect to Lebesgue measure on $\mathbb{R}^k$, such that

$$\int_\Theta w(\theta)d\theta = 1. \tag{2.19}$$

Let ϑ be a random vector with density w and denote the conditional density of X given $\vartheta = \theta$ by $p(x\,|\,\theta)$ and the joint density of X and ϑ by $f(x,\theta) = p(x\,|\,\theta)w(\theta)$. We will derive a spread inequality for

$$G(y) = P(T - q(\vartheta) \le y), \quad y \in \mathbb{R}^m, \tag{2.20}$$

but we will take $m = 1$ first. The crucial differences with (2.3) and Theorem 2.1 are the dimension k of Θ that might be more than 1, and the map q that need not be the identity.

Again, the basic step is differentiation of G as in (2.4). Extend $q : \Theta \to \mathbb{R}$ to $q : \mathbb{R}^k \to \mathbb{R}$. Fix $b \in \mathbb{R}^k$ and assume that for all $\eta \in \mathbb{R}^k$ and all $\epsilon \in \mathbb{R}$ with $|\epsilon|$ sufficiently small the equation

$$q(\eta) + \epsilon = q(\eta + \delta b) \tag{2.21}$$

has a solution $\delta = \delta(\epsilon, \eta)$. In case of multiple solutions we will take the one closest to 0. Given q this restricts the choice of $b \in \mathbb{R}^k$. Then, by the substitution $\theta = \eta + \delta(\epsilon,\eta)b$ we obtain (cf. Exercise 2.9)

$$\begin{aligned} & 1 - G(y-\epsilon) \\ &= P(T - q(\vartheta) + \epsilon > y) \\ &= \int_{\mathbb{R}^k}\int_{\mathcal{X}} \mathbf{1}_{[t(x)-q(\eta)>y]} f(x, \eta + \delta(\epsilon,\eta)b)\{1 + b^T\dot{\delta}(\epsilon,\eta)\}d\mu(x)d\eta, \end{aligned} \tag{2.22}$$

where $\dot{\delta}$ denotes the column k-vector of derivatives of $\delta(\epsilon,\eta)$ with respect to η. From (2.21) we get $\delta(0,\eta) = 0$ and by differentiation with respect to η, $\dot{\delta}(0,\eta) = 0$. Differentiating (2.21) with respect to ϵ at $\epsilon = 0$ we obtain

$$\frac{\partial}{\partial\epsilon}\,\delta(\epsilon,\eta)\Big|_{\epsilon=0} = \frac{1}{b^T\dot{q}(\eta)}, \tag{2.23}$$

where again $\dot{q}$ denotes the column k-vector of derivatives. Assuming that q is even twice continuously differentiable we denote the $k \times k$-matrix of second mixed partial derivatives of q by $\ddot{q}$. Differentiation of (2.21) with respect to η and ϵ at $\epsilon = 0$ yields the column k-vector

$$\frac{\partial}{\partial\epsilon}\,\dot{\delta}(\epsilon,\eta)\Big|_{\epsilon=0} = -\,\frac{\ddot{q}(\eta)b}{(b^T\dot{q}(\eta))^2}. \tag{2.24}$$

Combining (2.22) through (2.24) we get under regularity conditions that G has density

$$\begin{aligned} g(y) &= \lim_{\epsilon\to 0}\epsilon^{-1}\Big(1 - G(y-\epsilon) - (1 - G(y))\Big) \\ &= \lim_{\epsilon\to 0} \mathrm{E}\left(\mathbf{1}_{[T-q(\vartheta)>y]}\epsilon^{-1}\Big\{\frac{f(X,\vartheta+\delta(\epsilon,\vartheta)b)\{1+b^T\dot{\delta}(\epsilon,\vartheta)\}}{f(X,\vartheta)} - 1\Big\}\right) \\ &= \mathrm{E}\,\mathbf{1}_{[T-q(\vartheta)>y]}S \end{aligned} \tag{2.25}$$

with

$$
(2.26) \qquad \begin{aligned} S &= \frac{\partial}{\partial \epsilon} \log \Big(f(X, \vartheta + \delta(\epsilon, \vartheta) b)\{1 + b^T \dot{\delta}(\epsilon, \vartheta)\}\Big)\Big|_{\epsilon=0} \\ &= \frac{b^T}{b^T \dot{q}(\vartheta)} \Big\{ \frac{\dot{p}}{p}(X|\vartheta) + \frac{\dot{w}}{w}(\vartheta) - \frac{\ddot{q}(\vartheta) b}{b^T \dot{q}(\vartheta)} \Big\}. \end{aligned}
$$

Repeating the arguments (2.6) through (2.16) with S from (2.26) we arrive at the spread inequality (2.12) with G the distribution function of $T - q(\vartheta)$ with q one dimensional. Moreover, if this S has no pointmass equality holds iff

$$
(2.27) \qquad T - q(\vartheta) = G^{-1}(H(S)), \text{ a.s.}
$$

Note the dependence of S on the quite arbitrarily chosen $b \in \mathbb{R}^k$. Every appropriate b yields a lower bound K and typically, these bounds differ.

For general m-dimensional functions q we apply the preceding spread inequality to

$$
a^T (T - q(\vartheta)), \quad a \in \mathbb{R}^m,
$$

for every $a \in \mathbb{R}^m$. The resulting multidimensional spread inequality is

Theorem 2.3 (General Spread Inequality) *In the parametric model $\mathcal{P}$ from* (2.17) *with weight density w satisfying* (2.19)*, let $f(x, \theta) = p(x|\theta) w(\theta)$ be the density of (X, ϑ) with respect to $\mu \times$Lebesgue measure on $(\mathcal{X} \times \mathbb{R}^k, \mathcal{A} \times \mathcal{B}^k)$, $\mathcal{B}^k$ Borel. Let $\nu(P_\theta) = q(\theta)$ be the parameter of interest with $q : \mathbb{R}^k \to \mathbb{R}^m$ twice continuously differentiable. Fix $a \in \mathbb{R}^m$, $a \neq 0$, and let there exist $b \in \mathbb{R}^k$ such that*

$$
(2.28) \qquad a^T q(\eta + \delta b) - \epsilon = a^T q(\eta)
$$

(cf. (2.21)*) has a solution $\delta = \delta(\epsilon, \eta)$ for all $\eta \in \mathbb{R}^k$ and ϵ sufficiently small with $\delta(\epsilon, \eta)$ as close to 0 as possible, and such that $\epsilon \mapsto f(x, \eta + \delta(\epsilon, \eta) b)\{1 + b^T \dot{\delta}(\epsilon, \eta)\}$ is absolutely continuous in a neighborhood of 0 for almost all x and η. Define*

$$
(2.29) \qquad S_{a,b} = \frac{b^T}{b^T \dot{q}^T(\vartheta) a} \Big\{ \frac{\dot{p}}{p}(X|\vartheta) + \frac{\dot{w}}{w}(\vartheta) - \frac{\sum_{h=1}^m a_h \ddot{q}_h(\vartheta) b}{b^T \dot{q}^T(\vartheta) a} \Big\}
$$

with $\dot{q}(\theta)$ the $m \times k$-matrix with $\frac{\partial}{\partial \theta_j} q_i(\theta)$ in the i-th row and j-th column and with $\ddot{q}_h(\theta)$ the $k \times k$-matrix with $\frac{\partial^2}{\partial \theta_i \partial \theta_j} q_h(\theta)$ in the i-th row and j-th column. Assume that $S_{a,b}$ is well-defined with

$$
(2.30) \qquad \mathrm{E}\, |S_{a,b}| < \infty.
$$

Let $T = t(X)$, $t : \mathcal{X} \to \mathbb{R}^m$ measurable, be any estimator of $q(\theta)$. The distribution G_a of $a^T(T - q(\vartheta))$ is at least as spread out as $K_{a,b}$ in the sense of (2.12) *with $K_{a,b}$ defined as in* (2.11) *with H replaced by the distribution $H_{a,b}$ of $S_{a,b}$. Furthermore,*

$$
(2.31) \qquad G_a = K_{a,b}
$$

holds and $S_{a,b}$ has no pointmass iff

$$
(2.32) \qquad a^T (T - q(\vartheta)) = G_a^{-1}(H_{a,b}(S_{a,b})) \text{ a.s.}
$$

Note that the distribution of $T - q(\vartheta)$ is determined, once for each $a \in \mathbb{R}^m$ the distribution of $a^T(T - q(\vartheta))$ is given.

Remark 2.1 In the proof of Theorem 2.3, the absolute continuity and the related existence of $\dot{p}/p$ and $\dot{w}/w$ are used in checking the validity of (2.25) in a way similar to the proof of Theorem 2.1. In fact, one needs

$$\begin{aligned}\int_{\mathcal{X}} \int_{\mathbb{R}^k} \Big| f(x, \theta + \delta(\epsilon, \theta)b)\{1 + b^T \dot{\delta}(\epsilon, \theta)\} - f(x, \theta) \\ - b^T\{\dot{p}(x \mid \theta)w(\theta) + p(x \mid \theta)\dot{w}(\theta)\} \frac{\partial}{\partial \epsilon} \delta(\epsilon, \theta)\Big|_{\epsilon=0} \Big| d\mu(x) d\theta = o(\epsilon).\end{aligned} \tag{2.33}$$

Consequently, it also suffices to assume this $\mathcal{L}_1$-differentiability itself.

2.1 Exercises Chapter 2

Exercise 2.1 (Uniformity Generates It All) Let U have a uniform distribution on the unit interval and let F be an arbitrary distribution function. Define the inverse distribution function by

$$F^{-1}(u) = \inf\{x : F(x) \geq u\} \tag{2.34}$$

and prove that $F^{-1}(U)$ has distribution function F.

□

Exercise 2.2 (Implications Spread Ordering) The class of distribution functions on the real line may be ordered partially by the following spread order. The distribution function G is at least as spread out as F if all quantiles of G are at least as far apart as the corresponding quantiles of F, more precisely if

$$G^{-1}(v) - G^{-1}(u) \geq F^{-1}(v) - F^{-1}(u), \quad 0 < u < v < 1, \tag{2.35}$$

which we will denote by

$$G \geq_s F. \tag{2.36}$$

Prove that the variance of G equals at least the variance of F for all F and G satisfying (2.36).

□

Exercise 2.3 (Estimator Attaining the Normal Spread Lower Bound) Let ϑ have a normal distribution on $\mathbb{R}$ with mean μ and variance τ^2. Given $\vartheta = \theta$, let $X_1, \ldots, X_n$ be i.i.d. with a one-dimensional normal distribution with mean θ and variance σ^2. Compute the spread lower bound for estimators $T = t(X_1, \ldots, X_n)$ of θ, and determine the estimator that attains this bound.

□

Exercise 2.4 (Location Model) Let $X_1, \dots, X_n$ be i.i.d. with density $f(\cdot - \theta)$. Let $T_n = t_n(X_1, \dots, X_n)$ be a translation equivariant estimator of the location parameter θ, i.e.

$$t_n(x_1 + a, \cdots, x_n + a) = t_n(x_1, \dots, x_n) + a. \tag{2.37}$$

With $G(\cdot)$ the distribution function from (2.3) we define $G_n(y) = G(y/\sqrt{n}),\ y \in \mathbb{R}$. Note

$$G_n(y) = P_{f(\cdot - \theta)}(\sqrt{n}(T_n - \theta) \le y), \quad \text{all } \theta \in \mathbb{R}. \tag{2.38}$$

Choose ϑ normal with mean 0 and variance σ^2, and let f be the standard normal density φ. Compute the lower bound K from (2.11). This lower bound depends on σ, but is valid for all $\sigma > 0$. Take the limit for $\sigma \to \infty$. Which estimator attains the resulting bound?

□

Exercise 2.5 (More Spread Out Than Uniform) Let $\mathrm{E}\,|S| < \infty$ and note that (2.4) implies

$$g(G^{-1}(s)) = \mathrm{E}\,\mathbf{1}_{[T - \vartheta > G^{-1}(s)]} S \le \mathrm{E}\,|S|. \tag{2.39}$$

Prove that G is at least as spread out as the uniform distribution on $[0,\ 1/\mathrm{E}\,|S|]$.

□

Exercise 2.6 (Van Zwet Inequality) With $\mathrm{E}\,S^2 < \infty$, show that (2.4) implies

$$g(G^{-1}(s)) \le (\mathrm{E}\,S^2 \{s \wedge (1 - s)\})^{1/2}. \tag{2.40}$$

Prove that G is at least as spread out as the symmetric triangular distribution with support $[-\sqrt{2/\mathrm{E}\,S^2},\ \sqrt{2/\mathrm{E}\,S^2}]$.

□

Exercise 2.7 (Trigonometric Spread Inequality) With $\mathrm{E}\,S^2 < \infty$, show that (2.4) implies

$$g(G^{-1}(s)) \le (\mathrm{E}\,S^2 \{s(1 - s)\})^{1/2}, \tag{2.41}$$

which sharpens the Van Zwet inequality. Prove that G is at least as spread out as the trigonometric distribution with distribution function $(1 + \sin(\sqrt{\mathrm{E}\,S^2}\,x))/2$, $|x| \le \pi/(2\sqrt{\mathrm{E}\,S^2})$.

□

Exercise 2.8 (Strong Unimodality) A density f with respect to Lebesgue measure on the real line is called strongly unimodal if the convolution of any unimodal density with f yields a unimodal density. Ibragimov (1956) has shown that f is strongly unimodal iff f has a log-concave version, i.e. iff the logarithm of an appropriate version of f is concave. The distribution function of a strongly unimodal density is called strongly unimodal as well. Prove that the lower bound K of the spread inequality (2.12) is strongly unimodal.

□

Exercise 2.9 (Jacobian) In (2.22) the transformation $\eta \mapsto \eta + \delta(\epsilon/c, \eta)b$ has been applied. Prove that the corresponding Jacobian $|J_k + b\dot{\delta}^T(\epsilon/c, \eta)|$ equals $1 + b^T\dot{\delta}(\epsilon/c, \eta)$, where J_k denotes the kxk identity matrix and the $\dot{}$ denotes differentiation with respect to η.

□

Exercise 2.10 (Multivariate Normal Spread Lower Bound) Let ϑ have a multivariate normal distribution on $\mathbb{R}^k$ with mean vector μ and as covariance matrix Π. Let the conditional distribution of X given $\vartheta = \theta$ be multivariate normal with mean vector θ and covariance matrix Σ. Take $q(\theta) = \theta$ and compute the score statistic $S_{a,b}$ from Theorem 2.3 for arbitrary a and b in $\mathbb{R}^k$. Subsequently, given a determine b such that the variance of the lower bound $K_{a,b}$ in the General Spread Inequality is maximal. Call the corresponding lower bound K_a. Does there exist a random vector Z such that for all $a \in \mathbb{R}^k$ the random variable $a^T Z$ has distribution function K_a?

$\square$

Exercise 2.11 (Implication Spread Equality) If there exists an estimator that attains equality in the spread inequality, then the score statistic S from (2.5) cannot have atoms.

$\square$

Chapter 3

Regular Parametric Models

In this chapter we will consider the situation with $X,\ X_1,\ldots,X_n$ i.i.d. with unknown distribution from the class $\mathcal{P}=\{P_\theta : \theta\in\Theta\},\ \Theta\subset\mathbb{R}^k$. This is called a parametric model, because Θ is finite dimensional. We will assume the existence of a σ-finite measure μ dominating $\mathcal{P}$, and we will represent the elements of $\mathcal{P}$ in $\mathcal{L}_2(\mu)$ by the square roots $s(\theta)=p^{1/2}(\theta)$ of their densities $p(\theta)=dP_\theta/d\mu$ with respect to μ.

Definition 3.1 (Regularity) The parametrization $\theta\mapsto P_\theta$ is regular if

(i) Θ is an open subset of $\mathbb{R}^k$

(ii) for every $\theta_0\in\Theta$ there exists a k-vector $\dot\ell(\theta_0)$ of score functions in $\mathcal{L}_2(P_{\theta_0})$ such that

$$s(\theta)=s(\theta_0)+\frac{1}{2}(\theta-\theta_0)^T\dot\ell(\theta_0)s(\theta_0)+o(|\theta-\theta_0|) \tag{3.1}$$

in $\mathcal{L}_2(\mu)$ as $|\theta-\theta_0|\to 0$,

(iii) for every $\theta_0\in\Theta$ the $k\times k$ Fisher information matrix

$$I(\theta_0)=\int \dot\ell(\theta_0)\dot\ell^T(\theta_0)dP_{\theta_0} \tag{3.2}$$

is nonsingular,

(iv) the map $\theta\mapsto\dot\ell(\theta)s(\theta)$ is continuous from Θ to $\mathcal{L}_2^k(\mu)$.

Typically, the Fréchet-differentiability property of (3.1) is verified via pointwise differentiability. A result to this extent is the following

Proposition 3.1 *Let Θ be open and let for all $\theta\in\Theta$*

(i) $p(x;\theta)=p(\theta)(x)$ be continuously differentiable in θ for μ-almost all x with gradient $\dot p(x;\theta)$,

(ii) $|\dot\ell(\theta)|\in\mathcal{L}_2(P_\theta)$ with

$$\dot\ell(\theta)=\dot p(\theta)/p(\theta)\mathbf{1}_{[p(\theta)>0]}, \tag{3.3}$$

(iii) $I(\theta)$ defined by (3.2) *with $\dot\ell(\theta)$ as in* (3.3)*, is nonsingular and continuous in θ.*

Then, with $\dot\ell(\theta)$ as in (3.3)*, the parametrization*

$$\theta \to P_\theta$$

is regular.

Proof of Proposition 3.1.

Fix $\theta_0 \in \Theta$. In view of (i) we have for μ-almost all x

$$\begin{aligned} & s(x;\theta) - s(x;\theta_0) \\ & \quad = \int_0^1 \frac{1}{2}(\theta-\theta_0)^T \dot\ell(x;\theta_0+\lambda(\theta-\theta_0))p^{1/2}(x;\theta_0+\lambda(\theta-\theta_0))d\lambda \end{aligned} \tag{3.4}$$

provided $|\theta-\theta_0|$ is sufficiently small. The continuity of $\dot p(\theta)$ and (3.4) imply

$$(s(x;\theta)-s(x;\theta_0))\mathbf{1}_{[p(x;\theta_0)>0]} = \frac{1}{2}(\theta-\theta_0)^T\dot\ell(x;\theta_0)s(x;\theta_0) + o(|\theta-\theta_0|). \tag{3.5}$$

By (3.4), (ii) and (iii) it follows that (see notes 2, 3 and 4 in the appendix)

$$\begin{aligned} & \int |s(x;\theta)-s(x;\theta_0)|^2 d\mu(x) \\ \le\ & \frac{1}{4}(\theta-\theta_0)^T \int_0^1 \int \dot\ell\dot\ell^T p(x;\theta_0+\lambda(\theta-\theta_0))d\mu(x)d\lambda(\theta-\theta_0) \\ =\ & \frac{1}{4}(\theta-\theta_0)^T \int_0^1 I(\theta_0+\lambda(\theta-\theta_0))d\lambda(\theta-\theta_0) \\ =\ & \frac{1}{4}(\theta-\theta_0)^T I(\theta_0)(\theta-\theta_0) + o(|\theta-\theta_0|^2) \\ =\ & \frac{1}{4}(\theta-\theta_0)^T \int_{p(\theta_0)>0} \dot\ell\dot\ell^T p(\theta_0)d\mu(\theta-\theta_0) + o(|\theta-\theta_0|^2). \end{aligned} \tag{3.6}$$

Without loss of generality let $(\theta-\theta_0)/|\theta-\theta_0|$ converge (see note 5). Applying Vitali's theorem to (3.5) and (3.6) we obtain

$$\int_{p(x;\theta_0)>0} |s(x;\theta)-s(x;\theta_0)-\frac{1}{2}(\theta-\theta_0)^T\dot\ell s(x;\theta_0)|^2 d\mu(x) = o(|\theta-\theta_0|^2), \tag{3.7}$$

which combined with (3.6) implies

$$\int_{p(x;\theta_0)=0} |s(x;\theta)-s(x;\theta_0)|^2 d\mu(x) = o(|\theta-\theta_0|^2). \tag{3.8}$$

Together, (3.7) and (3.8) imply the Fréchet differentiability (3.1).

From (i) and (iii) we obtain

$$\lim_{\theta\to\theta_0} \dot\ell_i(x;\theta)s(x;\theta) = \dot\ell_i(x;\theta_0)s(x;\theta_0) \tag{3.9}$$

for μ-almost every x with $p(x;\theta_0) > 0$, and

$$\begin{aligned}
&\limsup_{\theta\to\theta_0} \int_{p(x;\theta_0)>0} \dot\ell_i^2(x;\theta)p(x;\theta)d\mu(x) \\
\leq\ & \limsup_{\theta\to\theta_0} \int \dot\ell_i^2(x;\theta)p(x;\theta)d\mu(x) \\
=\ & \int \dot\ell_i^2 p(x;\theta_0)d\mu(x) = \int_{p(x;\theta_0)>0} \dot\ell_i^2 p(x;\theta_0)d\mu(x).
\end{aligned} \tag{3.10}$$

The continuity from (iv) of Definition 3.1 follows from (3.9) and (3.10) by another application of Vitali's theorem. □

Example 3.1 (Normal distribution) Let us show that the family of normal distributions is a regular parametric family. Here the densities are given by

$$p_{(\mu,\sigma)}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad -\infty < x < \infty. \tag{3.11}$$

Note that the parameter θ is equal to the vector $(\theta_1, \theta_2) = (\mu, \sigma)$ and that the parameter space is given by $\Theta = \mathbb{R} \times (0, \infty)$. We will check the conditions of Proposition 3.1.

First we see that the density is continuously differentiable in θ for all x. The gradient is equal to

$$\dot p(x;(\mu,\sigma)) = \begin{pmatrix} \frac{1}{\sqrt{2\pi\sigma^3}}(x-\mu)e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \\ \frac{1}{\sqrt{2\pi\sigma^2}}\left(\frac{(x-\mu)^2}{\sigma^2} - 1\right)e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}} \end{pmatrix}. \tag{3.12}$$

As score function we get

$$\dot\ell(x;(\mu,\sigma)) = \frac{\dot p(x;(\mu,\sigma))}{p(x;(\mu,\sigma))} = \begin{pmatrix} \frac{(x-\mu)}{\sigma^2} \\ \frac{1}{\sigma}\left(\frac{(x-\mu)^2}{\sigma^2} - 1\right) \end{pmatrix}. \tag{3.13}$$

Computing the Fisher information (3.2) we get

$$I((\mu,\sigma)) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{pmatrix}. \tag{3.14}$$

Since all the conditions of Proposition 3.1 hold this family is a regular parametric model.

The proposition cannot be applied to the family of uniform distributions on $(0, \theta)$ and to the Laplace(μ) distributions. In the first case this is essential. A Local Asymptotic Spread Theorem as we will prove it, is valid for regular parametric models, but it does not hold for uniform distributions. For the Laplace distributions regularity can be shown by adapting the proof of the proposition.

The notation

$$\ell(\theta) = \log p(\theta) = 2 \log s(\theta) \tag{3.15}$$

explains the notation $\dot{\ell}(\theta)$ in (3.1). We define the log-likelihood and the score function of $(X_1, \dots, X_n)$ by

$$L_n(\theta) = \sum_{i=1}^{n} \ell(X_i; \theta) \quad \text{and} \quad S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}(X_i; \theta) \tag{3.16}$$

respectively. A fundamental consequence of regularity is *Local Asymptotic Normality*, formulated in

Theorem 3.1 (LAN) *Write*

$$L_n(\theta + \frac{t}{\sqrt{n}}) - L_n(\theta) = t^T S_n(\theta) - \frac{1}{2} t^T I(\theta) t + R_n(\theta, t). \tag{3.17}$$

If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model, then for any compact $K \subset \Theta$, any $M \in (0, \infty)$, and any $\epsilon > 0$

$$\lim_{n \to \infty} \sup_{|t| \leq M} \sup_{\theta \in K} P_\theta(|R_n(\theta, t)| > \epsilon) = 0 \tag{3.18}$$

holds. Moreover

$$S_n(\theta) \xrightarrow{\mathcal{D}}_{\theta + \frac{t}{\sqrt{n}}} \mathcal{N}(I(\theta)t,\, I(\theta)) \tag{3.19}$$

uniformly in $\theta \in K$ and in $|t| \leq M$, for every compact $K \subset \Theta$ and for every finite $M \geq 0$. Finally,

$$\lim_{n \to \infty} \sup_{|t| \leq M} \sup_{\theta \in K} P_\theta(|S_n(\theta + \frac{t}{\sqrt{n}}) - S_n(\theta) + I(\theta)t| > \epsilon) = 0, \tag{3.20}$$

for all K, M, and ϵ as above.

By the uniform convergence of (3.19) we mean that for any compact $K \subset \Theta$, any $M \in (0, \infty)$ and for every bounded, continuous function g on $\mathbb{R}$ we have

$$\lim_{n \to \infty} \sup_{|t| \leq M} \sup_{\theta \in K} |\mathrm{E}_{\theta + \frac{t}{\sqrt{n}}}\, g(S_n(\theta)) - \mathrm{E}_{\mathcal{N}(I(\theta)t, I(\theta))}\, g(S)| = 0. \tag{3.21}$$

For θ and t fixed (3.21) is equivalent to (3.19) itself, see for instance Grimmett and Stirzaker (1992) Theorem 7.2.19.

Proof of Theorem 3.1

We will only sketch the main ideas. A precise proof may be found on pages 509–513 of Appendix A.9 of Bickel, Klaassen, Ritov and Wellner (1993) (henceforth BKRW). Define

$$T_{ni} = 2\{s(X_i;\, \theta + \frac{t}{\sqrt{n}})/s(X_i; \theta) - 1\}, \quad i = 1, \dots, n. \tag{3.22}$$

First, one shows that $P_\theta(\max_{1\le i\le n}|T_{ni}|\ge\epsilon)\to 0$ as $n\to\infty$, uniformly in $\theta\in K$ and $|t|\le M$ (see note 7). Then, on the event $\{\max_{1\le i\le n}|T_{ni}|<\epsilon\}$, one uses

$$\log(1+x)=\int_0^x(1-y+\frac{y^2}{1+y})dy=x-\frac{1}{2}x^2+\frac{2}{3}\alpha|x|^3 \tag{3.23}$$

with $|\alpha|\le 1$ for $|x|\le\frac{1}{2}$, to write

$$\begin{aligned}
L_n(\theta+\frac{t}{\sqrt{n}})-L_n(\theta) &= 2\sum_{i=1}^n\log(1+\frac{1}{2}T_{ni})\\
&= 2\sum_{i=1}^n\Big\{\frac{1}{2}T_{ni}-\frac{1}{8}T_{ni}^2+\frac{1}{12}\alpha_{ni}\Big|T_{ni}\Big|^3\Big\}\\
&= \sum_{i=1}^n T_{ni}-\frac{1}{4}\sum_{i=1}^n T_{ni}^2+\frac{1}{6}\sum_{i=1}^n\alpha_{ni}\Big|T_{ni}\Big|^3\\
&= t^TS_n(\theta)-\frac{1}{2}t^TI(\theta)t \qquad (3.24)\\
&\quad+\Big\{\sum_{i=1}^n T_{ni}-\Big(t^TS_n(\theta)-\frac{1}{4}t^TI(\theta)t\Big)\Big\}\\
&\quad-\frac{1}{4}\Big\{\sum_{i=1}^n T_{ni}^2-t^TI(\theta)t\Big\}+\frac{1}{6}\sum_{i=1}^n\alpha_{ni}\Big|T_{ni}\Big|^3
\end{aligned}$$

with $|\alpha_{ni}|\le 1$ a.s. Subsequently, one proves that the last three terms in (3.24) all converge to 0 in probability under θ uniformly in $\theta\in K$, $|t|\le M$(on the event $\{\max_{1\le i\le n}|T_{ni}|<\epsilon\}$) (see note 7).

One needs a uniform central limit theorem to prove (3.19) for $M=0$ (see BKRW, p.513). To prove (3.19) in its full generality we need (3.20). To prove (3.20) we first have to study contiguity. □

Definition 3.2 (Contiguity) The sequence of probability measures $\{Q_n\}$ is contiguous with respect to $\{P_n\}$ (both P_n and Q_n on $(\mathcal{X}_n,\mathcal{A}_n)$), if for all $\{A_n\}$ with $P_n(A_n)\to 0$ also $Q_n(A_n)\to 0$ holds.

Corollary 3.1 *The Local Asymptotic Normality as formulated in* (3.18) *and* (3.19) *with* $M=0$ *and hence* $t=0$*, implies that* $\{P_{\theta_n}^n\}$ *and* $\{P_{\theta_n+t_n/\sqrt{n}}^n\}$ *are mutually contiguous for all convergent sequences* $\{\theta_n\}$ *and* $\{t_n\}$*.*

Proof

This follows from Le Cam's first and third lemma; see p.17 of BKRW. □

We will use Corollary 3.1 to complete the proof of Theorem 3.1.

Proof of Theorem 3.1. Continued

Contiguity as in Corollary 3.1, continuity and symmetry of $I(\theta)$ and (3.18) show

$$\begin{aligned} 0 &= L_n(\theta + \frac{t}{\sqrt{n}}) - L_n(\theta) + L_n\Big(\theta + \frac{t+h}{\sqrt{n}}\Big) - L_n\Big(\theta + \frac{t}{\sqrt{n}}\Big) \\ &\quad -\Big\{L_n\Big(\theta + \frac{t+h}{\sqrt{n}}\Big) - L_n(\theta)\Big\} \\ &= t^T S_n(\theta) - \frac{1}{2}t^T I(\theta)t + h^T S_n\Big(\theta + \frac{t}{\sqrt{n}}\Big) - \frac{1}{2}h^T I(\theta + \frac{t}{\sqrt{n}}\Big)h \\ &\quad -(t+h)^T S_n(\theta) + \frac{1}{2}(t+h)^T I(\theta)(t+h) + o_P(1) \\ &= h^T\Big\{S_n(\theta + \frac{t}{\sqrt{n}}\Big) - S_n(\theta) + I(\theta)t\Big\} + o_P(1). \end{aligned} \tag{3.25}$$

Since (3.25) holds for any $h \in \mathbb{R}^k$, this yields (3.20). This enables us to complete the proof of (3.19) to the case $M > 0$. Consider $\theta_n \to \theta,\ t_n \to t$. It suffices to prove

$$S_n(\theta_n) \xrightarrow{\mathcal{D}}_{\theta_n + \frac{t_n}{\sqrt{n}}} \mathcal{N}(I(\theta)t, I(\theta)) \tag{3.26}$$

in view of the continuity of $\theta \to I(\theta)$.

Now, (3.19) with $M = 0$ implies

$$S_n(\theta_n + \frac{t_n}{\sqrt{n}}) \xrightarrow{\mathcal{D}}_{\theta_n + \frac{t_n}{\sqrt{n}}} \mathcal{N}(0, I(\theta)), \tag{3.27}$$

whereas (3.20) yields

$$S_n\Big(\theta_n + \frac{t_n}{\sqrt{n}}\Big) = S_n(\theta_n) - I(\theta_n)t_n + o_P(1) \tag{3.28}$$

both under θ_n and, by contiguity, under $\theta_n + t_n/\sqrt{n}$. Combining (3.27) and (3.28) we obtain (3.26) and complete the proof of the LAN theorem. □

Example 3.2 (Normal Shift) Let I be nonsingular and let X be $\mathcal{N}(It, I)$, $t \in \mathbb{R}^k$. Note that in view of (3.19) the log-likelihood ratio

$$\begin{aligned} &\ell(X;t) - \ell(X;0) \\ &= -\frac{1}{2}(X - It)^T I^{-1}(X - It) + \frac{1}{2}X^T I^{-1}X \\ &= t^T X - \frac{1}{2}t^T It \end{aligned} \tag{3.29}$$

has the same structure as the right-hand side of (3.17). This explains the name LAN.

Clearly, the best (shift equivariant) estimator of t based on X is $I^{-1}X$. Comparing (3.17) and (3.29) we see that this indicates that good estimators of t should be equal to $I^{-1}(\theta)S_n(\theta)$ approximately and hence good estimators of θ to

$$\theta + \frac{1}{\sqrt{n}}I^{-1}(\theta)S_n(\theta) = \theta + \frac{1}{n}\sum_{i=1}^{n} I^{-1}(\theta)\dot{\ell}(X_i;\theta). \tag{3.30}$$

This is exactly what we will prove in Chapter 6. □

3.1 Exercises Chapter 3

Exercise 3.1 (Location Model) Let g be a density on $\mathbb{R}$, which is absolutely continuous with respect to Lebesgue measure with Radon-Nikodym derivative g', such that the Fisher information for location is finite, that is

$$I(g) = \int (g'/g)^2 g < \infty. \tag{3.31}$$

An example of such a density is the Laplace or double exponential density $g(x) = \frac{1}{2}\exp(-|x|)$. Note that this density does *not* yield a location model satisfying the conditions of Proposition 3.1. Nevertheless, it can be shown along the lines of the proof of Proposition 3.1, that the location model is regular for every g satisfying (3.31). See Example 2.1.2, p.15 of BKRW. □

Exercise 3.2 (Linear Regression) Compute score function $\dot{\ell}$ and Fisher information matrix for the linear regression model of Example 1.2 with G fixed, and formulate conditions needed for regularity.

□

Exercise 3.3 (Cox's Proportional Hazards Model) Do the same as in Exercise 3.2, for the Cox model of Example 1.3 with λ fixed.

Hint Introduce the cumulative hazard function Λ, $\Lambda(y) = \int_0^y \lambda$, to simplify computations and to make them more transparent. Verify that the conditional distribution of $\exp(\nu^T Z)\Lambda(Y)$ given $Z = z$ is exponential with shape parameter 1. □

Exercise 3.4 (Exponential Shift Model) Let $X_1, X_2, \ldots, X_n$ be i.i.d. exponential random variables with scale parameter 1 and location parameter θ, i.e. with density $p(\theta)$ with respect to Lebesgue measure on $\mathbb{R}$ given by

$$p(x;\theta) = e^{-(x-\theta)}\mathbf{1}_{[x>\theta]}.$$

Study the asymptotic behavior of the log-likelihood ratio

$$L_n(\theta + \frac{t}{n}) - L_n(\theta)$$

under θ for $t \in \mathbb{R}$ with $L_n(\theta)$ defined as in (3.16) and note that the limit experiment may be described via one exponential random variable with scale parameter 1 and location parameter t. □

Chapter 4

Local Asymptotic Spread Theorem

The Local Asymptotic Normality derived in the previous chapter for regular parametric models suggests a local asymptotic approach as follows. Fix $\theta_0 \in \Theta$ and consider parameter values at distance $O(1/\sqrt{n})$ from θ_0. In the framework of the spread inequality this can be implemented by choosing

$$\vartheta = \theta_0 + \frac{\sigma}{\sqrt{n}}\,\varsigma, \tag{4.1}$$

where the random vector ς has density w_0 on $(\mathbb{R}^k, \mathcal{B}^k)$ and $\sigma > 0$ will be taken to tend to infinity after the limit for $n \to \infty$ has been taken. As we will prove, the General Spread Inequality from Theorem 2.3 implies

Theorem 4.1 (Local Asymptotic Spread Theorem) *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a regular parametric model and let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of estimators of $\nu(P_\theta) = q(\theta)$ with $q : \mathbb{R}^k \to \mathbb{R}^m$ twice continuously differentiable. Fix $\theta_0 \in \Theta$ and let ϑ be defined as in (4.1), where w_0 is a differentiable density with bounded support and with derivative vector $\dot{w}_0$ satisfying*

$$\int |\dot{w}_0|^2/w_0 < \infty. \tag{4.2}$$

Assume that $\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)$ is nonsingular. For $a \in \mathbb{R}^m$, denote the distribution function of $\sqrt{n}\, a^T(T_n - q(\vartheta))$ by $G_{n,\sigma,a}$. Asymptotically, as $n \to \infty$ and subsequently $\sigma \to \infty$, $G_{n,\sigma,a}$ is at least as spread out as $\mathcal{N}(0, a^T\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)a)$, more precisely

$$\begin{aligned} &\limsup_{\sigma\to\infty}\limsup_{n\to\infty}\; G^{-1}_{n,\sigma,a}(v) - G^{-1}_{n,\sigma,a}(u) \\ \geq\; &\liminf_{\sigma\to\infty}\liminf_{n\to\infty}\; G^{-1}_{n,\sigma,a}(v) - G^{-1}_{n,\sigma,a}(u) \\ \geq\; &(a^T\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)a)^{1/2}(\Phi^{-1}(v) - \Phi^{-1}(u)), \quad 0 < u < v < 1. \end{aligned} \tag{4.3}$$

Moreover, equalities hold in (4.3) for all $a \in \mathbb{R}^m$ iff under θ_0

$$\sqrt{n}\,(T_n - q(\theta_0) - \frac{1}{n}\sum_{i=1}^{n}\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{\ell}(X_i;\theta_0)) \xrightarrow{P} 0. \tag{4.4}$$

Proof. As shown in note 8 in Appendix B.3 it follows from the regularity of $\mathcal{P}$ and the properties of w_0 and q that the General Spread Inequality, Theorem 2.3, may be applied here, if given $\sigma > 0$, the sample size n is large enough. This yields as a lower bound to $G_{n,\sigma,a}$ the distribution function $K_{n,\sigma,a}$, the score statistic of which is (take $a \neq 0$)

$$
\begin{aligned}
S_{n,\sigma,a} = {} & \frac{a^T \dot{q}(\theta_0) I^{-1}(\theta_0)}{a^T \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{q}^T(\vartheta) a} \\
\times & \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}(X_i; \vartheta) + \frac{1}{\sigma} \frac{\dot{w}_0}{w_0}(\varsigma) - \frac{1}{\sqrt{n}} \frac{\sum_{h=1}^{m} a_h \ddot{q}_h(\vartheta) I^{-1}(\theta_0) \dot{q}^T(\theta_0) a}{a^T \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{q}^T(\vartheta) a} \right\}
\end{aligned} \tag{4.5}
$$

and has distribution function $H_{n,\sigma,a}$. We have chosen here $b = I^{-1}(\theta_0)\dot{q}^T(\theta_0)a$ and we have replaced a from Theorem 2.3 by $\sqrt{n}a$. Since the support of w_0 is bounded, $\dot{q}$ and I are continuous, and $\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)$ is nonsingular, any choice of $b \neq 0$ is allowed in the sense of (2.21), provided n is large enough.

By the LAN Theorem 3.1, in particular (3.19) with $t = 0$, by the finiteness of $\mathrm{E}\,|\dot{w}_0(\varsigma)/w_0(\varsigma)|^2$, by the boundedness of the support of w_0, and by the continuity of $\dot{q}$ and $\ddot{q}$ we see that $S_{n,\sigma,a}$ converges in distribution to $\mathcal{N}(0, (a^T\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)a)^{-1})$ as $n \to \infty$ and subsequently $\sigma \to \infty$. Since (recall $a \neq 0$ and the nonsingularity of $\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)$)

$$
\lim_{\sigma\to\infty} \lim_{n\to\infty} \mathrm{E}\, S^2_{n,\sigma,a} = (a^T \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{q}^T(\theta_0) a)^{-1} < \infty \tag{4.6}
$$

holds, $|S_{n,\sigma,a}|$ is uniformly integrable and this suffices to show that $\int_s^1 H^{-1}_{n,\sigma,a}(t)dt$, $K^{-1}_{n,\sigma,a}$, and $K_{n,\sigma,a}$ converge (cf. Lemma 2.1 and its proof of Klaassen (1989a)). A straightforward computation shows that the limit of $K_{n,\sigma,a}$ is $\mathcal{N}(0, a^T\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)a)$, thus completing the proof of (4.3).

By an asymptotic version of the argument leading to the Spread Equality Theorem 2.2, we see that the equalities hold in (4.3) for all $a \in \mathbb{R}^m$ iff

$$
\sqrt{n}\, a^T(T_n - q(\vartheta)) - G^{-1}_{n,\sigma,a}(H_{n,\sigma,a}(S_{n,\sigma,a})) \xrightarrow{P} 0, \text{ as } n \to \infty,\ \sigma \to \infty. \tag{4.7}
$$

However, in view of the smoothness (3.20) and the contiguity in Corollary 3.1, both following from the LAN Theorem 3.1, and in view of the asymptotic normality of both $G_{n,\sigma,a}$ and $H_{n,\sigma,a}$ the convergence in (4.7) is equivalent to the one in (4.4). □

Remark 4.1 Let the random m-vector Z be $\mathcal{N}(0, \dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0))$ distributed. The random variable a^TZ has a normal distribution with variance $a^T\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)a$. Therefore, we might say that asymptotically as $n \to \infty$ and subsequently $\sigma \to \infty$, the random vector

$$
\sqrt{n}\,(T_n - q(\vartheta)) \tag{4.8}
$$

is at least as spread out as a normal distribution with covariance matrix $\dot{q}(\theta_0)I^{-1}(\theta_0)\dot{q}^T(\theta_0)$; see Exercise 2.10. It is an open question if in the General Spread Inequality there exist choices $b \in \mathbb{R}^k$ for each $a \in \mathbb{R}^m$ such that the bounds $K_{a,b}$ can be viewed as stemming from an m-dimensional random vector Z via the linear combinations a^TZ for all $a \in \mathbb{R}^m$.

Remark 4.2 If we would not have chosen $b = I^{-1}(\theta_0)\dot{q}^T(\theta_0)a$ in (4.5), then we would have gotten

$$\left(\frac{(b^T\dot{q}^T(\theta_0)a)^2}{b^T I(\theta_0)b}\right)^{1/2}(\Phi^{-1}(v) - \Phi^{-1}(u)) \tag{4.9}$$

as a lower bound in (4.3). If we differentiate, supervised by Lagrange, the expression $b^T I(\theta_0)b + \lambda(b^T\dot{q}^T(\theta_0)a - 1)$ with respect to b we obtain the right hand side of (4.3) as the maximum of (4.9).

The Local Asymptotic Spread Theorem renders meaning to the following definitions.

Definition 4.1 (Linearity) The sequence of estimators $\{T_n\}$ is *asymptotically linear* at P_0 in the influence function $\psi : \mathcal{X} \times \mathcal{P} \to \mathbb{R}^m$ with $\mathrm{E}_P|\psi(X,P)|^2 < \infty$, $\mathrm{E}_P\psi(X,P) = 0$, if

$$\lim_{n\to\infty} P_0\Big(\sqrt{n}\Big|T_n - \Big\{\nu(P_0) + \frac{1}{n}\sum_{i=1}^n \psi(X_i, P_0)\Big\}\Big| > \epsilon\Big) = 0 \tag{4.10}$$

for every positive ϵ.

Definition 4.2 (Efficient Influence Function, Information Bound)
The *efficient influence function* at P_θ in estimating ν within the model $\mathcal{P}$ is defined by

$$\tilde{\ell}(x; P_\theta \mid \nu, \mathcal{P}) = \dot{q}(\theta)I^{-1}(\theta)\dot{\ell}(x;\theta), \quad x \in \mathcal{X}. \tag{4.11}$$

The *information bound* at P_θ for estimating ν within the model $\mathcal{P}$ is defined as

$$I^{-1}(P_\theta \mid \nu, \mathcal{P}) = \dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta). \tag{4.12}$$

Note that

$$\mathrm{E}_\theta \tilde{\ell}\tilde{\ell}^T(X; P_\theta \mid \nu, \mathcal{P}) = I^{-1}(P_\theta \mid \nu, \mathcal{P}). \tag{4.13}$$

Definition 4.3 (Efficiency) The sequence of estimators $\{T_n\}_{n\in\mathbb{N}}$ is (locally asymptotically) *efficient* at P_0 in estimating $q(\theta)$ if it is asymptotically linear at P_0 in the efficient influence function, i.e. if it satisfies (4.4), or equivalently, if for each $\{\theta_n\}$ with $\sqrt{n}(\theta_n - \theta_0)$ bounded

$$\sqrt{n}(T_n - q(\theta_n) - \frac{1}{n}\sum_{i=1}^n \dot{q}(\theta_n)I^{-1}(\theta_n)\dot{\ell}(X_i;\theta_n)) \xrightarrow{P} 0 \tag{4.14}$$

under θ_n (cf. the last sentence of the proof of Theorem 4.1).

4.1 Exercises Chapter 4

Exercise 4.1 (Efficient Influence Functions) Compute the efficient influence functions for estimation of the Euclidean parameter ν in the models of Exercises 3.1, 3.2, and 3.3 with G fixed. □

Exercise 4.2 (Weak Convergence of Quantile Functions) Let $F_n, n = 1, 2, \ldots,$ and F be distribution functions with F_n weakly converging to F. Prove that for all u at which $F^{-1}(\cdot)$ is continuous, $F_n^{-1}(u)$ converges to $F^{-1}(u)$. We will say that F_n^{-1} converges weakly to F^{-1}. Also prove that the weak convergence of F_n^{-1} to F^{-1} implies that F_n converges weakly to F.

$\square$

Exercise 4.3 (Weak Convergence of Spread Bounds) Let $K_n^{-1}, n = 1, 2, \ldots,$ and K^{-1} be quantile functions as defined by (2.11) with H replaced by H_n for $K_n^{-1}, n = 1, 2, \ldots$. Here H and H_n are distribution functions with mean 0. If H_n converges weakly to H, then K_n converges weakly to K. Prove this.

$\square$

Chapter 5

Geometric Interpretation

The main issue in semiparametrics is estimation in the presence of infinite dimensional nuisance parameters. In the remainder of this section we will prepare for this by studying some consequences of the Local Asymptotic Spread Theorem for estimation under finite dimensional nuisance parameters.

Consider the regular parametric model $\mathcal{P}$ from Definition 3.1 with parameter $\theta \in \mathbb{R}^k$. This parameter $\theta = \binom{\nu}{\eta}$ is split up into the parameter of interest $\nu \in \mathbb{R}^m$ and the nuisance parameter $\eta \in \mathbb{R}^{k-m}$. Accordingly, we write the score function $\dot{\ell}$ as $\dot{\ell}(\theta) = \dot{\ell} = \binom{\dot{\ell}_1}{\dot{\ell}_2}$ and the Fisher information matrix as

$$I(\theta) = I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} \text{ with } I_{ij} = E_\theta \dot{\ell}_i \dot{\ell}_j (X;\theta).$$

In the situation with η known we have a submodel of $\mathcal{P}$ which is again regular parametric with score function $\dot{\ell}_1$, and Fisher information matrix I_{11}. In view of the Local Asymptotic Spread theorem an estimator sequence $\{T_n\}$ of ν is uniformly asymptotically efficient within this submodel if it is uniformly asymptotically linear in the efficient influence function $I_{11}^{-1}\dot{\ell}_1$, that is, if for all ν with $\theta = (\nu^T, \eta^T)^T \in \Theta$ and every sequence ν_n with $\nu_n \to \nu$ and $\theta_n = \left(\nu_n^T, \eta^T\right)^T \in \Theta$,

$$\sqrt{n}\Big(T_n - \nu_n - \frac{1}{n}\sum_{i=1}^n I_{11}^{-1}(\theta_n)\dot{\ell}_1(X_i;\theta_n)\Big) \xrightarrow{P}_{\theta_n} 0.$$

Furthermore, the information bound is I_{11}^{-1}.

Within the full model $\mathcal{P}$ with η unknown the Local Asymptotic Spread theorem shows that $\{T_n\}$ is uniformly asymptotically efficient in estimating ν, if it is uniformly asymptotically linear in the efficient influence function

$$\tilde{\ell}_1(\theta) = (J\ 0) I^{-1}(\theta)\dot{\ell}(\theta), \tag{5.1}$$

where J is the $m \times m$ identity matrix and 0 the $m \times (k-m)$ null matrix. By computing $I^{-1}I$ it is straightforward to check that we may write

$$I^{-1}(\theta) = I^{-1} = \begin{pmatrix} I_{11.2}^{-1} & -I_{11.2}^{-1} I_{12} I_{22}^{-1} \\ -I_{22.1}^{-1} I_{21} I_{11}^{-1} & I_{22.1}^{-1} \end{pmatrix} \tag{5.2}$$

with

(5.3) $$I_{11.2} = I_{11} - I_{12}I_{22}^{-1}I_{21},\ I_{22.1} = I_{22} - I_{21}I_{11}^{-1}I_{12}.$$

Furthermore, we will write

(5.4) $$\ell_1^\star = \dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2.$$

In this notation (5.1) becomes

$$\begin{aligned}\tilde{\ell}_1(\theta) &= \tilde{\ell}_1 = I_{11.2}^{-1}(J\,,\,-I_{12}I_{22}^{-1})\dot{\ell} = I_{11.2}^{-1}(\dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2) = I_{11.2}^{-1}\ell_1^\star \\ &= (E\ell_1^\star\ell^{\star T}_1)^{-1}\ell_1^\star,\end{aligned}$$

since $\mathrm{E}\,\ell_1^\star\ell^{\star T}_1 = \mathrm{E}\,(\dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2)(\dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2)^T = I_{11} - I_{12}I_{22}^{-1}I_{21} = I_{11.2}$ holds. Summarizing, we have: In the restricted model the efficient influence function equals

(5.5) $$(\mathrm{E}\,\dot{\ell}_1\dot{\ell}_1^T)^{-1}\dot{\ell}_1 \text{ with } (\mathrm{E}\,\dot{\ell}_1\dot{\ell}_1^T)^{-1} = I_{11}^{-1}$$

as information bound. In the full model the efficient influence function equals

(5.6) $$(\mathrm{E}\,\ell_1^\star\ell^{\star T}_1)^{-1}\ell_1^\star \text{ with } (\mathrm{E}\,\ell_1^\star\ell^{\star T}_1)^{-1} = (I^*)^{-1}$$

as information bound. In view of the similarities between (5.5) and (5.6) we will call $\ell_1^\star$ the *efficient score function* for estimation of ν within $\mathcal{P}$ and I^* the *efficient Fisher information matrix*.

In the space of mean 0 random variables of functions of X with $X \sim P_\theta$ we define an inner product by

(5.7) $$\langle f, g\rangle = \mathrm{E}_\theta f(X)g(X).$$

In the resulting Hilbert space each component of $\ell_1^\star$ may be viewed as the projection on $[\dot{\ell}_2]^\perp$ of the corresponding components of $\dot{\ell}_1$. Here, $[\dot{\ell}_2]$ is the (closed) linear span of all components of $\dot{\ell}_2$ and $[\dot{\ell}_2]^\perp$ is its orthocomplement. Indeed,

$$\langle \ell_1^\star,\, \dot{\ell}_2^T\rangle = \langle \dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2,\, \dot{\ell}_2^T\rangle = \mathrm{E}\,(\dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2)\dot{\ell}_2^T = 0$$

and hence $\ell_1^\star \in [\dot{\ell}_2]^\perp$. Furthermore,

$$\dot{\ell}_1 - \ell_1^\star = I_{12}I_{22}^{-1}\dot{\ell}_2 \perp [\dot{\ell}_2]^\perp,$$

since each component of $I_{12}I_{22}^{-1}\dot{\ell}_2$ belongs to $[\dot{\ell}_2]^{\perp\perp} = [\dot{\ell}_2]$. In formula,

(5.8) $$\ell_1^\star = \Pi\Big(\dot{\ell}_1 \mid [\dot{\ell}_2]^\perp\Big) = \dot{\ell}_1 - \Pi\Big(\dot{\ell}_1 \mid [\dot{\ell}_2]\Big).$$

Furthermore, we have

(5.9) $$I_{11}^{-1}\dot{\ell}_1 = \Pi\Big(\tilde{\ell}_1 \mid [\dot{\ell}_1]\Big),$$

as can be seen as follows

(5.10) $$\begin{aligned}&\langle \tilde{\ell}_1 - I_{11}^{-1}\dot{\ell}_1,\, \dot{\ell}_1^T\rangle \\ &= \mathrm{E}\,I_{11.2}^{-1}\Big(\dot{\ell}_1 - I_{12}I_{22}^{-1}\dot{\ell}_2\Big)\dot{\ell}_1^T - I_{11}^{-1}\mathrm{E}\,\dot{\ell}_1\dot{\ell}_1^T \\ &= I_{11.2}^{-1}I_{11.2} - I_{11}^{-1}I_{11} = 0.\end{aligned}$$

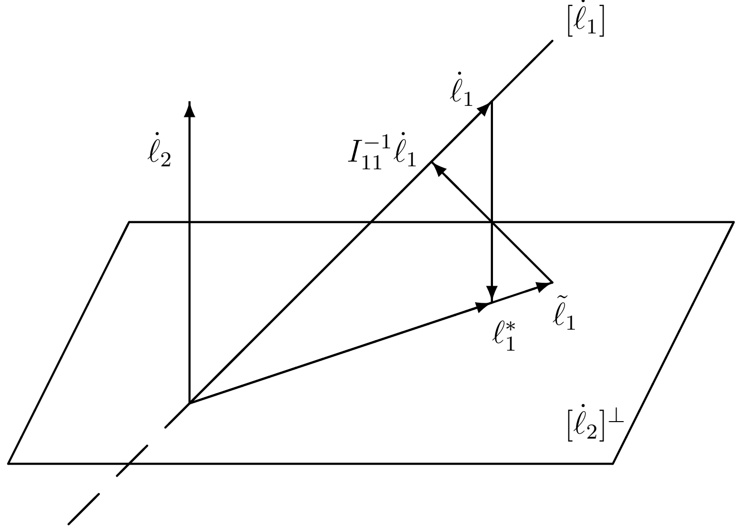

Figure 5.1: Efficient influence functions and efficient score functions

Formula (5.8) shows that, knowing the score function within the smaller model, we can obtain the score function of the same parameter within the bigger model by projection. Similarly, (5.9) shows that the efficient influence function for the smaller model may be obtained by projection of the efficient influence function of the bigger model; see Figure 5.1. Both devices are useful in semiparametrics as we will see in the next Chapter 6

Example 5.1 (Normal distribution)
Consider the normal family discussed in Example 3.1. The parameter θ equals (μ, σ). For the inverse of the Fisher information matrix for one observation we get

$$I(\theta)^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{pmatrix}. \tag{5.11}$$

The efficient influence function for estimating μ with σ known, as given by (5.5), equals (cf. (3.13))

$$\sigma^2 \frac{(x-\mu)}{\sigma^2} = x - \mu. \tag{5.12}$$

The efficient influence function for estimating μ with σ unknown (5.6) is also equal to $x - \mu$ since the off diagonal elements of the Fisher information matrix vanish. The information lower bound for estimating μ is equal to σ^2 in both cases. Here μ can be estimated just as well whether we know σ or not. The bounds are attained by the sample mean. □

Example 5.2 (Cox's Proportional Hazards Model (parametric))
The semiparametric version of the model is described in Example 1.3. One observes i.i.d. copies of $X = (Y, Z)$ with Z an m-vector of covariates. Given $Z = z$ the survival time Y has hazard function

$$\lambda(y \mid z, \nu) = \exp(z^T \nu)\lambda(y), \quad y > 0. \tag{5.13}$$

Here $\lambda(y) = g(y)/(1 - G(y))$ is the unknown baseline hazard function with G a distribution function with density g. We get a parametric version of this model by assuming that g is a density from a parametric family. Suppose for instance that g is equal to an $\mathrm{Exp}(\lambda)$ density, for some constant $\lambda > 0$. The corresponding hazard function is constant λ. We will be interested in estimation of ν in the presence of the nuisance parameter λ. Given $Z = z$ the hazard function of Y is

$$\lambda(y \mid z, \nu) = \exp(z^T\nu)\lambda, \quad y > 0. \tag{5.14}$$

Since this hazard is independent of y the conditional distribution of the survival time Y given $Z = z$ is $\mathrm{Exp}(\exp(z^T\nu)\lambda)$. Hence the unconditional density of the pair (Y, Z) is

$$p_\theta(y, z) = \exp(z^T\nu)\lambda \exp\Big(-\exp(z^T\nu)\lambda y\Big)p_Z(z), \tag{5.15}$$

where p_Z is the known density of the covariate vector Z and $\theta = (\nu^T, \lambda)^T$.

For convenience we assume that Z and ν are one dimensional. Let us consider the case of one observation first. We get

$$\ell(y, z) = z\nu + \log(\lambda) - \exp(z\nu)\lambda y + \log(p_Z(z))$$

and

$$\begin{aligned}
\dot{\ell}_1(y, z) &= \frac{\partial}{\partial \nu}\ell(y, z) = z(1 - \exp(z\nu)\lambda y), \\
\dot{\ell}_2(y, z) &= \frac{\partial}{\partial \lambda}\ell(y, z) = \frac{1}{\lambda} - \exp(z\nu)y.
\end{aligned}$$

Note that indeed the expectations of $\dot{\ell}_1(Y, Z)$ and $\dot{\ell}_2(Y, Z)$ vanish.

Computing the Fisher information matrix for one observation we get

$$I(\theta) = I((\nu, \lambda)) = \begin{pmatrix} \mathrm{E}\, Z^2 & \frac{1}{\lambda}\mathrm{E}\, Z \\ \frac{1}{\lambda}\mathrm{E}\, Z & \frac{1}{\lambda^2} \end{pmatrix}. \tag{5.16}$$

Its inverse is given by

$$I(\theta)^{-1} = \frac{1}{\mathrm{Var} Z} \begin{pmatrix} 1 & -\lambda \mathrm{E}\, Z \\ -\lambda \mathrm{E}\, Z & \lambda^2 \mathrm{E}\, Z^2 \end{pmatrix}. \tag{5.17}$$

The efficient influence function for estimating ν in the restricted model where λ is known, is given by (5.5) and equals

$$\frac{1}{\mathrm{E}\, Z^2} z(1 - \exp(z\nu)\lambda y). \tag{5.18}$$

In the larger model with both ν and λ unknown the score function ℓ_1^* for estimating ν equals (cf. (5.6))

$$\begin{aligned}
& z(1 - \exp(z\nu)\lambda y) - \lambda \mathrm{E}\, Z(\frac{1}{\lambda} - \exp(z\nu)y) \\
&= (z - \mathrm{E}\, Z)(1 - \exp(z\nu)\lambda y).
\end{aligned}$$

The information lower bounds for the restricted and unrestricted model are equal to

$$\frac{1}{\mathrm{E}\, Z^2} \quad \text{and} \quad \frac{1}{\mathrm{Var} Z}. \tag{5.19}$$

Since $\mathrm{Var} Z \leq \mathrm{E}\, Z^2$ the lower bound in the unrestricted model is indeed larger than the lower bound for the restricted model. Equality occurs if and only if $\mathrm{E}\, Z = 0$. It can be shown that maximum likelihood estimators attain the lower bounds. $\square$

5.1 Exercises Chapter 5

Exercise 5.1 (Bivariate Normality) Consider the regular parametric model $\mathcal{P}$ with $\theta = (\nu, \Sigma)$ and P_θ the normal distribution with mean vector $\nu = \binom{\nu_1}{\nu_2}$ and covariance matrix $\Sigma = \left(\begin{smallmatrix} \eta_1^2 & \rho\eta_1\eta_2 \\ \rho\eta_1\eta_2 & \eta_2^2 \end{smallmatrix}\right)$. Show that knowledge of Σ does not help asymptotically in estimating ν, as compared to the situation with Σ unknown. Prove that the information lower bound in estimating the correlation coefficient ρ is $(1-\rho^2)^2$; see Example 2.4.6 of BKRW (1993). $\square$

Exercise 5.2 (Plug-in Estimators) Consider the regular parametric model $\mathcal{P}$ from Definition 3.1 with the parameter $\theta = \binom{\nu}{\eta}$ split up into the parameter of interest $\nu \in \mathbb{R}^m$ and the nuisance parameter $\eta \in \mathbb{R}^{k-m}$. Let $\hat{\eta}_n$ be an efficient estimator of η based on n i.i.d. observations $X_1, \ldots, X_n$ from P_θ. For η known let $\hat{\nu}_n(\eta)$ be an efficient estimator of ν within the restricted model $\mathcal{P}_1(\eta) = \{P_\theta : \nu \in N\}, \eta$ fixed. Use (3.20) to show heuristically that $\hat{\nu}_n(\hat{\eta}_n)$ is an efficient estimator of ν within the full model $\mathcal{P}$.

A sample splitting technique as will be discussed in Chapter 8 enables one to construct a version of $\hat{\nu}_n(\hat{\eta}_n)$ that can be proved rigorously to be efficient. $\square$

Chapter 6

Semiparametric Models

In Chapter 4 we have discussed the Local Asymptotic Spread Theorem as a bound on the asymptotic performance of estimators of a parameter $\nu(P_\theta) = q(\theta)$ in regular parametric models. The regularity condition we needed on $\nu : \mathcal{P} \to \mathbb{R}^m$, was that $q : \Theta \to \mathbb{R}^m$ be twice (continuously) differentiable. For the more general nonparametric and semiparametric models $\mathcal{P}$ that we will study in this chapter, we will need a generalization of the differentiability of $q(\cdot)$, to wit pathwise differentiability of $\nu(\cdot)$. To this end we need the concept of tangent space.

Definition 6.1 (Paths and Tangent Space) Let $\mathcal{P}$ be a model and fix $P_0 \in \mathcal{P}$. A one-dimensional regular parametric submodel $\{P_\eta : |\eta| < \epsilon\}$, $\epsilon > 0$, containing P_0 is called a *curve* or *path* through P_0. Its score function h at P_0 is called a tangent and the collection of all such possible tangents h is called the *tangent set* $\dot{\mathcal{P}}^0$ of $\mathcal{P}$ at P_0. The *tangent space* $\dot{\mathcal{P}}$ is defined as the closed linear span $[\dot{\mathcal{P}}^0]$ of $\dot{\mathcal{P}}^0$ in $\mathcal{L}_2(P_0)$.

For $m = 1$ our definition of pathwise differentiability reads as follows.

Definition 6.2 (Pathwise Differentiability) The parameter $\nu : \mathcal{P} \to \mathbb{R}$ is *pathwise differentiable on $\mathcal{P}$ at $P_0 \in \mathcal{P}$*, if there exists a bounded linear functional $\dot{\nu}(P_0) : \dot{\mathcal{P}} \to \mathbb{R}$ such that for every path $\{P_\eta : |\eta| < \epsilon\}$, $\epsilon > 0$, with tangent h at P_0

$$\nu(P_\eta) = \nu(P_0) + \eta\dot{\nu}(P_0)(h) + o(\eta), \text{ as } \eta \to 0. \tag{6.1}$$

We will denote the inner product and norm of the Hilbert space $\mathcal{L}_2(P_0)$ by $\langle\cdot,\cdot\rangle_0$ and $\|\cdot\|_0$ respectively. Here $\langle f, g\rangle_0 = \mathrm{E}_{P_0} f(X)g(X)$. In fact, we are dealing with $\mathcal{L}_2^0(P_0) = \{f \in \mathcal{L}_2(P_0) : \mathrm{E}_{P_0} f(X) = 0\}$ usually. Within this subspace $\langle f, g\rangle_0$ may be interpreted as $\mathrm{Cov}_{P_0}(f(X), g(X))$ and $\|f\|_0$ becomes the standard deviation of $f(X)$ under P_0.

By the Riesz representation theorem, see for example Rudin (1966), there exists a unique $\dot{\nu} \in \dot{\mathcal{P}} \subset \mathcal{L}_2^0(P_0)$ such that

$$\dot{\nu}(P_0)(h) = \langle\dot{\nu}, h\rangle_0, \quad \text{for all } h \in \dot{\mathcal{P}}. \tag{6.2}$$

We will call this $\dot{\nu} \in \dot{\mathcal{P}}$ the *derivative* of ν.

Of course, $\nu = (\nu_1, \ldots, \nu_m)^T : \mathcal{P} \to \mathbb{R}^m$ will be called *pathwise differentiable on $\mathcal{P}$ at P_0* with derivative $\dot{\nu} = (\dot{\nu}_1, \ldots, \dot{\nu}_m)^T$ if each $\nu_i : \mathcal{P} \to \mathbb{R}$ is pathwise differentiable with derivative $\dot{\nu}_i \in \dot{\mathcal{P}} \subset \mathcal{L}_2^0(P_0)$, $i = 1, \ldots, m$.

Example 6.1 (Regular Parametric Models) Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, be a regular parametric model with score function $\dot{\ell}$ at $P_0 = P_{\theta_0}$, θ_0 fixed in Θ. The tangent space $\dot{\mathcal{P}} = [\dot{\ell}]$ at P_0 is the (k-dimensional) closed linear span of the components of $\dot{\ell}$. Let $\nu(P_\theta) = q(\theta)$, where $q(\cdot)$ is differentiable at θ_0 with (mxk)-matrix

$$\dot{q}(\theta_0) = \left(\frac{\partial q_i(\theta)}{\partial \theta_j}\right)_{j=1,\ldots,k}^{i=1,\ldots,m}$$

of derivatives. Choose $a \in \mathbb{R}^k$, $a \neq 0$. Then, $P_\eta = P_{\theta_0+\eta a}$ defines a one-dimensional regular parametric submodel of $\mathcal{P}$ with score function $a^T\dot{\ell} = \dot{\ell}^T a$ at P_0. Note

$$\begin{aligned} \nu(P_\eta) &= q(\theta_0 + \eta a) = q(\theta_0) + \eta \dot{q}(\theta_0)a + o(\eta) \\ &= \nu(P_0) + \eta\langle \dot{q}(\theta_0) I^{-1}(\theta_0)\dot{\ell},\ \dot{\ell}^T a\rangle_0 + o(\eta). \end{aligned} \tag{6.3}$$

This shows that $\nu : \mathcal{P} \to \mathbb{R}^m$ is pathwise differentiable on $\mathcal{P}$ at P_0 with derivative

$$\dot{\nu} = \dot{q}(\theta_0) I^{-1}(\theta_0)\dot{\ell} = \tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{P}) \tag{6.4}$$

the efficient influence function! □

In fact, $\dot{\nu}$ may be called the efficient influence function at P_0 for estimating ν within $\mathcal{P}$, whether $\mathcal{P}$ is regular parametric or not. As an illustration we prove

Theorem 6.1 (Regular Parametric Submodels) *Let $\nu : \mathcal{P} \to \mathbb{R}^m$ be pathwise differentiable at P_0 with derivative $\dot{\nu}$. Let $\mathcal{Q}$ be any regular parametric submodel of $\mathcal{P}$ with $I^{-1}(P_0 \mid \nu, \mathcal{Q})$ well-defined. If $P_0 \in \mathcal{Q}$, then*

$$\tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{Q}) = \Pi_0(\dot{\nu} \mid \dot{\mathcal{Q}}), \tag{6.5}$$

where Π_0 denotes projection under $\langle \cdot, \cdot \rangle_0$ within $\mathcal{L}_2^0(P_0)$. Furthermore,

$$I^{-1}(P_0 \mid \nu, \mathcal{Q}) \leq \langle \dot{\nu}, \dot{\nu}^T \rangle_0 \tag{6.6}$$

with equality iff $\dot{\nu} \in \dot{\mathcal{Q}}^m$.

Proof.

Let $\mathcal{Q}$ be *any* submodel of $\mathcal{P}$ and let $\nu_{\mathcal{Q}} : \mathcal{Q} \to \mathbb{R}^m$ denote the restriction of ν to $\mathcal{Q}$. Then $\nu_{\mathcal{Q}}$ is pathwise differentiable with derivative $\dot{\nu}_{\mathcal{Q}} \in \dot{\mathcal{Q}} \subset \dot{\mathcal{P}}$. Note that (6.1) implies that $\dot{\nu}_{\mathcal{Q}}(P_0) : \dot{\mathcal{Q}} \to \mathbb{R}^m$ is unique and equals the restriction of $\dot{\nu}(P_0) : \dot{\mathcal{P}} \to \mathbb{R}^m$ to $\dot{\mathcal{Q}}$. Consequently, we have

$$\begin{aligned} &\langle \dot{\nu}_{\mathcal{Q}} - \Pi_0(\dot{\nu} \mid \dot{\mathcal{Q}}), h\rangle_0 = \langle \dot{\nu}_{\mathcal{Q}} - \dot{\nu}, h\rangle_0 \\ &\quad = \dot{\nu}_{\mathcal{Q}}(P_0)(h) - \dot{\nu}(P_0)(h) = 0, \text{ for all } h \in \dot{\mathcal{Q}}. \end{aligned} \tag{6.7}$$

Together with $\dot{\nu}_{\mathcal{Q}} - \Pi_0(\dot{\nu} \mid \dot{\mathcal{Q}}) \in \dot{\mathcal{Q}}$ this yields

$$\dot{\nu}_{\mathcal{Q}} = \Pi_0(\dot{\nu} \mid \dot{\mathcal{Q}}). \tag{6.8}$$

If $\mathcal{Q}$ is a regular parametric submodel, then combination of (6.8) and (6.4) of Example 6.1 proves (6.5).

Furthermore, for any $a \in \mathbb{R}^m$ this implies

$$
\begin{aligned}
a^T I^{-1}(P_0 \mid \nu, \mathcal{Q}) a &= a^T \mathrm{E}_{P_0} \left(\tilde{l}_Q \tilde{l}_Q^T \right) a = \mathrm{E}_{P_0} \left(a^T \tilde{l}_Q \tilde{l}_Q^T a \right) = \mathrm{E}_{P_0} \left(a^T \tilde{l}_Q (\tilde{l}_Q a)^T \right) \\
&= \mathrm{E}_{P_0} \left((a^T \tilde{l}_Q)^2 \right) = \|a^T \Pi_0(\dot{\nu} \mid \dot{\mathcal{Q}})\|_0^2 = \|\Pi_0(a^T \dot{\nu} \mid \dot{\mathcal{Q}})\|_0^2 \\
&\leq \|a^T \dot{\nu}\|_0^2 = a^T \langle \dot{\nu}, \dot{\nu}^T \rangle_0 a.
\end{aligned} \tag{6.9}
$$

The resulting inequality with a arbitrary is the meaning and content of inequality (6.6). □

Theorem 6.1 suggests the following definitions.

Definition 6.3 (Efficient Influence Function and Information Bound) If $\nu : \mathcal{P} \to \mathbb{R}^m$ is pathwise differentiable with derivative $\dot{\nu}$ at P_0, then $\dot{\nu}$ is called the *efficient influence function* at P_0 for estimating ν within $\mathcal{P}$; notation

$$\tilde{\ell} = \tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{P}) = \dot{\nu}. \tag{6.10}$$

The *information bound* at P_0 for estimating ν within $\mathcal{P}$ is defined as

$$I^{-1}(P_0 \mid \nu, \mathcal{P}) = \langle \dot{\nu}, \dot{\nu}^T \rangle_0. \tag{6.11}$$

A justification for these definitions is given by the following generalization of Theorem 4.1.

Theorem 6.2 (LAS Theorem) *Let $\nu : \mathcal{P} \to \mathbb{R}^m$ be pathwise differentiable on $\mathcal{P}$ at P_0 and let T_n be any estimator of ν. Let $G_{\mathcal{Q},n,\sigma,a}$ for $a \in \mathbb{R}^m$ denote the distribution function of $\sqrt{n}\, a^T(T_n - q(\vartheta))$ for any regular parametric submodel $\mathcal{Q} = \{P_\theta : \theta \in \Theta\}$ with $q(\theta) = \nu(P_\theta)$ and with ϑ and σ as in (4.1). If*

$$[\dot{\nu}] \subset \overline{\dot{\mathcal{P}}^0}, \tag{6.12}$$

holds with $\overline{\dot{\mathcal{P}}^0}$ the closure of the tangent set $\dot{\mathcal{P}}^0$, then asymptotically, as $n \to \infty$ and subsequently $\sigma \to \infty$, $G_{\mathcal{Q},n,\sigma,a}$ is at least as spread out as $\mathcal{N}(0, I^{-1}(P_0|\nu, \mathcal{P}))$, i.e.

$$
\begin{aligned}
&\sup_{\mathcal{Q}} \limsup_{\sigma\to\infty} \limsup_{n\to\infty} \; G^{-1}_{\mathcal{Q},n,\sigma,a}(v) - G^{-1}_{\mathcal{Q},n,\sigma,a}(u) \\
&\quad \geq \sup_{\mathcal{Q}} \liminf_{\sigma\to\infty} \liminf_{n\to\infty} \; G^{-1}_{\mathcal{Q},n,\sigma,a}(v) - G^{-1}_{\mathcal{Q},n,\sigma,a}(u) \\
&\quad \geq (a^T I^{-1}(P_0|\nu, \mathcal{P}) a)^{1/2} (\Phi^{-1}(v) - \Phi^{-1}(u)), \quad 0 < u < v < 1.
\end{aligned} \tag{6.13}
$$

Moreover, equalities hold iff T_n is asymptotically linear with influence function $\tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{P})$ under P_0, and T_n is called efficient then at P_0.

The proof of this LAS Theorem is similar to the proof of the LAS Theorem 4.1 and involves an approximation procedure based on (6.12).

In the remainder of this chapter we will study consequences of the above results for semiparametric models $\mathcal{P} = \{P_{\nu,G} : \nu \in N,\ G \in \mathcal{G}\}$, where $N \subset \mathbb{R}^m$ is open and $\mathcal{G}$ is general. Fix $\nu_0 \in N$ and $G_0 \in \mathcal{G}$ and denote $P_0 = P_{\nu_0,G_0}$. We will assume that the parametric submodel

$$\mathcal{P}_1 = \mathcal{P}_1(G_0) = \{P_{\nu,G_0} : \nu \in N\} \tag{6.14}$$

of $\mathcal{P}$ is regular with score function $\dot{\ell}_1$. Furthermore, we will denote the submodel with $\nu = \nu_0$ fixed by

$$\mathcal{P}_2 = \mathcal{P}_2(\nu_0) = \{P_{\nu_o,G} : G \in \mathcal{G}\} \tag{6.15}$$

and the tangent space of $\mathcal{P}_2$ at P_0 by $\dot{\mathcal{P}}_2$. Analogously to (5.8) we introduce the *efficient score function* for ν by

$$\ell_1^\star(\cdot, P_0 \mid \nu, \mathcal{P}) = \ell_1^\star = \dot{\ell}_1 - \Pi(\dot{\ell}_1 \mid \dot{\mathcal{P}}_2). \tag{6.16}$$

This terminology is motivated by the following result.

Theorem 6.3 (Efficient Score Function) *Let $\mathcal{P}_1$ be regular and $\ell_1^\star$ be defined by* (6.16)*. If $\mathcal{Q} = \{P_{\nu,G_\gamma} : \nu \in N,\ \gamma \in \Gamma\}$ is a regular parametric submodel of $\mathcal{P}$ containing P_0 and hence $\mathcal{P}_1$, then*

$$I(P_0 \mid \nu, \mathcal{Q}) \geq \mathrm{E}\, \ell_1^\star \ell_1^{\star T} \tag{6.17}$$

with equality iff $[\ell_1^\star] \subset \dot{\mathcal{Q}}$. Furthermore, if $\dot{\mathcal{P}} = \dot{\mathcal{P}}_1 + \dot{\mathcal{P}}_2 = [\dot{\ell}_1] + \dot{\mathcal{P}}_2$ and (6.16) *are fulfilled for $\nu(P_{\nu,G}) = \nu$, then*

$$\dot{\nu} = \tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{P}) = \tilde{\ell}_1 = (\mathrm{E}\, \ell_1^\star \ell_1^{\star T})^{-1} \ell_1^\star \tag{6.18}$$

and the information bound equals

$$I^{-1}(P_0 \mid \nu, \mathcal{P}) = (\mathrm{E}\, \ell_1^\star \ell^{\star T}_1)^{-1}. \tag{6.19}$$

Proof.

In view of (5.6), (5.4) and (5.3), the Pythagorean Theorem yields for all $a \in \mathbb{R}^m$

$$\begin{aligned} a^T I(P_0 \mid \nu, \mathcal{Q}) a &= \|a^T(\dot{\ell}_1 - \Pi(\dot{\ell}_1 \mid \dot{\mathcal{Q}}_2))\|_0^2 \\ &= \|a^T \ell_1^\star + \Pi(a^T \dot{\ell}_1 \mid \dot{\mathcal{P}}_2) - \Pi(a^T \dot{\ell}_1 \mid \dot{\mathcal{Q}}_2)\|_0^2 \\ &= \|a^T \ell_1^\star\|_0^2 + \|\Pi(a^T \dot{\ell}_1 \mid \dot{\mathcal{P}}_2) - \Pi(a^T \dot{\ell}_1 \mid \dot{\mathcal{Q}}_2)\|_0^2 \\ &\geq a^T(\mathrm{E}\, \ell_1^\star \ell^{\star T}_1) a, \end{aligned} \tag{6.20}$$

and hence (6.17).

The pathwise differentiability of ν with derivative $\dot{\nu}$ at P_0 yields

$$\nu - \nu_0 = \nu(P_{\nu,G_0}) - \nu(P_0) = \langle \dot{\nu}, \dot{\ell}_1^T \rangle_0 (\nu - \nu_0) + o(|\nu - \nu_0|) \tag{6.21}$$

and hence

$$\langle \dot{\nu}, \dot{\ell}_1^T \rangle_0 \text{ equals the } m \times m \text{ identity matrix.} \tag{6.22}$$

The differentiability also yields for any path $\{P_{\nu_0,G_\gamma} : |\gamma| < \epsilon\} \subset \mathcal{P}_2$ with tangent h,

$$0 = \nu(P_{\nu_o,G_\gamma}) - \nu(P_0) = \gamma \langle \dot{\nu}, h \rangle_0 + o(|\gamma|) \tag{6.23}$$

and hence

$$[\dot\nu] \perp \dot{\mathcal{P}}_2. \tag{6.24}$$

Since the efficient influence function $\tilde\ell = \dot\nu$ belongs to $\dot{\mathcal{P}} = [\dot\ell_1] + \dot{\mathcal{P}}_2$ there exists an $m \times m$ matrix A and $h \in \dot{\mathcal{P}}_2^m$, such that

$$\dot\nu = A\ell_1^\star + h. \tag{6.25}$$

By (6.24) this implies

$$0 = \langle \dot\nu, h^T\rangle_0 = A\langle \ell_1^\star, h^T\rangle_0 + \langle h, h^T\rangle_0 = \langle h, h^T\rangle_0, \tag{6.26}$$

that is $h = 0$ or in other words

$$\dot\nu = A\ell_1^\star. \tag{6.27}$$

But, by (6.22) this shows that the identity matrix equals

$$\langle \dot\nu, \dot\ell_1^T\rangle_0 = A\langle \ell_1^\star, \dot\ell_1^T\rangle_0 = A\langle \ell_1^\star, \ell_1^{\star T}\rangle_0 \tag{6.28}$$

and hence $A = (\langle \ell_1^\star, \ell_1^{\star T}\rangle_0)^{-1}$ or

$$\dot\nu = (\langle \ell_1^\star, \ell_1^{\star T}\rangle_0)^{-1}\ell_1^\star. \tag{6.29}$$

□

Like we noted about (5.8), the projection in (6.16) and Theorem 6.3 show that, knowing the score function within the smaller, parametric submodel, we can obtain the score function of the same parameter within the bigger, semiparametric model by projection. Similarly, as we noted about (5.9), Theorem 6.1 and the LAS Theorem 6.2 show that the efficient influence function for the smaller (parametric) submodel may be obtained by projection of the efficient influence function of the bigger, nonparametric model. □

Example 6.2 (Construction of one dimensional families) Let $\mathcal{G}$ be the family of *all probability measures dominated by* μ. Let G_0 be a fixed element of $\mathcal{G}$ with density g_0 and let h be an element of $\mathcal{L}_2(G_0)$ such that $\int h(t)g_0(t)dt = 0$. We construct a one dimensional regular parametric subfamily with tangent h as follows.

Suppose that the function $\Psi : \mathbb{R} \to (0, \infty)$ is bounded and continuously differentiable with a bounded derivative Ψ'. Furthermore let $\Psi(0) = \Psi'(0) = 1$ and let Ψ'/Ψ be bounded. An example of such a function is

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

Define

$$g_\eta(x) = \frac{g_0(x)\Psi(\eta h(x))}{\int_{-\infty}^{\infty} g_0(t)\Psi(\eta h(t))dt}, -\infty < \eta < \infty. \tag{6.30}$$

We have

$$\frac{d}{d\eta} g_0(x)\Psi(\eta h(x)) \quad = \quad g_0(x)h(x)\Psi'(\eta h(x)) \tag{6.31}$$

$$\frac{d}{d\eta}\int_{-\infty}^{\infty} g_0(t)\Psi(\eta h(t))dt \quad = \quad \int_{-\infty}^{\infty} g_0(t)h(t)\Psi'(\eta h(t))dt. \tag{6.32}$$

Note that for $\eta = 0$ the derivatives (6.31) and (6.32) equal $g_0(x)h(x)$ and zero respectively. The derivative at zero of $g_\eta(x)$ equals

$$\frac{d}{d\eta} g_\eta(x)\Big|_{\eta=0} = \frac{1.g_0(x)h(x) - g_0(x).0}{1^2} = g_0(x)h(x). \tag{6.33}$$

Computing the score function for $\eta = 0$ we get

$$\dot{l}_0(x) = \frac{d}{d\eta}\log(g_\eta(x))\Big|_{\eta=0} = \frac{\frac{d}{d\eta} g_\eta(x)\Big|_{\eta=0}}{g_0(x)} = h(x). \tag{6.34}$$

By checking the conditions of Proposition 3.1 it follows that this one dimensional model is regular, with score function h at $\eta = 0$. i.e. at G_0. By Lemma B.1 all score functions at $\eta = 0$ satisfy $\int h(t)g_0(t)dt = 0$. Hence the tangent set at G_0 is equal to

$$\dot{\mathcal{P}}^0 = \{h \in \mathcal{L}_2(G_0) : \int h(t)g_0(t)dt = 0\} = \mathcal{L}_2^0(G_0). \tag{6.35}$$

Since the tangent set is already a closed linear space here we also have

$$\dot{\mathcal{P}} = \mathcal{L}_2^0(G_0). \tag{6.36}$$

Next consider $\mathcal{G}$ to be the family of *all probability measures dominated by* μ *which are symmetric around* ν_0. For each h in $\mathcal{L}_2(G_0)$ which is symmetric around ν_0, the construction above yields a one dimensional regular submodel of $\mathcal{G}$. On the other hand it is easy to see that all score functions for this model have to be symmeric around ν_0. Hence here we have

$$\begin{aligned} \dot{\mathcal{P}} = \dot{\mathcal{P}}^0 = \{h \in \mathcal{L}_2(G_0) : \int h(t)g_0(t)dt = 0, \\ h(\nu_0 - x) = h(\nu_0 + x), a.e.\}. \end{aligned} \tag{6.37}$$

Example 6.3 (Symmetric Location) Let $\mathcal{P} = \{P_{\nu,G} : \nu \in \mathbb{R}, G \in \mathcal{G}\}$ with $P_{\nu,G}$ the distribution with density $g(\cdot - \nu)$; here $g(-x) = g(x)$ and g has finite Fisher information for location (cf. (3.31)).

For this model the submodels $\mathcal{P}_1$ and $\mathcal{P}_2$ are defined by

$$\mathcal{P}_1 = \{P_{\nu,G_0} : \nu \in \mathbb{R}\}, \tag{6.38}$$

$$\mathcal{P}_2 = \{P_{\nu_0,G} : G \in \mathcal{G},\} \tag{6.39}$$

where $\mathcal{G}$ denotes the family of distribution functions of distributions which are symmetric around zero and have finite Fisher information for location. We can compute the tangent space $\dot{\mathcal{P}}_2$ by the same construction as in the previous example for the symmetric distributions, see Exercise 6.1. It then turns out that the efficient influence function for estimating ν equals $\ell_1^\star = \dot{l}_1$ and that the information lower bound equals

$$(\mathrm{E}\,\ell_1^{\star 2})^{-1} = I(g_0)^{-1}, \tag{6.40}$$

where $I(g) = \int (g'/g)^2 g$ is the Fisher information for location. The lower bounds for estimating ν with G known or unknown coincide.

6.1 Exercises Chapter 6

Exercise 6.1 (Symmetric Location) Prove the following statements for the symmetric location model of Example 6.3 (see Example 3.2.4 of BKRW (1993)

$$\begin{aligned} \dot{\mathcal{P}} &= [\dot{\ell}_1] + \dot{\mathcal{P}}_2, \\ \dot{\ell}_1(x) &= -\frac{g_0'}{g_0}(x-\nu_0), \\ \dot{\mathcal{P}}_2 &= \{h : h(\nu_0 - x) = h(\nu_0 + x), \int h g_0 = 0, \int h^2 g_0 < \infty\}, \\ \dot{\ell}_1 &\perp \dot{\mathcal{P}}_2. \end{aligned} \tag{6.41}$$

Determine $\ell_1^\star$ and the information lower bound for estimating ν. □

Exercise 6.2 (Linear Regression) Compute the efficient influence function for estimating ν within the semiparametric linear regression model of Example 1.2; see also Exercise 3.2 □

Chapter 7

Convolution Theorem

In general, if we don't impose restrictions on the class of estimators considered, the lower bound for estimating a parameter will be equal to zero. For instance an estimator with constant value θ_0 has zero loss under the parameter value θ_0. Under the other parameter values this estimator has a nonvanishing bias, independent of the sample size n. To avoid this problem we can either restrict the class of estimators or consider an average loss instead of the loss at a fixed parameter value. Indeed, we have considered the average distribution of an arbitrary estimator in the spread inequality. The approach of restriction is used in the Convolution Theorem 7.1 below where the estimators are required to be uniformly regular.

Suppose that, instead of fixed deterministic, the value θ is random, that the resulting random variable is denoted by ϑ as before, and that it has a probability density w. Further assume that, given $\vartheta = \theta$ the observation X has probability density $p(\cdot \mid \theta)$. The squared error loss of an estimator $T = t(X)$ is then equal to

$$\begin{aligned} \mathrm{E}\,(T-\vartheta)^2 &= \int \mathrm{E}\,((T-\theta)^2|\vartheta=\theta)w(\theta)d\theta \\ &= \int \Big(\int (t(x)-\theta)^2 p(x \mid \theta)d\mu(x)\Big)w(\theta)d\theta. \end{aligned} \tag{7.1}$$

So averaging over the parameter value can be viewed as considering the parameter random just as we have done in the preceding chapters. Before we state the Convolution Theorem we will adopt this approach in some heuristics.

Consider the normal shift experiment of Example 3.2. We will choose a weight function for t, that is, we will view t as random. To that end we rename t into U with U normally distributed with mean vector 0 and nonsingular covariance matrix Σ. Since the conditional distribution of X given $U = u$ is $\mathcal{N}(Iu, I)$, we obtain writing $X = IU + V, V \sim \mathcal{N}(0, I)$ with U and V independent,

$$\begin{pmatrix} X \\ U \end{pmatrix} \sim \mathcal{N}\Big(0, \begin{pmatrix} I\Sigma I + I & I\Sigma \\ \Sigma I & \Sigma \end{pmatrix}\Big) \tag{7.2}$$

and hence

$$\begin{pmatrix} X \\ CX - U \end{pmatrix} \sim \mathcal{N}\Big(0, \begin{pmatrix} I\Sigma I + I & 0 \\ 0 & C \end{pmatrix}\Big), \quad C = \Sigma I(I\Sigma I + I)^{-1}. \tag{7.3}$$

The proofs of (7.2) and (7.3) are given in note 9. Note that (7.3) implies that $CX - U$ and any function of X are independent. Let $T = t(X)$ be an estimator of u. Writing

$$T - U = (t(X) - CX) + (CX - U) \tag{7.4}$$

we see that $T - U$ is the sum of two independent terms, namely $CX - U$ and $t(X) - CX$. Consequently, the distribution of $T - U$ is the convolution of $\mathcal{N}(0, C)$ and another distribution. Moreover, $T = t(X) = CX$ is the best estimator of u here. For the particular choice $\Sigma = \sigma^2 I^{-1}$ we obtain $C = \sigma^2/(\sigma^2 + 1)I^{-1}$. Consequently, as $\sigma^2 \to \infty$ the distribution of $T - U$ is the convolution of $\mathcal{N}(0, I^{-1})$ and another distribution, and the best estimator is $T = I^{-1}X$.

This argument is based on (7.3) and (7.4), and it shows the origin of the convolution structure in a normal shift limit experiment. In view of (3.29) and (3.30) of Example 3.2 it is clear that LAN should imply a convolution structure. A precise formulation of this has been given by Hájek (1970). This result is called the Hájek-Le Cam convolution theorem. A proof along the lines of the above argument may be found in Chapter 3 of Van den Heuvel (1996). However, we will formulate and prove it along classical lines for regular parametric models and with $\nu : \mathcal{P} \to \mathbb{R}^m$ as parameter of interest. We will metrize $\mathcal{P}$ with the total variation distance, see Section A.1.2.

Definition 7.1 (Regularity) The sequence of estimators $\{T_n\}$ with $T_n = t_n(X_1, \dots, X_n)$ is *uniformly regular* in estimating $\nu(P)$ if $\Big\{\mathcal{L}_P\Big(\sqrt{n}(T_n - \nu(P))\Big)\Big\}$ $= \Big\{\mathcal{L}_P(Z_n)\Big\}$ converges uniformly on compact subsets K of $\mathcal{P}$. In other words, there exists a family $\{\mathcal{L}_P(Z)\}$ of distributions on $\mathbb{R}^m$ such that for every bounded, continuous function g on $\mathbb{R}^m$ and for every compact $K \subset \mathcal{P}$,

$$\lim_{n\to\infty} \sup_{P\in K} |\mathrm{E}_P g(Z_n) - \mathrm{E}_P g(Z)| = 0. \tag{7.5}$$

Example 7.1 Returning to the example of an estimator T_n of θ which has a constant value θ_0 we see that such an estimator is not regular. Even for fixed P the random variable $Z_n = \sqrt{n}(T_n - \theta) = \sqrt{n}(\theta_0 - \theta)$ does not converge in distribution for all θ values. By imposing uniform regularity on the estimators these constant estimators will therefore be excluded.

Definition 7.2 (Linearity) The sequence of estimators $\{T_n\}$ is *uniformly asymptotically linear* in the influence function $\psi : \mathcal{X} \times \mathcal{P} \to \mathbb{R}^m$ with $\mathrm{E}_P|\psi(X,P)|^2 < \infty$, $\mathrm{E}_P\psi(X,P) = 0$, if

$$\lim_{n\to\infty} \sup_{P\in K} P\Big(\sqrt{n}\Big|T_n - \{\nu(P) + \frac{1}{n}\sum_{i=1}^{n}\psi(X_i, P)\Big\}\Big| > \epsilon\Big) = 0 \tag{7.6}$$

for every positive ϵ and all compacts $K \subset \mathcal{P}$.

Note that this definition extends the pointwise asymptotic linearity introduced in Definition 4.1. The map $\nu : \mathcal{P} \to \mathbb{R}^m$ and the parametrization $\theta \to P_\theta$ together constitute a map $q : \Theta \to \mathbb{R}^m$. As before, we will assume that q is differentiable with $\dot{q}(\theta)$ the $m \times k$-matrix of partial derivatives $\partial q_i(\theta)/\partial\theta_j,\ i = 1, \dots, m,\ j = 1, \dots, k$.

These definitions enable us to formulate the convolution theorem and the resulting concept of efficiency as follows.

Theorem 7.1 (Convolution Theorem) *Let $\{T_n\}$ be a uniformly regular sequence of estimators of $q(\theta) = \nu(P_\theta)$ in the regular parametric model $\mathcal{P}$. Let $A(\theta)$ be an $\ell \times k$ matrix which is continuous in θ and let $S_n(\theta)$ be the score function defined in* (3.16)*. If $\dot{q}(\theta)$ is continuous in θ, then uniformly on compact subsets K of Θ*

$$\begin{pmatrix} \sqrt{n}(T_n - \nu(P_\theta) - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i; P_\theta \mid \nu, \mathcal{P})) \\ A(\theta)S_n(\theta) \end{pmatrix} \xrightarrow{\mathcal{D}}_\theta \begin{pmatrix} \Delta_\theta \\ S_\theta \end{pmatrix}, \tag{7.7}$$

where the random m-vectors Δ_θ and ℓ-vectors S_θ are independent. Consequently, the limit distribution of $\sqrt{n}(T_n - \nu(P_\theta))$ under P_θ is the convolution of the normal $\mathcal{N}(0, I^{-1}(P_\theta \mid \nu, \mathcal{P}))$ distribution and the distribution of Δ_θ.

Remark 7.1 (Efficiency) Adding up the two components of the vectors in (7.7) with $A(\theta) = \dot{q}(\theta)I^{-1}(\theta)$, we see that

$$\sqrt{n}\Big(T_n - \nu(P_\theta)\Big) \xrightarrow{\mathcal{D}}_\theta S_\theta + \Delta_\theta \tag{7.8}$$

with S_θ normal $\mathcal{N}\Big(0, I^{-1}(P_\theta \mid \nu, \mathcal{P})\Big)$ and with S_θ and Δ_θ independent. In terms of covariance matrices this independence yields

$$\mathrm{Var}(S_\theta + \Delta_\theta) = \mathrm{Var}S_\theta + \mathrm{Var}\Delta_\theta \geq \mathrm{Var}S_\theta \tag{7.9}$$

in the partial ordering of matrices defined by $A \geq B$ iff $A - B$ is positive semidefinite. This shows that a sequence $\{T_n\}$ of estimators is asymptotically optimal if and only if Δ_θ is degenerate at 0. In view of (7.7) this is equivalent to uniform asymptotic linearity of $\{T_n\}$ in the efficient influence function $\tilde{\ell}(\cdot; P_\theta \mid \nu, \mathcal{P})$. Such estimators are called *uniformly asymptotically efficient* and this also explains the terminology of Definition 4.2 once more.

Proof of Theorem 7.1

The concept of tightness and Prohorov's theorem which are needed in this proof are given in Section A.1.3.

Consider a sequence $\{\theta_n\}$ converging to a fixed θ. Denote

$$(U_n, V_n) = \Big(\sqrt{n}\Big(T_n - q(\theta_n)\Big),\ S_n(\theta_n)\Big). \tag{7.10}$$

By (7.5) and (3.19) $\{U_n\}$ and $\{V_n\}$ are marginally convergent in distribution, hence marginally tight, and consequently jointly tight. In view of Prohorov's theorem any subsequence $\{n'\}$ of $\{n\}$ has a further subsequence $\{n''\}$ such that

$$(U_{n''}, V_{n''}) \to (U, V). \tag{7.11}$$

By an abuse of notation we will write n instead of n''. Let

$$W_n = L_n\Big(\theta_n + \frac{t}{\sqrt{n}}\Big) - L_n(\theta_n),$$

and note that by (3.18) of the LAN-Theorem 3.1

$$(U_n, W_n) \to \Big(U, t^T V - \frac{1}{2} t^T I(\theta) t\Big) = (U, W). \tag{7.12}$$

By the continuous differentiability of $q(\cdot)$ and the uniform regularity of $\{T_n\}$ we have, for all $a \in \mathbb{R}^m$,

$$\begin{aligned} & \lim_{n\to\infty} \mathrm{E}_{\theta_n + t/\sqrt{n}} \exp\{ia^T U_n\} \\ = & \lim_{n\to\infty} \mathrm{E}_{\theta_n + t\sqrt{n}} \exp\Big\{ia^T \sqrt{n}\Big(T_n - q(\theta_n + t/\sqrt{n})\Big) + ia^T \dot{q}(\theta) t\Big\} \\ = & \ \mathrm{E} \exp\Big\{ia^T U + ia^T \dot{q}(\theta) t\Big\}. \end{aligned} \tag{7.13}$$

On the other hand we have

$$\mathrm{E}_{\theta_n + t/\sqrt{n}} \exp\{ia^T U_n\} = \mathrm{E}_{\theta_n} \exp\{ia^T U_n + W_n \wedge \lambda\} + R_{n\lambda} + R_n \tag{7.14}$$

with

$$\Big|R_{n\lambda}\Big| = \Big|\mathrm{E}_{\theta_n} \exp\{ia^T U_n\}\Big(e^{W_n} - e^{W_n \wedge \lambda}\Big)\Big| \le 1 - \mathrm{E}_{\theta_n} e^{W_n \wedge \lambda}, \tag{7.15}$$

$$\begin{aligned} \Big|R_n\Big| &= \Big|\mathrm{E}_{\theta_n + t/\sqrt{n}} \exp\Big\{ia^T U_n\Big\} \mathbf{1}_{[L_n(\theta_n) = -\infty]}\Big| \\ &\le P_{\theta_n + t/\sqrt{n}}\Big(L_n(\theta) = -\infty\Big). \end{aligned} \tag{7.16}$$

By contiguity (Corollary 3.1) $P_{\theta_n}\Big(L_n(\theta_n) = -\infty\Big) = 0$ implies

$$P_{\theta_n + t/\sqrt{n}}\Big(L_n(\theta_n) = -\infty\Big) \to 0$$

and hence

$$|R_n| \to 0, \quad \text{as } n \to \infty. \tag{7.17}$$

Note that (3.19), (7.12) and (3.29) imply $\mathrm{E}\, e^W = 1$. Consequently, (7.15) yields

$$\lim_{\lambda\to\infty} \lim_{n\to\infty} |R_{n\lambda}| \le \lim_{\lambda\to\infty} 1 - \mathrm{E}\, e^{W \wedge \lambda} = 0, \tag{7.18}$$

which together with (7.17) shows

$$\lim_{n\to\infty} \mathrm{E}_{\theta_n + t/\sqrt{n}} \exp\{ia^T U_n\} = \lim_{\lambda\to\infty} \mathrm{E} \exp\{ia^T U + W \wedge \lambda\} = \mathrm{E} \exp\{ia^T U + W\}. \tag{7.19}$$

Combining (7.13) and (7.19) we obtain

$$\mathrm{E} \exp\{ia^T U + t^T V - \frac{1}{2} t^T I(\theta) t\} = \exp\{ia^T \dot{q}(\theta) t\} \mathrm{E} \exp\{ia^T U\}, \ t \in \mathbb{R}^k. \tag{7.20}$$

Since both sides of (7.20) are analytic functions (cf. e.g. Theorem 2.9, p.52 of Lehmann (1959)), (7.20) also holds by analytic continuation (cf. e.g. for $t^T = -ia^T\dot{q}(\theta)I^{-1}(\theta) + ib^T A(\theta)$ with $b \in \mathbb{R}^m$ arbitrary. This results in

$$\text{(7.21)} \qquad \begin{aligned} & \mathrm{E}\,\exp\{ia^T(U - \dot{q}(\theta)I^{-1}(\theta)V + ib^T A(\theta)V\} \\ = \;& \mathrm{E}\,\exp\{ia^T U + \frac{1}{2}a^T\dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)a - \frac{1}{2}b^T A(\theta)I(\theta)A^T(\theta)b\} \end{aligned}$$

and in particular for $b = 0$, in

$$\text{(7.22)} \qquad \mathrm{E}\,\exp\{ia^T(U - \dot{q}(\theta)I^{-1}(\theta)V)\} = \mathrm{E}\,\exp\{ia^T U + \frac{1}{2}a^T\dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)a\}.$$

Combining (7.21) and (7.22), using (7.11) and recalling the notation in (7.10) and (7.7), we see that the characteristic function of $\left(\Delta_\theta^T, S_\theta^T\right)^T$ at $\left(a^T, b^T\right)^T$ can be written as

$$\text{(7.23)} \qquad \mathrm{E}\,\exp\left\{ia^T\Delta_\theta + ib^T S_\theta\right\} = \mathrm{E}\,\exp\left\{ia^T\Delta_\theta\right\}\exp\left\{-\frac{1}{2}b^T A(\theta)I(\theta)A^T(\theta)b\right\},$$

which implies the independence of Δ_θ and S_θ.

Note that (7.22) shows that the distribution of Δ_θ is the same for all subsequences $\{n''\}$. Since the distribution of S_θ does not depend on the subsequence either and since Δ_θ and S_θ are independent we obtain

$$\text{(7.24)} \qquad \begin{pmatrix} \sqrt{n}(T_n - \nu(P_{\theta_n}) - \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i; P_{\theta_n} \mid \nu, \mathcal{P})) \\ A(\theta_n)S_n(\theta_n) \end{pmatrix} \xrightarrow{\mathcal{D}}_{\theta_n} \begin{pmatrix} \Delta_\theta \\ S_\theta \end{pmatrix}.$$

However, (7.22) also shows that the distribution of Δ_θ is continuous in θ. Combining this with the continuity of the distribution of S_θ, with the independence of Δ_θ and S_θ and with (7.24) we obtain (7.7). The second part of Theorem 7.1 has been proved in Remark 7.1. □

For nonparametric (and semiparametric) models the previous classical Convolution Theorem can be extended as follows (cf. the LAS Theorem 6.2).

Theorem 7.2 (Convolution Theorem) *Let $\nu : \mathcal{P} \to \mathbb{R}^m$ be pathwise differentiable on $\mathcal{P}$ at P_0 and let T_n be a locally regular estimator of ν (on all regular parametric submodels $\mathcal{Q} = \{P_\theta : \theta \in \mathbb{R}^k,\ |\theta| < \epsilon\}$ of $\mathcal{P}$ containing P_0), that is,*

$$\text{(7.25)} \qquad \sqrt{n}(T_n - \nu\left(P_{\theta_n}\right)) \xrightarrow{\mathcal{D}}_{\theta_n} Z$$

whenever $\theta_n = \mathcal{O}(n^{-1/2})$. If

$$\text{(7.26)} \qquad [\dot{\nu}] \subset \overline{\dot{\mathcal{P}}^0},$$

the closure of the tangent set $\dot{\mathcal{P}}^0$, then for any $h \in (\overline{\dot{\mathcal{P}}^0})^\ell$

$$\text{(7.27)} \qquad \begin{pmatrix} \sqrt{n}\left(T_n - \nu(P_0) - \frac{1}{n}\sum_{i=1}^n \tilde{\ell}(X_i, P_0 \mid \nu, \mathcal{P})\right) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n h(X_i) \end{pmatrix} \xrightarrow{\mathcal{D}}_{P_0} \begin{pmatrix} \Delta_0 \\ W_0 \end{pmatrix},$$

where Δ_0 and W_0 are independent. Moreover, Δ_0 is degenerate at 0 iff T_n is regular and asymptotically linear with influence function $\tilde{\ell}(\cdot, P_0 \mid \nu, \mathcal{P})$ under P_0, and T_n is called efficient then.

The proof of this Convolution Theorem is similar to the proof of Convolution Theorem 7.1 and involves an approximation procedure based on (7.26). It may be found in Section 3.3, pp. 64, 65, of BKRW (1993).

As we have seen, LAS Theorems are valid for any sequence of estimators and Convolution Theorems for regular sequences only. For the still more specific regular sequences of asymptotically linear estimators we obtain via still a different approach asymptotic optimality as follows.

Theorem 7.3 (Regular Linear Estimators) *Let $\{T_n\}$ be a regular sequence of asymptotically linear estimators of $\nu : \mathcal{P} \to \mathbb{R}^m$ with influence function ψ at P_0. Then ν is pathwise differentiable at P_0 with derivative $\dot{\nu}$ and*

(7.28) $$\psi - \dot{\nu} = \psi - \tilde{\ell} \perp \dot{\mathcal{P}}.$$

In particular, the limit covariance matrix $< \psi, \psi^T >_0$ equals at least $I^{-1}(P_0 \mid \nu, \mathcal{P})$. Moreover, T_n is efficient at P_0 in the sense of

(7.29) $$< \psi, \psi^T >_0 = I^{-1}(P_0 \mid \nu, \mathcal{P})$$

iff

(7.30) $$\psi \in \dot{\mathcal{P}}^m$$

and then $\psi = \tilde{\ell}$.

Sketch of Proof.

Let $\{P_\eta\}$ be a path through P_0 with score function h and choose $\eta_n = t_n^{-1/2}$. Regularity and linearity of T_n imply

(7.31) $$\sqrt{n}(T_n - \nu(P_{\eta_n})) \xrightarrow{\mathcal{D}}_{P_{\eta_n}} \mathcal{N}(0, \langle \psi, \psi^T \rangle_0).$$

On the other hand linearity of T_n, LAN, and Le Cam's third lemma (e.g. lemma A.9.3 pp. 503, 504 in BKRW (1993)) yield

(7.32) $$\sqrt{n}(T_n - \nu(P_0)) \xrightarrow{\mathcal{D}}_{P_{\eta_n}} \mathcal{N}(t\langle \psi, h \rangle_0,\ \langle \psi, \psi^T \rangle_0).$$

Combining (7.31) and (7.32) we obtain

(7.33) $$\nu(P_{\eta_n}) = \nu(P_0) + \eta_n \langle \psi, h \rangle_0 + o(\eta_n).$$

Comparing (7.33) to (6.1) we arrive at (7.28). Because of $\tilde{\ell} \in \dot{\mathcal{P}}$, (7.28) implies that $< \psi, \psi^T >_0 - < \tilde{\ell}, \tilde{\ell}^T >_0$ is positive semidefinite. Finally, note that (7.28) and (7.30) together imply $\psi = \tilde{\ell}$. □

Chapter 8

Construction of Estimators

In the preceding two chapters bounds have been constructed on the asymptotic behavior of estimators of pathwise differentiable parameters $\nu : \mathcal{P} \to \mathbb{R}^m$ in non- and semiparametric models. A complete theory on construction of efficient estimators in these models, that is, estimators attaining these bounds in the limit, is lacking.

For regular parametric models Le Cam has proved that efficient estimators exist indeed. In fact, he gave a construction as follows. First, he constructed a $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of the parameter θ in the regular parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$; see Theorem 2.5.1 of BKRW (1993). Then, he discretized $\tilde{\theta}_n$ to an estimator θ_n^* taking its values in a grid in $\mathbb{R}^k$ with mesh width $cn^{-1/2}$, such that $|\theta_n^* - \tilde{\theta}_n| \le Cn^{-1/2}$ a.s. Motivated by a Newton-Raphson procedure he defined the one-step efficient estimator $\hat{\theta}_n$ by

$$\hat{\theta}_n = \theta_n^* + \frac{1}{n}\sum_{i=1}^{n} \tilde{\ell}(X_i; \theta_n^*), \tag{8.1}$$

where the efficient influence function $\tilde{\ell}(\cdot;\theta)$ is defined by (cf. (4.11))

$$\tilde{\ell}(x;\theta) = \tilde{\ell}(x; P_\theta \mid \theta, \mathcal{P}) = I^{-1}(\theta)\dot{\ell}(x;\theta), \quad x \in \mathcal{X}. \tag{8.2}$$

Theorem 8.1 (Discretization) *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^k$, be a regular parametric model and let $\tilde{\theta}_n$ be a locally $\sqrt{n}$-consistent estimator of θ based on $X_1, \dots, X_n$, that is, for every $\theta \in \mathbb{R}^k$ and every sequence $\{\theta_n\}$ with $\theta_n = \theta + \mathcal{O}(n^{-1/2})$*

$$\lim_{M\to\infty} \limsup_{n\to\infty} P_{\theta_n}\Big(|\sqrt{n}(\tilde{\theta}_n - \theta_n)| > M\Big) = 0. \tag{8.3}$$

If θ_n^* is a discretized version of $\tilde{\theta}_n$, then $\hat{\theta}_n$ as defined by (8.1) *and* (8.2) *is locally efficient in the sense (cf.* Definition 4.3 *and* Remark 7.1*)*

$$\lim_{n\to\infty} P_{\theta_n}\Big(|\sqrt{n}(\hat{\theta}_n - \{\theta_n + \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n)\})| > \epsilon\Big) = 0, \quad \epsilon > 0. \tag{8.4}$$

Proof.

Fix $\epsilon > 0$ and $\delta > 0$. In view of (8.3) and the definition of θ_n^* we may choose M sufficiently large such that for n large

$$P_{\theta_n}\left(|\sqrt{n}(\theta_n^* - \theta_n)| > M\right) < \delta. \tag{8.5}$$

Let B_n be the collection of grid points at distance at most $Mn^{-1/2}$ from θ_n. By (8.5) we have

$$\begin{aligned} & P_{\theta_n}\Big(|\sqrt{n}(\hat{\theta}_n - \{\theta_n + \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n)\})| > \epsilon\Big) \leq P_{\theta_n}(\theta_n^* \notin B_n) \\ & + \ P_{\theta_n}\Big(|\sqrt{n}(\theta_n^* - \theta_n + \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n^*) - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n))| > \epsilon, \theta_n^* \in B_n\Big) \\ & \leq \ \delta + \sum_{\theta_n' \in B_n} P_{\theta_n}\Big(|\sqrt{n}(\theta_n' - \theta_n + \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n') - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n))| > \epsilon\Big). \end{aligned} \tag{8.6}$$

Since the number of points in B_n is bounded (uniformly in n) and since δ is arbitrary, it suffices to prove that for every sequence $\{\theta_n'\}$ with $\theta_n' = \theta + \mathcal{O}(n^{-1/2})$ the probability at the right-hand side of (8.6) converges to 0. However, this holds in view of the smoothness property (3.20), the continuity of the Fisher information $I(\cdot)$ and the definition of $\tilde{\ell}(\cdot;\cdot)$ in (8.2). □

The estimator $\tilde{\theta}_n$ is called a preliminary estimator. The proof makes clear that the discretization is used to force (a kind of) independence between the preliminary estimator and the observations. Another technique to force this is by splitting the sample into two parts. One part is used for the preliminary estimator and an average is constructed of the efficient influence function at the observations from the other part. Interchanging the roles of these two parts we obtain an estimator that is asymptotically linear in the observations of the first part. An appropriate convex combination of these two estimators is efficient.

Theorem 8.2 (Sample Splitting) *Let $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ be a regular parametric model and let $\tilde{\theta}_n = t_n(X_1, \ldots, X_n)$ be a locally $\sqrt{n}$-consistent estimator of θ. With $\lambda_n \in \mathbb{N}$ we define*

$$\tilde{\theta}_{n1} = t_{\lambda_n}(X_1, \ldots, X_{\lambda_n}),\ \tilde{\theta}_{n2} = t_{n-\lambda_n}(X_{\lambda_n+1}, \ldots, X_n).$$

If

$$\lambda_n/n \to \lambda \in (0,1), \tag{8.7}$$

then the estimator

$$\hat{\theta}_n = \frac{\lambda_n}{n}\Big(\tilde{\theta}_{n2} + \frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n}\tilde{\ell}(X_i;\tilde{\theta}_{n2})\Big) + \frac{n-\lambda_n}{n}\Big(\tilde{\theta}_{n1} + \frac{1}{n-\lambda_n}\sum_{i=\lambda_n+1}^{n}\tilde{\ell}(X_i;\tilde{\theta}_{n1})\Big) \tag{8.8}$$

is locally efficient.

Proof.

First note that for (8.4) to hold with $\hat{\theta}_n$ as in (8.8), it suffices to prove

$$\lim_{n\to\infty} P_{\theta_n}\Big(|\sqrt{\lambda_n}(\tilde{\theta}_{n2} - \theta_n + \frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n}\tilde{\ell}(X_i;\tilde{\theta}_{n2}) - \frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n}\tilde{\ell}(X_i;\theta_n))| > \epsilon\Big) = 0,\ \epsilon > 0. \tag{8.9}$$

Let M be sufficiently large and let B_n be the ball of radius $M\lambda_n^{-1/2}$ about θ_n. Then the probability in (8.9) is bounded from above by

$$P_{\theta_n}(\tilde{\theta}_{n2} \notin B_n) + E_{\theta_n}\Big(\mathbf{1}_{B_n}(\tilde{\theta}_{n2})P_{\theta_n}(|\sqrt{\lambda_n}(\tilde{\theta}_{n2} - \theta_n +$$
$$+\frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n}\tilde{\ell}(X_i;\tilde{\theta}_{n2}) - \frac{1}{\lambda_n}\sum_{i=1}^{\lambda_n}\tilde{\ell}(X_i;\theta_n))| > \epsilon \mid \tilde{\theta}_{n2}\Big)).$$

But, the conditional probability converges to 0, for $\tilde{\theta}_{n2} \in B_n$, by the same argument as at the end of the proof of Theorem 8.1 in view of (8.7) and the independence of $\tilde{\theta}_{n2}$ and $(X_1, \ldots, X_{\lambda_n})$. □

Of course, both discretization and sample splitting are artificial techniques used only to make proofs work. The estimator $\hat{\theta}_n$ from (8.1) with θ_n^* replaced by $\tilde{\theta}_n$ itself would be more natural, since it is very close to a one-step Newton-Raphson approximation of the maximum likelihood equation

$$\sum_{i=1}^{n}\dot{\ell}(X_i;\theta) = 0. \tag{8.10}$$

Indeed, for $k = 1$ with $\tilde{\theta}_n$ close to a root of this equation one-step Newton-Raphson results in

$$\tilde{\theta}_n - \frac{1}{n}\sum_{i=1}^{n}\dot{\ell}(X_i;\tilde{\theta}_n)\Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\dot{\ell}(X_i;\tilde{\theta}_n)\Big\}^{-1}. \tag{8.11}$$

By the law of large numbers and under extra smoothness conditions on $\dot{\ell}(\cdot;\theta)$ the denominator should converge to

$$\begin{aligned} \mathrm{E}_\theta\frac{\partial}{\partial\theta}\dot{\ell}(X;\theta) &= \int\Big\{\frac{\partial}{\partial\theta}(\dot{\ell}(x;\theta)p(x;\theta)) - \dot{\ell}(x;\theta)\frac{\partial}{\partial\theta}p(x;\theta)\Big\}d\mu(x) \\ &= \frac{\partial}{\partial\theta}\mathrm{E}_\theta\dot{\ell}(X;\theta) - \int\dot{\ell}^2(x;\theta)p(x;\theta)d\mu(x) = -I(\theta) \end{aligned} \tag{8.12}$$

and hence (8.11) resembles (8.1). Under additional regularity conditions it can be shown that the maximum likelihood estimator itself is efficient; see pp. 500–504 of Cramér (1946) and Le Cam (1970).

Here we will apply the sample splitting technique to construct efficient estimators of the Euclidean parameter θ in the semiparametric model

$$\mathcal{P} = \{P_{\theta,G} : \theta \in \Theta,\ G \in \mathcal{G}\}, \tag{8.13}$$

where $\Theta \subset \mathbb{R}^k$ is open and $\mathcal{G}$ is general. Under the conditions of Theorem 6.3 with $\nu = \theta$ the efficient influence function is well-defined and we will write (cf. (6.18))

$$\tilde{\ell}(x;\theta;G) = (\mathrm{E}\,\ell_1^\star \ell_1^{\star T})^{-1}\ell_1^\star. \tag{8.14}$$

Since $\tilde{\ell}$ depends on G and since G is unknown, the estimator (8.8) cannot be used here. We will introduce an extra splitting step to handle this problem under the assumption that there exists an estimator $\tilde{\ell}_n(\cdot;\theta;\underline{X}) = \tilde{\ell}_n(\cdot;\theta;\,X_1,\ldots,X_n)$ of the efficient influence function $\tilde{\ell}(\cdot;\theta;G)$ satisfying (consistency)

$$\int \Big|\tilde{\ell}_n(x;\theta_n;\underline{X}) - \tilde{\ell}(x;\theta_n;G)\Big|^2 dP_{\theta_n,G}(x) = o_{\theta_n,G}(1) \tag{8.15}$$

and ($\sqrt{n}$-unbiasedness)

$$\sqrt{n}\int \tilde{\ell}_n(x;\theta_n;\underline{X})dP_{\theta_n,G}(x) = o_{\theta_n,G}(1), \tag{8.16}$$

for all $(\theta,G) \in \Theta\times\mathcal{G}$ and all sequences $\{\theta_n\}$ with $\theta_n = \theta + \mathcal{O}(n^{-1/2})$. To interpret (8.16) recall that

$$\int \tilde{\ell}(x;\theta_n;G)dP_{\theta_n,G}(x) = 0.$$

Furthermore, we assume the existence of a $\sqrt{n}$-consistent preliminary estimator of θ denoted by $\tilde{\theta}_n = t_n(X_1,\ldots,X_n)$ satisfying (8.3) under $P_{\theta_n,G}$ for every $G \in \mathcal{G}$. Note that in the parametric case such a $\sqrt{n}$-consistent estimator always exists according to Le Cam (cf. Exercise 8.1), but in a semiparametric model we do not know this beforehand (cf. Exercise 8.3).

Let $\{\lambda_n\}, \{\mu_n\}$, and $\{\nu_n\}$ be sequences of integers with $0 < \lambda_n < \mu_n < \nu_n < n$ and

$$\frac{\lambda_n}{n} \to \lambda,\ \frac{\mu_n}{n} \to \mu,\ \frac{\nu_n}{n} \to \nu, \tag{8.17}$$

with $0 < \lambda < \mu < \nu < 1$. We define (cf. Theorem 6.2)

$$\tilde{\theta}_{n1} = t_{\lambda_n}(X_1,\ldots,X_{\lambda_n}),\ \tilde{\theta}_{n2} = t_{\nu_n-\mu_n}(X_{\mu_n+1},\ldots,X_{\nu_n}), \tag{8.18}$$

$$\begin{aligned} \tilde{\ell}_{n1}(x;\theta) &= \tilde{\ell}_{\mu_n-\lambda_n}(x;\theta;X_{\lambda_n+1},\ldots,X_{\mu_n}), \\ \tilde{\ell}_{n2}(x;\theta) &= \tilde{\ell}_{n-\nu_n}(x;\theta;X_{\nu_n+1},\ldots,X_n), \end{aligned} \tag{8.19}$$

and we notice that these four estimators are independent, because they are based on the four independent blocks of observations into which our sample of n observations is split up. Analogously to (8.8) we define

$$\hat{\theta}_n = \frac{\mu_n}{n}\Big(\tilde{\theta}_{n2} + \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}_{n2}(X_i;\tilde{\theta}_{n2})\Big) + \frac{n-\mu_n}{n}\Big(\tilde{\theta}_{n1} + \frac{1}{n-\mu_n}\sum_{i=\mu_n+1}^{n}\tilde{\ell}_{n1}(X_i;\tilde{\theta}_{n1})\Big). \tag{8.20}$$

To prove efficiency of this estimator $\hat{\theta}_n$ in the semiparametric model (8.13) we assume smoothness of the efficient influence functions as follows,

$$\sqrt{n}\Big(\theta_n - \theta + \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n;G) - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta;G)\Big) = o_{\theta,G}(1), \tag{8.21}$$

for every $(\theta, G) \in \Theta \times \mathcal{G}$ and for all sequences $\{\theta_n\}$ with $\theta_n + \mathcal{O}(n^{-1/2})$. Note that in a least favorable regular parametric submodel this follows from the LAN-Theorem 3.1 via (3.20) as in the proof of Theorem 8.1.

Theorem 8.3 (Semiparametric Sample Splitting) *In the semiparametric model* (8.13) *let* $\mathcal{P}_1(G) = \{P_{\theta,G} : \theta \in \Theta\}$ *be a regular parametric submodel for every* $G \in \mathcal{G}$ *and let* $\tilde{\ell}$ *be well-defined by* (8.14)*. Let* $\tilde{\theta}_n$ *be a* $\sqrt{n}$*-consistent estimator of* θ *and let* $\tilde{\ell}_n(\cdot;\cdot;\underline{X})$ *satisfy* (8.15) *and* (8.16)*. If* (8.21) *holds, then the estimator* $\hat{\theta}_n$ *from* (8.20) *is locally uniformly efficient, i.e.*

$$\sqrt{n}\Big(\hat{\theta}_n - \theta_n - \frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}(X_i;\theta_n;G)\Big) = o_{\theta_n,G}(1), \tag{8.22}$$

for every $(\theta, G) \in \Theta \times \mathcal{G}$ *and every sequence* $\{\theta_n\}$ *with* $\theta_n = \theta + \mathcal{O}(n^{-1/2})$.

Proof. Fix $\theta \in \Theta$ and $G \in \mathcal{G}$. In view of Theorem 3.1 and Corollary 3.1 the regularity of $\mathcal{P}_1(G)$ implies the mutual contiguity of $\{P^n_{\theta,G}\}$ and $\{P^n_{\theta_n,G}\}$ for every sequence $\{\theta_n\}$ with $\theta_n = \theta + \mathcal{O}(n^{-1/2})$. Together with the smoothness condition (8.21) this shows that it suffices to prove (8.22) with $\theta_n = \theta$. We will just prove

$$P_{\theta,G}\Big(\Big|\sqrt{\mu}_n\Big(\tilde{\theta}_{n2} - \theta + \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}_{n2}(X_i;\tilde{\theta}_{n2}) - \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}(X_i;\theta;G)\Big)\Big| > \epsilon\Big) \to 0,\ \epsilon > 0. \tag{8.23}$$

As in the proof of Theorem 8.2, let M be sufficiently large and let B_n be the ball of radius $Mn^{-1/2}$ about θ. Then the probability in (8.23) is bounded from above by

$$P_{\theta,G}(\tilde{\theta}_{n2} \notin B_n) + E_{\theta,G}\Big(\mathbf{1}_{B_n}(\tilde{\theta}_{n2})P_{\theta,G}\Big(\Big|\sqrt{\mu}_n\Big(\tilde{\theta}_{n2} - \theta + \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}_{n2}(X_i;\tilde{\theta}_{n2})$$
$$-\frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}(X_i;\theta;G)\Big)\Big| > \epsilon \mid \tilde{\theta}_{n2}\Big)\Big),$$

and by the $\sqrt{n}$-consistency of $\tilde{\theta}_{n2}$, by the independence of $\tilde{\theta}_{n2}$ and $(X_1, \ldots, X_{\mu_n},\ X_{\nu_n+1}, \ldots, X_n)$ and by the dominated convergence theorem it suffices to prove that for $\theta_n = \theta + \mathcal{O}(n^{-1/2})$

$$P_{\theta,G}\Big(\Big|\sqrt{\mu}_n\Big(\theta_n - \theta + \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}_{n2}(X_i;\theta_n) - \frac{1}{\mu_n}\sum_{i=1}^{\mu_n}\tilde{\ell}(X_i;\theta,G)\Big)\Big| > \epsilon\Big) \to 0. \tag{8.24}$$

In view of (8.21) and (8.17), and because of contiguity this holds if for all $a \in \mathbb{R}^k$

$$\frac{1}{\sqrt{\mu}_n}\sum_{i=1}^{\mu_n} a^T\Big(\tilde{\ell}_{n2}(X_i;\theta_n) - \tilde{\ell}(X_i;\theta_n,G)\Big) = o_{\theta_n,G}(1). \tag{8.25}$$

Computing the conditional expectation and conditional variance of the left-hand side of (8.25) given $(X_{\nu_n+1}, \ldots, X_n)$, and using (8.16) and (8.15) we see that (8.25) holds indeed. □

Remark 8.1 (Necessity) Theorem 8.3 may be extended to general linear estimators as in Klaassen (1987). Our proof is taken from this paper. The proof as given for Theorem 7.8.1 of BKRW (1993) has been messed up at the type setting stage but has been corrected in the 1998 edition. In Klaassen (1987) it has also been proved under a uniform integrability condition on the efficient influence functions that existence of an estimator of the efficient influence function satisfying (8.15) and (8.16), is necessary for existence of an efficient estimator of θ.

Remark 8.2 For a broad class of semiparametric models estimation of the efficient influence function reduces to estimation of the score function for location $-g'/g$ for a density g. A method for this has been given by Bickel; see Bickel (1982), Bickel and Klaassen (1986) and Proposition 7.8.1, p. 400, of BKRW (1993). For other methods of estimation of θ, efficiently or just $\sqrt{n}$-consistently, we refer to Chapter 7 of BKRW (1993).

8.1 Exercises Chapter 8

Exercise 8.1 ($\sqrt{n}$-consistency) In a regular parametric model $\tilde{\theta}_n$ satisfies (8.3) iff for every $\theta \in \mathbb{R}^k$

(8.26)
$$\lim_{M\to\infty} \limsup_{n\to\infty} P_\theta\Big(\Big|\sqrt{n}(\tilde{\theta}_n - \theta)\Big| > M\Big) = 0.$$

Exercise 8.2 (Smoothness) If $\mathcal{P}_1(G)$ is regular, (8.21) is necessary for the existence of efficient estimators of θ.

Exercise 8.3 (Symmetric Location) Verify that Theorem 8.3 may be applied to the symmetric location case. To construct a $\sqrt{n}$-consistent estimator of the location parameter, consider $\psi : \mathbb{R} \to \mathbb{R}$, strictly increasing, uneven, bounded and with bounded first and second derivatives. Let $\tilde{\theta}_n$ be the M-estimator solving

(8.27)
$$\sum_{i=1}^{n} \psi(X_i - \theta) = 0.$$

Note that

(8.28)
$$P_{\theta,G}\Big(\sqrt{n}(\tilde{\theta}_n - \theta) \le y\Big) = P_{\theta,G}\Big(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(X_i - \theta - \frac{y}{\sqrt{n}}) \le 0\Big)$$

and show that $\tilde{\theta}_n$ is $\sqrt{n}$-consistent.

Exercise 8.4 (Some Semiparametric Models) *We have to cope with the following estimation problem. $X_1, \dots, X_n$ are independent and identically distributed random quantities with distribution $P_{\nu,g}$, where $\nu \in \mathbb{R}$ and $g \in \mathcal{G}$ are unknown. We want to estimate ν in the presence of the nuisance parameter g and based on $X_1, \dots, X_n$. The following questions (A. through F.) have to be answered for the estimation problems from the list following the list of questions. The degree of mathematical detail in your answers may vary and typically decreases when you proceed from question A to F. For example, to answer question B you have to determine tangent spaces on which to project. This is quite hard to do exactly, but your efforts will be appreciated!*

Furthermore, in answering question F it suffices to indicate the lines along which an efficient estimator could be constructed: recall Section 8. Answering question F in full mathematical rigor would be equivalent to writing a scientific paper about the model.

A. *Fix g_0. If $g = g_0$ is known, then what does the Local Asymptotic Spread Theorem tell us?*

B. *Determine a lower bound on the asymptotic performance of estimator sequences of ν in case g is unknown.*

C. *What is the information loss due to g being unknown?*

D. *Construct an efficient estimator of ν in case g is known.*

E. *Can you determine a $\sqrt{n}$-consistent estimator of ν when g is unknown?*

F. *(When possible!) Construct an efficient estimator of ν if g is unknown.*

Formulate the regularity conditions that you need in your answers and solutions; you are free to choose them yourself.

1. **Symmetric Location**
   Let $\mathcal{G}$ be the class of densities g on $\mathbb{R}$ with respect to Lebesgue measure λ, that are symmetric about 0, that are absolutely continuous with Radon-Nikodym derivative g', and that have finite Fisher information for location

$$I(g) = \int (\frac{g'}{g})^2 g \,.$$

   $P_{\nu,g}$ has density $g(\cdot - \nu)$ on $(\mathbb{R}, \mathcal{B}, \lambda)$.

2. **Linear Regression**
   The observation X is defined by

$$X = (Y, Z)\,,\; Y = \mu + \nu Z + \epsilon.$$

   Here, ϵ and Z are independent random variables, ϵ has unknown density g_1 on $\mathbb{R}$ with respect to Lebesgue measure λ, that is absolutely continuous with derivative g_1' and that has finite Fisher information for location

$$I(g_1) = \int (\frac{g_1'}{g_1})^2 g_1 \,,$$

   Z has unknown density g_2 on $(\mathbb{R}, \mathcal{B}, \lambda)$, and $\mu \in \mathbb{R}$ is unknown as well. Of course, $\mathcal{G}$ is the class of triplets (μ, g_1, g_2), where μ, g_1, and g_2 satisfy the above description.

3. **Heteroscedasticity**
   The generic observation X is defined by

$$X = (Y, Z)\,,\; Y = \mu + e^{\nu Z}\epsilon.$$

   Here, ϵ and Z are independent random variables, ϵ has the standard normal distribution, Z has finite second moment and an unknown density g on $(\mathbb{R}, \mathcal{B}, \lambda)$, and $\mu \in \mathbb{R}$ is unknown as well. Consequently, $\mathcal{G}$ is the class of pairs (μ, g).

4. **Partial Splines**
The generic observation X *is defined by*

$$X = (Y, Z)\,,\, Y = \nu + g(Z) + \epsilon.$$

Here, ϵ *and* Z *are independent,* ϵ *is standard normally distributed, and* Z *is uniformly distributed on* $(0,1)$. *Furthermore,* $\mathcal{G}$ *is the class of functions* $g : (0,1) \to \mathbb{R}$ *with the properties*

$$\int_0^1 g(z)dz = 0, \int_0^1 g^2(z)dz < \infty.$$

A model that comes closer to the one of Engle, Granger, Rice, and Weiss (1986), is the following. $X = (W, Y, Z), Y = \mu + \nu^T W + g(Z) + \epsilon$, *with* W, Z, *and* ϵ *independent,* $EW = 0$, $\mu \in \mathbb{R}$ *unknown, and* $\mathcal{G}$ *as above. When possible, treat this problem as well.*

5. **Projection Pursuit Regression**
The observation X *is defined by*

$$X = (Y, Z_1, Z_2)\,,\, Y = g(Z_1 + \nu Z_2) + \epsilon.$$

Here, ϵ, Z_1, *and* Z_2 *are independent standard normal random variables. Furthermore,* $g : \mathbb{R} \to \mathbb{R}$ *is an unknown differentiable function, and hence* $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} \text{ differentiable}\}$.

6. **Logistic Regression**
The generic observation X *is defined by* $X = (W, Y, Z)$, *where the random variable* Y *has a Bernoulli distribution with success probability*

$$\frac{1}{1 + e^{\nu Z + g(W)}}.$$

Here, W *and* Z *are independent standard normal random variables and* $\mathcal{G} = \{g : \mathbb{R} \to \mathbb{R} \text{ measurable}\}$.

7. **Errors in variables**
The observation X *is defined by*

$$X = (Y, Z)\,,\, Z = Z' + \epsilon_1\,,\, Y = \nu Z' + \epsilon_2.$$

Here, Z', ϵ_1, *and* ϵ_2 *are independent random variables,* ϵ_1 *and* ϵ_2 *have a standard normal distribution, and* Z' *has unknown density* g *on* $(\mathbb{R}, \mathcal{B}, \lambda)$.
Hint. $\dot{\mathcal{P}}_2$ *exists of all random variables with mean 0 and finite variance that can be written as a function of* $Z + \nu Y$.

Chapter 9

Time Series

In the preceding Chapters we have studied asymptotically efficient estimation of Euclidean parameters in semiparametric models for i.i.d. random variables. Crucial in our set-up has been Local Asymptotic Normality of regular parametric models and submodels as formulated in the LAN-Theorem 3.1. Via the LAS Theorems 4.1 and 6.2 and the Convolution Theorems 7.1 and 7.2 LAN has provided a lower bound on the asymptotic performance of estimators both in parametric and in nonparametric models, and via the smoothness properties (3.20) and (8.21) it has been essential in the proof of asymptotic linearity in the efficient influence function of the one-step estimators we have constructed; cf. the proof of Theorem 8.1, the sentence after (8.21), and Theorem 8.3.

In many applications time-series models play an important part, e.g. in econometrics. They constitute a large group of models for non-i.i.d. observations, $Y_1, \dots, Y_n$. Typically, an i.i.d. structure is hidden in a time-series as follows. Let $\epsilon_1, \dots, \epsilon_n$ be i.i.d. observations from a distribution with density g, which are independent of the random vector X_n. Assume that the σ-fields $\mathcal{F}_0^n = \mathcal{F}(X_n),\ \mathcal{F}_t^n = \mathcal{F}_{t-1}^n \vee \mathcal{F}(\epsilon_t),\ t = 1, \dots, n$, define a filtration. One observes $X_n \in \mathcal{F}_0^n$ and $Y_1, \dots, Y_n$ with

$$Y_t = \mu_t(\theta) + \sigma_t(\theta)\epsilon_t, \quad t = 1, \dots, n, \tag{9.1}$$

where the time-dependent location-scale parameter $(\mu_t(\theta),\ \sigma_t(\theta))^T \in (\mathcal{F}_{t-1}^n)^2$ is supposed to depend on θ, the observed starting values $X_n \in \mathcal{F}_0^n$, and the first $t-1$ observations $Y_1, \dots, Y_{t-1}$. So, Y_t depends on the "past" and the new independent innovation ϵ_t.

As in the i.i.d. case it is crucial to obtain LAN. Fix g and $\theta \in \Theta \subset \mathbb{R}^k$. Let θ_n denote the true parameter point, suppose $\theta_n \to \theta$, and let $\tilde{\theta}_n$ be such that $\sqrt{n}(\tilde{\theta}_n - \theta_n) \to \lambda$. We denote the log-likelihood ratio statistic of the observations $X_n, Y_1, \dots, Y_n$ for $\tilde{\theta}_n$ with respect to θ_n by Λ_n. Let

$$\mathcal{Q} = \{Q_{\mu,\sigma} :\ \mu \in \mathbb{R}, \sigma > 0\} \tag{9.2}$$

with Lebesgue densities $q(\zeta) = \sigma^{-1} g(\sigma^{-1}(\cdot - \mu)),\ \zeta = (\mu, \sigma)^T$, of $Q_{\mu,\sigma}$, be the location-scale family of g and denote

$$\ell(\zeta) = \log q(\zeta). \tag{9.3}$$

We assume that the innovations can be reconstructed from the observations if θ is known and we write

$$\epsilon_t(\theta) = \epsilon(X_n, Y_1, \dots, Y_t, \theta). \tag{9.4}$$

Under θ we have $\epsilon_t(\theta) = \epsilon_t, \ t = 1, \dots, n$. We define $k \times 2$-matrices W_{nt} by their rows, $j = 1, \dots, k$,

$$W_{ntj} = \begin{cases} \sigma_t^{-1}(\theta_n)(\tilde\theta_{nj} - \theta_{nj})^{-1}\Big(\mu_t(\tilde\theta_n^j) - \mu_t(\tilde\theta_n^{j-1}), \sigma_t(\tilde\theta_n^j) - \sigma_t(\tilde\theta_n^{j-1})\Big) & \text{if} \quad \tilde\theta_{nj} \neq \theta_{nj} \\ \sigma_t^{-1}(\theta_n)\frac{\partial}{\partial\theta_j}(\mu_t(\theta), \sigma_t(\theta)\Big)|_{\theta=\tilde\theta_n^j} & \text{if} \quad \tilde\theta_{nj} = \theta_{nj}, \end{cases} \tag{9.5}$$

where $\tilde\theta_n^j = (\tilde\theta_{n1}, \dots, \tilde\theta_{nj}, \ \theta_{n,j+1}, \dots, \theta_{nk})^T$. Note that

$$W_{nt}^T\Big(\tilde\theta_n - \theta_n\Big) = \sigma_t^{-1}(\theta_n)\Big(\mu_t(\tilde\theta_n) - \mu_t(\theta_n), \ \sigma_t(\tilde\theta_n) - \sigma_t(\theta_n)\Big)^T. \tag{9.6}$$

Now, the log-likelihood ratio statistic Λ_n may be written with $\zeta_0 = (0,1)^T$ as

$$\Lambda_n = \Lambda_n^s + \sum_{t=1}^n \Big\{\ell\Big(\zeta_0 + W_{nt}^T(\tilde\theta_n - \theta_n)\Big)\Big(\epsilon_t(\theta_n)\Big) - \ell(\zeta_0)\Big(\epsilon_t(\theta_n)\Big)\Big\}. \tag{9.7}$$

Here $\Lambda_n^s \in \mathcal{F}_0^n$ depends on the starting observations X_n only, and we assume that they have a negligible effect, i.e.

$$\Lambda_n^s = o_{\theta_n, g}(1). \tag{9.8}$$

Furthermore, we assume that $\mathcal{Q}$ is a regular parametric model with score function

$$\dot\ell(x) = \begin{pmatrix} -g'/g(x) \\ -1 - xg'/g(x) \end{pmatrix}, \quad x \in \mathbb{R}, \tag{9.9}$$

and Fisher information matrix

$$J = \mathrm{E}_g \dot\ell\dot\ell^T(X) \tag{9.10}$$

at $\zeta_0 = (0,1)^T$. Finally, we assume that there exists a continuous nonsingular Fisher information matrix $I(\theta)$ and square integrable $k \times 2$-matrices $W_t(\theta_n) \in \mathcal{F}_{t-1}^n$ satisfying

$$\frac{1}{n}\sum_{i=1}^n W_t(\theta_n) J W_t^T(\theta_n) \xrightarrow{P} I(\theta), \tag{9.11}$$

and

$$\frac{1}{n}\sum_{t=1}^n \Big|W_t(\theta_n)\Big|^2 \mathbf{1}_{[|W_t(\theta_n)| > \delta\sqrt{n}]} \xrightarrow{P} 0 \tag{9.12}$$

under θ_n for all $\delta > 0$, such that

$$\sum_{t=1}^n \Big|(W_{nt} - W_t(\theta_n))^T(\tilde\theta_n - \theta_n)\Big|^2 \xrightarrow{P} 0 \tag{9.13}$$

under θ_n. We will denote the score functions of our time-series model by

$$\dot\ell_t(\theta) = W_t(\theta)\dot\ell(\epsilon_t(\theta)), \ t = 1, \dots, n. \tag{9.14}$$

Theorem 9.1 (LAN) *Under the above conditions, write*

$$\Lambda_n = \lambda^T \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \dot{\ell}_t(\theta_n) - \frac{1}{2n} \sum_{t=1}^{n} \left| \lambda^T \dot{\ell}_t(\theta_n) \right|^2 + R_n. \tag{9.15}$$

Under θ_n and g

$$R_n \xrightarrow{P} 0, \ \ \Lambda_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(-\frac{1}{2} \lambda^T I(\theta) \lambda, \ \lambda^T I(\theta) \lambda \right). \tag{9.16}$$

The distributions of $(X_n, Y_1, \ldots, Y_n)$ under θ_n and $\tilde{\theta}_n$ are contiguous. In case $\theta_n = \theta + \mathcal{O}(n^{-1/2})$, the smoothness condition holds:

$$\sqrt{n}\Big(\tilde{\theta}_n - \theta_n + \frac{1}{n} \sum_{t=1}^{n} I^{-1}(\tilde{\theta}_n) \dot{\ell}_t(\tilde{\theta}_n) - \frac{1}{n} \sum_{t=1}^{n} I^{-1}(\theta_n) \dot{\ell}_t(\theta_n) \Big) \xrightarrow{P} 0. \tag{9.17}$$

Proof.
A proof is given in Drost, Klaassen and Werker (1997). It is along similar lines as the proof of Theorem 3.1, to wit Taylor expansion of (9.7), but with ergodic theorems and martingale difference central limit theorems replacing the law of large numbers and the "ordinary" central limit theorem. □

The classical convolution theorem of Hájek (1970) shows that LAN suffices for the convolution structure of the limit distribution of regular estimators. (Note that in the proof of Theorem 7.1 we only used (3.18), (3.19) and contiguity, hence only LAN, and not anything else about regular parametric models). Therefore, in our time-series models, estimators T_n of $q(\theta)$ are asymptotically efficient if they satisfy

$$\sqrt{n}\Big(T_n - q(\theta_n) - \frac{1}{n} \sum_{t=1}^{n} \dot{q}(\theta_n) I^{-1}(\theta_n) \dot{\ell}_t(\theta_n) \Big) \xrightarrow{P} 0 \tag{9.18}$$

under θ_n and fixed g for $\theta_n = \theta + \mathcal{O}(n^{-1/2})$. Note that we have considered a parametric model, in fact, since we kept g fixed. Of course, the convolution lower bound is still valid in the semiparametric model with g unknown. Often, under the "right" parametrization and for the "right" function $q(\cdot)$ estimators of $q(\theta)$ may be given for the semiparametric model that attain this bound from the parametric submodel with g fixed. We call such estimators adaptive. The geometry of this situation is studied in Drost, Klaassen and Werker (1994).

In Drost, Klaassen and Werker (1997), a general method for the construction of such adaptive estimators is given based on sample splitting *and* discretization for the present time-series models. We present two examples here.

Example 9.1 (ARMA (p, q)) In the Auto Regressive Moving Average model of orders p and q, we observe realizations of $Y_1, \ldots, Y_n$ with

$$Y_t = \rho_1 Y_{t-1} + \cdots + \rho_p Y_{t-p} + \varphi_1 \epsilon_{t-1} + \cdots + \varphi_q \epsilon_{t-q} + \epsilon_t. \tag{9.19}$$

As in the general model (9.1) the innovations ϵ_t are i.i.d. with unknown density g. Furthermore, $\theta = (\rho_1, \ldots, \rho_p, \varphi_1, \ldots, \varphi_q)^T$. Comparing (9.19) to (9.1) we note that $\sigma_t(\theta) = 1$

here. Consequently, we do not need the full location-scale model (9.2), but only its restriction to location. Therefore, we will assume that g has finite Fisher information for location $I_\ell(g) = \int (g'/g)^2 g$, thus obtaining regularity of the location model; cf. Exercises 3.1 and 9.1.

We will also assume $\mathrm{E}_g \epsilon_t = 0,\ \mathrm{E}_g \epsilon_t^2 < \infty$. The starting observations are collected into the vector $X_n = (Y_0, \dots, Y_{1-p},\ \epsilon_0, \dots, \epsilon_{1-q})^T$ and we assume (partially unrealistically) that X_n will be observed too. It may be checked straightforwardly that the innovations can be recovered from the observations $X_n,\ Y_1, \dots, Y_n$ as in (9.4).

Roughly speaking, an adaptive estimator of θ can be constructed as follows. Via the first $p+q+1$ sample autocovariances θ is estimated $\sqrt{n}$-consistently and this estimator is discretized. Given this discretized estimator the innovations can be recovered approximately from the observations; usually, one calls these "estimated" innovations the residuals. Via these residuals the score function for location can be estimated by the same methods as mentioned in Remark 8.2. Based on these estimators for θ and the score function an efficient, adaptive estimator of θ can be constructed based on sample splitting as in (8.20). This sample splitting and the discretization yield enough "independence" to be able to prove adaptivity (and hence efficiency) of this estimator of θ. For technical details we refer to Example 4.2, pp. 809–811, of Drost, Klaassen and Werker (1997). $\square$

Example 9.2 (GARCH (p, q)) Consider the time series model

$$Y_t = \mu h_t^{1/2} + \sigma h_t^{1/2} \epsilon_t \tag{9.20}$$

with

$$h_t = 1 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p} + \alpha_1 Y_{t-1}^2 + \cdots + \alpha_q Y_{t-q}^2. \tag{9.21}$$

Originally, this model was introduced with $\alpha_1 = \cdots = \alpha_q = 0$; since (9.21) exhibits an autoregression structure for the h_t's and since the innovations are multiplied by a random factor, it was called an AutoRegressive Conditional Heteroskedasticity model: ARCH. Therefore, the generalized model above is called GARCH. Note that it fits into the framework of (9.1) via

$$\mu_t(\theta) = \mu h_t^{1/2}, \sigma_t(\theta) = \sigma h_t^{1/2}, \quad \theta = \Big(\alpha_1, \dots \alpha_q, \beta_1, \dots, \beta_p, \mu, \sigma\Big)^T. \tag{9.22}$$

Under appropriate regularity conditions and via a construction as in Example 9.1, an estimator of $\nu(\theta) = (\alpha_1, \dots, \alpha_q,\ \beta_1, \dots, \beta_p)^T$ can be constructed which is efficient in the semiparametric model with the density g of the innovations ϵ_t unknown. This estimator is in fact adaptive in the presence of the nuisance parameters μ and σ. This means that, with μ and σ as nuisance parameters, it is performing asymptotically as well under g unknown as the best estimator of $\nu(\theta)$ under g known (but still with μ and σ unknown nuisance parameters). Details may be found in Drost and Klaassen (1997). $\square$

9.1 Exercises Chapter 9

Exercise 9.1 (Location-Scale) Let g be a density on $\mathbb{R}$ which is absolutely continuous with respect to Lebesgue measure with derivative g'. Define the Fisher information for location and for scale respectively, by

$$I_\ell(g) = \int \left(\frac{g'}{g}\right)^2 g, \;\; I_s(g) = \int \left(1 + x\frac{g'}{g}(x)\right)^2 g(x)dx. \tag{9.23}$$

Use the technique of pp. 211–212 and p. 214 of Hájek and Šidák (1967) to show that finiteness of $I_\ell(g)$ and $I_s(g)$ imply that the location-scale family $\mathcal{Q}$ from (9.2) is a regular parametric model. $\square$

Chapter 10

Banach Parameters

Up to now, we have considered only Euclidean (finite dimensional) parameters, i.e. parameterfunctions $\nu : \mathcal{P} \to \mathbb{R}^m$. Here we will discuss generalization to Banach (infinite dimensional) parameters, i.e. to parameterfunctions $\nu : \mathcal{P} \to \mathbb{B}$ with $\mathbb{B}$ a Banach space with norm $\|\cdot\|_{\mathbb{B}}$. The connection between Euclidean and Banach parameters is made via the dual space $\mathbb{B}^*$. Recall that $\mathbb{B}^*$ is the Banach space of real valued bounded linear functions $b^* : \mathbb{B} \to \mathbb{R}$. So, for any $b^* \in \mathbb{B}^*$ the parameterfunction $b^*\nu : \mathcal{P} \to \mathbb{R}$ is Euclidean. Not surprisingly the convolution theorem for estimators of Euclidean parameters thus gives rise to convolution theorems for Banach parameters. To formulate them accurately we need some definitions similar to those in Chapter 6.

Definition 10.1 (Pathwise (Weak-)Differentiability) The parameter $\nu : \mathcal{P} \to \mathbb{B}$ is *pathwise differentiable* on $\mathcal{P}$ at $P_0 \in \mathcal{P}$, if there exists a bounded linear operator $\dot{\nu}(P_0) = \dot{\nu} : \dot{\mathcal{P}} \to \mathbb{B}$ such that for all one-dimensional regular parametric submodels $\{P_\eta : |\eta| < \epsilon\}$, $\epsilon > 0$, with score function h at P_0

$$\|\nu(P_\eta) - \nu(P_0) - \eta\dot{\nu}(h)\|_{\mathbb{B}} = o(\eta), \text{ as } \eta \to 0. \tag{10.1}$$

The parameter ν is *pathwise weak-differentiable* if

$$b^*\nu(P_\eta) = b^*\nu(P_0) + \eta b^*\dot{\nu}(h) + o(\eta), \tag{10.2}$$

for all $b^* \in \mathbb{B}^*$.

The transpose of $\dot{\nu} : \dot{\mathcal{P}} \to \mathbb{B}$ is the map $\dot{\nu}^T : \mathbb{B}^* \to \dot{\mathcal{P}}$ defined by

$$\langle \dot{\nu}^T b^*, h\rangle_0 = b^*\dot{\nu}(h), \quad h \in \dot{\mathcal{P}}. \tag{10.3}$$

By the Riesz representation theorem (cf. (6.2)) there exists a unique $\dot{\nu}_{b^*} \in \dot{\mathcal{P}}$ with

$$b^*\dot{\nu}(h) = \langle \dot{\nu}_{b^*}, h\rangle_0, \tag{10.4}$$

and hence we have

$$\dot{\nu}_{b^*} = \dot{\nu}^T b^*, \quad b^* \in \mathbb{B}^*. \tag{10.5}$$

Definition 10.2 (Efficient Influence Operator and Function) If ν is pathwise (weak-) differentiable with derivative operator $\dot{\nu} : \dot{\mathcal{P}} \to \mathbb{B}$, then the *efficient influence operator* $\tilde{\ell}(P_0 \mid \nu, \mathcal{P}) = \tilde{\ell}_\nu : \mathbb{B}^* \to \dot{\mathcal{P}}$ is defined as its transpose

$$\tilde{\ell}_\nu(b^*) = \dot{\nu}^T b^* = \dot{\nu}_{b*}, \tag{10.6}$$

and the *inverse information covariance functional* $I^{-1}(P_0 \mid \nu, \mathcal{P}) = I_\nu^{-1} : \mathbb{B}^* \times \mathbb{B}^* \to \mathbb{R}$ by

$$I_\nu^{-1}(b_1^*, b_2^*) = \langle \tilde{\ell}_\nu(b_1^*),\, \tilde{\ell}_\nu(b_2^*) \rangle_0. \tag{10.7}$$

If there exists a map $\tilde{\ell} : \mathcal{X} \to \mathbb{B}$ such that for all $b^* \in \mathbb{B}^*$

$$b^* \tilde{\ell}(\cdot) = \dot{\nu}_{b^*}(\cdot) = \tilde{\ell}_\nu(b^*)(\cdot), \tag{10.8}$$

then we call $\tilde{\ell}$ the *efficient influence function*.

Definition 10.3 (Weak Regularity) The estimator sequence $\{T_n\}$ of $\nu(P)$ is said to be *weakly regular* at $P_0 \in \mathcal{P}$ if there exists a process $\{b^*\mathbb{Z} :\ b^* \in \mathbb{B}^*\}$ on $(\mathbb{R}^{\mathbb{B}^*}, \mathcal{B}^{\mathbb{B}^*})$ such that for all one-dimensional regular parametric submodels $\{P_\eta :\ |\eta| < \epsilon\},\ \epsilon > 0$, sequences $\eta_n = \mathcal{O}(n^{-1/2})$, and $b^* \in \mathbb{B}^*$

$$\sqrt{n}\Big(b^* T_n - b^* \nu(P_{\eta_n})\Big) \to b^*\mathbb{Z},\ \text{ as } n \to \infty. \tag{10.9}$$

With these definitions we may formulate a generalization of Theorem 7.2.

Theorem 10.1 (Convolution Theorem) *Let $\nu : \mathcal{P} \to \mathbb{B}$ be pathwise weak-differentiable at $P_0 \in \mathcal{P}$, let $\{T_n\}$ be weakly regular with limit $\mathbb{Z}$. If the tangent set $\dot{\mathcal{P}}^0$ is linear, then there exist processes $\{b^*\mathbb{Z}_0 : b^* \in \mathbb{B}^*\}$ and $\{b^*\Delta_0 : b^* \in \mathbb{B}^*\}$ on $\{\mathbb{R}^{\mathbb{B}^*}, \mathcal{B}^{\mathbb{B}^*}\}$ such that:*

$$b^*\mathbb{Z} \stackrel{\mathrm{D}}{=} b^*\mathbb{Z}_0 + b^*\Delta_0, \quad b^* \in \mathbb{B}^*, \tag{10.10}$$

$\mathbb{Z}_0$ and Δ_0 are independent, $\mathbb{Z}_0$ is Gaussian with mean 0 *and*

$$\mathrm{Cov}(b_1^*\mathbb{Z}_0,\, b_2^*\mathbb{Z}_0) = I_\nu^{-1}(b_1^*, b_2^*), \tag{10.11}$$

and for every $h \in \dot{\mathcal{P}}$ and $b^* \in \mathbb{B}^*$

$$\begin{pmatrix} \sqrt{n}\Big(b^* T_n - b^*\nu(P_0) - \frac{1}{n}\sum_{i=1}^n \tilde{\ell}_\nu(b^*)(X_i)\Big) \\ \frac{1}{\sqrt{n}}\sum_{i=1}^n h(X_i) \end{pmatrix} \xrightarrow{D}_{P_0} \begin{pmatrix} b^*\Delta_0 \\ W_0 \end{pmatrix} \tag{10.12}$$

with Δ_0 and W_0 independent. Furthermore, Δ_0 is degenerate at $0 \in \mathbb{R}^{\mathbb{B}^*}$ iff for every $b^* \in \mathbb{B}^*$, $b^* T_n$ is asymptotically linear with influence function $\tilde{\ell}_\nu(b^*)$, and $\{T_n\}$ is called weakly efficient at P_0 in this case.

This theorem is proved in Section 5.2 of BKRW (1993) together with a further generalization for regular estimators.

Example 10.1 (Estimation of Distribution Functions) Let $\mathcal{P}$ be the collection of all probability measures on $\mathbb{R}$ and $\nu(P)$ the distribution function F of P. With $\mathbb{B}$ the space of cadlag functions with the supnorm, $\nu : \mathcal{P} \to \mathbb{B}$ is pathwise differentiable at P_0 with $\dot{\nu} : \dot{\mathcal{P}} \to \mathbb{B}$ given by

$$\dot{\nu}(h)(t) = \int \Big(\mathbf{1}_{[x \leq t]} - F_0(t)\Big) h(x) dP_0(x), \quad t \in \mathbb{R}. \tag{10.13}$$

If $b^* : \mathcal{B} \to \mathbb{R}$ is such that $b^*(f) = f(t)$, then

$$\tilde{\ell}_\nu(b^*)(x) = \mathbf{1}_{[x \leq t]} - F_0(t), \tag{10.14}$$

and from the convolution theorem we conclude that the empirical distribution function

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{[X_i \leq t]}, \quad t \in \mathbb{R}, \tag{10.15}$$

is a weakly efficient estimator of $\nu(P) = F(\cdot)$, with efficient influence function $\tilde{\ell} : \mathcal{X} \to \mathbb{B}$ satifying $\tilde{\ell}(x) : t \mapsto \mathbf{1}_{[x \leq t]} - F_0(t)$. □

We close this chapter with a Cramér-Rao inequality for unbiased estimation of distribution functions.

Theorem 10.2 (Cramér-Rao Inequality for Distribution Functions) *Let $\mathcal{P}$ be the collection of all probability measures on $\mathbb{R}$ and $\nu(P) = F(\cdot)$ the distribution function of P. For every unbiased estimator $\hat{F}_n(\cdot)$ of $\nu(P) = F(\cdot)$ and every $P_0 \in \mathcal{P}$ with distribution function $F_0(\cdot)$ we have*

$$\mathrm{Var}_{P_0}\Big(\hat{F}_n(t)\Big) \geq \frac{1}{n} F_0(t) \left(1 - F_0(t)\right), \quad t \in \mathbb{R}. \tag{10.16}$$

Chapter 11

Cross Sectional Sampling

11.1 The Core Survival Model

- D is a survival time, for instance the duration of a disease of an individual from a homogeneous group of patients with a particular disease,
- W is a vector of covariates of dimension k with density h with respect to a measure ν. We assume that h is known.
- $\theta \in \Theta$ is an unknown k-vector of regression parameters.

We will assume that the random vector (D, W) satisfies a standard survival model, the Accelerated Failure Time model, and consider estimation of θ under length biased and cross sectional sampling.

11.2 The Accelerated Failure Time Model (AFT)

The semiparametric AFT model is given by

$$D = e^{-\theta^T W} V, \tag{11.1}$$

where V is a nondegenerate random variable on $[0, \infty)$ with unknown absolutely continuous distribution function G_0, with density g_0 and hazard function λ_0, and where V and W are independent.

Conditional survival, density and hazard function of D given $W = w$: for $t > 0$

- $\bar{G}_\theta(t|w) = 1 - G_\theta(t|w) = \bar{G}_0(e^{\theta^T w} t)$,
- $g_\theta(t|w) = e^{\theta^T w} g_0(e^{\theta^T w} t)$,
- $\lambda_\theta(t|w) = e^{\theta^T w} \lambda_0(e^{\theta^T w} t)$.

The function λ_0 serves as baseline hazard in this scale model.

11.3 Cross Sectional Sample

Observe the durations and their covariates at a *specific point in time.*

- Y is the total length of a sampled survival time,
- X is the time from onset until the time of sampling, the *present* say, of a sampled survival time.

What is the gain in being patient, i.e. using the full survival time ?

11.4 No Covariates

If f and F are the density and distribution function of the survival times D in the core model then under suitable assumptions

1. $f_Y(y) = \frac{yf(y)}{\mu}$,
2. $f_X(x) = \frac{\bar{F}(x)}{\mu}$, with $\bar{F}(x) = 1 - F(x)$,

where

$$\mu = \int_0^\infty uf(u)du.$$

Use $X = YU$ with $U \sim Un(0,1)$ and Y and U independent.

Formula (1) follows from the length bias in the sampling and (2) from length bias and multiplicative censoring !

11.5 Survival Analysis Setting

We observe n i.i.d. realizations of (Y, Z) and (X, Z) of survival times (in total or from onset to present) and the sampled covariates.

Under the AFT model assumptions for the core model it turns out that

- Given the covariate Z the distributions of both Y and X are scale families,
- The distribution of Z, the observed covariate, only depends on θ. It does not depend on g_0 !

For $x > 0, y > 0$ and $z \in \mathbb{R}^k$ we have

1. Total survival time:

$$\begin{aligned} f_{Y,Z}(y,z) &= \frac{e^{\theta^T z} y g_0(e^{\theta^T z} y) h(z)}{\mathrm{E}_{g_0} V \mathrm{E}_h e^{-\theta^T W}}, \\ f_Z(z) &= \frac{e^{-\theta^T z} h(z)}{\mathrm{E}_h e^{-\theta^T W}} \\ f_{Y|Z}(y|z) &= \frac{e^{\theta^T z} e^{\theta^T z} y g_0(e^{\theta^T z} y)}{\mathrm{E}_{g_0} V}, \end{aligned}$$

2. From onset to present:

$$\begin{aligned}
f_{X,Z}(x,z) &= \frac{\bar{G}_0(e^{\theta^T z}x)h(z)}{\mathrm{E}_{g_0,h}T}, \\
f_Z(z) &= \frac{e^{-\theta^T z}h(z)}{\mathrm{E}_h e^{-\theta^T W}}, \\
f_{X|Z}(x|z) &= \frac{e^{\theta^T z}\bar{G}_0(e^{\theta^T z}x)}{\mathrm{E}_{g_0}V}.
\end{aligned}$$

11.6 Efficient Estimation

Construction of efficient estimators of the Euclidian parameter in semi parametric models:

- Find a $\sqrt{n}$ consistent estimator of θ. Here we can use the observed covariates for that purpose,
- Compute the efficient influence function for θ with g_0 as nuisance parameter and construct a suitably consistent estimator,
- Use sample splitting to construct an efficient estimator.

This can be done under both sampling schemes. See Van Es, Klaassen and Oudshoorn (2000) and Klaassen, Mokveld and Van Es (2003) for the "impatient" sampling scheme.

11.7 Information Bounds

The information bounds under the two sampling schemes are:

- $\Sigma_Z(I_s(e^{\theta^T Z}Y)+1)$,
- $\Sigma_Z(I_s(e^{\theta^T Z}X)+1)$,

where for a random variable U with density f

$$I_s(U) = \int \Big(1 + u\,\frac{f_U'(u)}{f_U(u)}\Big)^2 f_U(u)du.$$

Here

$$I_s(e^{\theta^T Z}Y) = \int \Big(2 + u\,\frac{g_0'(u)}{g_0(u)}\Big)^2 \frac{ug_0(u)}{\mu}du$$

and

$$I_s(e^{\theta^T Z}X) = \int \Big(1 - u\,\frac{g_0'(u)}{\bar{G}_0(u)}\Big)^2 \frac{\bar{G}_0(u)}{\mu}du$$

11.8 An example: gamma densities

Consider the case where g_0 is a Gamma(m) density, i.e.

$$g_0(t) = \frac{t^{m-1}e^{-t}}{(m-1)!}, \quad t \geq 0, \ m = 1, 2, \dots .$$

Then $I_s(e^{\theta^T Z}Y) = m + 1$.

By numerical computation of $I_s(e^{\theta^T Z}X)$ we get the following table.

m	$I_s(e^{\theta^T Z}X) + 1$	$I_s(e^{\theta^T Z}Y) + 1$
1	2	3
2	2.30	4
3	2.55	5
4	2.77	6
5	2.96	7
10	3.75	12

Appendix A

Additions

A.1 Measure Theory and Probability

A.1.1 Vitali's theorem

Pointwise convergence together with convergence of the norms implies convergence *in* norm; more precisely:

Theorem A.1 (Vitali) *If $f_n \to f$, μ-almost everywhere, and*

$$\limsup_{n\to\infty} \int |f_n|^p d\mu \leq \int |f|^p d\mu, \quad 0 < p < \infty,$$

then we have

$$\int |f_n - f|^p d\mu \to 0. \tag{A.1}$$

Proof Note that we have $(a+b)^p \leq 2^p(a^p + b^p)$ for all $a, b \geq 0$. So

$$|f_n - f|^p \leq (|f_n| + |f|)^p \leq 2^p(|f_n|^p + |f|^p). \tag{A.2}$$

The convergence of f_n to f implies

$$2^{p+1}|f|^p = \lim_{n\to\infty} \Big(2^p(|f_n|^p + |f|^p) - |f_n - f|^p \Big),$$

μ-almost everywhere, and by Fatou's lemma

$$\begin{aligned} & \int 2^{p+1}|f|^p d\mu \\ = & \int \lim_{n\to\infty} \Big(2^p(|f_n|^p + |f|^p) - |f_n - f|^p \Big) d\mu \\ \leq & \liminf_{n\to\infty} \int \Big(2^p(|f_n|^p + |f|^p) - |f_n - f|^p \Big) d\mu \\ \leq & \limsup_{n\to\infty} 2^p \int |f_n|^p d\mu + 2^p \int |f|^p d\mu - \limsup_{n\to\infty} \int |f_n - f|^p d\mu \\ \leq & \int 2^{p+1}|f|^p d\mu - \limsup_{n\to\infty} \int |f_n - f|^p d\mu. \end{aligned} \tag{A.3}$$

This implies (A.1). Note that by (A.2) the second integrand in (A.3) is positive which is essential in an application of Fatou's lemma. □

A.1.2 Total variation distance

Consider a set of measures $\mathcal{P}$ on a separable metric space $\mathcal{X}$ with Borel sets $\mathcal{B}$. We define the *total variation distance* between two measures P and Q as

$$d(P,Q) = \sup_{A\in\mathcal{B}} |P(A) - Q(A)|. \tag{A.4}$$

If both P and Q are dominated by μ then $d(P,Q)$ is half the $\mathcal{L}_1(\mu)$ distance between the densities

$$d(P,Q) = \int_{A_0} |p-q|d\mu = \frac{1}{2}\int_{\mathcal{X}} |p-q|d\mu, \tag{A.5}$$

where $A_0 = \{x \in \mathcal{X} : p(x) > q(x)\}$.

Examples of sets of measures K which are compact in this metric are for instance sets existing of finitely many elements, or sets $\{P, P_1, P_2, \ldots\}$ where $\{P_n, n = 1, 2, \ldots\}$ is a sequence of measures converging to P in total variation. If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model then for $\theta_n \to \theta$ and $t_n \to t$ the sequence $\{P_{\theta_n + \frac{t_n}{\sqrt{n}}}\}$ converges in total variation to P_θ. This follows from the fact that the Fréchet differentiability implies $\mathcal{L}_2$ convergence of the densities, and hence $\mathcal{L}_1$ convergence too. Hence $K = \{P_{\theta_n + \frac{t_n}{\sqrt{n}}}, n = 1, 2, \ldots\}$ is also a compact set.

A.1.3 Tightness and Prohorov's theorem

We need the concept of *tightness* of a sequence of distributions.

Definition A.1 (Tigthness) A sequence of probability measures $\{P_n\}_{n=1}^\infty$ on a measurable space $(\mathcal{X}, \mathcal{B})$ is called tight if for every $\epsilon > 0$ there exists a compact set K_ϵ such that $P_n(K_\epsilon) > 1 - \epsilon$, for all n.

The next theorem establishes a correspondence between tightness and weak convergence of subsequences. For a proof we refer to Billingsley (1968) page 37.

Theorem A.2 (Prohorov) *Let $\{P_n\}_{n=1}^\infty$ denote a sequence of probability measures on a complete separable measurable space $(\mathcal{X}, \mathcal{B})$. Then $\{P_n\}_{n=1}^\infty$ is tight if and only if for each subsequence there exists a further subsequence which converges weakly to a probability measure P. This limit measure P may depend on the specific subsequences.*

Weak convergence of the whole sequence can also be characterized in a similar manner. The proof is straightforward.

Theorem A.3 *Let $\{P_n\}_{n=1}^\infty$ denote a sequence of probability measures on a measurable space $(\mathcal{X}, \mathcal{B})$. Then $\{P_n\}_{n=1}^\infty$ converges weakly to a probabilty measure P if and only if for each subsequence there exists a further subsequence which converges weakly to P.*

Appendix B

Notes

B.1 Notes Chapter 2

Note 1

Proof of the "if" part of Theorem 2.2.
Under (2.16) equality in (2.10) can be violated only if $G^{-1}(v) - G^{-1}(u) > \int_u^v \frac{1}{g(G^{-1}(s))} ds$. This can happen only if G^{-1} has jumps, that is, if g vanishes on $[y_0, y_1]$, say, with $y_0 < y_1$ and with g positive at some points $y < y_0$ and some $y > y_1$. However, by (2.4)

$$\begin{aligned} g(y) &= \mathrm{E}\, S I_{[T-\vartheta>y]} \\ &= \mathrm{E}\, S I_{[G^{-1}(H(S))>y]} \\ &= \mathrm{E}\,(S|G^{-1}(H(S)) > y) P(G^{-1}(H(S)) > y) \end{aligned}$$

holds and the first factor at the right hand side is nonnegative and nondecreasing in y. Now, assume $g(y_2) = 0$ for some y_2. If $P(G^{-1}(H(S)) > y_2) = 0$ then $P(G^{-1}(H(S)) > y) = 0$ and hence $g(y)$ vanishes for all $y \geq y_2$. If $\mathrm{E}\,(S|G^{-1}(H(S)) > y_2) = 0$, then $\mathrm{E}\,(S|G^{-1}(H(S)) > y)$ and hence $g(y)$ vanish for all $y \leq y_2$. It follows that g cannot vanish on an interval strictly within its support. Consequently, G^{-1} cannot have jumps. □

B.2 Notes Chapter 3

Note 2

Recall that the directional derivative of a function $g : \mathbb{R}^k \to \mathbb{R}$ on the line segment from $a \in \mathbb{R}^k$ to $b \in \mathbb{R}^k$ is given by

$$\frac{d}{d\lambda} g(a + \lambda(b-a)) = (b-a)^T g'(a + \lambda(b-a)), \quad \lambda \in [0,1],$$

where g' is the gradient vector of partial derivatives of g.
In the context of the proof of Proposition 3.1 we get, since $\dot{s} = \frac{1}{2}\dot{\ell} s$,

$$\begin{aligned} & s(x;\theta) - s(x;\theta_0) \\ = & \int_0^1 \frac{d}{d\lambda} s(\theta_0 + \lambda(\theta - \theta_0)) d\lambda \\ = & \int_0^1 (\theta - \theta_0)^T \dot{s}(\theta_0 + \lambda(\theta - \theta_0)) d\lambda \\ = & \int_0^1 \frac{1}{2} (\theta - \theta_0)^T \dot{\ell} s(\theta_0 + \lambda(\theta - \theta_0)) d\lambda. \end{aligned}$$

Note 3

Writing

$$g(\lambda) = \frac{1}{2} (\theta - \theta_0)^T \dot{\ell} s(\theta_0 + \lambda(\theta - \theta_0))$$

use the inequality (Jensen)

$$\Big(\int_0^1 g(\lambda) d\lambda \Big)^2 \leq \int_0^1 g(\lambda)^2 d\lambda.$$

Note 4

Let a and b be k-vectors. The following equality will be used repeatedly

$$(a^T b)^2 = a^T b (a^T b) = a^T (a^T b)^T = a^T b (b^T a) = a^T (b b^T) a. \tag{B.1}$$

By this equality we obtain

$$\begin{aligned} & \int (t^T \dot{\ell} s(x;\theta))^2 d\mu(x) = \int (t^T \dot{\ell}(x;\theta))^2 s(x;\theta)^2 d\mu(x) \\ = & \int (t^T \dot{\ell}\dot{\ell}^T (x;\theta) t) p(x;\theta) d\mu(x) = t^T \int \dot{\ell}\dot{\ell}^T (x;\theta) p(x;\theta) d\mu(x)\ t \\ = & t^T I(\theta) t. \end{aligned}$$

Now substitute $\theta - \theta_0$ for t and $\theta_0 + \lambda(\theta - \theta_0)$ for θ.

Note 5

Suppose that (3.7) is not true. Then there exists a sequence $\{\theta_n\}$ such that

$$\text{(B.2)} \qquad \frac{1}{|\theta_n - \theta_0|^2} \int\limits_{p(x;\theta_0)>0} |s(x;\theta_n) - s(x;\theta_0) - \frac{1}{2}(\theta_n - \theta_0)^T \dot{\ell} s(x;\theta_0)|^2 d\mu(x)$$

is at least at distance ϵ from 0 for some $\epsilon > 0$. Now consider the sequence $\{(\theta_n - \theta_0)/|\theta_n - \theta_0|\}$. This is a sequence of points on the unit ball of $\mathbb{R}^k$. So there is a convergent subsequence for which (B.2) does not converge. Hence it suffices to prove (3.7) for all sequences $\{\theta_n\}$ for which $(\theta_n - \theta_0)/|\theta_n - \theta_0|$ converges.

Now consider an arbitrary sequence $\{\theta_n\}$ with $(\theta_n - \theta_0)/|\theta_n - \theta_0| \to c \in \mathbb{R}^k$. Then (B.2) equals

$$\text{(B.3)} \qquad \int\limits_{p(x;\theta_0)>0} |\frac{s(x;\theta_n) - s(x;\theta_0)}{|\theta_n - \theta_0|} - \frac{1}{2} c^T \dot{\ell} s(x;\theta_0)|^2 d\mu(x) + o(1)$$

From (3.5) we get

$$\frac{s(x;\theta_n) - s(x;\theta_0)}{|\theta_n - \theta_0|} \to \frac{1}{2} c\dot{\ell} s(x;\theta_0)$$

and by (3.6) we have (use $c^T \dot{\ell}\dot{\ell}^T c = (c^T \dot{\ell})^2$)

$$\limsup_{n\to\infty} \int\limits_{p(x;\theta_0)>0} \left|\frac{s(x;\theta_n) - s(x;\theta_0)}{|\theta_n - \theta_0|}\right|^2 d\mu(x) \leq \int\limits_{p(x;\theta_0)>0} \left|\frac{1}{2} c^T \dot{\ell} s(x;\theta_0)\right|^2 d\mu(x).$$

By Vitali's theorem A.1 we can now conclude that (B.2) converges to zero. This proves (3.7).

Note 6

Lemma B.1 *If $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model then*

$$\text{(B.4)} \qquad \int \dot{\ell}(x;\theta) dP_\theta(x) = 0.$$

Proof By the Fréchet differentiability we have

$$\|s(\theta + h) - s(\theta)\| \leq \|s(\theta + h) - s(\theta) - \frac{1}{2} h^T \dot{\ell}(\theta) s(\theta)\| + \|\frac{1}{2} h^T \dot{\ell} s\| = \mathcal{O}(|h|).$$

Now note that, working in $\mathcal{L}_2(\mu)$, we have

$$< s(\theta), s(\theta) >= \int s(\theta) s(\theta) d\mu = \int p(\theta) d\mu = 1.$$

This implies

$$\begin{aligned}
&| < h^T\dot{\ell}(\theta)s(\theta), s(\theta) > | \\
&= \ | < s(\theta+h), s(\theta+h) > - < s(\theta), s(\theta) > \\
&\qquad - < \frac{1}{2}h^T\dot{\ell}(\theta)s(\theta), 2s(\theta) > | \\
&= \ | < s(\theta+h) - s(\theta) - \frac{1}{2}h^T\dot{\ell}(\theta)s(\theta), 2s(\theta) > \\
&\qquad + < s(\theta+h) - s(\theta), s(\theta+h) - s(\theta) > | \\
&\leq \ | < s(\theta+h) - s(\theta) - \frac{1}{2}h^T\dot{\ell}(\theta)s(\theta), 2s(\theta) > | \\
&\qquad +| < s(\theta+h) - s(\theta), s(\theta+h) - s(\theta) > | \\
&\leq \ 2\|s(\theta+h) - s(\theta) - \frac{1}{2}h^T\dot{\ell}(\theta)s(\theta)\| \\
&\qquad +\|s(\theta+h) - s(\theta)\|^2 \\
&= \ \mathcal{O}(|h|), \quad \text{as } h \to 0.
\end{aligned}$$

But, by taking $h = (0, \ldots, 0, h_i, 0, \ldots, 0)^T$ with $h_i \to 0$, this implies

$$\int \dot{\ell}_i(x;\theta)dP_\theta(x) =< \dot{\ell}(\theta)s(\theta), s(\theta) >= 0,$$

for $i = 1, \ldots, k$ □

Note 7

Lemma B.2 *If the parametrization $\theta \longrightarrow P_\theta$ is regular and T_n is defined by*

$$T_n = T_n(X,t) = 2\left(\frac{s(X;\theta + \frac{t}{\sqrt{n}}) - s(X;\theta)}{s(X;\theta)}\right)$$

then we have uniformly for $\theta \in K$ and $|t| \leq M$

1. $\mathrm{E}_\theta T_n = -\frac{1}{4n}t^T I(\theta)t + o(\frac{1}{n})$
2. $\mathrm{E}_\theta T_n^2 = \frac{1}{n}t^T I(\theta)t + o(\frac{1}{n})$
3. $\mathrm{E}_\theta \left(T_n - \frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right)^2 = o(\frac{1}{n})$
4. $P_\theta\left(\{|T_n| > \varepsilon\}\right) = o(\frac{1}{n})$ voor alle $\varepsilon > 0$.

Proof

We give the proof for θ and t fixed. The generalization to convergent sequences θ_n and t_n, to prove uniformity, is straightforward.
Write $A(\theta) = \{x : \ s(x;\theta) = 0\}$. We need the following bound

$$\int_{A(\theta)} s^2(x;\theta+h)d\mu(x) = o(|h|^2), \quad \text{if } h \to 0. \tag{B.5}$$

This follows from the regularity of the parametrization, since we have

$$\begin{aligned}
&\int_{A(\theta)} s^2(x;\theta+h)d\mu(x)\\
&\leq \int_{A(\theta)} (s(x;\theta+h)+s(x;\theta-h))^2 d\mu(x)\\
&= \int_{A(\theta)} (s(x;\theta+h) - \frac{1}{2}h^T\dot{\ell}s(x;\theta) + \frac{1}{2}h^T\dot{\ell}s(x;\theta) + s(x;\theta-h))^2 d\mu(x)\\
&\leq 2\int_{A(\theta)} (s(x;\theta+h) - \frac{1}{2}h^T\dot{\ell}s(x;\theta))^2 d\mu(x)\\
&\quad + 2\int_{A(\theta)} (s(x;\theta-h) + \frac{1}{2}h^T\dot{\ell}s(x;\theta))^2 d\mu(x)\\
&= o(|h|^2) \quad \text{as } h \to 0.
\end{aligned}$$

Here we have used the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

Next we prove $\mathrm{E}_\theta T_n^2 = -4\mathrm{E}_\theta T_n + o(1/n)$. Define the set $B(\theta) = A(\theta)^c = \{x : \ s(x;\theta) > 0\}$, then, using $\int s(x;\theta)^2 d\mu = \int s(x;\theta+\frac{t}{\sqrt{n}})^2 d\mu = 1$, we get

$$\begin{aligned}
\mathrm{E}_\theta T_n^2 &= 4\int_{\mathbf{X}} \left(\frac{s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)}{s(x;\theta)}\right)^2 dP_\theta(x)\\
&= 4\int_{B(\theta)} \left(\frac{s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)}{s(x;\theta)}\right)^2 dP_\theta(x)\\
&= 4\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)\right)^2 d\mu(x)\\
&= 4\int_{B(\theta)} \left(s^2(x;\theta+\frac{t}{\sqrt{n}}) - 2s(x;\theta+\frac{t}{\sqrt{n}})s(x;\theta) + s^2(x;\theta)\right) d\mu(x)\\
&= -8\int_{B(\theta)} s(x;\theta)\left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)\right) d\mu(x)\\
&\quad -4\int_{A(\theta)} s^2(x;\theta+\frac{t}{\sqrt{n}})d\mu(x)\\
&\overset{(B.5)}{=} -8\int_{B(\theta)} \left(\frac{s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)}{s(x;\theta)}\right) dP_\theta(x) + o(\frac{1}{n})\\
&= -4\mathrm{E}_\theta T_n + o(\frac{1}{n}).
\end{aligned}$$

Next we prove 2., which in its turn proves 1. We have

$$
\begin{aligned}
\mathrm{E}_\theta T_n^2 = 4\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta)\right)^2 d\mu(x) \\
&= 4\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta) + \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\right)^2 d\mu(x) \\
&= 4\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\right)^2 d\mu(x) \\
&\quad + \frac{1}{n}\int_{B(\theta)} (t^T\dot{\ell}s(x;\theta))^2 d\mu(x) \\
&\quad + \int_{B(\theta)} \frac{4t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t}{\sqrt{n}}\dot{\ell}s(x;\theta)\right) d\mu(x) \\
&\overset{(B.1)}{=} \frac{1}{n}t^T I(\theta)t + o(\frac{1}{n}),
\end{aligned}
$$

since by the Cauchy-Schwarz inequality

$$
\begin{aligned}
&\left(\int_{B(\theta)} t^T\dot{\ell}s(x;\theta)\left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\right) d\mu(x)\right)^2 \\
&\le \left(\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\right)^2 d\mu(x)\right) \\
&\qquad \times \left(\int_{B(\theta)} (t^T\dot{\ell}s(x;\theta))^2 d\mu(x)\right) \\
&= o(\frac{1}{n}).
\end{aligned}
$$

Statement 3. follows from

$$
\begin{aligned}
&\mathrm{E}_\theta\left(T_n - \frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right)^2 \\
&= 4\int_{B(\theta)} \left(s(x;\theta+\frac{t}{\sqrt{n}}) - s(x;\theta) - \frac{1}{2}\frac{t^T}{\sqrt{n}}\dot{\ell}s(x;\theta)\right)^2 d\mu(x) \\
&= o(\frac{1}{n}).
\end{aligned}
$$

Finally we prove 4. We have

$$
\begin{aligned}
&P_\theta\left(\{|T_n| \ge \varepsilon\}\right) \\
&\le P_\theta\left(\left\{\left|T_n - \frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right| \ge \frac{\varepsilon}{2}\right\}\right) + P_\theta\left(\left\{\left|\frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right| \ge \frac{\varepsilon}{2}\right\}\right) \\
&\le \frac{4}{\varepsilon^2}\mathrm{E}_\theta\left(T_n - \frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right)^2 + \frac{4}{\varepsilon^2}\mathrm{E}_\theta\left(\frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)\right)^2 1_{\{|\frac{t^T}{\sqrt{n}}\dot{\ell}(\theta)| \ge \frac{\varepsilon}{2}\}}
\end{aligned}
$$

$$
\begin{aligned}
&= o(\frac{1}{n}) + \frac{4}{\varepsilon^2 n}\mathrm{E}_\theta(t^T\dot{\ell}(\theta))^2 1_{\{|t^T\dot{\ell}(\theta)|\geq\frac{\sqrt{n}\varepsilon}{2}\}} \\
&= o(\frac{1}{n}).
\end{aligned}
$$

The last equality holds since $t^T\dot{\ell}(\theta)$ is P_θ-a.e. finite. So $(t^T\dot{\ell}(\theta))^2 1_{\{|t^T\dot{\ell}(\theta)|\geq\frac{\sqrt{n}\varepsilon}{2}\}}$ vanishes P_θ almost surely and by the dominated convergence theorem we get

$$
\mathrm{E}_\theta\dot{\ell}^2(\theta)1_{\{|t^T\dot{\ell}(\theta)|\geq\frac{\sqrt{n}\varepsilon}{2}\}} \longrightarrow 0.
$$

This completes the proof. □

We now prove the claims in the proof of Theorem 3.1

1. $P_\theta\left(\{\max_{1\leq k\leq n}|T_{nk}|>\varepsilon\}\right)=o(1)$
2. $\sum_{k=1}^n T_{nk} - \left(t^T S_n(\theta) - \frac{1}{4}t^T I(\theta)t\right) = o_{P_\theta}(1)$
3. $\sum_{k=1}^n T_{nk}^2 - t^T I(\theta)t = o_{P_\theta}(1)$
4. $\sum_{k=1}^n \alpha_{nk}|T_{nk}|^3 = o_{P_\theta}(1)$

We have

$$
P_\theta\left(\{\max_{1\leq k\leq n}|T_{nk}|>\varepsilon\}\right) \leq \sum_{k=1}^n P_\theta\left(\{|T_{nk}|>\varepsilon\}\right) = nP_\theta\left(\{|T_{n1}|>\varepsilon\}\right) = o(1)
$$

by the previous lemma. This proves 1.
Next we prove 2. We have

$$
\begin{aligned}
&\mathrm{E}_\theta\left(\sum_{k=1}^n T_{nk} - \left(t^T S_n(\theta) - \frac{1}{4}t^T I(\theta)t\right)\right)^2 \\
&= \mathrm{E}_\theta\left(\sum_{k=1}^n\left(T_{nk} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_k;\theta) + \frac{1}{4n}t^T I(\theta)t\right)\right)^2 \\
&= \sum_{k=1}^n\sum_{j=1}^n \mathrm{E}_\theta\left(T_{nk} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_k;\theta) + \frac{1}{4n}t^T I(\theta)t\right) \\
&\qquad\times\left(T_{nj} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_j;\theta) + \frac{1}{4n}t^T I(\theta)t\right) \\
&\overset{(a)}{=} n\mathrm{E}_\theta\left(T_{n1} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta) + \frac{1}{4n}t^T I(\theta)t\right)^2 \\
&\quad+n(n-1)\mathrm{E}_\theta\left(T_{n1} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta) + \frac{1}{4n}t^T I(\theta)t\right) \\
&\qquad\times\mathrm{E}_\theta\left(T_{n2} - \frac{t^T}{\sqrt{n}}\dot{\ell}(X_2;\theta) + \frac{1}{4n}t^T I(\theta)t\right)
\end{aligned}
$$

$$\begin{aligned}
&\stackrel{(b)}{=} \quad n\mathrm{E}_\theta\left(T_{n1}-\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right)^2+\frac{1}{2}t^TI(\theta)t\mathrm{E}_\theta\left(T_{n1}-\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right)\\
&\qquad +\frac{1}{16n}(t^TI(\theta)t)^2+o(1)\\
&\stackrel{(c)}{=} \quad o(1).
\end{aligned}$$

(a) because the terms are i.i.d.
(b) and (c) because of the previous lemma and lemma B.1

By the Chebyshev inequality we have

$$P_\theta\left(\left|\sum_{k=1}^n T_{nk}-t^TS_n(\theta)+\frac{1}{4}t^TI(\theta)t\right|\geq\varepsilon\right)\longrightarrow 0 \qquad \forall\varepsilon>0.$$

Next we prove 3. Note that

$$\begin{aligned}
\sum_{k=1}^n T_{nk}^2-t^TI(\theta)t &= \sum_{k=1}^n T_{nk}^2-\frac{1}{n}\sum_{k=1}^n t^T\dot{\ell}\dot{\ell}^T(X_k;\theta)t\\
&+ \quad \frac{1}{n}\sum_{k=1}^n t^T\dot{\ell}\dot{\ell}^T(X_k;\theta)t-t^TI(\theta)t.
\end{aligned}$$

By the weak law of large numbers

$$\frac{1}{n}\sum_{k=1}^n t^T\dot{\ell}\dot{\ell}^T(X_k;\theta)t\xrightarrow{P_\theta}t^TI(\theta)t$$

and by the previous lemma and the Cauchy-Schwarz inequality

$$\begin{aligned}
&\mathrm{E}_\theta\left|\sum_{k=1}^n T_{nk}^2-\frac{1}{n}\sum_{k=1}^n t^T\dot{\ell}\dot{\ell}^T(X_k;\theta)t\right|\\
&\leq \quad n\mathrm{E}_\theta\left|T_{n1}^2-\left(\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right)^2\right|\\
&\leq \quad n\mathrm{E}_\theta\left(T_{n1}-\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right)^2+2n\mathrm{E}_\theta\left|\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right|\left|T_{n1}-\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\right|\\
&\leq \quad o(1)+2n\left(\mathrm{E}_\theta(\frac{t^T}{n}\dot{\ell}(X_1;\theta))^2\right)^{\frac{1}{2}}\left(\mathrm{E}_\theta\Big(T_{n1}-\frac{t^T}{\sqrt{n}}\dot{\ell}(X_1;\theta)\Big)^2\right)^{\frac{1}{2}}\\
&= \quad o(1).
\end{aligned}$$

Here we have used the inequality $|a^2-b^2|\leq(a-b)^2+2|b(a-b)|$.
So for all $\varepsilon>0$

$$P_\theta\left(\left|\sum_{k=1}^n T_{nk}^2-t^TI(\theta)t\right|\geq\varepsilon\right)$$

$$
\begin{aligned}
&\leq P_\theta\left(\left|\sum_{k=1}^n T_{nk}^2 - \frac{1}{n}\sum_{k=1}^n (t^T\dot{\ell}(X_k;\theta))^2\right| \geq \frac{\varepsilon}{2}\right) \\
&\qquad + P_\theta\left(\left|\frac{1}{n}\sum_{k=1}^n (t^T\dot{\ell}(X_k;\theta))^2 - t^T I(\theta) t\right| \geq \frac{\varepsilon}{2}\right) \\
&\leq \frac{2}{\varepsilon}\,\mathrm{E}_\theta\left|\sum_{k=1}^n T_{nk}^2 - \frac{1}{n}\sum_{k=1}^n (t^T\dot{\ell}(X_k;\theta))^2\right| \\
&\qquad + P_\theta\left(\left|\frac{1}{n}\sum_{k=1}^n t^T\dot{\ell}\dot{\ell}^T(X_k;\theta)t - t^T I(\theta) t\right| \geq \frac{\varepsilon}{2}\right) \\
&= o(1)
\end{aligned}
$$

It remains to prove 4. Since $|\alpha_{nk}| \leq 1$ we have

$$
\begin{aligned}
&P_\theta\left(\left|\sum_{k=1}^n \alpha_{nk}|T_{nk}|^3\right| \geq \varepsilon\right) \\
&\leq P_\theta\left(\sum_{k=1}^n |T_{nk}|^3 \geq \varepsilon\right) \\
&\leq P_\theta\left(\max_{1\leq k\leq n}|T_{nk}|\sum_{k=1}^n |T_{nk}|^2 \geq \varepsilon\right) \\
&= P_\theta\left(\max_{1\leq k\leq n}|T_{nk}|\sum_{k=1}^n |T_{nk}|^2 \geq \varepsilon, \sum_{k=1}^n |T_{nk}|^2 \leq 1 + t^T I(\theta) t\right) \\
&\qquad + P_\theta\left(\max_{1\leq k\leq n}|T_{nk}|\sum_{k=1}^n |T_{nk}|^2 \geq \varepsilon, \sum_{k=1}^n |T_{nk}|^2 > 1 + t^T I(\theta) t\right) \\
&\leq P_\theta\left(\max_{1\leq k\leq n}|T_{nk}| \geq \frac{\varepsilon}{1 + t^T I(\theta) t}\right) + P_\theta\left(\sum_{k=1}^n |T_{nk}|^2 > 1 + t^T I(\theta) t\right) \\
&= o(1),
\end{aligned}
$$

by statements 1. and 3. of this proof. This proves 4. □

B.3 Notes Chapter 4

Note 8

In a regular parametric model, (3.1) yields

$$
\begin{aligned}
&\int |s^2(\theta) - s^2(\theta_0) - 2(\theta-\theta_0)^T \dot{\ell}(\theta_0) s^2(\theta_0)| d\mu \\
&\quad \leq \int |s(\theta) - s(\theta_0) - (\theta-\theta_0)^T \dot{\ell}(\theta_0) s(\theta_0)||s(\theta)+s(\theta_0)| d\mu \\
&\qquad\quad + \int |(\theta-\theta_0)^T \dot{\ell}(\theta_0) s(\theta_0)||s(\theta)-s(\theta_0)| d\mu \\
&\quad \leq \|s(\theta) - s(\theta_0) - (\theta-\theta_0)^T \dot{\ell}(\theta_0) s(\theta_0)\|_\mu \ \{\|s(\theta)\|_\mu + \|s(\theta_0)\|_\mu\} \\
&\qquad\quad + |\theta-\theta_0| \ \||\dot{\ell}(\theta_0)|s(\theta_0)\|_\mu \|s(\theta)-s(\theta_0)\|_\mu = o(|\theta-\theta_0|)
\end{aligned}
$$

with $\|\cdot|\|_\mu$ denoting the norm in $\mathcal{L}_2(\mu)$. It follows that for n and σ fixed

$$
\begin{aligned}
&\int_{\mathbb{R}^k}\int_{\mathcal{X}^n} \Big| \Big\{ \prod_{i=1}^n p(x_i;\theta+\delta b) \Big\} \Big\{ 1 + b^T \dot{\delta} \Big\} \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta+\delta b-\theta_0)}{\sigma} \Big) \\
&\qquad - \Big\{ \prod_{i=1}^n p(x_i;\theta) \Big\} \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} \Big) \\
&\qquad - \Big\{ \delta b^T \sum_{i=1}^n \dot{\ell}(x_i;\theta) + \delta b^T \frac{\sqrt{n}}{\sigma} \frac{\dot{w}_0}{w_0} \Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} \Big) + b^T \dot{\delta} \Big\} \\
&\qquad \times \Big\{ \prod_{i=1}^n p(x_i;\theta) \Big\} \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} \Big) \Big| \ d\mu(x_1)\cdots d\mu(x_n) d\theta \\
&\leq \sum_{i=1}^n \int_{\mathbb{R}^k}\int_{\mathcal{X}^n} \Big| p(x_i;\theta+\delta b) - p(x_i;\theta) - \delta b^T \dot{\ell}(x_i;\theta) p(x_i;\theta) \Big| \\
&\qquad \times \Big\{ \prod_{j=1}^{i-1} p(x_j;\theta) \prod_{j=i+1}^n p(x_j;\theta+\delta b) \Big\} \Big\{ 1 + b^T \dot{\delta} \Big\} \\
&\qquad \times \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta+\delta b-\theta_0)}{\sigma} \Big) \ d\mu(x_1)\cdots d\mu(x_n) d\theta \\
&\quad + \int_{\mathbb{R}^k}\int_{\mathcal{X}^n} \Big\{ \prod_{i=1}^n p(x_i;\theta) \Big\} \Big| \Big\{ 1 + b^T \dot{\delta} \Big\} \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta+\delta b-\theta_0)}{\sigma} \Big) \\
&\qquad - \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} \Big) - \Big(\frac{\sqrt{n}}{\sigma} \Big)^2 \delta b^T \dot{w}_0\Big(\sqrt{n}(\theta-\theta_0) \Big) \\
&\qquad - b^T \dot{\delta} \frac{\sqrt{n}}{\sigma} w_0\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma} \Big) \Big| \ d\mu(x_1)\cdots d\mu(x_n) d\theta \\
&\quad + \sum_{i=1}^n \int_{\mathbb{R}^k}\int_{\mathcal{X}^n} \Big| \delta b^T \dot{\ell}(x_i;\theta) p(x_i;\theta) \Big| \Big\{ \prod_{j=1}^{i-1} p(x_j;\theta) \Big\}
\end{aligned}
$$

$$
\begin{aligned}
&\times\Big|\Big\{\prod_{j=i+1}^{n} p(x_j;\theta+\delta b)\Big\}\Big\{1+b^T\dot{\delta}\Big\}\frac{\sqrt{n}}{\sigma}\,w_0\Big(\frac{\sqrt{n}(\theta+\delta b-\theta_0)}{\sigma}\Big)\\
&-\Big\{\prod_{j=i+1}^{n} p(x_j;\theta)\Big\}\frac{\sqrt{n}}{\sigma}\,w_0\Big(\frac{\sqrt{n}(\theta-\theta_0)}{\sigma}\Big)\Big|\;d\mu(x_1)\cdots d\mu(x_n)d\theta\\
\rightarrow\;\; & 0
\end{aligned}
$$

as δ and $\dot{\delta}$ converge to 0. Together with formulas (2.23) and (2.24) this proves the first sentence of the proof of Theorem 4.1. Note that for given σ, the sample size n has to be large enough such that $\sqrt{n}\sigma^{-1}w_0(\sqrt{n}\sigma^{-1}(\theta-\theta_0))$ puts all its mass within Θ. This can be done since w_0 has bounded support and Θ is open.

B.4 Notes Chapter 7

Note 9

Let the random vector U have a multivariate normal distribution with mean 0 and covariance matrix Σ. Then its density is equal to

$$f_U(u) = (2\pi)^{-\frac{1}{2}k}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}u^T\Sigma^{-1}u}, \quad u \in \mathbb{R}^k.$$

Now, if X, given $U = u$, is multivariate normally distributed with mean vector Iu and covariance matrix I, then its density is equal to

$$f_{X|U=u}(x) = (2\pi)^{-\frac{1}{2}k}|I|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-Iu)^T I^{-1}(x-Iu)}, \quad x, u \in \mathbb{R}^k,$$

The unconditional density of X can be computed from these two distributions

$$\begin{aligned} f_{X,U}(x,u) = f_U(u)\frac{f_{X,U}(x,u)}{f_U(u)} &= f_U(u)f_{X|U=u}(x) \\ &= (2\pi)^{-k}|I|^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}[(x-Iu)^T I^{-1}(x-Iu)+u^T\Sigma^{-1}u]} \\ &= (2\pi)^{-k}|I|^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}[x^T I^{-1}x-2u^T x+u^T(\Sigma^{-1}+I)u]} \\ &= (2\pi)^{-k}|I|^{-\frac{1}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}\binom{x}{u}^T A\binom{x}{u}}, \end{aligned}$$

where we write

$$A = \begin{pmatrix} I^{-1} & -\mathbf{1} \\ -\mathbf{1} & \Sigma^{-1}+I \end{pmatrix} = \begin{pmatrix} I\Sigma I + I & I\Sigma \\ \Sigma I & \Sigma \end{pmatrix}^{-1}$$

with $\mathbf{1}$ denoting the $k \times k$ identity matrix. This proves (7.2)

To prove (7.3), since the distributions are multivariate normal, we only have to check the expectation and covariances of X and $CX - U$. Clearly the expectations vanish. We get

$$\begin{aligned} &\mathrm{E}\,X(CX-U)^T = \\ &= \mathrm{E}\,X(CX)^T - \mathrm{E}\,XU^T \\ &= \mathrm{E}\,(XX^T)C^T - \mathrm{E}\,XU^T \\ &= (I\Sigma I + I)C^T - I\Sigma = \mathbf{0}, \end{aligned}$$

and (all matrices are symmetric and so matrix multiplication is commutative)

$$\begin{aligned} &\mathrm{E}\,(CX-U)(CX-U)^T = \\ &= \mathrm{E}\,(CX)(CX)^T - \mathrm{E}\,(CX)U^T - \mathrm{E}\,U(CX)^T + \mathrm{E}\,UU^T \\ &= C\mathrm{E}\,XX^TC^T - 2C\mathrm{E}\,XU^T + \mathrm{E}\,UU^T \\ &= CI\Sigma - 2CI\Sigma + \Sigma \\ &= C(-I\Sigma + (I\Sigma I + I)I^{-1}) \\ &= C. \end{aligned}$$

Together with $\mathrm{E}\,XX^T = I\Sigma I + I$ this proves (7.3).

Bibliography

[1] Begun, J.M., Hall, W. J., Huang, W.-M., and Wellner, J. A. (1983), Information and asymptotic efficiency in parametric–nonparametric models, *Ann. Statist.* **11**, 432–452.

[2] Beran, R. (1974), Asymptotically efficient adaptive rank estimates in location models, *Ann. Statist.* **2**, 63–74.

[3] Bickel, P.J. (1982), On adaptive estimation, *Ann. Statist.* **10**, 647–671.

[4] Bickel, P.J. and Klaassen, C.A.J. (1986), Empirical bayes estimation in functional and structural models, and uniformly adaptive estimation of location, *Adv. Appld. Math.* **7**, 55–69.

[5] Bickel, P.J., Klaassen, C.A.J., Ritov, Y., and Wellner, J.A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore. Paperbound edition 1998, Springer, New York.

[6] Billingsley, P. (1968), *Convergence of Probability Measures*, Wiley, New York.

[7] Box, G.E.P., Jenkins, G.M., and Reinsel, G.C. (1994), *Time series analysis: Forecasting and control*, third edition, Prentice Hall, Englewood Cliffs, NJ.

[8] Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton.

[9] Drost, F.C. and Klaassen, C.A.J. (1997), Efficient estimation in semiparametric GARCH models, *J. Econometrics* **81**, 193–221.

[10] Drost, F.C., Klaassen, C.A.J., and Werker, B.J.M. (1994), Adaptiveness in time-series models, in P. Mandl and M. Hušková, editors, *Asymptotic Statistics*, 203–211, Physica Verlag, New York.

[11] Drost, F.C., Klaassen, C.A.J., and Werker, B.J.M. (1997), Adaptive estimation in time-series models, *Ann. Statist* **25**, 786–817.

[12] van Eeden, C. (1970), Efficiency-robust estimation of location, *Ann. Math. Statist.* **41**, 172–181.

[13] van Es, B., Klaassen, C.A.J., and Oudshoorn, K. (2000), Survival analysis under cross sectional sampling, *J. Statist. Plann. Inf.* **91**, 295–312.

[14] Grimmett, G. and Stirzaker, D.R. (1992), *Probability and Random Processes*, Clarendon Press, Oxford.

[15] Hájek, J. (1962), Asymptotically most powerful rank-order tests, *Ann. Math. Statist.* **33**, 1124–1147.

[16] Hájek, J. (1970), A characterization of limiting distributions of regular estimates, *Z. Wahrsch. Gebiete* **14**, 323–330.

[17] Hájek, J. and Šidák, Z. (1967), *Theory of Rank Tests*, Academia, Prague.

[18] van den Heuvel, E.R. (1996), *Bounds for Statistical Estimation in Semiparametric Models*, PhD thesis, University of Amsterdam.

[19] van den Heuvel, E.R. and Klaassen, C.A.J. (1997), Spread, estimators and nuisance parameters, *Bernoulli* **3**, 323–328.

[20] Klaassen, C.A.J. (1987), Consistent estimation of the influence of locally asymptotically linear estimates, *Ann. Statist.* **15**, 1548–1562.

[21] Klaassen, C.A.J. (1989a), Estimators and spread, *Ann. Statist.* **17**, 859–867.

[22] Klaassen, C.A.J. (1989b), The asymptotic spread of estimators, *J. Statist. Plann. Inference* **23**, 267–285.

[23] Klaassen, C.A.J., Mokveld, Ph.J., and van Es, B. (2003), Efficient estimation in the accelerated failure time model under cross sectional sampling, *Math. Preprint Series 03-06, Korteweg-de Vries Instituut voor Wiskunde, Universiteit van Amsterdam.*

[24] Klaassen, C.A.J. and Putter, H. (2005). Efficient estimation of Banach parameters in semiparametric models, *The Annals of Statistics* **33**, 307–346.

[25] LeCam, L. (1970), On the asssumptions used to prove asymptotic normality of maximum likelihood estimates, *Ann. Math. Statist.* **41**, 802–828.

[26] Lehmann, E.L. (1959), *Testing Statistical Hypotheses*, Wiley, New York.

[27] Pfanzagl, J. and Wefelmeyer, W. (1982), *Contributions to a General Asymptotic Statistical Theory*, volume 13 of *Lecture Notes in Statistics*, Springer Verlag, Berlin.

[28] Rudin, W. (1966), *Real and Complex Analysis, second edition*, Tata McGraw Hill, New Delhi.

[29] Stein, C. (1956), Efficient nonparametric testing and estimation, in *Proc. Third Berkeley Symp. Math. Statist. Prob.* **1** 187–195, Univ. California Press, Berkeley.

[30] Stone, C.J. (1975), Adaptive maximum likelihood estimators of a location parameter, *Ann. Statist.* **3**, 267–284.