research
reports

No. 404 May 1999

Lars Korsholm

Likelihood Methods in Semiparametric Models

department of
theoretical
statistics

university of
aarhus

Likelihood Methods in Semiparametric Models

Lars Korsholm,
Department of Theoretical Statistics
and
Centre for Labour Market and Social Research
University of Aarhus

Abstract

The purpose of the present manuscript is to give an overview of the state of the fundamental statistical theory for semiparametric models, in particular the likelihood methods in such models. The focus is on the main ideas and where the problems occur in comparison with the archetypal likelihood theory of parametric models. We refer to the literature for further details and proofs. Finally, we discuss the applications in a number of examples.

Keywords: information bound; asymptotic normality; efficient estimation; inference; maximum likelihood estimation; information estimation; likelihood ratio test.

1 Introduction

In many practical applications of statistics it is unreasonable or undesirable to make full finite dimensional parametric assumptions on the probability distributions of the phenomena we observe. On the other hand a nonparametric model might loose "too much" of the structure that nevertheless is at hand. Then a semiparametric model is a competitive alternative that should be considered.

In general, semiparametric models are statistical models indexed by a parameter in an infinite dimensional set. There are two ways to formulate such models, even though a strict definition of a semiparametric model is not given. We denote by "the first type" infinite dimensional models with

a parameter map, i.e. we consider a model $\mathcal{P}$ on a measurable space $(\mathcal{X}, \mathcal{A})$ given of the form

$$\mathcal{P} = \{P_\psi \mid \psi \in \mathcal{H}\}, \tag{1}$$

where P_ψ is a probability distribution on $(\mathcal{X}, \mathcal{A})$ and $\mathcal{H}$ is an infinite dimensional set (expressing some conditions on P_ψ such that $\mathcal{P}$ is not all probability measures on $(\mathcal{X}, \mathcal{A})$). On the model (1) we have a map $\vartheta : \mathcal{P} \mapsto \Theta$, which yields the parameter of interest $\vartheta(P)$. Here Θ is a subset of a Euclidean space*. Typical parameter maps include $\vartheta(P) = \int \phi(X) dP$ for $\phi(x) = x, x^2, \mathbf{1}_{\{x \leq t\}}$ etc. In the second formulation we have (genuine) semiparametric models in the sense that they are naturally parametrized by two components (θ, η), where, typically, $\theta \in \Theta$ is a Euclidean interest parameter and η is a nuisance parameter ranging over an abstract set H of infinite dimension. I.e. the model is of the form

$$\mathcal{P} = \{P_{\theta,\eta} \mid \theta \in \Theta,\ \eta \in H\}, \tag{2}$$

where $\Theta \subseteq \mathbb{R}^d$ for some natural number d, H is an abstract set, and θ is the interest parameter (In this formulation the term *semi*-parametric becomes clear). Even though the setup is quite abstract in both formulations, we shall see that the majority of concepts studied here are given in terms of finite dimensional submodels, so we will be on familiar ground mostly. Example 11 below illustrates that the two formulations overlap but they are in general not equivalent.

Even though the semiparametric formulation in (2) is often the most appealing to keep in mind, the infinite dimensional model formulation in (1) with a parameter map is the formulation that works for most purposes. The goal here is to discuss the possibilities of estimating $\vartheta(P)$ (optimally) at the 'classical' $\sqrt{n}$–rate. In order to achieve this goal we need to reconsider some of the concepts used in regular parametric models. First we define differentiability, since coordinatewise differentiability is insufficient in abstract spaces. In Section 3 we introduce two new concepts of differentiability in order to define score functions (by Hellinger differentiability) and differentiable parameters (by pathwise differentiability). Other tools are introduced in Appendix B.1 on Banach and Hilbert spaces and in Appendix C on the theory of empirical processes, which generalize the notion of convergence in distribution to infinite dimensional spaces and non–measurable random elements. In Sections 4–6 we will discuss the estimation theory. One of the general result

*In the literature abstract parameter spaces Θ are also considered, see e.g. Chapter 5 in Bickel, Klaassen, Ritov and Wellner (1993).

for semiparametric models is the convolution theorem presented in Section 4, which gives a lower bound on the asymptotic variance of any regular estimator. This is a generalization of the well known Cramér–Rao bound for parametric models. At the end of the section we briefly mentions alternative results. Section 5 considers estimation techniques. The efficient score function can be used to express the information bound found in Section 4, and in Subsection 5.1 we discuss how it can be used for estimation as well. The construction of estimates based on estimating equations is considered in Section 5.3. In Section 6 we touch upon the test theory. The final major part of this introduction is given in Section 7, where we consider the cases when maximum likelihood methods are successful in these models, this includes parameter estimation, the use of the observed information and the likelihood ratio test. In the next section we shall look at the motivation for studying semiparametric models and list some examples that are typical for the area. We follow up on the applications in Section 8.

As the title and the outline above indicate we want to imitate the classical likelihood theory for regular parametric models "as much as possible". When is it possible to estimate the interest parameter at the usual order of accuracy? How can we construct estimators which reach the lower bound on the asymptotic variance? When will the maximum likelihood method work? Can we estimate the asymptotic variance matrix? By the observed information? What can be said about the likelihood ratio test? In answering these questions we will aim to make the differences to the classic theory and the arising problems clear to the reader.

The theory in the area is not complete. The information bound in Section 4 is well studied, but the remaining areas leave much work to be done and knowledge to be found. At present the literature in the area is dominated by papers with exact deduction in specific examples and hence it might appear confusing to the novice. However, the monograph of Bickel et al. (1993) presents the information calculations in general and gives detailed computations in many interesting examples and the authors discuss methods for constructing estimators. Chapter 25 in van der Vaart (1998) gives a general account with a clear view on both the estimation theory and the maximum likelihood methods, with proofs. In survival analysis Andersen, Borgan, Gill and Keiding (1993) show how counting process and martingale theory can be fruitfully used. For further and up to date results see the recent journals in statistics and econometrics — also note the suggestions for new students in Appendix A.

The chapter on semiparametric models in van der Vaart (1998) has been of great inspiration in preparing this account of the theory.

2 Motivation and Examples

As mentioned in the introduction semiparametric models are intermediate between parametric models and nonparametric models. We should choose a semiparametric model if we want the flexibility of the nonparametric model and if we want to answer the questions that a parametric model allow us to ask. An example might be a study of some aspect of the behavior of humans, where many incontrollable factors including the intellect of the individual reacting to its own situation might bring the nature of the experiment "too far" away from a closed laboratory trail. Another occasion where a semiparametric model is useful is the case where the model given from the scientific context only partly describes the phenomena under study, e.g. it is given that a certain relation holds in mean but no distributional form is given. This type of models occurs frequently in econometrics, for such models see Example 17 and 7 below.

Semiparametric models have roots long back in the history of statistics. In the statistical literature the theory has evolved dramatically over the last three decades, since the proposal by Cox in the early seventies to model survival times by proportional hazard functions. Since this model is a particularly nice and well known example, we will briefly discuss it here and emphasize what makes it (and survival analysis in general) so distinctive among semiparametric models.

Example 1 *(Cox model)* We observe a pair (T, Z), where T is a survival time and Z is a covariate. The conditional hazard of T given Z is given by $\lambda(t|z) = \lambda_0(t)e^{\theta^\top z}$, where λ_0 is an unknown baseline hazard function, θ is a real parameter of interest that expresses the proportional difference between hazard functions, and the distribution of Z is unrestricted. □

Example (1) was introduced in Cox (1972, 1975) and estimated by Cox's partial likelihood which is constructed to be a function of θ only. Aalen (1978) reformulated the model in terms of counting processes and this formulation was used in Andersen and Gill (1982) to give a rigorous proof for the asymptotic distributional results based on martingale theory, see also the readable survey in Gill (1984). The connection between the partial maximum likelihood estimator and the non-parametric maximum likelihood estimator was first obtained in Bailey (1979), see also Bailey (1984). Jacobsen (1984) discusses maximum likelihood estimation in a general counting process setup. Later, several other issues of the model have been studied, e.g. Jacobsen (1989) ensures uniqueness of the partial maximum likelihood estimator in absence of co-linearity in the covariates, Bartlett adjustments of the partial

likelihood ratio test have been considered in Gu and Zheng (1993), model control by test for non-proportional hazard in e.g. Murphy (1993), etc. Furthermore, the model has been extended in a variety of directions including several types of censored observations, introducing unobserved frailties, etc.

Example 1 captures the basic idea in and the strength of a semiparametric model. We parameterize the part of the model we have interest in (how different covariates influence the survival hazard) and we leave the remaining part of the model as unspecified as possible (the baseline pattern by which individuals fail). However, the partial likelihood function is special to the Cox model and it does not have a universal counterpart, and the martingale property of the likelihood is unique for survival analysis. Unfortunately, it has not been possible to generalize these strong features to a version of a likelihood for semiparametric models. The partial likelihood is replaced by various types of likelihood–like or pseudo–likelihood functions (Section 7). To obtain a convenient terminology we refer to all such functions as 'likelihoods'. The martingale tool is substituted by empirical process methods (Appendix C), which seem to give the tools that can be utilized to prove asymptotic results.

The intention with the following list of examples is to give an impression of the range of semiparametric models. We will return to these and related examples in Section 8. Each example include a reference to a starting point for further details.

Example 2 *(Parametric models)* Let μ be a fixed σ-finite measure on a sample space $(\mathcal{X}, \mathcal{A})$. We observe X with distribution P from the class $\mathcal{P} = \{P_\theta \ll \mu \mid \theta \in \Theta\}$, where Θ is an open subset of $\mathbb{R}^d$ and the parametrization $\theta \mapsto P_\theta$ satisfies the following. The map $\theta \to \sqrt{\frac{dP_\theta}{d\mu}}$ from Θ to $L^2(\mu)$ is Fréchet differentiable with derivative $s(\theta) \in \mathbf{R}^d$. The Fisher $d \times d$ information matrix for θ given by $I(\theta) = \int s(\theta)^\top s(\theta) d\mu$ is nonsingular. Finally, the map $\theta \to s_i(\theta)$ is continuous from Θ to $L^2(\mu)$ for $i = 1, \ldots, d$. Then $\mathcal{P}$ is a (finite dimensional) regular parametric model. Such a model is of course a special case of a semiparametric model. (Bickel et al. (1993, Chapter 2)) □

Example 3 *(Mixture models)* We observe a random variable X. Given an unobserved random variable Z the distribution of X has density $q_\theta(\cdot|Z)$, which belongs to a regular parametric family of densities with respect to some σ-finite measure μ. The random variable Z is assumed to have completely unknown distribution G on a measurable space $(\mathcal{T}, \mathcal{C})$. The density of X with respect to μ is given by

$$p(x; \theta, G) = \int_{\mathcal{T}} q_\theta(x|z) G(dz).$$

The probability distribution G is called the *mixing distribution*, and the function $q_\theta(\cdot|Z)$ is called the *kernel* or the *mixture density* which is known up to the parameter θ in an open subset Θ of $\mathbb{R}^d$. (Lindsay and Lesperance (1995) and van der Vaart (1996)) □

Example 4 *(Paired exponential)* In the mixture model above we observe $X = (X_1, X_2)$, which conditional on Z is a pair of independent exponentially distributed random variables with parameter Z and θZ. The interest parameter θ describes the ratio of conditional hazard rates. To model Z with an unknown distribution function yields a more flexible model for X than if z simply was an unknown parameter. That Z is random allows for unobserved heterogeneity in the population. (Bickel et al. (1993, Example 4.5)) □

Example 5 *(Errors-in-variables)* We observe a pair (X_1, X_2), where $X_1 = Z + \epsilon_1$ and $X_2 = \alpha + \beta Z + \epsilon_2$, and $\epsilon = (\epsilon_1, \epsilon_2)$ has a two dimensional normal distribution with mean zero and unknown covariance matrix, i.e. X_2 is a linear regression on Z which we observe with error ϵ_1. The distribution of Z is unknown. (Murphy and van der Vaart (1996)) □

Example 6 *(Convolution models)* Let the observation X have the same distribution as the sum $Y + Z$, where Y has a known fixed distribution G and Z has an unknown distribution F. Estimation in this type of model is in particular difficult with low rates of convergence of estimators of F. Particular examples occur when Y is an exponential variable and Z is a positive random variable, or when Y is standard Gaussian and Z has an unrestricted distribution on the real line. (Groeneboom and Wellner (1992) and Groeneboom (1996)) □

Example 7 *(Regression)* Let Z and ϵ be two independent random vectors and suppose that $Y = \mu(Z;\theta) + \sigma(Z;\theta)\epsilon$ for known functions μ and σ. We observe the pair $X = (Y, Z)$. If ϵ has a parametric distribution and the observed value of Z is treated as a constant, then this is just a classical regression model. When the distribution of ϵ belongs to an infinite dimensional set, such as all mean zero distributions, we obtain a semiparametric version of the regression model. (Horowitz (1998, Chapter 3)) □

Example 8 *(Partly linear regression)* Consider the setup from Example 7 with Z decomposed into (W, T), $\sigma \equiv 1$, and $\mu(Z, \theta) = h(\beta^\top W + \eta(T))$. Here h is a known fixed function, β is a parameter vector, and η belongs to an infinite dimensional set of "smooth" functions. We might also relax the independence assumption on Z and ϵ to require that the conditional mean of ϵ given Z is zero. In other words, we observe $X = (Y, W, T)$ and assume that $\mathbf{E}(Y|W, T) = h(\beta^\top W + \eta(T))$. (Chen (1995)) □

Example 9 *(Transformation models)* Assume that $X = (Y, Z)$ satisfies $\eta(Y) = \theta^\top Z + \epsilon$ for an unknown map η and independent random vectors Z and ϵ with known or parametrized distributions. The transformation map η is restricted in some simple way, e.g. it belongs to the set of all monotone functions. (Wang and Ruppert (1996) and Bickel et al. (1993, Section 4.7 and 6.7)) □

Example 10 *(Projection pursuit regression)* Let Z and ϵ be two independent random vectors and suppose that $Y = \eta(\theta^\top Z) + \epsilon$ for a function η ranging over a set of smooth functions, and ϵ having a mean zero normal distribution. Obviously, θ and η are confounded, but θ is estimable up to a constant. This type of model is also called a *single–index* model. (Horowitz (1998)) □

Example 11 *(Symmetric location)* Suppose that the random variable X has a symmetric distribution with unknown centre of symmetry. We want to estimate the centre of symmetry. If we assume that the probability distribution for X is dominated by the Lebesgue measure, then we parameterize the density of X by $p(x; \theta, g) = g(x - \theta)$, where g is a density with respect to the Lebesgue measure on $\mathbb{R}$ which is symmetric at zero. (Bickel et al. (1993)) □

Example 12 *(Copula model)* We observe $X = (X_1, X_2)$ with two–dimensional distribution $F_X(x_1, x_2) = G_\theta\left(G_1(x_1), G_2(x_2)\right)$, where G_θ is a bivariate distribution function known up to the parameter θ and with uniform marginals. The marginal distribution functions G_i can both be unknown or one can be known. The purpose of the Copula model is to model the covariance structure between X_1 and X_2 by the parameter θ without affecting the marginal distributions. (Klaassen and Wellner (1997)) □

Example 13 *(Missing at random)* Suppose that the second coordinate of (Y_1, Y_2) sometimes is missing. If the conditional probability that Y_2 is observed depends only on Y_1, then we say that Y_2 is *missing at random* (MAR). The interest parameter is typically a function of the distribution of Y. (van der Vaart (1998)) □

Example 14 *(Random censoring)* We observe a survival time T if it occurs before an independent censoring time C, otherwise C is observed. If Δ is the indicator variable for observing T, then the observation is the pair $X = (T \wedge C, \Delta)$. The distribution of T and C may be completely unknown or T might follow the Cox model in Example 1. (Andersen et al. (1993)) □

Example 15 *(Interval censoring)* At the random censoring time C we observe whether the "death" time T has occurred, i.e. we observe $X = (C, \Delta)$ where Δ is the indicator of the event $\{T \leq C\}$. The distribution of T and C may be as in the previous example. (van der Laan and Robins (1998)) □

Example 16 *(Frailty)* Let two survival times T_1 and T_2 conditional on the random variable (W, Z) be independent with conditional hazards of the form $\lambda(t|z) = w\lambda_0(t)e^{\beta^\top z}$. However, the variable W is not observed but independently of Z it follows a gamma distribution with mean one and variance θ. Thus W and θ model the unobserved heterogeneity and we observe $X = (T_1, T_2, Z)$. (Nielsen, Gill, Andersen and Sørensen (1992)) □

Example 17 *(Conditional moment restrictions)* Let $g(y, z; \theta)$ be a given vector function. We observe $X = (Y, Z)$ and assume that the distribution of X satisfies $\mathbf{E}_P(g(Y, Z; \theta) \mid Z) = 0$ for a unique value of θ. Except from this condition the distribution of (Y, Z) is unrestricted. (Newey (1993)) □

3 Differentiability

In this section we introduce some of the technical tools for semiparametric models. Typically, the infinite dimensional parameter belongs to a Banach or Hilbert space. A *Banach space* is an abstract linear space with a norm $\| \ \|$, such that every Cauchy sequence has a limit point in that space, i.e. a complete normed linear space. A *Hilbert space* is an abstract linear space with an inner product $\langle \cdot, \cdot \rangle$, where every Cauchy sequence is convergent with respect to the norm $\|x\| = \sqrt{\langle x, x \rangle}$, i.e. a complete inner product space. In Appendix B.1 we list some of the properties of Hilbert and Banach spaces which will be used in the sequel.

We write $\epsilon_n = o_P(\alpha_n)$ for random variables ϵ_n and real numbers α_n if $\epsilon_n/\alpha_n \to 0$ in P-probability. We write $\hat{\theta} - \theta_0 = O_P(n^{-1/2})$, if for all $\epsilon > 0$ there exists $K > 0$ such that $\limsup_{n\to\infty} P(\sqrt{n}|\hat{\theta} - \theta_0| \geq K) < \epsilon$, and we say that $\hat{\theta}$ is $\sqrt{n}$–consistent for θ_0. We use the operator notation for integrals $Pf(\cdot; \theta) = \int f(x; \theta)P(dx)$ also when the parameter θ is random. From a sample $X_1, \ldots, X_n$ we denote the empirical measure by $\mathbb{P}_n = \sum_{i=1}^n \delta_{X_i}$. Given a probability space $(\mathcal{X}, \mathcal{A}, P)$, we write $L^2(P)$ for the set of all measurable functions $f : \mathcal{X} \mapsto \mathbb{R}$ with finite second moment $Pf^2 < \infty$. If we equip $L^2(P)$ with the usual inner product $\langle f, g \rangle_P = Pfg$ and the associated norm $\|f\|_P = \sqrt{Pf^2}$ and identify functions that are equal alomost surely, then it is well known that $L^2(P)$ is a Hilbert space.

In $\mathbb{R}^n$ there is one definition of differentiability that works. In a normed linear space there are several definitions with individual advantages. In regular parametric models the derivative of the log-likelihood (the score function) is used both for construction of an estimator and for calculating the Cramér–Rao lower bound of information. Assume (for the moment) that the semiparametric model $\mathcal{P}$ is dominated by a σ-finite measure μ. Then the set $\{\sqrt{\frac{dP}{d\mu}} \mid P \in \mathcal{P}\}$ is a subset on the unit sphere in the Hilbert space $L^2(\mu)$. The size and shape of the tangent set at a given point expresses how well we can estimate a given parameter. Estimators that solve the average of a certain function (associated with the derivative of the parameter map) equal to zero, have nice properties, just like the solution to the score equation. Here follow the strict definitions.

An important metric on the set of probability distributions is the *Hellinger distance* $d_H(P,Q)$ given by

$$d_H(P,Q) \;=\; \left(\int \left| \sqrt{\frac{dP}{d\mu}} - \sqrt{\frac{dQ}{d\mu}} \right|^2 d\mu \right)^{1/2} \tag{3}$$

for μ dominating both P and Q. We define the *score functions* at a point P in $\mathcal{P}$ as follows. Let $t \mapsto P_t$ be a map from a neighbourhood of 0 in $[0,\infty)$ into $\mathcal{P}$ with $P_0 = P$ such that there exists a measurable function $g : \mathcal{X} \mapsto \mathbb{R}$, for which[†]

$$\int \left[\frac{\sqrt{dP_t} - \sqrt{dP}}{t} - \frac{1}{2} g \sqrt{dP} \right]^2 \to 0. \tag{4}$$

This is Hellinger differentiability (differentiability in quadratic mean) along the path $\{P_t\}_{0<t<\epsilon}$ at $t = 0$ with score function g. If $\{P_t\}$ ranges over a collection of submodels, we obtain a set of score functions, which we call the *tangent set* of the model $\mathcal{P}$ at the point P, denoted by $\dot{\mathcal{P}}_P$. From the lemma below we see that any score function will belong to $L^2(P)$.

In principle we want to use all submodels $\{P_t\}$ through P, but sometimes it is wiser to consider a subset. The results derived later will be relative to the tangent set determined by the chosen set of submodels. Since the path $t \mapsto P_{at}$ for $a \geq 0$ has score function ag when the submodel $t \mapsto P_t$ has score function g we see that the maximal tangent set is a cone. Hence we will assume that the tangent set always is a cone. If the tangent set is a linear space, we call it the *tangent space*. However, the literature does not agree

[†]We write the definition in this abstract way with $\sqrt{dP_t}$ and $\sqrt{dP}$ because the model $\mathcal{P}$ is not always dominated, and because the choice of dominating measure for the path $\{P_t\}$ is irrelevant.

on these definitions of tangent set and space, e.g. Bickel et al. (1993) use all 'two-sided' submodels to define the tangent set and define the tangent space as the closure of linear span of the tangent set. Usually, one has a good idea of what the tangent set 'should be', but it might require an effort to verify the conjecture.[‡] A useful method to construct the score function g for a submodel $t \mapsto P_t$ is for each x to compute the score function in the usual manner

$$g(x) = \frac{\partial}{\partial t} \log dP_t(x)\Big|_{t=0},$$

and then verify the L^2 condition (4).

Lemma 1 *Let g be the score function of the map $t \mapsto P_t$ satisfying (4). Then we have $Pg = 0$, $Pg^2 < \infty$, and the submodel $\{P_t\}$ has the local asymptotic normality* (LAN) *property*[§]

$$\log \prod_{i=1}^{n} \frac{dP_{1/\sqrt{n}}}{dP}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \frac{1}{2} Pg^2 + o_p(1).$$

Proof: See van der Vaart (1998) page 363. □

To define smooth parameter maps we will need a pathwise version of Hadamard differentiability with respect to the Hellinger metric in (3). A map $\vartheta : \mathcal{P} \mapsto \mathbb{R}^d$ is said to be *differentiable* at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear map[¶] $\dot{\vartheta}_P : L^2(P) \mapsto \mathbb{R}^d$ such that for every submodel $\{P_t\}$ with score function $g \in \dot{\mathcal{P}}_P$,

$$\frac{\vartheta(P_t) - \vartheta(P)}{t} \to \dot{\vartheta}_P g \,.$$

Observe that this differentiability requirement is twofold: the derivative must exist and be of the form $\dot{\vartheta}_P g$. Also note that if we reduce the tangent set $\dot{\mathcal{P}}_P$ by considering fewer submodels more parameter maps become differentiable.

From Hilbert space theory we know that the map $\dot{\vartheta}_P$ can be represented as an inner product with a random vector $\bar{\vartheta}_P : \mathcal{X} \mapsto \mathbb{R}^d$,

$$\dot{\vartheta}_P g = \langle \bar{\vartheta}_P, g \rangle_P = \int \bar{\vartheta}_P g dP \,.$$

The function $\bar{\vartheta}_P$ is called an influence function of the parameter ϑ and it is clear that we may add elements orthogonal to the tangent set to this

[‡]See Appendix B.3 for a comment.

[§]Again dP and dP_t denote the density under any dominating measure.

[¶]See Appendix B.1 for a precise definition.

function. However, there exists (again from the theory for Hilbert spaces) a unique function $\tilde{\vartheta}_P$ with coordinate functions in $\overline{\text{span}}\dot{\mathcal{P}}_P$, the closure of the linear span of the tangent set $\dot{\mathcal{P}}_P$. We name the function $\tilde{\vartheta}_P$ the *efficient influence function* and it can be found by projecting any influence function $\bar{\vartheta}_P$ onto $\overline{\text{span}}\dot{\mathcal{P}}_P$.

The two differentiability notions introduced above for a score function and a differential parameter map can both be seen as Hadamard differentiability tangentially to the given path, see Bickel et al. (1993, Appendix 5).

We close this section with a definition of regular estimators. Let $\{P_{t,g}\}$ denote a submodel with score function $g \in \dot{\mathcal{P}}_P$, such that the parameter ϑ is differentiable. An estimator T_n is a measurable function of the observations $X_1, \ldots, X_n$. An estimator sequence $\{T_n\}$ is called *regular* (or *locally regular*) at P for estimating $\vartheta(P)$ (with respect to the tangent set $\dot{\mathcal{P}}_P$) if there exists a probability distribution L such that for every $g \in \dot{\mathcal{P}}_P$ with corresponding submodel $\{P_{t,g}\}$ we have[‖]

$$\sqrt{n}\left(T_n - \vartheta(P_{1/\sqrt{n},g})\right) \overset{P_{1/\sqrt{n},g}}{\Rightarrow} L.$$

Here we use the notation $\Rightarrow$ for weak convergence, and $X_n \overset{P_n}{\Rightarrow} X$ denotes weak convergence under $\{P_n\}$, i.e. $P_n f(X_n) \to Pf(X)$ for every bounded continuous function f.

Note that two different notions of differentiability are used, one for score functions and one for the parameter map. These are pathwise defined and closely related to the Hadamard differentiability.

4 Information Bound

In the following three sections we consider general estimation results for semiparametric models. In the present section we consider the most clear and complete result in the area concerning the asymptotic information contained in any regular estimator sequence. The result is given in term of an upper bound on the information and any particular estimator at hand must be evaluated against this bound to determine the efficiency of the estimator. The idea here and later is to consider regular estimators and measure the "noisiness" by the asymptotic variance.

The problem of estimating the d–dimensional parameter $\vartheta(P)$ given that P belongs to the semiparametric model $\mathcal{P}$ is obviously more difficult than

[‖]This local uniformity might seem unnecessarily technical, but it is required to exclude super efficient estimators like Stein's shrinkage estimator, see Bickel et al. (1993) example 2.2.1.

estimating $\vartheta(P)$ given that P belongs to a smooth parametric submodel $\mathcal{P}_0 = \{P_\alpha \mid \alpha \in A\} \subset \mathcal{P}$. Let $I(P \mid \vartheta, \mathcal{P}_0)$ denote the Cramér–Rao information bound for estimating ϑ in the regular parametric model $\mathcal{P}_0$ containing P. Any regular estimator T in a parametric model $\mathcal{P}_0$ has asymptotic variance $\Sigma \geq I^{-1}(P \mid \vartheta, \mathcal{P}_0)$.

Definition 2 *(Information bound)* We define the lower bound on asymptotic variance in a semiparametric model $\mathcal{P}$ for estimating ϑ at P by

$$I^{-1}(P \mid \vartheta, \mathcal{P}) = \sup \left\{ I^{-1}(P \mid \vartheta, \mathcal{P}_0) \;\middle|\; P \in \mathcal{P}_0 \subset \mathcal{P} \right\},$$

where $\mathcal{P}_0$ is a regular finite dimensional submodel and the supremum is taken in the class of positive semidefinite matrices. Moreover, we define the *information* in the semiparametric model $\mathcal{P}$ for estimating θ at P by $I(P \mid \vartheta, \mathcal{P}) = (I^{-1}(P \mid \vartheta, \mathcal{P}))^{-1}$.

The connection between the "efficient influence function" and the "information" is illuminated by the following result. For simplicity, assume that the parameter $\vartheta(P)$ is real. The Cramér–Rao bound for the parameter $t \mapsto \vartheta(P_{t,g})$ in the model $\{P_{t,g}\}$ is

$$\frac{(d\vartheta(P_{t,g})/dt)^2}{Pg^2} = \frac{\langle \tilde{\vartheta}_P, g \rangle_P^2}{\langle g, g \rangle_P}.$$

From the Cauchy-Schwartz inequality and since $\tilde{\vartheta}_P \in \overline{\text{span}}\dot{\mathcal{P}}_P$, we see that taking supremum over all submodels in the display above or equivalently over all g in the tangent set gives

$$\sup_{g \in \text{span}\dot{\mathcal{P}}_P} \frac{\langle \tilde{\vartheta}_P, g \rangle_P^2}{\langle g, g \rangle_P} = P\tilde{\vartheta}_P^2 .$$

This shows that the second moment of the efficient influence function expresses the lower bound on the asymptotic variance. This can be generalized to an arbitrary Euclidean parameter to show that the lower bound on asymptotic variance is given by $P\tilde{\vartheta}_P\tilde{\vartheta}_P^\top = I^{-1}(P \mid \vartheta, \mathcal{P})$. The following theorem is a key point in semiparametric models.

Theorem 3 (Convolution)[**] *Let the parameter map $\vartheta : \mathcal{P} \mapsto \mathbb{R}^d$ be differentiable at P with respect to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\vartheta}_P$. Then we have that the asymptotic variance matrix of every regular sequence of estimators is bounded below by $P\tilde{\vartheta}_P\tilde{\vartheta}_P^\top$. Furthermore, if the tangent set $\dot{\mathcal{P}}_P$ is a convex cone, then every limit distribution L of a regular sequence of estimators can be written as the distribution of $L_0 + Z$, where $L_0 \sim N(0, P\tilde{\vartheta}_P\tilde{\vartheta}_P^\top)$ and independent of Z with an arbitrary distribution.*

[**] For a review on the convolution theorem see the end of this section.

Proof: See van der Vaart (1998) page 366. □

The interpretation of the theorem is that among regular estimators the matrix $P\tilde{\vartheta}_P\tilde{\vartheta}_P^\top$ is the "optimal" asymptotic variance matrix for estimators of the parameter $\vartheta(P)$ in the model $\mathcal{P}$. Therefore, we call an estimator sequence which is regular at P with limit distribution $L = N(0, P\tilde{\vartheta}_P\tilde{\vartheta}_P^\top)$ *asymptotically efficient* at P. It is possible to prove that a sequence of estimators $\{T_n\}$ is asymptotically efficient at P if and only if

$$\sqrt{n}\,(T_n - \vartheta(P)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\vartheta}_P(X_i) \;+\; o_P(1). \tag{5}$$

Hence we say that the influence function $\tilde{\vartheta}_P$ is efficient.

For a 'genuine' semiparametric model $\mathcal{P} = \{P_{\theta,\eta} \mid \theta \in \Theta, \eta \in H\}$ with interest parameter $\vartheta(P_{\theta,\eta}) = \theta$ the results above can be simplified. The information bound in Theorem 3 can be expressed in terms of an efficient score function. The approach is similar in spirit to the parametric case with a finite dimensional nuisance parameter.

Let $\mathcal{P}_\theta$ denote the submodel $\{P_{\theta,\eta} \mid \eta \in H\}$ where θ is fixed and let $\dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}$ denote the corresponding tangent set for η. We define $\mathcal{P}_\eta$ similarly to $\mathcal{P}_\theta$. In the parametric family $\mathcal{P}_\eta$ we have the ordinary score vector function $\dot{\ell}_{\theta,\eta} = \frac{\partial}{\partial\theta}\log dP_{\theta,\eta}$. Typically, we use submodels of the form $t \mapsto P_{\theta+ta,\eta_t}$ for given paths $t \mapsto \eta_t$, which gives score functions as a sum,

$$\frac{\partial}{\partial t}\log dP_{\theta+ta,\eta_t}\Big|_{t=0} = a^\top\dot{\ell}_{\theta,\eta} \;+\; g,$$

where g is a nuisance score function in $\dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}$. Ordinary differentiability of the parameter $\vartheta(P_{\theta+ta,\eta}) = \theta + ta$ is obvious, but our definition requires a certain form. Let $\Pi_{\theta,\eta}(f \mid S)$ denote the orthogonal projection of f in $L^2(P_{\theta,\eta})$ on the linear space S. Define the *efficient score function* for estimating θ by

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}(\dot{\ell}_{\theta,\eta} \mid \overline{\operatorname{span}}\,\dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}), \tag{6}$$

and the *efficient information matrix* by its variance matrix $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top$. The following lemma justifies this terminology.

Lemma 4 [††] *Suppose that for every $a \in \mathbb{R}^d$ and every $g \in \dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}$, there exists a path $t \mapsto \eta_t$ in H such that*

$$\int\left[\frac{\sqrt{dP_{\theta+ta,\eta_t}} - \sqrt{dP_{\theta,\eta}}}{t} - \frac{1}{2}(a^\top\dot{\ell}_{\theta,\eta} + g)\sqrt{dP_{\theta,\eta}}\right]^2 \to 0. \tag{7}$$

[††]For the same result in the alternative setup, see Theorem 3.4.1 in Bickel et al. (1993).

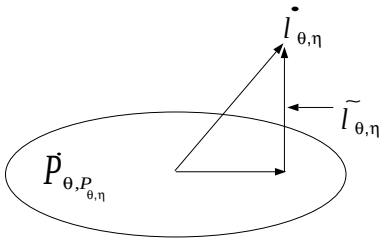

Figure 1: The geometric picture of the efficient score function $\tilde{\ell}_{\theta,\eta}$.

If $\tilde{I}_{\theta,\eta}$ is nonsingular, then the function $\vartheta(P_{\theta,\eta}) = \theta$ is differentiable at $P_{\theta,\eta}$ with respect to the tangent set $\dot{\mathcal{P}}_{P_{\theta,\eta}} = \operatorname{span} \dot{\ell}_{\theta,\eta} + \dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}$ with efficient influence function $\tilde{\vartheta}_{\theta,\eta} = \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$.

Proof: It is straightforward to verify the necessary conditions. □

From our definition of the efficient score function in (6) we may or may not have $\tilde{\ell}_{\theta,\eta}$ on closed form. The interpretation of the efficient score function is that from the score function for θ we subtract the part that can be accounted for by nuisance score functions. A part of the information for θ is lost when we do not know the nuisance parameter η. A particular nice case, called adaptation, occurs when $\dot{\ell}_{\theta,\eta} \perp \dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}$. This means that we can estimate θ equally well (up to the classical order) knowing the value of η and only knowing that the nuisance parameter belongs to the set H.

A further refinement can be achieved if the model has finite, positive information for the nuisance parameter η. Let T_η be a subset of a Hilbert space which constitutes "directions", say b, in which we can approximate η within H. Suppose that there exists a continuous linear operator $B_{\theta,\eta} : \operatorname{span} T_\eta \mapsto L^2(P_{\theta,\eta})$, and for every $a \in \mathbb{R}^d$ and $b \in T_\eta$ there exists a path $t \mapsto \eta_t$ such that the path $t \mapsto P_{\theta+ta,\eta_t}$ is Hellinger differentiable with score function of the form

$$A_{\theta,\eta}(a,b) = a^\top \dot{\ell}_{\theta,\eta} + B_{\theta,\eta}(b).$$

Here the "score operator" $B_{\theta,\eta}$ generates the score functions g for the nuisance parameter. The domain of the operator $A_{\theta,\eta} : \mathbb{R}^d \times \operatorname{span} T_\eta \mapsto L^2(P_{\theta,\eta})$ is a Hilbert space with respect to the inner product given by the sum

$$\langle (a,b),(\alpha,\beta)\rangle_\eta = a^\top \alpha + \langle b,\beta\rangle_{T_\eta}$$

The score operator $A_{\theta,\eta}$ has adjoint operator[‡‡] $A^*_{\theta,\eta} : L^2(P_{\theta,\eta}) \mapsto \mathbb{R}^d \times \overline{\operatorname{span}}\, T_\eta$, and corresponding information operator $A^*_{\theta,\eta}A_{\theta,\eta} : \mathbb{R}^d \times T_\eta \mapsto \mathbb{R}^d \times \overline{\operatorname{span}}\, T_\eta$ given by

$$\begin{aligned} A^*_{\theta,\eta} g &= (P_{\theta,\eta} g \dot{\ell}_{\theta,\eta}\ ,\ B^*_{\theta,\eta} g) \\ A^*_{\theta,\eta} A_{\theta,\eta}(a,b) &= \begin{pmatrix} I_{\theta,\eta} & P_{\theta,\eta}\dot{\ell}_{\theta,\eta} B_{\theta,\eta} \\ B^*_{\theta,\eta}\dot{\ell}^\top_{\theta,\eta} & B^*_{\theta,\eta} B_{\theta,\eta} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}, \end{aligned}$$

where $B^*_{\theta,\eta} : L^2(P_{\theta,\eta}) \mapsto \overline{\operatorname{span}} T_\eta$ is the adjoint of $B_{\theta,\eta}$. The first diagonal element in the matrix is ordinary Fisher information matrix $I_{\theta,\eta}$ for θ and the second diagonal element is the information operator for η. From Lemma 4 we know that the efficient influence function for estimating $\vartheta(P_{\theta,\eta}) = \theta$ is expressed in the efficient score function. If the information operator $B^*_{\theta,\eta}B_{\theta,\eta}$ is continuously invertible, then the orthogonal projection on the nuisance tangent set is given by the operator $B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta}$, and we obtain the following simple formula for the efficient score function

$$\tilde{\ell}_{\theta,\eta} = (I - B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta})\dot{\ell}_{\theta,\eta}\,.$$

This implies that the submodel $\{P_{\theta+t\mathbf{1},\eta_t}\}$, where $t \mapsto \eta_t$ has tangent $b = (B^*_{\theta,\eta}B_{\theta,\eta})^{-1}\ B^*_{\theta,\eta}\dot{\ell}_{\theta,\eta}$, is a "least favourable submodel", i.e. the submodel is least informative about θ.

Furthermore, under the same conditions and in the case where the parameter of interest $\vartheta(P_{\theta,\eta}) = \chi(\eta)$ is a function of the nuisance parameter η (despite the name), and there exists a continuous linear operator $\dot{\chi}_\eta : \operatorname{span} T_\eta \mapsto \mathbb{R}^d$ such that for every $b \in T_\eta$ there exists a path $t \mapsto \eta_t$ with

$$\frac{\chi(\eta_t) - \chi(\eta)}{t} \to \dot{\chi}_\eta b$$

for $t \downarrow 0$. Then the parameter map ϑ is pathwise differentiable with respect to the tangent set $A_{\theta,\eta}(\mathbb{R}^d \times T_\eta)$ with efficient influence function $\tilde{\vartheta}_{\theta,\eta}$ given by

$$P_{\theta,\eta}\tilde{\vartheta}_{\theta,\eta}\dot{\ell}_{\theta,\eta} = 0, \quad \text{and} \quad B^*_{\theta,\eta}\tilde{\vartheta}_{\theta,\eta} = \tilde{\chi}_\eta\,,$$

where $\tilde{\chi}_\eta$ is the efficient influence function for χ. If $\tilde{I}_{\theta,\eta}$ is nonsingular and $\tilde{\chi}_\eta$ belongs to the range of $B^*_{\theta,\eta}B_{\theta,\eta}$, we obtain

$$\tilde{\vartheta}_{\theta,\eta} = B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^-\tilde{\chi}_\eta - \langle B_{\theta,\eta}(B^*_{\theta,\eta}B_{\Theta,\eta})^-\tilde{\chi}_\eta, \dot{\ell}_{\theta,\eta}\rangle^\top_{P_{\theta,\eta}}\ \tilde{I}^{-1}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\,,$$

where $z = C^-y$ means nothing more than a solution to $Cz = y$. For an infinite dimensional interest parameter and no Euclidean parameter the formula

[‡‡] See Appendix B.1 for a definition.

above remains true without the second term.

It often happens that the operator $B^*_{\theta,\eta}B_{\theta,\eta}$ is **not** continuously invertible. This means that the nuisance parameter η does not have a $\sqrt{n}$–consistent estimator. In this case the maximum likelihood estimator of η is on the boundary of the parameter set H to such an extent that the usual rate breaks down and (maybe) $\tilde{l}$ is no longer a score function. However, in some of these difficult cases it is possible to estimate the interest parameter via the efficient score function with the usual $\sqrt{n}$–rate, as we shall see in the following section, since the efficient information is still well defined and nonsingular.

An information bound of the type studied here first appeared in the literature in Begun, Hall, Huang and Wellner (1983) and has later been improved and simplified in van der Vaart (1989, 1991b); Groeneboom and Wellner (1992) and Bickel et al. (1993). A presentation using differential geometry is given in Amari and Kawanabe (1997). If the efficient information is zero, then there does not exist a $\sqrt{n}$–consistent estimator, see Newey (1990), Chamberlain (1986) or van der Vaart (1991b). However, the information bound is not necessarily sharp, since Ritov and Bickel (1990) have given examples with finite and positive information where there does not exist any estimator which converges at rate $n^{-\alpha}$ for any $\alpha > 0$. It is possible to avoid the regularity assumption on the estimator sequence by using the local asymptotic minimax (LAM) bound (see van der Vaart (1998)). Both the convolution theorem and the LAM theorem have been generalized in a result concerning convergence of experiments, see van der Vaart (1991a).

5 Estimation Methods

Under regularity conditions the method of maximum likelihood can be applied successfully in finite dimensional parametric models. Unfortunately, in infinite dimensional models there is not a widely usable method for constructing the estimates. In some examples ingenious but ad hoc methods have been applied, in other cases estimating equations are successful, and in some smooth models maximum likelihood estimation is possible. In this section we will discuss some of the most general results available. We will leave the discussion of maximum likelihood methods to Section 7. For methods used in particular models references can be found in Section 8. Bickel (1982) studies estimation in adaptive models and his ideas are further developed in Schick (1986). In a semiparametric model with a sufficient statistic for η for each fixed θ van der Vaart (1988) constructs asymptotically regular estimators, which are efficient under further conditions. The moral seems to

be that in each model one needs to carefully consider which method to apply in the case at hand.

5.1 The Efficient Score Function

In a semiparametric model where θ is the interest parameter and the efficient score function is explicitly known we can use an estimating equation with the efficient score function just as the ordinary score equation is used in maximum likelihood estimation. Under conditions stated below such estimators are efficient.

Suppose that a consistent estimator $\hat{\eta}_n$ of η is given, and let $\tilde{\ell}_{\theta,\eta}(x)$ be the efficient score function for θ. Then we say that $\hat{\theta}_n$ is an *efficient score estimator* if it solves the equation in θ

$$\sum_{i=1}^{n} \tilde{\ell}_{\theta,\hat{\eta}_n}(X_i) \;=\; 0\,. \tag{8}$$

For the asymptotic results given below it suffices, in fact, that the left hand side evaluated at $\hat{\theta}_n$ is $o_P(\sqrt{n})$.

In classical theory asymptotic results for a solution to an equation of this type are proved by a linearization argument. However, such a scheme is not possible here because the estimator $\hat{\eta}_n$ does not, in general, have the usual $\sqrt{n}$–rate of convergence (when we consider the 'natural' parametrization). Instead we apply results about Donsker classes from the theory for empirical processes.

Theorem 5 *Suppose that the model $\{P_{\theta,\eta} \mid \theta \in \Theta\}$ is Hellinger differentiable in the sense of (4) with respect to θ at (θ,η), let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular, and let the efficient score estimator $\hat{\theta}_n$ be consistent for θ. Assume that the given estimator $\hat{\eta}_n$ is consistent with respect to a metric d on H and that it satisfies*

$$P_{\hat{\theta}_n,\eta}\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} \;=\; o_P(n^{-1/2} + \|\hat{\theta}_n - \theta\|)\,, \tag{9}$$

$$P_{\theta,\eta}\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} - \tilde{\ell}_{\theta,\eta}\|^2 \xrightarrow{P} 0\,, \quad P_{\hat{\theta}_n,\eta}\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\|^2 \;= O_P(1)\,. \tag{10}$$

Furthermore, suppose that there exists a $\delta > 0$ such that the set of functions $\{\tilde{\ell}_{\theta',\eta'} \mid \|\theta'-\theta\| < \delta\,,\, d(\eta',\eta) < \delta\}$ is a P–Donsker class with square-integrable envelope function. Then the sequence $\hat{\theta}_n$ is asymptotically efficient at (θ,η).

Above we estimate the efficient score function by plugging in an estimator $\hat{\eta}_n$ for the nuisance parameter. Since $P_{\hat{\theta}_n,\eta}\tilde{\ell}_{\hat{\theta}_n,\eta} = 0$, condition (9) requires

that the "bias" of the plug-in estimator $\tilde{\ell}_{\theta,\hat{\eta}_n}$ of the true efficient score function $\tilde{\ell}_{\theta,\eta}$ converges to zero at a rate faster than $1/\sqrt{n}$. This can be expected to be true, since the efficient score function is orthogonal to the score functions for the nuisance parameter. Then small changes in η should not effect its expectation. The condition is necessary within the setting of the theorem. If it fails $\sqrt{n}$ times the left hand side is added to the asymptotic linear expansion for $\hat{\theta}_n$, see van der Vaart (1998, Theorem 25.59). The second condition (10) states that the plug-in estimator must be consistent for the true efficient score function. Naturally, Theorem 5 can be generalized to the case where the plug-in estimator $\tilde{\ell}_{\theta,\hat{\eta}_n}$ for the efficient score function is substituted by other types of data dependent estimators. For further discussion of these assumptions and some generalizations and for a proof, see van der Vaart (1998) page 391pp. Of particular interest is the case where the maximum likelihood estimator $(\hat{\theta}_n, \hat{\eta}_n)$ is a solution to the efficient score equation, cf. Section 7. The limitation of this method is that the efficient score function is not always explicitly known.

5.2 The One-step Method

The purpose here is to discuss the case where by some means we have obtained a $\sqrt{n}$–consistent estimator which does not have minimal asymptotic variance. The conclusion is that we can obtain an efficient estimator based on this initial estimator by, loosely speaking, one iteration in the Newton–Raphson algorithm for solving the efficient score equation (8). We might expect this to work since a second order polynomial is maximized by one step in a Newton–Raphson algorithm.

Let $\tilde{\theta}_n$ be the given $\sqrt{n}$–consistent estimator of θ and let $\hat{\eta} = \hat{\eta}(X_1, \ldots, X_n)$ be given estimators of η. Assume without loss of generality that the initial estimators are discretized on a grid with step size $n^{-1/2}$. The basic tool is sample splitting. Let m be the integer part of $n/2$ and define

$$\hat{\eta}_{n,i} = \begin{cases} \hat{\eta}_m(X_1, \ldots, X_m) & \text{for } i > m \\ \hat{\eta}_{n-m}(X_{m+1}, \ldots, X_n) & \text{for } i \leq m \end{cases} \tag{11}$$

Then we define the one–step estimator as

$$\hat{\theta}_n = \tilde{\theta}_n - \left(\sum_{i=1}^{n} \tilde{\ell}_{\tilde{\theta}_n,\hat{\eta}_{n,i}} \tilde{\ell}^{\top}_{\tilde{\theta}_n,\hat{\eta}_{n,i}}(X_i) \right)^{-1} \sum_{i=1}^{n} \tilde{\ell}_{\tilde{\theta}_n,\hat{\eta}_{n,i}}(X_i)\,. \tag{12}$$

The discretization of $\tilde{\theta}_n$ and the sample splitting technique are for mathematical convenience. They allow simple conditions and the use of independence between parameter estimates and the data argument of the functions.

Theorem 6 *Suppose that the model* $\{P_{\theta,\eta} \mid \theta \in \Theta\}$ *is Hellinger differentiable in the sense of (4) with respect to* θ *at* (θ, η) *and let the efficient information matrix* $\tilde{I}_{\theta,\eta}$ *be nonsingular. Assume that for every deterministic sequence* $\theta_n = \theta + O(n^{-1/2})$

$$\sqrt{n} P_{\theta_n,\eta} \tilde{\ell}_{\theta_n,\hat{\eta}_n} \xrightarrow{P} 0, \quad P_{\theta_n,\eta} \|\tilde{\ell}_{\theta_n,\hat{\eta}_n} - \tilde{\ell}_{\theta_n,\eta}\|^2 \xrightarrow{P} 0. \tag{13}$$

$$\int \|\tilde{\ell}_{\theta_n,\eta} \sqrt{dP_{\theta_n,\eta}} - \tilde{\ell}_{\theta,\eta} \sqrt{dP_{\theta,\eta}}\|^2 \to 0. \tag{14}$$

Then the sequence $\hat{\theta}_n$ *is asymptotically efficient at* (θ, η).

Proof: See van der Vaart (1998) p 394. □

This idea of improving a given initial estimator can also be carried over to the general estimating equation setup considered below for mean zero square integrable functions. For a formulation of the one–step improvement result in that setup see Section 7.8 in Bickel et al. (1993).

5.3 Estimating Equations

The methods in the preceding subsections aim at efficiency, which might be too difficult to achieve in some situations. This subsection considers estimation of the interest parameter by the solution to an estimation equation (also called M– or Z–estimates in the literature). This method ignores the remaining part of the parameter in $\mathcal{P}$. We denote the interest parameter $\vartheta(P)$ by θ. This work is due to Huber (1967) and Pollard (1985) and the setup is the following.

Suppose that the following maps are given

$$\begin{aligned} \phi_n : \Theta &\to I\!R^d, \quad n = 1, 2, \ldots && \text{(random)} \\ \phi : \Theta &\to I\!R^d && \text{(deterministic).} \end{aligned}$$

We construct an estimator $\hat{\theta}_n$ of θ by solving the equation $\phi_n(\theta) = 0$.* We need the basic assumption that in the population Θ there is an element which zeroes out ϕ, i.e. (C0) there exists a unique θ_0 in Θ such that $\phi(\theta_0) = 0$. We require that the maps ϕ_n and ϕ satisfy the following four smoothness conditions. (C1 - *Convergence at the true model*) The random maps ϕ_n converge to ϕ at θ_0 in the sense that

$$\sqrt{n}\,(\phi_n - \phi)\,(\theta_0) \Rightarrow Z_0\,,$$

*Such an estimator is called a generalized M–estimator. If $\phi_n(\hat{\theta}_n)$ is only $o_P(n^{-1/2})$, we call $\hat{\theta}_n$ an asymptotic generalized M–estimator.

where Z_0 is a random variable. (C2 - *Asymptotic equicontinuity*) Assume that

$$\sup_{|\theta-\theta_0|\leq\delta_n} \frac{|\sqrt{n}(\phi_n-\phi)(\theta)-\sqrt{n}(\phi_n-\phi)(\theta_0)|}{1+\sqrt{n}|\theta-\theta_0|} \;=\; o_P(1)$$

for all sequences $\delta_n \downarrow 0$. (C3 - *Differentiability*) The map ϕ is ordinary differentiable at θ_0 with

$$\phi(\theta)-\phi(\theta_0)-\dot{\phi}_{\theta_0}(\theta-\theta_0) = o(|\theta-\theta_0|)\,,$$

where the derivative is a linear map $\dot{\phi}_{\theta_0}(\theta-\theta_0) = \dot{\phi}(\theta_0)(\theta-\theta_0)$ acting on the difference $\theta-\theta_0$. (C4 - *Nonsingular inverse*) The derivative map $\dot{\phi}_0 \equiv \dot{\phi}(\theta_0)$ is a nonsingular $d\times d$ matrix.

Theorem 7 *Suppose that C0–C4 hold. Let $\hat{\theta}_n$ be random maps in $\Theta\subseteq\mathbb{R}^d$ such that $\hat{\theta}_n \to \theta_0$ in probability , and $\phi_n(\hat{\theta}_n) \;=\; o_P(n^{-1/2})$. Then we have that*

$$\sqrt{n}(\hat{\theta}_n-\theta_0) \Rightarrow -\dot{\phi}(\theta_0)^{-1}Z_0\,.$$

Proof: See page 310 in van der Vaart and Wellner (1996). □

The assumptions of Theorem 7 are very weak. Occasionally condition C2 can be verified without the $\sqrt{n}|\theta-\theta_0|$ term in the denominator. The two conditions C1 and C2 imply that $\phi_n \to \phi$ in probability locally around θ_0. Typically we have maps which are linear in the empirical part, i.e. random maps of the form $\phi_n(\theta) = \mathbb{P}_n f(\cdot,\theta)$. In this case the term

$$\sqrt{n}(\phi_n-\phi)(\theta) = \sqrt{n}(\mathbb{P}_n - P)f(\cdot,\theta)$$

in C2 is the empirical process indexed by $\mathcal{F} = \{f_1(\cdot,\theta),\ldots,f_d(\cdot,\theta) \mid |\theta-\theta_0| \leq \delta\}$. Conditions on $\mathcal{F}$ that insure the set to be P–Donsker also imply condition C2, see e.g. Lemma 3.3.5 in van der Vaart and Wellner (1996). Furthermore, if we may choose f as the efficient score function we are within the framework of subsection 5.1, see also corollary 7.8.1 in Bickel et al. (1993).

Construction of the estimating equations can be done by the following heuristic outline. A reasonable requirement for an estimator T_n is that it is *asymptotically linear*, i.e.

$$\sqrt{n}\,(T_n-\vartheta(P)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\bar{\vartheta}_P(X_i) \;+\; o_P(1)\,.$$

If T_n is regular also, then the parameter $\vartheta(P)$ is differentiable† with influence function $\bar{\vartheta}_P$, and the difference $\bar{\vartheta}_P - \tilde{\vartheta}_P$ is orthogonal to the tangent set $\dot{\mathcal{P}}_P$. Now, consider a class of influence functions corresponding to different asymptotically linear estimator sequences and suppose that $\{\bar{\vartheta}_{\theta,\tau} \mid \tau \in T\}$ is a nice parametrization of this set, where θ is the interest parameter and τ becomes a control parameter. Then we might estimate θ from the estimating equation

$$\phi_{n,\tau}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \bar{\vartheta}_{\theta,\tau}(X_i) = 0$$

for given τ. The value of τ determines the asymptotic precision of the estimator $\hat{\theta}_n$, hence we should try to make it data dependent to obtain a "best" estimator within this class. For further discussion on estimating equations in semiparametric models see Pfanzagl (1990) and Amari and Kawanabe (1997). A result on estimating equations for infinite dimensional parameters can be found in van der Vaart (1995).

6 Test Theory

We have two things to say about testing in semiparametric models. The discussion of the likelihood ratio test is postponed to the following section on maximum likelihood methods. Here we discuss the power at local alternatives and show that a test based on an efficient estimator is "asymptotically optimal."

Let $\mathcal{P}$ be a given model with real interest parameter $\vartheta(P)$. We want to test the null hypothesis $H_0 : \vartheta(P) = 0$ against the alternative $H_1 : \vartheta(P) > 0$. Let P be a measure on the boundary of the null hypothesis, i.e. $\vartheta(P) = 0$. We want to analyze the "local asymptotic power" in a neighbourhood of P by the means of the submodels $\{P_{t,g}\}$, where g belongs to the tangent set $\dot{\mathcal{P}}_P$ and ϑ is differentiable. For $g \in \dot{\mathcal{P}}_P$ with $P\tilde{\vartheta}_P g > 0$ the differentiability of ϑ gives that $\vartheta(P_{t,g}) = tP\tilde{\vartheta}_P g + o_P(t)$. Thus for t sufficiently small, $P_{t,g}$ belongs to the alternative hypothesis $H_1 : \vartheta(P) > 0$. The following theorem gives an upper bound on the asymptotic power at the alternatives $P_{h/\sqrt{n},g}$.

Theorem 8 *Let the parameter map $\vartheta : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\vartheta}_P$. Suppose that $\vartheta(P) = 0$. Then for every sequence of power functions $P \mapsto \pi_n(P)$ of level–α tests for the hypothesis $H_0 : \vartheta(P) \leq 0$, for every $g \in \dot{\mathcal{P}}_P$ with $P\tilde{\vartheta}_P g > 0$,*

†See Bickel et al. (1993) Proposition 3.3.1.

and every $h > 0$, we have that

$$\limsup_{n\to\infty} \pi_n(P_{h/\sqrt{n},g}) \le 1 - \Phi\left(z_\alpha - h\frac{P\tilde{\vartheta}_P g}{(P\tilde{\vartheta}_P^2)^{1/2}}\right),$$

where z_α is the α–quantile and Φ is the standard normal distribution function.

The preceding theorem is accompanied by the following result, which states that the power of a test based on an efficient estimator T_n is asymptotically locally uniformly most powerful.

Corollary 9 *Let the parameter map $\vartheta : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\vartheta}_P$. Suppose that $\vartheta(P) = 0$ and that the estimator T_n is regular at P with limit distribution $N(0, P\tilde{\vartheta}_P^2)$. Furthermore, suppose that S_n^2 is a consistent estimator of the asymptotic variance of T_n, i.e. $S_n^2 \to P\tilde{\vartheta}_P^2$ in probability. Then, for every $h \ge 0$ and $g \in \dot{\mathcal{P}}_P$ we have that*

$$\limsup_{n\to\infty} P_{h/\sqrt{n},g}\left(\frac{\sqrt{n}T_n}{S_n} \ge z_\alpha\right) = 1 - \Phi\left(z_\alpha - h\frac{P\tilde{\vartheta}_P g}{(P\tilde{\vartheta}_P^2)^{1/2}}\right).$$

The results above are as one would hope and expect, i.e. tests based on efficient estimators are asymptotically efficient. These results were given in Andersen et al. (1993) and Choi, Hall and Schick (1996) who also consider two sided alternatives and multidimensional hypotheses. The present formulation follows from Section 25.6 in van der Vaart (1998).

7 Maximum Likelihood Estimation

A particularly interesting type of semiparametric models are those where maximum likelihood methods can be applied. In parametric models we derive three important conclusions from the likelihood function, the value of the argument that maximizes the likelihood function is a good estimator of the unknown parameter, the second–order derivative at the maximum is used to estimate the Fisher information, and finally we construct a confidence set of parameter values where the likelihood ratio exceeds a certain level. These goals are pursued for semiparametric models in the three subsections below. The ideas from parametric likelihood theory can be generalized to either maximizing an appropriate "likelihood" over the infinite dimensional space or we can require that infinitely many "score" equations are satisfied. At first such a task might seem impossible, but that is not true. In the Cox model

it is actually easy, in other models it is numerical laborious but possible. The computational methods change from case to case; for a review of some numerical procedures see Appendix B.2.

Even if the semiparametric model $\mathcal{P}$ is dominated it is not given that a maximizer exists. Hence the first difficulty with applying likelihood methods in semiparametric models is to choose an appropriate (pseudo–) likelihood for the experiment. It happens that different choices of likelihoods result in different maximum likelihood estimators. At present it appears that in each concrete example one needs to determine carefully the likelihood that should be used. To illustrate this point consider the model of all measures on the real line dominated by the Lebesgue measure. Then the (ordinary) likelihood function

$$L_n(P) = \prod_{i=1}^{n} p(X_i)\,, \tag{15}$$

where p is the Lebesgue density of P, is unbounded and has no maximizer. One way to extend the likelihood is to allow for discrete distributions, i.e. to work with the *empirical* likelihood function given by

$$L_n(P) = \prod_{i=1}^{n} P\{X_i\}\,, \tag{16}$$

where $P\{x\}$ denotes the mass of P at x. This likelihood function has the well-known empirical distribution $\mathbb{P}_n$ with mass $1/n$ at each observation as its maximizer. As one would expect this estimator is optimal. Other modifications of the likelihood function are: A sieved likelihood, where we maximize the function (15) over a finite dimensional sieve $\mathcal{P}_m$ in $\mathcal{P}$ with a dimension $m = m(n)$ that grows with the sample size such that $\mathcal{P}_m$ becomes dense in $\mathcal{P}$. A penalized likelihood, where we subtract from (15) a disappearing penalty

$$L_n(p) = \prod_{i=1}^{n} p(X_i) \;-\; \lambda_n J(p)\,,$$

which (typically) forces the likelihood to prefer smooth parameters. Quasi likelihoods, where the density p is substituted by another reasonable "likelihood–like" object. A weighted likelihood utilizes the idea in kernel estimation to the semiparametric setting.[‡] Finally, combinations of these methods are also used. Strictly our (pseudo–) likelihood in (17) is only a criteria function but we will use the 'likelihood' terminology in accordance with some part of the literature. The resulting estimator is often called a nonparametric

[‡]See Hunsberger (1994).

maximum likelihood estimator (NPMLE) due to the large dimension, but we shall refer to it as the maximum likelihood estimator or just as the MLE[§].

In the sequel we will assume that a likelihood $\mathrm{lik}(\psi)(x)$ of an observation x and a parameter ψ is given. For an i.i.d. sample the likelihood function is given by

$$L_n(\psi) = \prod_{i=1}^{n} \mathrm{lik}(\psi)(X_i)\,. \tag{17}$$

In a semiparametric model with a decomposed parameter $\psi = (\theta, \eta) \in \Theta \times H$ a central object is the *profile* likelihood function

$$PL_n(\theta) \;=\; \sup_{\eta \in H} L_n(\theta, \eta)\,, \tag{18}$$

where we maximize the likelihood function over η for each θ fixed. Let $\hat{\eta}(\theta)$ denote a maximizer of (18), so that $PL_n(\theta) = L_n(\theta, \hat{\eta}(\theta))$. The maximizer $\hat{\theta}$ of PL_n is the first argument of the overall MLE $(\hat{\theta}, \hat{\eta}) = (\hat{\theta}, \hat{\eta}(\hat{\theta}))$. Under regularity conditions the profile likelihood function can be used, largely as an ordinary parametric likelihood function, see Subsection 7.4.

The concept of a least favourable submodel is important in the field of maximum likelihood estimation, it is a finite dimensional submodel where the score function equals the efficient score function, see (20) and Appendix B.3. The heuristic argument and picture one should keep in mind follow. The MLE is also the maximum likelihood estimator in any parametric submodel passing through the MLE. If we have a least favourable submodel at any relevant point in the parameter space, we know that the (ordinary) score in the least favourable submodel at the MLE is zero. By continuity the least favourable submodel at the maximizer approximates the least favourable submodel at the true value of the parameter, hence we might expect that the MLE is not fare away from the maximizer in the least favourable submodel at the true value of the parameter. That is we expect that the MLE will be asymptotically equivalent to the parametric maximum likelihood estimator in the least favourable submodel at the true value of the parameter. The reduction to least favourable submodels have been used in Severini and Wong (1992) and goes back to Levit (1978) and Stein (1956).

7.1 Asymptotic Normality and Efficiency

As outlined above there are two ways of proving asymptotic normality and efficiency of the MLE in semiparametric models. The method based on gener-

[§] The name MLE is also used if the estimator is the solution to an infinite dimensional set of equations.

alized score equations or likelihood equations has been considered in a series of papers, where Gill (1989), Gill and van der Vaart (1993) and van der Vaart (1995) are the most outstanding. The focus in the first two references (which are part I and II of the same paper) is on explaining the underlining idea. The authors consider cases where the MLE is determined as the solution of the likelihood equations for a collection of smooth parametric submodels. Efficiency in the semiparametric sense of such infinite dimensional estimators is proved and the result is transformed to be valid for interest parameters by the delta method. The last reference gives a simple proof of efficiency within the same setup.

The score equation method has the complication that the classical pointwise Taylor expansion fails to work, so advanced tools including empirical process theory have to be applied and the method also implies that the entire parameter ψ or (θ, η) is $\sqrt{n}$–consistent. Here we outline the typical derivation of the asymptotic distribution of the MLE defined by a set of likelihood equations with a decomposed parameter $\psi = (\theta, \eta)$.[¶] The situation is quite close to the finite dimensional nuisance parameter case. The true value of the parameters is denoted by a subscript zero and for convenience the true data generating measure is denoted $P_0 = P_{\theta_0, \eta_0}$.

Suppose that the MLE $(\hat{\theta}, \hat{\eta})$ maximizes the likelihood function given in (17). Let $\dot{l}_{\theta,\eta}(x)$ denote the ordinary derivative of $\log \mathrm{lik}(\theta, \eta)(x)$ with respect to θ. Hence, by varying θ, the MLE satisfies the equation

$$\mathbb{P}_n \dot{\ell}_{\theta,\eta} \;=\; 0.$$

For the nuisance parameter η we typically consider the submodels $t \mapsto \eta_t$ which are used to define the nuisance tangent set. If, for each direction h in an index set $\mathcal{H}$, the score for η takes the operator form $B_{\theta,\eta}$ working on the direction h, then the likelihood equation for the nuisance parameter becomes

$$\mathbb{P}_n B_{\hat{\theta},\hat{\eta}} h \;=\; 0$$

for all $h \in \mathcal{H}$. This is true for the MLE if for every (θ, η) there exists a path $t \mapsto \eta_t(\theta, \eta)$ with $\eta_0(\theta, \eta) = \eta$, such that

$$B_{\theta,\eta} h(x) = \frac{\partial}{\partial t} \log \mathrm{lik}(\theta, \eta_t(\theta, \eta))\Big|_{t=0}.$$

Suppose that the index set $\mathcal{H}$ is chosen in such a manner that the map $h \mapsto B_{\theta,\eta} h(x)$ is uniformly bounded on $\mathcal{H}$, for every x and (θ, η). Then we

[¶]Basically, we transform the MLE into an estimator defined by an infinite dimensional set of equations and apply a result for generalized estimating equations together with some efficiency considerations.

define random maps $\Psi_n : \mathbb{R}^d \times H \mapsto \mathbb{R} \times \ell^\infty(\mathcal{H})$ by $\Psi_n = (\Psi_{n1}, \Psi_{n2})$, where

$$\begin{aligned} \Psi_{n1}(\theta,\eta) &= \mathbb{P}_n \dot{\ell}_{\theta,\eta} \\ \Psi_{n2}(\theta,\eta)h &= \mathbb{P}_n B_{\theta,\eta} h, \qquad h \in \mathcal{H}\,. \end{aligned}$$

The asymptotic version of these maps is the expectation under the parameter (θ_0, η_0) which gives a deterministic map $\Psi = (\Psi_1, \Psi_2)$

$$\begin{aligned} \Psi_1(\theta,\eta) &= P_0 \dot{\ell}_{\theta,\eta} \\ \Psi_2(\theta,\eta)h &= P_0 B_{\theta,\eta} h, \qquad h \in \mathcal{H}\,. \end{aligned}$$

These maps are constructed such that Ψ_n is zero at the MLE and Ψ is zero at the true value of the parameter (θ_0, η_0). The following theorem is based on a linearization argument.

Theorem 10 *Suppose that the following four conditions are satisfied. (i) The set of "score functions"* $\left\{\dot{\ell}_{\theta,\eta}, B_{\theta,\eta}h \mid h \in \mathcal{H},\ (\theta,\eta) \in U\right\}$*, where* U *is a neighbourhood of* (θ_0, η_0)*, is contained in a* P_0*–Donsker class. (ii) The score functions are continuous along the MLE sequence*

$$P_0\|\dot{\ell}_{\theta,\eta} - \dot{\ell}_{\theta_0,\eta_0}\|^2 \to 0, \qquad \sup_{h\in\mathcal{H}} P_0|B_{\hat{\theta},\hat{\eta}}h - B_{\theta_0,\eta_0}h|^2 \to 0$$

in P_0*–probability. (iii) The map* $\Psi : \Theta \times H \mapsto \mathbb{R}^d \times \ell^\infty(\mathcal{H})$ *is Fréchet differentiable at* (θ_0, η_0)*, with derivative* $\dot{\Psi}_0 : \mathbb{R}^d \times \mathrm{span}H \mapsto \mathbb{R}^d \times \ell^\infty(\mathcal{H})$*. (iv) The derivative* $\dot{\Psi}_0$ *has a continuous inverse on its range. If the sequence of maximum likelihood estimators is consistent for* (θ_0, η_0) *and satisfies* $\sqrt{n}\Psi_n(\hat{\theta}, \hat{\eta}) = o_{P_0}(1)$*, then*

$$\dot{\Psi}_0 \sqrt{n}\left(\hat{\theta} - \theta_0, \hat{\eta} - \eta_0\right) = -\sqrt{n}\Psi_n(\theta_0, \eta_0) + o_P(1)\,. \tag{19}$$

Proof: See page 420 in van der Vaart (1998). □

This theorem contains the joint asymptotic distribution of the MLE. Since $\sqrt{n}\Psi_n(\theta_0, \eta_0)$ is the empirical process of the Donsker class including the functions $\dot{\ell}_0$ and $B_0 h$ for $h \in \mathcal{H}$ the right hand side is asymptotically Gaussian. Since the continuous linear inverse of $\dot{\Psi}_0$ preserves normality, we see that the sequence $\sqrt{n}(\hat{\theta} - \theta_0, \hat{\eta} - \eta_0)$ is asymptotically normally distributed.

For known θ_0 the information operator for η is $B_0^* B_0$. If this operator is continuously invertible and $h = (B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$ is a potential direction in $\mathcal{H}$, then the efficient score function is given by $\tilde{\ell}_0 = \dot{\ell}_0 - B_0(B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$. Furthermore, by Theorem 5 or by direct manipulation of the system of equations in (19) we obtain that

$$\tilde{I}_{\theta_0,\eta_0}\sqrt{n}(\hat{\theta} - \theta_0) = \sqrt{n}(\mathbb{P}_n - P_0)\tilde{\ell}_{\theta_0,\eta_0} + o_P(1)$$

Thus the MLE is asymptotically linear in the efficient influence function and hence asymptotically efficient. This is a complete mimic of the parametric case with a nuisance parameter, only the notation is more difficult to handle and the mathematics more abstract to manage.

The condition (vi) in the theorem above is often the most difficult to verify. In the Euclidean case the matrix $\dot{\Psi}_0^{-1}$ is automatically continuous if it exists, i.e. if $\dot{\Psi}_0$ is one–to–one. However, in this theorem we have some freedom in the choice of the set of directions $\mathcal{H}$. A larger set makes $\dot{\Psi}_0^{-1}$ more likely to be continuous, but makes the differentiability of $\dot{\Psi}_0$ and the Donsker condition more restrictive. For a discussion on how to compute the operator $\dot{\Psi}_0$ and to prove that it is continuously invertible, see p. 421-424 in van der Vaart (1998).

If the nuisance parameter cannot be estimated with a $\sqrt{n}$–rate of convergence, we can use a general result in Huang (1996) within the setup from above. The result requires that a positive rate of convergence, say $\beta > 0$, of the estimator of η is given, i.e. $\|\hat{\eta} - \eta_0\| = O_P(n^{-\beta})$, and that the differentiability condition on the generalized score map Ψ can be strengthened to a smoothness condition of order $\alpha > 1$ (with $\alpha\beta > 1/2$) in the nuisance parameter. Together with assumptions similar to those in the previous theorem these imply asymptotic normality and efficiency of the MLE for θ.

Another way to avoid the $\sqrt{n}$–consistency of the estimator for η is to employ least favourable submodels. If the MLE satisfies the equation $\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{\eta}} = 0$, then efficiency follows from Theorem 5. Despite the name, the efficient score function is not, in general, a score function; it is defined as the projection of the score function for the interest parameter on the orthogonal complement of the nuisance tangent set. Occasionally there exist paths $t \mapsto \eta_t(\hat{\theta}, \hat{\eta})$ such that $\eta_0(\hat{\theta}, \hat{\eta}) = \hat{\eta}$, and for every x

$$\tilde{\ell}_{\hat{\theta},\hat{\eta}}(x) = \frac{\partial}{\partial t} \log \operatorname{lik}\left(\hat{\theta} + t, \eta_t(\hat{\theta}, \hat{\eta})\right)(x)\Big|_{t=0}.$$

Then the MLE satisfies the efficient score equation. It is far from obvious how one should construct these (exact) least favourable submodels, in particular at the MLE , since $(\hat{\theta}, \hat{\eta})$ often belongs to the "boundary" of the parameter space. However, inspection of the conditions in Theorem 5 reveals that the MLE does not have to solve the efficient score equation exactly, an $o_P(n^{-1/2})$ error term is sufficient, and the function $\tilde{\ell}_{\theta_0,\eta_0}$ needs to be the efficient score function only at the true value of the parameter, in order to maintain the same asymptotic result.

This space for refinement is utilized by "approximately least favourable submodels", which are defined as maps $t \mapsto \eta_t(\theta, \eta)$ from a neighbourhood of $0 \in \mathbb{R}^d$ to the parameter set for η with $\eta_0(\theta, \eta) = \eta$ (for every (θ, η)) such

that

$$\tilde{\kappa}_{\theta,\eta}(x) := \frac{\partial}{\partial t} \log \operatorname{lik}\left(\theta + t, \eta_t(\theta,\eta)\right)(x)\Big|_{t=0} \tag{20}$$

exists and at (θ_0, η_0) equals $\tilde{\ell}_{\theta_0,\eta_0}$. The submodel $\{\eta_t(\theta,\eta)\}$ in the nuisance parameter space passes through η at $t = 0$, and at the true value of the parameter $\tilde{\kappa}$ is the efficient score function. As a minimum we need these submodels at (θ_0, η_0) and at any possible value of $(\hat{\theta}, \hat{\eta})$.

When $(\hat{\theta}, \hat{\eta})$ is the MLE it also maximizes the likelihood over the approximately least favourable submodel, i.e. the function $t \mapsto \mathbb{P}_n \log \operatorname{lik}\Big(\hat{\theta} + t, \eta_t(\hat{\theta}, \hat{\eta})\Big)$ is maximized at $t = 0$. Hence the MLE satisfies the equation $\mathbb{P}_n \tilde{\kappa}_{\hat{\theta},\hat{\eta}} = 0$ and Theorem 5 can be reformulated to give the relevant asymptotic efficiency result.

Theorem 11 *Suppose that the model $\{P_{\theta,\eta} \mid \theta \in \Theta\}$ is Hellinger differentiable in the sense of (4) with respect to θ at (θ, η), let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular, and assume that the maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$ is consistent. Assume that $\tilde{\kappa}_{\theta,\eta}$ are score functions of approximately least favourable submodels in the sense of (20) and that they satisfy*

$$P_{\hat{\theta}_n,\eta_0} \tilde{\kappa}_{\hat{\theta}_n,\hat{\eta}_n} = o_P(n^{-1/2} + \|\hat{\theta}_n - \theta_0\|)\,, \tag{21}$$

$$P_{\theta_0,\eta_0} \|\tilde{\kappa}_{\hat{\theta}_n,\hat{\eta}_n} - \tilde{\kappa}_{\theta_0,\eta_0}\|^2 \xrightarrow{P} 0\,, \quad P_{\hat{\theta}_n,\eta_0} \|\tilde{\kappa}_{\hat{\theta}_n,\hat{\eta}_n}\|^2 = O_P(1)\,. \tag{22}$$

Furthermore, suppose that there exists a $\delta > 0$ such that the set of functions $\{\tilde{\kappa}_{\theta,\eta} \mid \|\theta - \theta_0\| < \delta\,,\, d(\eta,\eta_0) < \delta\}$ is a P_0–Donsker class with square-integrable envelope function. Then the sequence $\hat{\theta}_n$ is asymptotically efficient at (θ_0, η_0).

The no bias condition (21) can be verified by permuting the left hand side around $\tilde{\kappa}_{\theta_0,\eta_0}$ and establishing rate of convergence of $\hat{\eta}$. This is usually completed by results for empirical processes, cf. Theorem 25.81 in van der Vaart (1998)

7.2 Estimation of the Efficient Information

The results in the previous subsection state that we may have strong confidence in our estimator, i.e. the estimator approaches the true value with an error $1/\sqrt{n}$ times a "minimal" disturbance. However, if we have an estimator of the asymptotic variance we increase our benefit of the previous theorems with standard errors and confidence regions. Partly, the results discussed here have applications beyond the MLE.

In the case where the efficient score function is a known function on closed form and we have used a result similar to Theorem 11 for proving asymptotic efficiency of the estimator $(\hat{\theta}, \hat{\eta})$, then a natural estimator of $\tilde{I}_{\theta_0,\eta_0} = P_0 \tilde{\ell}_{\theta_0,\eta_0} \tilde{\ell}^{\top}_{\theta_0,\eta_0}$ is the sample average of $\tilde{\ell}_{\theta,\eta} \tilde{\ell}^{\top}_{\theta,\eta}$ evaluated at the estimator.

Lemma 12 *Assume that (i) for some neighbourhood U of (θ_0, η_0) the class $\{\tilde{\ell}_{\theta,\eta} \mid (\theta,\eta) \in U\}$ is P–Donsker with square-integrable envelope function, (ii) the estimator $(\hat{\theta}, \hat{\eta})$ is consistent, and (iii)*

$$P_0 \| \tilde{\ell}_{\hat{\theta},\hat{\eta}} - \tilde{\ell}_{\theta_0,\eta_0} \|^2 \xrightarrow{P} 0 \,. \tag{23}$$

Then we have that

$$\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{\eta}} \tilde{\ell}^{\top}_{\hat{\theta},\hat{\eta}} \xrightarrow{P} \tilde{I}_{\theta_0,\eta_0} \,. \tag{24}$$

Proof: See Appendix B. □

Observe that the assumptions of the lemma typically have been proved when establishing the asymptotic distribution of the estimator. Unfortunately, in important examples the efficient score function is not known on closed form, whence alternative measures are needed. Here we discuss the ideas of using the observed information obtained from the profile likelihood PL_n given in (18), or estimating the efficient score function by a nonparametric least square regression, and finally by using the bootstrap method.

Since the profile likelihood $\theta \to PL_n(\theta)$ may not be given explicitly and existence of a second order derivative matrix is in general unclear, several authors have proposed to use a discretized version of the observed profile information. Under regularity conditions and existence of approximately least favourable submodels Murphy and van der Vaart (1997a) prove that for every $h_n \xrightarrow{P} 0$ such that $(\sqrt{n} h_n)^{-1} = O_P(1)$, the sequence

$$-2 \frac{\log PL_n(\hat{\theta} + h_n v_n) - \log PL_n(\hat{\theta})}{n h_n^2} \xrightarrow{P} v^{\top} \tilde{I}_{\theta_0,\eta_0} v \,, \tag{25}$$

for every sequence of "directions" $v_n \xrightarrow{P} v \in \mathbb{R}^d$. Applying this with $v_n = e_i$, e_j and $e_i + e_j$ (where e_i is the i'th unit vector) gives the (i,j) coordinate in $\tilde{I}_{\theta_0,\eta_0}$. Following the same paper "a heuristic explanation that this method might provide a consistent estimator of the inverse of the asymptotic covariance matrix is as follows. If $\hat{\eta}_\theta$ achieves the supremum in (18), then the map $\theta \mapsto (\theta, \hat{\eta}_\theta)$ ought to be an estimator of a least favourable submodel for the estimation of θ (See Severini and Wong (1992)). By definition, differentiation

of the likelihood along the least favourable submodel (if the derivative exists) yields the efficient score function for θ. The efficient information matrix is the covariance matrix of the efficient score function, and, as usual, the expectation of minus the second derivative along this submodel should yield the same matrix."

Next, suppose that the nuisance scores are in the operator form $B_{\theta,\eta}(b)$ for some direction $b \in T_\eta$ by which we can approximate η within H. Assume that the efficient score function is given by $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - B_{\theta,\eta}(b_0)$, where b_0 minimizes the function

$$\rho(b) = P_0 \|\dot{\ell}_{\theta,\eta} - B_{\theta,\eta}(b)\|^2$$

over T_η (recall that the efficient score function is, in general, defined as the projection of the score for θ on the closure of the linear span of the nuisance scores). Let $(\hat{\theta}, \hat{\eta})$ be a consistent estimator. Then Huang (1998) proposes to estimate the least favourable direction b_0 by the minimum value of

$$\rho_n(b) = \frac{1}{n} \sum_{i=1}^{n} \|\dot{\ell}_{\hat{\theta},\hat{\eta}}(X_i) - B_{\hat{\theta},\hat{\eta}}(b)(X_i)\|^2 . \tag{26}$$

Furthermore, let $\hat{b}$ denote the minimizer of ρ_n, then Huang proposes $\tilde{\mathcal{I}}_n = \rho_n(\hat{b})$ as a natural estimator of $\tilde{I}_{\theta,\eta}$ (see Huang (1998) for further details and a proof of consistency of this estimator).

Finally, the bootstrap method is proposed in Wellner and Zhan (1996) for estimators solving a set of infinite dimensional score equations as in Theorem 10. Under an additional hypothesis on the continuity in (θ, η) of the score map Ψ, then appropriate bootstrap estimators $\widehat{(\hat{\theta}, \hat{\eta})}$ are consistent in the sense that

$$\dot{\Psi}_0 \sqrt{n} \left(\widehat{(\hat{\theta}, \hat{\eta})} - (\hat{\theta}, \hat{\eta}) \right) = -c\sqrt{n} \Psi_n(\theta_0, \eta_0) + o_P(1) \tag{27}$$

for a constant c that depends on the bootstrap sampling scheme. For the complete notation, a proof, and some examples see Wellner and Zhan (1996).

7.3 Inference

With the asymptotic distribution result for the MLE from subsection 7.1 and an estimator from subsection 7.2 of the asymptotic variance we can, by Corollary 9, perform efficient tests for the interest parameter θ based on these two estimators. In parametric models the likelihood ratio test is a popular alternative. One may ask whether such a test is meaningful and/or

possible in an infinite dimensional context. However, (once again) we shall see that approximative least favourable submodels allow to draw on finite-dimensional-type arguments. The following theorem is due to Murphy and van der Vaart (1997c) and gives the usual χ^2 distribution as the limit of the likelihood ratio test, where the degrees of freedom equal the number of parameters that are fixed.

Let $L_n(\psi)$ be the likelihood function in (17) with interest parameter $\theta(\psi) \in \mathbb{R}$ and consider the hypothesis $H_0 : \theta(\psi) = \theta_0$. Let $\hat{\psi}$ be the unrestricted MLE and let $\hat{\psi}_0$ be the maximizer under the restriction $\theta(\psi) = \theta_0$. Then the likelihood ratio statistic for testing H_0 is

$$\begin{aligned} -2\ln Q &= 2\sup_{\psi\in\Psi} \log L_n(\psi) - 2 \sup_{\psi\in\Psi,\theta(\psi)=\theta_0} \log L_n(\psi) \\ &= 2n\mathbb{P}_n \log \mathrm{lik}(\hat{\psi}) - 2n\mathbb{P}_n \log \mathrm{lik}(\hat{\psi}_0) \end{aligned} \tag{28}$$

The setup and assumptions that guarantee an asymptotic χ^2 distribution of $-2\ln Q$ are the following. Assume that the MLE $\hat{\theta} = \theta(\hat{\psi})$ is asymptotically linear,

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n}\,\mathbb{P}_n\tilde{\ell}/\tilde{I} \;+\; o_P(1)\,, \tag{29}$$

where $\tilde{I}$ is the variance of the mean zero function $\tilde{\ell}$ under P_0. Typically, $\tilde{\ell}$ is the efficient score function $\tilde{\ell}_{\theta_0,\eta_0}$ for estimating θ. In this setting the "approximately least favourable submodels" are of the form $t \mapsto \boldsymbol{\psi}_t(\psi) \in \Psi$ for every t and ψ, and passing through ψ at $t = \theta(\psi)$, i.e.

$$\theta(\boldsymbol{\psi}_t(\psi)) = t, \qquad \text{and} \qquad \boldsymbol{\psi}_t(\psi)\Big|_{t=\theta(\psi)} = \psi\,. \tag{30}$$

The corresponding submodel of log–densities

$$t \mapsto \ell(x;t,\psi) = \log \mathrm{lik}(\boldsymbol{\psi}_t(\psi))(x) \tag{31}$$

should be twice continuously differentiable for every x, with derivatives $\dot{\ell}$ and $\ddot{\ell}$. That the submodels are least favourable is expressed in the following conditions. The first condition is a double statement that in the sample average the Bartlett identity holds and that the scores in the submodel approximate the efficient score function in an L^2 sense. The second assumption is related to how well the scores approximate the efficient score function under the hypothesis. Assume that for any random sequences $\tilde{\theta} \xrightarrow{P} \theta_0$ and $\psi \xrightarrow{P} \psi_0$ we have

$$-\mathbb{P}_n\ddot{\ell}(\cdot;\tilde{\theta},\tilde{\psi}) \xrightarrow{P} P_0\tilde{\ell}^2 \tag{32}$$

$$\sqrt{n}\,\mathbb{P}_n\left(\dot{\ell}(\cdot,\theta_0,\hat{\psi}_0) - \tilde{\ell}\right) \xrightarrow{P} 0\,. \tag{33}$$

Theorem 13 *Suppose that the maps* $t \mapsto \ell(x;t,\psi)$ *satisfy (29)–(33) and that* $\hat{\psi}$ *and* $\hat{\psi}_0$ *are consistent. Then the likelihood ratio statistic is asymptotically* $\chi^2(1)$ *distributed with one degree of freedom, i.e.*

$$-2\ln Q \Rightarrow \chi^2(1)\,. \tag{34}$$

Proof: See Murphy and van der Vaart (1997c). □

7.4 Profile Likelihood and Least Favourable Submodels Revisited

Readers might recall that the results in parametric likelihood theory corresponding to the three previous subsections all build on an asymptotic expansion of the likelihood around the true value θ_0. Is such expansion possible in semiparametric models? If we have smooth least favourable submodels we may expand the profile likelihood function given in (18) around θ_0 of the form

$$\begin{aligned}\log PL_n(\tilde{\theta}) &= \log PL_n(\theta_0) + (\tilde{\theta}-\theta_0)^\top \sum_{i=1}^n \tilde{\ell}_{\theta_0,\eta_0}(X_i) \\ &\quad -\frac{1}{2}n(\tilde{\theta}-\theta_0)^\top \tilde{I}_{\theta_0,\eta_0}(\tilde{\theta}-\theta_0) + o_{P_0}\left(\sqrt{n}\|\tilde{\theta}-\theta_0\|+1\right)^2\end{aligned} \tag{35}$$

for any random sequence $\tilde{\theta} \to \theta_0$ in probability.

This asymptotic expansion is built on approximately least favourable submodels as in (30)-(31) with $\boldsymbol{\psi}_t(\psi) = (t, \eta_t(\theta,\eta))$. We require that the submodels are twice differentiable in t with $(t,\theta,\eta) \mapsto \dot{\ell}(t,\theta,\eta)$ and $(t,\theta,\eta) \mapsto \ddot{\ell}(t,\theta,\eta)$ continuous at $(\theta_0,\theta_0,\eta_0)$ a.s. and $\dot{\ell}(\theta_0,\theta_0,\eta_0)$ equal to the efficient score function for θ. Assume that there exists a neighbourhood V of $(\theta_0,\theta_0,\eta_0)$ such that $\{\dot{\ell}(t,\theta,\eta) \mid (t,\theta,\eta) \in V\}$ is P_{θ_0,η_0}–Donsker and $\{\ddot{\ell}(t,\theta,\eta) \mid (t,\theta,\eta) \in V\}$ is P_{θ_0,η_0}–Glivenko-Cantelli with envelope functions $F_1 \in L^2(P_{\theta_0,\eta_0})$ and $F_2 \in L^1(P_{\theta_0,\eta_0})$ such that $|\dot{\ell}_i(t,\theta,\eta)| \le F_1$ and $|\ddot{\ell}_{i,j}(t,\theta,\eta)| \le F_2$ for all $(t,\theta,\eta) \in V$ and $i,j = 1,\ldots,d$. Furthermore, we need to strengthen the consistency of the MLE of the nuisance parameter and retain the no-bias condition already discussed, i.e. for any sequence $\tilde{\theta} \xrightarrow{P} \theta_0$

$$\hat{\eta}(\tilde{\theta}) \to \eta_0 \tag{36}$$

$$P_0\dot{\ell}(\theta_0,\tilde{\theta},\hat{\eta}(\tilde{\theta})) = o_P(n^{-1/2} + \|\tilde{\theta}-\theta_0\|)\,, \tag{37}$$

where $\hat{\eta}(\theta)$ is the argument that maximizes the likelihood for θ fixed.

The submodels yield the Bartlett identity $P_0\ddot{\ell}(\theta_0,\theta_0,\eta_0) = -\tilde{I}_{\theta_0,\eta_0}$ and allow us to sandwich $\log PL_n(\tilde{\theta}) - \log PL_n(\theta_0)$ between appropriate terms

from the submodels $(t, \eta_t(\theta_0, \hat{\eta}_{\theta_0}))$ and $(t, \eta_t(\tilde{\theta}, \hat{\eta}_{\tilde{\theta}}))$ (for t around θ_0 and $\tilde{\theta}$, respectively), where we apply a two-term Taylor expansion in t. The Donsker and the Glivenko-Cantelli properties provide empirical process methods to control the error terms in the presence of the infinite dimensional nuisance parameter, also the no-bias condition appears in these approximations. For a rigorous proof of the asymptotic expansion under these assumptions and methods for verifying the conditions, see Murphy and van der Vaart (1997b).

From the asymptotic expansion in (35) we obtain results similar to those of the previous subsections. If $\tilde{I}_{\theta_0,\eta_0}$ is invertible and the MLE $\hat{\theta}$ of θ is consistent, then (35) may be manipulated such that the MLE has the asymptotic linear expansion

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}^{-1}_{\theta_0,\eta_0} \tilde{\ell}_{\theta_0,\eta_0}(X_i) + o_P(1) \,, \tag{38}$$

which implies asymptotic normality and efficiency of $\hat{\theta}$. Combining (35) and (38) gives an expansion of the log profile likelihood function around $\hat{\theta}$ as

$$\begin{aligned} \log PL_n(\tilde{\theta}) \;=\; \log PL_n(\hat{\theta}) \;&-\; \frac{1}{2} n (\tilde{\theta} - \hat{\theta})^\top \tilde{I}_{\theta_0,\eta_0} (\tilde{\theta} - \hat{\theta}) \\ &+\; o_P\left(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1\right)^2 . \end{aligned} \tag{39}$$

These two expansions of the log profile likelihood function justify the use of the profile likelihood as an ordinary likelihood. Murphy and van der Vaart summarize the following conclusions.

Corollary 14 *If (35) holds, the maximum likelihood estimator $\hat{\theta}$ is consistent, and the efficient Fisher information matrix $\tilde{I}_{\theta_0,\eta_0}$ is invertible, then (38) and (39) also hold. In particular, $\hat{\theta}$ is an efficient estimator with asymptotic distribution*

$$\sqrt{n}(\hat{\theta} - \theta_0) \Rightarrow N(0, \tilde{I}^{-1}_{\theta_0,\eta_0}) \,. \tag{40}$$

Corollary 15 *If (35) holds, and $\tilde{I}_{\theta_0,\eta_0}$ is invertible, then under the null hypothesis $H_0 : \theta = \theta_0$, the likelihood ratio statistic $-2 \ln Q$ is asymptotically χ^2–distributed with d degrees of freedom, i.e.*

$$-2 \ln Q = -2 \log \frac{PL_n(\theta_0)}{PL_n(\hat{\theta})} \Rightarrow \chi^2(d) \,. \tag{41}$$

Corollary 16 *If (35) holds, then the discretized observed profile information is a consistent estimator of the efficient information matrix, i.e. for all sequences* $v_n \xrightarrow{P} v \in \mathbb{R}^d$ *and* $h_n \xrightarrow{P} 0$ *such that* $(\sqrt{n}h_n)^{-1} = O_P(1)$*, we have*

$$-2\frac{\log PL_n(\hat{\theta} + h_n v_n) - \log PL_n(\hat{\theta})}{nh_n^2} \xrightarrow{P} v^\top \tilde{I}_{\theta_0,\eta_0} v \,, \tag{42}$$

We see that under these slightly stronger conditions we have "all the usual" results. However, the crucial point is the construction of the submodels which takes a creative mind. When the submodels have been established we use empirical process methods to verify the conditions.

8 Applications to Examples

The abstract theory of the previous sections has been used in a multitude of examples. Here we give a few selected examples from the literature to give the reader a flavour of what has and can be proved in concrete cases within the area. Since the theory in the previous sections is developed recently, it should be seen as an extraction of what has been shown in examples, rather than the examples should be seen as pure applications. We have gathered the examples in three groups. First we discuss some examples that are common in the semiparametric statistical literature. In the second group there are models from survival analysis, where the results from semiparametric maximum likelihood theory have been applied. Finally, we discuss examples from the econometric literature. This partition is, however, not very natural due to the large overlap between the three groups. Wellner (1985) gives a systematic review of semiparametric models, see also the many examples treated in Bickel et al. (1993) and van der Vaart (1998).

In the first group we have mixture models, Example 3, which is one of the major and most successful groups of models studied in the semiparametric literature. Lindsay and Lesperance (1995) review important areas of application and summarize some of the theoretical results including inference for the mixing distribution and the structural parameter θ. Several results on a general level are known for semiparametric mixture models. Lindsay (1983a) gives an ingenious treatment of how maximum likelihood estimation can be achieved in general mixture models, see also the discussion in Appendix B.2 on numerical methods. From Section 4.5 in Bickel et al. (1993) we know that, if the mixture density $q_\theta(x|z)$ is of the exponential form

$$q_\theta(x|z) = \exp\{z^\top T(x,\theta) + S(x,\theta) - b(\theta,z)\},$$

if the support of the mixing distribution G has a non-empty interior and some regularity conditions are satisfied, then the efficient score function is known and given by

$$\tilde{\ell} = \left[\dot{T} - \mathbf{E}(\dot{T} \mid T)\right] \mathbf{E}(Z \mid T) + \dot{S} - \mathbf{E}(\dot{S} \mid T),$$

where $\dot{S}$ and $\dot{T}$ denote the derivative with respect to θ of S and T, respectively. Thus, mixtures of exponential families have a known efficient score function. The MLE for θ is asymptotically normally distributed and efficient by van der Vaart (1996). The examples in the latter paper include the paired exponential model Example 4, the errors–in–variables model Example 5, and a scale mixture model, where the mixture density q_θ corresponds to a normal density with mean θ and variance z^2.

In mixture models with observations from the marginal distribution of X the mixing distribution G is not estimable at $\sqrt{n}$–rate. However, in *upgraded mixture models*, we observe along with the X_i's independently Z_j's with distribution G or (X_j, Z_j) with distribution $q_\theta(x|z)G\{z\}$. In these models efficient maximum likelihood estimation of G is possible with positive information, see van der Vaart and Wellner (1992) and van der Vaart (1994).

The deconvolution problem in Example 6 is also a difficult one since we have low rate of convergence of any estimator of the convoluting distribution. However, maximum likelihood estimation is possible for some classes of examples, see Groeneboom and Wellner (1992) and Groeneboom (1996). The errors–in–variables model Example 5 is extensively studied in Murphy and van der Vaart (1996) who consider maximum likelihood estimation and confidence regions based on the likelihood ratio test. See also the treatment of the Weibull mixture in Ishwaran (1996a,b). Roeder (1992) proposes a consistent estimator in the normal mean mixture model with a fixed lower bound on the variance parameter.

Another natural class in the first group of semiparametric models is formed by extensions of the classical regression model. These extensions can be carried out in two directions. The first alternative (Example 7) lets the error distribution belong to a set of infinite dimension and/or allows for correlation with the covariate. These models are discussed further below in the last group of examples. The second alternative is a semiparametric regression or a partly linear regression, introduced in Example 8, where the conditional mean of the response variable is linear in the first part and nonlinear in the second part of the covariates. To avoid over–fitting we need some smoothness assumption on the nonparametric regression function η. Typically we assume that η belongs to the Sobolev class of k–times differentiable functions with square integrable k'th derivative or we use a sieve of

spline functions of order k with an increasing number of split points. In these models ordinary likelihood methods are not usable. For a recent treatment of Example 8 see Mammen and van de Geer (1997b) who consider penalized quasi-likelihood estimation. See also Cuzick (1992); Hunsberger (1994); Severini and Staniswalis (1994); Müller and Zhao (1995) and Chen (1995). For work on the non–parametric regression model (without a linear part) see van de Geer (1990); van de Geer and Wegkamp (1996) and Mammen and van de Geer (1997a).

The second group of models are used in survival analysis. In models for survival data counting processes and martingale theory have traditionally been used to prove asymptotic results. Recently, empirical process methods and the principles discussed there have been applied to examples in the area with success. Huang and Wellner (1995) study the proportional hazard model with "case 2" interval censoring, cf. Example 1 and 15. With "case 2" censoring we have two censoring times $C_1 < C_2$ and we observe one of the events $\{T \leq C_1\}$, $\{C_1 < T \leq C_2\}$, or $\{C_2 < T\}$. The authors use the profile likelihood to estimate the regression parameter efficiently and to estimate the efficient information.

An alternative to the Cox model is the proportional odds model where the hazard ratio approaches unity when the time increases (i.e. the regressors have a disappearing effect). The model assumes that the survival function $S(\cdot\,|Z)$ given the covariate Z satisfies

$$-\text{logit}(S(t\,|Z)) = \eta(t) + \theta^\top Z\,,$$

where $\text{logit}(x) = \log(x/(1-x))$ and $\eta(t)$ is the baseline log odds of failure at time t. Murphy, Rossini and van der Vaart (1997) consider the maximum likelihood method for this model and prove $\sqrt{n}$–consistency of the estimator of the baseline odds of the failure function and asymptotic normality and efficiency of the estimator of the regression parameter. Furthermore, they verify the use of the likelihood ratio test for inference for the regression parameter. The Gamma-Frailty model occurs when the frailty in Example 16 has a Gamma distribution. This model has been considered by several authors. Parner (1998) proves that the full MLE is consistent, asymptotically Gaussian and efficient. The likelihood ratio test for the regression parameter is justified in Korsholm (1998b) by the method in subsection 7.4. See also van der Laan and Robins (1998) and Rotnitzky and Robins (1995) for related work.

The last group of models that we consider here is frequently used in econometrics. Semiparametric methods are important in that area because economic theory typically only states that a certain relation should hold (e.g. in

conditional mean) and does not yield restrictions on the distributional form of the object under study. The models from survival analysis are suitable for duration data e.g. from (un)employment spells.[‖] A key reference in the econometric literature on duration models is Heckman and Singer (1984), see also Goto (1996). Mixture models, in particular errors–in–variable models, are frequently used due to covariates being contaminated.[**] The theory from section 4 and 5 is presented in the econometric literature by Newey (1990) and applications in the econometric literature are considered in Robinson (1988), and Horowitz (1998). A version of the profile likelihood method is considered in Ai (1997). It is common in the literature to use ad hoc methods to construct estimators that achieve the semiparametric efficiency bound.

One of the favourite models in econometrics is an extension of Example 7 called the conditional moment model, which is given by the equation

$$\mathbf{E}\left(g(Y,Z;\theta) \mid Z\right) = 0 \tag{43}$$

for a unique value of the interest parameter θ, where g is a known vector function. The information bound for this model was found in Chamberlain (1987) and Newey (1993) cleverly constructs an efficient estimator using the method of moments proposed in Hansen (1982). A treatment of this model from the point of view of the present account can be found in Korsholm (1998a).

The binary response model is another example frequently used in econometrics. The model is a censored version of the regression model Example 7, where[††] $Y = \mathbf{1}_{\{\theta^\top Z+\epsilon>0\}}$. Manski (1985) proposes "the maximum score estimator" for θ in this model and Kim and Pollard (1990) show that this estimator converges with cube root rate to a non standard distribution, see also the thorough presentation in Horowitz (1998) Chapter 3.

The final model we will consider here is the Single–Index model, also known as the projection pursuit regression, in Example 10. In the single index model Y is a real variable and Z is a vector of covariates which satisfy the conditional mean condition

$$\mathbf{E}(Y \mid Z) = \eta(\theta^\top Z)\,,$$

where θ is an unknown parameter and $\eta : I\!R \mapsto I\!R$ is an unknown function. The quantity $\theta^\top Z$ is called an index and the purpose with the model is to

[‖]However, such models will describe the data set rather than explaining which factors in the market that generate the distributional form.

[**]E.g. income data from tax reports have a tendency to underestimate the true income.

[††]E.g. Y could be the decision variable for whether a person should participate in the labour market and $\theta^\top Z + \epsilon$ is the possible net earnings given the skills Z.

subtract the information in a simple and presentable way (in particular if the dimension of Z is high). For a discussion on identification, the information bound, efficient estimation of θ by weighted nonlinear least square and alternative methods see Horowitz (1998) and the references therein.

Appendix

A A Guideline for future students

The intention with this section is to list some of the main prerequisites that I wish I had known before studying semiparametric models and to list the main references to the literature. Thus this section is solely built on my own experience and gives my personal recommendations.

Beside the pre-graduate level the mathematical background should include knowledge of basic Banach and Hilbert space theory. In particular, one must be familiar with continuous linear operators, know how to perform projections (in order to compute the efficient score function), be able to determine when an operator is continuously invertible (in order to identify whether we have positive finite information in a given model), and be familiar with the three definitions of differentiability: Gâteaux (pointwise), Hadamard (compact), and Fréchet (bounded). As a possible reference we give Appendix 5 in Bickel et al. (1993).

In probability theory the usual measurability theory must be extended with the methods for empirical processes, which extend the law of large numbers and the central limit theorem to processes indexed by a class of functions. See the short introduction in Appendix C below and the recent and readable book by van der Vaart and Wellner (1996).

The statistical background should cover regular parametric models and the Cramér–Rao bound on the asymptotic information for such models (since this is one of the main objects that we can generalize to semiparametric models). The local asymptotic normality (LAN) condition and its implications also have important counterparts in semiparametric models. Finally, underneath the convolution theorems is Le Cam's three Lemmas.

Having come this far one should be well equipped for studying and doing research in semiparametric models. An overview paper as the present thesis is a recommended place to start; then follow the references to the literature from here. At present most developments appear in the statistics and econometrics journals. For a textbook in survival analysis we refer to Andersen et al. (1993). The comprehensive monograph of Bickel et al. (1993) gives a detailed account of the convolution theorem in general and information calculations

in many important examples. A broad and readable account of the theory can be found in Chapter 25 of van der Vaart (1998). Groeneboom and Wellner (1992) discuss some of the numerical methods that are used. For the maximum likelihood methods in semiparametric models see Gill (1989); Gill and van der Vaart (1993); van der Vaart (1995) and Murphy and van der Vaart (1997a,c,b).

We close this section by giving a 'user manual' with the intention to show in principle how a given semiparametric model could be analyzed. Suppose we have a semiparametric model $\{ P_{\theta,\eta} \mid \theta \in \Theta,\ \eta \in H \}$ with a real interest parameter $\theta \in \Theta \subseteq I\!R$. First we find the score functions. For θ the score function usually is the ordinary score. If the nuisance parameter is a distribution function, then we construct paths $\{\eta_t\}$ by $d\eta_t = (1 + th)d\eta$, where h is a bounded mean zero square integrable function (also paths with $d\eta_t = d\eta + th$ are used, see also the technique in Example 3.2.1 in Bickel et al. (1993)). A candidate for the nuisance score function is obtained by

$$g(x) = \frac{\partial}{\partial t} \log dP_{\theta,\eta_t}\Big|_{t=0} .$$

This suggestion must then be verified by the condition in (4) and we obtain the nuisance tangent set $\dot{\mathcal{P}}_\theta$. From here there are two possibilities. If the score function for the nuisance parameter is on the operator form $g = B_{\theta,\eta}h$ for a continuous, linear operator $B_{\theta,\eta}$ and if $B^*_{\theta,\eta}B_{\theta,\eta}$ is continuously invertible, then one can expect that both θ and η are estimable at rate $\sqrt{n}$, the efficient score function is given by

$$\tilde{\ell}_{\theta,\eta} = \left(I - B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta}\right)\dot{\ell}_{\theta,\eta} ,$$

and the submodel in the direction $h = (B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta})\dot{\ell}_{\theta,\eta}$ is least favourable. In this case we typically also have a (pseudo–) likelihood function as well, and Kiefer and Wolfowitz (1956) can be used to show consistency of the MLE. Then we apply the results in Subsection 7.4 or alternatively we use Theorem 10, one of the methods in Subsection 7.2, and Theorem 13 to obtain asymptotic normality and efficiency of the MLE $\hat{\theta}$ (or of the full MLE $(\hat{\theta},\hat{\eta})$), a consistent estimator for the efficient information, and we conduct inference by the likelihood ratio test. If $B^*_{\theta,\eta}B_{\theta,\eta}$ is not continuously invertible, then we find the efficient score function by the definition

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}(\dot{\ell}_{\theta,\eta} \mid \overline{\text{span}}\,\dot{\mathcal{P}}_{\theta,P_{\theta,\eta}}) .$$

And if this function is non–zero the efficient information for estimating θ is given by its variance. Then we have three alternatives. If we can establish

the existence of approximately least favourable submodels and a (pseudo–) likelihood function exists, then we are basically in the same situation as above with maximum likelihood estimation, except Theorem 10 has to be substituted by Theorem 11. If the efficient score function is on closed form and a preliminary estimator of the nuisance parameter is available, then we could use the efficient score estimator from Subsection 5.1. And if everything else fails, we always have the estimating equation approach in Subsection 5.3, where we probably lose some efficiency.

B Techniques

In this section we review some of the techniques that are frequently used in semiparametric models. First we consider results for Banach and Hilbert spaces. Secondly, we list some numerical methods that can be utilized in computing infinite dimensional estimators. Finally, we discuss some technical issues that are useful when working with semiparametric models.

B.1 Banach and Hilbert Spaces

A *Banach space* is an abstract linear space equipped with a norm $\| \ \|$, where every Cauchy sequence has a limit point in that space, i.e. a complete normed linear space. A *Hilbert space* is an abstract linear space with an inner product $\langle \cdot, \cdot \rangle$, where every Cauchy sequence with respect to the norm $\|x\| = \sqrt{\langle x, x \rangle}$ is convergent, i.e. a complete inner product space.

For a Hilbert space $\mathbb{H}$, a convex and closed subset $C \subset \mathbb{H}$, and every $g \in \mathbb{H}$ there exists a unique projection $\Pi(g \mid C)$ of g on C, which minimizes the distance $c \mapsto \|g - c\|$ over C. If C is a closed, linear subspace, then the projection can be found by the equation

$$\langle g - \Pi(g \mid C), c \rangle \ = \ 0 \,, \quad \text{for every } c \in C \,.$$

For every Banach space $\mathbb{B}$ there exists a *dual space* $\mathbb{B}^*$, which is the set of all continuous, linear maps $b^* : \mathbb{B} \mapsto I\!R$. The Riesz Representation Theorem states that continuous, linear maps on a Hilbert space are of the form

$$h \mapsto \langle h, h^* \rangle \,,$$

for some $h^* \in \mathbb{H}$. Hence, we may identify the dual space of any Hilbert space with the Hilbert space itself, $\mathbb{H}^* \equiv \mathbb{H}$.

A linear map[‡‡] $A : \mathbb{B}_1 \mapsto \mathbb{B}_2$ between two Banach spaces is continuous if and only if $\|Ab_1\|_2 \leq c\|b_1\|_1$ for every $b_1 \in \mathbb{B}_1$ and some number c. To

[‡‡] Also called an *operator*.

such an operator there exists an *adjoint* operator $A^* : \mathbb{B}_2^* \mapsto \mathbb{B}_1^*$ given by $(A^* b_2^*) b_1 = b_2^*(A b_1)$ or $A^* : b_2^* \mapsto \{(A^* b_2^*) : b_1 \mapsto b_2^*(A b_1)\}$. For operators between Hilbert spaces the definition of the adjoint simplifies to a map $A^* : \mathbb{H}_2 \mapsto \mathbb{H}_1$ satisfying

$$\langle A h_1, h_2 \rangle_2 = \langle h_1, A^* h_2 \rangle_1 , \qquad \text{for every } h_1 \in \mathbb{H}_1 , h_2 \in \mathbb{H}_2 .$$

An operator between Euclidean spaces is represented by a matrix and the transposed matrix represent the corresponding adjoint operator.

Finally, we need to know when an operator between Banach spaces is continuously invertible. In Euclidean spaces a one–to–one linear operator is invertible with continuous inverse. This is not automatically the case in Banach spaces. A one–to–one continuous, linear operator A has a continuous inverse if and only if the range $R(A) = \{A b_1 \mid b_1 \in \mathbb{B}_1\}$ is closed.

If $A^* A : \mathbb{H} \mapsto \mathbb{H}$ is continuously invertible, then the projection on the range of A in $\mathbb{H}$ is given by $\Pi(\cdot \mid R(A)) = A(A^* A)^{-1} A^* : \mathbb{H} \mapsto R(A)$.

B.2 Numerical Methods

The numerical methods used to obtain estimators of infinite dimensional parameters differ from case to case. In the Cox model, Example 1, simple analytic considerations simplify the problem to an estimation problem over $\{\Lambda(t_i), \beta \mid i = 1, \ldots, n\}$ of dimension $d + n$, where n is the sample size and d is the dimension of the interest parameter. Then standard maximization procedures yield the MLE.

For mixture models (cf. Example 3) we have the VDM–algorithm from Lindsay (1983a,b), who reformulates the problem as a maximization over a convex set in $\mathbb{R}^n$. The idea is, for a given estimator of the mixture distribution, G_m, to find a new support point in the "direction" of which the likelihood has the steepest derivative and then in the second step to adjust the weights between G_m and the new support point. The method has been improved by several authors. Probably, the ISDM–algorithm is the most efficient version, because it makes use of all the local maxima in the first step (which have to be identified anyway) and because it does not have to keep track of all the support point visited, see Lesperance and Kalbfleisch (1992) and the references therein.

The EM–algorithm has been proposed and used in numerous models in survival analysis. For a discussion of the EM–algorithm in the interval censored model, Example 15, see Groeneboom and Wellner (1992), and in frailty models, Example 16, see Parner (1997). The EM–algorithm often corresponds to a "self-consistency" equation where the MLE is a fixed point.

The idea of the Iterative Convex Minorant (ICM) algorithm proposed in Groeneboom (1991) goes back to Grenander (1956), who showed that the MLE of a decreasing density is the density corresponding to the smallest concave distribution function larger than the empirical distribution function. This method has been applied to the interval censoring problem, Example 15, and to convolution models, Example 6, where the random variable Y has a decreasing density. For an introduction to the ICM–algorithm see Groeneboom (1996).

Wellner and Zhan (1997) propose a hybrid algorithm for censored survival data. The algorithm alternates between the EM– and the ICM–algorithm and thereby rapidly approaches the optimum.

B.3 Useful Remarks

There are two different definitions in the literature of a least favourable submodel. The monograph of Bickel et al. (1993) defines such a submodel of a semiparametric model as a regular finite dimensional model with Cramér–Rao bound equal to the inverse of the semiparametric information bound, i.e. there exists a submodel, denoted B, such that $I_B = \tilde{I}$. Alternatively, Aad van der Vaart defines a least favourable submodel as a regular finite dimensional model such that the ordinary score function in the submodel equals the efficient score function for the semiparametric model, i.e. there exists a submodel, denoted V, such that $\dot{\ell}_V = \tilde{\ell}$ pointwise. From Lemma 4 the latter definition implies the first. In a semiparametric model with a decomposed parameter $\psi = (\theta, \eta)$ we expect that the score function in the submodel B can be written as $\dot{\ell}_B = \dot{\ell}_\theta - g$, where g is the nuisance score function. Then we have that

$$P(\dot{\ell}_B - \dot{\ell}_V)^2 = P\dot{\ell}_B^2 + P\dot{\ell}_V^2 - 2P\dot{\ell}_B\dot{\ell}_V = \tilde{I} + \tilde{I} - 2\tilde{I} = 0\,.$$

The second equality follows since the efficient score function or $\dot{\ell}_V$ is orthogonal to any nuisance score and the inner product with $\dot{\ell}_\theta$ is equal to the efficient information. For an interest parameter of dimension d these considerations hold for each coordinate. Thus the two definitions are congruent in the sense that $I_B = \tilde{I} = I_V$ and $\dot{\ell}_B = \dot{\ell}_V$ almost surely. In the present account of semiparametric models we have used the second definition.

The next remark concerns the efficiency bound and the tangent set. Suppose that we have a candidate T for the full tangent set $\dot{\mathcal{P}}$ with $T \subset \dot{\mathcal{P}}$. Instead of proving the reverse inclusion, which might be difficult, we may ignore the problem in the following case. Suppose $\hat{\vartheta}$ is a regular estimator of the parameter $\vartheta(P)$ with a Gaussian limit distribution and asymptotic

variance Σ. Then the convolution theorem, Theorem 3, implies that

$$\Sigma \geq \tilde{I}^{-1} = \|\Pi(\dot{\vartheta} \mid \overline{span}\dot{\mathcal{P}})\|^2 \geq \|\Pi(\dot{\vartheta} \mid \overline{span}T)\|^2 .$$

If the estimator $\hat{\vartheta}$ achieves the bound given by T, i.e. $\Sigma = \|\Pi(\dot{\vartheta} \mid \overline{span}T)\|^2$, then the inequalities above are identities and $T = \dot{\mathcal{P}}$.

Finally we provide a proof of the naive estimator of the efficient information in Lemma 12. Probably this is known, but I have not seen a proof before.

Proof of Lemma 12: For any $h \in I\!R^d$ fixed the class of real functions $h\mathcal{F} = \{h^\top \tilde{\ell}_{\theta,\eta} \mid (\theta,\eta) \in U\}$ is P–Donsker. From Lemma 2.10.14 in van der Vaart and Wellner (1996) the Donsker property on $h\mathcal{F}$ implies that the class $(h\mathcal{F})^2 = \{h^\top \tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top h \mid (\theta,\eta) \in U\}$ is Glivenko-Cantelli, i.e. $\sup_{f\in(h\mathcal{F})^2} |(I\!P_n - P_0)f| \to 0$. Together with (23) this yields that

$$\begin{aligned} h^\top \left(I\!P_n \tilde{\ell}_{\hat{\theta},\hat{\eta}}\tilde{\ell}_{\hat{\theta},\hat{\eta}}^\top\right) h - h^\top \tilde{I}_{\theta_0,\eta_0} h &= (I\!P_n - P_0) h^\top \tilde{\ell}_{\hat{\theta},\hat{\eta}}\tilde{\ell}_{\hat{\theta},\hat{\eta}}^\top h \\ &\quad + h^\top P_0(\tilde{\ell}_{\hat{\theta},\hat{\eta}}\tilde{\ell}_{\hat{\theta},\hat{\eta}}^\top - \tilde{\ell}_{\theta_0,\eta_0}\tilde{\ell}_{\theta_0,\eta_0}^\top) h \xrightarrow{P} 0 . \end{aligned}$$

Applying this with $h = e_i$, e_j and $e_i + e_j$, where e_i is the ith unit vector in $I\!R^d$, we obtain (24) for each entry $i, j = 1, \ldots, d$. □

C Empirical Processes

Among the useful tools in modern semiparametric theory the methods of empirical processes call upon attention. From a semiparametric point of view we study such processes for two reasons, either because of their usefulness, which is apparent in this thesis, or for the technical reason that the methods handle objects that are not measurable, which frequently occurs, and hence substitute the usual convergence theory. We refer to van der Vaart and Wellner (1996) for a readable textbook on the theory. The reader is probably familiar with a stochastic process of the form $\{X_t\}_{t\in I}$ for an interval $I \subset I\!R$. Here we consider a different type of stochastic processes, where we presume more structure of the random variables for fixed index but the index set is a collection of functions rather than a simple interval on the real line.

Let $X_1, \ldots, X_n$ be an i.i.d. sample on the measure space $(\mathcal{X}, \mathcal{A})$ with common distribution P. We study the empirical process

$$\mathbb{G}_n[f] = \sqrt{n}\,(I\!P_n - P)[f] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f(X_i) - Pf) ,$$

where $\mathbb{P}_n$ is the empirical distribution from the i.i.d. sample. The index f is an element from the class $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$. The randomness (or dependence on ω) is through the empirical measure $\mathbb{P}_n$. A simple example is to take $\mathcal{F} = \left\{\mathbf{1}_{]-\infty;t]} \mid t \in \mathbb{R}\right\}$, the collection of indicator functions. Then $\mathbb{P}_n$ becomes the empirical distribution function $\mathbb{F}_n$. Our goal is to give conditions under which $\mathbb{G}_n$ converges in distribution as a process. The space $l^\infty(\mathcal{F})$ is defined as the set of all uniformly bounded, real functions on $\mathcal{F}$. Under suitable conditions on $\mathcal{F}$ the empirical process $\mathbb{G}_n$ belongs to $l^\infty(\mathcal{F})$. E.g. assume that there exists a finite function F such that $|f(x)| \leq F(x)$ for all $f \in \mathcal{F}$ and all $x \in \mathcal{X}$. Recall that $\mathbb{Z}_n$ converges in distribution to $\mathbb{Z}$ in $l^\infty(\mathcal{F})$, if for any function h, which is continuous and bounded and takes as argument an element in $l^\infty(\mathcal{F})$, we have that

$$P^*h(\mathbb{Z}_n) \to Ph(\mathbb{Z})\,.$$

This is weak convergence in the sense of Hoffmann-Jørgensen and Dudley and we denote it by $\mathbb{Z}_n \Rightarrow \mathbb{Z}$ in $l^\infty(\mathcal{F})$.

Definition 17 *A collection of functions $\mathcal{F}$ is called a P–Glivenko–Cantelli class if the uniform version of the law of large numbers holds in outer probability or outer almost surely, i.e. if*

$$\sup_{f\in\mathcal{F}} |\mathbb{P}_n f - Pf| \to 0\,.$$

Definition 18 *A collection of functions $\mathcal{F}$ is called a P–Donsker class if there exists a tight Borel measurable element $\mathbb{G}_P$ in $l^\infty(\mathcal{F})$ such that*

$$\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P) \Rightarrow \mathbb{G}_P \qquad \text{in } l^\infty(\mathcal{F})$$

A collection of functions is P–Donsker if the sample paths of the empirical process are sufficiently continuous in the index. For a probability measure P define the seminorm $\rho_P(f)$ by

$$\rho_P(f) = \sqrt{P(f - Pf)^2} = \|f - Pf\|_{L^2(P)}\,. \tag{44}$$

A set $\mathcal{F}$ is ρ_P–totally bounded if for every $\epsilon > 0$ it can be covered with finitely many balls of ρ_P–radius ϵ. This is equivalent to the closure of the space being compact.

Theorem 19 *A class $\mathcal{F}$ of measurable functions is P–Donsker if and only if the following two conditions are met*

(i) The empirical process is asymptotically continuous: for every $\epsilon > 0$,

$$\lim_{\delta \downarrow 0} \limsup_{n \to \infty} P^* \left(\sup_{\rho_P(f-g)<\delta} \left| \mathbb{G}_n[f] - \mathbb{G}_n[g] \right| > \epsilon \right) = 0 \,. \tag{45}$$

(ii) $\mathcal{F}$ *is* ρ_P*–totally bounded.*

For more useful theorems to prove the Donsker property in practical situations we need covering numbers and entropy conditions. These theorems state that if the class $\mathcal{F}$ can be covered with a sufficiently small number of balls then the class is Donsker. Two useful tools here are the concept of VC-classes and smoothness of functions in $\mathcal{F}$. In particular, chapters 2.5 through 2.7 of van der Vaart and Wellner (1996) give such results. The usefulness of Donsker theorems is clear from this overview of semiparametric models.

References

Aalen, O. (1978) Nonparametric inference for a family of counting processes. *Ann. Statist.*, **6**, 701–726.

Ai, C. (1997) A semiparametric maximum likelihood estimator. *Econometrica*, **65**, 933–963.

Amari, S.i. and Kawanabe, M. (1997) Information geometry of estimating functions in semi-parametric statistical models. *Bernoulli*, **3**, 29–54.

Andersen, P.K. and Gill, R.D. (1982) Cox's regression model for counting processes: a large sample study. *Ann. Statist.*, **10**, 1100–1120.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993) *Statistical Models Based on Counting Processes*. New York: Springer Verlag.

Bailey, K.R. (1979) *The general maximum-likelihood approach to the Cox regression model*. Ph.D. thesis, University of Chicago.

Bailey, K.R. (1984) Asymptotic equivalence between the Cox estimator and the general ML estimators of regression and survival parameters in the Cox model. *Ann. Statist.*, **12**, 730–736.

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.

Bickel, P.J. (1982) On adaptive estimation. *Ann. Statist.*, **10**, 647–671.

Bickel, P.J., Klaassen, C.A., Ritov, Y. and Wellner, J.A. (1993) *Efficient and adaptive estimation for semiparametric models*. Baltimore: John Hopkins.

Chamberlain, G. (1986) Asymptotic efficiency in semiparametric models with censoring. *J. Econometrics*, **32**, 189–218.

Chamberlain, G. (1987) Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, **34**, 305–334.

Chen, H. (1995) Asymptotically efficient estimation in semiparametric generalized linear models. *Ann. Statist.*, **23**, 1102–1129.

Choi, S., Hall, W.J. and Schick, A. (1996) Asymptotically uniformly most powerful tests in parametric and semiparametric models. *Ann. Statist.*, **24**, 841–861.

Cox, D.R. (1972) Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34**, 187–220. (with discussion).

Cox, D.R. (1975) Partial likelihood. *Biometrika*, **62**, 269–276.

Cuzick, J. (1992) Efficient estimates in semiparametric additive regression models with unknown error distribution. *Ann. Statist.*, **20**, 1129–1136.

Gill, R.D. (1984) Understanding Cox's regression model: a martingale approach. *J. Amer. Statist. Assoc.*, **79**, 441–447.

Gill, R.D. (1989) Non- and semi-parametric maximum likelihood estimators and the von Mises method. I. *Scand. J. Statist.*, **16**, 97–128. With a discussion by J. A. Wellner and J. Præstgaard and a reply by the author.

Gill, R.D. and van der Vaart, A.W. (1993) Non- and semi-parametric maximum likelihood estimators and the von Mises method. II. *Scand. J. Statist.*, **20**, 271–288.

Goto, F. (1996) Achieving semiparametric efficiency bounds in left-censored duration models. *Econometrica*, **64**, 439–442.

Gourieroux, C. and Monfort, A. (1995) *Statistics and Econometric Models*, vol. 1 and 2. Cambridge University Press.

Grenander, U. (1956) On the theory of mortality measurement. II. *Skand. Aktuarietidskr.*, **39**, 125–153 (1957).

Groeneboom, P. (1991) Nonparametric maximum likelihood estimators for interval censoring and deconvolution. Technical Report 378, Department of Statistics, Stanford University.

Groeneboom, P. (1996) Lectures on inverse problems. In *Lectures on probability theory and statistics (Saint-Flour, 1994)*, pp. 67–164, Berlin: Springer.

Groeneboom, P. and Wellner, J.A. (1992) *Information bounds and nonparametric maximum likelihood estimation*. Basel: Birkhäuser Verlag.

Gu, M. and Zheng, Z.K. (1993) On the Bartlett adjustment for the partial likelihood ratio test in the Cox regression model. *Statist. Sinica*, **3**, 543–555.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Heckman, J. and Singer, B. (1984) A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271–320.

Horowitz, J.L. (1998) *Semiparametric methods in econometrics*. New York: Springer-Verlag.

Huang, J. (1996) Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.*, **24**, 540–568.

Huang, J. (1998) A Least Squares Approach to Consistent Information Estimation in Semiparametric Models. Submitted, available from http://-www.stat.uiowa.edu/∼jian/publication.html.

Huang, J. and Wellner, J.A. (1995) Efficient estimation for the proportional hazards model with "case 2" interval censoring. Technical report 290, Dept. Statist., University of Washington.

Huber, P.J. (1967) The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fith Berk. Symp. Math. Statist. and Prob.*, vol. 1, pp. 221–233. Univ. California Press, Berkeley.

Hunsberger, S. (1994) Semiparametric regression in likelihood-based models. *J. Amer. Statist. Assoc.*, **89**, 1354–1365.

Ishwaran, H. (1996a) Identifiability and rates of estimation for scale parameters in location mixture models. *Ann. Statist.*, **24**, 1560–1571.

Ishwaran, H. (1996b) Uniform rates of estimation in the semiparametric Weibull mixture model. *Ann. Statist.*, **24**, 1572–1585.

Jacobsen, M. (1984) Maximum likelihood estimation in the multiplicative intensity model: a survey. *Internat. Statist. Rev.*, **52**, 193–207.

Jacobsen, M. (1989) Existence and unicity of MLEs in discrete exponential family distributions. *Scand. J. Statist.*, **16**, 335–349.

Kiefer, J. and Wolfowitz, J. (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, **27**, 887–906.

Kim, J. and Pollard, D. (1990) Cube root asymptotics. *Ann. Statist.*, **18**, 191–219.

Klaassen, C.A.J. and Wellner, J.A. (1997) Efficient estimation in the bivariate normal copula model: normal margins are least favourable. *Bernoulli*, **3**, 55–77.

Korsholm, L. (1998a) The GMM estimator versus the semiparametric efficient score estimator under conditional moment restrictions. Unpublished.

Korsholm, L. (1998b) Likelihood ratio test in the correlated gamma-frailty model. Research Reports 396, Dept. Theor. Statist., Aarhus University.

Lesperance, M.L. and Kalbfleisch, J.D. (1992) An Algorithm for Computing the Nonparametric MLE of a Mixing Distribution. *J. Amer. Statist. Assoc.*, **87**, 120–126.

Levit, B.J. (1978) Infinite-dimensional informational bounds. *Teor. Verojatnost. i Primenen.*, **23**, 388–394.

Lindsay, B.G. (1983a) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

Lindsay, B.G. (1983b) The geometry of mixture likelihoods. II: The exponential family. *Ann. Statist.*, **11**, 783–792.

Lindsay, B.G. and Lesperance, M.L. (1995) A review of semiparametric mixture models. *J. Statist. Plann. Inference*, **47**, 29–39.

Mammen, E. and van de Geer, S. (1997a) Locally adaptive regression splines. *Ann. Statist.*, **25**, 387–413.

Mammen, E. and van de Geer, S. (1997b) Penalized quasi-likelihood estimation in partial linear models. *Ann. Statist.* To appear.

Manski, C.F. (1985) Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator. *J. Econometrics*, **27**, 313–333.

Müller, H.G. and Zhao, P.L. (1995) On a semiparametric variance function model and a test for heteroscedasticity. *Ann. Statist.*, **23**, 946–967.

Murphy, S.A. (1993) Testing for a time dependent coefficient in Cox's regression model. *Scand. J. Statist.*, **20**, 35–50.

Murphy, S.A. and van der Vaart, A.W. (1996) Likelihood Inference in the Errors-in-Variables Model. *Journal of Multivariate Analysis*, **59**, 81–108.

Murphy, S.A. and van der Vaart, A.W. (1997a) Observed information in semiparametric models. Unpublished.

Murphy, S.A. and van der Vaart, A.W. (1997b) On profile likelihood. Unpublished, http://www.stat.lsa.umich.edu/~samurphy/research.html.

Murphy, S.A. and van der Vaart, A.W. (1997c) Semiparametric likelihood ratio inference. *Ann. Statist.*, **25**, 1471–1509.

Murphy, S.A., Rossini, A.J. and van der Vaart, A.W. (1997) Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.*, **92**, 968–976.

Newey, W.K. (1990) Semiparametric efficiency bounds. *Journal of applied econometrics*, **5**, 99–135.

Newey, W.K. (1993) Efficient estimation of models with conditional moment restrictions. In *Econometrics*, vol. 11 of *Handbook of Statistics*, pp. 419–454, Amsterdam: North-Holland.

Nielsen, G.G., Gill, R.D., Andersen, P.K. and Sørensen, T.I.A. (1992) A counting process approach to maximum likelihood estimation in frailty models. *Scand. J. Statist.*, **19**, 25–44.

Parner, E. (1997) *Inference in Semiparametric Frailty Models*. Ph.d. dissertation, Dept. Theor. Statist., Aarhus University.

Parner, E. (1998) Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, **26**, 183–214.

Pfanzagl, J. (1990) *Estimation in Semiparametric Models*. Lecture Notes in Statistics, 63. Berlin: Springer-Verlag.

Pollard, D. (1985) New ways to prove central limit theorems. *Econometric Theory*, **1**, 295–314.

Ritov, Y. and Bickel, P.J. (1990) Achieving information bounds in non and semiparametric models. *Ann. Statist*, **18**, 925–938.

Robinson, P.M. (1988) Semiparametric Econometrics: A Survey. *J. Appl. Econometrics*, **3**, 35–51.

Roeder, K. (1992) Semiparametric estimation of normal mixture densities. *Ann. Statist.*, **20**, 929–943.

Rotnitzky, A. and Robins, J.M. (1995) Semi-parametric estimation of models for means and covariances in the presence of missing data. *Scand. J. Statist.*, **22**, 323–333.

Schick, A. (1986) On asymptotically efficient estimation in semiparametric models. *Ann. Statist.*, **14**, 1139–1151.

Severini, T.A. and Staniswalis, J.G. (1994) Quasi-likelihood estimation in semiparametric models. *J. Amer. Statist. Assoc.*, **89**, 501–511.

Severini, T.A. and Wong, W.H. (1992) Profile likelihood and conditionally parametric models. *Ann. Statist.*, **20**, 1768–1802.

Stein, C. (1956) Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pp. 187–195. Berkeley and Los Angeles: University of California Press.

van de Geer, S. (1990) Estimating a regression function. *Ann. Statist.*, **18**, 907–924.

van de Geer, S. and Wegkamp, M. (1996) Consistency for the least squares estimator in nonparametric regression. *Ann. Statist.*, **24**, 2513–2523.

van der Laan, M.J. and Robins, J.M. (1998) Locally efficient estimation with current status data and time-dependent covariates. *J. Amer. Statist. Assoc.*, **93**, 693–701.

van der Vaart, A.W. (1988) Estimating a real parameter in a class of semiparametric models. *Ann. Statist.*, **16**, 1450–1474.

van der Vaart, A.W. (1989) On the asymptotic information bound. *Ann. Statist.*, **17**, 1487–1500.

van der Vaart, A.W. (1991a) An asymptotic representation theorem. *International Statistical Review*, **59**, 97–121.

van der Vaart, A.W. (1991b) On differentiable functionals. *Ann. Statist.*, **19**, 178–204.

van der Vaart, A.W. (1994) Maximum likelihood estimation with partially censored data. *Ann. Statist.*, **22**, 1896–1916.

van der Vaart, A.W. (1995) Efficiency of infinite dimensional M-estimators. *Statist. Neerlandica*, **49**, 9–30.

van der Vaart, A.W. (1996) Efficient maximum likelihood estimation in semiparametric mixture models. *Ann. Statist.*, **24**, 862–878.

van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge: Cambridge University Press.

van der Vaart, A.W. and Wellner, J.A. (1992) Existence and consistency of maximum likelihood in upgraded mixture models. *J. Multivariate Anal.*, **43**, 133–146.

van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. New York: Springer-Verlag.

Wang, N. and Ruppert, D. (1996) Estimation of regression parameters in a semiparametric transformation model. *J. Statist. Plann. Inference*, **52**, 331–351.

Wellner, J.A. (1985) Semiparametric models: progress and problems. Proceedings of the 45th session of the International Statistical Institute, Vol. 4 (Amsterdam, 1985). *Bull. Inst. Internat. Statist.*, **51**, no. 4, No. 23.1, 20 pp.; Vol. V: pp. 175–180. With discussion.

Wellner, J.A. and Zhan, Y. (1996) Bootstrapping Z-estimators. Technical Report 308, University of Washington.

Wellner, J.A. and Zhan, Y. (1997) A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *J. Amer. Statist. Assoc.*, **92**, 945–959.