25

Semiparametric Models

This chapter is concerned with statistical models that are indexed by infinite-dimensional parameters. It gives an introduction to the theory of asymptotic efficiency, and discusses methods of estimation and testing.

25.1 Introduction

Semiparametric models are statistical models in which the parameter is not a Euclidean vector but ranges over an "infinite-dimensional" parameter set. A different name is "model with a large parameter space." In the situation in which the observations consist of a random sample from a common distribution P, the *model* is simply the set $\mathcal{P}$ of all possible values of P: a collection of probability measures on the sample space. The simplest type of infinite-dimensional model is the *nonparametric model*, in which we observe a random sample from a completely unknown distribution. Then $\mathcal{P}$ is the collection of all probability measures on the sample space, and, as we shall see and as is intuitively clear, the empirical distribution is an asymptotically efficient estimator for the underlying distribution. More interesting are the intermediate models, which are not "nicely" parametrized by a Euclidean parameter, as are the standard classical models, but do restrict the distribution in an important way. Such models are often parametrized by infinite-dimensional parameters, such as distribution functions or densities, that express the structure under study. Many aspects of these parameters are estimable by the same order of accuracy as classical parameters, and efficient estimators are asymptotically normal. In particular, the model may have a natural parametrization $(\theta, \eta) \mapsto P_{\theta,\eta}$, where θ is a Euclidean parameter and η runs through a nonparametric class of distributions, or some other infinite-dimensional set. This gives a semiparametric model in the strict sense, in which we aim at estimating θ and consider η as a *nuisance parameter*. More generally, we focus on estimating the value $\psi(P)$ of some function $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ on the model.

In this chapter we extend the theory of asymptotic efficiency, as developed in Chapters 8 and 15, from parametric to semiparametric models and discuss some methods of estimation and testing. Although the efficiency theory (lower bounds) is fairly complete, there are still important holes in the estimation theory. In particular, the extent to which the lower bounds are sharp is unclear. We limit ourselves to parameters that are $\sqrt{n}$-estimable, although in most semiparametric models there are many "irregular" parameters of interest that are outside the scope of "asymptotically normal" theory. Semiparametric testing theory has

little more to offer than the comforting conclusion that tests based on efficient estimators are efficient. Thus, we shall be brief about it.

We conclude this introduction with a list of examples that shows the scope of semiparametric theory. In this description, X denotes a typical observation. Random vectors Y, Z, e, and f are used to describe the model but are not necessarily observed. The parameters θ and ν are always Euclidean.

25.1 ***Example (Regression).*** Let Z and e be independent random vectors and suppose that $Y = \mu_\theta(Z) + \sigma_\theta(Z)e$ for functions μ_θ and σ_θ that are known up to θ. The observation is the pair $X = (Y, Z)$. If the distribution of e is known to belong to a certain parametric family, such as the family of $N(0, \sigma^2)$-distributions, and the independent variables Z are modeled as constants, then this is just a classical regression model, allowing for heteroscedasticity. Semiparametric versions are obtained by letting the distribution of e range over all distributions on the real line with mean zero, or, alternatively, over all distributions that are symmetric about zero. □

25.2 ***Example (Projection pursuit regression).*** Let Z and e be independent random vectors and let $Y = \eta(\theta^T Z) + e$ for a function η ranging over a set of (smooth) functions, and e having an $N(0, \sigma^2)$-distribution. In this model θ and η are confounded, but the direction of θ is estimable up to its sign. This type of regression model is also known as a *single-index* model and is intermediate between the classical regression model in which η is known and the nonparametric regression model $Y = \eta(Z) + e$ with η an unknown smooth function. An extension is to let the error distribution range over an infinite-dimensional set as well. □

25.3 ***Example (Logistic regression).*** Given a vector Z, let the random variable Y take the value 1 with probability $1/(1 + e^{-r(Z)})$ and be 0 otherwise. Let $Z = (Z_1, Z_2)$, and let the function r be of the form $r(z_1, z_2) = \eta(z_1) + \theta^T z_2$. Observed is the pair $X = (Y, Z)$. This is a semiparametric version of the logistic regression model, in which the response is allowed to be nonlinear in part of the covariate. □

25.4 ***Example (Paired exponential).*** Given an unobservable variable Z with completely unknown distribution, let $X = (X_1, X_2)$ be a vector of independent exponentially distributed random variables with parameters Z and $Z\theta$. The interest is in the ratio θ of the conditional hazard rates of X_1 and X_2. Modeling the "baseline hazard" Z as a random variable rather than as an unknown constant allows for heterogeneity in the population of all pairs (X_1, X_2), and hence ensures a much better fit than the two-dimensional parametric model in which the value z is a parameter that is the same for every observation. □

25.5 ***Example (Errors-in-variables).*** The observation is a pair $X = (X_1, X_2)$, where $X_1 = Z + e$ and $X_2 = \alpha + \beta Z + f$ for a bivariate normal vector (e, f) with mean zero and unknown covariance matrix. Thus X_2 is a linear regression on a variable Z that is observed with error. The distribution of Z is unknown. □

25.6 ***Example (Transformation regression).*** Suppose that $X = (Y, Z)$, where the random vectors Y and Z are known to satisfy $\eta(Y) = \theta^T Z + e$ for an unknown map η and independent random vectors e and Z with known or parametrically specified distributions.

The transformation η ranges over an infinite-dimensional set, for instance the set of all monotone functions. □

25.7 ***Example (Cox).*** The observation is a pair $X = (T, Z)$ of a "survival time" T and a covariate Z. The distribution of Z is unknown and the conditional hazard function of T given Z is of the form $e^{\theta^T Z}\lambda(t)$ for λ being a completely unknown hazard function. The parameter θ has an interesting interpretation in terms of a ratio of hazards. For instance, if the ith coordinate Z_i of the covariate is a 0-1 variable then e^{θ_i} can be interpreted as the ratio of the hazards of two individuals whose covariates are $Z_i = 1$ and $Z_i = 0$, respectively, but who are identical otherwise. □

25.8 ***Example (Copula).*** The observation X is two-dimensional with cumulative distribution function of the form $C_\theta\big(G_1(x_1), G_2(x_2)\big)$, for a parametric family of cumulative distribution functions C_θ on the unit square with uniform marginals. The marginal distribution functions G_i may both be completely unknown or one may be known. □

25.9 ***Example (Frailty).*** Two survival times Y_1 and Y_2 are conditionally independent given variables (Z, W) with hazard function of the form $We^{\theta^T Z}\lambda(y)$. The random variable W is not observed, possesses a gamma(ν, ν) distribution, and is independent of the variable Z which possesses a completely unknown distribution. The observation is $X = (Y_1, Y_2, Z)$. The variable W can be considered an unobserved regression variable in a Cox model. □

25.10 ***Example (Random censoring).*** A "time of death" T is observed only if death occurs before the time C of a "censoring event" that is independent of T; otherwise C is observed. A typical observation X is a pair of a survival time and a 0-1 variable and is distributed as $\big(T \wedge C, 1\{T \le C\}\big)$. The distributions of T and C may be completely unknown. □

25.11 ***Example (Interval censoring).*** A "death" that occurs at time T is only observed to have taken place or not at a known "check-up time" C. The observation is $X = \big(C, 1\{T \le C\}\big)$, and T and C are assumed independent with completely unknown or partially specified distributions. □

25.12 ***Example (Truncation).*** A variable of interest Y is observed only if it is larger than a censoring variable C that is independent of Y; otherwise, nothing is observed. A typical observation $X = (X_1, X_2)$ is distributed according to the conditional distribution of (Y, C) given that $Y > C$. The distributions of Y and C may be completely unknown. □

25.2 Banach and Hilbert Spaces

In this section we recall some facts concerning Banach spaces and, in particular, Hilbert spaces, which play an important role in this chapter.

Given a probality space $(\mathcal{X}, \mathcal{A}, P)$, we denote by $L_2(P)$ the set of measurable functions $g : \mathcal{X} \mapsto \mathbb{R}$ with $Pg^2 = \int g^2\, dP < \infty$, where almost surely equal functions are identified. This is a *Hilbert space*, a complete inner-product space, relative to the inner product

and norm

$$\langle g_1, g_2 \rangle = Pg_1 g_2, \qquad \|g\| = \sqrt{Pg^2}.$$

Given a Hilbert space $\mathbb{H}$, the *projection lemma* asserts that for every $g \in \mathbb{H}$ and convex, closed subset $C \subset \mathbb{H}$, there exists a unique element $\Pi g \in C$ that minimizes $c \mapsto \|g - c\|$ over C. If C is a closed, linear subspace, then the projection Πg can be characterized by the orthogonality relationship

$$\langle g - \Pi g, c \rangle = 0, \qquad \text{every } c \in C.$$

The proof is the same as in Chapter 11. If $C_1 \subset C_2$ are two nested, closed subspaces, then the projection onto C_1 can be found by first projecting onto C_2 and next onto C_1. Two subsets C_1 and C_2 are *orthogonal*, notation $C \perp C_2$, if $\langle c_1, c_2 \rangle = 0$ for every pair of $c_i \in C_i$. The projection onto the sum of two orthogonal closed subspaces is the sum of the projections. The *orthocomplement* $C^\perp$ of a set C is the set of all $g \perp C$.

A *Banach space* is a complete, normed space. The *dual space* $\mathbb{B}^*$ of a Banach space $\mathbb{B}$ is the set of all continuous, linear maps $b^* : \mathbb{B} \mapsto \mathbb{R}$. Equivalently, all linear maps such that $|b^*(b)| \leq \|b^*\| \, \|b\|$ for every $b \in \mathbb{B}$ and some number $\|b^*\|$. The smallest number with this property is denoted by $\|b^*\|$ and defines a norm on the dual space. According to the *Riesz representation theorem* for Hilbert spaces, the dual of a Hilbert space $\mathbb{H}$ consists of all maps

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

where h^* ranges over $\mathbb{H}$. Thus, in this case the dual space $\mathbb{H}^*$ can be identified with the space $\mathbb{H}$ itself. This identification is an isometry by the Cauchy-Schwarz inequality $|\langle h, h^* \rangle| \leq \|h\| \, \|h^*\|$.

A linear map $A : \mathbb{B}_1 \mapsto \mathbb{B}_2$ from one Banach space into another is continuous if and only if $\|Ab_1\|_2 \leq \|A\| \, \|b_1\|$ for every $b_1 \in \mathbb{B}_1$ and some number $\|A\|$. The smallest number with this property is denoted by $\|A\|$ and defines a norm on the set of all continuous, linear maps, also called *operators*, from $\mathbb{B}_1$ into $\mathbb{B}_2$. Continuous, linear operators are also called "bounded," even though they are only bounded on bounded sets. To every continuous, linear operator $A : \mathbb{B}_1 \mapsto \mathbb{B}_2$ exists an *adjoint map* $A^* : \mathbb{B}_2^* \mapsto \mathbb{B}_1^*$ defined by $(A^* b_2^*) b_1 = b_2^* A b_1$. This is a continuous, linear operator of the same norm $\|A^*\| = \|A\|$. For Hilbert spaces the dual space can be identified with the original space and then the adjoint of $A : \mathbb{H}_1 \mapsto \mathbb{H}_2$ is a map $A^* : \mathbb{H}_2 \mapsto \mathbb{H}_1$. It is characterized by the property

$$\langle Ah_1, h_2 \rangle_2 = \langle h_1, A^* h_2 \rangle_1, \qquad \text{every } h_1 \in \mathbb{H}_1, h_2 \in \mathbb{H}_2.$$

An operator between Euclidean spaces can be identified with a matrix, and its adjoint with the transpose. The adjoint of a restriction $A_0 : \mathbb{H}_{1,0} \subset \mathbb{H}_1 \mapsto \mathbb{H}_2$ of A is the composition $\Pi \circ A^*$ of the projection $\Pi : \mathbb{H}_1 \mapsto \mathbb{H}_{1,0}$ and the adjoint of the original A.

The range $\mathrm{R}(A) = \{Ab_1 : b_1 \in \mathbb{B}_1\}$ of a continuous, linear operator is not necessarily closed. By the "bounded inverse theorem" the range of a 1-1 continuous, linear operator between Banach spaces is closed if and only if its inverse is continuous. In contrast the kernel $\mathrm{N}(A) = \{b_1 : Ab_1 = 0\}$ of a continuous, linear operator is always closed. For an operator between Hilbert spaces the relationship $\mathrm{R}(A)^\perp = \mathrm{N}(A^*)$ follows readily from the

characterization of the adjoint. The range of A is closed if and only if the range of A^* is closed if and only if the range of A^*A is closed. In that case $R(A^*) = R(A^*A)$.

If $A^*A : \mathbb{H}_1 \mapsto \mathbb{H}_1$ is continuously invertible (i.e., is 1-1 and onto with a continuous inverse), then $A(A^*A)^{-1}A^* : \mathbb{H}_2 \mapsto \mathrm{R}(A)$ is the orthogonal projection onto the range of A, as follows easily by checking the orthogonality relationship.

25.3 Tangent Spaces and Information

Suppose that we observe a random sample $X_1, \ldots, X_n$ from a distribution P that is known to belong to a set $\mathcal{P}$ of probability measures on the sample space $(\mathcal{X}, \mathcal{A})$. It is required to estimate the value $\psi(P)$ of a functional $\psi : \mathcal{P} \mapsto \mathbb{R}^k$. In this section we develop a notion of information for estimating $\psi(P)$ given the model $\mathcal{P}$, which extends the notion of Fisher information for parametric models.

To estimate the parameter $\psi(P)$ given the model $\mathcal{P}$ is certainly harder than to estimate this parameter given that P belongs to a submodel $\mathcal{P}_0 \subset \mathcal{P}$. For every smooth parametric submodel $\mathcal{P}_0 = \{P_\theta : \theta \in \Theta\} \subset \mathcal{P}$, we can calculate the Fisher information for estimating $\psi(P_\theta)$. Then the information for estimating $\psi(P)$ in the whole model is certainly not bigger than the infimum of the informations over all submodels. We shall simply define the information for the whole model as this infimum. A submodel for which the infimum is taken (if there is one) is called *least favorable* or a "hardest" submodel.

In most situations it suffices to consider one-dimensional submodels $\mathcal{P}_0$. These should pass through the "true" distribution P of the observations and be differentiable at P in the sense of Chapter 7 on local asymptotic normality. Thus, we consider maps $t \mapsto P_t$ from a neighborhood of $0 \in [0, \infty)$ to $\mathcal{P}$ such that, for some measurable function $g : \mathcal{X} \mapsto \mathbb{R}$,[†]

$$\int \left[\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g \, dP^{1/2} \right]^2 \to 0. \tag{25.13}$$

In other words, the parametric submodel $\{P_t : 0 < t < \varepsilon\}$ is differentiable in quadratic mean at $t = 0$ with *score function* g. Letting $t \mapsto P_t$ range over a collection of submodels, we obtain a collection of score functions, which we call a *tangent set* of the model $\mathcal{P}$ at P and denote by $\dot{\mathcal{P}}_P$. Because Ph^2 is automatically finite, the tangent space can be identified with a subset of $L_2(P)$, up to equivalence classes. The tangent set is often a linear space, in which case we speak of a *tangent space*.

Geometrically, we may visualize the model $\mathcal{P}$, or rather the corresponding set of "square roots of measures" $dP^{1/2}$, as a subset of the unit ball of $L_2(P)$, and $\dot{\mathcal{P}}_P$, or rather the set of all objects $\frac{1}{2} g \, dP^{1/2}$, as its tangent set.

Usually, we construct the submodels $t \mapsto P_t$ such that, for every x,

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

[†] If P and every one of the measures P_t possess densities p and p_t with respect to a measure μ_t, then the expressions dP and dP_t can be replaced by p and p_t and the integral can be understood relative to μ_t (add $d\mu_t$ on the right). We use the notations dP_t and dP, because some models $\mathcal{P}$ of interest are not dominated, and the choice of μ_t is irrelevant. However, the model could be taken dominated for simplicity, and then dP_t and dP are just the densities of P_t and P.

However, the differentiability (25.13) is the correct definition for defining information, because it ensures a type of local asymptotic normality. The following lemma is proved in the same way as Theorem 7.2.

25.14 ***Lemma.*** *If the path $t \mapsto P_t$ in $\mathcal{P}$ satisfies* (25.13)*, then $Pg = 0$, $Pg^2 < \infty$, and*

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

For defining the information for estimating $\psi(P)$, only those submodels $t \mapsto P_t$ along which the parameter $t \mapsto \psi(P_t)$ is differentiable are of interest. Thus, we consider only submodels $t \mapsto P_t$ such that $t \mapsto \psi(P_t)$ is differentiable at $t = 0$. More precisely, we define $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ to be *differentiable* at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear map $\dot{\psi}_P : L_2(P) \mapsto \mathbb{R}^k$ such that for every $g \in \dot{\mathcal{P}}_P$ and a submodel $t \mapsto P_t$ with score function g,

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

This requires that the derivative of the map $t \mapsto \psi(P_t)$ exists in the ordinary sense, and also that it has a special representation. (The map $\dot{\psi}_P$ is much like a Hadamard derivative of ψ viewed as a map on the space of "square roots of measures.") Our definition is also relative to the submodels $t \mapsto P_t$, but we speak of "relative to $\dot{\mathcal{P}}_P$" for simplicity.

By the Riesz representation theorem for Hilbert spaces, the map $\dot{\psi}_P$ can always be written in the form of an inner product with a fixed vector-valued, measurable function $\tilde{\psi}_P : \mathcal{X} \mapsto \mathbb{R}^k$,

$$\dot{\psi}_P g = \langle \tilde{\psi}_P, g \rangle_P = \int \tilde{\psi}_P \, g \, dP.$$

Here the function $\tilde{\psi}_P$ is not uniquely defined by the functional ψ and the model $\mathcal{P}$, because only inner products of $\tilde{\psi}_P$ with elements of the tangent set are specified, and the tangent set does not span all of $L_2(P)$. However, it is always possible to find a candidate $\tilde{\psi}_P$ whose coordinate functions are contained in $\overline{\text{lin}}\,\dot{\mathcal{P}}_P$, the closure of the linear span of the tangent set. This function is unique and is called the *efficient influence function*. It can be found as the projection of any other "influence function" onto the closed linear span of the tangent set.

In the preceding set-up the tangent sets $\dot{\mathcal{P}}_P$ are made to depend both on the model $\mathcal{P}$ and the functional ψ. We do not always want to use the "maximal tangent set," which is the set of all score functions of differentiable submodels $t \mapsto P_t$, because the parameter ψ may not be differentiable relative to it. We consider every subset of a tangent set a tangent set itself.

The maximal tangent set is a cone: If $g \in \dot{\mathcal{P}}_P$ and $a \geq 0$, then $ag \in \dot{\mathcal{P}}_P$, because the path $t \mapsto P_{at}$ has score function ag when $t \mapsto P_t$ has score function g. It is rarely a loss of generality to assume that the tangent set we work with is a cone as well.

25.15 ***Example (Parametric model).*** Consider a parametric model with parameter θ ranging over an open subset Θ of $\mathbb{R}^m$ given by densities p_θ with respect to some measure μ. Suppose that there exists a vector-valued measurable map $\dot{\ell}_\theta$ such that, as $h \to 0$,

$$\int \left[p_{\theta+h}^{1/2} - p_\theta^{1/2} - \frac{1}{2} h^T \dot{\ell}_\theta \, p_\theta^{1/2} \right]^2 d\mu = o\big(\|h\|^2\big).$$

Then a tangent set at P_θ is given by the linear space $\{h^T\dot{\ell}_\theta : h \in \mathbb{R}^m\}$ spanned by the score functions for the coordinates of the parameter θ.

If the Fisher information matrix $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ is invertible, then every map $\chi : \Theta \mapsto \mathbb{R}^k$ that is differentiable in the ordinary sense as a map between Euclidean spaces is differentiable as a map $\psi(P_\theta) = \chi(\theta)$ on the model relative to the given tangent space. This follows because the submodel $t \mapsto P_{\theta+th}$ has score $h^T\dot{\ell}_\theta$ and

$$\frac{\partial}{\partial t}_{|t=0} \chi(\theta + th) = \dot{\chi}_\theta h = P_\theta\Big[\big(\dot{\chi}_\theta\, I_\theta^{-1}\,\dot{\ell}_\theta\big)\, h^T\dot{\ell}_\theta\Big].$$

This equation shows that the function $\tilde{\psi}_{P_\theta} = \dot{\chi}_\theta\, I_\theta^{-1}\,\dot{\ell}_\theta$ is the efficient influence function. In view of the results of Chapter 8, asymptotically efficient estimator sequences for $\chi(\theta)$ are asymptotically linear in this function, which justifies the name "efficient influence function." □

25.16 ***Example (Nonparametric models).*** Suppose that $\mathcal{P}$ consists of all probability laws on the sample space. Then a tangent set at P consists of all measurable functions g satisfying $\int g\,dP = 0$ and $\int g^2\,dP < \infty$. Because a score function necessarily has mean zero, this is the maximal tangent set.

It suffices to exhibit suitable one-dimensional submodels. For a bounded function g, consider for instance the exponential family $p_t(x) = c(t)\,\exp\big(tg(x)\big)\,p_0(x)$ or, alternatively, the model $p_t(x) = \big(1 + tg(x)\big)\,p_0(x)$. Both models have the property that, for every x,

$$g(x) = \frac{\partial}{\partial t}_{|t=0} \log p_t(x).$$

By a direct calculation or by using Lemma 7.6, we see that both models also have score function g at $t = 0$ in the L_2-sense (25.13). For an unbounded function g, these submodels are not necessarily well-defined. However, the models have the common structure $p_t(x) = c(t)\,k\big(tg(x)\big)\,p_0(x)$ for a nonnegative function k with $k(0) = k'(0) = 1$. The function $k(x) = 2(1 + e^{-2x})^{-1}$ is bounded and can be used with any g. □

25.17 ***Example (Cox model).*** The density of an observation in the Cox model takes the form

$$(t, z) \mapsto e^{-e^{\theta^T z}\Lambda(t)}\,\lambda(t)\,e^{\theta^T z}\,p_Z(z).$$

Differentiating the logarithm of this expression with respect to θ gives the score function for θ,

$$z - ze^{\theta^T z}\Lambda(t).$$

We can also insert appropriate parametric models $s \mapsto \lambda_s$ and differentiate with respect to s. If a is the derivative of $\log \lambda_s$ at $s = 0$, then the corresponding score for the model for the observation is

$$a(t) - e^{\theta^T z}\int_{[0,t]} a\,d\Lambda.$$

Finally, scores for the density p_Z are functions $b(z)$. The tangent space contains the linear span of all these functions. Note that the scores for Λ can be found as an "operator" working on functions a. □

25.18 ***Example (Transformation regression model).*** If the transformation η is increasing and e has density ϕ, then the density of the observation can be written $\phi\big(\eta(y) - \theta^T z\big)\,\eta'(y)\,p_Z(z)$. Scores for θ and η take the forms

$$-z\frac{\phi'}{\phi}\big(\eta(y) - \theta^T z\big), \qquad \frac{\phi'}{\phi}\big(\eta(y) - \theta^T z\big)a(y) + \frac{a'}{\eta'}(y),$$

where a is the derivative for η. If the distributions of e and Z are (partly) unknown, then there are additional score functions corresponding to their distributions. Again scores take the form of an operator acting on a set of functions. □

To motivate the definition of information, assume for simplicity that the parameter $\psi(P)$ is one-dimensional. The Fisher information about t in a submodel $t \mapsto P_t$ with score function g at $t = 0$ is Pg^2. Thus, the "optimal asymptotic variance" for estimating the function $t \mapsto \psi(P_t)$, evaluated at $t = 0$, is the Cramér-Rao bound

$$\frac{\big(d\psi(P_t)/dt\big)^2}{Pg^2} = \frac{\langle \tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P}.$$

The supremum of this expression over all submodels, equivalently over all elements of the tangent set, is a lower bound for estimating $\psi(P)$ given the model $\mathcal{P}$, if the "true measure" is P. This supremum can be expressed in the norm of the efficient influence function $\tilde{\psi}_P$.

25.19 ***Lemma.*** *Suppose that the functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ is differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$. Then*

$$\sup_{g \in \operatorname{lin} \dot{\mathcal{P}}_P} \frac{\langle \tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P} = P\tilde{\psi}_P^2.$$

Proof. This is a consequence of the Cauchy-Schwarz inequality $(P\tilde{\psi}_P g)^2 \le P\tilde{\psi}_P^2 Pg^2$ and the fact that, by definition, the efficient influence function $\tilde{\psi}_P$ is contained in the closure of $\operatorname{lin} \dot{\mathcal{P}}_P$. ■

Thus, the squared norm $P\tilde{\psi}_P^2$ of the efficient influence function plays the role of an "optimal asymptotic variance," just as does the expression $\dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^T$ in Chapter 8. Similar considerations (take linear combinations) show that the "optimal asymptotic covariance" for estimating a higher-dimensional parameter $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ is given by the covariance matrix $P\tilde{\psi}_P \tilde{\psi}_P^T$ of the efficient influence function.

In Chapter 8, we developed three ways to give a precise meaning to optimal asymptotic covariance: the convolution theorem, the almost-everywhere convolution theorem, and the minimax theorem. The almost-everywhere theorem uses the Lebesgue measure on the Euclidean parameter set, and does not appear to have an easy parallel for semiparametric models. On the other hand, the two other results can be generalized.

For every g in a given tangent set $\dot{\mathcal{P}}_P$, write $P_{t,g}$ for a submodel with score function g along which the function ψ is differentiable. As usual, an estimator T_n is a measurable function $T_n(X_1, \ldots, X_n)$ of the observations. An estimator sequence T_n is called *regular* at P for estimating $\psi(P)$ (relative to $\dot{\mathcal{P}}_P$) if there exists a probability measure L such that

$$\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big) \overset{P_{1/\sqrt{n},g}}{\rightsquigarrow} L, \qquad \text{every } g \in \dot{\mathcal{P}}_P.$$

25.20 ***Theorem (Convolution).*** *Let the function $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ be differentiable at P relative to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. Then the asymptotic covariance matrix of every regular sequence of estimators is bounded below by $P\tilde{\psi}_P\tilde{\psi}_P^T$. Furthermore, if $\dot{\mathcal{P}}_P$ is a convex cone, then every limit distribution L of a regular sequence of estimators can be written $L = N\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big) * M$ for some probability distribution M.*

25.21 ***Theorem (LAM).*** *Let the function $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ be differentiable at P relative to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. If $\dot{\mathcal{P}}_P$ is a convex cone, then, for any estimator sequence $\{T_n\}$ and subconvex function $\ell : \mathbb{R}^k \mapsto [0, \infty)$,*

$$\sup_I \liminf_{n\to\infty} \sup_{g\in I} \mathrm{E}_{P_{1/\sqrt{n},g}}\, \ell\Big(\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big)\Big) \geq \int \ell\, dN\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big).$$

Here the first supremum is taken over all finite subsets I of the tangent set.

Proofs. These results follow essentially by applying the corresponding theorems for parametric models to sufficiently rich finite-dimensional submodels. However, because we have defined the tangent set using one-dimensional submodels $t \mapsto P_{t,g}$, it is necessary to rework the proofs a little.

Assume first that the tangent set is a linear space, and fix an orthonormal base $g_P = (g_1, \ldots, g_m)^T$ of an arbitrary finite-dimensional subspace. For every $g \in \operatorname{lin} g_P$ select a submodel $t \mapsto P_{t,g}$ as used in the statement of the theorems. Each of the submodels $t \mapsto P_{t,g}$ is locally asymptotically normal at $t = 0$ by Lemma 25.14. Therefore, because the covariance matrix of g_P is the identity matrix,

$$\big(P^n_{1/\sqrt{n},h^Tg_P} : h \in \mathbb{R}^m\big) \rightsquigarrow \big(N_m(h, I) : h \in \mathbb{R}^m\big)$$

in the sense of convergence of experiments. The function $\psi_n(h) = \psi(P_{1/\sqrt{n},h^Tg_P})$ satisfies

$$\sqrt{n}\big(\psi_n(h) - \psi_n(0)\big) \to \dot{\psi}_P h^T g_P = \big(P\tilde{\psi}_P g_P^T\big)h =: Ah.$$

For the same $(k \times m)$ matrix the function Ag_P is the orthogonal projection of $\tilde{\psi}_P$ onto $\operatorname{lin} g_P$, and it has covariance matrix AA^T. Because $\tilde{\psi}_P$ is, by definition, contained in the closed linear span of the tangent set, we can choose g_P such that $\tilde{\psi}_P$ is arbitrarily close to its projection and hence AA^T is arbitrarily close to $P\tilde{\psi}_P\tilde{\psi}_P^T$.

Under the assumption of the convolution theorem, the limit distribution of the sequence $\sqrt{n}\big(T_n - \psi_n(h)\big)$ under $P_{1/\sqrt{n},h^Tg_P}$ is the same for every $h \in \mathbb{R}^m$. By the asymptotic representation theorem, Proposition 7.10, there exists a randomized statistic T in the limit experiment such that the distribution of $T - Ah$ under h does not depend on h. By Proposition 8.4, the null distribution of T contains a normal $N(0, AA^T)$-distribution as a convolution factor. The proof of the convolution theorem is complete upon letting AA^T tend to $P\tilde{\psi}_P\tilde{\psi}_P^T$.

Under the assumption that the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ is tight, the minimax theorem is proved similarly, by first bounding the left side by the minimax risk relative to the submodel corresponding to g_P, and next applying Proposition 8.6. The tightness assumption can be dropped by a compactification argument. (see, e.g., [139], or [146]).

If the tangent set is a convex cone but not a linear space, then the submodel constructed previously can only be used for h ranging over a convex cone in $\mathbb{R}^m$. The argument can

remain the same, except that we need to replace Propositions 8.4 and 8.6 by stronger results that refer to convex cones. These extensions exist and can be proved by the same Bayesian argument, now choosing priors that flatten out inside the cone (see, e.g., [139]).

If the tangent set is a cone that is not convex, but the estimator sequence is regular, then we use the fact that the matching randomized estimator T in the limit experiment satisfies $\mathrm{E}_h T = Ah + \mathrm{E}_0 T$ for every eligible h, that is, every h such that $h^T g_P \in \dot{\mathcal{P}}_P$. Because the tangent set is a cone, the latter set includes parameters $h = th_i$ for $t \geq 0$ and directions h_i spanning $\mathbb{R}^m$. The estimator T is unbiased for estimating $Ah + \mathrm{E}_0 T$ on this parameter set, whence the covariance matrix of T is bounded below by AA^T, by the Cramér-Rao inequality. ■

Both theorems have the interpretation that the matrix $P\tilde{\psi}_P\tilde{\psi}_P^T$ is an optimal asymptotic covariance matrix for estimating $\psi(P)$ given the model $\mathcal{P}$. We might wish that this could be formulated in a simpler fashion, but this is precluded by the problem of superefficiency, as is already the case for the parametric analogues, discussed in Chapter 8. That the notion of asymptotic efficiency used in the present interpretation should not be taken absolutely is shown by the shrinkage phenomena discussed in section 8.8, but we use it in this chapter. We shall say that an estimator sequence is *asymptotically efficient* at P, if it is regular at P with limit distribution $L = N(0, P\tilde{\psi}_P\tilde{\psi}_P^T)$.†

The efficient influence function $\tilde{\psi}_P$ plays the same role as the normalized score function $I_\theta^{-1}\dot{l}_\theta$ in parametric models. In particular, a sequence of estimators T_n is asymptotically efficient at P if

$$\sqrt{n}\big(T_n - \psi(P)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_P(X_i) + o_P(1). \tag{25.22}$$

This justifies the name "efficient influence function."

25.23 *Lemma.* *Let the function $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ be differentiable at P relative to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. A sequence of estimators T_n is regular at P with limiting distribution $N(0, P\tilde{\psi}_P\tilde{\psi}_P^T)$ if and only if it satisfies* (25.22).

Proof. Because the submodels $t \mapsto P_{t,g}$ are locally asymptotically normal at $t = 0$, "if" follows with the help of Le Cam's third lemma, by the same arguments as for the analogous result for parametric models in Lemma 8.14.

To prove the necessity of (25.22), we adopt the notation of the proof of Theorem 25.20. The statistics $S_n = \psi(P) + n^{-1}\sum_{i=1}^n \tilde{\psi}_P(X_i)$ depend on P but can be considered a true estimator sequence in the local subexperiments. The sequence S_n trivially satisfies (25.22) and hence is another asymptotically efficient estimator sequence. We may assume for simplicity that the sequence $\sqrt{n}\big(S_n - \psi(P_{1/\sqrt{n},g}), T_n - \psi(P_{1/\sqrt{n},g})\big)$ converges under every local parameter g in distribution. Otherwise, we argue along subsequences, which can be

† If the tangent set is not a linear space, then the situation becomes even more complicated. If the tangent set is a convex cone, then the minimax risk in the left side of Theorem 25.21 cannot fall below the normal risk on the right side, but there may be nonregular estimator sequences for which there is equality. If the tangent set is not convex, then the assertion of Theorem 25.21 may fail. Convex tangent cones arise frequently; fortunately, nonconvex tangent cones are rare.

selected with the help of Le Cam's third lemma. By Theorem 9.3, there exists a matching randomized estimator $(S, T) = (S, T)(X, U)$ in the normal limit experiment. By the efficiency of both sequences S_n and T_n, the variables $S - Ah$ and $T - Ah$ are, under h, marginally normally distributed with mean zero and covariance matrix $P\tilde{\psi}_P\tilde{\psi}_P^T$. In particular, the expectations $\mathrm{E}_h S = \mathrm{E}_h T$ are identically equal to Ah. Differentiate with respect to h at $h = 0$ to find that

$$\mathrm{E}_0 S X^T = \mathrm{E}_0 T X^T = A.$$

It follows that the orthogonal projections of S and T onto the linear space spanned by the coordinates of X are identical and given by $\Pi S = \Pi T = AX$, and hence

$$\mathrm{Cov}_0(S - T) = \mathrm{Cov}_0(\Pi^\perp S - \Pi^\perp T) \le 2\,\mathrm{Cov}_0\,\Pi^\perp S + 2\,\mathrm{Cov}_0\,\Pi^\perp T.$$

(The inequality means that the difference of the matrices on the right and the left is nonnegative-definite.) We have obtained this for a fixed orthonormal set $g_P = (g_1, \ldots, g_m)$. If we choose g_P such that AA^T is arbitrarily close to $P\tilde{\psi}_P\tilde{\psi}_P^T$, equivalently $\mathrm{Cov}_0\,\Pi T = AA^T = \mathrm{Cov}_0\,\Pi S$ is arbitrarily close to $\mathrm{Cov}_0\,T = P\tilde{\psi}_P\tilde{\psi}_P^T = \mathrm{Cov}_0\,S$, and then the right side of the preceding display is arbitrarily close to zero, whence $S - T \approx 0$. The proof is complete on noting that $\sqrt{n}(S_n - T_n) \overset{0}{\rightsquigarrow} S - T$. ■

25.24 ***Example (Empirical distribution).*** The empirical distribution is an asymptotically efficient estimator if the underlying distribution P of the sample is completely unknown. To give a rigorous expression to this intuitively obvious statement, fix a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$ with $Pf^2 < \infty$, for instance an indicator function $f = 1_A$, and consider $\mathbb{P}_n f = n^{-1}\sum_{i=1}^n f(X_i)$ as an estimator for the function $\psi(P) = Pf$.

In Example 25.16 it is seen that the maximal tangent space for the nonparametric model is equal to the set of all $g \in L_2(P)$ such that $Pg = 0$. For a general function f, the parameter ψ may not be differentiable relative to the maximal tangent set, but it is differentiable relative to the tangent space consisting of all bounded, measurable functions g with $Pg = 0$. The closure of this tangent space is the maximal tangent set, and hence working with this smaller set does not change the efficient influence functions. For a bounded function g with $Pf = 0$ we can use the submodel defined by $dP_t = (1 + tg)\,dP$, for which $\psi(P_t) = Pf + tPfg$. Hence the derivative of ψ is the map $g \mapsto \dot{\psi}_P g = Pfg$, and the efficient influence function relative to the maximum tangent set is the function $\tilde{\psi}_P = f - Pf$. (The function f is an influence function; its projection onto the mean zero functions is $f - Pf$.)

The optimal asymptotic variance for estimating $P \mapsto Pf$ is equal to $P\tilde{\psi}_P^2 = P(f - Pf)^2$. The sequence of empirical estimators $\mathbb{P}_n f$ is asymptotically efficient, because it satisfies (25.22), with the $o_P(1)$-remainder term identically zero. □

25.4 Efficient Score Functions

A function $\psi(P)$ of particular interest is the parameter θ in a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$. Here Θ is an open subset of $\mathbb{R}^k$ and H is an arbitrary set, typically of infinite dimension. The information bound for the functional of interest $\psi(P_{\theta,\eta}) = \theta$ can be conveniently expressed in an "efficient score function."

As submodels, we use paths of the form $t \mapsto P_{\theta+ta,\eta_t}$, for given paths $t \mapsto \eta_t$ in the parameter set H. The score functions for such submodels (if they exist) typically have the form of a sum of "partial derivatives" with respect to θ and η. If $\dot{\ell}_{\theta,\eta}$ is the ordinary score function for θ in the model in which η is fixed, then we expect

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

The function g has the interpretation of a score function for η if θ is fixed and runs through an infinite-dimensional set if we are concerned with a "true" semiparametric model. We refer to this set as the *tangent set for* η, and denote it by ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$.

The parameter $\psi(P_{\theta+ta,\eta_t}) = \theta + ta$ is certainly differentiable with respect to t in the ordinary sense but is, by definition, differentiable as a parameter on the model if and only if there exists a function $\tilde{\psi}_{\theta,\eta}$ such that

$$a = \frac{\partial}{\partial t}_{|t=0} \psi(P_{\theta+ta,\eta_t}) = \langle \tilde{\psi}_{\theta,\eta}, a^T \dot{\ell}_{\theta,\eta} + g \rangle_{P_{\theta,\eta}}, \qquad a \in \mathbb{R}^k, g \in {}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}.$$

Setting $a = 0$, we see that $\tilde{\psi}_{\theta,\eta}$ must be orthogonal to the tangent set ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ for the nuisance parameter. Define $\Pi_{\theta,\eta}$ as the orthogonal projection onto the closure of the linear span of ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ in $L_2(P_{\theta,\eta})$.

The function defined by

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$$

is called the *efficient score function* for θ, and its covariance matrix $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^T$ is the *efficient information matrix*.

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

$$\int \left[\frac{dP_{\theta+ta,\eta_t}^{1/2} - dP_{\theta,\eta}^{1/2}}{t} - \frac{1}{2}\left(a^T \dot{\ell}_{\theta,\eta} + g\right) dP_{\theta,\eta}^{1/2} \right]^2 \to 0. \tag{25.26}$$

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

Proof. The given set $\dot{\mathcal{P}}_{P_{\theta,\eta}}$ is a tangent set by assumption. The function ψ is differentiable with respect to this tangent set because

$$\langle \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}, a^T \dot{\ell}_{\theta,\eta} + g \rangle_{P_{\theta,\eta}} = \tilde{I}_{\theta,\eta}^{-1} \langle \tilde{\ell}_{\theta,\eta}, \dot{\ell}_{\theta,\eta}^T \rangle_{P_{\theta,\eta}} a = a.$$

The last equality follows, because the inner product of a function and its orthogonal projection is equal to the square length of the projection. Thus, we may replace $\dot{\ell}_{\theta,\eta}$ by $\tilde{\ell}_{\theta,\eta}$. ■

Consequently, an estimator sequence is asymptotically efficient for estimating θ if

$$\sqrt{n}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}(X_i) + o_{P_{\theta,\eta}}(1).$$

This equation is very similar to the equation derived for efficient estimators in parametric models in Chapter 8. It differs only in that the ordinary score function $\dot{\ell}_{\theta,\eta}$ has been replaced by the efficient score function (and similarly for the information). The intuitive explanation is that a part of the score function for θ can also be accounted for by score functions for the nuisance parameter η. If the nuisance parameter is unknown, a part of the information for θ is "lost," and this corresponds to a loss of a part of the score function.

25.27 ***Example (Symmetric location).*** Suppose that the model consists of all densities $x \mapsto \eta(x-\theta)$ with $\theta \in \mathbb{R}$ and the "shape" η symmetric about 0 with finite Fisher information for location I_η. Thus, the observations are sampled from a density that is symmetric about θ.

By the symmetry, the density can equivalently be written as $\eta\big(|x-\theta|\big)$. It follows that any score function for the nuisance parameter η is necessarily a function of $|x-\theta|$. This suggests a tangent set containing functions of the form $a(\eta'/\eta)(x-\theta) + b\big(|x-\theta|\big)$. It is not hard to show that all square-integrable functions of this type with mean zero occur as score functions in the sense of (25.26).†

A symmetric density has an asymmetric derivative and hence an asymmetric score function for location. Therefore, for every b,

$$\mathrm{E}_{\theta,\eta}\frac{\eta'}{\eta}(X-\theta)\,b\big(|X-\theta|\big) = 0.$$

Thus, the projection of the θ-score onto the set of nuisance scores is zero and hence the efficient score function coincides with the ordinary score function. This means that there is no difference in information about θ whether the form of the density is known or not known, as long as it is known to be symmetric. This surprising fact was discovered by Stein in 1956 and has been an important motivation in the early work on semiparametric models.

Even more surprising is that the information calculation is not misleading. There exist estimator sequences for θ whose definition does not depend on η that have asymptotic variance I_η^{-1} under any true η. See section 25.8. Thus a symmetry point can be estimated as well if the shape is known as if it is not, at least asymptotically. □

25.28 ***Example (Regression).*** Let g_θ be a given set of functions indexed by a parameter $\theta \in \mathbb{R}^k$, and suppose that a typical observation (X, Y) follows the regression model

$$Y = g_\theta(X) + e, \qquad \mathrm{E}(e \mid X) = 0.$$

This model includes the logistic regression model, for $g_\theta(x) = 1/(1+e^{-\theta^T x})$. It is also a version of the ordinary linear regression model. However, in this example we do not assume that X and e are independent, but only the relations in the preceding display, apart from qualitative smoothness conditions that ensure existence of score functions, and the existence of moments. We shall write the formulas assuming that (X, e) possesses a density η. Thus, the observation (X, Y) has a density $\eta\big(x, y - g_\theta(x)\big)$, in which η is (essentially) only restricted by the relations $\int e\eta(x,e)\,de \equiv 0$.

Because any perturbation η_t of η within the model must satisfy this same relation $\int e\eta_t(x,e)\,de = 0$, it is clear that score functions for the nuisance parameter η are functions

† That no other functions can occur is shown in, for example, [8, p. 56–57] but need not concern us here.

$a\big(x, y - g_\theta(x)\big)$ that satisfy

$$\mathrm{E}\big(ea(X,e) \mid X\big) = \frac{\int ea(X,e)\,\eta(X,e)\,de}{\int \eta(X,e)\,de} = 0.$$

By the same argument as for nonparametric models all bounded square-integrable functions of this type that have mean zero are score functions. Because the relation $\mathrm{E}\big(ea(X,e) \mid X\big) = 0$ is equivalent to the orthogonality in $L_2(\eta)$ of $a(x,e)$ to all functions of the form $eh(x)$, it follows that the set of score functions for η is the orthocomplement of the set $e\mathcal{H}$, of all functions of the form $(x,y) \mapsto \big(y - g_\theta(x)\big)h(x)$ within $L_2(P_{\theta,\eta})$, up to centering at mean zero.

Thus, we obtain the efficient score function for θ by projecting the ordinary score function $\dot\ell_{\theta,\eta}(x,y) = -\eta_2/\eta(x,e)\dot g_\theta(x)$ onto $e\mathcal{H}$. The projection of an arbitrary function $b(x,e)$ onto the functions $e\mathcal{H}$ is a function $eh_0(x)$ such that $\mathrm{E}b(X,e)eh(X) = \mathrm{E}eh_0(X)eh(X)$ for all measurable functions h. This can be solved for h_0 to find that the projection operator takes the form

$$\Pi_{e\mathcal{H}} b(X,e) = e\frac{\mathrm{E}\big(b(X,e)e \mid X\big)}{\mathrm{E}(e^2 \mid X)}.$$

This readily yields the efficient score function

$$\tilde\ell_{\theta,\eta}(X,Y) = -\frac{e\dot g_\theta(X)}{\mathrm{E}(e^2 \mid X)}\frac{\int \eta_2(X,e)e\,de}{\int \eta(X,e)\,de} = \frac{\big(Y - g_\theta(X)\big)\dot g_\theta(X)}{\mathrm{E}(e^2 \mid X)}.$$

The efficient information takes the form $\tilde I_{\theta,\eta} = \mathrm{E}\big(\dot g_\theta \dot g_\theta^T(X)/\mathrm{E}(e^2 \mid X)\big)$. □

25.5 Score and Information Operators

The method to find the efficient influence function of a parameter given in the preceding section is the most convenient method if the model can be naturally partitioned in the parameter of interest and a nuisance parameter. For many parameters such a partition is impossible or, at least, unnatural. Furthermore, even in semiparametric models it can be worthwhile to derive a more concrete description of the tangent set for the nuisance parameter, in terms of a "score operator."

Consider first the situation that the model $\mathcal{P} = \{P_\eta : \eta \in H\}$ is indexed by a parameter η that is itself a probability measure on some measurable space. We are interested in estimating a parameter of the type $\psi(P_\eta) = \chi(\eta)$ for a given function $\chi : H \mapsto \mathbb{R}^k$ on the model H.

The model H gives rise to a tangent set $\dot H_\eta$ at η. If the map $\eta \mapsto P_\eta$ is differentiable in an appropriate sense, then its derivative maps every score $b \in \dot H_\eta$ into a score g for the model $\mathcal{P}$. To make this precise, we assume that a smooth parametric submodel $t \mapsto \eta_t$ induces a smooth parametric submodel $t \mapsto P_{\eta_t}$, and that the score functions b of the submodel $t \mapsto \eta_t$ and g of the submodel $t \mapsto P_{\eta_t}$ are related by

$$g = A_\eta b.$$

Then $A_\eta \dot H_\eta$ is a tangent set for the model $\mathcal{P}$ at P_η. Because A_η turns scores for the model H into scores for the model $\mathcal{P}$ it is called a *score operator*. It is seen subsequently here that if η

and P_η are the distributions of an unobservable Y and an observable $X = m(Y)$, respectively, then the score operator is a conditional expectation. More generally, it can be viewed as a derivative of the map $\eta \mapsto P_\eta$. We assume that A_η, as a map $A_\eta : \operatorname{lin} \dot{H}_\eta \subset L_2(\eta) \mapsto L_2(P_\eta)$, is continuous and linear.

Next, assume that the function $\eta \mapsto \chi(\eta)$ is differentiable with influence function $\tilde{\chi}_\eta$ relative to the tangent set $\dot{H}_\eta$. Then, by definition, the function $\psi(P_\eta) = \chi(\eta)$ is pathwise differentiable relative to the tangent set $\dot{\mathcal{P}}_{P_\eta} = A_\eta \dot{H}_\eta$ if and only if there exists a vector-valued function $\tilde{\psi}_{P_\eta}$ such that

$$\langle \tilde{\psi}_{P_\eta}, A_\eta b \rangle_{P_\eta} = \frac{\partial}{\partial t}_{|t=0} \psi(P_{\eta_t}) = \frac{\partial}{\partial t}_{|t=0} \chi(\eta_t) = \langle \tilde{\chi}_\eta, b \rangle_\eta, \qquad b \in \dot{H}_\eta.$$

This equation can be rewritten in terms of the *adjoint score operator* $A_\eta^* : L_2(P_\eta) \mapsto \overline{\operatorname{lin}}\, \dot{H}_\eta$. By definition this satisfies $\langle h, A_\eta b \rangle_{P_\eta} = \langle A_\eta^* h, b \rangle_\eta$ for every $h \in L_2(P_\eta)$ and $b \in \dot{H}_\eta$.[†] The preceding display is equivalent to

$$A_\eta^* \tilde{\psi}_{P_\eta} = \tilde{\chi}_\eta. \tag{25.29}$$

We conclude that the function $\psi(P_\eta) = \chi(\eta)$ is differentiable relative to the tangent set $\dot{\mathcal{P}}_{P_\eta} = A_\eta \dot{H}_\eta$ if and only if this equation can be solved for $\tilde{\psi}_{P_\eta}$; equivalently, if and only if $\tilde{\chi}_\eta$ is contained in the range of the adjoint A_η^*. Because A_η^* is not necessarily onto $\overline{\operatorname{lin}}\, \dot{H}_\eta$, not even if it is one-to-one, this is a condition.

For multivariate functionals (25.29) is to be understood coordinate-wise. Two solutions $\tilde{\psi}_{P_\eta}$ of (25.29) can differ only by an element of the kernel $\mathrm{N}(A_\eta^*)$ of A_η^*, which is the orthocomplement $\mathrm{R}(A_\eta)^\perp$ of the range of $A_\eta : \operatorname{lin} \dot{H}_\eta \mapsto L_2(P_\eta)$. Thus, there is at most one solution $\tilde{\psi}_{P_\eta}$ that is contained in $\overline{\mathrm{R}}(A_\eta) = \overline{\operatorname{lin}}\, A_\eta \dot{H}_\eta$, the closure of the range of A_η, as required.

If $\tilde{\chi}_\eta$ is contained in the smaller range of $A_\eta^* A_\eta$, then (25.29) can be solved, of course, and the solution can be written in the attractive form

$$\tilde{\psi}_{P_\eta} = A_\eta (A_\eta^* A_\eta)^- \tilde{\chi}_\eta. \tag{25.30}$$

Here $A_\eta^* A_\eta$ is called the *information operator*, and $(A_\eta^* A_\eta)^-$ is a "generalized inverse." (Here this will not mean more than that $b = (A_\eta^* A_\eta)^- \tilde{\chi}_\eta$ is a solution to the equation $A_\eta^* A_\eta b = \tilde{\chi}_\eta$.) In the preceding equation the operator $A_\eta^* A_\eta$ performs a similar role as the matrix $X^T X$ in the least squares solution of a linear regression model. The operator $A_\eta (A_\eta^* A_\eta)^{-1} A_\eta^*$ (if it exists) is the orthogonal projection onto the range space of A_η.

So far we have assumed that the parameter η is a probability distribution, but this is not necessary. Consider the more general situation of a model $\mathcal{P} = \{P_\eta : \eta \in H\}$ indexed by a parameter η running through an arbitrary set H. Let $\mathbb{H}_\eta$ be a subset of a Hilbert space that indexes "directions" b in which η can be approximated within H. Suppose that there exist continuous, linear operators $A_\eta : \operatorname{lin} \mathbb{H}_\eta \mapsto L_2(P_\eta)$ and $\dot{\chi}_\eta : \operatorname{lin} \mathbb{H}_\eta \mapsto \mathbb{R}^k$, and for every $b \in \mathbb{H}_\eta$ a path $t \mapsto \eta_t$ such that, as $t \downarrow 0$,

$$\int \left[\frac{dP_{\eta_t}^{1/2} - dP_\eta^{1/2}}{t} - \frac{1}{2} A_\eta b \, dP_\eta^{1/2} \right]^2 \to 0, \qquad \frac{\chi(\eta_t) - \chi(\eta)}{t} \to \dot{\chi}_\eta b.$$

[†] Note that we define A_η^* to have range $\overline{\operatorname{lin}}\, \dot{H}_\eta$, so that it is the adjoint of $A_\eta : \dot{H}_\eta \mapsto L_2(P_\eta)$. This is the adjoint of an extension $A_\eta : L_2(\eta) \mapsto L_2(P_\eta)$ followed by the orthogonal projection onto $\overline{\operatorname{lin}}\, \dot{H}_\eta$.

By the Riesz representation theorem for Hilbert spaces, the "derivative" $\dot{\chi}_\eta$ has a representation as an inner product $\dot{\chi}_\eta b = \langle \tilde{\chi}_\eta, b\rangle_{\mathbb{H}_\eta}$ for an element $\tilde{\chi}_\eta \in \overline{\text{lin}}\,\mathbb{H}_\eta^k$. The preceding discussion can be extended to this abstract set-up.

25.31 ***Theorem.*** *The map $\psi : \mathcal{P} \mapsto \mathbb{R}^k$ given by $\psi(P_\eta) = \chi(\eta)$ is differentiable at P_η relative to the tangent set $A_\eta \mathbb{H}_\eta$ if and only if each coordinate function of $\tilde{\chi}_\eta$ is contained in the range of $A_\eta^* : L_2(P_\eta) \mapsto \overline{\text{lin}}\,\mathbb{H}_\eta$. The efficient influence function $\tilde{\psi}_{P_\eta}$ satisfies* (25.29). *If each coordinate function of $\tilde{\chi}_\eta$ is contained in the range of $A_\eta^* A_\eta : \overline{\text{lin}}\,\mathbb{H}_\eta \mapsto \overline{\text{lin}}\,\mathbb{H}_\eta$, then it also satisfies* (25.30).

Proof. By assumption, the set $A_\eta \mathbb{H}_\eta$ is a tangent set. The map ψ is differentiable relative to this tangent set (and the corresponding submodels $t \mapsto P_{\eta_t}$) by the argument leading up to (25.29). ■

The condition (25.29) is odd. By definition, the influence function $\tilde{\chi}_\eta$ is contained in the closed linear span of $\mathbb{H}_\eta$ and the operator A_η^* maps $L_2(P_\eta)$ into $\overline{\text{lin}}\,\mathbb{H}_\eta$. Therefore, the condition is certainly satisfied if A_η^* is onto. There are two reasons why it may fail to be onto. First, its range $\mathrm{R}(A_\eta^*)$ may be a proper subspace of $\overline{\text{lin}}\,\mathbb{H}_\eta$. Because $b \perp \mathrm{R}(A_\eta^*)$ if and only if $b \in \mathrm{N}(A_\eta)$, this can happen only if A_η is not one-to-one. This means that two different directions b may lead to the same score function $A_\eta b$, so that the information matrix for the corresponding two-dimensional submodel is singular. A rough interpretation is that the parameter is not locally identifiable. Second, the range space $\mathrm{R}(A_\eta^*)$ may be dense but not closed. Then for any $\tilde{\chi}_\eta$ there exist elements in $\mathrm{R}(A_\eta^*)$ that are arbitrarily close to $\tilde{\chi}_\eta$, but (25.29) may still fail. This happens quite often. The following theorem shows that failure has serious consequences.†

25.32 ***Theorem.*** *Suppose that $\eta \mapsto \chi(\eta)$ is differentiable with influence function $\tilde{\chi}_\eta \notin \mathrm{R}(A_\eta^*)$. Then there exists no estimator sequence for $\chi(\eta)$ that is regular at P_η.*

25.5.1 *Semiparametric Models*

In a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, the pair (θ, η) plays the role of the single η in the preceding general discussion. The two parameters can be perturbed independently, and the score operator can be expected to take the form

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

Here $B_{\theta,\eta} : \mathbb{H}_\eta \mapsto L_2(P_{\theta,\eta})$ is the score operator for the nuisance parameter. The domain of the operator $A_{\theta,\eta} : \mathbb{R}^k \times \text{lin}\,\mathbb{H}_\eta \mapsto L_2(P_{\theta,\eta})$ is a Hilbert space relative to the inner product

$$\big\langle (a, b), (\alpha, \beta) \big\rangle_\eta = a^T \alpha + \langle b, \beta \rangle_{\mathbb{H}_\eta}.$$

Thus this example fits in the general set-up, with $\mathbb{R}^k \times \mathbb{H}_\eta$ playing the role of the earlier $\mathbb{H}_\eta$. We shall derive expressions for the efficient influence functions of θ and η.

The efficient influence function for estimating θ is expressed in the *efficient score function* for θ in Lemma 25.25, which is defined as the ordinary score function minus its projection

† For a proof, see [140].

onto the score-space for η. Presently, the latter space is the range of the operator $B_{\theta,\eta}$. If the operator $B^*_{\theta,\eta}B_{\theta,\eta}$ is continuously invertible (but in many examples it is not), then the operator $B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta}$ is the orthogonal projection onto the nuisance score space, and

$$\tilde{\ell}_{\theta,\eta} = \big(I - B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta}\big)\dot{\ell}_{\theta,\eta}. \tag{25.33}$$

This means that $b = -(B^*_{\theta,\eta}B_{\theta,\eta})^{-1}B^*_{\theta,\eta}\dot{\ell}_{\theta,\eta}$ is a "least favorable direction" in H, for estimating θ. If θ is one-dimensional, then the submodel $t \mapsto P_{\theta+t,\eta_t}$ where η_t approaches η in this direction, has the least information for estimating t and score function $\tilde{\ell}_{\theta,\eta}$, at $t = 0$.

A function $\chi(\eta)$ of the nuisance parameter can, despite the name, also be of interest. The efficient influence function for this parameter can be found from (25.29). The adjoint of $A_{\theta,\eta}: \mathbb{R}^k \times \mathbb{H}_\eta \mapsto L_2(P_{\theta,\eta})$, and the corresponding information operator $A^*_{\theta,\eta}A_{\theta,\eta}: \mathbb{R}^k \times \mathbb{H}_\eta \mapsto \mathbb{R}^k \times \overline{\text{lin}}\,\mathbb{H}_\eta$ are given by, with $B^*_{\theta,\eta}: L_2(P_{\theta,\eta} \mapsto \overline{\text{lin}}\,\mathbb{H}_\eta$ the adjoint of $B_{\theta,\eta}$,

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

The diagonal elements in the matrix are the information operators for the parameters θ and η, respectively, the former being just the ordinary Fisher information matrix $I_{\theta,\eta}$ for θ. If $\eta \mapsto \chi(\eta)$ is differentiable as before, then the function $(\theta, \eta) \mapsto \chi(\eta)$ is differentiable with influence function $(0, \tilde{\chi}_\eta)$. Thus, for a real parameter $\chi(\eta)$, equation (25.29) becomes

$$P_{\theta,\eta}\tilde{\psi}_{P_{\theta,\eta}}\dot{\ell}_{\theta,\eta} = 0, \qquad B^*_{\theta,\eta}\tilde{\psi}_{P_{\theta,\eta}} = \tilde{\chi}_\eta.$$

If $\tilde{I}_{\theta,\eta}$ is invertible and $\tilde{\chi}_\eta$ is contained in the range of $B^*_{\theta,\eta}B_{\theta,\eta}$, then the solution $\tilde{\psi}_{P_{\theta,\eta}}$ of these equations is

$$B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^-\tilde{\chi}_\eta - \big\langle B_{\theta,\eta}(B^*_{\theta,\eta}B_{\theta,\eta})^-\tilde{\chi}_\eta, \dot{\ell}_{\theta,\eta}\big\rangle^T_{P_{\theta,\eta}}\tilde{I}^{-1}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}.$$

The second part of this function is the part of the efficient score function for $\chi(\eta)$ that is "lost" due to the fact that θ is unknown. Because it is orthogonal to the first part, it adds a positive contribution to the variance.

25.5.2 *Information Loss Models*

Suppose that a typical observation is distributed as a measurable transformation $X = m(Y)$ of an unobservable variable Y. Assume that the form of m is known and that the distribution η of Y is known to belong to a class H. This yields a natural parametrization of the distribution P_η of X. A nice property of differentiability in quadratic mean is that it is preserved under "censoring" mechanisms of this type: If $t \mapsto \eta_t$ is a differentiable submodel of H, then the induced submodel $t \mapsto P_{\eta_t}$ is a differentiable submodel of $\{P_\eta : \eta \in H\}$. Furthermore, the score function $g = A_\eta b$ (at $t = 0$) for the induced model $t \mapsto P_{\eta_t}$ can be obtained from the score function b (at $t = 0$) of the model $t \mapsto \eta_t$ by taking a conditional expectation:

$$A_\eta b(x) = \mathrm{E}_\eta\big(b(Y) \mid X = x\big).$$

If we consider the scores b and g as the carriers of information about t in the variables Y with law η_t and X with law P_{η_t}, respectively, then the intuitive meaning of the conditional expectation operator is clear. The information contained in the observation X is the information contained in Y diluted (and reduced) through conditioning.†

25.34 ***Lemma.*** *Suppose that $\{\eta_t : 0 < t < 1\}$ is a collection of probability measures on a measurable space $(\mathcal{Y}, \mathcal{B})$ such that for some measurable function $b : \mathcal{Y} \mapsto \mathbb{R}$*

$$\int \left[\frac{d\eta_t^{1/2} - d\eta^{1/2}}{t} - \frac{1}{2} b \, d\eta^{1/2} \right]^2 \to 0.$$

For a measurable map $m : \mathcal{Y} \mapsto \mathcal{X}$ let P_η be the distribution of $m(Y)$ if Y has law η and let $A_\eta b(x)$ be the conditional expectation of $b(Y)$ given $m(Y) = x$. Then

$$\int \left[\frac{dP_{\eta_t}^{1/2} - dP_\eta^{1/2}}{t} - \frac{1}{2} A_\eta b \, dP_\eta^{1/2} \right]^2 \to 0.$$

If we consider A_η as an operator $A_\eta : L_2(\eta) \mapsto L_2(P_\eta)$, then its adjoint $A_\eta^* : L_2(P_\eta) \mapsto L_2(\eta)$ is a conditional expectation operator also, reversing the roles of X and Y,

$$A_\eta^* g(y) = \mathrm{E}_\eta\big(g(X) \mid Y = y\big).$$

This follows because, by the usual rules for conditional expectations, $\mathrm{E}\mathrm{E}\big(g(X) \mid Y\big)b(Y) = \mathrm{E}g(X)b(Y) = \mathrm{E}g(X)\mathrm{E}\big(b(Y) \mid X\big)$. In the "calculus of scores" of Theorem 25.31 the adjoint is understood to be the adjoint of $A_\eta : \mathbb{H}_\eta \mapsto L_2(P_\eta)$ and hence to have range $\overline{\mathrm{lin}}\, \mathbb{H}_\eta \subset L_2(\eta)$. Then the conditional expectation in the preceding display needs to be followed by the orthogonal projection onto $\overline{\mathrm{lin}}\, \mathbb{H}_\eta$.

25.35 ***Example (Mixtures).*** Suppose that a typical observation X possesses a conditional density $p(x \mid z)$ given an unobservable variable $Z = z$. If the unobservable Z possesses an unknown probability distribution η, then the observations are a random sample from the mixture density

$$p_\eta(x) = \int p(x \mid z) \, d\eta(z).$$

This is a missing data problem if we think of X as a function of the pair $Y = (X, Z)$. A score for the mixing distribution η in the model for Y is a function $b(z)$. Thus, a score space for the mixing distribution in the model for X consists of the functions

$$A_\eta b(x) = \mathrm{E}_\eta\big(b(Z) \mid X = x\big) = \frac{\int b(z) \, p(x \mid z) \, d\eta(z)}{\int p(x \mid z) \, d\eta(z)}.$$

If the mixing distribution is completely unknown, which we assume, then the tangent set $\dot{H}_\eta$ for η can be taken equal to the maximal tangent set $\big\{b \in L_2(\eta) : \eta b = 0\big\}$.

In particular, consider the situation that the kernel $p(x \mid z)$ belongs to an exponential family, $p(x \mid z) = c(z)d(x) \exp\big(z^T x\big)$. We shall show that the tangent set $A_\eta \dot{H}_\eta$ is dense

† For a proof of the following lemma, see, for example, [139, pp. 188–193].

in the maximal tangent set $\{g \in L_2(P_\eta) : P_\eta g = 0\}$, for every η whose support contains an interval. This has as a consequence that empirical estimators $\mathbb{P}_n g$, for a fixed squared-integrable function g, are efficient estimators for the functional $\psi(\eta) = P_\eta g$. For instance, the sample mean is asymptotically efficient for estimating the mean of the observations.

Thus nonparametric mixtures over an exponential family form very large models, which are only slightly smaller than the nonparametric model. For estimating a functional such as the mean of the observations, it is of relatively little use to know that the underlying distribution is a mixture. More precisely, the additional information does not decrease the asymptotic variance, although there may be an advantage for finite n. On the other hand, the mixture structure may express a structure in reality and the mixing distribution η may define the functional of interest.

The closure of the range of the operator A_η is the orthocomplement of the kernel $N(A_\eta^*)$ of its adjoint. Hence our claim is proved if this kernel is zero. The equation

$$0 = A_\eta^* g(z) = E\big(g(X) \mid Z = z\big) = \int g(x)\, p(x \mid z)\, d\nu(x)$$

says exactly that $g(X)$ is a zero-estimator under $p(x \mid z)$. Because the adjoint is defined on $L_2(\eta)$, the equation $0 = A_\eta^* g$ should be taken to mean $A_\eta^* g(Z) = 0$ almost surely under η. In other words, the display is valid for every z in a set of η-measure 1. If the support of η contains a limit point, then this set is rich enough to conclude that $g = 0$, by the completeness of the exponential family.

If the support of η does not contain a limit point, then the preceding approach fails. However, we may reach almost the same conclusion by using a different type of scores. The paths $\eta_t = (1 - ta)\eta + ta\eta_1$ are well-defined for $0 \le at \le 1$, for any fixed $a \ge 0$ and η_1, and lead to scores

$$\frac{\partial}{\partial t}_{|t=0} \log p_{\eta_t}(x) = a\left(\frac{p_{\eta_1}}{p_\eta}(x) - 1\right).$$

This is certainly a score in a pointwise sense and can be shown to be a score in the L_2-sense provided that it is in $L_2(P_\eta)$. If $g \in L_2(P_\eta)$ has $P_\eta g = 0$ and is orthogonal to all scores of this type, then

$$0 = P_{\eta_1} g = P_\eta g\left(\frac{p_{\eta_1}}{p_\eta} - 1\right), \qquad \text{every } \eta_1.$$

If the set of distributions $\{P_\eta : \eta \in H\}$ is complete, then we can typically conclude that $g = 0$ almost surely. Then the closed linear span of the tangent set is equal to the nonparametric, maximal tangent set. Because this set of scores is also a convex cone, Theorems 25.20 and 25.21 next show that nonparametric estimators are asymptotically efficient. □

25.36 ***Example (Semiparametric mixtures).*** In the preceding example, replace the density $p(x \mid z)$ by a parametric family $p_\theta(x \mid z)$. Then the model $p_\theta(x \mid z)\, d\eta(z)$ for the unobserved data $Y = (X, Z)$ has scores for both θ and η. Suppose that the model $t \mapsto \eta_t$ is differentiable with score b, and that

$$\int\int \left[p_{\theta+a}^{1/2}(x \mid z) - p_\theta^{1/2}(x \mid z) - \frac{1}{2} a^T \dot{\ell}_\theta(x \mid z)\, p_\theta^{1/2}(x \mid z)\right]^2 d\mu(x)\, d\eta(z) = o\big(\|a\|^2\big).$$

Then the function $a^T \dot{\ell}_\theta(x \mid z) + b(z)$ can be shown to be a score function corresponding to the model $t \mapsto p_{\theta+ta}(x \mid z)\, d\eta_t(z)$. Next, by Lemma 25.34, the function

$$\mathrm{E}_{\theta,\eta}\big(a^T \dot{\ell}_\theta(X \mid Z) + b(Z) \mid X = x\big) = \frac{\int \big(\dot{\ell}_\theta(x \mid z) + b(z)\big)\, p_\theta(x \mid z)\, d\eta(z)}{\int p_\theta(x \mid z)\, d\eta(z)}$$

is a score for the model corresponding to observing X only. □

25.37 ***Example (Random censoring).*** Suppose that the time T of an event is only observed if the event takes place before a censoring time C that is generated independently of T; otherwise we observe C. Thus the observation $X = (Y, \Delta)$ is the pair of transformations $Y = T \wedge C$ and $\Delta = 1\{T \leq C\}$ of the "full data" (T, C). If T has a distribution function F and $t \mapsto F_t$ is a differentiable path with score function a, then the submodel $t \mapsto P_{F_t,G}$ for X has score function

$$A_{F,G}a(x) = \mathrm{E}_F\big(a(T) \mid X = (y, \delta)\big) = (1 - \delta)\frac{\int_{(y,\infty)} a\, dF}{1 - F(y)} + \delta a(y).$$

A score operator for the distribution of C can be defined similarly, and takes the form, with G the distribution of C,

$$B_{F,G}b(x) = (1 - \delta)b(y) + \delta\frac{\int_{[y,\infty)} b\, dG}{1 - G_-(y)}.$$

The scores $A_{F,G}a$ and $B_{F,G}b$ form orthogonal spaces, as can be checked directly from the formulas, because $\mathrm{E}A_F a(X)B_G b(X) = FaGb$. (This is also explained by the product structure in the likelihood.) A consequence is that knowing G does not help for estimating F in the sense that the information for estimating parameters of the form $\psi(P_{F,G}) = \chi(F)$ is the same in the models in which G is known or completely unknown, respectively. To see this, note first that the influence function of such a parameter must be orthogonal to every score function for G, because $d/dt\, \psi(P_{F,G_t}) = 0$. Thus, due to the orthogonality of the two score spaces, an influence function of this parameter that is contained in the closed linear span of $\mathrm{R}(A_{F,G}) + \mathrm{R}(B_{F,G})$ is automatically contained in $\mathrm{R}(A_{F,G})$. □

25.38 ***Example (Current status censoring).*** Suppose that we only observe whether an event at time T has happened or not at an observation time C. Then we observe the transformation $X = \big(C, 1\{T \leq C\}\big) = (C, \Delta)$ of the pair (C, T). If T and C are independent with distribution functions F and G, respectively, then the score operators for F and G are given by, with $x = (c, \delta)$,

$$\begin{aligned} A_{F,G}a(x) &= \mathrm{E}_F\big(a(T) \mid C = c, \Delta = \delta\big) = (1 - \delta)\frac{\int_{(c,\infty)} a\, dF}{1 - F(c)} + \delta\frac{\int_{[0,c]} a\, dF}{F(c)}, \\ B_{F,G}b(x) &= \mathrm{E}\big(b(C) \mid C = c, \Delta = \delta\big) = b(c). \end{aligned}$$

These score functions can be seen to be orthogonal with the help of Fubini's theorem. If we take F to be completely unknown, then the set of a can be taken all functions in $L_2(F)$ with $Fa = 0$, and the adjoint operator $A^*_{F,G}$ restricted to the set of mean-zero functions in $L_2(P_{F,G})$ is given by

$$A^*_{F,G}h(c) = \int_{[c,\infty)} h(u, 1)\, dG(u) + \int_{[0,c)} h(u, 0)\, dG(u).$$

For simplicity assume that the true F and G possess continuous Lebesgue densities, which are positive on their supports. The range of $A^*_{F,G}$ consists of functions as in the preceding display for functions h that are contained in $L_2(P_{F,G})$, or equivalently

$$\int h^2(u,0)(1-F)(u)\,dG(u) < \infty \quad \text{and} \quad \int h^2(u,1)F(u)\,dG(u) < \infty.$$

Thus the functions $h(u,1)$ and $h(u,0)$ are square-integrable with respect to G on any interval inside the support of F. Consequently, the range of the adjoint $A^*_{F,G}$ contains only absolutely continuous functions, and hence (25.29) fails for every parameter $\chi(F)$ with an influence function $\tilde{\chi}_F$ that is discontinuous. More precisely, parameters $\chi(F)$ with influence functions that are not almost surely equal under F to an absolutely continuous function. Because this includes the functions $1_{[0,t]} - F(t)$, the distribution function $F \mapsto \chi(F) = F(t)$ at a point is not a differentiable functional of the model. In view of Theorem 25.32 this means that this parameter is not estimable at $\sqrt{n}$-rate, and the usual normal theory does not apply to it.

On the other hand, parameters with a smooth influence function $\tilde{\chi}_F$ may be differentiable. The score operator for the model $P_{F,G}$ is the sum $(a,b) \mapsto A_{F,G}a + B_{F,G}b$ of the score operators for F and G separately. Its adjoint is the map $h \mapsto (A^*_{F,G}h, B^*_{F,G}h)$. A parameter of the form $(F,G) \mapsto \chi(F)$ has an influence function of the form $(\tilde{\chi}_F, 0)$. Thus, for a parameter of this type equation (25.29) takes the form

$$A^*_{F,G}\tilde{\psi}_{P_{F,G}} = \tilde{\chi}_F, \qquad B^*_{F,G}\tilde{\psi}_{P_{F,G}} = 0.$$

The kernel $\mathrm{N}(A^*_{F,G})$ consists of the functions $h \in L_2(P_{F,G})$ such that $h(u,0) = h(u,1)$ almost surely under F and G. This is precisely the range of $B_{F,G}$, and we can conclude that

$$\mathrm{R}(A_{F,G})^\perp = \mathrm{N}(A^*_{F,G}) = \mathrm{R}(B_{F,G}) = \mathrm{N}(B^*_{F,G})^\perp.$$

Therefore, we can solve the preceding display by first solving $A^*_{F,G}h = \tilde{\chi}_F$ and next projecting a solution h onto the closure of the range of $A_{F,G}$. By the orthogonality of the ranges of $A_{F,G}$ and $B_{F,G}$, the latter projection is the identity minus the projection onto $\mathrm{R}(B_{F,G})$. This is convenient, because the projection onto $\mathrm{R}(B_{F,G})$ is the conditional expectation relative to C.

For example, consider a function $\chi(F) = Fa$ for some fixed known, continuously differentiable function a. Differentiating the equation $a = A^*_{F,G}h$, we find $a'(c) = \big(h(c,0) - h(c,1)\big)g(c)$. This can happen for some $h \in L_2(P_{F,G})$ only if, for any τ such that $0 < F(\tau) < 1$,

$$\int_\tau^\infty \left(\frac{a'}{g}\right)^2 (1-F)\,dG = \int_\tau^\infty \big(h(u,0) - h(u,1)\big)^2 (1-F)(u)\,dG(u) < \infty,$$
$$\int_0^\tau \left(\frac{a'}{g}\right)^2 F\,dG = \int_0^\tau \big(h(u,0) - h(u,1)\big)^2 F(u)\,dG(u) < \infty.$$

If the left sides of these equations are finite, then the parameter $P_{F,G} \mapsto Fa$ is differentiable. An influence function is given by the function h defined by

$$h(c,0) = \frac{a'(c)1_{[\tau,\infty)}(c)}{g(c)}, \quad \text{and} \quad h(c,1) = -\frac{a'(c)1_{[0,\tau)}(c)}{g(c)}.$$

The efficient influence function is found by projecting this onto $\overline{\mathrm{R}}(A_{F,G})$, and is given by

$$\begin{aligned} h(c,\delta) - \mathrm{E}_{F,G}\big(h(C,\Delta) \mid C = c\big) &= \big(h(c,1) - h(c,0)\big)\big(\delta - F(c)\big) \\ &= -\delta\frac{1-F(c)}{g(c)}a'(c) + (1-\delta)\frac{F(c)}{g(c)}a'(c). \end{aligned}$$

For example, for the mean $\chi(F) = \int u\,dF(u)$, the influence function certainly exists if the density g is bounded away from zero on the compact support of F. □

*25.5.3 *Missing and Coarsening at Random*

Suppose that from a given vector (Y_1, Y_2) we sometimes observe only the first coordinate Y_1 and at other times both Y_1 and Y_2. Then Y_2 is said to be *missing at random (MAR)* if the conditional probability that Y_2 is observed depends only on Y_1, which is always observed. We can formalize this definition by introducing an indicator variable Δ that indicates whether Y_2 is missing ($\Delta = 0$) or observed ($\Delta = 1$). Then Y_2 is missing at random if $\mathrm{P}(\Delta = 0 \mid Y)$ is a function of Y_1 only.

If next to $\mathrm{P}(\Delta = 0 \mid Y)$ we also specify the marginal distribution of Y, then the distribution of (Y, Δ) is fixed, and the observed data are the function $X = \big(\phi(Y,\Delta), \Delta\big)$ defined by (for instance)

$$\phi(y,0) = y_1, \qquad \phi(y,1) = y.$$

The tangent set for the model for X can be derived from the tangent set for the model for (Y, Δ) by taking conditional expectations. If the distribution of (Y, Δ) is completely unspecified, then so is the distribution of X, and both tangent spaces are the maximal "nonparametric tangent space". If we restrict the model by requiring MAR, then the tangent set for (Y, Δ) is smaller than nonparametric. Interestingly, provided that we make no further restrictions, the tangent set for X remains the nonparametric tangent set.

We shall show this in somewhat greater generality. Let Y be an arbitrary unobservable "full observation" (not necessarily a vector) and let Δ be an arbitrary random variable. The distribution of (Y, Δ) can be determined by specifying a distribution Q for Y and a conditional density $r(\delta \mid y)$ for the conditional distribution of Δ given Y.[†] As before, we observe $X = \big(\phi(Y,\Delta), \Delta\big)$, but now ϕ may be an arbitrary measurable map. The observation X is said to be *coarsening at random (CAR)* if the conditional densities $r(\delta \mid y)$ depend on $x = \big(\phi(y,\delta), \delta\big)$ only, for every possible value (y, δ). More precisely, $r(\delta \mid y)$ is a measurable function of x.

25.39 ***Example (Missing at random).*** If $\Delta \in \{0, 1\}$ the requirements are both that $\mathrm{P}(\Delta = 0 \mid Y = y)$ depends only on $\phi(y, 0)$ and 0 and that $\mathrm{P}(\Delta = 1 \mid Y = y)$ depends only on $\phi(y, 1)$ and 1. Thus the two functions $y \mapsto \mathrm{P}(\Delta = 0 \mid Y = y)$ and $y \mapsto \mathrm{P}(\Delta = 1 \mid Y = y)$ may be different (fortunately) but may depend on y only through $\phi(y, 0)$ and $\phi(y, 1)$, respectively.

If $\phi(y, 1) = y$, then $\delta = 1$ corresponds to observing y completely. Then the requirement reduces to $\mathrm{P}(\Delta = 0 \mid Y = y)$ being a function of $\phi(y, 0)$ only. If $Y = (Y_1, Y_2)$ and $\phi(y, 0) = y_1$, then CAR reduces to MAR as defined in the introduction. □

[†] The density is relative to a dominating measure ν on the sample space for Δ, and we suppose that $(\delta, y) \mapsto r(\delta \mid y)$ is a Markov kernel.

Denote by $\mathcal{Q}$ and $\mathcal{R}$ the parameter spaces for the distribution Q of Y and the kernels $r(\delta \mid y)$ giving the conditional distribution of Δ given Y, respectively. Let $\mathcal{Q} \times \mathcal{R} = (Q \times R : Q \in \mathcal{Q}, R \in \mathcal{R})$ and $\mathcal{P} = (P_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R})$ be the models for (Y, Δ) and X, respectively.

25.40 ***Theorem.*** *Suppose that the distribution Q of Y is completely unspecified and the Markov kernel $r(\delta \mid y)$ is restricted by CAR, and only by CAR. Then there exists a tangent set $\dot{\mathcal{P}}_{P_{Q,R}}$ for the model $\mathcal{P} = (P_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R})$ whose closure consists of all mean-zero functions in $L_2(P_{Q,R})$. Furthermore, any element of $\dot{\mathcal{P}}_{P_{Q,R}}$ can be orthogonally decomposed as*

$$\mathrm{E}_{Q,R}\big(a(Y) \mid X = x\big) + b(x),$$

where $a \in \dot{\mathcal{Q}}_Q$ and $b \in \dot{\mathcal{R}}_R$. The functions a and b range exactly over the functions $a \in L_2(Q)$ with $Qa = 0$ and $b \in L_2(P_{Q,R})$ with $\mathrm{E}_R\big(b(X) \mid Y\big) = 0$ almost surely, respectively.

Proof. Fix a differentiable submodel $t \mapsto Q_t$ with score a. Furthermore, for every fixed y fix a differentiable submodel $t \mapsto r_t(\cdot \mid y)$ for the conditional density of Δ given $Y = y$ with score $b_0(\delta \mid y)$ such that

$$\int\int \left[\frac{r_t^{1/2}(\delta \mid y) - r^{1/2}(\delta \mid y)}{t} - \frac{1}{2} b_0(\delta \mid y) r^{1/2}(\delta \mid y)\right]^2 d\nu(\delta)\, dQ(y) \to 0.$$

Because the conditional densities satisfy CAR, the function $b_0(\delta \mid y)$ must actually be a function $b(x)$ of x only. Because it corresponds to a score for the conditional model, it is further restricted by the equations $\int b_0(\delta \mid y)\, r(\delta \mid y)\, d\nu(\delta) = \mathrm{E}_R\big(b(X) \mid Y = y\big) = 0$ for every y. Apart from this and square integrability, b_0 can be chosen freely, for instance bounded.

By a standard argument, with $Q \times R$ denoting the law of (Y, Δ) under Q and r,

$$\int \left[\frac{(dQ_t \times R_t)^{1/2} - (dQ \times R)^{1/2}}{t} - \frac{1}{2}\big(a(y) + b(x)\big)\,(dQ \times R)^{1/2}\right]^2 \to 0.$$

Thus $a(y) + b(x)$ is a score function for the model of (Y, Δ), at $Q \times R$. By Lemma 25.34 its conditional expectation $\mathrm{E}_{Q,R}\big(a(Y) + b(X) \mid X = x\big)$ is a score function for the model of X.

This proves that the functions as given in the theorem arise as scores. To show that the set of all functions of this type is dense in the nonparametric tangent set, suppose that some function $g \in L_2(P_{Q,R})$ is orthogonal to all functions $\mathrm{E}_{Q,R}\big(a(Y) \mid X = x\big) + b(x)$. Then $\mathrm{E}_{Q,R} g(X) a(Y) = \mathrm{E}_{Q,R} g(X) \mathrm{E}_{Q,R}\big(a(Y) \mid X\big) = 0$ for all a. Hence g is orthogonal to all functions of Y and hence is a function of the type b. If it is also orthogonal to all b, then it must be 0. ■

The interest of the representation of scores given in the preceding theorem goes beyond the case that the models $\mathcal{Q}$ and $\mathcal{R}$ are restricted by CAR only, as is assumed in the theorem. It shows that, under CAR, any tangent space for $\mathcal{P}$ can be decomposed into two orthogonal pieces, the first part consisting of the conditional expectations $\mathrm{E}_{Q,R}\big(a(Y) \mid X\big)$ of scores a for the model of Y (and their limits) and the second part being scores b for the model $\mathcal{R}$

describing the "missingness pattern." CAR ensures that the latter are functions of x already and need not be projected, and also that the two sets of scores are orthogonal. By the product structure of the likelihood $q(y)r(\delta \mid y)$, scores a and b for q and r in the model $\mathcal{Q} \times \mathcal{R}$ are always orthogonal. This orthogonality may be lost by projecting them on the functions of x, but not so under CAR, because b is equal to its projection.

In models in which there is a positive probability of observing the complete data, there is an interesting way to obtain all influence functions of a given parameter $P_{Q,R} \mapsto \chi(Q)$. Let C be a set of possible values of Δ leading to a complete observation, that is, $\phi(y, \delta) = y$ whenever $\delta \in C$, and suppose that $R(C \mid y) = \mathrm{P}_R(\Delta \in C \mid Y = y)$ is positive almost surely. Suppose for the moment that R is known, so that the tangent space for X consists only of functions of the form $\mathrm{E}_{Q,R}\big(a(Y) \mid X\big)$. If $\dot{\chi}_Q(y)$ is an influence function of the parameter $Q \mapsto \chi(Q)$ on the model $\mathcal{Q}$, then

$$\dot{\psi}_{P_{Q,R}}(x) = \frac{1\{\delta \in C\}}{R(C \mid y)} \dot{\chi}_Q(y)$$

is an influence function for the parameter $\psi(P_{F,G}) = \chi(Q)$ on the model $\mathcal{P}$. To see this, first note that, indeed, it is a function of x, as the indicator $1\{\delta \in C\}$ is nonzero only if $(y, \delta) = x$. Second,

$$\begin{aligned} \mathrm{E}_{Q,R} \dot{\psi}_{P_{Q,R}}(X) \mathrm{E}_{Q,R}\big(a(Y) \mid X\big) &= \mathrm{E}_{Q,R} \frac{1\{\Delta \in C\}}{R(C \mid Y)} \dot{\chi}_Q(Y)\, a(Y) \\ &= \mathrm{E}_{Q,R} \dot{\chi}_Q(Y)\, a(Y). \end{aligned}$$

The influence function we have found is just one of many influence functions, the other ones being obtainable by adding the orthocomplement of the tangent set. This particular influence function corresponds to ignoring incomplete observations altogether but reweighting the influence function for the full model to eliminate the bias caused by such neglect. Usually, ignoring all partial observations does not yield an efficient procedure, and correspondingly this influence function is usually not the efficient influence function.

All other influence functions, including the efficient influence function, can be found by adding the orthocomplement of the tangent set. An attractive way of doing this is:
– by varying $\dot{\chi}_Q$ over all possible influence functions for $Q \mapsto \chi(Q)$, combined with
– by adding all functions $b(x)$ with $\mathrm{E}_R\big(b(X) \mid Y\big) = 0$.
This is proved in the following lemma. We still assume that R is known; if it is not, then the resulting functions need not even be influence functions.

25.41 ***Lemma.*** *Suppose that the parameter $Q \mapsto \chi(Q)$ on the model $\mathcal{Q}$ is differentiable at Q, and that the conditional probability $R(C \mid Y) = \mathrm{P}(\Delta \in C \mid Y)$ of having a complete observation is bounded away from zero. Then the parameter $P_{Q,R} \mapsto \chi(Q)$ on the model $(P_{Q,R} : Q \in \mathcal{Q})$ is differentiable at $P_{Q,R}$ and any of its influence functions can be written in the form*

$$\frac{1\{\delta \in C\}}{R(C \mid y)} \dot{\chi}_Q(y) + b(x),$$

for $\dot{\chi}_Q$ an influence function of the parameter $Q \mapsto \chi(Q)$ on the model $\mathcal{Q}$ and a function $b \in L_2(P_{Q,R})$ satisfying $\mathrm{E}_R\big(b(X) \mid Y\big) = 0$. This decomposition is unique. Conversely, every function of this form is an influence function.

Proof. The function in the display with $b = 0$ has already been seen to be an influence function. (Note that it is square-integrable, as required.) Any function $b(X)$ such that $\mathrm{E}_R\big(b(X) \mid Y\big) = 0$ satisfies $\mathrm{E}_{Q,R} b(X) \mathrm{E}_{Q,R}\big(a(Y) \mid X\big) = 0$ and hence is orthogonal to the tangent set, whence it can be added to any influence function.

To see that the decomposition is unique, it suffices to show that the function as given in the lemma can be identically zero only if $\dot\chi_Q = 0$ and $b = 0$. If it is zero, then its conditional expectation with respect to Y, which is $\dot\chi_Q$, is zero, and reinserting this we find that $b = 0$ as well.

Conversely, an arbitrary influence function $\dot\psi_{P_{Q,R}}$ of $P_{Q,R} \mapsto \chi(Q)$ can be written in the form

$$\dot\psi_{P_{Q,R}}(x) = \frac{1\{\delta \in C\}}{R(C \mid y)} \dot\chi(y) + \left[\dot\psi_{P_{Q,R}}(x) - \frac{1\{\delta \in C\}}{R(C \mid y)} \dot\chi(y)\right].$$

For $\dot\chi(Y) = \mathrm{E}_R\big(\dot\psi_{P_{Q,R}}(X) \mid Y\big)$, the conditional expectation of the part within square brackets with respect to Y is zero and hence this part qualifies as a function b. This function $\dot\chi$ is an influence function for $Q \mapsto \chi(Q)$, as follows from the equality $\mathrm{E}_{Q,R}\mathrm{E}_R\big(\dot\psi_{P_{Q,R}}(X) \mid Y\big)a(Y) = \mathrm{E}_{Q,R}\dot\psi_{P_{Q,R}}(X)\mathrm{E}_{Q,R}\big(a(Y) \mid X\big)$ for every a. ■

Even though the functions $\dot\chi_Q$ and b in the decomposition given in the lemma are uniquely determined, the decomposition is not orthogonal, and (even under CAR) the decomposition does not agree with the decomposition of the (nonparametric) tangent space given in Theorem 25.40. The second term is as the functions b in this theorem, but the leading term is not in the maximal tangent set for Q.

The preceding lemma is valid without assuming CAR. Under CAR it obtains an interesting interpretation, because in that case the functions b range exactly over all scores for the parameter r that we would have had if R were completely unknown. If R is known, then these scores are in the orthocomplement of the tangent set and can be added to any influence function to find other influence functions.

A second special feature of CAR is that a similar representation becomes available in the case that R is (partially) unknown. Because the tangent set for the model $(P_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R})$ contains the tangent set for the model $(P_{Q,R} : Q \in \mathcal{Q})$ in which R is known, the influence functions for the bigger model are a subset of the influence functions of the smaller model. Because our parameter $\chi(Q)$ depends on Q only, they are exactly those influence functions in the smaller model that are orthogonal to the set ${}_R\dot{\mathcal{P}}_{P_{Q,R}}$ of all score functions for R. This is true in general, also without CAR. Under CAR they can be found by subtracting the projections onto the set of scores for R.

25.42 ***Corollary.*** *Suppose that the conditions of the preceding lemma hold and that the tangent space $\dot{\mathcal{P}}_{P_{Q,R}}$ for the model $(P_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R})$ is taken to be the sum ${}_Q\dot{\mathcal{P}}_{P_{Q,R}} + {}_R\dot{\mathcal{P}}_{P_{Q,R}}$ of tangent spaces of scores for Q and R separately. If ${}_Q\dot{\mathcal{P}}_{P_{Q,R}}$ and ${}_R\dot{\mathcal{P}}_{P_{Q,R}}$ are orthogonal, in particular under CAR, any influence function of $P_{Q,R} \mapsto \chi(Q)$ for the model $(P_{Q,R} : Q \in \mathcal{Q}, R \in \mathcal{R})$ can be obtained by taking the functions given by the preceding lemma and subtracting their projection onto $\overline{\mathrm{lin}}\,{}_R\dot{\mathcal{P}}_{P_{Q,R}}$.*

Proof. The influence functions for the bigger model are exactly those influence functions for the model in which R is known that are orthogonal to ${}_R\dot{\mathcal{P}}_{P_{Q,R}}$. These do not change

by subtracting their projection onto this space. Thus we can find all influence functions as claimed.

If the score spaces for Q and R are orthogonal, then the projection of an influence function onto $\overline{\text{lin}}\, {}_R\dot{\mathcal{P}}_{P_{Q,R}}$ is orthogonal to ${}_Q\dot{\mathcal{P}}_{P_{Q,R}}$, and hence the inner products with elements of this set are unaffected by subtracting it. Thus we necessarily obtain an influence function. ■

The efficient influence function $\tilde{\psi}_{P_{Q,R}}$ is an influence function and hence can be written in the form of Lemma 25.41 for some $\dot{\chi}_Q$ and b. By definition it is the unique influence function that is contained in the closed linear span of the tangent set. Because the parameter of interest depends on Q only, the efficient influence function is the same (under CAR or, more generally, if ${}_Q\dot{\mathcal{P}}_{P_{Q,R}} \perp {}_R\dot{\mathcal{P}}_{P_{Q,R}}$), whether we assume R known or not. One way of finding the efficient influence function is to minimize the variance of an arbitrary influence function as given in Lemma 25.41 over $\dot{\chi}_Q$ and b.

25.43 ***Example (Missing at random).*** In the case of MAR models there is a simple representation for the functions $b(x)$ in Lemma 25.41. Because MAR is a special case of CAR, these functions can be obtained by computing all the scores for R in the model for (Y, Δ) under the assumption that R is completely unknown, by Theorem 25.40. Suppose that Δ takes only the values 0 and 1, where 1 indicates a full observation, as in Example 25.39, and set $\pi(y) := \mathrm{P}(\Delta = 1 \mid Y = y)$. Under MAR $\pi(y)$ is actually a function of $\phi(y, 0)$ only. The likelihood for (Y, Δ) takes the form

$$q(y)r(\delta \mid y) = q(y)\pi(y)^{\delta}\big(1 - \pi(y)\big)^{1-\delta}.$$

Insert a path $\pi_t = \pi + tc$, and differentiate the log likelihood with respect to t at $t = 0$ to obtain a score for R of the form

$$\frac{\delta}{\pi(y)}c(y) - \frac{1-\delta}{1-\pi(y)}c(y) = \frac{\delta - \pi(y)}{\pi(y)(1-\pi)(y)}c(y).$$

To remain within the model the functions π_t and π, whence c, may depend on y only through $\phi(y, 0)$. Apart from this restriction, the preceding display gives a candidate for b in Lemma 25.41 for any c, and it gives all such b.

Thus, with a slight change of notation any influence function can be written in the form

$$\frac{\delta}{\pi(y)}\dot{\chi}_Q(y) - \frac{\delta - \pi(y)}{\pi(y)}c(y).$$

One approach to finding the efficient influence function in this case is first to minimize the variance of this influence function with respect to c and next to optimize over $\dot{\chi}_Q$. The first step of this plan can be carried out in general. Minimizing with respect to c is a weighted least-squares problem, whose solution is given by

$$\tilde{c}(Y) = \mathrm{E}_{Q,R}\big(\dot{\chi}_Q(Y) \mid \phi(Y, 0)\big).$$

To see this it suffices to verify the orthogonality relation, for all c,

$$\frac{\delta}{\pi(y)}\dot{\chi}_Q(y) - \frac{\delta - \pi(y)}{\pi(y)}\tilde{c}(y) \perp \frac{\delta - \pi(y)}{\pi(y)}c(y).$$

Splitting the inner product of these functions on the first minus sign, we obtain two terms, both of which reduce to $\mathrm{E}_{Q,R}\dot{\chi}_Q(Y)c(Y)(1-\pi)(Y)/\pi(Y)$. □

25.6 Testing

The problem of testing a null hypothesis $H_0 : \psi(P) \le 0$ versus the alternative $H_1 : \psi(P) > 0$ is closely connected to the problem of estimating the function $\psi(P)$. It ought to be true that a test based on an asymptotically efficient estimator of $\psi(P)$ is, in an appropriate sense, asymptotically optimal. For real-valued parameters $\psi(P)$ this optimality can be taken in the absolute sense of an asymptotically (locally) uniformly most powerful test. With higher-dimensional parameters we run into the same problem of defining a satisfactory notion of asymptotic optimality as encountered for parametric models in Chapter 15. We leave the latter case undiscussed and concentrate on real-valued functionals $\psi : \mathcal{P} \mapsto \mathbb{R}$.

Given a model $\mathcal{P}$ and a measure P on the boundary of the hypotheses, that is, $\psi(P) = 0$, we want to study the "local asymptotic power" in a neighborhood of P. Defining a local power function in the present infinite-dimensional case is somewhat awkward, because there is no natural "rescaling" of the parameter set, such as in the Euclidean case. We shall utilize submodels corresponding to a tangent set. Given an element g in a tangent set $\dot{\mathcal{P}}_P$, let $t \mapsto P_{t,g}$ be a differentiable submodel with score function g along which ψ is differentiable. For every such g for which $\dot{\psi}_P g = P\tilde{\psi}_P g > 0$, the submodel $P_{t,g}$ belongs to the alternative hypothesis H_1 for (at least) every sufficiently small, positive t, because $\psi(P_{t,g}) = tP\tilde{\psi}_P g + o(t)$ if $\psi(P) = 0$. We shall study the power at the alternatives $P_{h/\sqrt{n},g}$.

25.44 *Theorem.* *Let the functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. Suppose that $\psi(P) = 0$. Then for every sequence of power functions $P \mapsto \pi_n(P)$ of level-α tests for $H_0 : \psi(P) \le 0$, and every $g \in \dot{\mathcal{P}}_P$ with $P\tilde{\psi}_P g > 0$ and every $h > 0$,*

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

Proof. This theorem is essentially Theorem 15.4 applied to sufficiently rich submodels. Because the present situation does not fit exactly in the framework of Chapter 15, we rework the proof. Fix arbitrary h_1 and g_1 for which we desire to prove the upper bound. For notational convenience assume that $Pg_1^2 = 1$.

Fix an orthonormal base $g_P = (g_1, \ldots, g_m)^T$ of an arbitrary finite-dimensional subspace of $\dot{\mathcal{P}}_P$ (containing the fixed g_1). For every $g \in \operatorname{lin} g_P$, let $t \mapsto P_{t,g}$ be a submodel with score g along which the parameter ψ is differentiable. Each of the submodels $t \mapsto P_{t,g}$ is locally asymptotically normal at $t = 0$ by Lemma 25.14. Therefore, with S^{m-1} the unit sphere of $\mathbb{R}^m$,

$$\left(P^n_{h/\sqrt{n},a^T g_P} : h > 0, a \in S^{m-1}\right) \rightsquigarrow \left(N_m(ha, I) : h > 0, a \in S^{m-1}\right),$$

in the sense of convergence of experiments. Fix a subsequence along which the limsup in the statement of the theorem is taken for $h = h_1$ and $g = g_1$. By contiguity arguments, we can extract a further subsequence along which the functions $\pi_n(P_{h/\sqrt{n},a^T g})$ converge pointwise to a limit $\pi(h, a)$ for every (h, a). By Theorem 15.1, the function $\pi(h, a)$ is the power function of a test in the normal limit experiment. If it can be shown that this test is of level α for testing $H_0 : a^T P\tilde{\psi}_P g_P = 0$, then Proposition 15.2 shows that, for every (a, h)

with $a^T P\tilde{\psi}_P g_P > 0$,

$$\pi(h,a) \le 1 - \Phi\left(z_\alpha - h\frac{a^T P\tilde{\psi}_P g_P}{\left(P\tilde{\psi}_P g_P^T P\tilde{\psi}_P g_P\right)^{1/2}}\right).$$

The orthogonal projection of $\tilde{\psi}_P$ onto lin g_P is equal to $(P\tilde{\psi}_P g_P^T)g_P$, and has length $P\tilde{\psi}_P g_P^T P\tilde{\psi}_P g_P$. By choosing lin g_P large enough, we can ensure that this length is arbitrarily close to $P\tilde{\psi}_P^2$. Choosing $(h,a) = (h_1, e_1)$ completes the proof, because $\limsup \pi_n(P_{h_1/\sqrt{n},g_1}) \le \pi(h_1, e_1)$, by construction.

To complete the proof, we show that π is of level α. Fix any $h > 0$ and an $a \in S^{m-1}$ such that $a^T P\tilde{\psi}_P g_P < 0$. Then

$$\psi(P_{h/\sqrt{n},a^T g}) = \psi(P) + \frac{h}{\sqrt{n}}\left(a^T P\tilde{\psi}_P g_P + o(1)\right)$$

is negative for sufficiently large n. Hence $P_{h/\sqrt{n},a^T g}$ belongs to H_0 and

$$\pi(h,a) = \lim \pi_n(P_{h/\sqrt{n},a^T g}) \le \alpha.$$

Thus, the test with power function π is of level α for testing $H_0 : a^T P\tilde{\psi}_P g_P < 0$. By continuity it is of level α for testing $H_0 : a^T P\tilde{\psi}_P g_P \le 0$. ■

As a consequence of the preceding theorem, a test based on an efficient estimator for $\psi(P)$ is automatically "locally uniformly most powerful": Its power function attains the upper bound given by the theorem. More precisely, suppose that the sequence of estimators T_n is asymptotically efficient at P and that S_n is a consistent sequence of estimators of its asymptotic variance. Then the test that rejects $H_0 : \psi(P) = 0$ for $\sqrt{n}T_n/S_n \ge z_\alpha$ attains the upper bound of the theorem.

25.45 ***Lemma.*** *Let the functional $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P with $\psi(P) = 0$. Suppose that the sequence T_n is regular at P with a $N(0, P\tilde{\psi}_P^2)$-limit distribution. Furthermore, suppose that $S_n^2 \xrightarrow{\mathrm{P}} P\tilde{\psi}_P^2$. Then, for every $h \ge 0$ and $g \in \dot{\mathcal{P}}_P$,*

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

Proof. By the efficiency of T_n and the differentiability of ψ, the sequence $\sqrt{n}T_n$ converges under $P_{h/\sqrt{n},g}$ to a normal distribution with mean $hP\tilde{\psi}_P g$ and variance $P\tilde{\psi}_P^2$. ■

25.46 ***Example (Wilcoxon test).*** Suppose that the observations are two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ from distribution functions F and G, respectively. To fit this two-sample problem in the present i.i.d. set-up, we pair the two samples and think of (X_i, Y_i) as a single observation from the product measure $F \times G$ on $\mathbb{R}^2$. We wish to test the null hypothesis $H_0 : \int F\,dG \le \frac{1}{2}$ versus the alternative $H_1 : \int F\,dG > \frac{1}{2}$. The Wilcoxon test, which rejects for large values of $\int \mathbb{F}_n\,d\mathbb{G}_n$, is asymptotically efficient, relative to the model in which F and G are completely unknown. This gives a different perspective on this test, which in Chapters 14 and 15 was seen to be asymptotically efficient for testing location in the logistic location-scale family. Actually, this finding is an

example of the general principle that, in the situation that the underlying distribution of the observations is completely unknown, empirical-type statistics are asymptotically efficient for whatever they naturally estimate or test (also see Example 25.24 and section 25.7). The present conclusion concerning the Wilcoxon test extends to most other test statistics.

By the preceding lemma, the efficiency of the test follows from the efficiency of the Wilcoxon statistic as an estimator for the function $\psi(F \times G) = \int F\,dG$. This may be proved by Theorem 25.47, or by the following direct argument.

The model $\mathcal{P}$ is the set of all product measures $F \times G$. To generate a tangent set, we can perturb both F and G. If $t \mapsto F_t$ and $t \mapsto G_t$ are differentiable submodels (of the collection of all probability distributions on $\mathbb{R}$) with score functions a and b at $t = 0$, respectively, then the submodel $t \mapsto F_t \times G_t$ has score function $a(x) + b(y)$. Thus, as a tangent space we may take the set of all square-integrable functions with mean zero of this type. For simplicity, we could restrict ourselves to bounded functions a and b and use the paths $dF_t = (1 + ta)\,dF$ and $dG_t = (1 + tb)\,dG$. The closed linear span of the resulting tangent set is the same as before. Then, by simple algebra,

$$\dot{\psi}_{F\times G}(a, b) = \frac{\partial}{\partial t}\psi(F_t \times G_t)_{|t=0} = \int (1 - G_-)a\,dF + \int Fb\,dG.$$

We conclude that the function $(x, y) \mapsto (1 - G_-)(x) + F(y)$ is an influence function of ψ. This is of the form $a(x) + b(y)$ but does not have mean zero; the efficient influence function is found by subtracting the mean.

The efficiency of the Wilcoxon statistic is now clear from Lemma 25.23 and the asymptotic linearity of the Wilcoxon statistic, which is proved by various methods in Chapters 12, 13, and 20. □

*25.7 Efficiency and the Delta Method

Many estimators can be written as functions $\phi(T_n)$ of other estimators. By the delta method asymptotic normality of T_n carries over into the asymptotic normality of $\phi(T_n)$, for every differentiable map ϕ. Does efficiency of T_n carry over into efficiency of $\phi(T_n)$ as well? With the right definitions, the answer ought to be affirmative. The matter is sufficiently useful to deserve a discussion and turns out to be nontrivial. Because the result is true for the functional delta method, applications include the efficiency of the product-limit estimator in the random censoring model and the sample median in the nonparametric model, among many others.

If T_n is an estimator of a Euclidean parameter $\psi(P)$ and both ϕ and ψ are differentiable, then the question can be answered by a direct calculation of the normal limit distributions. In view of Lemma 25.23, efficiency of T_n can be defined by the asymptotic linearity (25.22). By the delta method,

$$\begin{aligned}\sqrt{n}\big(\phi(T_n) - \phi \circ \psi(P)\big) &= \phi'_{\psi(P)}\sqrt{n}\big(T_n - \psi(P)\big) + o_P(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\phi'_{\psi(P)}\tilde{\psi}_P(X_i) + o_P(1).\end{aligned}$$

The asymptotic efficiency of $\phi(T_n)$ follows, provided that the function $x \mapsto \phi'_{\psi(P)}\tilde{\psi}_P(x)$ is the efficient influence function of the parameter $P \mapsto \phi \circ \psi(P)$. If the coordinates of

$\tilde{\psi}_P$ are contained in the closed linear span of the tangent set, then so are the coordinates of $\phi'_{\psi(P)}\tilde{\psi}_P$, because the matrix multiplication by $\phi'_{\psi(P)}$ means taking linear combinations. Furthermore, if ψ is differentiable at P (as a statistical parameter on the model $\mathcal{P}$) and ϕ is differentiable at $\psi(P)$ (in the ordinary sense of calculus), then

$$\frac{\phi \circ \psi(P_t) - \phi \circ \psi(P)}{t} \to \phi'_{\psi(P)}\dot{\psi}_P g = P\phi'_{\psi(P)}\tilde{\psi}_P g.$$

Thus the function $\phi'_{\psi(P)}\tilde{\psi}_P$ is an influence function and hence the efficient influence function.

More involved is the same question, but with T_n an estimator of a parameter in a Banach space, for instance a distribution in the space $D[-\infty, \infty]$ or in a space $\ell^\infty(\mathcal{F})$. The question is empty until we have defined efficiency for this situation. A definition of asymptotic efficiency of Banach-valued estimators can be based on generalizations of the convolution and minimax theorems to general Banach spaces.† We shall avoid this route and take a more naive approach.

The *dual space* $\mathbb{D}^*$ of a Banach space $\mathbb{D}$ is defined as the collection of all continuous, linear maps $d^* : \mathbb{D} \mapsto \mathbb{R}$. If T_n is a $\mathbb{D}$-valued estimator for a parameter $\psi(P) \in \mathbb{D}$, then d^*T_n is a real-valued estimator for the parameter $d^*\psi(P) \in \mathbb{R}$. This suggests to defining T_n to be *asymptotically efficient* at $P \in \mathcal{P}$ if $\sqrt{n}\big(T_n - \psi(P)\big)$ converges under P in distribution to a tight limit and d^*T_n is asymptotically efficient at P for estimating $d^*\psi(P)$, for every $d^* \in \mathbb{D}^*$.

This definition presumes that the parameters $d^*\psi$ are differentiable at P in the sense of section 25.3. We shall require a bit more. Say that $\psi : \mathcal{P} \mapsto \mathbb{D}$ is *differentiable* at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear map $\dot{\psi}_P : L_2(P) \mapsto \mathbb{D}$ such that, for every $g \in \dot{\mathcal{P}}_P$ and a submodel $t \mapsto P_t$ with score function g,

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

This implies that every parameter $d^*\psi : \mathcal{P} \mapsto \mathbb{R}$ is differentiable at P, whence, for every $d^* \in \mathbb{D}^*$, there exists a function $\tilde{\psi}_{P,d^*} : \mathcal{X} \mapsto \mathbb{R}$ in the closed linear span of $\dot{\mathcal{P}}_P$ such that $d^*\dot{\psi}_P(g) = P\tilde{\psi}_{P,d^*}g$ for every $g \in \dot{\mathcal{P}}_P$. The efficiency of d^*T_n for $d^*\psi$ can next be understood in terms of asymptotic linearity of $d^*\sqrt{n}\big(T_n - \psi(P)\big)$, as in (25.22), with influence function $\tilde{\psi}_{P,d^*}$.

To avoid measurability issues, we also allow nonmeasurable functions $T_n = T_n(X_1, \ldots, X_n)$ of the data as estimators in this section. Let both $\mathbb{D}$ and $\mathbb{E}$ be Banach spaces.

25.47 *Theorem.* *Suppose that $\psi : \mathcal{P} \mapsto \mathbb{D}$ is differentiable at P and takes its values in a subset $\mathbb{D}_\phi \subset \mathbb{D}$, and suppose that $\phi : \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$ is Hadamard-differentiable at $\psi(P)$ tangentially to $\overline{\operatorname{lin}}\,\dot{\psi}_P(\dot{\mathcal{P}}_P)$. Then $\phi \circ \psi : \mathcal{P} \mapsto \mathbb{E}$ is differentiable at P. If T_n is a sequence of estimators with values in $\mathbb{D}_\phi$ that is asymptotically efficient at P for estimating $\psi(P)$, then $\phi(T_n)$ is asymptotically efficient at P for estimating $\phi \circ \psi(P)$.*

Proof. The differentiability of $\phi \circ \psi$ is essentially a consequence of the chain rule for Hadamard-differentiable functions (see Theorem 20.9) and is proved in the same way. The derivative is the composition $\phi'_{\psi(P)} \circ \dot{\psi}_P$.

† See for example, Chapter 3.11 in [146] for some possibilities and references.

First, we show that the limit distribution L of the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ concentrates on the subspace $\overline{\operatorname{lin}}\,\dot\psi_P(\dot{\mathcal{P}}_P)$. By the Hahn-Banach theorem, for any $S \subset \mathbb{D}$,

$$\overline{\operatorname{lin}}\,\dot\psi_P(\dot{\mathcal{P}}_P) \cap S = \cap_{d^* \in \mathbb{D}^*: d^*\dot\psi_P = 0} \big\{d \in S : d^* d = 0\big\}.$$

For a separable set S, we can replace the intersection by a countable subintersection. Because L is tight, it concentrates on a separable set S, and hence L gives mass 1 to the left side provided $L(d : d^* d = 0) = 1$ for every d^* as on the right side. This probability is equal to $N\big(0, \|\tilde\psi_{d^*P}\|_P^2\big)\{0\} = 1$.

Now we can conclude that under the assumptions the sequence $\sqrt{n}\big(\phi(T_n) - \phi\circ\psi(P)\big)$ converges in distribution to a tight limit, by the functional delta method, Theorem 20.8. Furthermore, for every $e^* \in \mathbb{E}^*$

$$\sqrt{n}\big(e^*\phi(T_n) - e^*\phi\circ\psi(P)\big) = e^*\phi'_{\psi(P)}\sqrt{n}\big(T_n - \psi(P)\big) + o_P(1),$$

where, if necessary, we can extend the definition of $d^* = e^*\phi'_{\psi(P)}$ to all of $\mathbb{D}$ in view of the Hahn-Banach theorem. Because $d^* \in \mathbb{D}^*$, the asymptotic efficiency of the sequence T_n implies that the latter sequence is asymptotically linear in the influence function $\tilde\psi_{P,d^*}$. This is also the influence function of the real-valued map $e^*\phi\circ\psi$, because

$$e^*\phi'_{\psi(P)} \circ \dot\psi_P g = d^*\dot\psi_P g = P\tilde\psi_{P,d^*} g, \qquad g \in \dot{\mathcal{P}}_P.$$

Thus, $e^*\phi(T_n)$ is asymptotically efficient at P for estimating $e^*\phi\circ\psi(P)$, for every $e^* \in \mathbb{E}^*$. ■

The proof of the preceding theorem is relatively simple, because our definition of an efficient estimator sequence, although not unnatural, is relatively involved.

Consider, for instance, the case that $\mathbb{D} = \ell^\infty(S)$ for some set S. This corresponds to estimating a (bounded) function $s \mapsto \psi(P)(s)$ by a random function $s \mapsto T_n(s)$. Then the "marginal estimators" d^*T_n include the estimators $\pi_s T_n = T_n(s)$ for every fixed s – the coordinate projections $\pi_s : d \mapsto d(s)$ are elements of the dual space $\ell^\infty(S)^*$–, but include many other, more complicated functions of T_n as well. Checking the efficiency of every marginal of the general type d^*T_n may be cumbersome.

The deeper result of this section is that this is not necessary. Under the conditions of Theorem 17.14, the limit distribution of the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ in $\ell^\infty(S)$ is determined by the limit distributions of these processes evaluated at finite sets of "times" $s_1, \ldots, s_k$. Thus, we may hope that the asymptotic efficiency of T_n can also be characterized by the behavior of the marginals $T_n(s)$ only. Our definition of a differentiable parameter $\psi : \mathcal{P} \mapsto \mathbb{D}$ is exactly right for this purpose.

25.48 *Theorem (Efficiency in $\ell^\infty(S)$).* *Suppose that $\psi : \mathcal{P} \mapsto \ell^\infty(S)$ is differentiable at P, and suppose that $T_n(s)$ is asymptotically efficient at P for estimating $\psi(P)(s)$, for every $s \in S$. Then T_n is asymptotically efficient at P provided that the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ converges under P in distribution to a tight limit in $\ell^\infty(S)$.*

The theorem is a consequence of a more general principle that obtains the efficiency of T_n from the efficiency of d^*T_n for a sufficient number of elements $d^* \in \mathbb{D}^*$. By definition, efficiency of T_n means efficiency of d^*T_n for all $d^* \in \mathbb{D}^*$. In the preceding theorem the efficiency is deduced from efficiency of the estimators $\pi_s T_n$ for all coordinate projections π_s

on $\ell^\infty(S)$. The coordinate projections are a fairly small subset of the dual space of $\ell^\infty(S)$. What makes them work is the fact that they are of norm 1 and satisfy $\|z\|_S = \sup_s |\pi_s z|$.

25.49 ***Lemma.*** *Suppose that $\psi : \mathcal{P} \mapsto \mathbb{D}$ is differentiable at P, and suppose that $d'T_n$ is asymptotically efficient at P for estimating $d'\psi(P)$ for every d' in a subset $\mathbb{D}' \subset \mathbb{D}^*$ such that, for some constant C,*

$$\|d\| \le C \sup_{d' \in \mathbb{D}', \|d'\| \le 1} \left|d'(d)\right|.$$

Then T_n is asymptotically efficient at P provided that the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ is asymptotically tight under P.

Proof. The efficiency of all estimators $d'T_n$ for every $d' \in \mathbb{D}'$ implies their asymptotic linearity. This shows that $d'T_n$ is also asymptotically linear and efficient for every $d' \in \operatorname{lin} \mathbb{D}'$. Thus, it is no loss of generality to assume that $\mathbb{D}'$ is a linear space.

By Prohorov's theorem, every subsequence of $\sqrt{n}\big(T_n - \psi(P)\big)$ has a further subsequence that converges weakly under P to a tight limit T. For simplicity, assume that the whole sequence converges; otherwise argue along subsequences. By the continuous-mapping theorem, $d^*\sqrt{n}\big(T_n - \psi(P)\big)$ converges in distribution to d^*T for every $d^* \in \mathbb{D}^*$. By the assumption of efficiency, the sequence $d^*\sqrt{n}\big(T_n - \psi(P)\big)$ is asymptotically linear in the influence function $\tilde{\psi}_{P,d^*}$ for every $d^* \in \mathbb{D}'$. Thus, the variable d^*T is normally distributed with mean zero and variance $P\tilde{\psi}^2_{P,d^*}$ for every $d^* \in \mathbb{D}'$. We show below that this is then automatically true for every $d^* \in \mathbb{D}^*$.

By Le Cam's third lemma (which by inspection of its proof can be seen to be valid for general metric spaces), the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ is asymptotically tight under $P_{1/\sqrt{n}}$ as well, for every differentiable path $t \mapsto P_t$. By the differentiability of ψ, the sequence $\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n}})\big)$ is tight also. Then, exactly as in the preceding paragraph, we can conclude that the sequence $d^*\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n}})\big)$ converges in distribution to a normal distribution with mean zero and variance $P\tilde{\psi}^2_{P,d^*}$, for every $d^* \in \mathbb{D}^*$. Thus, d^*T_n is asymptotically efficient for estimating $d^*\psi(P)$ for every $d^* \in \mathbb{D}^*$ and hence T_n is asymptotically efficient for estimating $\psi(P)$, by definition.

It remains to prove that a tight, random element T in $\mathbb{D}$ such that d^*T has law $N\big(0, \|d^*\dot{\psi}_P\|^2\big)$ for every $d^* \in \mathbb{D}'$ necessarily verifies this same relation for every $d^* \in \mathbb{D}^*$.[†] First assume that $\mathbb{D} = \ell^\infty(S)$ and that $\mathbb{D}'$ is the linear space spanned by all coordinate projections.

Because T is tight, there exists a semimetric ρ on S such that S is totally bounded and almost all sample paths of T are contained in $UC(S, \rho)$ (see Lemma 18.15). Then automatically the range of $\dot{\psi}_P$ is contained in $UC(S, \rho)$ as well.

To see the latter, we note first that the map $s \mapsto \mathrm{E}T(s)T(u)$ is contained in $UC(S, \rho)$ for every fixed u: If $\rho(s_m, t_m) \to 0$, then $T(s_m) - T(t_m) \to 0$ almost surely and hence in second mean, in view of the zero-mean normality of $T(s_m) - T(t_m)$ for every m, whence $\left|\mathrm{E}T(s_m)T(u) - \mathrm{E}T(t_m)T(u)\right| \to 0$ by the Cauchy-Schwarz inequality. Thus, the map

$$s \mapsto \dot{\psi}_P\big(\tilde{\psi}_{P,\pi_u}\big)(s) = \pi_s\dot{\psi}_P\big(\tilde{\psi}_{P,\pi_u}\big) = \big\langle \tilde{\psi}_{P,\pi_u}, \tilde{\psi}_{P,\pi_s}\big\rangle_P = \mathrm{E}T(u)T(s)$$

[†] The proof of this lemma would be considerably shorter if we knew already that there exists a tight random element T with values in $\mathbb{D}$ such that d^*T has a $N\big(0, \|d^*\dot{\psi}_P\|^2_{P,2}\big)$-distribution for every $d^* \in \mathbb{D}^*$. Then it suffices to show that the distribution of T is uniquely determined by the distributions of d^*T for $d^* \in \mathbb{D}'$.

is contained in the space $UC(S, \rho)$ for every u. By the linearity and continuity of the derivative $\dot{\psi}_P$, the same is then true for the map $s \mapsto \dot{\psi}_P(g)(s)$ for every g in the closed linear span of the gradients $\tilde{\psi}_{P,\pi_u}$ as u ranges over S. It is even true for every g in the tangent set, because $\dot{\psi}_P(g)(s) = \dot{\psi}_P(\Pi g)(s)$ for every g and s, and Π the projection onto the closure of lin $\tilde{\psi}_{P,\pi_u}$.

By a minor extension of the Riesz representation theorem for the dual space of $C(\overline{S}, \rho)$, the restriction of a fixed $d^* \in \mathbb{D}^*$ to $UC(S, \rho)$ takes the form

$$d^*z = \int_{\overline{S}} \overline{z}(s)\, d\overline{\mu}(s),$$

for $\overline{\mu}$ a signed Borel measure on the completion $\overline{S}$ of S, and $\overline{z}$ the unique continuous extension of z to $\overline{S}$. By discretizing $\overline{\mu}$, using the total boundedness of S, we can construct a sequence d_m^* in lin $\{\pi_s : s \in S\}$ such that $d_m^* \to d^*$ pointwise on $UC(S, \rho)$. Then $d_m^*\dot{\psi}_P \to d^*\dot{\psi}_P$ pointwise on $\dot{\mathcal{P}}_P$. Furthermore, $d_m^*T \to d^*T$ almost surely, whence in distribution, so that d^*T is normally distributed with mean zero. Because $d_m^*T - d_n^*T \to 0$ almost surely, we also have that

$$\mathrm{E}(d_m^*T - d_n^*T)^2 = \|d_m^*\dot{\psi}_P - d_n^*\dot{\psi}_P\|_{P,2}^2 \to 0,$$

whence $d_m^*\dot{\psi}_P$ is a Cauchy sequence in $L_2(P)$. We conclude that $d_m^*\dot{\psi}_P \to d^*\dot{\psi}_P$ also in norm and $\mathrm{E}(d_m^*T)^2 = \|d_m^*\dot{\psi}_P\|_{P,2}^2 \to \|d^*\dot{\psi}_P\|_{P,2}^2$. Thus, d^*T is normally distributed with mean zero and variance $\|d^*\dot{\psi}_P\|_{P,2}^2$.

This concludes the proof for $\mathbb{D}$ equal to $\ell^\infty(S)$. A general Banach space $\mathbb{D}$ can be embedded in $\ell^\infty(\mathbb{D}_1')$, for $\mathbb{D}_1' = \{d' \in \mathbb{D}', \|d'\| \le 1\}$, by the map $d \to z_d$ defined as $z_d(d') = d'(d)$. By assumption, this map is a norm homeomorphism, whence T can be considered to be a tight random element in $\ell^\infty(\mathbb{D}_1')$. Next, the preceding argument applies. ■

Another useful application of the lemma concerns the estimation of functionals $\psi(P) = (\psi_1(P), \psi_2(P))$ with values in a product $\mathbb{D}_1 \times \mathbb{D}_2$ of two Banach spaces. Even though marginal weak convergence does not imply joint weak convergence, marginal efficiency implies joint efficiency!

25.50 ***Theorem (Efficiency in product spaces).*** *Suppose that $\psi_i : \mathcal{P} \mapsto \mathbb{D}_i$ is differentiable at P, and suppose that $T_{n,i}$ is asymptotically efficient at P for estimating $\psi_i(P)$, for $i = 1, 2$. Then $(T_{n,1}, T_{n,2})$ is asymptotically efficient at P for estimating $(\psi_1(P), \psi_2(P))$ provided that the sequences $\sqrt{n}(T_{n,i} - \psi_i(P))$ are asymptotically tight in $\mathbb{D}_i$ under P, for $i = 1, 2$.*

Proof. Let $\mathbb{D}'$ be the set of all maps $(d_1, d_2) \mapsto d_i^*(d_i)$ for d_i^* ranging over $\mathbb{D}_i^*$, and $i = 1, 2$. By the Hahn-Banach theorem, $\|d_i\| = \sup\{|d_i^*(d_i)| : \|d_i^*\| = 1, d_i^* \in \mathbb{D}_i^*\}$. Thus, the product norm $\|(d_1, d_2)\| = \|d_1\| \vee \|d_2\|$ satisfies the condition of the preceding lemma (with $C = 1$ and equality). ■

25.51 ***Example (Random censoring).*** In section 25.10.1 it is seen that the distribution of $X = (C \wedge T, 1\{T \le C\})$ in the random censoring model can be any distribution on the sample space. It follows by Example 20.16 that the empirical subdistribution functions $\mathbb{H}_{0n}$ and $\mathbb{H}_{1n}$ are asymptotically efficient. By Example 20.15 the product limit estimator is a Hadamard-differentiable functional of the empirical subdistribution functions. Thus, the product limit-estimator is asymptotically efficient. □

25.8 Efficient Score Equations

The most important method of estimating the parameter in a parametric model is the method of maximum likelihood, and it can usually be reduced to solving the score equations $\sum_{i=1}^n \dot{\ell}_\theta(X_i) = 0$, if necessary in a neighborhood of an initial estimate. A natural generalization to estimating the parameter θ in a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$ is to solve θ from the *efficient score equations*

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

Here we use the efficient score function instead of the ordinary score function, and we substitute an estimator $\hat{\eta}_n$ for the unknown nuisance parameter. A refinement of this method has been applied successfully to a number of examples, and the method is likely to work in many other examples. A disadvantage is that the method requires an explicit form of the efficient score function, or an efficient algorithm to compute it. Because, in general, the efficient score function is defined only implicitly as an orthogonal projection, this may preclude practical implementation.

A variation on this approach is to obtain an estimator $\hat{\eta}_n(\theta)$ of η for each given value of θ, and next to solve θ from the equation

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

If $\hat{\theta}_n$ is a solution, then it is also a solution of the estimating equation in the preceding display, for $\hat{\eta}_n = \hat{\eta}_n(\hat{\theta}_n)$. The asymptotic normality of $\hat{\theta}_n$ can therefore be proved by the same methods as applying to this estimating equation. Due to our special choice of estimating function, the nature of the dependence of $\hat{\eta}_n(\theta)$ on θ should be irrelevant for the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$. Informally, this is because the partial derivative of the estimating equation relative to the θ inside $\hat{\eta}_n(\theta)$ should converge to zero, as is clear from our subsequent discussion of the "no-bias" condition (25.52). The dependence of $\hat{\eta}_n(\theta)$ on θ does play a role for the consistency of $\hat{\theta}_n$, but we do not discuss this in this chapter, because the general methods of Chapter 5 apply. For simplicity we adopt the notation as in the first estimating equation, even though for the construction of $\hat{\theta}_n$ the two-step procedure, which "profiles out" the nuisance parameter, may be necessary.

In a number of applications the nuisance parameter η, which is infinite-dimensional, cannot be estimated within the usual order $O(n^{-1/2})$ for parametric models. Then the classical approach to derive the asymptotic behavior of Z-estimators – linearization of the equation in both parameters – is impossible. Instead, we utilize the notion of a Donsker class, as developed in Chapter 19. The auxiliary estimator for the nuisance parameter should satisfy†

$$P_{\hat{\theta}_n,\eta}\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} = o_P\big(n^{-1/2} + \|\hat{\theta}_n - \theta\|\big), \tag{25.52}$$

$$P_{\theta,\eta}\big\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} - \tilde{\ell}_{\theta,\eta}\big\|^2 \xrightarrow{P} 0, \qquad P_{\hat{\theta}_n,\eta}\big\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\big\|^2 = O_P(1). \tag{25.53}$$

† The notation $P\ell_{\hat{\eta}}$ is an abbreviation for the integral $\int \ell_{\hat{\eta}}(x)\,dP(x)$. Thus the expectation is taken with respect to x only and not with respect to $\hat{\eta}$.

The second condition (25.53) merely requires that the "plug-in" estimator $\tilde{\ell}_{\theta,\hat{\eta}_n}$ is a consistent estimator for the true efficient influence function. Because $P_{\hat{\theta}_n,\eta}\tilde{\ell}_{\hat{\theta}_n,\eta} = 0$, the first condition (25.52) requires that the "bias" of the plug-in estimator, due to estimating the nuisance parameter, converge to zero faster than $1/\sqrt{n}$. Such a condition comes out naturally of the proofs. A partial motivation is that the efficient score function is orthogonal to the score functions for the nuisance parameter, so that its expectation should be insensitive to changes in η.

25.54 *Theorem.* *Suppose that the model $\{P_{\theta,\eta} : \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ, η) and let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular. Assume that* (25.52) *and* (25.53) *hold. Let $\hat{\theta}_n$ satisfy $\sqrt{n}\,\mathbb{P}_n\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} = o_{\hat{P}}(1)$ and be consistent for θ. Furthermore, suppose that there exists a Donsker class with square-integrable envelope function that contains every function $\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}$ with probability tending to* 1. *Then the sequence $\hat{\theta}_n$ is asymptotically efficient at (θ, η).*

Proof. Let $G_n(\theta', \eta') = \sqrt{n}(\mathbb{P}_n - P_{\theta,\eta})\tilde{\ell}_{\theta',\eta'}$ be the empirical process indexed by the functions $\tilde{\ell}_{\theta',\eta'}$. By the assumption that the functions $\tilde{\ell}_{\hat{\theta},\hat{\eta}}$ are contained in a Donsker class, together with (25.53),

$$G_n(\hat{\theta}_n, \hat{\eta}_n) = G_n(\theta, \eta) + o_P(1).$$

(see Lemma 19.24.) By the defining relationship of $\hat{\theta}_n$ and the "no-bias" condition (25.52), this is equivalent to

$$\sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} = G_n(\theta, \eta) + o_P\bigl(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|\bigr).$$

The remainder of the proof consists of showing that the left side is asymptotically equivalent to $\bigl(\tilde{I}_{\theta,\eta} + o_P(1)\bigr)\sqrt{n}(\hat{\theta}_n - \theta)$, from which the theorem follows. Because $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}_{\theta,\eta}^T$, the difference of the left side of the preceding display and $\tilde{I}_{\theta,\eta}\sqrt{n}(\hat{\theta}_n - \theta)$ can be written as the sum of three terms:

$$\begin{aligned}
&\sqrt{n}\int \tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\bigl(p_{\hat{\theta}_n,\eta}^{1/2} + p_{\theta,\eta}^{1/2}\bigr)\Bigl[\bigl(p_{\hat{\theta}_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2}\bigr) - \frac{1}{2}(\hat{\theta}_n - \theta)^T\dot{\ell}_{\theta,\eta}\,p_{\theta,\eta}^{1/2}\Bigr]\,d\mu \\
&\quad + \int \tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\bigl(p_{\hat{\theta}_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2}\bigr)\frac{1}{2}\dot{\ell}_{\theta,\eta}^T\,p_{\theta,\eta}^{1/2}\,d\mu\,\sqrt{n}(\hat{\theta}_n - \theta) \\
&\quad - \int \bigl(\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} - \tilde{\ell}_{\theta,\eta}\bigr)\dot{\ell}_{\theta,\eta}^T\,p_{\theta,\eta}\,d\mu\,\sqrt{n}(\hat{\theta}_n - \theta).
\end{aligned}$$

The first and third term can easily be seen to be $o_P\bigl(\sqrt{n}\|\hat{\theta}_n - \theta\|\bigr)$ by applying the Cauchy-Schwarz inequality together with the differentiability of the model and (25.53). The square of the norm of the integral in the middle term can for every sequence of constants $m_n \to \infty$ be bounded by a multiple of

$$\begin{aligned}
&m_n^2\int \bigl\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\bigr\|\,p_{\theta,\eta}^{1/2}\bigl|p_{\hat{\theta}_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2}\bigr|\,d\mu^2 \\
&\quad + \int \bigl\|\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}\bigr\|^2(p_{\hat{\theta}_n,\eta} + p_{\theta,\eta})\,d\mu \int_{\|\dot{\ell}_{\theta,\eta}\| > m_n} \|\dot{\ell}_{\theta,\eta}\|^2\,p_{\theta,\eta}\,d\mu.
\end{aligned}$$

In view of (25.53), the differentiability of the model in θ, and the Cauchy-Schwarz inequality, the first term converges to zero in probability provided $m_n \to \infty$ sufficiently slowly

to ensure that $m_n\|\hat{\theta}_n - \theta\| \xrightarrow{P} 0$. (Such a sequence exists. If $Z_n \xrightarrow{P} 0$, then there exists a sequence $\varepsilon_n \downarrow 0$ such that $P(|Z_n| > \varepsilon_n) \to 0$. Then $\varepsilon_n^{-1/2} Z_n \xrightarrow{P} 0$.) In view of the last part of (25.53), the second term converges to zero in probability for every $m_n \to \infty$. This concludes the proof of the theorem. ■

The preceding theorem is best understood as applying to the efficient score functions $\tilde{\ell}_{\theta,\eta}$. However, its proof only uses this to ensure that, at the true value (θ, η),

$$\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}_{\theta\eta}^T.$$

The theorem remains true for arbitrary, mean-zero functions $\tilde{\ell}_{\theta,\eta}$ provided that this identity holds. Thus, if an estimator $(\hat{\theta}, \hat{\eta})$ only approximately satisfies the efficient score equation, then the latter can be replaced by an approximation.

The theorem applies to many examples, but its conditions may be too stringent. A modification that can be theoretically carried through under minimal conditions is based on the one-step method. Suppose that we are given a sequence of initial estimators $\tilde{\theta}_n$ that is $\sqrt{n}$-consistent for θ. We can assume without loss of generality that the estimators are discretized on a grid of meshwidth $n^{-1/2}$, which simplifies the constructions and proof. Then the one-step estimator is defined as

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

The estimator $\hat{\theta}_n$ can be considered a one-step iteration of the Newton-Raphson algorithm for solving the equation $\sum \tilde{\ell}_{\theta,\hat{\eta}}(X_i) = 0$ with respect to θ, starting at the initial guess $\tilde{\theta}_n$. For the benefit of the simple proof, we have made the estimators $\hat{\eta}_{n,i}$ for η dependent on the index i. In fact, we shall use only two different values for $\hat{\eta}_{n,i}$, one for the first half of the sample and another for the second half. Given estimators $\hat{\eta}_n = \hat{\eta}_n(X_1, \ldots, X_n)$ define $\hat{\eta}_{n,i}$ by, with $m = \lfloor n/2 \rfloor$,

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

Thus, for X_i belonging to the first half of the sample, we use an estimator $\hat{\eta}_{n,i}$ based on the second half of the sample, and vice versa. This sample-splitting trick is convenient in the proof, because the estimator of η used in $\tilde{\ell}_{\theta,\eta}(X_i)$ is always independent of X_i, simultaneously for X_i running through each of the two halves of the sample.

The discretization of $\tilde{\theta}_n$ and the sample-splitting are mathematical devices that rarely are useful in practice. However, the conditions of the preceding theorem can now be relaxed to, for every deterministic sequence $\theta_n = \theta + O(n^{-1/2})$,

$$\sqrt{n}P_{\theta_n,\eta}\tilde{\ell}_{\theta_n,\hat{\eta}_n} \xrightarrow{P} 0, \qquad P_{\theta_n,\eta}\left\|\tilde{\ell}_{\theta_n,\hat{\eta}_n} - \tilde{\ell}_{\theta_n,\eta}\right\|^2 \xrightarrow{P} 0. \tag{25.55}$$

$$\int \left\|\tilde{\ell}_{\theta_n,\eta} dP_{\theta_n,\eta}^{1/2} - \tilde{\ell}_{\theta,\eta} dP_{\theta,\eta}^{1/2}\right\|^2 \to 0. \tag{25.56}$$

25.57 Theorem. *Suppose that the model $\{P_{\theta,\eta} : \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ, η), and let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular.*

Assume that (25.55) *and* (25.56) *hold. Then the sequence* $\hat{\theta}_n$ *is asymptotically efficient at* (θ, η).

Proof. Fix a deterministic sequence of vectors $\theta_n = \theta + O(n^{-1/2})$. By the sample-splitting, the first half of the sum $\sum \tilde{\ell}_{\theta_n, \hat{\eta}_{n,i}}(X_i)$ is a sum of conditionally independent terms, given the second half of the sample. Thus,

$$\mathrm{E}_{\theta_n,\eta}\Big(\sqrt{m}\mathbb{P}_m\big(\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - \tilde{\ell}_{\theta_n,\eta}\big) \mid X_{m+1}, \ldots, X_n\Big) = \sqrt{m}P_{\theta_n,\eta}\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}},$$

$$\mathrm{var}_{\theta_n,\eta}\Big(\sqrt{m}\mathbb{P}_m\big(\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - \tilde{\ell}_{\theta_n,\eta}\big) \mid X_{m+1}, \ldots, X_n\Big) \le P_{\theta_n,\eta}\big\|\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - \tilde{\ell}_{\theta_n,\eta}\big\|^2.$$

Both expressions converge to zero in probability by assumption (25.55). We conclude that the sum inside the conditional expectations converges conditionally, and hence also unconditionally, to zero in probability. By symmetry, the same is true for the second half of the sample, whence

$$\sqrt{n}\mathbb{P}_n\big(\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - \tilde{\ell}_{\theta_n,\eta}\big) \xrightarrow{\mathrm{P}} 0.$$

We have proved this for the probability under (θ_n, η), but by contiguity the convergence is also under (θ, η).

The second part of the proof is technical, and we only report the result. The condition of differentiabily of the model and (25.56) imply that

$$\sqrt{n}\mathbb{P}_n\big(\tilde{\ell}_{\theta_n,\eta} - \tilde{\ell}_{\theta,\eta}\big) + \sqrt{n}P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}^T_{\theta,\eta}(\theta_n - \theta) \xrightarrow{\mathrm{P}} 0$$

(see [139], p. 185). Under stronger regularity conditions, this can also be proved by a Taylor expansion of $\tilde{\ell}_{\theta,\eta}$ in θ.) By the definition of the efficient score function as an orthogonal projection, $P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}^T_{\theta,\eta} = \tilde{I}_{\theta,\eta}$. Combining the preceding displays, we find that

$$\sqrt{n}\mathbb{P}_n\big(\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - \tilde{\ell}_{\theta,\eta}\big) + \tilde{I}_{\theta,\eta}\sqrt{n}(\theta_n - \theta) \xrightarrow{\mathrm{P}} 0.$$

In view of the discretized nature of $\tilde{\theta}_n$, this remains true if the deterministic sequence θ_n is replaced by $\tilde{\theta}_n$; see the argument in the proof of Theorem 5.48.

Next we study the estimator for the information matrix. For any vector $h \in \mathbb{R}^k$, the triangle inequality yields

$$\Big|\sqrt{\mathbb{P}_m(h^T\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}})^2} - \sqrt{\mathbb{P}_m(h^T\tilde{\ell}_{\theta_n,\eta})^2}\Big|^2 \le \mathbb{P}_m\big(h^T\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}} - h^T\tilde{\ell}_{\theta_n,\eta}\big)^2.$$

By (25.55), the conditional expectation under (θ_n, η) of the right side given $X_{m+1}, \ldots, X_n$ converges in probability to zero. A similar statement is valid for the second half of the observations. Combining this with (25.56) and the law of large numbers, we see that

$$\mathbb{P}_n\tilde{\ell}_{\theta_n,\hat{\eta}_{n,i}}\tilde{\ell}^T_{\theta_n,\hat{\eta}_{n,i}} \xrightarrow{\mathrm{P}} \tilde{I}_{\theta,\eta}.$$

In view of the discretized nature of $\tilde{\theta}_n$, this remains true if the deterministic sequence θ_n is replaced by $\tilde{\theta}_n$.

The theorem follows combining the results of the last two paragraphs with the definition of $\hat{\theta}_n$. ■

A further refinement is not to restrict the estimator for the efficient score function to be a plug-in type estimator. Both theorems go through if $\tilde{\ell}_{\theta,\hat{\eta}}$ is replaced by a general estimator $\hat{\ell}_{n,\theta} = \hat{\ell}_{n,\theta}(\cdot \mid X_1, \ldots, X_n)$, provided that this satisfies the appropriately modified conditions of the theorems, and in the second theorem we use the sample-splitting scheme. In the generalization of Theorem 25.57, condition (25.55) must be replaced by

$$\sqrt{n} P_{\theta_n,\eta} \hat{\ell}_{n,\theta_n} \xrightarrow{\mathrm{P}} 0, \qquad P_{\theta_n,\eta} \left\| \hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\theta_n,\eta} \right\|^2 \xrightarrow{\mathrm{P}} 0. \tag{25.58}$$

The proofs are the same. This opens the door to more tricks and further relaxation of the regularity conditions. An intermediate theorem concerning one-step estimators, but without discretization or sample-splitting, can also be proved under the conditions of Theorem 25.54. This removes the conditions of existence and consistency of solutions to the efficient score equation.

The theorems reduce the problem of efficient estimation of θ to estimation of the efficient score function. The estimator of the efficient score function must satisfy a "no-bias" and a consistency conditions. The consistency is usually easy to arrange, but the no-bias condition, such as (25.52) or the first part of (25.58), is connected to the structure and the size of the model, as the bias of the efficient score equations must converge to zero at a rate faster than $1/\sqrt{n}$. Within the context of Theorem 25.54 condition (25.52) is necessary. If it fails, then the sequence $\hat{\theta}_n$ is not asymptotically efficient and may even converge at a slower rate than $\sqrt{n}$. This follows by inspection of the proof, which reveals the following adaptation of the theorem. We assume that $\tilde{\ell}_{\theta,\eta}$ is the efficient score function for the true parameter (θ, η) but allow it to be arbitrary (mean-zero) for other parameters.

25.59 *Theorem.* *Suppose that the conditions of Theorem* 25.54 *hold except possibly condition* (25.52). *Then*

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_{\theta,\eta}^{-1} \tilde{\ell}_{\theta,\eta}(X_i) + \sqrt{n} P_{\hat{\theta}_n,\eta} \tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} + o_P(1).$$

Because by Lemma 25.23 the sequence $\hat{\theta}_n$ can be asymptotically efficient (regular with $N(0, \tilde{I}_{\theta,\eta}^{-1})$-limit distribution) only if it is asymptotically equivalent to the sum on the right, condition (25.52) is seen to be necessary for efficiency.

The verification of the no-bias condition may be easy due to special properties of the model but may also require considerable effort. The derivative of $P_{\theta,\eta}\tilde{\ell}_{\theta,\hat{\eta}}$ with respect to θ ought to converge to $\partial/\partial\theta P_{\theta,\eta}\tilde{\ell}_{\theta,\eta} = 0$. Therefore, condition (25.52) can usually be simplified to

$$\sqrt{n} P_{\theta,\eta} \tilde{\ell}_{\theta,\hat{\eta}_n} \xrightarrow{\mathrm{P}} 0.$$

The dependence on $\hat{\eta}$ is more interesting and complicated. The verification may boil down to a type of Taylor expansion of $P_{\theta,\eta}\tilde{\ell}_{\theta,\hat{\eta}}$ in $\hat{\eta}$ combined with establishing a rate of convergence for $\hat{\eta}$. Because η is infinite-dimensional, a Taylor series may be nontrivial. If $\hat{\eta} - \eta$ can

occur as a direction of approach to η that leads to a score function $B_{\theta,\eta}(\hat{\eta}-\eta)$, then we can write

$$P_{\theta,\eta}\tilde{\ell}_{\theta,\hat{\eta}} = (P_{\theta,\eta}-P_{\theta,\hat{\eta}})(\tilde{\ell}_{\theta,\hat{\eta}}-\tilde{\ell}_{\theta,\eta}) - P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\left[\frac{p_{\theta,\hat{\eta}}-p_{\theta,\eta}}{p_{\theta,\eta}} - B_{\theta,\eta}(\hat{\eta}-\eta)\right]. \tag{25.60}$$

We have used the fact that $P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}B_{\theta,\eta}h=0$ for every h, by the orthogonality property of the efficient score function. (The use of $B_{\theta,\eta}(\hat{\eta}-\eta)$ corresponds to a score operator that yields scores $B_{\theta,\eta}h$ from paths of the form $\eta_t=\eta+th$. If we use paths $d\eta_t=(1+th)\,d\eta$, then $B_{\theta,\eta}(d\hat{\eta}/d\eta-1)$ is appropriate.) The display suggests that the no-bias condition (25.52) is certainly satisfied if $\|\hat{\eta}-\eta\|=O_P(n^{-1/2})$, for $\|\cdot\|$ a norm relative to which the two terms on the right are both of the order $o_P\big(\|\hat{\eta}-\eta\|\big)$. In cases in which the nuisance parameter is not estimable at $\sqrt{n}$-rate the Taylor expansion must be carried into its second-order term. If the two terms on the right are both $O_P\big(\|\hat{\eta}-\eta\|^2\big)$, then it is still sufficient to have $\|\hat{\eta}-\eta\|=o_P(n^{-1/4})$. This observation is based on a crude bound on the bias, an integral in which cancellation could occur, by norms and can therefore be too pessimistic (See [35] for an example.) Special properties of the model may also allow one to take the Taylor expansion even further, with the lower order derivatives vanishing, and then a slower rate of convergence of the nuisance parameter may be sufficient, but no examples of this appear to be known. However, the extreme case that the expression in (25.52) is identically zero occurs in the important class of models that are convex-linear in the parameter.

25.61 ***Example (Convex-linear models).*** Suppose that for every fixed θ the model $\{P_{\theta,\eta}:\eta\in H\}$ is convex-linear: H is a convex subset of a linear space, and the dependence $\eta\mapsto P_{\theta,\eta}$ is linear. Then for every pair (η_1,η) and number $0\le t\le 1$, the convex combination $\eta_t=t\eta_1+(1-t)\eta$ is a parameter and the distribution $tP_{\theta,\eta_1}+(1-t)P_{\theta,\eta}=P_{\theta,\eta_t}$ belongs to the model. The score function at $t=0$ of the submodel $t\mapsto P_{\theta,\eta_t}$ is

$$\frac{\partial}{\partial t}_{|t=0}\log dP_{\theta,t\eta_1+(1-t)\eta}=\frac{dP_{\theta,\eta_1}}{dP_{\theta,\eta}}-1.$$

Because the efficient score function for θ is orthogonal to the tangent set for the nuisance parameter, it should satisfy

$$0=P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\left(\frac{dP_{\theta,\eta_1}}{dP_{\theta,\eta}}-1\right)=P_{\theta,\eta_1}\tilde{\ell}_{\theta,\eta}.$$

This means that the unbiasedness conditions in (25.52) and (25.55) are trivially satisfied, with the expectations $P_{\theta,\eta}\tilde{\ell}_{\theta,\hat{\eta}}$ even equal to 0.

A particular case in which this convex structure arises is the case of estimating a linear functional in an information-loss model. Suppose we observe $X=m(Y)$ for a known function m and an unobservable variable Y that has an unknown distribution η on a measurable space $(\mathcal{Y},\mathcal{A})$. The distribution $P_\eta=\eta\circ m^{-1}$ of X depends linearly on η. Furthermore, if we are interested in a linear function $\theta=\chi(\eta)$, then the nuisance-parameter space $H_\theta=\{\eta:\chi(\eta)=\theta\}$ is a convex subset of the set of probability measures on $(\mathcal{Y},\mathcal{A})$. □

25.8.1 Symmetric Location

Suppose that we observe a random sample from a density $\eta(x-\theta)$ that is symmetric about θ. In Example 25.27 it was seen that the efficient score function for θ is the ordinary score function,

$$\tilde{\ell}_{\theta,\eta}(x) = -\frac{\eta'}{\eta}(x-\theta).$$

We can apply Theorem 25.57 to construct an asymptotically efficient estimator sequence for θ under the minimal condition that the density η has finite Fisher information for location.

First, as an initial estimator $\tilde{\theta}_n$, we may use a discretized Z-estimator, solving $\mathbb{P}_n\psi(x-\theta) = 0$ for a well-behaved, symmetric function ψ. For instance, the score function of the logistic density. The $\sqrt{n}$-consistency can be established by Theorem 5.21.

Second, it suffices to construct estimators $\hat{\ell}_{n,\theta}$ that satisfy (25.58). By symmetry, the variables $T_i = |X_i - \theta|$ are, for a fixed θ, sampled from the density $g(s) = 2\eta(s)1\{s > 0\}$. We use these variables to construct an estimator $\hat{k}_n$ for the function g'/g, and next we set

$$\hat{\ell}_{n,\theta}(x; X_1, \ldots, X_n) = -\hat{k}_n\big(|x-\theta|; T_1, \ldots, T_n\big)\,\mathrm{sign}(x-\theta).$$

Because this function is skew-symmetric about the point θ, the bias condition in (25.58) is satisfied, with a bias of zero. Because the efficient score function can be written in the form

$$\tilde{\ell}_{\theta,\eta}(x) = -\frac{g'}{g}\big(|x-\theta|\big)\,\mathrm{sign}(x-\theta),$$

the consistency condition in (25.58) reduces to consistency of $\hat{k}_n$ for the function g'/g in that

$$\int \left(\hat{k}_n - \frac{g'}{g}\right)^2(s)\, g(s)\, ds \xrightarrow{\mathrm{P}} 0. \tag{25.62}$$

Estimators $\hat{k}_n$ can be constructed by several methods, a simple one being the kernel method of density estimation. For a fixed twice continuously differentiable probability density ω with compact support, a bandwidth parameter σ_n, and further positive tuning parameters α_n, β_n, and γ_n, set

$$\begin{aligned}
\hat{g}_n(s) &= \frac{1}{\sigma_n}\sum_{i=1}^n \omega\left(\frac{s-T_i}{\sigma_n}\right),\\
\hat{k}_n(s) &= \frac{\hat{g}_n'}{\hat{g}_n}(s)1_{\hat{B}_n}(s),\\
\hat{B}_n &= \big\{s : |\hat{g}_n'(s)| \le \alpha_n, \hat{g}_n(s) \ge \beta_n, s \ge \gamma_n\big\}.
\end{aligned} \tag{25.63}$$

Then (25.58) is satisfied provided $\alpha_n \uparrow \infty$, $\beta_n \downarrow 0$, $\gamma_n \downarrow 0$, and $\sigma_n \downarrow 0$ at appropriate speeds. The proof is technical and is given in the next lemma.

This particular construction shows that efficient estimators for θ exist under minimal conditions. It is not necessarily recommended for use in practice. However, any good initial estimator $\tilde{\theta}_n$ and any method of density or curve estimation may be substituted and will lead to a reasonable estimator for θ, which is theoretically efficient under some regularity conditions.

25.64 ***Lemma.*** *Let $T_1, \ldots, T_n$ be a random sample from a density g that is supported and absolutely continuous on $[0, \infty)$ and satisfies $\int (g'/\sqrt{g})^2(s)\, ds < \infty$. Then $\hat{k}_n$ given by (25.63) for a probability density ω that is twice continuously differentiable and supported on $[-1, 1]$ satisfies (25.62), if $\alpha_n \uparrow \infty$, $\gamma_n \downarrow 0$, $\beta_n \downarrow 0$, and $\sigma_n \downarrow 0$ in such a way that $\sigma_n \le \gamma_n$, $\alpha_n^2 \sigma_n / \beta_n^2 \to 0$, $n\sigma_n^4 \beta_n^2 \to \infty$.*

Proof. Start by noting that $\|g\|_\infty \le \int |g'(s)|\, ds \le \sqrt{I_g}$, by the Cauchy-Schwarz inequality. The expectations and variances of $\hat{g}_n$ and its derivative are given by

$$\begin{aligned} g_n(s) := \mathrm{E}\hat{g}_n(s) &= \mathrm{E}\frac{1}{\sigma}\omega\left(\frac{s - T_1}{\sigma}\right) = \int g(s - \sigma y)\, \omega(y)\, dy, \\ \mathrm{var}\hat{g}_n(s) &= \frac{1}{n\sigma^2}\mathrm{var}\omega\left(\frac{s - T_1}{\sigma}\right) \le \frac{1}{n\sigma^2}\|\omega\|_\infty^2, \\ \mathrm{E}\hat{g}_n'(s) &= g_n'(s) = \int g'(s - \sigma y)\omega(y)\, dy, \qquad (s \ge \gamma), \\ \mathrm{var}\hat{g}_n'(s) &\le \frac{1}{n\sigma^4}\|\omega'\|_\infty^2. \end{aligned}$$

By the dominated-convergence theorem, $g_n(s) \to g(s)$, for every $s > 0$. Combining this with the preceding display, we conclude that $\hat{g}_n(s) \xrightarrow{\mathrm{P}} g(s)$. If g' is sufficiently smooth, then the analogous statement is true for $\hat{g}_n'(s)$. Under only the condition of finite Fisher information for location, this may fail, but we still have that $\hat{g}_n'(s) - g_n'(s) \xrightarrow{\mathrm{P}} 0$ for every s; furthermore, $g_n' 1_{[\sigma,\infty)} \to g'$ in L_1, because

$$\int_\sigma^\infty |g_n' - g'|(s)\, ds \le \iint |g'(s - \sigma y) - g'(s)|\, ds\, \omega(y)\, dy \to 0,$$

by the L_1-continuity theorem on the inner integral, and next the dominated-convergence theorem on the outer integral.

The expectation of the integral in (25.62) restricted to the complement of the set $\hat{B}_n$ is equal to

$$\int \left(\frac{g'}{g}\right)^2 (s)\, g(s)\, \mathrm{P}\big(|\hat{g}_n'|(s) > \alpha \text{ or } \hat{g}_n(s) < \beta \text{ or } s < \gamma\big)\, ds.$$

This converges to zero by the dominated-convergence theorem. To see this, note first that $\mathrm{P}\big(\hat{g}_n(s) < \beta\big)$ converges to zero for all s such that $g(s) > 0$. Second, the probability $\mathrm{P}\big(|\hat{g}_n'|(s) > \alpha\big)$ is bounded above by $1\big\{|g_n'|(s) > \alpha/2\big\} + o(1)$, and the Lebesgue measure of the set $\big\{s : |g_n'|(s) > \alpha/2\big\}$ converges to zero, because $g_n' \to g'$ in L_1.

On the set $\hat{B}_n$ the integrand in (25.62) is the square of the function $(\hat{g}_n'/\hat{g}_n - g'/g)g^{1/2}$. This function can be decomposed as

$$\frac{\hat{g}_n'}{\hat{g}_n}\big(g^{1/2} - g_n^{1/2}\big) + \frac{(\hat{g}_n' - g_n')g_n^{1/2}}{\hat{g}_n} - \frac{g_n'(\hat{g}_n - g_n)}{g_n^{1/2}\hat{g}_n} + \left(\frac{g_n'}{g_n^{1/2}} - \frac{g'}{g^{1/2}}\right).$$

On $\hat{B}_n$ the sum of the squares of the four terms on the right is bounded above by

$$\frac{\alpha^2}{\beta^2}|g_n - g| + \frac{1}{\beta^2}(\hat{g}_n' - g_n')^2 g_n + \frac{1}{\beta^2}\left(\frac{g_n'}{g_n^{1/2}}\right)^2 (\hat{g}_n - g_n)^2 + \left(\frac{g_n'}{g_n^{1/2}} - \frac{g'}{g^{1/2}}\right)^2.$$

The expectations of the integrals over $\hat{B}_n$ of these four terms converge to zero. First, the integral over the first term is bounded above by

$$\frac{\alpha^2}{\beta^2}\int\int_{s>\gamma}\bigl|g(s-\sigma t)-g(s)\bigr|\,\omega(t)\,dt\,ds \le \frac{\alpha^2\sigma}{\beta^2}\int\bigl|g'(t)\bigr|\,dt\int|t|\omega(t)\,dt.$$

Next, the sum of the second and third terms gives the contribution

$$\frac{1}{n\sigma^4\beta^2}\|\omega'\|_\infty^2\int g_n(s)\,ds+\frac{1}{n\sigma^2\beta^2}\|\omega\|_\infty^2\int\left(\frac{g_n'}{g_n^{1/2}}\right)^2 ds.$$

The first term in this last display converges to zero, and the second as well, provided the integral remains finite. The latter is certainly the case if the fourth term converges to zero. By the Cauchy-Schwarz inequality,

$$\frac{\bigl(\int g'(s-\sigma y)\,\omega(y)\,dy\bigr)^2}{\int g(s-\sigma y)\,\omega(y)\,dy}\le\int\left(\frac{g'}{g^{1/2}}\right)^2(s-\sigma y)\,\omega(y)\,dy.$$

Using Fubini's theorem, we see that, for any set B, and B^σ its σ-enlargement,

$$\int_B\left(\frac{g_n'}{g_n^{1/2}}\right)^2(s)\,ds\le\int_{B^\sigma}\left(\frac{g'}{g^{1/2}}\right)^2 ds.$$

In particular, we have this for $B=B^\sigma=\mathbb{R}$, and $B=\bigl\{s:g(s)=0\bigr\}$. For the second choice of B, the sets B^σ decrease to B, by the continuity of g. On the complement of B, $g_n'/g_n^{1/2}\to g'/g^{1/2}$ in Lebesgue measure. Thus, by Proposition 2.29, the integral of the fourth term converges to zero. ■

25.8.2 *Errors-in-Variables*

Let the observations be a random sample of pairs (X_i, Y_i) with the same distribution as

$$\begin{aligned}X&=Z+e\\Y&=\alpha+\beta Z+f,\end{aligned}$$

for a bivariate normal vector (e, f) with mean zero and covariance matrix Σ and a random variable Z with distribution η, independent of (e, f). Thus Y is a linear regression on a variable Z which is observed with error. The parameter of interest is $\theta=(\alpha,\beta,\Sigma)$ and the nuisance parameter is η. To make the parameters identifiable one can put restrictions on either Σ or η. It suffices that η is not normal (if a degenerate distribution is considered normal with variance zero); alternatively it can be assumed that Σ is known up to a scalar.

Given (θ,Σ) the statistic $\psi_\theta(X,Y)=(1,\beta)\Sigma^{-1}(X,Y-\alpha)^T$ is sufficient (and complete) for η. This suggests to define estimators for (α,β,Σ) as the solution of the "conditional score equation" $\mathbb{P}_n\tilde{\ell}_{\theta,\hat{\eta}}=0$, for

$$\tilde{\ell}_{\theta,\eta}(X,Y)=\dot{\ell}_{\theta,\eta}(X,Y)-\mathrm{E}_\theta\bigl(\dot{\ell}_{\theta,\eta}(X,Y)\,|\,\psi_\theta(X,Y)\bigr).$$

This estimating equation has the attractive property of being unbiased in the nuisance parameter, in that

$$P_{\theta,\eta}\tilde{\ell}_{\theta,\eta'}=0,\qquad\text{every }\theta,\eta,\eta'.$$

Therefore, the no-bias condition is trivially satisfied, and the estimator $\hat{\eta}$ need only be consistent for η (in the sense of (25.53)). One possibility for $\hat{\eta}$ is the maximum likelihood estimator, which can be shown to be consistent by Wald's theorem, under some regularity conditions.

As the notation suggests, the function $\tilde{\ell}_{\theta,\eta}$ is equal to the efficient score function for θ. We can prove this by showing that the closed linear span of the set of nuisance scores contains all measurable, square-integrable functions of $\psi_\theta(x, y)$, because then projecting on the nuisance scores is identical to taking the conditional expectation.

As explained in Example 25.61, the functions $p_{\theta,\eta_1}/p_{\theta,\eta} - 1$ are score functions for the nuisance parameter (at (θ, η)). As is clear from the factorization theorem or direct calculation, they are functions of the sufficient statistic $\psi_\theta(X, Y)$. If some function $b\big(\psi_\theta(x, y)\big)$ is orthogonal to all scores of this type and has mean zero, then

$$\mathrm{E}_{\theta,\eta_1} b\big(\psi_\theta(X, Y)\big) = \mathrm{E}_{\theta,\eta} b\big(\psi_\theta(X, Y)\big)\left(\frac{p_{\theta,\eta_1}}{p_{\theta,\eta}} - 1\right) = 0.$$

Consequently, $b = 0$ almost surely by the completeness of $\psi_\theta(X, Y)$.

The regularity conditions of Theorem 25.54 can be shown to be satisfied under the condition that $\int |z|^9\, d\eta(z) < \infty$. Because all coordinates of the conditional score function can be written in the form $Q_\theta(x, y) + P_\theta(x, y)\mathrm{E}_\eta\big(Z \mid \psi_\theta(X, Y)\big)$ for polynomials Q_θ and P_θ of orders 2 and 1, respectively, the following lemma is the main part of the verification.†

25.65 ***Lemma.*** *For every $0 < \alpha \leq 1$ and every probability distribution η_0 on $\mathbb{R}$ and compact $K \subset (0, \infty)$, there exists an open neighborhood U of η_0 in the weak topology such that the class $\mathcal{F}$ of all functions*

$$(x, y) \mapsto (a_0 + a_1 x + a_2 y)\frac{\int z\, e^{z(b_0+b_1x+b_2y)}\, e^{-cz^2}\, d\eta(z)}{\int e^{z(b_0+b_1x+b_2y)}\, e^{-cz^2}\, d\eta(z)},$$

with η ranging over U, c ranging over K, and a and b ranging over compacta in $\mathbb{R}^3$, satisfies

$$\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \leq C\left(\frac{1}{\varepsilon}\right)^V \Big(P\big(1 + |x| + |y|\big)^{5+2\alpha+4/V+\delta}\Big)^{V/2},$$

for every $V \geq 1/\alpha$, every measure P on $\mathbb{R}^2$ and $\delta > 0$, and a constant C depending only on α, η_0, U, V, the compacta, and δ.

25.9 General Estimating Equations

Taking the efficient score equation as the basis for estimating a parameter is motivated by our wish to construct asymptotically efficient estimators. Perhaps, in certain situations, this is too much to ask, and it is better to aim at estimators that come close to attaining efficiency or are efficient only at the elements of a certain "ideal submodel." The pay off could be a gain in robustness, finite-sample properties, or computational simplicity. The information bounds then have the purpose of quantifying how much efficiency has possibly been lost.

† See [108] for a proof.

We retain the requirement that the estimator is $\sqrt{n}$-consistent and regular at every distribution P in the model. A somewhat stronger but still reasonable requirement is that it be *asymptotically linear* in that

$$\sqrt{n}\big(T_n - \psi(P)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{\psi}_P(X_i) + o_P(1).$$

This type of expansion and regularity implies that $\dot{\psi}_P$ is an influence function of the parameter $\psi(P)$, and the difference $\dot{\psi}_P - \tilde{\psi}_P$ must be orthogonal to the tangent set $\dot{\mathcal{P}}_P$.

This suggests that we compute the set of all influence functions to obtain an indication of which estimators T_n might be possible. If there is a nice parametrization $\dot{\psi}_{\theta,\tau}$ of these sets of functions in terms of a parameter of interest θ and a nuisance parameter τ, then a possible estimation procedure is to solve θ from the estimating equation, for given τ,

$$\sum_{i=1}^{n}\dot{\psi}_{\theta,\tau}(X_i) = 0.$$

The choice of the parameter τ determines the efficiency of the estimator $\hat{\theta}$. Rather than fixing it at some value we also can make it data-dependent to obtain efficiency at every element of a given submodel, or perhaps even the whole model. The resulting estimator can be analyzed with the help of, for example, Theorem 5.31.

If the model is parametrized by a partitioned parameter (θ, η), then any influence function for θ must be orthogonal to the scores for the nuisance parameter η. The parameter τ might be indexing both the nuisance parameter η and "position" in the tangent set at a given (θ, η). Then the unknown η (or the aspect of it that plays a role in τ) must be replaced by an estimator. The same reasoning as for the "no-bias" condition discussed in (25.60) allows us to hope that the resulting estimator for θ behaves as if the true η had been used.

25.66 ***Example (Regression).*** In the regression model considered in Example 25.28, the set of nuisance scores is the orthocomplement of the set $e\mathcal{H}$ of all functions of the form $(x, y) \mapsto \big(y - g_\theta(x)\big)h(x)$, up to centering at mean zero. The efficient score function for θ is equal to the projection of the score for θ onto the set $e\mathcal{H}$, and an arbitrary influence function is obtained, up to a constant, by adding any element from $e\mathcal{H}$ to this. The estimating equation

$$\sum_{i=1}^{n}\big(Y_i - g_\theta(X_i)\big)h(X_i) = 0$$

leads to an estimator with influence function in the direction of $\big(y - g_\theta(x)\big)h(x)$. Because the equation is unbiased for any h, we easily obtain $\sqrt{n}$-consistent estimators, even for data-dependent h. The estimator is more efficient if h is closer to the function $\dot{g}_\theta(x)/\mathrm{E}_\eta(e^2 \mid X = x)$, which gives the efficient influence function. For full efficiency it is necessary to estimate the function $x \mapsto \mathrm{E}_\eta(e^2 \mid X = x)$ nonparametrically, where consistency (for the right norm) suffices. □

25.67 ***Example (Missing at random).*** In Lemma 25.41 and Example 25.43 the influence functions in a MAR model are characterized as the sums of reweighted influence functions in the original model and the influence functions obtained from the MAR specification. If

the function π is known, then this leads to estimating equations of the form

$$\sum_{i=1}^{n} \frac{\Delta_i}{\pi(Y_i)} \dot{\psi}_{\theta,\tau}(X_i) - \sum_{i=1}^{n} \frac{\Delta_i - \pi(Y_i)}{\pi(Y_i)} c(Y_i) = 0.$$

For instance, if the original model is the regression model in the preceding example, then $\dot{\psi}_{\theta,\tau}(y)$ is $\big(y - g_\theta(x)\big)h(x)$. The efficiency of the estimator is influenced by the choice of c (the optimal choice is given in Example 25.43) and the choice of $\dot{\psi}_{\theta,\tau}$. (The efficient influence function of the original model need not be efficient here.) If π is correctly specified, then the second part of the estimating equation is unbiased for any c, and the asymptotic variance when using a random c should be the same as when using the limiting value of c. □

25.10 Maximum Likelihood Estimators

Estimators for parameters in semiparametric models can be constructed by any method – for instance, M-estimation or Z-estimation. However, the most important method to obtain asymptotically efficient estimators may be the method of maximum likelihood, just as in the case of parametric models. In this section we discuss the definitions of likelihoods and give some examples in which maximum likelihood estimators can be analyzed by direct methods. In Sections 25.11 and 25.12 we discuss two general approaches for analyzing these estimators.

Because many semiparametric models are not dominated or are defined in terms of densities that maximize to infinity, the functions that are called the "likelihoods" of the models must be chosen with care. For some models a likelihood can be taken equal to a density with respect to a dominating measure, but for other models we use an "empirical likelihood." Mixtures of these situations occur as well, and sometimes it is fruitful to incorporate a "penalty" in the likelihood, yielding a "penalized likelihood estimator"; maximize the likelihood over a set of parameters that changes with n, yielding a "sieved likelihood estimator"; or group the data in some way before writing down a likelihood. To bring out this difference with the classical, parametric maximum likelihood estimators, our present estimators are sometimes referred to as "nonparametric maximum likelihood estimators" (NPMLE), although semiparametric rather than nonparametric seems more correct. Thus we do not give an abstract definition of "likelihood," but describe "likelihoods that work" for particular examples. We denote the likelihood for the parameter P given one observation x by $\mathrm{lik}(P)(x)$.

Given a measure P, write $P\{x\}$ for the measure of the one-point set $\{x\}$. The function $x \mapsto P\{x\}$ may be considered the density of P, or its absolutely continuous part, with respect to counting measure. The *empirical likelihood* of a sample $X_1, \ldots, X_n$ is the function,

$$P \mapsto \prod_{i=1}^{n} P\{X_i\}.$$

Given a model $\mathcal{P}$, a maximum likelihood estimator could be defined as the distribution $\hat{P}$ that maximizes the empirical likelihood over $\mathcal{P}$. Such an estimator may or may not exist.

25.68 ***Example (Empirical distribution).*** Let $\mathcal{P}$ be the set of all probability distributions on the measurable space $(\mathcal{X}, \mathcal{A})$ (in which one-point sets are measurable). Then, for n

fixed different values $x_1, \ldots, x_n$, the vector $\big(P\{x_1\}, \ldots, P\{x_n\}\big)$ ranges over all vectors $p \geq 0$ such that $\sum p_i \leq 1$ when P ranges over $\mathcal{P}$. To maximize $p \mapsto \prod_i p_i$, it is clearly best to choose p maximal: $\sum_i p_i = 1$. Then, by symmetry, the maximizer must be $p = (1/n, \ldots, 1/n)$. Thus, the empirical distribution $\mathbb{P}_n = n^{-1} \sum \delta_{X_i}$ maximizes the empirical likelihood over the nonparametric model, whence it is referred to as the nonparametric maximum likelihood estimator.

If there are ties in the observations, this argument must be adapted, but the result is the same.

The empirical likelihood is appropriate for the nonparametric model. For instance, in the case of a Euclidean space, even if the model is restricted to distributions with a continuous Lebesgue density p, we still cannot use the map $p \mapsto \prod_{i=1}^n p(X_i)$ as a likelihood. The supremum of this likelihood is infinite, for we could choose p to have an arbitrarily high, very thin peak at some observation. □

Given a partitioned parameter (θ, η), it is sometimes helpful to consider the *profile likelihood*. Given a likelihood $\text{lik}_n(\theta, \eta)(X_1, \ldots, X_n)$, the profile likelihood for θ is defined as the function

$$\theta \mapsto \sup_{\eta} \text{lik}_n(\theta, \eta)(X_1, \ldots, X_n).$$

The supremum is taken over all possible values of η. The point of maximum of the profile likelihood is exactly the first coordinate of the maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$. We are simply computing the maximum of the likelihood over (θ, η) in two steps.

It is rarely possible to compute a profile likelihood explicitly, but its numerical evaluation is often feasible. Then the profile likelihood may serve to reduce the dimension of the likelihood function. Profile likelihood functions are often used in the same way as (ordinary) likelihood functions of parametric models. Apart from taking their points of maximum as estimators $\hat{\theta}$, the second derivative at $\hat{\theta}$ is used as an estimate of minus the inverse of the asymptotic covariance matrix of $\hat{\theta}$. Recent research appears to validate this practice.

25.69 ***Example (Cox model).*** Suppose that we observe a random sample from the distribution of $X = (T, Z)$, where the conditional hazard function of the "survival time" T with covariate Z takes the form

$$\lambda_{T|Z}(t) = e^{\theta Z} \lambda(t).$$

The hazard function λ is completely unspecified. The density of the observation $X = (T, Z)$ is equal to

$$e^{\theta z} \lambda(t) e^{-e^{\theta z} \Lambda(t)},$$

where Λ is the primitive function of λ (with $\Lambda(0) = 0$). The usual estimator for (θ, Λ) based on a sample of size n from this model is the maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$, where the likelihood is defined as, with $\Lambda\{t\}$ the jump of Λ at t,

$$(\theta, \Lambda) \mapsto \prod_{i=1}^n e^{\theta z_i} \Lambda\{t_i\} e^{-e^{\theta z_i} \Lambda(t_i)}.$$

This is the product of the density at the observations, but with the hazard function $\lambda(t)$ replaced by the jumps $\Lambda\{t\}$ of the cumulative hazard function. (This likelihood is close

but not exactly equal to the empirical likelihood of the model.) The form of the likelihood forces the maximizer $\hat{\Lambda}$ to be a jump function with jumps at the observed "deaths" t_i, only and hence the likelihood can be reduced to a function of the unknowns $\Lambda\{t_1\}, \ldots, \Lambda\{t_n\}$. It appears to be impossible to derive the maximizers $(\hat{\theta}, \hat{\Lambda})$ in closed-form formulas, but we can make some headway in characterizing the maximum likelihood estimators by "profiling out" the nuisance parameter Λ. Elementary calculus shows that, for a fixed θ, the function

$$(\lambda_1, \ldots, \lambda_n) \mapsto \prod_{i=1}^{n} e^{\theta z_i} \lambda_i e^{-e^{\theta z_i} \sum_{j: t_j \le t_i} \lambda_j}$$

is maximal for

$$\frac{1}{\lambda_k} = \sum_{i: t_i \ge t_k} e^{\theta z_i}.$$

The profile likelihood for θ is the supremum of the likelihood over Λ for fixed θ. In view of the preceding display this is given by

$$\theta \mapsto \prod_{i=1}^{n} \frac{e^{\theta z_i}}{\sum_{j: t_j \ge t_i} e^{\theta z_j}} e^{-1}.$$

The latter expression is known as the *Cox partial likelihood*. The original motivation for this criterion function is that the terms in the product are the conditional probabilities that the ith subject dies at time i given that one of the subjects at risk dies at that time. The maximum likelihood estimator for Λ is the step function with jumps

$$\hat{\Lambda}\{t_k\} = \frac{1}{\sum_{i: t_i \ge t_k} e^{\hat{\theta} z_i}}.$$

The estimators $\hat{\theta}$ and $\hat{\Lambda}$ are asymptotically efficient, under some restrictions. (See section 25.12.1.) We note that we have ignored the fact that jumps of hazard functions are smaller than 1 and have maximized over all measures Λ. □

25.70 ***Example (Scale mixture).*** Suppose we observe a sample from the distribution of $X = \theta + Z\varepsilon$, where the unobservable variables Z and ε are independent with completely unknown distribution η and a known density ϕ, respectively. Thus, the observation has a mixture density $\int p_\theta(x \mid z)\, d\eta(z)$ for the kernel

$$p_\theta(x \mid z) = \frac{1}{z} \phi\left(\frac{x - \theta}{z}\right).$$

If ϕ is symmetric about zero, then the mixture density is symmetric about θ, and we can estimate θ asymptotically efficiently with a fully adaptive estimator, as discussed in Section 25.8.1. Alternatively, we can take the mixture form of the underlying distribution into account and use, for instance, the maximum likelihood estimator, which maximizes the likelihood

$$(\theta, \eta) \mapsto \prod_{i=1}^{n} \int p_\theta(X_i \mid z)\, d\eta(z).$$

Under some conditions this estimator is asymptotically efficient.

Because the efficient score function for θ equals the ordinary score function for θ, the maximum likelihood estimator satisfies the efficient score equation $\mathbb{P}_n \tilde{\ell}_{\theta,\eta} = 0$. By the convexity of the model in η, this equation is unbiased in η. Thus, the asymptotic efficiency of the maximum likelihood estimator $\hat{\theta}$ follows under the regularity conditions of Theorem 25.54. Consistency of the sequence of maximum likelihood estimators $(\hat{\theta}_n, \hat{\eta}_n)$ for the product of the Euclidean and the weak topology can be proved by the method of Wald. The verification that the functions $\tilde{\ell}_{\theta,\eta}$ form a Donsker class is nontrivial but is possible using the techniques of Chapter 19. □

25.71 ***Example (Penalized logistic regression).*** In this model we observe a random sample from the distribution of $X = (V, W, Y)$, for a 0-1 variable Y that follows the logistic regression model

$$\mathrm{P}_{\theta,\eta}(Y = 1 \mid V, W) = \Psi\big(\theta V + \eta(W)\big),$$

where $\Psi(u) = 1/(1 + e^{-u})$ is the logistic distribution function. Thus, the usual linear regression of (V, W) has been replaced by the partial linear regression $\theta V + \eta(W)$, in which η ranges over a large set of "smooth functions." For instance, η is restricted to the Sobolev class of functions on $[0, 1]$ whose $(k-1)$st derivative exists and is absolutely continuous with $J(\eta) < \infty$, where

$$J^2(\eta) = \int_0^1 \big(\eta^{(k)}(w)\big)^2 \, dw.$$

Here $k \geq 1$ is a fixed integer and $\eta^{(k)}$ is the kth derivative of η with respect to z.

The density of an observation is given by

$$p_{\theta,\eta}(x) = \Psi\big(\theta v + \eta(w)\big)^y \big(1 - \Psi(\theta v + \eta(w)\big)^{1-y} f_{V,W}(v, w).$$

We cannot use this directly for defining a likelihood. The resulting maximizer $\hat{\eta}$ would be such that $\hat{\eta}(w_i) = \infty$ for every w_i with $y_i = 1$ and $\hat{\eta}(w_i) = -\infty$ when $y_i = 0$, or at least we could construct a sequence of finite, smooth η_m approaching this extreme choice. The problem is that qualitative smoothness assumptions such as $J(\eta) < \infty$ do not restrict η on a finite set of points $w_1, \ldots, w_n$ in any way.

To remedy this situation we can restrict the maximization to a smaller set of η, which we allow to grow as $n \to \infty$; for instance, the set of all η such that $J(\eta) \leq M_n$ for $M_n \uparrow \infty$ at a slow rate, or a sequence of spline approximations.

An alternative is to use a penalized likelihood, of the form

$$(\theta, \eta) \mapsto \mathbb{P}_n \log p_{\theta,\eta} - \hat{\lambda}_n^2 J^2(\eta).$$

Here $\hat{\lambda}_n$ is a "smoothing parameter" that determines the importance of the penalty $J^2(\eta)$. A large value of $\hat{\lambda}_n$ leads to smooth maximizers $\hat{\eta}$, for small values the maximizer is more like the unrestricted maximum likelihood estimator. Intermediate values are best and are often chosen by a data-dependent scheme, such as cross-validation. The penalized estimator $\hat{\theta}$ can be shown to be asymptotically efficient if the smoothing parameter is constructed to satisfy $\hat{\lambda}_n = o_P(n^{-1/2})$ and $\hat{\lambda}_n^{-1} = O_P(n^{k/(2k+1)})$ (see [102]). □

25.72 ***Example (Proportional odds).*** Suppose that we observe a random sample from the distribution of the variable $X = \big(T \wedge C, 1\{T \leq C\}, Z\big)$, in which, given Z, the variables

T and C are independent, as in the random censoring model, but with the distribution function $F(t \mid z)$ of T given Z restricted by

$$\frac{F(t \mid z)}{1 - F(t \mid z)} = e^{z^T\theta}\eta(t).$$

In other words, the conditional *odds* given z of survival until t follows a Cox-type regression model. The unknown parameter η is a nondecreasing, cadlag function from $[0, \infty)$ into itself with $\eta(0) = 0$. It is the odds of survival if $\theta = 0$ and T is independent of Z.

If η is absolutely continuous, then the density of $X = (Y, \Delta, Z)$ is

$$\left(\frac{e^{-z^T\theta}\eta'(y)\left(1 - F_C(y - \mid z)\right)}{\left(\eta(y) + e^{-z^T\theta}\right)^2}\right)^{\delta}\left(\frac{e^{-z^T\theta} f_C(y \mid z)}{\eta(y) + e^{-z^T\theta}}\right)^{1-\delta} f_Z(z).$$

We cannot use this density as a likelihood, for the supremum is infinite unless we restrict η in an important way. Instead, we view η as the distribution function of a measure and use the empirical likelihood. The probability that $X = x$ is given by

$$\left(\frac{e^{-z^T\theta}\eta\{y\}\left(1 - F_C(y - \mid z)\right)}{\left(\eta(y) + e^{-z^T\theta}\right)\left(\eta(y-) + e^{-z^T\theta}\right)}\right)^{\delta}\left(\frac{e^{-z^T\theta} F_C(\{y\} \mid z)}{\eta(y) + e^{-z^T\theta}}\right)^{1-\delta} F_Z\{z\},$$

For likelihood inference concerning (θ, η) only, we may drop the terms involving F_C and F_Z and define the likelihood for one observation as

$$\operatorname{lik}(\theta, \eta)(x) = \left(\frac{e^{-z^T\theta}\eta\{y\}}{\left(\eta(y) + e^{-z^T\theta}\right)\left(\eta(y-) + e^{-z^T\theta}\right)}\right)^{\delta}\left(\frac{e^{-z^T\theta}}{\eta(y) + e^{-z^T\theta}}\right)^{1-\delta}.$$

The presence of the jumps $\eta\{y\}$ causes the maximum likelihood estimator $\hat{\eta}$ to be a step function with support points at the observed survival times (the values y_i corresponding to $\delta_i = 1$). First, it is clear that each of these points must receive a positive mass. Second, mass to the right of the largest y_i such that $\delta_i = 1$ can be deleted, meanwhile increasing the likelihood. Third, mass assigned to other points can be moved to the closest y_i to the right such that $\delta_i = 1$, again increasing the likelihood. If the biggest observation y_i has $\delta_i = 1$, then $\hat{\eta}\{y_i\} = \infty$ and that observation gives a contribution 1 to the likelihood, because the function $p \mapsto p/(p + r)$ attains for $p \geq 0$ its maximal value 1 at $p = \infty$. On the other hand, if $\delta_i = 0$ for the largest y_i, then all jumps of $\hat{\eta}$ must be finite.

The maximum likelihood estimators have been shown to be asymptotically efficient under some conditions in [105]. □

25.10.1 *Random Censoring*

Suppose that we observe a random sample $(X_1, \Delta_1), \ldots, (X_n, \Delta_n)$ from the distribution of $\left(T \wedge C, 1\{T \leq C\}\right)$, in which the "survival time" T and the "censoring time" C are independent with completely unknown distribution functions F and G, respectively. The distribution of a typical observation (X, Δ) satisfies

$$\mathrm{P}_{F,G}(X \leq x, \Delta = 0) = \int_{[0,x]} (1 - F)\, dG,$$

$$\mathrm{P}_{F,G}(X \leq x, \Delta = 1) = \int_{[0,x]} (1 - G_-)\, dF.$$

Consequently, if F and G have densities f and g (relative to some dominating measures), then (X, Δ) has density

$$(x, \delta) \mapsto \big((1-F)(x)g(x)\big)^{\delta}\big((1-G_-)(x)f(x)\big)^{1-\delta}.$$

For f and g interpreted as Lebesgue densities, we cannot use this expression as a factor in a likelihood, as the resulting criterion would have supremum infinity. (Simply choose f or g to have a very high, thin peak at an observation X_i with $\Delta_i = 1$ or $\Delta_i = 0$, respectively.) Instead, we may take f and g as densities relative to counting measure. This leads to the empirical likelihood

$$(F, G) \mapsto \prod_{i=1}^{n}\big((1-F)(X_i)G\{X_i\}\big)^{1-\Delta_i}\prod_{i=1}^{n}\big((1-G_-)(X_i)F\{X_i\}\big)^{\Delta_i}.$$

In view of the product form, this factorizes in likelihoods for F and G separately. The maximizer $\hat{F}$ of the likelihood $F \mapsto \prod_{i=1}^{n}(1-F)(X_i)^{1-\Delta_i}F\{X_i\}^{\Delta_i}$ turns out to be the *product limit estimator*, given in Example 20.15.

That the product limit estimator maximizes the likelihood can be seen by direct arguments, but a slight detour is more insightful. The next lemma shows that under the present model the distribution $P_{F,G}$ of (X, Δ) can be any distribution on the sample space $[0, \infty) \times \{0, 1\}$. In other words, if F and G range over all possible probability distributions on $[0, \infty]$, then $P_{F,G}$ ranges over all distributions on $[0, \infty) \times \{0, 1\}$. Moreover, the relationship $(F, G) \leftrightarrow P_{F,G}$ is one-to-one on the interval where $(1-F)(1-G) > 0$. As a consequence, there exists a pair $(\hat{F}, \hat{G})$ such that $P_{\hat{F},\hat{G}}$ is the empirical distribution $\mathbb{P}_n$ of the observations

$$P_{\hat{F},\hat{G}}\{X_i, \Delta_i\} = \mathbb{P}_n\{X_i, \Delta_i\}, \qquad 1 \le i \le n.$$

Because the empirical distribution maximizes $P \mapsto \prod_{i=1}^{n}P\{X_i, \Delta_i\}$ over all distributions, it follows that $(\hat{F}, \hat{G})$ maximizes $(F, G) \mapsto \prod_{i=1}^{n}P_{F,G}\{X_i, \Delta_i\}$ over all (F, G). That $\hat{F}$ is the product limit estimator next follows from Example 20.15.

To complete the discussion, we study the map $(F, G) \leftrightarrow P_{F,G}$. A probability distribution on $[0, \infty) \times \{0, 1\}$ can be identified with a pair (H_0, H_1) of subdistribution functions on $[0, \infty)$ such that $H_0(\infty) + H_1(\infty) = 1$, by letting $H_i(x)$ be the mass of the set $[0, x] \times \{i\}$. A given pair of distribution functions (F_0, F_1) on $[0, \infty)$ yields such a pair of subdistribution functions (H_0, H_1), by

$$H_0(x) = \int_{[0,x]}(1-F_1)\,dF_0, \qquad H_1(x) = \int_{[0,x]}(1-F_{0-})\,dF_1. \tag{25.73}$$

Conversely, the pair (F_0, F_1) can be recovered from a given pair (H_0, H_1) by, with ΔH_i the jump in H_i, $H = H_0 + H_1$ and Λ_i^c the continuous part of Λ_i,

$$\Lambda_0(x) = \int_{[0,x]}\frac{dH_0}{1-H_- - \Delta H_1}, \qquad \Lambda_1(x) = \int_{[0,x]}\frac{dH_1}{1-H_-},$$

$$1 - F_i(x) = \prod_{0 \le s \le x}\big(1-\Lambda_i\{s\}\big)e^{-\Lambda_i^c(x)}.$$

25.74 ***Lemma.*** *Given any pair (H_0, H_1) of subdistribution functions on $[0, \infty)$ such that $H_0(\infty) + H_1(\infty) = 1$, the preceding display defines a pair (F_0, F_1) of subdistribution functions on $[0, \infty)$ such that* (25.73) *holds.*

Proof. For any distribution function A and cumulative hazard function B on $[0, \infty)$, with B^c the continuous part of B,

$$1 - A(t) = \prod_{0 \le s \le t} \bigl(1 - B\{s\}\bigr) e^{-B^c(t)} \text{ iff } B(t) = \int_{[0,t]} \frac{dA}{1 - A_-}.$$

To see this, rewrite the second equality as $(1 - A_-)\, dB = dA$ and $B(0) = A(0)$, and integrate this to rewrite it again as the *Volterra equation*

$$(1 - A) = 1 + \int_{[0,\cdot]} (1 - A_-)\, d(-B).$$

It is well known that the Volterra equation has the first equation of the display as its unique solution.†

Combined with the definition of F_i, the equivalence in the preceding display implies immediately that $d\Lambda_i = dF_i/(1 - F_{i-})$. Secondly, as immediate consequences of the definitions,

$$(1 - F_0)(1 - F_1)(t) = \prod_{s \le t} (1 - \Delta\Lambda_0 - \Delta\Lambda_1 + \Delta\Lambda_0 \Delta\Lambda_1)(s) e^{-(\Lambda_0 + \Lambda_1)^c(t)},$$

$$(\Lambda_0 + \Lambda_1)(t) - \sum_{s \le t} \Delta\Lambda_0(s) \Delta\Lambda_1(s) = \int_{[0,t]} \frac{dH}{1 - H_-}.$$

(Split $dH_0/(1 - H_- - \Delta H_1)$ into the parts corresponding to dH_0^c and ΔH_0 and note that ΔH_1 may be dropped in the first part.) Combining these equations with the Volterra equation, we obtain that $1 - H = (1 - F_0)(1 - F_1)$. Taken together with $dH_1 = (1 - H_-)\, d\Lambda_1$, we conclude that $dH_1 = (1 - F_{0-})(1 - F_{1-})\, d\Lambda_1 = (1 - F_{0-})\, dF_1$, and similarly $dH_0 = (1 - F_1)\, dF_0$. ■

25.11 Approximately Least-Favorable Submodels

If the maximum likelihood estimator satisfies the efficient score equation $\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{\eta}} = 0$, then Theorem 25.54 yields its asymptotic normality, provided that its conditions can be verified for the maximum likelihood estimator $\hat{\eta}$. Somewhat unexpectedly, the efficient score function may not be a "proper" score function and the maximum likelihood estimator may not satisfy the efficient score equation. This is because, by definition, the efficient score function is a projection, and nothing guarantees that this projection is the derivative of the log likelihood along some submodel. If there exists a "least favorable" path $t \mapsto \eta_t(\hat{\theta}, \hat{\eta})$ such that $\eta_0(\hat{\theta}, \hat{\eta}) = \hat{\eta}$, and, for every x,

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

then the maximum likelihood estimator satisfies the efficient score equation; if not, then this is not clear. The existence of an exact least favorable submodel appears to be particularly uncertain at the maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$, as this tends to be on the "boundary" of the parameter set.

† See, for example, [133, p. 206] or [55] for an extended discussion.

A method around this difficulty is to replace the efficient score equation by an approximation. First, it suffices that $(\hat\theta, \hat\eta)$ satisfies the efficient score equation approximately, for Theorem 25.54 goes through provided $\sqrt{n}\,\mathbb{P}_n\tilde\ell_{\hat\theta,\hat\eta} = o_P(1)$. Second, it was noted following the proof of Theorem 25.54 that this theorem is valid for estimating equations of the form $\mathbb{P}_n\tilde\ell_{\theta,\hat\eta} = 0$ for arbitrary mean-zero functions $\tilde\ell_{\theta,\eta}$; its assertion remains correct provided that at the true value of (θ, η) the function $\tilde\ell_{\theta,\eta}$ is the efficient score function. This suggests to replace, in our proof, the function $\tilde\ell_{\theta,\eta}$ by functions $\tilde\kappa_{\theta,\eta}$ that are proper score functions and are close to the efficient score function, at least for the true value of the parameter. These are derived from "approximately-least favorable submodels."

We define such submodels as maps $t \mapsto \eta_t(\theta, \eta)$ from a neighborhood of $0 \in \mathbb{R}^k$ to the parameter set for η with $\eta_0(\theta, \eta) = \eta$ (for every (θ, η)) such that

$$\tilde\kappa_{\theta,\eta}(x) = \frac{\partial}{\partial t}_{|t=0} \log \operatorname{lik}\big(\theta + t, \eta_t(\theta, \eta)\big)(x),$$

exists (for every x) and is equal to the efficient score function at $(\theta, \eta) = (\theta_0, \eta_0)$. Thus, the path $t \mapsto \eta_t(\theta, \eta)$ must pass through η at $t = 0$, and at the true parameter (θ_0, η_0) the submodel is truly least favorable in that its score is the efficient score for θ. We need such a submodel for every fixed (θ, η), or at least for the true value (θ_0, η_0) and every possible value of $(\hat\theta, \hat\eta)$.

If $(\hat\theta, \hat\eta)$ maximizes the likelihood, then the function $t \mapsto \mathbb{P}_n \log \operatorname{lik}\big(\theta + t, \eta_t(\hat\theta, \hat\eta)\big)$ is maximal at $t = 0$ and hence $(\hat\theta, \hat\eta)$ satisfies the stationary equation $\mathbb{P}_n\tilde\kappa_{\hat\theta,\hat\eta} = 0$. Now Theorem 25.54, with $\tilde\ell_{\theta,\eta}$ replaced by $\tilde\kappa_{\theta,\eta}$, yields the asymptotic efficiency of $\hat\theta_n$. For easy reference we reformulate the theorem.

$$P_{\hat\theta_n,\eta_0}\tilde\kappa_{\hat\theta_n,\hat\eta_n} = o_P\big(n^{-1/2} + \|\hat\theta_n - \theta_0\|\big) \tag{25.75}$$

$$P_{\theta_0,\eta_0}\big\|\tilde\kappa_{\hat\theta_n,\hat\eta_n} - \tilde\kappa_{\theta_0,\eta_0}\big\|^2 \xrightarrow{\mathrm{P}} 0, \qquad P_{\hat\theta_n,\eta_0}\big\|\tilde\kappa_{\hat\theta_n,\hat\eta_n}\big\|^2 = O_P(1). \tag{25.76}$$

25.77 *Theorem.* *Suppose that the model $\{P_{\theta,\eta} : \theta \in \Theta\}$, is differentiable in quadratic mean with respect to θ at (θ_0, η_0) and let the efficient information matrix $\tilde I_{\theta_0,\eta_0}$ be nonsingular. Assume that $\tilde\kappa_{\theta,\eta}$ are the score functions of approximately least-favorable submodels (at (θ_0, η_0)), that the functions $\tilde\kappa_{\hat\theta,\hat\eta}$ belong to a P_{θ_0,η_0}-Donsker class with square-integrable envelope with probability tending to 1, and that (25.75) and (25.76) hold. Then the maximum likelihood estimator $\hat\theta_n$ is asymptotically efficient at (θ_0, η_0) provided that it is consistent.*

The no-bias condition (25.75) can be analyzed as in (25.60), with $\tilde\ell_{\theta,\hat\eta}$ replaced by $\tilde\kappa_{\theta,\hat\eta}$. Alternatively, it may be useful to avoid evaluating the efficient score function at $\hat\theta$ or $\hat\eta$, and (25.60) may be adapted to

$$\begin{aligned} P_{\hat\theta,\eta_0}\tilde\kappa_{\hat\theta,\hat\eta} &= (P_{\hat\theta,\eta_0} - P_{\hat\theta,\hat\eta})(\tilde\kappa_{\hat\theta,\hat\eta} - \tilde\kappa_{\theta_0,\eta_0}) \\ &\quad - \int \tilde\kappa_{\theta_0,\eta_0}\big[p_{\hat\theta,\hat\eta} - p_{\hat\theta,\eta_0} - B_{\theta_0,\eta_0}(\hat\eta - \eta_0)\, p_{\theta_0,\eta_0}\big]\,d\mu. \end{aligned} \tag{25.78}$$

Replacing $\hat\theta$ by θ_0 should make at most a difference of $o_P\big(\|\hat\theta - \theta_0\|\big)$, which is negligible in the preceding display, but the presence of $\hat\eta$ may require a rate of convergence for $\hat\eta$. Theorem 5.55 yields such rates in some generality and can be translated to the present setting as follows.

Consider estimators $\hat{\tau}_n$ contained in a set H_n that, for a given $\hat{\lambda}_n$ contained in a set $\Lambda_n \subset \mathbb{R}$, maximize a criterion $\tau \mapsto \mathbb{P}_n m_{\tau,\hat{\lambda}_n}$, or at least satisfy $\mathbb{P}_n m_{\tau,\hat{\lambda}_n} \geq \mathbb{P}_n m_{\tau_0,\hat{\lambda}_n}$. Assume that for every $\lambda \in \Lambda_n$, every $\tau \in H_n$ and every $\delta > 0$,

$$P(m_{\tau,\lambda} - m_{\tau_0,\lambda}) \lesssim -d_\lambda^2(\tau, \tau_0) + \lambda^2, \tag{25.79}$$

$$\mathrm{E}^* \sup_{\substack{d_\lambda(\tau,\tau_0)<\delta \\ \lambda\in\Lambda_n,\tau\in H_n}} \left|\mathbb{G}_n(m_{\tau,\lambda} - m_{\tau_0,\lambda})\right| \lesssim \phi_n(\delta). \tag{25.80}$$

25.81 ***Theorem.*** *Suppose that* (25.79) *and* (25.80) *are valid for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is decreasing for some $\alpha < 2$ and sets $\Lambda_n \times H_n$ such that $\mathrm{P}(\hat{\lambda}_n \in \Lambda_n, \hat{\tau}_n \in H_n) \to 1$. Then $d_{\hat{\lambda}}(\hat{\tau}_n, \tau_0) \leq O_P^*(\delta_n + \hat{\lambda}_n)$ for any sequence of positive numbers δ_n such that $\phi_n(\delta_n) \leq \sqrt{n}\,\delta_n^2$ for every n.*

25.11.1 *Cox Regression with Current Status Data*

Suppose that we observe a random sample from the distribution of $X = (C, \Delta, Z)$, in which $\Delta = 1\{T \leq C\}$, that the "survival time" T and the observation time C are independent given Z, and that T follows a Cox model. The density of X relative to the product of $F_{C,Z}$ and counting measure on $\{0, 1\}$ is given by

$$p_{\theta,\Lambda}(x) = \left(1 - \exp\left(-e^{\theta^T z}\Lambda(c)\right)\right)^\delta \left(\exp\left(-e^{\theta^T z}\Lambda(c)\right)\right)^{1-\delta}.$$

We define this as the likelihood for one observation x. In maximizing the likelihood we restrict the parameter θ to a compact in $\mathbb{R}^k$ and restrict the parameter Λ to the set of all cumulative hazard functions with $\Lambda(\tau) \leq M$ for a fixed large constant M and τ the end of the study.

We make the following assumptions. The observation times C possess a Lebesgue density that is continuous and positive on an interval $[\sigma, \tau]$ and vanishes outside this interval. The true parameter Λ_0 is continuously differentiable on this interval, satisfies $0 < \Lambda_0(\sigma-) \leq \Lambda_0(\tau) < M$, and is continuously differentiable on $[\sigma, \tau]$. The covariate vector Z is bounded and $\mathrm{E}\,\mathrm{cov}(Z \mid C) > 0$. The function h_{θ_0,Λ_0} given by (25.82) has a version that is differentiable with a bounded derivative on $[\sigma, \tau]$. The true parameter θ_0 is an inner point of the parameter set for θ.

The score function for θ takes the form

$$\dot{\ell}_{\theta,\Lambda}(x) = z\Lambda(c)Q_{\theta,\Lambda}(x),$$

for the function $Q_{\theta,\Lambda}$ given by

$$Q_{\theta,\Lambda}(x) = e^{\theta^T z}\left[\delta\frac{e^{-e^{\theta^T z}\Lambda(c)}}{1 - e^{-e^{\theta^T z}\Lambda(c)}} - (1 - \delta)\right].$$

For every nondecreasing, nonnegative function h and positive number t, the submodel $\Lambda_t = \Lambda + th$ is well defined. Inserting this in the log likelihood and differentiating with respect to t at $t = 0$, we obtain a score function for Λ of the form

$$B_{\theta,\Lambda}h(x) = h(c)Q_{\theta,\Lambda}(x).$$

The linear span of these score functions contains $B_{\theta,\Lambda}h$ for all bounded functions h of bounded variation. In view of the similar structure of the scores for θ and Λ, projecting $\dot{\ell}_{\theta,\Lambda}$ onto the closed linear span of the nuisance scores is a weighted least-squares problem with weight function $Q_{\theta,\Lambda}$. The solution is given by the vector-valued function

$$h_{\theta,\Lambda}(c) = \Lambda(c)\frac{\mathrm{E}_{\theta,\Lambda}\big(ZQ^2_{\theta,\Lambda}(X)\,|\,C=c\big)}{\mathrm{E}_{\theta,\Lambda}\big(Q^2_{\theta,\Lambda}(X)\,|\,C=c\big)}. \tag{25.82}$$

The efficient score function for θ takes the form

$$\tilde{\ell}_{\theta,\Lambda}(x) = \big(z\Lambda(c) - h_{\theta,\Lambda}(c)\big)Q_{\theta,\Lambda}(x).$$

Formally, this function is the derivative at $t = 0$ of the log likelihood evaluated at $(\theta + t, \Lambda - t^T h_{\theta,\Lambda})$. However, the second coordinate of the latter path may not define a nondecreasing, nonnegative function for every t in a neighborhood of 0 and hence cannot be used to obtain a stationary equation for the maximum likelihood estimator. This is true in particular for discrete cumulative hazard functions Λ, for which $\Lambda + th$ is nondecreasing for both $t < 0$ and $t > 0$ only if h is constant between the jumps of Λ.

This suggests that the maximum likelihood estimator does not satisfy the efficient score equation. To prove the asymptotic normality of $\hat{\theta}$, we replace this equation by an approximation, obtained from an approximately least favorable submodel.

For fixed (θ, Λ), and a fixed bounded, Lipschitz function ϕ, define

$$\Lambda_t(\theta, \Lambda) = \Lambda - t^T\phi(\Lambda)\big(h_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1}\big)(\Lambda).$$

Then $\Lambda_t(\theta, \Lambda)$ is a cumulative hazard function for every t that is sufficiently close to zero, because for every $u \le v$,

$$\Lambda_t(\theta, \Lambda)(v) - \Lambda_t(\theta, \Lambda)(u) \ge \big(\Lambda(v) - \Lambda(u)\big)\Big(1 - \|t\|\,\big\|\phi\, h_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1}\big\|_{\mathrm{Lip}}\Big).$$

Inserting $\big(\theta + t, \Lambda_t(\theta, \Lambda)\big)$ into the log likelihood, and differentiating with respect to t at $t = 0$, yields the score function

$$\tilde{\kappa}_{\theta,\Lambda}(x) = \Big(z\Lambda(c) - \phi\big(\Lambda(c)\big)\big(h_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1}\big)\big(\Lambda(c)\big)\Big)Q_{\theta,\Lambda}(x).$$

If evaluated at (θ_0, Λ_0) this reduces to the efficient score function $\tilde{\ell}_{\theta_0,\Lambda_0}(x)$ provided $\phi(\Lambda_0) = 1$, whence the submodel is approximately least favorable. To prove the asymptotic efficiency of $\hat{\theta}_n$ it suffices to verify the conditions of Theorem 25.77.

The function ϕ is a technical device that has been introduced in order to ensure that $0 \le \Lambda_t(\theta, \Lambda) \le M$ for all t that are sufficiently close to 0. This is guaranteed if $0 \le y\,\phi(y) \le c\big(y \wedge (M - y)\big)$ for every $0 \le y \le M$, for a sufficiently large constant c. Because by assumption $\big[\Lambda_0(\sigma-), \Lambda_0(\tau)\big] \subset (0, M)$, there exists such a function ϕ that also fulfills $\phi(\Lambda_0) = 1$ on $[\sigma, \tau]$.

In order to verify the no-bias condition (25.52) we need a rate of convergence for $\hat{\Lambda}_n$.

25.83 ***Lemma.*** *Under the conditions listed previously, $\hat{\theta}_n$ is consistent and $\|\hat{\Lambda}_n - \Lambda_0\|_{P_0,2} = O_P(n^{-1/3})$.*

Proof. Denote the index (θ_0, Λ_0) by 0, and define functions

$$m_{\theta,\Lambda} = \log(p_{\theta,\Lambda} + p_0)/2.$$

The densities $p_{\theta,\Lambda}$ are bounded above by 1, and under our assumptions the density p_0 is bounded away from zero. It follows that the functions $m_{\theta,\Lambda}(x)$ are uniformly bounded in (θ, Λ) and x.

By the concavity of the logarithm and the definition of $(\hat\theta, \hat\Lambda)$,

$$\mathbb{P}_n m_{\hat\theta,\hat\Lambda} \geq \tfrac{1}{2}\,\mathbb{P}_n \log p_{\hat\theta,\hat\Lambda} + \tfrac{1}{2}\,\mathbb{P}_n \log p_0 \geq\ \mathbb{P}_n \log p_0 =\ \mathbb{P}_n m_0.$$

Therefore, Theorem 25.81 is applicable with $\tau = (\theta, \Lambda)$ and without λ. For technical reasons it is preferable first to establish the consistency of $(\hat\theta, \hat\Lambda)$ by a separate argument.

We apply Wald's proof, Theorem 5.14. The parameter set for θ is compact by assumption, and the parameter set for Λ is compact relative to the weak topology. Wald's theorem shows that the distance between $(\hat\theta, \hat\Lambda)$ and the set of maximizers of the Kullback-Leibler divergence converges to zero. This set of maximizers contains (θ_0, Λ_0), but this parameter is not fully identifiable under our assumptions: The parameter Λ_0 is identifiable only on the interval (σ, τ). It follows that $\hat\theta \xrightarrow{\mathrm{P}} \theta_0$ and $\hat\Lambda(t) \xrightarrow{\mathrm{P}} \Lambda_0(t)$ for every $\sigma < t < \tau$. (The convergence of $\hat\Lambda$ at the points σ and τ does not appear to be guaranteed.)

By the proof of Lemma 5.35 and Lemma 25.85 below, condition (25.79) is satisfied with $d\big((\theta, \Lambda), (\theta_0, \Lambda_0)\big)$ equal to $\|\theta - \theta_0\| + \|\Lambda - \Lambda_0\|_2$. By Lemma 25.84 below, the bracketing entropy of the class of functions $m_{\theta,\Lambda}$ is of the order $(1/\varepsilon)$. By Lemma 19.36 condition (25.80) is satisfied for

$$\phi_n(\delta) = \sqrt{\delta}\left(1 + \frac{\sqrt{\delta}}{\delta^2\sqrt{n}}\right).$$

This leads to a convergence rate of $n^{-1/3}$ for both $\|\hat\theta - \theta_0\|$ and $\|\hat\Lambda - \Lambda_0\|_2$. ■

To verify the no-bias condition (25.75), we use the decomposition (25.78). The integrands in the two terms on the right can both be seen to be bounded, up to a constant, by $(\hat\Lambda - \Lambda_0)^2$, with probability tending to one. Thus the bias $P_{\hat\theta,\eta_0}\tilde\kappa_{\hat\theta,\hat\eta}$ is actually of the order $O_P(n^{-2/3})$.

The functions $x \mapsto \tilde\kappa_{\theta,\Lambda}(x)$ can be written in the form $\psi\big(z, e^{\theta^T z}, \Lambda(c), \delta\big)$ for a function ψ that is Lipschitz in its first three coordinates, for $\delta \in \{0, 1\}$ fixed. (Note that $\Lambda \mapsto \Lambda Q_{\theta,\Lambda}$ is Lipschitz, as $\Lambda \mapsto h_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1}(\Lambda)/\Lambda = (h_{\theta_0,\Lambda_0}/\Lambda_0) \circ \Lambda_0^{-1}(\Lambda)$.) The functions $z \mapsto z$, $z \mapsto \exp\theta^T z$, $c \mapsto \Lambda(c)$ and $\delta \mapsto \delta$ form Donsker classes if θ and Λ range freely. Hence the functions $x \mapsto \Lambda(c)Q_{\theta,\Lambda}(x)$ form a Donsker class, by Example 19.20. The efficiency of $\hat\theta_n$ follows by Theorem 25.77.

25.84 ***Lemma.*** *Under the conditions listed previously, there exists a constant C such that, for every $\varepsilon > 0$,*

$$\log N_{[\,]}\Big(\varepsilon, \big\{m_{\theta,\Lambda}, (\theta, \Lambda)\big\}, L_2(P_0)\Big) \leq C\left(\frac{1}{\varepsilon}\right).$$

Proof. First consider the class of functions $m_{\theta,\Lambda}$ for a fixed θ. These functions depend on Λ monotonely if considered separately for $\delta = 0$ and $\delta = 1$. Thus a bracket $\Lambda_1 \leq \Lambda \leq \Lambda_2$ for Λ leads, by substitution, readily to a bracket for $m_{\theta,\Lambda}$. Furthermore, because this dependence is Lipschitz, there exists a constant D such that

$$\int (m_{\theta,\Lambda_1} - m_{\theta,\Lambda_2})^2\, dF_{C,Z} \leq D\int_\sigma^\tau \big(\Lambda_1(c) - \Lambda_2(c)\big)^2\, dc.$$

Thus, brackets for Λ of L_2-size ε translate into brackets for $m_{\theta,\Lambda}$ of $L_2(P_{\theta,\Lambda})$-size proportional to ε. By Example 19.11 we can cover the set of all Λ by $\exp C(1/\varepsilon)$ brackets of size ε.

Next, we allow θ to vary freely as well. Because θ is finite-dimensional and $\partial/\partial\theta\, m_{\theta,\Lambda}(x)$ is uniformly bounded in (θ, Λ, x), this increases the entropy only slightly. ■

25.85 ***Lemma.*** *Under the conditions listed previously there exist constants $C, \varepsilon > 0$ such that, for all Λ and all $\|\theta - \theta_0\| < \varepsilon$,*

$$\int \left(p_{\theta,\Lambda}^{1/2} - p_{\theta_0,\Lambda_0}^{1/2}\right)^2 d\mu \geq C \int_\sigma^\tau (\Lambda - \Lambda_0)^2(c)\, dc + C\|\theta - \theta_0\|^2.$$

Proof. The left side of the lemma can be rewritten as

$$\int \frac{(p_{\theta,\Lambda} - p_{\theta_0,\Lambda_0})^2}{\left(p_{\theta,\Lambda}^{1/2} + p_{\theta_0,\Lambda_0}^{1/2}\right)^2}\, d\mu.$$

Because p_0 is bounded away from zero, and the densities $p_{\theta,\Lambda}$ are uniformly bounded, the denominator can be bounded above and below by positive constants. Thus the Hellinger distance (in the display) is equivalent to the L_2-distance between the densities, which can be rewritten

$$2\int \left[e^{-e^{\theta^T z}\Lambda(c)} - e^{-e^{\theta_0^T z}\Lambda_0(c)}\right]^2 dF^{Y,Z}(c, z).$$

Let $g(t)$ be the function $\exp\bigl(-e^{\theta^T z}\Lambda(c)\bigr)$ evaluated at $\theta_t = t\theta + (1-t)\theta_0$ and $\Lambda_t = t\Lambda + (1-t)\Lambda_0$, for fixed (c, z). Then the integrand is equal to $\bigl(g(1) - g(0)\bigr)^2$, and hence, by the mean value theorem, there exists $0 \leq t = t(c, z) \leq 1$ such that the preceding display is equal to

$$P_0\left(e^{-\Lambda_t(c)e^{\theta_t^T z}} e^{\theta_t^T z}\left[(\Lambda - \Lambda_0)(c)\bigl(1 + t(\theta - \theta_0)^T z\bigr) + (\theta - \theta_0)^T z\Lambda_0(c)\right]\right)^2.$$

Here the multiplicative factor $e^{-\Lambda_t(c)e^{\theta_t^T z}} e^{\theta_t^T z}$ is bounded away from zero. By dropping this term we obtain, up to a constant, a lower bound for the left side of the lemma. Next, because the function Q_{θ_0,Λ_0} is bounded away from zero and infinity, we may add a factor Q_{θ_0,Λ_0}^2, and obtain the lower bound, up to a constant,

$$P_0\Bigl(\bigl(1 + t(\theta - \theta_0)^T z\bigr) B_{\theta_0,\Lambda_0}(\Lambda - \Lambda_0)(x) + (\theta - \theta_0)^T \dot{\ell}_{\theta_0,\Lambda_0}(x)\Bigr)^2.$$

Here the function $h = \bigl(1 + t(\theta - \theta_0)^T z\bigr)$ is uniformly close to 1 if θ is close to θ_0. Furthermore, for any function g and vector a,

$$\begin{aligned}\bigl(P_0(B_{\theta_0,\Lambda_0} g)a^T \dot{\ell}_{\theta_0,\Lambda_0}\bigr)^2 &= \bigl(P_0(B_{\theta_0,\Lambda_0} g)a^T(\dot{\ell}_{\theta_0,\Lambda_0} - \tilde{\ell}_0)\bigr)^2 \\ &\leq P_0(B_{\theta_0,\Lambda_0} g)^2\, a^T(I_0 - \tilde{I}_0)a,\end{aligned}$$

by the Cauchy-Schwarz inequality. Because the efficient information $\tilde{I}_0$ is positive-definite, the term $a^T(I_0 - \tilde{I}_0)a$ on the right can be written $a^T I_0 a c$ for a constant $0 < c < 1$. The lemma now follows by application of Lemma 25.86 ahead. ■

25.86 ***Lemma.*** *Let h, g_1 and g_2 be measurable functions such that $c_1 \le h \le c_2$ and $(Pg_1g_2)^2 \le cPg_1^2Pg_2^2$ for a constant $c < 1$ and constants $c_1 < 1 < c_2$ close to 1. Then*

$$P(hg_1 + g_2)^2 \ge C\big(Pg_1^2 + Pg_2^2\big),$$

for a constant C depending on c, c_1 and c_2 that approaches $1 - \sqrt{c}$ as $c_1 \uparrow 1$ and $c_2 \downarrow 1$.

Proof. We may first use the inequalities

$$\begin{aligned}(hg_1 + g_2)^2 &\ge c_1hg_1^2 + 2hg_1g_2 + c_2^{-1}hg_2^2\\ &= h(g_1 + g_2)^2 + (c_1 - 1)hg_1^2 + \big(1 - c_2^{-1}\big)hg_2^2\\ &\ge c_1\big(g_1^2 + 2g_1g_2 + g_2^2\big) + (c_1 - 1)c_2g_1^2 + \big(c_2^{-1} - 1\big)g_2^2.\end{aligned}$$

Next, we integrate this with respect to P, and use the inequality for Pg_1g_2 on the second term to see that the left side of the lemma is bounded below by

$$c_1\big(Pg_1^2 - 2\sqrt{cPg_1^2Pg_2^2} + Pg_2^2\big) + (c_1 - 1)c_2Pg_1^2 + \big(c_2^{-1} - 1\big)c_2Pg_2^2.$$

Finally, we apply the inequality $2xy \le x^2 + y^2$ on the second term. ■

25.11.2 *Exponential Frailty*

Suppose that the observations are a random sample from the density of $X = (U, V)$ given by

$$p_{\theta,\eta}(u, v) = \int ze^{-zu}\,\theta ze^{-\theta zv}\,d\eta(z).$$

This is a density with respect to Lebesgue measure on the positive quadrant of $\mathbb{R}^2$, and we may take the likelihood equal to just the joint density of the observations. Let $(\hat\theta_n, \hat\eta_n)$ maximize

$$(\theta, \eta) \mapsto \prod_{i=1}^n p_{\theta,\eta}(U_i, V_i).$$

This estimator can be shown to be consistent, under some conditions, for the Euclidean and weak topology, respectively, by, for instance, the method of Wald, Theorem 5.14.

The "statistic" $\psi_\theta(U, V) = U + \theta V$ is, for fixed and known θ, sufficient for the nuisance parameter. Because the likelihood depends on η only through this statistic, the tangent set ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ for η consists of functions of $U + \theta V$ only. Furthermore, because $U + \theta V$ is distributed according to a mixture over an exponential family (a gamma-distribution with shape parameter 2), the closed linear span of ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ consists of all mean-zero, square-integrable functions of $U + \theta V$, by Example 25.35. Thus, the projection onto the closed linear span of ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ is the conditional expectation with respect to $U + \theta V$, and the efficient score function for θ is the "conditional score," given by

$$\begin{aligned}\tilde\ell_{\theta,\eta}(x) &= \dot\ell_{\theta,\eta}(x) - \mathrm{E}_\theta\big(\dot\ell_{\theta,\eta}(X) \mid \psi_\theta(X) = \psi_\theta(x)\big)\\ &= \frac{\int \frac{1}{2}(u - \theta v)z^3e^{-z(u+\theta v)}\,d\eta(z)}{\int \theta z^2e^{-z(u+\theta v)}\,d\eta(z)},\end{aligned}$$

where we may use that, given $U + \theta V = s$, the variables U and θV are uniformly distributed on the interval $[0, s]$. This function turns out to be also an actual score function, in that there exists an exact least favorable submodel, given by

$$\eta_t(\theta, \eta)(B) = \eta\left(B\left(1 - \frac{t}{2\theta}\right)\right).$$

Inserting $\eta_t(\theta, \eta)$ in the log likelihood, making the change of variables $z\big(1 - t/(2\theta)\big) \to z$, and computing the (ordinary) derivative with respect to t at $t = 0$, we obtain $\tilde{\ell}_{\theta,\eta}(x)$. It follows that the maximum likelihood estimator satisfies the efficient score equation, and its asymptotic normality can be proved with the help of Theorem 25.54.

The linearity of the model in η (or the formula involving the conditional expectation) implies that

$$P_{\theta,\eta_0}\tilde{\ell}_{\theta,\eta} = 0, \qquad \text{every } \theta, \eta, \eta_0.$$

Thus, the "no-bias" condition (25.52) is trivially satisfied. The verification that the functions $\tilde{\ell}_{\theta,\eta}$ form a Donsker class is more involved but is achieved in the following lemma.†

25.87 ***Lemma.*** *Suppose that $\int (z^2 + z^{-5})\, d\eta_0(z) < \infty$. Then there exists a neighborhood V of η_0 for the weak topology such that the class of functions*

$$(x, y) \mapsto \frac{\int (a_1 + a_2 z x + a_3 z y)\, z^2\, e^{-z(b_1 x + b_2 y)}\, d\eta(z)}{\int z^2\, e^{-z(b_1 x + b_2 y)}\, d\eta(z)},$$

where $(a_1, \ldots, a_3)$ ranges over a bounded subset of $\mathbb{R}^3$, (b_1, b_2) ranges over a compact subset of $(0, \infty)^2$, and η ranges over V, is P_{θ_0,η_0}-Donsker with square-integrable envelope.

25.11.3 *Partially Linear Regression*

Suppose that we observe a random sample from the distribution of $X = (V, W, Y)$, in which for some unobservable error e independent of (V, W),

$$Y = \theta V + \eta(W) + e.$$

Thus, the independent variable Y is a regression on (V, W) that is linear in V with slope θ but may depend on W in a nonlinear way. We assume that V and W take their values in the unit interval $[0, 1]$, and that η is twice differentiable with $J(\eta) < \infty$, for

$$J^2(\eta) = \int_0^1 \eta''(w)^2\, dw.$$

This smoothness assumption should help to ensure existence of efficient estimators of θ and will be used to define an estimator.

If the (unobservable) error is assumed to be normal, then the density of the observation $X = (V, W, Y)$ is given by

$$p_{\theta,\eta}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(y - \theta v - \eta(w))^2/\sigma^2}\, p_{V,W}(v, w).$$

† For a proof see [106].

We cannot use this directly to define a maximum likelihood estimator for (θ, η), as a maximizer for η will interpolate the data exactly: A choice of η such that $\eta(w_i) = y_i - \theta v_i$ for every i maximizes $\prod p_{\theta,\eta}(x_i)$ but does not provide a useful estimator. The problem is that so far η has only been restricted to be differentiable, and this does not prevent it from being very wiggly. To remedy this we use a penalized log likelihood estimator, defined as the minimizer of

$$(\theta, \eta) \mapsto \mathbb{P}_n\big(y - \theta v - \eta(w)\big)^2 + \hat{\lambda}_n^2 J^2(\eta).$$

Here $\hat{\lambda}_n$ is a "smoothing parameter" that may depend on the data, and determines the weight of the "penalty" $J^2(\eta)$. A large value of $\hat{\lambda}_n$ gives much influence to the penalty term and hence leads to a smooth estimate of η, and conversely. Intermediate values are best. For the purpose of estimating θ we may use any values in the range

$$\hat{\lambda}_n^2 = o_P(n^{-1/2}), \qquad \hat{\lambda}_n^{-1} = O_P(n^{2/5}).$$

There are simple numerical schemes to compute the maximizer $(\hat{\theta}_n, \hat{\eta}_n)$, the function $\hat{\eta}_n$ being a natural cubic spline with knots at the values $w_1, \ldots, w_n$. The sequence $\hat{\theta}_n$ can be shown to be asymptotically efficient provided that the regression components involving V and W are not confounded or degenerate. More precisely, we assume that the conditional distribution of V given W is nondegenerate, that the distribution of W has at least two support points, and that $h_0(w) = \mathrm{E}(V \mid W = w)$ has a version with $J(h_0) < \infty$. Then, we have the following lemma on the behavior of $(\hat{\theta}_n, \hat{\eta}_n)$.

Let $\|\cdot\|_W$ denote the norm of $L_2(P_W)$.

25.88 ***Lemma.*** *Under the conditions listed previously, the sequence $\hat{\theta}_n$ is consistent for θ_0, $\|\hat{\eta}_n\|_\infty = O_P(1)$, $J(\hat{\eta}_n) = O_P(1)$, and $\|\hat{\eta}_n - \eta\|_W = O_P(\hat{\lambda}_n)$, under (θ_0, η_0).*

Proof. Write $g(v, w) = \theta v + \eta(w)$, let $\mathbb{P}_n$ and P_0 be the empirical and true distribution of the variables (e_i, V_i, W_i), and define functions

$$m_{g,\lambda}(e, v, w) = \big(y - g(v, w)\big)^2 + \lambda^2\big(J^2(\eta) - J^2(\eta_0)\big).$$

Then $\hat{g}(v, w) = \hat{\theta} v + \hat{\eta}(w)$ minimizes $g \mapsto \mathbb{P}_n m_{g,\hat{\lambda}}$, and

$$m_{g,\lambda} - m_{g_0,\lambda} = 2e(g_0 - g) + (g_0 - g)^2 + \lambda^2 J^2(\eta) - \lambda^2 J^2(\eta_0).$$

By the orthogonality property of a conditional expectation and the Cauchy-Schwarz inequality, $(\mathrm{E}V\eta(W))^2 \le \mathrm{E}\mathrm{E}(V \mid W)^2 \mathrm{E}\eta^2(W) < \mathrm{E}V^2 \|\eta\|_W^2$. Therefore, by Lemma 25.86,

$$P_0(g - g_0)^2 \gtrsim |\theta - \theta_0|^2 + \|\eta - \eta_0\|_W^2.$$

Consequently, because $P_0 e = 0$ and e is independent of (V, W),

$$P_0(m_{g,\lambda} - m_{g_0,\lambda}) \gtrsim |\theta - \theta_0|^2 + \|\eta - \eta_0\|_W^2 + \lambda^2 J^2(\eta) - \lambda^2.$$

This suggests to apply Theorem 25.81 with $\tau = (\theta, \eta)$ and $d_\lambda^2(\tau, \tau_0)$ equal to the sum of the first three terms on the right.

Because $\hat{\lambda}_n^{-1} = O_P(1/\lambda_n)$ for $\lambda_n = n^{-2/5}$, it is not a real loss of generality to assume that $\hat{\lambda}_n \in \Lambda_n = [\lambda_n, \infty)$. Then $d_\lambda(\tau, \tau_0) < \delta$ and $\lambda \in \Lambda_n$ implies that $|\theta - \theta_0| < \delta$, that $\|\eta - \eta_0\|_W < \delta$ and that $J(\eta) \le \delta/\lambda_n$. Assume first that it is known already that $|\hat{\theta}|$ and

$\|\hat\eta\|_\infty$ are bounded in probability, so that it is not a real loss of generality to assume that $|\hat\theta| \vee \|\hat\eta\|_\infty \le 1$. Then

$$P_0\big(e^{|e\eta|} - 1 - |e\eta|\big) = \sum_{m\ge 2} P_0 \frac{|e\eta|^m}{m!} \le P_0\eta^2 \mathrm{E} e^{|e|}.$$

Thus a bound on the $\|\cdot\|_W$-norm of η yields a bound on the "Bernstein norm" of $e\eta$ (given on the left) of proportional magnitude. A bracket $[\eta_1, \eta_2]$ for η induces a bracket $[e^+\eta_1 - e^-\eta_2, e^+\eta_2 - e^-\eta_1]$ for the functions $e\eta$. In view of Lemma 19.37 and Example 19.10, we obtain

$$\mathrm{E} \sup_{d_\lambda(\tau,\tau_0)<\delta} \big|\mathbb{G}_n e(\eta - \eta_0)\big| \lesssim \phi_n(\delta) := J_n(\delta)\left(1 + \frac{J_n(\delta)}{\delta^2\sqrt{n}}\right),$$

for

$$J_n(\delta) = \int_0^\delta \sqrt{\left(\frac{1 + \delta/\lambda_n}{\varepsilon}\right)^{1/2}}\, d\varepsilon \lesssim \delta^{3/4} + \frac{\delta}{\lambda_n^{1/4}}.$$

This bound remains valid if we replace $\eta - \eta_0$ by $g - g_0$, for the parametric part θv adds little to the entropy. We can obtain a similar maximal inequality for the process $\mathbb{G}_n(g - g_0)^2$, in view of the inequality $P_0(g - g_0)^4 \le 4P_0(g - g_0)^2$, still under our assumption that $|\theta| \vee \|\eta\|_\infty \le 1$. We conclude that Theorem 25.81 applies and yields the rate of convergence $|\hat\theta - \theta_0| + \|\hat\eta - \eta_0\|_W = O_P(n^{-2/5} + \hat\lambda_n) = O_P(\hat\lambda_n)$.

Finally, we must prove that $\hat\theta$ and $\|\hat\eta\|_\infty$ are bounded in probability. By the Cauchy-Schwarz inequality, for every w and η,

$$\big|\eta(w) - \eta(0) - \eta'(0)w\big| \le \int_0^w \int_0^u |\eta''|(s)\, ds\, du \le J(\eta).$$

This implies that $\|\eta\|_\infty \le |\eta(0)| + \big|\eta'(0)\big| + J(\eta)$, whence it sufficies to show that $\hat\theta$, $\hat\eta(0)$, $\hat\eta'(0)$, and $J(\hat\eta)$ remain bounded. The preceding display implies that

$$\big|\theta v + \eta(0) + \eta'(0)w\big| \le \big|g(v, w)\big| + J(\eta).$$

The empirical measure applied to the square of the left side is equal to $a^T A_n a$ for $a = \big(\theta, \eta(0), \eta'(0)\big)$ and $A_n = \mathbb{P}_n(v, 1, w)(v, 1, w)^T$ the sample second moment matrix of the variables $(V_i, 1, W_i)$. By the conditions on the distribution of (V, W), the corresponding population matrix is positive-definite, whence we can conclude that $\hat a$ is bounded in probability as soon as $\hat a^T A_n \hat a$ is bounded in probability, which is certainly the case if $\mathbb{P}_n \hat g^2$ and $J(\hat\eta)$ are bounded in probability.

We can prove the latter by applying the preceding argument conditionally, given the sequence $V_1, W_1, V_2, W_2, \ldots$. Given these variables, the variables e_i are the only random part in $m_{g,\lambda} - m_{g_0,\lambda}$ and the parts $(g - g_0)^2$ only contribute to the centering function. We apply Theorem 25.81 with square distance equal to

$$d_\lambda^2(\tau, \tau_0) = \mathbb{P}_n(g - g_0)^2 + \lambda^2 J^2(\eta).$$

An appropriate maximal inequality can be derived from, for example Corollary 2.2.8 in [146], because the stochastic process $\mathbb{G}_n eg$ is sub-Gaussian relative to the $L_2(\mathbb{P}_n)$-metric on the set of g. Because $d_\lambda(\tau, \tau_0) < \delta$ implies that $\mathbb{P}_n(g - g_0)^2 < \delta^2$, $J(\eta) \le \delta/\lambda_n$, and $|\theta|^2 \vee \|\eta\|_\infty^2 \le C\big(\mathbb{P}_n(g - g_0)^2 + J^2(\eta)\big)$ for C dependent on the smallest eigenvalue of

the second moment matrix A_n, the maximal inequality has a similar form as before, and we conclude that $\mathbb{P}_n(\hat{g} - g_0)^2 + \hat{\lambda}^2 J^2(\hat{\eta}) = O_P(\hat{\lambda}^2)$. This implies the desired result. ■

The normality of the error e motivates the least squares criterion and is essential for the efficiency of $\hat{\theta}$. However, the penalized least-squares method makes sense also for nonnormal error distributions. The preceding lemma remains true under the more general condition of exponentially small error tails: $\mathrm{E}e^{c|e|} < \infty$ for some $c > 0$.

Under the normality assumption (with $\sigma = 1$ for simplicity) the score function for θ is given by

$$\dot{\ell}_{\theta,\eta}(x) = \big(y - \theta v - \eta(w)\big)v.$$

Given a function h with $J(h) < \infty$, the path $\eta_t = \eta + th$ defines a submodel indexed by the nuisance parameter. This leads to the nuisance score function

$$B_{\theta,\eta}h(x) = \big(y - \theta v - \eta(w)\big)h(w).$$

On comparing these expressions, we see that finding the projection of $\dot{\ell}_{\theta,\eta}$ onto the set of η-scores is a weighted least squares problem. By the independence of e and (V, W), it follows easily that the projection is equal to $B_{\theta,\eta}h_0$ for $h_0(w) = \mathrm{E}(V \mid W = w)$, whence the efficient score function for θ is given by

$$\tilde{\ell}_{\theta,\eta}(x) = \big(y - \theta v - \eta(w)\big)\big(v - h_0(w)\big).$$

Therefore, an exact least-favorable path is given by $\eta_t(\theta, \eta) = \eta - th_0$.

Because $(\hat{\theta}_n, \hat{\eta}_n)$ maximizes a penalized likelihood rather than an ordinary likelihood, it certainly does not satisfy the efficient score equation as considered in section 25.8. However, it satisfies this equation up to a term involving the penalty. Inserting $\big(\hat{\theta} + t, \eta_t(\hat{\theta}, \hat{\eta})\big)$ into the least-squares criterion, and differentiating at $t = 0$, we obtain the stationary equation

$$\mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{\eta}} - 2\hat{\lambda}^2 \int_0^1 \hat{\eta}''(w)h_0''(w)\,dw = 0.$$

The second term is the derivative of $\hat{\lambda}^2 J^2\big(\eta_t(\hat{\theta}, \hat{\eta})\big)$ at $t = 0$. By the Cauchy-Schwarz inequality, it is bounded in absolute value by $2\hat{\lambda}^2 J(\hat{\eta})J(h_0) = o_P(n^{-1/2})$, by the first assumption on $\hat{\lambda}$ and because $J(\hat{\eta}) = O_P(1)$ by Lemma 25.88. We conclude that $(\hat{\theta}_n, \hat{\eta}_n)$ satisfies the efficient score equation up to a $o_P(n^{-1/2})$-term. Within the context of Theorem 25.54 a remainder term of this small order is negligible, and we may use the theorem to obtain the asymptotic normality of $\hat{\theta}_n$.

A formulation that also allows other estimators $\hat{\eta}$ is as follows.

25.89 ***Theorem.*** *Let $\hat{\eta}_n$ be any estimators such that $\|\hat{\eta}_n\|_\infty = O_P(1)$ and $J(\hat{\eta}_n) = O_P(1)$. Then any consistent sequence of estimators $\hat{\theta}_n$ such that $\sqrt{n}\ \mathbb{P}_n\tilde{\ell}_{\hat{\theta},\hat{\eta}} = o_P(1)$ is asymptotically efficient at (θ_0, η_0).*

Proof. It suffices to check the conditions of Theorem 25.54. Since

$$P_{\theta,\eta}\tilde{\ell}_{\theta,\hat{\eta}} = P_{\theta,\eta}\big(\eta(w) - \hat{\eta}(w)\big)\big(v - h_0(w)\big) = 0,$$

for every (θ, η), the no-bias condition (25.52) is satisfied.

That the functions $\tilde{\ell}_{\hat{\theta},\hat{\eta}}$ are contained in a Donsker class, with probability tending to 1, follows from Example 19.10 and Theorem 19.5.

The remaining regularity conditions of Theorem 25.54 can be seen to be satisfied by standard arguments. ■

In this example we use the smoothness of η to define a penalized likelihood estimator for θ. This automatically yields a rate of convergence of $n^{-2/5}$ for $\hat{\eta}$. However, efficient estimators for θ exist under weaker smoothness assumptions on η, and the minimal smoothness of η can be traded against smoothness of the function $g(w) = E(V|W = w)$, which also appears in the formula for the efficient score function and is unknown in practice. The trade-off is a consequence of the bias $P_{\theta,\eta,g}\tilde{\ell}_{\theta,\bar{\eta},\hat{g}}$ being equal to the cross product of the biases in $\hat{\eta}$ and $\hat{g}$. The square terms in the second order expansion (25.60), in which the derivative relative to (η, g) (instead of η) is a (2×2)-matrix, vanish. See [35] for a detailed study of this model.

25.12 Likelihood Equations

The "method of the efficient score equation" isolates the parameter θ of interest and characterizes an estimator $\hat{\theta}$ as the solution of a system of estimating equations. In this system the nuisance parameter has been replaced by an estimator $\hat{\eta}$. If the estimator $\hat{\eta}$ is the maximum likelihood estimator, then we may hope that a solution $\hat{\theta}$ of the efficient score equation is also the maximum likelihood estimator for θ, or that this is approximately true.

Another approach to proving the asymptotic normality of maximum likelihood estimators is to design a system of likelihood equations for the parameter of interest and the nuisance parameter jointly. For a semiparametric model, this necessarily is a system of infinitely many equations.

Such a system can be analyzed much in the same way as a finite-dimensional system. The system is linearized in the estimators by a Taylor expansion around the true parameter, and the limit distribution involves the inverse of the derivative applied to the system of equations. However, in most situations an ordinary pointwise Taylor expansion, the classical argument as employed in the introduction of section 5.3, is impossible, and the argument must involve some advanced tools, in particular empirical processes. A general scheme is given in Theorem 19.26, which is repeated in a different notation here. A limitation of this approach is that both $\hat{\theta}$ and $\hat{\eta}$ must converge at $\sqrt{n}$-rate. It is not clear that a model can always appropriately parametrized such that this is the case; it is certainly not always the case for the natural parametrization.

The system of estimating equations that we are looking for consists of stationary equations resulting from varying either the parameter θ or the nuisance parameter η. Suppose that our maximum likelihood estimator $(\hat{\theta}, \hat{\eta})$ maximizes the function

$$(\theta, \eta) \mapsto \prod \text{lik}(\theta, \eta)(X_i),$$

for $\text{lik}(\theta, \eta)(x)$ being the "likelihood" given one observation x.

The parameter θ can be varied in the usual way, and the resulting stationary equation takes the form

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

This is the usual maximum likelihood equation, except that we evaluate the score function at the joint estimator $(\hat\theta, \hat\eta)$, rather than at the single value $\hat\theta$. A precise condition for this equation to be valid is that the partial derivative of $\log \mathrm{lik}(\theta, \eta)(x)$ with respect to θ exists and is equal to $\dot\ell_{\theta,\eta}(x)$, for every x, (at least for $\eta = \hat\eta$ and at $\theta = \hat\theta$).

Varying the nuisance parameter η is conceptually more difficult. Typically, we can use a selection of the submodels $t \mapsto \eta_t$ used for defining the tangent set and the information in the model. If scores for η take the form of an "operator" $B_{\theta,\eta}$ working on a set of indices h, then a typical likelihood equation takes the form

$$\mathbb{P}_n B_{\hat\theta,\hat\eta} h = P_{\hat\theta,\hat\eta} B_{\hat\theta,\hat\eta} h.$$

Here we have made it explicit in our notation that a score function always has mean zero, by writing the score function as $x \mapsto B_{\theta,\eta}h(x) - P_{\theta,\eta}B_{\theta,\eta}h$ rather than as $x \mapsto B_{\theta,\eta}h(x)$. The preceding display is valid if, for every (θ, η), there exists some path $t \mapsto \eta_t(\theta, \eta)$ such that $\eta_0(\theta, \eta) = \eta$ and, for every x,

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

Assume that this is the case for every h in some index set $\mathcal{H}$, and suppose that the latter is chosen in such a way that the map $h \mapsto B_{\theta,\eta}h(x) - P_{\theta,\eta}B_{\theta,\eta}h$ is uniformly bounded on $\mathcal{H}$, for every x and every (θ, η).

Then we can define random maps $\Psi_n : \mathbb{R}^k \times H \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ by $\Psi_n = (\Psi_{n1}, \Psi_{n2})$ with

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

The expectation of these maps under the parameter (θ_0, η_0) is the deterministic map $\Psi = (\Psi_1, \Psi_2)$ given by

$$\begin{aligned} \Psi_1(\theta, \eta) &= P_{\theta_0,\eta_0}\dot\ell_{\theta,\eta}, \\ \Psi_2(\theta, \eta)h &= P_{\theta_0,\eta_0}B_{\theta,\eta}h - P_{\theta,\eta}B_{\theta,\eta}h, \qquad h \in \mathcal{H}. \end{aligned}$$

By construction, the maximum likelihood estimators $(\hat\theta_n, \hat\eta_n)$ and the "true" parameter (θ_0, η_0) are zeros of these maps,

$$\Psi_n(\hat\theta_n, \hat\eta_n) = 0 = \Psi(\theta_0, \eta_0).$$

The argument next proceeds by linearizing these equations. Assume that the parameter set H for η can be identified with a subset of a Banach space. Then an adaptation of Theorem 19.26 is as follows.

25.90 *Theorem.* *Suppose that the functions $\dot\ell_{\theta,\eta}$ and $B_{\theta,\eta}h$, if h ranges over $\mathcal{H}$ and (θ, η) over a neighborhood of (θ_0, η_0), are contained in a P_{θ_0,η_0}-Donsker class, and that*

$$P_{\theta_0,\eta_0}\big\|\dot\ell_{\hat\theta,\hat\eta} - \dot\ell_{\theta_0,\eta_0}\big\|^2 \xrightarrow{\mathrm{P}} 0, \qquad \sup_{h\in\mathcal{H}} P_{\theta_0,\eta_0}\big|B_{\hat\theta,\hat\eta}h - B_{\theta_0,\eta_0}h\big|^2 \xrightarrow{\mathrm{P}} 0.$$

Furthermore, suppose that the map $\Psi : \Theta \times H \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ is Fréchet-differentiable at (θ_0, η_0), with a derivative $\dot\Psi_0 : \mathbb{R}^k \times \mathrm{lin}\, H \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ that has a continuous inverse

on its range. If the sequence $(\hat\theta_n, \hat\eta_n)$ *is consistent for* (θ_0, η_0) *and satisfies* $\Psi_n(\hat\theta_n, \hat\eta_n) = o_P(n^{-1/2})$, *then*

$$\dot\Psi_0\sqrt{n}(\hat\theta_n - \theta_0, \hat\eta_n - \eta_0) = -\sqrt{n}\Psi_n(\theta_0, \eta_0) + o_P(1).$$

The theorem gives the joint asymptotic distribution of $\hat\theta_n$ and $\hat\eta_n$. Because $\sqrt{n}\Psi_n(\theta_0, \eta_0)$ is the empirical process indexed by the Donsker class consisting of the functions $\dot\ell_{\theta_0,\eta_0}$ and $B_{\theta_0,\eta_0}h$, this process is asymptotically normally distributed. Because normality is retained under a continuous, linear map, such as $\dot\Psi_0^{-1}$, the limit distribution of the sequence $\sqrt{n}(\hat\theta_n - \theta_0, \hat\eta_n - \eta_0)$ is Gaussian as well.

The case of a partitioned parameter (θ, η) is an interesting one and illustrates most aspects of the application of the preceding theorem. Therefore, we continue to write the formulas in the corresponding partitioned form. However, the preceding theorem, applies more generally. In Example 25.5.1 we wrote the score operator for a semiparametric model in the form

$$A_{\theta,\eta}(a, b) = a^T\dot\ell_{\theta,\eta} + B_{\theta,\eta}b.$$

Corresponding to this, the system of likelihood equations can be written in the form

$$\mathbb{P}_n A_{\theta,\eta}(a, b) = P_{\theta,\eta}A_{\theta,\eta}(a, b), \qquad \text{every } (a, b).$$

If the partitioned parameter (θ, η) and the partitioned "directions" (a, b) are replaced by a general parameter τ and general direction c, then this formulation extends to general models. The maps Ψ_n and Ψ then take the forms

$$\Psi_n(\tau)c = \mathbb{P}_n A_\tau c - P_\tau A_\tau g, \qquad \Psi(\tau)c = P_{\tau_0}A_\tau c - P_\tau A_\tau c.$$

The theorem requires that these can be considered maps from the parameter set into a Banach space, for instance a space $\ell^\infty(C)$.

To gain more insight, consider the case that η is a measure on a measurable space $(\mathcal{Z}, \mathcal{C})$. Then the directions h can often be taken equal to bounded functions $h : \mathcal{Z} \mapsto \mathbb{R}$, corresponding to the paths $d\eta_t = (1 + th)\,d\eta$ if η is a completely unknown measure, or $d\eta_t = \big(1 + t(h - \eta h)\big)\,d\eta$ if the total mass of each η is fixed to one. In the remainder of the discussion, we assume the latter. Now the derivative map $\dot\Psi_0$ typically takes the form

$$(\theta - \theta_0, \eta - \eta_0) \mapsto \begin{pmatrix} \dot\Psi_{11} & \dot\Psi_{12} \\ \dot\Psi_{21} & \dot\Psi_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix}$$

where

$$\begin{aligned} \dot\Psi_{11}(\theta - \theta_0) &= -P_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}\dot\ell^T_{\theta_0,\eta_0}(\theta - \theta_0), \\ \dot\Psi_{12}(\eta - \eta_0) &= -\int B^*_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}\,d(\eta - \eta_0), \\ \dot\Psi_{21}(\theta - \theta_0)h &= -P_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}h)\dot\ell^T_{\theta_0,\eta_0}(\theta - \theta_0), \\ \dot\Psi_{22}(\eta - \eta_0)h &= -\int B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}h\,d(\eta - \eta_0). \end{aligned} \tag{25.91}$$

For instance, to find the last identity in an informal manner, consider a path η_t in the direction of g, so that $d\eta_t - d\eta_0 = tg\, d\eta_0 + o(t)$. Then by the definition of a derivative

$$\Psi_2(\theta_0, \eta_t) - \Psi_2(\theta_0, \eta_0) \approx \dot{\Psi}_{22}(\eta_t - \eta_0) + o(t).$$

On the other hand, by the definition of Ψ, for every h,

$$\begin{aligned}\Psi_2(\theta_0, \eta_t)h - \Psi(\theta_0, \eta_0)h &= -(P_{\theta_0,\eta_t} - P_{\theta_0,\eta_0})B_{\theta_0,\eta_t}h \\ &\approx -t\,P_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}g)(B_{\theta_0,\eta_0}h) + o(t) \\ &= -\int (B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}h)\, tg\, d\eta_0 + o(t).\end{aligned}$$

On comparing the preceding pair of displays, we obtain the last line of (25.91), at least for $d\eta - d\eta_0 = g\, d\eta_0$. These arguments are purely heuristic, and this form of the derivative must be established for every example. For instance, within the context of Theorem 25.90, we may need to apply $\dot{\Psi}_0$ to η that are not absolutely continuous with respect to η_0. Then the validity of (25.91) depends on the version that is used to define the adjoint operator $B^*_{\theta_0,\eta_0}$. By definition, an adjoint is an operator between L_2-spaces and hence maps equivalence classes into equivalence classes.

The four partial derivatives $\dot{\Psi}_{ij}$ in (25.91) involve the four parts of the information operator $A^*_{\theta,\eta}A_{\theta,\eta}$, which was written in a partitioned form in Example 25.5.1. In particular, the map $\dot{\Psi}_{11}$ is exactly the Fisher information for θ, and the operator $\dot{\Psi}_{22}$ is defined in terms of the information operator for the nuisance parameter. This is no coincidence, because the formulas can be considered a version of the general identity "expectation of the second derivative is equal to minus the information." An abstract form of the preceding argument applied to the map $\Psi(\tau)c = P_{\tau_0}A_\tau c - P_\tau A_\tau c$ leads to the identity, with τ_t a path with derivative $\dot{\tau}_0$ at $t = 0$ and score function $A_{\tau_0}d$,

$$\dot{\Psi}_0(\dot{\tau}_0)c = \langle A^*_{\tau_0}A_{\tau_0}c, d\rangle_{\tau_0}.$$

In the case of a partitioned parameter $\tau = (\theta, \eta)$, the inner inner product on the right is defined as $\langle (a, b), (\alpha, \beta)\rangle_{\tau_0} = a^T\alpha + \int b\beta\, d\eta_0$, and the four formulas in (25.91) follow by Example 25.5.1 and some algebra. A difference with the finite-dimensional situation is that the derivatives $\dot{\tau}_0$ may not be dense in the domain of $\dot{\Psi}_0$, so that the formula determines $\dot{\Psi}_0$ only partly.

An important condition in Theorem 25.90 is the continuous invertibility of the derivative. Because a linear map between Euclidean spaces is automatically continuous, in the finite-dimensional set-up this condition reduces to the derivative being one-to-one. For infinite-dimensional systems of estimating equations, the continuity is far from automatic and may be the condition that is hardest to verify. Because it refers to the $\ell^\infty(\mathcal{H})$-norm, we have some control over it while setting up the system of estimating equations and choosing the set of functions $\mathcal{H}$. A bigger set $\mathcal{H}$ makes $\dot{\Psi}_0^{-1}$ more readily continuous but makes the differentiability of Ψ and the Donsker condition more stringent.

In the partitioned case, the continuous invertibility of $\dot{\Psi}_0$ can be verified by ascertaining the continuous invertibility of the two operators $\dot{\Psi}_{11}$ and $\dot{V} = \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}$. In that case we have

$$\dot{\Psi}_0^{-1} = \begin{pmatrix} \dot{\Psi}_{11}^{-1} + \dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}\dot{V}^{-1}\dot{\Psi}_{21}\dot{\Psi}_{11}^{-1} & -\dot{\Psi}_{11}^{-1}\Psi_{12}\dot{V}^{-1} \\ -\dot{V}^{-1}\dot{\Psi}_{21}\dot{\Psi}_{11}^{-1} & \dot{V}^{-1} \end{pmatrix}$$

The operator $\dot{\Psi}_{11}$ is the Fisher information matrix for θ if η is known. If this would not be invertible, then there would be no hope of finding asymptotically normal estimators for θ. The operator $\dot{V}$ has the form

$$\dot{V}(\eta - \eta_0)h = -\int \left(B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} + K\right)h \, d(\eta - \eta_0),$$

where the operator K is defined as

$$Kh = -\left(P_{\theta_0,\eta_0}\left(B_{\theta_0,\eta_0}h\right)\dot{\ell}^T_{\theta_0,\eta_0}\right) I^{-1}_{\theta_0,\eta_0} B^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0}.$$

The operator $\dot{V}: \operatorname{lin} H \mapsto \ell^\infty(\mathcal{H})$ is certainly continuously invertible if there exists a positive number ϵ such that

$$\sup_{h\in\mathcal{H}} \left|\dot{V}(\eta - \eta_0)h\right| \geq \varepsilon \|\eta - \eta_0\|.$$

In the case that η is identified with the map $h \mapsto \eta h$ in $\ell^\infty(\mathcal{H})$, the norm on the right is given by $\sup_{h\in\mathcal{H}}\left|(\eta - \eta_0)h\right|$. Then the display is certainly satisfied if, for some $\varepsilon > 0$,

$$\left\{\left(B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} + K\right)h : h \in \mathcal{H}\right\} \supset \epsilon\mathcal{H}.$$

This condition has a nice interpretation if $\mathcal{H}$ is equal to the unit ball of a Banach space $\mathbb{B}$ of functions. Then the preceding display is equivalent to the operator $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} + K : \mathbb{B} \mapsto \mathbb{B}$ being continuously invertible. The first part of this operator is the information operator for the nuisance parameter. Typically, this is continuously invertible if the nuisance parameter is regularly estimable at a $\sqrt{n}$-rate (relatively to the norm used) if θ is known. The following lemma guarantees that the same is then true for the operator $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} + K$ if the efficient information matrix for θ is nonsingular, that is, the parameters θ and η are not locally confounded.

25.92 *Lemma.* *Let $\mathbb{B}$ be a Banach space contained in $\ell^\infty(\mathcal{L})$. If $\tilde{I}_{\theta_0,\eta_0}$ is nonsingular, $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} : \mathbb{B} \mapsto \mathbb{B}$ is onto and continuously invertible and $B^*_{\theta_0,\eta_0}\dot{\ell}_{\theta_0,\eta_0} \in \mathbb{B}$, then $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} + K : \mathbb{B} \mapsto \mathbb{B}$ is onto and continuously invertible.*

Proof. Abbreviate the index (θ_0, η_0) to 0. The operator K is compact, because it has a finite-dimensional range. Therefore, by Lemma 25.93 below, the operator $B_0^* B_0 + K$ is continuously invertible provided that it is one-to-one.

Suppose that $(B_0^* B_0 + K)h = 0$ for some $h \in \mathbb{B}$. By assumption there exists a path $t \mapsto \eta_t$ with score function $\overline{B}_0 h = B_0 h - P_0 B_0 h$ at $t = 0$. Then the submodel indexed by $t \mapsto (\theta_0 + t a_0, \eta_t)$, for $a_0 = -I_0^{-1} P_0(B_0 h)\dot{\ell}_0$, has score function $a_0^T\dot{\ell}_0 + \overline{B}_0 h$ at $t = 0$, and information

$$a_0^T I_0 a_0 + P_0(\overline{B}_0 h)^2 + 2a_0^T P_0\dot{\ell}_0(B_0 h) = P_0(\overline{B}_0 h)^2 - a_0^T I_0 a_0.$$

Because the efficient information matrix is nonsingular, this information must be strictly positive, unless $a_0 = 0$. On the other hand,

$$0 = \eta_0 h(B_0^* B_0 + K)h = P_0(B_0 h)^2 + a_0^T P_0(B_0 h)\dot{\ell}_0.$$

This expression is at least the right side of the preceding display and is positive if $a_0 \neq 0$. Thus $a_0 = 0$, whence $Kh = 0$. Reinserting this in the equation $(B_0^* B_0 + K)h = 0$, we find that $B_0^* B_0 h = 0$ and hence $h = 0$. ■

The proof of the preceding lemma is based on the Fredholm theory of linear operators. An operator $K : \mathbb{B} \mapsto \mathbb{B}$ is *compact* if it maps the unit ball into a totally bounded set. The following lemma shows that for certain operators continuous invertibility is a consequence of their being one-to-one, as is true for matrix operators on Euclidean space.† It is also useful to prove the invertibility of the information operator itself.

25.93 ***Lemma.*** *Let $\mathbb{B}$ be a Banach space, let the operator $A : \mathbb{B} \mapsto \mathbb{B}$ be continuous, onto and continuously invertible and let $K : \mathbb{B} \mapsto \mathbb{B}$ be a compact operator. Then $\mathrm{R}(A + K)$ is closed and has codimension equal to the dimension of $\mathrm{N}(A + K)$. In particular, if $A + K$ is one-to-one, then $A + K$ is onto and continuously invertible.*

The asymptotic covariance matrix of the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ can be computed from the expression for $\dot{\Psi}_0$ and the covariance function of the limiting process of the sequence $\sqrt{n}\Psi_n(\theta_0, \eta_0)$. However, it is easier to use an asymptotic representation of $\sqrt{n}(\hat{\theta}_n - \theta_0)$ as a sum. For a continuously invertible information operator $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0}$ this can be obtained as follows.

In view of (25.91), the assertion of Theorem 25.90 can be rewritten as the system of equations, with a subscript 0 denoting (θ_0, η_0),

$$\begin{aligned} -I_0(\hat{\theta}_n - \theta_0) - (\hat{\eta}_n - \eta_0)B_0^* \dot{\ell}_0 &= -(\mathbb{P}_n - P_0)\dot{\ell}_0 + o_P(1/\sqrt{n}), \\ -P_0(B_0 h)\dot{\ell}_0^T(\hat{\theta}_n - \theta_0) - (\hat{\eta}_n - \eta_0)B_0^* B_0 h &= -(\mathbb{P}_n - P_0)B_0 h + o_P(1/\sqrt{n}). \end{aligned}$$

The $o_P(1/\sqrt{n})$-term in the second line is valid for every $h \in \mathcal{H}$ (uniformly in h). If we can also choose $h = (B_0^* B_0)^{-1} B_0^* \dot{\ell}_0$, and subtract the first equation from the second, then we arrive at

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

Here $\tilde{\ell}_{\theta_0,\eta_0}$ is the efficient score function for θ, as given by (25.33), and $\tilde{I}_{\theta_0,\eta_0}$ is the efficient information matrix. The representation shows that the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically linear in the efficient influence function for estimating θ. Hence the maximum likelihood estimator $\hat{\theta}$ is asymptotically efficient.‡ The asymptotic efficiency of the estimator $\hat{\eta}h$ for ηh follows similarly.

We finish this section with a number of examples. For each example we describe the general structure and main points of the verification of the conditions of Theorem 25.90, but we refer to the original papers for some of the details.

25.12.1 *Cox Model*

Suppose that we observe a random sample from the distribution of the variable $X = \big(T \wedge C, 1\{T \leq C\}, Z\big)$, where, given Z, the variables T and C are independent, as in the

† For a proof see, for example, [132, pp. 99–103].

‡ This conclusion also can be reached from general results on the asymptotic efficiency of the maximum likelihood estimator. See [56] and [143].

random censoring model, and T follows the Cox model. Thus, the density of $X = (Y, \Delta, Z)$ is given by

$$\left(e^{\theta z}\lambda(y)e^{-e^{\theta z}\Lambda(y)}\left(1 - F_{C|Z}(y-|z)\right)\right)^{\delta}\left(e^{-e^{\theta z}\Lambda(y)}f_{C|Z}(y|z)\right)^{1-\delta}p_Z(z).$$

We define a likelihood for the parameters (θ, Λ) by dropping the factors involving the distribution of (C, Z), and replacing $\lambda(y)$ by the pointmass $\Lambda\{y\}$,

$$\text{lik}(\theta, \Lambda)(x) = \left(e^{\theta z}\Lambda\{y\}e^{-e^{\theta z}\Lambda(y)}\right)^{\delta}\left(e^{-e^{\theta z}\Lambda(y)}\right)^{1-\delta}.$$

This likelihood is convenient in that the profile likelihood function for θ can be computed explicitly, exactly as in Example 25.69. Next, given the maximizer $\hat{\theta}$, which must be calculated numerically, the maximum likelihood estimator $\hat{\Lambda}$ is given by an explicit formula.

Given the general results put into place so far, proving the consistency of $(\hat{\theta}, \hat{\Lambda})$ is the hardest problem. The methods of section 5.2 do not apply directly, because of the empirical factor $\Lambda\{y\}$ in the likelihood. These methods can be adapted. Alternatively, the consistency can be proved using the explicit form of the profile likelihood function. We omit a discussion.

For simplicity we make a number of partly unnecessary assumptions. First, we assume that the covariate Z is bounded, and that the true conditional distributions of T and C given Z possess continuous Lebesgue densities. Second, we assume that there exists a finite number $\tau > 0$ such that $\mathrm{P}(C \geq \tau) = \mathrm{P}(C = \tau) > 0$ and $P_{\theta_0,\Lambda_0}(T > \tau) > 0$. The latter condition is not unnatural: It is satisfied if the survival study is stopped at some time τ at which a positive fraction of individuals is still "at risk" (alive). Third, we assume that, for any measurable function h, the probability that $Z \neq h(Y)$ is positive. The function Λ now matters only on $[0, \tau]$; we shall identify Λ with its restriction to this interval. Under these conditions the maximum likelihood estimator $(\hat{\theta}, \hat{\Lambda})$ can be shown to be consistent for the product of the Euclidean topology and the topology of uniform convergence on $[0, \tau]$.

The score function for θ takes the form

$$\dot{\ell}_{\theta,\Lambda}(x) = \delta z - ze^{\theta z}\Lambda(y).$$

For any bounded, measurable function $h : [0, \tau] \mapsto \mathbb{R}$, the path defined by $d\Lambda_t = (1 + th)\,d\Lambda$ defines a submodel passing through Λ at $t = 0$. Its score function at $t = 0$ takes the form

$$B_{\theta,\Lambda}h(x) = \delta h(y) - e^{\theta z}\int_{[0,y]} h\,d\Lambda.$$

The function $h \mapsto B_{\theta,\Lambda}h(x)$ is bounded on every set of uniformly bounded functions h, for any finite measure Λ, and is even uniformly bounded in x and in (θ, Λ) ranging over a neighborhood of (θ_0, Λ_0).

It is not difficult to find a formula for the adjoint $B^*_{\theta,\Lambda}$ of $B_{\theta,\Lambda} : L_2(\Lambda) \mapsto L_2(P_{\theta,\Lambda})$, but this is tedious and not insightful. The information operator $B^*_{\theta,\Lambda}B_{\theta,\Lambda} : L_2(\Lambda) \mapsto L_2(\Lambda)$ can be calculated from the identity $P_{\theta,\Lambda}(B_{\theta,\Lambda}g)(B_{\theta,\Lambda}h) = \Lambda g(B^*_{\theta,\Lambda}B_{\theta,\Lambda}h)$. For continuous Λ it takes the surprisingly simple form

$$B^*_{\theta,\Lambda}B_{\theta,\Lambda}h(y) = h(y)\,\mathrm{E}_{\theta,\Lambda}1_{Y\geq y}e^{\theta Z}.$$

To see this, write the product $B_{\theta,\Lambda} g B_{\theta,\Lambda} h$ as the sum of four terms

$$\delta h(y) g(y) - \delta h(y) e^{\theta z} \int_0^y g \, d\Lambda - \delta g(y) e^{\theta z} \int_0^y h \, d\Lambda + e^{2\theta z} \int_0^y g \, d\Lambda \int_0^y h \, d\Lambda.$$

Take the expectation under $P_{\theta,\Lambda}$ and interchange the order of the integrals to represent $B^*_{\theta,\Lambda} B_{\theta,\Lambda} h$ also as a sum of four terms. Partially integrate the fourth term to see that this cancels the second and third terms. We are left with the first term. The function $B^*_{\theta,\Lambda} \dot{\ell}_{\theta,\Lambda}$ can be obtained by a similar argument, starting from the identity $P_{\theta,\Lambda} \dot{\ell}_{\theta,\Lambda} B_{\theta,\Lambda} h = \Lambda(B^*_{\theta,\Lambda} \dot{\ell}_{\theta,\Lambda}) h$. It is given by

$$B^*_{\theta,\Lambda} \dot{\ell}_{\theta,\Lambda} = \mathrm{E}_{\theta,\Lambda} 1_{Y \geq y} Z e^{\theta Z}.$$

The calculation of the information operator in this way is instructive, but only to check (25.91) for this example. As in other examples a direct derivation of the derivative of the map $\Psi = (\Psi_1, \Psi_2)$ given by $\Psi_1(\theta, \Lambda) = P_0 \dot{\ell}_{\theta,\Lambda}$ and $\Psi_2(\theta, \Lambda) h = P_0 B_{\theta,\Lambda} h$ requires less work. In the present case this is almost trivial, for the map Ψ is already linear in Λ. Writing $G_0(y \mid Z)$ for the distribution function of Y given Z, this map can be written as

$$\Psi_1(\theta, \Lambda) = \mathrm{E} Z e^{\theta_0 Z} \int \bar{G}_0(y \mid Z) \, d\Lambda_0(y) - \mathrm{E} Z e^{\theta Z} \int \Lambda(y) \, dG_0(y \mid Z),$$

$$\Psi_2(\theta, \Lambda) h = \mathrm{E} e^{\theta_0 Z} \int h(y) \bar{G}_0(y \mid Z) \, d\Lambda_0(y) - \mathrm{E} e^{\theta Z} \iint_{[0,y]} h \, d\Lambda \, dG_0(y \mid Z).$$

If we take $\mathcal{H}$ equal to the unit ball of the space $\mathrm{BV}[0, \tau]$ of bounded functions of bounded variation, then the map $\Psi : \mathbb{R} \times \ell^\infty(\mathcal{H}) \mapsto \mathbb{R} \times \ell^\infty(\mathcal{H})$ is linear and continuous in Λ, and its partial derivatives with respect to θ can be found by differentiation under the expectation and are continuous in a neighborhood of (θ_0, Λ_0). Several applications of Fubini's theorem show that the derivative takes the form (25.91).

We can consider $B_0^* B_0$ as an operator of the space $\mathrm{BV}[0, \tau]$ into itself. Then it is continuously invertible if the function $y \mapsto \mathrm{E}_{\theta_0,\Lambda_0} 1_{Y \geq y} e^{\theta_0 Z}$ is bounded away from zero on $[0, \tau]$. This we have (indirectly) assumed. Thus, we can apply Lemma 25.92. The efficient score function takes the form (25.33), which, with $M_i(y) = \mathrm{E}_{\theta_0,\Lambda_0} 1_{Y \geq y} Z^i e^{\theta_0 Z}$, reduces to

$$\tilde{\ell}_{\theta_0,\Lambda_0}(x) = \delta\left(z - \frac{M_1}{M_0}(y)\right) - e^{\theta_0 z} \int_{[0,y]} \left(z - \frac{M_1}{M_0}(t)\right) d\Lambda_0(t).$$

The efficient information for θ can be computed from this as

$$\tilde{I}_{\theta_0,\Lambda_0} = \mathrm{E} e^{\theta_0 Z} \int \left(Z - \frac{M_1}{M_0}(y)\right)^2 \bar{G}_0(y \mid Z) \, d\Lambda_0(y).$$

This is strictly positive by the assumption that Z is not equal to a function of Y.

The class $\mathcal{H}$ is a universal Donsker class, and hence the first parts $\delta h(y)$ of the functions $B_{\theta,\Lambda} h$ form a Donsker class. The functions of the form $\int_{[0,y]} h \, d\Lambda$ with h ranging over $\mathcal{H}$ and Λ ranging over a collection of measures of uniformly bounded variation are functions of uniformly bounded variation and hence also belong to a Donsker class. Thus the functions $B_{\theta,\Lambda} h$ form a Donsker class by Example 19.20.

25.12.2 *Partially Missing Data*

Suppose that the observations are a random sample from a density of the form

$$(x, y, z) \mapsto \int p_\theta(x \mid s)\, d\eta(s)\, p_\theta(y \mid z)\, d\eta(z) =: p_\theta(x \mid \eta)\, p_\theta(y \mid z)\, d\eta(z).$$

Here the parameter η is a completely unknown distribution, and the kernel $p_\theta(\cdot \mid s)$ is a given parametric model indexed by the parameters θ and s, relative to some density μ. Thus, we obtain equal numbers of bad and good (direct) observations concerning η. Typically, by themselves the bad observations do not contribute positive information concerning the cumulative distribution function η, but along with the good observations they help to cut the asymptotic variance of the maximum likelihood estimators.

25.94 ***Example.*** This model can arise if we are interested in the relationship between a response Y and a covariate Z, but because of the cost of measurement we do not observe Z for a fraction of the population. For instance, a full observation $(Y, Z) = (D, W, Z)$ could consist of

– a logistic regression D on $\exp Z$ with intercept and slope β_0 and β_1, respectively, and
– a linear regression W on Z with intercept and slope α_0 and α_1, respectively, and an $N(0, \sigma^2)$-error.

Given Z the variables D and W are assumed independent, and Z has a completely unspecified distribution η on an interval in $\mathbb{R}$. The kernel is equal to, with Ψ denoting the logistic distribution function and ϕ denoting the standard normal density,

$$p_\theta(d, w \mid z) = \Psi(\beta_0 + \beta_1 e^z)^d \big(1 - \Psi(\beta_0 + \beta_1 e^z)\big)^{1-d} \frac{1}{\sigma} \phi\Big(\frac{w - \alpha_0 - \alpha_1 z}{\sigma}\Big).$$

The precise form of this density does not play a major role in the following.

In this situation the covariate Z is a gold standard, but, in view of the costs of measurement, for a selection of observations only the "surrogate covariate" W is available. For instance, Z corresponds to the LDL cholesterol and W to total cholesterol, and we are interested in heart disease $D = 1$. For simplicity, each observation in our set-up consists of one full observation $(Y, Z) = (D, W, Z)$ and one reduced observation $X = (D, W)$. □

25.95 ***Example.*** If the kernel $p_\theta(y \mid z)$ is equal to the normal density with mean z and variance θ, then the observations are a random sample $Z_1, \ldots, Z_n$ from η, a random sample $X_1, \ldots, X_n$ from η perturbed by an additive (unobserved) normal error, and a sample $Y_1, \ldots, Y_n$ of random variables that given $Z_1, \ldots, Z_n$ are normally distributed with means Z_i and variance θ. In this case the interest is perhaps focused on estimating η, rather than θ. □

The distribution of an observation (X, Y, Z) is given by two densities and a nonparametric part. We choose as likelihood

$$\mathrm{lik}(\theta, \eta)(x, y, z) = p_\theta(x \mid \eta)\, p_\theta(y \mid z)\, \eta\{z\}.$$

Thus, for the completely unknown distribution η of Z we use the empirical likelihood for the other part of the observations we use the density, as usual. It is clear that the maximum

likelihood estimator $\hat{\eta}$ charges all observed values $z_1, \ldots, z_n$, but the term $p_\theta(x \mid \eta)$ leads to some additional support points as well. In general, these are not equal to values of the observations.

The score function for θ is given by

$$\dot{\ell}_{\theta,\eta}(x, y, z) = \dot{\kappa}_{\theta,\eta}(x) + \dot{\kappa}_\theta(y \mid z) = \frac{\int \dot{\kappa}_\theta(x \mid s)\, p_\theta(x \mid s)\, d\eta(s)}{p_\theta(x \mid \eta)} + \dot{\kappa}_\theta(y \mid z).$$

Here $\dot{\kappa}_\theta(y \mid z) = \partial/\partial\theta \log p_\theta(y \mid z)$ is the score function for θ for the conditional density $p_\theta(y \mid z)$, and $\dot{\kappa}_{\theta,\eta}(x)$ is the score function for θ of the mixture density $p_\theta(x \mid \eta)$.

Paths of the form $d\eta_t = (1 + th)\, d\eta$ (with $\eta h = 0$) yield scores

$$B_{\theta,\eta}h(x, z) = C_{\theta,\eta}h(x) + h(z) = \frac{\int h(s) p_\theta(x \mid s)\, d\eta(s)}{p_\theta(x \mid \eta)} + h(z).$$

The operator $C_{\theta,\eta} : L_2(\eta) \mapsto L_2\big(p_\theta(\cdot \mid \eta)\big)$ is the score operator for the mixture part of the model. Its Hilbert-space adjoint is given by

$$C^*_{\theta,\eta} g(z) = \int g(x)\, p_\theta(x \mid z)\, d\mu(x).$$

The range of $B_{\theta,\eta}$ is contained in the subset G of $L_2\big(p_\theta(\cdot \mid \eta) \times \eta\big)$ consisting of functions of the form $(x, z) \mapsto g_1(x) + g_2(z) + c$. This representation of a function of this type is unique if both g_1 and g_2 are taken to be mean-zero functions. With $P_{\theta,\eta}$ the distribution of the observation (X, Y, Z),

$$P_{\theta,\eta}(g_1 \oplus g_2 \oplus c) B_{\theta,\eta}h = P_{\theta,\eta} g_1 C_{\theta,\eta}h + \eta g_2 h + 2\eta h c = \eta(C^*_{\theta,\eta} g_1 + g_2 + 2c)h.$$

Thus, the adjoint $B^*_{\theta,\eta} : G \mapsto L_2(\eta)$ of the operator $B_{\theta,\eta} : L_2(\eta) \mapsto G$ is given by

$$B^*_{\theta,\eta}(g_1 \oplus g_2 \oplus c) = C^*_{\theta,\eta} g_1 + g_2 + 2c.$$

Consequently, on the set of mean-zero functions in $L_2(\eta)$ we have the identity $B^*_{\theta,\eta} B_{\theta,\eta} = C^*_{\theta,\eta} C_{\theta,\eta} + I$. Because the operator $C^*_{\theta,\eta} C_{\theta,\eta}$ is nonnegative definite, the operator $B^*_{\theta,\eta} B_{\theta,\eta}$ is strictly positive definite and hence continuously invertible as an operator of $L_2(\eta)$ into itself. The following lemma gives a condition for continuous invertibility as an operator on the space $C^\alpha(\mathcal{Z})$ of all "α-smooth functions." For $\alpha_0 \le \alpha$ the smallest integer strictly smaller than α, these consist of the functions $h : \mathcal{Z} \subset \mathbb{R}^d \mapsto \mathbb{R}$ whose partial derivatives up to order α_0 exist and are bounded and whose α_0-order partial derivatives are Lipschitz of order $\alpha - \alpha_0$. These are Banach spaces relative to the norm, with D^k a differential operator $\partial^{k_1} \cdots \partial^{k^d} / \partial z_1^{k_1} \cdots z_k^{k_d}$,

$$\|h\|_\alpha = \max_{|k| < \alpha} \sup_{z \in \mathcal{Z}} |D^k h(z)| \vee \max_{|k| = \alpha_0} \sup_{z_1 \ne z_2 \in \mathcal{Z}} \frac{\big|D^k(z_1) - D^k(z_2)\big|}{\|z_1 - z_2\|^{\alpha - \alpha_0}}.$$

The unit ball of one of these spaces is a good choice for the set $\mathcal{H}$ indexing the likelihood equations if the maps $z \mapsto p_{\theta_0}(x \mid z)$ are sufficiently smooth.

25.96 ***Lemma.*** *Let $\mathcal{Z}$ be a bounded, convex subset of $\mathbb{R}^d$ and assume that the maps $z \mapsto p_0(x \mid z)$ are continuously differentiable for each x with partial derivatives $\partial/\partial z_i\, p_{\theta_0}(x \mid z)$*

satisfying, for all z, z' in $\mathcal{Z}$ and fixed constants K and $\alpha > 0$,

$$\int \left| \frac{\partial}{\partial z_i} p_0(x \mid z) - \frac{\partial}{\partial z_i} p_0(x \mid z') \right| d\mu(x) \le K \|z - z'\|^\alpha,$$

$$\int \left| \frac{\partial}{\partial z_i} p_0(x \mid z) \right| d\mu(x) \le K.$$

Then $B^*_{\theta_0,\eta_0} B_{\theta_0,\eta_0} : C^\beta(\mathcal{Z}) \mapsto C^\beta(\mathcal{Z})$ is continuously invertible for every $\beta < \alpha$.

Proof. By its strict positive-definiteness in the Hilbert-space sense, the operator $B_0^* B_0 : \ell^\infty(\mathcal{Z}) \mapsto \ell^\infty(\mathcal{Z})$ is certainly one-to-one in that $B_0^* B_0 h = 0$ implies that $h = 0$ almost surely under η_0. On reinserting this we find that $-h = C_0^* C_0 h = C_0^* 0 = 0$ everywhere. Thus $B_0^* B_0$ is also one-to-one in a pointwise sense. If it can be shown that $C_0^* C_0 : C^\beta(\mathcal{Z}) \mapsto C^\beta(\mathcal{Z})$ is compact, then $B_0^* B_0$ is onto and continuously invertible, by Lemma 25.93.

It follows from the Lipschitz condition on the partial derivatives that $C_0^* h(z)$ is differentiable for every bounded function $h : \mathcal{X} \mapsto \mathbb{R}$ and its partial derivatives can be found by differentiating under the integral sign:

$$\frac{\partial}{\partial z_i} C_0^* h(z) = \int h(x) \frac{\partial}{\partial z_i} p_0(x \mid z) \, d\mu(x).$$

The two conditions of the lemma imply that this function has Lipschitz norm of order α bounded by $K \|h\|_\infty$. Let h_n be a uniformly bounded sequence in $\ell^\infty(\mathcal{X})$. Then the partial derivatives of the sequence $C_0^* h_n$ are uniformly bounded and have uniformly bounded Lipschitz norms of order α. Because $\mathcal{Z}$ is totally bounded, it follows by a strengthening of the Arzela-Ascoli theorem that the sequences of partial derivatives are precompact with respect to the Lipschitz norm of order β for every $\beta < \alpha$. Thus there exists a subsequence along which the partial derivatives converge in the Lipschitz norm of order β. By the Arzela-Ascoli theorem there exists a further subsequence such that the functions $C_0^* h_n(z)$ converge uniformly to a limit. If both a sequence of functions itself and their continuous partial derivatives converge uniformly to limits, then the limit of the functions must have the limits of the sequences of partial derivatives as its partial derivatives. We conclude that $C_0^* h_n$ converges in the $\|\cdot\|_{1+\beta}$-norm, whence $C_0^* : \ell^\infty(\mathcal{X}) \mapsto C^\beta(\mathcal{Z})$ is compact. Then the operator $C_0^* C_0$ is certainly compact as an operator from $C^\beta(\mathcal{Z})$ into itself. ■

Because the efficient information for θ is bounded below by the information for θ in a "good" observation (Y, Z), it is typically positive. Then the preceding lemma together with Lemma 25.92 shows that the derivative $\dot{\Psi}_0$ is continuously invertible as a map from $\mathbb{R}^k \times \ell^\infty(\mathcal{H}) \times \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ for $\mathcal{H}$ the unit ball of $C^\beta(\mathcal{Z})$. This is useful in the cases that the dimension of $\mathcal{Z}$ is not bigger than 3, for, in view of Example 19.9, we must have that $\beta > d/2$ in order that the functions $B_{\theta,\eta} h = C_{\theta,\eta} h \oplus h$ form a Donsker class, as required by Theorem 25.90. Thus $\alpha > 1/2, 2, 3/2$ suffice in dimensions 1, 2, 3, but we need $\beta > 2$ if $\mathcal{Z}$ is of dimension 4.

Sets $\mathcal{Z}$ of higher dimension can be treated by extending Lemma 25.96 to take into account higher-order derivatives, or alternatively, by not using a $C^\alpha(\mathcal{Z})$-unit ball for $\mathcal{H}$. The general requirements for a class $\mathcal{H}$ that is the unit ball of a Banach space $\mathbb{B}$ are that $\mathcal{H}$ is η_0-Donsker, that $C_0^* C_0 \mathbb{B} \subset \mathbb{B}$, and that $C_0^* C_0 : \mathbb{B} \mapsto \mathbb{B}$ is compact. For instance, if $p_\theta(x \mid z)$ corresponds to a linear regression on z, then the functions $z \mapsto C_0^* C_0 h(z)$ are of the form $z \mapsto g(\alpha^T z)$

for functions g with a one-dimensional domain. Then the dimensionality of $\mathcal{Z}$ does not really play an important role, and we can apply similar arguments, under weaker conditions than required by treating $\mathcal{Z}$ as general higher dimensional, with, for instance, $\mathbb{B}$ equal to the Banach space consisting of the linear span of the functions $z \mapsto g(\alpha^T z)$ in $C_1^1(\mathcal{Z})$ and $\mathcal{H}$ its unit ball.

The second main condition of Theorem 25.92 is that the functions $i_{\theta,\eta}$ and $B_{\theta,\eta}h$ form a Donsker class. Dependent on the kernel $p_\theta(x \mid z)$, a variety of methods may be used to verify this condition. One possibility is to employ smoothness of the kernel in x in combination with Example 19.9. If the map $x \mapsto p_\theta(x \mid z)$ is appropriately smooth, then so is the map $x \mapsto C_{\theta,\eta}h(x)$. Straightforward differentiation yields

$$\frac{\partial}{\partial x_i} C_{\theta,\eta}h(x) = \operatorname{cov}_x\left(h(Z), \frac{\partial}{\partial x_i} \log p_\theta(x \mid Z)\right),$$

where for each x the covariance is computed for the random variable Z having the (conditional) density $z \mapsto p_\theta(x \mid z)\, d\eta(z)/p_\theta(x \mid \eta)$. Thus, for a given bounded function h,

$$\left|\frac{\partial}{\partial x_i} C_{\theta,\eta}h(x)\right| \le \|h\|_\infty \frac{\int \left|\frac{\partial}{\partial x_i} \log p_\theta(x \mid z)\right| p_\theta(x \mid z)\, d\eta(z)}{\int p_\theta(x \mid z)\, d\eta(z)}.$$

Depending on the function $\partial/\partial x_i \log p_\theta(x \mid z)$, this leads to a bound on the first derivative of the function $x \mapsto C_{\theta,\eta}h(x)$. If $\mathcal{X}$ is an interval in $\mathbb{R}$, then this is sufficient for applicability of Example 19.9. If $\mathcal{X}$ is higher dimensional, the we can bound higher-order partial derivatives in a similar manner.

If the main interest is in the estimation of η rather than θ, then there is also a nontechnical criterion for the choice of $\mathcal{H}$, because the final result gives the asymptotic distribution of $\hat{\eta}h$ for every $h \in \mathcal{H}$, but not necessarily for $h \notin \mathcal{H}$. Typically, a particular h of interest can be added to a set $\mathcal{H}$ that is chosen for technical reasons without violating the results as given previously. The addition of an infinite set would require additional arguments. Reference [107] gives more details concerning this example.

Notes

Most of the results in this chapter were obtained during the past 15 years, and the area is still in development. The monograph by Bickel, Klaassen, Ritov, and Wellner [8] gives many detailed information calculations, and heuristic discussions of methods to construct estimators. See [77], [101], [102], [113], [122], [145] for a number of other, also more recent, papers. For many applications in survival analysis, counting processes offer a flexible modeling tool, as shown in Andersen, Borgan, Gill, and Keiding [1], who also treat semiparametric models for survival analysis. The treatment of maximum likelihood estimators is motivated by (partially unpublished) joint work with Susan Murphy. Apparently, the present treatment of the Cox model is novel, although proofs using the profile likelihood function and martingales go back at least 15 years. In connection with estimating equations and CAR models we profited from discussions with James Robins, the representation in section 25.53 going back to [129]. The use of the empirical likelihood goes back a long way, in particular in survival analysis. More recently it has gained popularity as a basis for constructing likelihood ratio based confidence intervals. Limitations of the information

bounds and the type of asymptotics discussed in this chapter are pointed out in [128]. For further information concerning this chapter consult recent journals, both in statistics and econometrics.

PROBLEMS

1. Suppose that the underlying distribution of a random sample of real-valued observations is known to have mean zero but is otherwise unknown.
   (i) Derive a tangent set for the model.
   (ii) Find the efficient influence function for estimating $\psi(P) = P(C)$ for a fixed set C.
   (iii) Find an asymptotically efficient sequence of estimators for $\psi(P)$.
2. Suppose that the model consists of densities $p(x-\theta)$ on $\mathbb{R}^k$, where p is a smooth density with $p(x) = p(-x)$. Find the efficient influence function for estimating θ.
3. In the regression model of Example 25.28, assume in addition that e and X are independent. Find the efficient score function for θ.
4. Find a tangent set for the set of mixture distributions $\int p(x \mid z)\, dF(z)$ for $x \mapsto p(x \mid z)$ the uniform distribution on $[z, z+1]$. Is the linear span of this set equal to the nonparametric tangent set?
5. **(Neyman-Scott problem)** Suppose that a typical observation is a pair (X, Y) of variables that are conditionally independent and $N(Z, \theta)$-distributed given an unobservable variable Z with a completely unknown distribution η on $\mathbb{R}$. A natural approach to estimating θ is to "eliminate" the unobservable Z by taking the difference $X - Y$. The maximum likelihood estimator based on a sample of such differences is $T_n = \frac{1}{2} n^{-1} \sum_{i=1}^n (X_i - Y_i)^2$.
   (i) Show that the closed linear span of the tangent set for η contains all square-integrable, mean-zero functions of $X + Y$.
   (ii) Show that T_n is asymptotically efficient.
   (iii) Is T_n equal to the semiparametric maximum likelihood estimator?
6. In Example 25.72, calculate the score operator and the information operator for η.
7. In Example 25.12, express the density of an observation X in the marginal distributions F and G of Y and C and
   (i) Calculate the score operators for F and G.
   (ii) Show that the empirical distribution functions $\hat{F}^*$ and $\hat{G}^*$ of the Y_i and C_j are asymptotically efficient for estimating the marginal distributions F^* and G^* of Y and C, respectively;
   (iii) Prove the asymptotic normality of the estimator for F given by

$$\hat{F}(y) = 1 - \prod_{0 \le s \le y} \bigl(1 - \hat{\Lambda}\{s\}\bigr), \qquad \hat{\Lambda}(y) = \int_{[0,y]} \frac{d\hat{F}^*}{\hat{G}^* - \hat{F}^*};$$

   (iv) Show that this estimator is asymptotically efficient.
8. **(Star-shaped distributions)** Let $\mathcal{F}$ be the collection of all cumulative distribution functions on $[0, 1]$ such that $x \mapsto F(x)/x$ is nondecreasing. (This is a famous example in which the maximum likelihood estimator is inconsistent.)
   (i) Show that there exists a maximizer $\hat{F}_n$ (over $\mathcal{F}$) of the empirical likelihood $F \mapsto \prod_{i=1}^n F\{x_i\}$, and show that this satisfies $\hat{F}_n(x) \to xF(x)$ for every x.
   (ii) Show that at every $F \in \mathcal{F}$ there is a convex tangent cone whose closed linear span is the nonparametric tangent space. What does this mean for efficient estimation of F?

9. Show that a U-statistic is an asymptotically efficient estimator for its expectation if the model is nonparametric.

10. Suppose that the model consists of all probability distributions on the real line that are symmetric.
(i) If the symmetry point is known to be 0, find the maximum likelihood estimator relative to the empirical likelihood.
(ii) If the symmetry point is unknown, characterize the maximum likelihood estimators relative to the empirical likelihood; are they useful?

11. Find the profile likelihood function for the parameter θ in the Cox model with censoring discussed in Section 25.12.1.

12. Let $\mathcal{P}$ be the set of all probability distributions on $\mathbb{R}$ with a positive density and let $\psi(P)$ be the median of P.
(i) Find the influence function of ψ.
(ii) Prove that the sample median is asymptotically efficient.