Semiparametric Statistics

by

A.W. van der Vaart

Vrije Universiteit Amsterdam

Contents

Preface 4
Notation 5
1. Introduction, Tangent Sets 6
1.1. Introduction 6
1.2. Tangent Spaces and Information 9
2. Lower Bounds 15
2.1. Lower Bounds 15
2.2. Efficient Score Functions 22
3. Calculus of Scores 26
3.1. Score and Information Operators 26
3.2. Semiparametric Models 33
4. Gaussian Approximations 38
4.1. Contiguity 38
4.2. Gaussian Representations 41
5. Empirical Processes and Consistency of Z-Estimators 51
5.1. Empirical Measures and Entropy Numbers 51
5.2. Glivenko-Cantelli Classes 52
5.3. Consistency of M- and Z-estimators 54
5.4. Nuisance Parameters 61
6. Empirical Processes and Normality of Z-Estimators 63
6.1. Weak Convergence in Metric Spaces 63
6.2. Donsker Classes 67
6.3. Maximal Inequalities 69
6.4. Random Functions 70
6.5. Asymptotic Normality of Z-Estimators 73
6.6. Nuisance parameters 74
7. Efficient Score and One-step Estimators 79
7.1. Efficient Score Estimators 79
7.2. One-step Estimators 80
7.3. Symmetric location 85
7.4. Errors-in-Variables 86
8. Rates of Convergence 89

8.1. A General Result 89
8.2. Nuisance Parameters 92
8.3. Cox Regression with Current Status Data 93
9. Maximum and Profile Likelihood 98
9.1. Examples 98
9.2. Asymptotic Normality 101
9.3. Cox Regression with Current Status Data 103
9.4. Profile Likelihood 106
10. Infinite-dimensional Z-Estimators 110
10.1. General Result 110
10.2. Maximum Likelihood 112
References 118

Preface

These notes provide extended versions of my lectures in the St Flour meeting of 1999. The general subject are semiparametric models for replicated experiments, in particular the theory for functionals that are estimable at the rate equal to the square root of the number of replications. We discuss bounds on the efficiency of estimators and tests, and methods of constructing efficient or inefficient estimators and tests, with particular attention for maximum likelihood estimators. Furthermore, we discuss abstract empirical processes, which play an important role in the analysis of the estimators.

The ten lectures have a certain overlap with material earlier published in the books [41] and [42]. A number of proofs have been omitted, because they can be found in these works. On the other hand, these notes are an attempt to give a consistent and reasonably self-contained overview of (a part of) semiparametric statistics, including digressions into empirical process theory, new examples, and a number of more recent developments.

This area is certainly not complete. To illustrate this point, scattered through the text we pose some problems whose solutions are presently unknown (to me).

Our list of references is restricted to the references that are directly relevant to the lectures. In beginning 2000 the Mathematical Reviews gave 415 responses to a query on semiparametric models, so our list does not do justice to the great amount of work having been done. A general work covering the subject of semiparametric models, but from a somewhat different point of view with relatively little attention for the subject of Lectures 5–10, is the book [3] by Bickel, Klaassen, Ritov and Wellner. This book also has an extensive list of references.

Notation

We use the wiggly arrow $\rightsquigarrow$ for weak convergence, also for nonmeasurable maps: if X_n and X are maps defined on some probability spaces $(\Omega_n, \mathcal{U}_n, \mathrm{P}_n)$ with values in a metric space $\mathbb{D}$, then we say that $X_n \rightsquigarrow X$ if $\mathrm{E}^* f(X_n) \to \mathrm{E} f(X)$ for all bounded, continuous functions $f: \mathbb{D} \mapsto \mathbb{R}$. Here the limit X is always assumed Borel measurable, but the X_n may be arbitrary maps. The $*$ in $\mathrm{E}^* f(X_n)$ is for *outer expectation* on $(\Omega_n, \mathcal{U}_n, \mathrm{P}_n)$.

Given a measure space $(\mathcal{X}, \mathcal{A}, P)$ the set $L_r(P)$ (for $r \geq 1$) is the collection of all measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$ with $\|f\|_{P,r}^r := \int |f|^r \, dP < \infty$.

The wiggly inequality $\lesssim$ means "less than equal up to a constant". The range and kernel of an operator A are denoted by $\mathrm{R}(A)$ and $\mathrm{N}(A)$. The space of all bounded functions $z: T \mapsto \mathbb{R}$ on a set T is denoted by $\ell^\infty(T)$ and $\|z\|_T$ is the uniform norm. The set $UC(T, \rho)$ is the set of all ρ-uniformly continuous functions on T.

Lecture 1
Introduction, Tangent Sets

In this lecture we introduce basic notation, give a number of examples of semiparametric models, and define the tangent set of a model.

1.1 Introduction

Throughout the presentation of the general theory we denote by $X_1, \ldots, X_n$ the observations. These are measurable maps on some underlying probability space that we usually need not further specify, and take values in a measurable space $(\mathcal{X}, \mathcal{A})$. The observations are independent and identically distributed (i.i.d.), with a distribution P on $(\mathcal{X}, \mathcal{A})$. A *model* $\mathcal{P}$ is a collection of probability measures on the sample space, to be considered the set of all possible values of P.

A *semiparametric model* is one that is neither a parametric model nor a nonparametric model. This definition is not informative, but could be saved by giving precise definitions of parametric and nonparametric models. The *nonparametric model* $\mathcal{P}$ is the set of all probability distributions on $\mathcal{P}$. A *parametric model* is a model that can be smoothly indexed by a Euclidean vector ("the parameter"). We shall not attempt to make this definition more precise by specifying "smoothly", but note that this should cover all classical statistical models, including exponential families and the uniform distributions. The concept of a "nonparametric model" is often also used in a more vague sense of a model that does not essentially restrict the elements $P \in \mathcal{P}$. A model in which all P are assumed to have a second moment or a smooth density relative to Lebesgue measure is then also considered to be "nonparametric".

Thus the "definition" says that a semiparametric model is an infinite-dimensional model that is essentially smaller than the set of all possible distributions. Even this vague description is not universally accepted. For instance: the nonparametric model is often considered to be semiparametric if it is parametrized in an interesting way.

A few examples will give a better idea.

1.1 Example (Symmetric location). For a given $\theta \in \mathbb{R}$ and a probability density η on $\mathbb{R}$ that is symmetric about 0, let $P_{\theta,\eta}$ be the measure with density $x \mapsto \eta(x-\theta)$.

Then consider the semiparametric model $\mathcal{P}$ consisting of all measures $P_{\theta,\eta}$ when θ ranges over $\mathbb{R}$ and η ranges over all Lebesgue densities that are absolutely continuous with finite Fisher information for location: $I(\eta) := \int (\eta'/\eta)^2 \eta \, d\lambda < \infty$. This model arose naturally in the study of nonparametric testing theory (e.g. rank tests) and was studied long before the general subject of semiparametric models had been conceived. It turns out to be a very special model as regards the estimation of the center of symmetry θ. As we shall see there exist estimators for θ in this model (which cannot use the form of the unknown η) that are (asymptotically) of the same quality as the best estimators specially designed to work for a particular η (for instance as good as the sample mean in the case of normal η and as good as the median for Laplace η). □

1.2 Example (Partially linear regression). A classical regression model specifies that the conditional mean of a "response variable" Y given a covariate V is a linear function $\theta^T V$ of the covariate, or a fixed transformation $\Psi(\theta^T V)$ of it. A nonparametric regression model would replace the linear function by an arbitrary function, perhaps restricted by being "smooth". A typical semiparametric model would mix these two extremes, for instance by specifying that the conditional mean is of the form $\Psi(\theta^T V + \eta(W))$ for $\theta \in \mathbb{R}^d$ and η ranging over the class of all twice differentiable functions on the domain of W, and a fixed function Ψ.

To describe the full model we could specify that the observation is $X = (Y, V, W)$ and that (V, W) has an arbitrary distribution. Next there are several possibilities to complete the description by specifying the form of the conditional distribution of Y given (V, W). One possibility is to specify only that $\mathrm{E}(Y \,|\, V, W) = \theta^T V + \eta(W)$. This type of model is popular among econometricians. A smaller model is obtained by postulating that $Y = \theta^T V + \eta(W) + e$ for e independent of (V, W) and of mean zero, leaving the rest of the distribution of e unspecified, assuming it to be normal or symmetric. Third, we can also create semiparametric versions of the generalized linear model. For instance, the response Y could be a 0-1 variable and we could assume that $\mathrm{P}(Y = 1 \,|\, V, W)$ is of the form $\Psi(\theta^T V + \eta(W))$ for Ψ the logistic distribution function. □

1.3 Example (Cox). In the Cox model a typical observation is a pair $X = (T, Z)$ of a "survival time" T and a covariate Z. It is best described in terms of the conditional hazard function of T given Z.

Recall that the *hazard function* λ corresponding to a probability density f is the function $\lambda = f/(1 - F)$, for F the distribution function corresponding to f. Simple algebra shows that $1 - F = e^{-\Lambda}$ and hence $f = \lambda e^{-\Lambda}$, so that the relationship between f and λ is on-to-one.

In the Cox model the distribution of Z is arbitrary and the conditional hazard function of T given Z is postulated to be of the form $e^{\theta^T Z}\lambda(t)$ for $\theta \in \mathbb{R}^d$ and 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} is the ratio of the hazards of two individuals whose covariates are $Z_i = 1$ and $Z_i = 0$, respectively, and whose covariates are identical otherwise. This is one reason for the popularity of the model: the model gives a better fit to data than a parametric model (obtained for instance by assuming that the baseline hazard function is of Weibull form), but its parameters are still

easy to interpret. A second reason for its popularity is that statistical procedures for estimating the parameters take a simple form. They were originally found and motivated by ad-hoc arguments. We shall use the model throughout these lectures as an illustration and show how the standard estimators can be derived and analysed by principles that apply equally well to other semiparametric models. □

1.4 Example (Mixture models). Suppose that $x \mapsto p_\theta(x|\, z)$ is a probability density for every pair $(\theta, z) \in \Theta \times \mathcal{Z}$ for a subset Θ of a Euclidean space and a measurable space $(\mathcal{Z}, \mathcal{C})$. If the map $(x, z) \mapsto p_\theta(x|\, z)$ is jointly measurable, then

$$p_{\theta,\eta}(x) = \int p_\theta(x|\, z)\, d\eta(z)$$

defines a probability density for every probability measure η on $(\mathcal{Z}, \mathcal{C})$. This mixture density reduces to the density $p_\theta(\cdot|\, z)$ when η is degenerate at z. Hence the model consisting of all mixture densities of this type is considerably bigger than the "original model", which is parametric if z is Euclidean and the map $(\theta, z) \mapsto p_\theta(\cdot|\, z)$ is smooth.

A concrete example of a mixture model is the errors-in-variables model, which is most easily described structurally, as follows. The observation is a pair $X = (X_1, X_2)$, where $X_1 = Z + e$ and $X_2 = g_\theta(Z) + f$ for a bivariate normal vector (e, f) with mean zero and unknown covariance matrix, and a function g_θ that is known up to a parameter θ. Thus X_2 is a (possibly nonlinear) regression on a variable Z that is observed with error. The distribution of Z is unknown. The kernel $p_\theta(\cdot|\, z)$ is in this case a multivariate Gaussian density.

A particular example is the linear errors-in-variables model, for which $\theta = (\alpha, \beta)$ and $g_\theta(z) = \alpha + \beta z$. This linear model has been studied before the 1980s, but not from a semiparametric perspective. Semiparametric theory has led to new, more efficient estimators of the regression parameters. Surprisingly, for most of the nonlinear cases good estimators for θ are still unknown, and in fact it is unknown if the parameter θ is estimable at $\sqrt{n}$ rate in general. (See [35] and work in progress by the same author.) □

1.5 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. Thus, a typical observation X is a pair of a survival time and a 0-1 variable, and is distributed as $(T \wedge C, 1\{T \leq C\})$. If the distributions of T and C are allowed to range over all distributions on $[0, \infty]$, then the distribution of X can be shown to take an arbitrary form on the sample space $\mathcal{X} = [0, \infty) \times \{0, 1\}$. Therefore, this model is a nonparametric example. Because the interest is usually in the distribution of T, which is a complicated function of the distribution of X to which much of the semiparametric machinery applies, the model is usually also considered semiparametric. □

Our lectures aim at developing theory for the estimation and testing of functionals $\psi: \mathcal{P} \mapsto \mathbb{B}$ defined on a model $\mathcal{P}$ and taking values in some Banach space $\mathbb{B}$ (most often $\mathbb{R}^d$). An important examples is the functional $\psi(P_{\theta,\eta}) = \theta$ if the model $\mathcal{P} = \{P_{\theta,\eta}: \theta \in \Theta, \eta \in H\}$ is indexed by two "parameters" θ and η. In this case,

because apparently the prime interest is in θ, we refer to η as a "nuisance parameter". This will not stop us from also considering the estimation of η. Models with a partitioned parameter (θ, η), with θ finite-dimensional, are semiparametric models in a strict sense. Begun, Hall, Huang and Wellner in [1] called them parametric-nonparametric, having in mind that η would be an element of a nonparametric model.

Our main interest in these lectures is in functionals ψ that allow an asymptotic theory analogous to the theory for smooth parametric models. This comprises the asymptotic normality of the maximum likelihood estimator, rooted in the work by Fisher in the 1920s, the asymptotic chisquare distribution of the likelihood ratio statistic, rooted in the work by Wilks in 1930s, and the lower bound theory rooted in the work by Cramér and Rao in the 1940s. The dates might suggest that we are only setting out to a simple extension of "classical theory" of the first half of the 20th century. There is some truth to this, but as we shall see, apart from necessitating more mathematical sophistication (which word we mean to use in a positive sense), the theory of semiparametric models turns out to be much richer than the classical theory.

Unfortunately, not all problems have been solved. This is true for the problems in the restricted realm of the preceding paragraph. It is even more true for the general theory of semiparametric models, which also contains many so-called inverse problems. In later lectures we shall indicate some of the important open questions, to be solved in the next millennium.

In the following section we start by developing a notion of "information" for estimating $\psi(P)$ given the model $\mathcal{P}$, which extends the notion of Fisher information for parametric models.

1.2 Tangent Spaces and Information

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 favourable* 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 an appropriate way.

1.6 Definition. A *differentiable path* is a map $t \mapsto P_t$ from a neighbourhood of $0 \in [0, \infty)$ to $\mathcal{P}$ such that, for some measurable function $g: \mathcal{X} \mapsto \mathbb{R}$,

$$\int \Big[\frac{dP_t^{1/2} - dP^{1/2}}{t} - \tfrac{1}{2} g \, dP^{1/2} \Big]^2 \to 0. \tag{1.7}$$

The function g is called the *score function* of the submodel $\{P_t: t \geq 0\}$ at $t = 0$.

The notation in the preceding display is due to Le Cam. The objects $dP_t^{1/2}$ can be formalized by introducing an Hilbert space of "square roots of measures". Simpler and sufficient for our purposes is to read the display as

$$\int \Big[\frac{p_{tt}^{1/2} - p_t^{1/2}}{t} - \tfrac{1}{2} g \, p_t^{1/2} \Big]^2 d\mu_t \to 0,$$

where, for each t, μ_t is an arbitrary measure relative to which P and P_t possess densities p_t and p_{tt}. For instance, the measure $\mu_t = P_t + P$, or a fixed σ-finite dominating measure for $\mathcal{P}$ if it exists. The value of the integral does not depend on the choice of μ_t.

In words we say that a differentiable path is a parametric submodel $\{P_t: 0 \leq t < \varepsilon\}$ that 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 $\dot{\mathcal{P}}_P$.

1.8 Lemma. *Every score function satisfies $Pg = 0$ and $Pg^2 < \infty$.*

Proof. For given, arbitrary $t_n \downarrow 0$, let p_n and p be densities of P_{t_n} and P relative to a σ-finite dominating measure μ, for instance a convex combination of the countably many measures $P_{t_n} + P$. By (1.7) the sequence $(\sqrt{p_n} - \sqrt{p})/t_n$ converges in quadratic mean (i.e. in $L_2(\mu)$) to $\frac{1}{2} g \sqrt{p}$. This implies immediately that $g \in L_2(P)$. Furthermore, it implies that the sequence $\sqrt{p_n}$ converges in quadratic mean to $\sqrt{p}$. By the continuity of the inner product,

$$Pg = \int \tfrac{1}{2} g \sqrt{p} \, 2\sqrt{p} \, d\mu = \lim \int \frac{\big(\sqrt{p_n} - \sqrt{p}\big)}{t_n} \big(\sqrt{p_n} + \sqrt{p}\big) \, d\mu_n.$$

The right side equals $(1 - 1)/t_n = 0$ for every n, because both probability densities integrate to 1. Thus $Pg = 0$. ∎

It follows that a tangent set 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 a Hilbert space (the space $L_2(\mu)$ if the model is dominated), and $\dot{\mathcal{P}}_P$, or rather the set of all objects $\frac{1}{2} g \, dP^{1/2}$, as its tangent set. Note however that we have not defined a tangent set to be equal to the set of all score functions g that correspond to some differentiable submodel. For many purposes this "maximal tangent set" is too big, so that we have given ourselves the flexibility of calling any set of score

functions a tangent set. The drawback will be that in any result obtained later on we must specify which tangent set we are working with.

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

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

This pointwise differentiability is not required by (1.7). Conversely, given this pointwise differentiability we still need to be able to apply a convergence theorem for integrals to obtain (1.7). The following lemma solves most examples.

1.9 Lemma. *If p_t is a probability density relative to a fixed measure μ and $t \mapsto \sqrt{p_t(x)}$ is continuously differentiable in a neighbourhood of 0 and $t \mapsto \int \dot{p}_t^2/p_t \, d\mu$ is finite and continuous in this neighbourhood, then $t \mapsto P_t$ is a differentiable path.*

The differentiability (1.7) is the correct definition for defining information, because it ensures a type of local asymptotic normality, as shown by the following lemma.

1.10 Lemma. *If the path $t \mapsto P_t$ in $\mathcal{P}$ satisfies (1.7), then*

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

Proof. We adopt the notation of the preceding proof, but with $t_n = 1/\sqrt{n}$. The random variable $W_{ni} = 2\big[\sqrt{p_n/p}(X_i) - 1\big]$ is with P-probability 1 well-defined. By (1.7)

$$\operatorname{var}\Big(\sum_{i=1}^n W_{ni} - \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i)\Big) \le \mathrm{E}\big(\sqrt{n} W_{ni} - g(X_i)\big)^2 \to 0, \tag{1.11}$$

$$\mathrm{E} \sum_{i=1}^n W_{ni} = 2n\Big(\int \sqrt{p_n}\sqrt{p}\, d\mu - 1\Big) = -n \int \big[\sqrt{p_n} - \sqrt{p}\big]^2 d\mu \to -\tfrac{1}{4} P g^2.$$

Therefore, combining the preceding pair of displayed equations, we find

$$\sum_{i=1}^n W_{ni} = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(X_i) - \tfrac{1}{4} P g^2 + o_P(1). \tag{1.12}$$

Next, we express the log likelihood ratio in $\sum_{i=1}^n W_{ni}$ through a Taylor expansion of the logarithm. If we write $\log(1+x) = x - \frac{1}{2}x^2 + x^2 R(2x)$, then $R(x) \to 0$ as $x \to 0$, and

$$\begin{aligned} \log \prod_{i=1}^n \frac{p_n}{p}(X_i) &= 2 \sum_{i=1}^n \log\big(1 + \tfrac{1}{2} W_{ni}\big) \\ &= \sum_{i=1}^n W_{ni} - \tfrac{1}{4} \sum_{i=1}^n W_{ni}^2 + \tfrac{1}{2} \sum_{i=1}^n W_{ni}^2 R(W_{ni}). \end{aligned} \tag{1.13}$$

As a consequence of the right side of (1.11), it is possible to write $nW_{ni}^2 = g^2(X_i) + A_{ni}$ for random variables A_{ni} such that $\mathrm{E}|A_{ni}| \to 0$. The averages $\overline{A}_n$ converge in mean and hence in probability to zero. Combination with the law of large numbers yields

$$\sum_{i=1}^n W_{ni}^2 = \overline{(g^2)}_n + \overline{A}_n \xrightarrow{P} Pg^2.$$

By the triangle inequality followed by Markov's inequality,

$$\begin{aligned} n\mathrm{P}\big(|W_{ni}| > \varepsilon\sqrt{2}\big) &\le n\mathrm{P}\big(g^2(X_i) > n\varepsilon^2\big) + n\mathrm{P}\big(|A_{ni}| > n\varepsilon^2\big) \\ &\le \varepsilon^{-2} Pg^2\{g^2 > n\varepsilon^2\} + \varepsilon^{-2}\mathrm{E}|A_{ni}| \to 0. \end{aligned}$$

The left side is an upper bound for $\mathrm{P}\big(\max_{1\le i\le n}|W_{ni}| > \varepsilon\sqrt{2}\big)$. Thus the sequence $\max_{1\le i\le n}|W_{ni}|$ converges to zero in probability. By the property of the function R, the sequence $\max_{1\le i\le n}\big|R(W_{ni})\big|$ converges in probability to zero as well. The last term on the right in (1.13) is bounded by $\max_{1\le i\le n}\big|R(W_{ni})\big|\sum_{i=1}^n W_{ni}^2$. Thus it is $o_P(1)O_P(1)$, and converges in probability to zero. Combine to obtain that

$$\log\prod_{i=1}^n \frac{p_n}{p}(X_i) = \sum_{i=1}^n W_{ni} - \tfrac{1}{4}Pg^2 + o_P(1).$$

Together with (1.12) this yields the theorem. ∎

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. A minimal requirement is that the map $t \mapsto \psi(P_t)$ be differentiable at $t = 0$, but we need more.

1.14 Definition. A map $\psi\colon \mathcal{P} \mapsto \mathbb{B}$ is *differentiable* at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a continuous linear map $\dot{\psi}_P\colon L_2(P) \mapsto \mathbb{B}$ 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 definition 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.

In the case that $\mathbb{B} = \mathbb{R}^k$ the Riesz representation theorem for Hilbert spaces allows us to write the derivative map $\dot{\psi}_P$ in the form of an inner product. Precisely, there exists a fixed vector-valued, measurable function $\tilde{\psi}_P\colon \mathcal{X} \mapsto \mathbb{R}^k$,

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

The function $\tilde{\psi}_P$ is not uniquely defined by the functional ψ and the model $\mathcal{P}$, since 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 $\overline{\mathrm{lin}}\,\dot{\mathcal{P}}_P$, 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. Here an *influence function* will be any measurable function $\dot\psi_P\colon \mathcal{X} \mapsto \mathbb{R}$ whose projection on $\overline{\text{lin}}\,\dot{\mathcal{P}}_P$ is the efficient influence function.

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. According to our definition 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 loss of generality to assume that the tangent set we work with is a cone as well.

1.15 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 $h \to 0$,

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

Then a tangent set at P_θ is given by the linear space $\big\{h^T \dot\ell_\theta\colon h \in \mathbb{R}^m\big\}$ 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 $\chi\colon \Theta \mapsto \mathbb{R}^k$ 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. □

1.16 Example (Nonparametric model). 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 $\int g^2\, dP < \infty$. Since 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 $p_t(x) = c(t)\, \exp(tg(x))\, p_0(x)$ 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 1.9, we see that both models also have score function g at $t = 0$ in the L_2-sense (1.7). 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. □

1.17 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, with $x = (t, z)$,

$$\dot{\ell}_{\theta,\Lambda}(x) = z - ze^{\theta^T z}\Lambda(t).$$

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

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

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. □

Notes

Tangent spaces of statistical models as presented here were popularized as a general theory by Pfanzagl in [28], except that Pfanzagl initially did not define differentiable paths through root-densities, which is an idea going back to Le Cam in the 1960s (see [15], [16], [18]). The study of tangent spaces and information in infinite-dimensional models goes further back to Levit and Koshevnik and Levit (see [20] and [19]) in the mid 1970s, who however considered mostly nonparametric models.

We are going to use the Cox model as an illustration throughout the ten lectures. Cox introduced it in [7] and discussed the partial likelihood methods of estimation in [8].

Lecture 2
Lower Bounds

In this lecture we state a number of theorems giving lower bounds on the asymptotic performance of estimators and tests, and make these concrete for the estimation of a parameter θ in a strict semiparametric model. Some of the proofs are deferred to Lecture 4.

2.1 Lower Bounds

A "lower bound theorem" in statistics is an assertion that something, estimation or testing, cannot be done better than in some way. The best known bound is the Cramér-Rao bound for the case of independent sampling from a parametric model $\{P_\theta : \theta \in \Theta \subset \mathbb{R}\}$, which is taught in most introductory statistics courses.

2.1 Fact. *If $\theta \mapsto P_\theta$ is differentiable at θ with score function $\dot{\ell}_\theta$ and $T_n = T_n(X_1, \ldots, X_n)$ is un unbiased estimator of $\chi(\theta)$ for a differentiable function $\chi : \mathbb{R} \mapsto \mathbb{R}$, then under regularity conditions $\operatorname{var}_\theta(\sqrt{n}T_n) \geq \chi'(\theta)^2/I_\theta$ for $I_\theta = \operatorname{var}_\theta \dot{\ell}_\theta(X_1)$ the "Fisher information" for θ.*

The *Cramér-Rao bound* is the number $\chi'(\theta)^2/I_\theta$, which depends solely on the functional χ to be estimated and on the model $\{P_\theta : \theta \in \mathbb{R}\}$, through its Fisher information. It turns out that this bound is often not sharp, in the sense that there may not exist unbiased estimators T_n for which n^{-1} their variance is equal to the bound. However, the bound is sharp in a certain asymptotic sense, as $n \to \infty$. One purpose of this lecture is to state the deep theorems that allow a precise formulation of what it means to be "asymptotically sharp", in a semiparametric context.

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

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

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 $\mathcal{P}$, when the "true measure" is P. This supremum can be expressed in the norm of the efficient influence function $\tilde{\psi}_P$.

2.2 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 $(P\tilde{\psi}_P g)^2 \leq P\tilde{\psi}_P^2 P g^2$ and the fact that, by definition, the efficient influence function $\tilde{\psi}_P$ is contained in the closure of $\operatorname{lin} \dot{\mathcal{P}}_P$. We obtain equality by choosing g equal to $\tilde{\psi}_P$. ∎

Thus, the squared norm $P\tilde{\psi}_P^2$ of the efficient influence function plays the role of a "smallest variance". Similar considerations (take linear combinations) show that the "smallest 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. The following example shows that the Cramér-Rao parametric set-up is a special case.

2.3 Example (Parametric model). Consider a parametric model as in Example 1.15. If the Fisher information matrix is invertible and the map χ is differentiable, then the efficient influence function is given by

$$\tilde{\psi}_{P_\theta} = \chi_\theta' I_\theta^{-1} \dot{\ell}_\theta.$$

Thus the appropriate covariance matrix is $P_\theta \tilde{\psi}_{P_\theta} \tilde{\psi}_{P_\theta}^T = \chi_\theta' I_\theta^{-1} (\chi_\theta')^T$. This is precisely the Cramér-Rao bound. □

It is time to give a precise meaning to "smallest covariance". We shall state two theorems regarding the estimation problem and one theorem regarding testing.

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 functional ψ is differentiable.

As usual, an estimator T_n is a measurable function $T_n(X_1, \ldots, X_n)$ of the observations.

2.4 Definition. A function $\ell: \mathbb{R}^k \mapsto [0, \infty)$ is subconvex if for all $c > 0$ the set $\{y: \ell(y) \leq c\}$ is convex, symmetric and closed.

2.5 Theorem (LAM). *Let the functional $\psi: \mathcal{P} \mapsto \mathbb{R}^k$ be differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$ 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 $\ell: \mathbb{R}^k \mapsto [0, \infty)$,*

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

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

The purpose of the theorem is to give a lower bound, depending only on the model and the functional to be estimated, for the liminf of the *risk* $\mathrm{E}_P\ell\big(\sqrt{n}(T_n - \psi(P)\big)$, for an arbitrary estimator T_n. A "best" estimator T_n can then be defined as one that attains equality (of the limsup and for every $P \in \mathcal{P}$). The theorem is more complicated than that and involves a supremum over the risk over shrinking neighbourhoods of P. A slightly weaker assertion makes this clearer. Let $\|\cdot\|$ be the variation norm.

2.6 Corollary.

$$\inf_{\delta>0} \liminf_{n\to\infty} \sup_{\|Q-P\|<\delta} \mathrm{E}_Q \ell\big(\sqrt{n}(T_n - \psi(Q))\big) \geq \int \ell \, dN\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big).$$

Without taking the (local) maximum risk the theorem would fail.

It is attractive that the LAM theorem applies to any estimator, even though it may blur the distinction between two estimator sequences by evaluating only a maximum risk. The next theorem avoids the maximum, but at the strong price of restricting itself to regular estimator sequences. An estimator sequence T_n is *regular* at P for estimating $\psi(P)$ (relative to $\dot{\mathcal{P}}_P$) if there exists a probability measure L such that

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

It follows from the definition of weak convergence (or the portmanteau lemma), that for a regular estimator sequence and bounded, continuous function ℓ the limiting local maximum risk in the left side of the LAM theorem reduces to $\int \ell \, dL$. Thus if ℓ is subconvex, this is bounded by $\int \ell \, dN\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big)$ for any such ℓ. The convolution theorem shows that this discrepancy between limit and lower bound always results from L being more dispersed than the normal measure.

2.7 Theorem (Convolution). *Let the functional $\psi: \mathcal{P} \mapsto \mathbb{R}^k$ be differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. Let T_n be regular at P with limit distribution L.*

(i) if $\dot{\mathcal{P}}_P$ is a cone, then $\int yy^T\,dL(y) - P\tilde{\psi}_P\tilde{\psi}_P^T$ is nonnegative definite.

(ii) if $\dot{\mathcal{P}}_P$ is a convex cone, then there exists a probability measure M such that $L = N\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big) * M$.

Both theorems give the message that the normal distribution with mean zero and covariance matrix $P\tilde{\psi}_P\tilde{\psi}_P^T$ is a best limiting distribution for an estimator sequence. We should not take this in a too absolute sense. For instance, shrinkage estimators as first invented by Stein in the 1950s are not regular, and hence are not in the realm of the convolution theorem, and are LAM for certain loss functions ℓ, because in fact better than the usual estimators (which are best regular and LAM), but are not asymptotically normal. A second reason to be careful is that both the LAM and convolution theorems need assumptions on the form of the tangent set. Nevertheless, so far best regular estimator sequences have been considered "best" in semiparametric theory. We adopt the same convention in the following definition.

2.8 Definition. An estimator sequence is *asymptotically efficient* at P for estimating the differentiable parameter $\psi(P)$, if it is regular at P with limit distribution $L = N(0, P\tilde{\psi}_P\tilde{\psi}_P^T)$.

We note that our definition of asymptotic efficiency is not absolute, because it is relative to a given tangent set, and we permit a variety of tangent sets. In "practice" one hunts for a pair of a tangent set and estimator sequence such that the tangent set is "big enough" and the estimator sequence "efficient enough" so that the latter is asymptotically efficient according to the preceding definition. Next one strongly believes that this is all that need to be said about the problem.

The following lemma shows that efficient estimator sequences must be asymptotically approximable by an average of the efficient influence function evaluated at the observations. Because given a sequence $y_1, y_2, \ldots$ the difference of the averages $\overline{y}_{n+1} - \overline{y}_n$ is proportional to the additional term y_{n+1}, the lemma explains the name "influence function" for $\tilde{\psi}_P$.

2.9 Lemma. *Let the functional $\psi: \mathcal{P} \mapsto \mathbb{R}^k$ 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 $N\big(0, P\tilde{\psi}_P\tilde{\psi}_P^T\big)$ if and only if it satisfies*

$$\sqrt{n}\big(T_n - \psi(P)\big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_P(X_i) + o_P(1). \tag{2.10}$$

2.11 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 1.16 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 certainly 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 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 closed, linear span of $\dot{\mathcal{P}}_P$ 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 (2.10), with the $o_P(1)$-remainder term identically zero. □

The problem of testing a null hypothesis $H_0\colon \psi(P) \le 0$ versus the alternative $H_1\colon \psi(P) > 0$ 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. For higher-dimensional parameters it is difficult to define a satisfactory notion of asymptotic optimality. We therefore first concentrate on real-valued functionals $\psi\colon \mathcal{P} \mapsto \mathbb{R}$.

Given a model $\mathcal{P}$ and a measure P on the boundary of the hypotheses, i.e. $\psi(P) = 0$, we shall study the "local asymptotic power" in a neighbourhood of P. For every score function g for which $\dot\psi_P g = P\tilde\psi_P g > 0$, the corresponding submodel $P_{t,g}$ belongs to the alternative hypothesis H_1 for (at least) every sufficiently small, positive t, since $\psi(P_{t,g}) = tP\tilde\psi_P g + o(t)$ if $\psi(P) = 0$. Thus the measures $P_{h/\sqrt{n},g}$ can be viewed as "local alternatives".

A test function ϕ_n is an estimator that takes is values in $[0, 1]$. The interpretation is that we reject the null hypothesis if the observed value of ϕ_n is 1, do not reject if it is 0, and reject it with probability ϕ_n (performing an additional experiment) if it is between 0 and 1. The following theorem shows that tests whose probabilities of the first kind (rejecting H_0 if it is true) are bounded above by some level α necessarily have probabilities of the second kind (not rejecting H_0 if it is false) bounded below by a certain Gaussian integral. Let $z_\alpha = \Phi^{-1}(1 - \alpha)$ be the upper α-quantile of the standard normal distribution.

2.12 Theorem. *Let the functional $\psi\colon \mathcal{P} \mapsto \mathbb{R}$ 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 tests ϕ_n such that*

$$\sup_{Q:\psi(Q)\le 0} Q^n\phi_n \le \alpha \in (0, 1),$$

and every $g \in \dot{\mathcal{P}}_P$ with $P\tilde\psi_P g > 0$ and every $h > 0$,

$$\limsup_{n\to\infty} P^n_{h/\sqrt{n},g}\phi_n \le 1 - \Phi\Big(z_\alpha - h\frac{P\tilde\psi_P g}{(P\tilde\psi_P^2)^{1/2}}\Big).$$

It is reasonable to expect that a test based on an efficient estimator is efficient as a test, and this is true, as we now show using the preceding theorem. Suppose that the sequence of estimators T_n is asymptotically efficient for $\psi(P)$ at P and that S_n is a consistent sequence of estimators of its asymptotic variance $P\tilde{\psi}_P^2$. Then the test that rejects $H_0: \psi(P) = 0$ for $\sqrt{n}T_n/S_n \geq \Phi^{-1}(1-\alpha)$ attains the upper bound of the theorem. The critical value z_α is chosen exactly so that the asymptotic probability of an error of the first kind is α: $\mathrm{P}_P(\sqrt{n}T_n/S_n \geq z_\alpha) \to \alpha$.

2.13 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 $h \geq 0$ and $g \in \dot{\mathcal{P}}_P$,*

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

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 $P\tilde{\psi}_P^2$. Thus the lemma follows by simple algebra. ∎

2.14 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 examples 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 $H_0: \int F\,dG \leq \frac{1}{2}$ versus the alternative $H_1: \int F\,dG > \frac{1}{2}$. The Wilcoxon test rejects H_0 for large values of $\int \mathbb{F}_n\,d\mathbb{G}_n$, where $\mathbb{F}_n$ and $\mathbb{G}_n$ are the empirical distribution functions of the two samples. This test is asymptotically efficient relative to the model in which F and G are completely unknown. This gives a different perspective on this test, which is usually presented as being asymptotically optimal for testing a difference of 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. 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 $\psi(F \times G) = \int F\,dG$. We do not give the complete argument for this, but note that it could be derived from the efficiency of the $\mathbb{F}_n$ for F and of $\mathbb{G}_n$ for G, which we noted in Example 2.11, either by applying a preservation theorem of efficiency, or by similar arguments. □

All three theorems presented in this section give a special role to normal distributions with covariance matrix $P\tilde{\psi}_P\tilde{\psi}_P^T$. We have motivated the covariance matrix by the Cramér-Rao theorem, but the normality is a new element. That "normal limit distributions are best" was proved for parametric models in the 1970s by Hájek, and is best explained from Le Cam's theory of limiting experiments. This theory shows that the sequence of statistical experiments

$$(P^n_{1/\sqrt{n},g}: g \in \dot{\mathcal{P}}_P)$$

converges in the weak sense of Le Cam to a Gaussian location experiment, indexed by the tangent set $\dot{\mathcal{P}}_P$. We do not discuss this convergence theory here, but do present a fourth theorem that is more in its spirit.

2.15 Theorem. *Suppose that T_n are estimators with values in a separable, Banach space $\mathbb{D}$ such that, for every $g \in \dot{\mathcal{P}}_P$ and a probability measure L_g,*

$$\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n},g})\big) \rightsquigarrow L_g, \qquad \text{under } P_{1/\sqrt{n},g}.$$

If ψ is differentiable at P, relative to $\dot{\mathcal{P}}_P$, then for any orthonormal sequence $g_1, \ldots, g_m$ in $L_2(P)$ there exists a measurable map $T\colon \mathbb{R}^m \times [0,1] \mapsto \mathbb{D}$ such that $T - \dot{\psi}_P(g)$ is distributed as L_g if the law of T is calculated under the product of the normal measure with mean $\big(\langle g, g_1\rangle_P, \ldots, \langle g, g_m\rangle_P\big)$ and covariance the identity and the uniform measure on $[0,1]$.

The measurable map T in this theorem should be regarded as a randomized estimator $T = T(X,U)$ in a statistical experiment that consists of observing a vector $X = (X_1, \ldots, X_m)$ of m independent normal variables, with means $\langle g_i, g\rangle_P$ depending on an unknown parameter g and unit variance. The estimator is allowed to depend also on an auxiliary uniform variable U that can be generated by the statistician. (For many purposes it is not helpful to use randomization, but sometimes, as with nonconvex loss functions, it may be.) The theorem shows that asymptotically the problem of statistical inference about $\psi(P_{1/\sqrt{n},g})$ based on a sample of size n from $P_{1/\sqrt{n},g}$, where g is unknown, is matched by the problem of estimating $\dot{\psi}_P(g)$ based on X. Here we could restrict $g = \sum_{i=1} a_i g_i$ to the linear span of $g_1, \ldots, g_m$ and develop the parameter of interest $\dot{\psi}_P(g) = \sum_{i=1}^m a_i \dot{\psi}_P(g_i)$. Then we are to make inference about a linear function $\sum a_i d_i$ based on a normal $N_m(a, I)$-distributed vector, which is a well-studied problem with simple solutions. The preceding theorems are merely specifications to particular problems (minimax estimation, equivariant estimation, or uniformly most powerful testing) of this Gaussian approximation. Using the preceding theorem we could obtain a load of other concrete statements on asymptotic lower bounds, provided we can solve the particular question in the Gaussian experiment. For instance, we can derive statements for tangent sets that do not satisfy the convexity or linearity requirements of the preceding theorems; we can consider loss functions that are not subconvex; or we can consider testing of higher-dimensional functionals. The problem with testing a parameter of dimension 2 or higher is that no uniformly most powerful, unbiased test does exist and hence an optimal test can only be defined through restricting the class of tests or working with envelope power functions. Appropriate restriction through invariance will of course lead to the same conclusion that tests best on best regular estimator sequences are best invariant tests.

Rather than using finitely many functions $g_1, \ldots, g_m$, we could have used an infinite sequence $g_1, g_2, \ldots$ (unless $L_2(P)$ is finite dimensional). The analogous result will be true. However, the analysis of an infinite-dimensional Gaussian experiment will proceed by finite-dimensional approximation, so not much is gained by this formulation. We have a similar reservation against a representation of the Gaussian experiment using a Brownian motion with drift (as in [22]). It is impossible to perform direct calculations on risks of estimators which are measurable functions of

Brownian motion and hence it will be necessary to approximate the experiment by finite-dimensional ones in any case.

Proofs of generalizations of the preceding theorems are given in Lecture 4.

2.2 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) will 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 where 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 when θ is fixed, and will run 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 $L_2(P_{\theta,\eta})$.

2.16 Definition.

(i) The *efficient score function* for θ is $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$.

(ii) The *efficient information matrix* for θ is $\tilde{I}_{\theta,\eta} = P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^T$.

2.17 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*

$$\int \Big[\frac{dP^{1/2}_{\theta+ta,\eta_t} - dP^{1/2}_{\theta,\eta}}{t} - \tfrac{1}{2}(a^T\dot\ell_{\theta,\eta} + g)\, dP^{1/2}_{\theta,\eta}\Big]^2 \to 0. \tag{2.18}$$

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 $\tilde\psi_{\theta,\eta} = \tilde I^{-1}_{\theta,\eta}\tilde\ell_{\theta,\eta}$.

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 since

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

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

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

This is very similar to the situation for efficient estimators in parametric models. The only difference is that the ordinary score function $\dot\ell_{\theta,\eta}$ is replaced by the efficient score function (and similarly for the informations). 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. When 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.

2.19 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 (2.18).

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! We shall see this in Lecture 8. □

2.20 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

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

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 $\eta\big(x, y - g_\theta(x)\big)$, where η is (essentially) only restricted by the relations $\int e\eta(x, e)\, de \equiv 0$.

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

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

By the same argument as for nonparametric models all square-integrable functions of this type that have mean zero are score functions. Since the relation $\mathrm{E}\big(ea(X, e)|\, X\big) = 0$ 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

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

This readily yields the efficient score function

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

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

Notes

The study of the symmetric location model has a long history. That the scores for the location parameter and the shape parameter were orthogonal was first noted by Stein in [34]. Several authors subsequently worked on defining adaptive estimators. A summary approach was given by Bickel in [2], which provided a starting point to extensions to more general models.

The convolution and minimax theorems for parametric models are due to Hájek, see [10] and [11]. The semiparametric versions given here are, in a way, simple extensions of these theorems. The role of convexity or linearity of tangent spaces for these theorems was investigated in [36], which is also the basis of Theorem 2.15.

Efficient score functions were presented by Begun, Hall, Huang and Wellner, in [1], as an alternative to the (more general) presentations by Levit and Pfanzagl.

Lecture 3
Calculus of Scores

In this lecture we introduce a "calculus of scores", which is a useful way of finding efficient influence functions in models that are parametrized.

3.1 Score and Information Operators

The method to find the efficient influence function of a parameter given in the preceding lecture 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 will map 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. Since A_η turns scores for the model H into scores for the model $\mathcal{P}$ it is called a *score operator*. Ahead it is seen 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* $A_\eta^*: L_2(P_\eta) \mapsto \overline{\text{lin}}\, \dot{H}_\eta$. 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 $b \in \dot{H}_\eta$. Note that we define A_η^* to have range $\overline{\text{lin}}\, \dot{H}_\eta$, 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 $\overline{\text{lin}}\, \dot{H}_\eta$.

3.1 Fact. *Every continuous, linear map $A: \mathbb{H}_1 \mapsto \mathbb{H}_2$ between two Hilbert spaces has an adjoint map $A^*: \mathbb{H}_2 \mapsto \mathbb{H}_1$, which is a continuous, linear map that satisfies and is uniquely determined by the equations $\langle A^* h_2, h_1\rangle_1 = \langle h_2, A h_1\rangle_2$ for every $h_i \in \mathbb{H}_i$. If A is considered the restriction to $\mathbb{H}_1 \subset \tilde{\mathbb{H}}_1$ of a continuous, linear map $\tilde{A}: \tilde{\mathbb{H}}_1 \mapsto \mathbb{H}_2$ with domain a Hilbert space that contains $\mathbb{H}_1$ isometrically, then $A^* = \Pi \tilde{A}^*$ for $\Pi: \tilde{\mathbb{H}}_1 \mapsto \mathbb{H}_1$ the orthogonal projection of $\tilde{\mathbb{H}}_1$ onto $\mathbb{H}_1$.*

The preceding display is equivalent to

$$A_\eta^* \tilde{\psi}_{P_\eta} = \tilde{\chi}_\eta. \tag{3.2}$$

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^*. Since A_η^* is not necessarily onto $\overline{\text{lin}}\, \dot{H}_\eta$, not even when it is one-to-one, this is a condition!

For multivariate functionals equation (3.2) is to be understood coordinate-wise. Two solutions $\tilde{\psi}_{P_\eta}$ of (3.2) can differ only by an element of the kernel $\text{N}(A_\eta^*)$ of A_η^*, which is the orthocomplement $\text{R}(A_\eta)^\perp$ of the range of $A_\eta: \text{lin}\, \dot{H}_\eta \mapsto L_2(P_\eta)$. Thus, there is at most one solution $\tilde{\psi}_{P_\eta}$ that is contained in $\overline{\text{R}}(A_\eta) = \overline{\text{lin}}\, A_\eta \dot{H}_\eta$, 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 equation (3.2) can be solved, of course, and the solution can be written in the attractive form

$$\tilde{\psi}_{P_\eta} = A_\eta (A_\eta^* A_\eta)^- \tilde{\chi}_\eta. \tag{3.3}$$

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 $A_\eta^* A_\eta b = \tilde{\chi}_\eta$.) The following lemma shows that this attractive form is available for any functional χ if the range of the score operator is closed, a situation which unfortunately fails often.

3.4 Fact. *Let $A: \mathbb{H}_1 \mapsto \mathbb{H}_2$ be a continuous linear map between two Hilbert spaces. Then equivalent are:*

(i) $\text{R}(A)$ *is closed.*
(ii) $\text{R}(A^*)$ *is closed.*
(iii) $\text{R}(A^* A)$ *is closed.*
(iv) $\text{R}(A^* A) = \text{R}(A^*)$.

3.5 Fact. *Let $A: \mathbb{H}_1 \mapsto \mathbb{H}_2$ be a continuous linear map between two Hilbert spaces. Then*
(i) $\mathrm{N}(A) = \mathrm{R}(A^*)^\perp$.
(ii) $\mathrm{N}(A^*) = \mathrm{R}(A)^\perp$.
Furthermore, the map $A^*A: \mathbb{H}_1 \mapsto \mathbb{H}_1$ is one-to-one, onto and has a continuous inverse if and only if A is one-to-one and $\mathrm{R}(A)$ is closed if and only if A^*A is one-to-one and onto.

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 $t \downarrow 0$,

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

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

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 $\tilde\chi_\eta \in \overline{\operatorname{lin}}\, \mathbb{H}_\eta^k$. The preceding discussion can be extended to this abstract set-up.

3.6 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 $A_\eta^*: L_2(P_\eta) \mapsto \overline{\operatorname{lin}}\, \mathbb{H}_\eta$. The efficient influence function $\tilde\psi_{P_\eta}$ satisfies (3.2). If each coordinate function of $\tilde\chi_\eta$ is contained in the range of $A_\eta^* A_\eta: \overline{\operatorname{lin}}\, \mathbb{H}_\eta \mapsto \overline{\operatorname{lin}}\, \mathbb{H}_\eta$, then it also satisfies (3.3).*

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 (3.2). ∎

The condition (3.2) 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 $\overline{\operatorname{lin}}\, \mathbb{H}_\eta$. 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 $\overline{\operatorname{lin}}\, \mathbb{H}_\eta$. Since $b \perp \mathrm{R}(A_\eta^*)$ 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 and it is not surprising that we have a problem. 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 (3.2) may still fail. This is harder to understand, but it happens quite often. The following theorem shows that failure has serious consequences.

3.7 Theorem. *In the above setting, if* $\tilde{\chi}_\eta \notin \mathrm{R}(A_\eta^*)$, *then*
(i) there exists no estimator sequence for $\chi(\eta)$ *that is regular at* P_η.
(ii)

$$\sup_{b \in \mathbb{H}_\eta} \frac{\langle \tilde{\chi}_\eta, b \rangle_\eta^2}{\|A_\eta b\|_{P_\eta}^2} = \infty.$$

Proof. We shall only give the proof of (ii). (See [38] for a proof of (i).) The proof of (ii) can be carried out using the spectral decomposition of the information operator and spectral calculus. (See for instance [33] for this background.) For simplicity of notation, we drop the index η throughout the proof. The spectral decomposition takes the form $A^*A = \int \lambda \, dP_\lambda$ for $\lambda \mapsto P_\lambda$ the spectral resolution of the nonnegative, self-adjoint operator A^*A. (In simple cases, the formal integral is a sum over the (countable many, nonnegative) eigenvalues of A^*A and the P_λ are orthogonal projections on the corresponding eigenspaces. In general, the spectral resolution may be continuous.) Next the operator

$$(A^*A)^{1/2} = \int \sqrt{\lambda} \, dP_\lambda$$

is a square root in that it is nonnegative, self-adjoint and has A^*A as its square. The adjoint A^* can be expressed in this square root through the polar decomposition $A^* = (A^*A)^{1/2}U$, for $U \colon L_2(P) \mapsto \overline{\mathrm{R}}((A^*A)^{1/2}) = \mathrm{N}((A^*A)^{1/2})^\perp$ an operator whose restriction to $\overline{\mathrm{R}}(A)$ is an isometry and has $\mathrm{R}(A)^\perp$ as its kernel. It follows that the ranges of A^* and $(A^*A)^{1/2}$ are identical.

The spectral calculus also gives a meaning to integrals of the type $\int f(\lambda) \, dP_\lambda$ for general functions f. Such expressions define operators, which can be manipulated with rules such as $\int f(\lambda) \, dP_\lambda \int g(\lambda) \, dP_\lambda = \int f(\lambda) g(\lambda) \, dP_\lambda$.

Furthermore, to every $b \in \mathbb{H}$ corresponds a spectral measure μ_b, which is a measure on the interval $[0, \|A^*A\|]$ containing the spectrum with the property that $\langle \int f(\lambda) \, dP_\lambda b, b \rangle = \int |f|^2 \, d\mu_b$ for every function f that is well-behaved on the spectrum of A^*A.

We then obtain that

$$\mathrm{R}(A^*) = \mathrm{R}\big((A^*A)^{1/2}\big) = \Big\{ b \colon \int \lambda^{-1} \, d\mu_b(\lambda) < \infty \Big\}.$$

Therefore, if $\tilde{\chi}$ is not contained in the range of A^*, then we must have at least one of

$$\mu_{\tilde{\chi}}\{0\} > 0, \qquad \text{or} \qquad \int 1_{\lambda > 0} \lambda^{-1} \, d\mu_{\tilde{\chi}}(\lambda) = \infty.$$

In the first case we evaluate the quotient in (ii) at $b = P_{\{0\}}\tilde{\chi} = \int_{\{0\}} dP_\lambda \tilde{\chi}$. For this choice we have by spectral calculus that $A^*Ab = \int \lambda \, dP_\lambda P_{\{0\}} \tilde{\chi} = \int 1_{\{0\}}(\lambda) \lambda \, dP_\lambda \tilde{\chi} = 0$, whereas $\langle \tilde{\chi}, b \rangle = \mu_{\tilde{\chi}}\{0\} > 0$. This yields the quotient $(> 0)^2/0$ in (ii). In the case that the second possibility in the preceding display is valid we evaluate the quotient in (ii) at the sequence $b_n = \int 1_{\lambda \geq 1/n} \lambda^{-1} \, dP_\lambda \tilde{\chi}$. For this choice we have by spectral calculus that $A^*Ab_n = \int 1_{\lambda \geq 1/n} \, dP_\lambda \tilde{\chi} \to \tilde{\chi}$, whereas $\langle \tilde{\chi}, b_n \rangle = \int 1_{\lambda \geq 1/n} \lambda^{-1} \, d\mu_{\tilde{\chi}}(\lambda) \to \infty$. Thus the quotient in (ii) is infinite. ∎

3.1.1 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:

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

If we consider the scores b and g as the carriers of information about t in the variables $Y \sim \eta_t$ and $X \sim P_{\eta_t}$, 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.

3.8 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 $b : \mathcal{Y} \mapsto \mathbb{R}$*

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

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

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

Proof. For simplicity of notation we assume that the measures η_t and η have densities h_t and h relative to a fixed probability measure ν. (If this is not the case, choose $\nu = \nu_t = \frac{1}{2}(\eta_t + \eta)$ dependent on t and add t's throughout the following.) Furthermore, we assume that b is uniformly bounded by M. (If this is not the case truncate b at $M_t \to \infty$ and add t's in the following.) Then $\nu u_t^2 \to 0$ for

$$u_t = \frac{h_t^{1/2} - h^{1/2}}{t} - \tfrac{1}{2} b h^{1/2}.$$

Define μ to be the law of $X = m(Y)$ if Y is distributed according to ν. Then

$$p_t(x) = \mathrm{E}_\nu\big(h_t(Y) \,|\, X = x\big), \qquad \text{and} \qquad p(x) = \mathrm{E}_\nu\big(h(Y) \,|\, X = x\big)$$

are densities of P_{η_t} and P_η with respect to μ. In case of the second one, this follows from the equations

$$\begin{aligned} P_\eta(A) = \int 1_A\big(m(y)\big)\, d\eta(y) &= \mathrm{E}_\nu 1_A\big(m(Y)\big) h(Y) \\ &= \mathrm{E}_\nu 1_A(X) \mathrm{E}_\nu\big(h(Y) \,|\, X\big) = \int_A p(x)\, d\mu(x). \end{aligned}$$

By a similar argument, we have, almost surely under P_η,

$$A_\eta b(X)\, p(X) = \mathrm{E}_\nu\big(b(Y)h(Y)\big|\, X\big).$$

From the definition of u_t we obtain that

$$h_t = h + tbh + t^2u_t^2 + t(tu_tbh^{1/2} + 2u_th^{1/2} + \tfrac{1}{4}tb^2h).$$

Evaluating these functions at Y and taking conditional expectations with respect to X, we find

$$p_t = p + t(A_\eta b)\, p + c + d,$$

where c and d satisfy

$$\begin{aligned} c(X) &= t^2\mathrm{E}_\nu\big(u_t^2(Y)\big|\, X\big),\\ |d(X)|^2 &= t^2\mathrm{E}_\nu\Big(\big(tu_tbh^{1/2} + 2u_th^{1/2} + \tfrac{1}{4}tb^2h\big)(Y)\big|\, X\Big)^2\\ &\lesssim t^2\mathrm{E}_\nu\Big(\big(u_th^{1/2}(tM+1) + tM^2h\big)(Y)\big|\, X\Big)^2\\ &\lesssim t^2\Big(\mathrm{E}_\nu\big(u_t^2(Y)\big|\, X\big)(tM+1)^2 + t^2M^4p(X)\Big)p(X), \end{aligned}$$

by the Cauchy-Schwarz inequality. By a Taylor expansion (see Lemma 3.9), we conclude that on the set $A = \{p > 0\}$

$$\begin{aligned} \Big[\frac{p_t^{1/2} - p^{1/2}}{t} - \tfrac{1}{2}(A_\eta b)p^{1/2}\Big]^2 &\lesssim \mathrm{E}_\nu\big(u_t^2(Y)\big|\, X\big)(tM+1)^2 + t^2M^4p(X)\\ &\quad + \mathrm{E}_\nu\big(u_t^2(Y)\big|\, X\big) + \Big|\frac{1}{\sqrt{1-Mt}} - 1\Big|^2M^2p(X). \end{aligned}$$

The integral over the set A of this function relative to μ converges to zero as $t \to 0$.

Finally, the equation $\eta\big(m^{-1}(A^c)\big) = P_\eta(A^c) = 0$ implies that $P_{\eta_t}(A^c) = \eta_t\big(m^{-1}(A^c)\big) = o(t^2)$, because $\nu(tu_t^2 1_B) = \eta_t(B)$ if $\eta(B) = 0$. Thus the integral of the preceding display over the set A^c converges to zero as well. ∎

3.9 Lemma. *For any real numbers a, b, c, d with $a > 0$, $b/a \le \varepsilon < 1$, $c \ge 0$ and $a + b + c + d \ge 0$*

$$\Big|\sqrt{a+b+c+d} - \sqrt{a} - \tfrac{1}{2}\frac{b}{\sqrt{a}}\Big|^2 \le \frac{3d^2}{a(1-\varepsilon)} + 3c + \Big|\frac{1}{\sqrt{1-\varepsilon}} - 1\Big|^2\frac{b^2}{a}.$$

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

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

This follows since, by the usual rules for conditional expectations,

$$\mathrm{E}\mathrm{E}\big(g(X)\big|\, Y\big)b(Y) = \mathrm{E}g(X)b(Y) = \mathrm{E}g(X)\mathrm{E}\big(b(Y)\big|\, X\big).$$

In the "calculus of scores" of Theorem 3.6 the adjoint is understood to be the adjoint of $A_\eta\colon \mathbb{H}_\eta \mapsto L_2(P_\eta)$ 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 $\overline{\text{lin}}\,\mathbb{H}_\eta$.

3.10 Example (Mixtures). Suppose that a typical observation X possesses a conditional density $p(x|\,z)$ 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

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

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

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

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 $\{b \in L_2(\eta)\colon \eta b = 0\}$.

In particular, consider the situation that the kernel $p(x|\,z)$ belongs to an exponential family, i.e. $p(x|\,z) = c(z)d(x)\,\exp\big(z^T x\big)$. Mixtures over exponential families of this type give relatively large models. In fact, if the interior of the support of η is nonempty, than the tangent set $A_\eta\dot{H}_\eta$ is dense in the maximal tangent set $\{g \in L_2(P_\eta)\colon P_\eta g = 0\}$. We show this below.

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. This is somewhat surprising, because the mixture densities may still possess very special properties. For instance, mixtures over the exponential scale family $p(x|\,z) = ze^{zx}1_{x>0}$ are monotone densities, and mixtures over the normal location family are extremely smooth. In terms of entropy the second collection of mixtures is almost finite-dimensional and there exist estimators $p_{\hat{\eta}}$ that obtain a rate of convergence in the Hellinger distance of the order $\log n/\sqrt{n}$. Thus the set of all exponential mixtures can be far from being equal to the nonparametric model.

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

$$0 = A_\eta^* g(z) = \mathrm{E}\big(g(X)|\,Z = z\big) = \int g(x)\,p(x|\,z)\,dx$$

says exactly that $g(X)$ is a zero-estimator under $p(x|\,z)$. Since 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.

The same argument shows also that the range of the score operator, equivalently the range of its adjoint, is not closed in this example. This has as a consequence that many functionals $y \mapsto \chi(\eta)$ are not in the realm of the $\sqrt{n}$-theory of estimation. As an example consider the functional $\chi(\eta) = \eta(A)$ for a given set A. This has influence

function $\tilde{\chi} = 1_A - \eta(A)$, which is contained in the range of A_η^* if and only if there exists a measurable function g such that

$$1_A(z) = \int g(x)c(z)d(x)e^{z^T x}\, d\mu(x), \qquad \eta\text{-a.e..}$$

The completeness of the exponential family shows that the A must have probability 0 or 1 under η. Functionals as these belong to the realm of inverse problems. Not much is known about them today. The deconvolution problem (i.e. $p(x|\, z)$ a location family) has best been studied, with a characterization of rates for estimating the mixing distribution function and its derivatives using Fourier inversion methods. Even in this case very little is known concerning standard methods of estimation, such as maximum likelihood.

If the support of η does not contain a limit point, then the preceding approach to show that the tangent set is dense 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 $0 \leq at \leq 1$, for any fixed $a \geq 0$ and η_1, and lead to scores

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

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

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

If the set of distributions $\{P_\eta\colon \eta \in H\}$ 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. Since this set of scores is also a convex cone, Theorems 2.7 and 2.5 next show that nonparametric estimators are asymptotically efficient. □

3.2 Semiparametric Models

In a semiparametric model $\{P_{\theta,\eta}\colon \theta \in \Theta, \eta \in H\}$, 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 $B_{\theta,\eta}\colon \mathbb{H}_\eta \mapsto L_2(P_{\theta,\eta})$ is the "score operator" for the nuisance parameter. The domain of the operator $A_{\theta,\eta}\colon \mathbb{R}^k \times \operatorname{lin} \mathbb{H}_\eta \mapsto L_2(P_{\theta,\eta})$ is a Hilbert space relative to the inner product

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

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.

3.11 Fact. *Given a continuous, linear map $A: \mathbb{H}_1 \mapsto \mathbb{H}_2$ between Hilbert spaces, the operator $A(A^*A)^{-1}A^*$ (if it exists) is the orthogonal projection of $\mathbb{H}_1$ onto the range space of A.*

The efficient influence function for estimating θ is expressed in the *efficient score function* for θ in Lemma 2.17, which is defined as the ordinary score function minus its projection 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

$$\tilde{\ell}_{\theta,\eta} = \big(I - B_{\theta,\eta}(B_{\theta,\eta}^* B_{\theta,\eta})^{-1} B_{\theta,\eta}^*\big)\dot{\ell}_{\theta,\eta}. \tag{3.12}$$

This means that $b = -(B_{\theta,\eta}^* B_{\theta,\eta})^{-1} B_{\theta,\eta}^* \dot{\ell}_{\theta,\eta}$ is a "least favourable 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 (3.2). The adjoint of $A_{\theta,\eta}: \mathbb{R}^k \times \mathbb{H}_\eta \mapsto \underline{L_2}(P_{\theta,\eta})$, and the corresponding information operator $A_{\theta,\eta}^* A_{\theta,\eta}: \mathbb{R}^k \times \mathbb{H}_\eta \mapsto \mathbb{R}^k \times \overline{\operatorname{lin}}\, \mathbb{H}_\eta$ are given by, with $B_{\theta,\eta}^*: L_2(P_{\theta,\eta} \mapsto \overline{\operatorname{lin}}\, \mathbb{H}_\eta$ the adjoint of $B_{\theta,\eta}$,

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

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 (3.2) becomes

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

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

$$B_{\theta,\eta}(B_{\theta,\eta}^* B_{\theta,\eta})^- \tilde{\chi}_\eta - \langle B_{\theta,\eta}(B_{\theta,\eta}^* B_{\theta,\eta})^- \tilde{\chi}_\eta, \dot{\ell}_{\theta,\eta}\rangle_{P_{\theta,\eta}}^T \tilde{I}_{\theta,\eta}^{-1} \tilde{\ell}_{\theta,\eta}.$$

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. Since it is orthogonal to the first part, it adds a positive contribution to the variance.

3.13 Example (Cox model). We illustrate the general formulas by explicit calculations for the Cox model. This model is appropriate for this purpose, because the information operator can be obtained in a simple form, whereas in other models not much progress can be made beyond writing out formulas for the score operator and its adjoint.

For later reference we consider the Cox model under right censoring. In this model we observe a random sample from the distribution of the variable $X = (T \wedge C, 1\{T \leq C\}, Z)$, where, given Z, the variables T and C are independent, as in the right censoring model, and (Z, T) follows the Cox model. Thus, the density of $X = (Y, \Delta, Z)$ is given by

$$\Big(e^{\theta z}\lambda(y)e^{-e^{\theta z}\Lambda(y)}\big(1 - F_{C|Z}(y-|\, z)\big)\Big)^{\delta}\Big(e^{-e^{\theta z}\Lambda(y)}f_{C|Z}(y|\, z)\Big)^{1-\delta}p_Z(z).$$

We make a number of assumptions, whose main purpose is to simplify the formulas and to ensure the existence of the inverse of the information operator. First, we assume that the covariate Z is bounded, and that the true conditional distribution of T given Z possesses a continuous Lebesgue density. 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.

The score function for θ takes the form, with $x = (y, \delta, z)$

$$\dot{\ell}_{\theta,\Lambda}(x) = \delta z - ze^{\theta z}\Lambda(y).$$

For any bounded, measurable function $a\colon [0, \tau] \mapsto \mathbb{R}$, the path defined by $d\Lambda_t = (1 + ta)\, d\Lambda$ defines a submodel passing through Λ at $t = 0$. Its score function at $t = 0$ takes the form

$$B_{\theta,\Lambda}a(x) = \delta a(y) - e^{\theta z}\int_{[0,y]} a\, d\Lambda.$$

For unbounded functions a we could employ paths of the form $d\Lambda_t = \chi(ta)\, d\Lambda$ and obtain a score of the same form. The score operator can be viewed as an operator $B_{\theta,\Lambda}\colon L_2(\Lambda) \mapsto L_2(P_{\theta,\Lambda})$, so we can take $\mathbb{H}_\Lambda = L_2(\Lambda)$ or take $\mathbb{H}_\Lambda$ equal to the subset of all bounded functions in $L_2(\Lambda)$.

To find a formula for the adjoint $B^*_{\theta,\Lambda}$ of $B_{\theta,\Lambda}\colon L_2(\Lambda) \mapsto L_2(P_{\theta,\Lambda})$, we write

$$\begin{aligned}
&\langle B^*_{\theta,\Lambda}g, a\rangle = \langle g, B_{\theta,\Lambda}a\rangle_{P_{\theta,\Lambda}}\\
&= E_Z\int g(y,1,z)\Big(a(y) - e^{\theta z}\int_0^y a\, d\Lambda\Big)e^{\theta z}e^{-e^{\theta z}\Lambda(y)}\\
&\qquad\qquad \big(1 - F_C(y-|\, z)\big)\, d\Lambda(y)\\
&+ \mathrm{E}_Z\int g(y,0,z)\Big(-e^{\theta z}\int_0^y a\, d\Lambda\Big)e^{\theta z}e^{-e^{\theta z}\Lambda(y)}\, dF_C(y|\, z).
\end{aligned}$$

Next we use Fubini's theorem to change the order of integration in the two terms, rewriting the right side as $\int a\big[\cdots\big]\, d\Lambda$. By definition the term appearing inside the square brackets is then $B^*_{\theta,\Lambda}g$. It is given by

$$\begin{aligned}
B^*_{\theta,\Lambda}g(y) = {}& E_Z g(y,1,z)e^{\theta z}e^{-e^{\theta z}\Lambda(y)}\big(1 - F_C(y-|\, z)\big)\\
&- \mathrm{E}_Z\int g(s,1,z)e^{2\theta z}1_{y\leq s}e^{-e^{\theta z}\Lambda(s)}\big(1 - F_C(s-|\, z)\big)\, d\Lambda(s)\\
&- \mathrm{E}_Z\int g(y,0,z)e^{2\theta z}e^{-e^{\theta z}\Lambda(s)}1_{y\leq s}\, dF_C(s|\, z).
\end{aligned}$$

This is not the simple formula promised in the introduction, though it has the benefit of being obtainable by simple mechanical manipulations. Now, $B^*_{\theta,\Lambda}g$ for an arbitrary function g is not really what interests us: rather we would like to obtain formulas for the information operator $B^*_{\theta,\Lambda}B_{\theta,\Lambda}g$ and for $B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}$. For this we can continue our mechanical work by combining the formulas obtained so far. This is straightforward again, and for most examples this would be the end of the story. The Cox model is special in that clever partial integrations can next simply the formulas considerably.

We shall not pursue this approach, as it is tedious and not insightful. Rather we obtain the desired formulas using a statistical principle: minus the mean of the observed information is the Fisher information. A preciser formulation of this principle is that, given probability densities $x \mapsto p_{s,t}(x)$ that depend smoothly on a parameter $(s,t) \in \mathbb{R}^2$, we have

$$\mathrm{E}_{s,t}\Big(\frac{\partial}{\partial s}\log p_{s,t}\Big)\Big(\frac{\partial}{\partial t}\log p_{s,t}\Big) = -\mathrm{E}_{s,t}\frac{\partial^2}{\partial s\partial t}\log p_{s,t}.$$

We apply this to the submodels $(s,t) \mapsto P_{\theta,\Lambda_{s,t}}$ for $d\Lambda_{s,t} = (1+sa+tb+stab)\,d\Lambda = (1+sa)\,d\Lambda_{0,t}$ at $(s,t)=(0,0)$. This gives

$$\begin{aligned}
\mathrm{E}_{\theta,\Lambda}(B_{\theta,\Lambda}a)(B_{\theta,\Lambda}b) &= -\mathrm{E}_{\theta,\Lambda}\frac{\partial^2}{\partial s\partial t}_{|s,t=0} p_{\theta,\Lambda_{s,t}} \\
&= -\mathrm{E}_{\theta,\Lambda}\frac{\partial}{\partial t}_{|t=0} B_{\theta,\Lambda_{0,t}}a \\
&= \mathrm{E}_{\theta,\Lambda}e^{\theta Z}\int_{[0,Y]} ab\,d\Lambda \\
&= \int b(s)\mathrm{E}_{\theta,\Lambda}e^{\theta Z}1_{s\le Y}\,a(s)\,d\Lambda(s).
\end{aligned}$$

By definition of the adjoint, the left side of this display is also equal to the inner product of b and $B^*_{\theta,\Lambda}B_{\theta,\Lambda}a$ in $L_2(\Lambda)$. Thus we read off that the information operator is the multiplication operator given by

$$B^*_{\theta,\Lambda}B_{\theta,\Lambda}a(s) = \Big(\mathrm{E}_{\theta,\Lambda}e^{\theta Z}1_{s\le Y}\Big)a(s).$$

The function $B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}$ can be obtained by a similar argument, using the submodel $(s,t) \mapsto P_{\theta+s,\Lambda_t}$ with $d\Lambda_t = (1+tb)d\Lambda$. It is given by

$$B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda}(s) = \mathrm{E}_{\theta,\Lambda}1_{s\le Y}Ze^{\theta Z}.$$

It is remarkable that the information operator is already in its spectral form. It is a theorem in Hilbert space theory that every self-adjoint operator can be written as a multiplication operator, relative to an appropriate coordinate system. In the present case the information operator already takes the form of a multiplication operator relative to the original coordinate system.

It is easy to invert a multiplication operator. In the present situation, if (θ,Λ) is a pair of parameters that satisfies the assumptions we have made, the multiplier

function $y \mapsto \mathrm{E}_{\theta,\Lambda} 1_{y \le Y} e^{\theta Z}$ is bounded away from zero on $[0, \tau]$. Thus the inverse of the information operator exists as a continuous operator and is given by

$$(B^*_{\theta,\Lambda} B_{\theta,\Lambda})^{-1} a(s) = \Big(\mathrm{E}_{\theta,\Lambda} e^{\theta Z} 1_{s \le Y} \Big)^{-1} a(s).$$

The efficient score function takes the general form (3.12), which, with the functions $L_{i,\theta}(y) = \mathrm{E} 1_{Y \ge y} Z^i e^{\theta Z}$, reduces to

$$\tilde{\ell}_{\theta,\Lambda}(x) = \delta\Big(z - \frac{L_{1,\theta}}{L_{0,\theta}}(y)\Big) - e^{\theta z} \int_{[0,y]} \Big(z - \frac{L_{1,\theta}}{L_{0,\theta}}(t)\Big)\, d\Lambda(t).$$

The efficient information for θ can be computed from this as

$$\tilde{I}_{\theta,\Lambda} = \mathrm{E} e^{\theta Z} \int \Big(Z - \frac{L_{1,\theta}}{L_{0,\theta}}(y)\Big)^2 \overline{G}_{\theta,\Lambda}(y|\, Z)\, d\Lambda(y),$$

where $\overline{G}(y|\, Z) = \mathrm{P}(Y \ge y|\, Z)$. This is strictly positive by the assumption that Z is not almost surely equal to a function of Y.

The formula for $\tilde{I}_{\theta,\Lambda}$ can be obtained by direct (but tedious, if not difficult) computations. Alternatively, we can use martingale theory. The process

$$M_t = 1_{T \le t} - \int_{[0,t]} 1_{s \le T} e^{\theta Z}\, d\Lambda(s)$$

is a martingale relative to the filtration generated by (Z, C) and $1_{T \le s}$ for $s \le t$, with predictable quadratic variation process

$$\langle M_t \rangle = \int_{[0,t]} 1_{s \le T} e^{\theta Z}\, d\Lambda(s).$$

(We have assumed that Λ is continuous.) The efficient score can be written as the integral

$$\tilde{\ell}_{\theta,\Lambda}(X) = \int 1_{t \le C} \Big(Z - \frac{L_{1,\theta}}{L_{0,\theta}}(t)\Big)\, dM_s.$$

Because the integrand is predictable, the integral can be viewed as both an ordinary Stieltjes integral and a stochastic integral. By the second interpretation we have that

$$\mathrm{E} \tilde{\ell}^2_{\theta,\Lambda}(X) = \mathrm{E} \int 1_{s \le C} \Big(Z - \frac{L_{1,\theta}}{L_{0,\theta}}(t)\Big)^2\, d\langle M_t \rangle.$$

This can be seen to reduce to the formula obtained previously. □

Notes

This lecture is based on the papers [1] and [37].

Lecture 4
Gaussian Approximations

In this lecture we give proofs of the lower bound theorems stated in Lecture 2, in a more general setting. For completeness we start by a crash course on contiguity.

4.1 Contiguity

Suppose we are given two probability measures P and Q on a measurable space $(\Omega, \mathcal{U})$, with densities p and q relative to some measure μ. We denote by dQ/dP the ratio q/p, which is with P-probability one well-defined and not depending on μ. In fact, it is a density of the absolutely continuous part of Q relative to P. (Note that we do no write dQ^a/dP and we do not assume that $Q \ll P$.) Let $X: \Omega \mapsto \mathbb{D}$ be a measurable map in a metric space. Then $(X, dQ/dP)$ is a measurable map into $\mathbb{D} \times \mathbb{R}$, and it induces a law L on this space if we equip $(\Omega, \mathcal{U})$ with P. If $Q \ll P$, then this law determines the law of X under Q, because in this case

$$Q(X \in B) = \mathrm{E}_P 1_B(X) \frac{dQ}{dP} = \int_{B \times \mathbb{R}} v \, dL(x, v).$$

The validity of this formula depends essentially on the absolute continuity of Q with respect to P, because a part of Q that is orthogonal with respect to P cannot be recovered from any P-law.

Consider an asymptotic version of the problem. Let $(\Omega_n, \mathcal{A}_n)$ be measurable spaces, each equipped with a pair of probability measures P_n and Q_n. Under what conditions can a Q_n-limit law of random vectors $X_n: \Omega_n \mapsto \mathbb{R}^k$ be obtained from suitable P_n-limit laws? In view of the above it is necessary that Q_n is "asymptotically absolutely continuous" with respect to P_n in a suitable sense. The right concept is contiguity.

4.1 Definition. The sequence Q_n is *contiguous* with respect to the sequence P_n if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ for every sequence of measurable sets A_n. This is denoted $Q_n \lhd P_n$. The sequences P_n and Q_n are *mutually contiguous* if both $P_n \lhd Q_n$ and $Q_n \lhd P_n$. This is denoted $P_n \lhd \rhd Q_n$.

The name "contiguous" is standard, but perhaps conveys a wrong image. "Contiguity" suggests sequences of probability measures living next to each other, while the correct image is "on top of each other" (in the limit).

Before answering the question of interest, we give two characterizations of contiguity in terms of the asymptotic behaviour of the likelihood ratios of P_n and Q_n. The likelihood ratios dQ_n/dP_n and dP_n/dQ_n are nonnegative and satisfy

$$\mathrm{E}_{P_n} \frac{dQ_n}{dP_n} \leq 1 \qquad \text{and} \qquad \mathrm{E}_{Q_n} \frac{dP_n}{dQ_n} \leq 1.$$

Thus, the sequences of likelihood ratios dQ_n/dP_n and dP_n/dQ_n are uniformly tight under P_n and Q_n, respectively. By Prohorov's theorem, every subsequence has a further weakly converging subsequence. The next lemma shows that the properties of the limit points determine contiguity.

4.2 Lemma (Le Cam's first lemma). *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Then the following statements are equivalent:*

(i) $Q_n \lhd P_n$;

(ii) if $dP_n/dQ_n \overset{Q_n}{\rightsquigarrow} U$ *along a subsequence, then* $\mathrm{P}(U > 0) = 1$;

(iii) if $dQ_n/dP_n \overset{P_n}{\rightsquigarrow} V$ *along a subsequence, then* $\mathrm{E}V = 1$;

(iv) for any statistics $T_n : \Omega_n \mapsto \mathbb{R}^k$*: if* $T_n \overset{P_n}{\to} 0$, *then* $T_n \overset{Q_n}{\to} 0$.

We do not include a proof of this lemma, but note that the lemma is easy if the sequences P_n and Q_n are constant. If $(\Omega_n, \mathcal{U}_n) = (\Omega, \mathcal{U})$, $P_n = P$ and $Q_n = Q$ for each n, then contiguity is equivalent to absolute continuity, and the lemma reduces to the equivalence of the three statements

$$Q \ll P, \qquad Q\Big(\frac{dP}{dQ} = 0\Big) = 0, \qquad \mathrm{E}_P \frac{dQ}{dP} = 1.$$

The lemma shows that these equivalences persist if the three statements are replaced by their asymptotic counterparts.

According to Lemma 1.10 the likelihood ratios of the measures $P^n_{1/\sqrt{n}}$ and P^n for a given differentiable path $t \mapsto P_t$ are asymptotically log-normally distributed with mean $-\frac{1}{2}Pg^2$ and variance Pg^2. This makes these sequences of measures mutually contiguous.

4.3 Example (Asymptotic log normality). Let P_n and Q_n be probability measures on arbitrary measurable spaces such that

$$\frac{dP_n}{dQ_n} \overset{Q_n}{\rightsquigarrow} e^{N(\mu,\sigma^2)}.$$

Then $Q_n \lhd P_n$. Furthermore, $Q_n \lhd \rhd P_n$ if and only if $\mu = -\frac{1}{2}\sigma^2$.

Since the (log normal) variable on the right is positive, the first assertion is immediate from (ii) of the theorem. The second follows from (iii) with the roles of P_n and Q_n switched, upon noting that $\mathrm{E}\exp N(\mu,\sigma^2) = 1$ if and only if $\mu = -\frac{1}{2}\sigma^2$. □

The following theorem solves the problem of obtaining a Q_n-limit law from a P_n-limit law that we posed in the introduction. The result, a version of *Le Cam's third lemma*, is in perfect analogy with the nonasymptotic situation.

4.4 Theorem. *Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$, and let $X_n\colon \Omega_n \mapsto \mathbb{R}^k$ be a sequence of maps. Suppose that $Q_n \triangleleft P_n$ and*

$$\left(X_n, \frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} (X, V).$$

Then $L(B) = \mathrm{E}1_B(X)\,V$ defines a probability measure, and $X_n \overset{Q_n}{\rightsquigarrow} L$.

Proof. Since $V \geq 0$, it follows with the help of the monotone convergence theorem that L defines a measure. By contiguity, $\mathrm{E}V = 1$ and hence L is a probability measure. It is immediate from the definition of L that $\int f\,dL = \mathrm{E}f(X)\,V$ for every measurable indicator function f. Conclude, in steps, that the same is true for every simple function f, any nonnegative measurable function, and every integrable function.

If f is continuous and nonnegative, then so is the function $(x, v) \mapsto f(x)\,v$ on $\mathbb{R}^k \times [0,\infty)$. Thus

$$\liminf \mathrm{E}_{Q_n,*} f(X_n) \geq \liminf \int_* f(X_n)\,\frac{dQ_n}{dP_n}\,dP_n \geq \mathrm{E}f(X)V,$$

by the portmanteau lemma. Apply the portmanteau lemma in the converse direction to conclude the proof that $X_n \overset{Q_n}{\rightsquigarrow} L$. ■

4.5 Example (Le Cam's third lemma). The name *Le Cam's third lemma* is often reserved for the following result. If

$$\left(X_n, \log\frac{dQ_n}{dP_n}\right) \overset{P_n}{\rightsquigarrow} N_{k+1}\left(\begin{pmatrix}\mu\\ -\frac{1}{2}\sigma^2\end{pmatrix}, \begin{pmatrix}\Sigma & \tau\\ \tau^T & \sigma^2\end{pmatrix}\right),$$

then

$$X_n \overset{Q_n}{\rightsquigarrow} N_k(\mu + \tau, \Sigma).$$

In this situation the asymptotic covariance matrices of the sequence X_n are the same under P_n and Q_n, but the mean vectors differ by the asymptotic covariance τ between X_n and the log likelihood ratios.

The statement is a special case of the preceding theorem. Let (X, W) have the given $(k+1)$-dimensional normal distribution. By the continuous mapping theorem, the sequence $(X_n, dQ_n/dP_n)$ converges in distribution under P_n to (X, e^W). Since W is $N(-\frac{1}{2}\sigma^2, \sigma^2)$-distributed, the sequences P_n and Q_n are mutually contiguous. According to the abstract version of Le Cam's third lemma, $X_n \overset{Q_n}{\rightsquigarrow} L$ with $L(B) = \mathrm{E}1_B(X)e^W$. The characteristic function of L is $\int e^{it^Tx}\,dL(x) = \mathrm{E}e^{it^TX}\,e^W$. This is

the characteristic function of the given normal distribution at the vector $(t, -i)$. Thus

$$\int e^{it^Tx}\,dL(x) = e^{it^T\mu - \frac{1}{2}\sigma^2 - \frac{1}{2}(t^T,-i)\begin{pmatrix}\Sigma & \tau\\ \tau^T & \sigma^2\end{pmatrix}\begin{pmatrix}t\\ -i\end{pmatrix}} = e^{it^T(\mu+\tau)-\frac{1}{2}t^T\Sigma t}.$$

The right side is the characteristic function of the $N_k(\mu+\tau,\Sigma)$ distribution. □

4.6 Example. Let $t \mapsto P_t$ be a differentiable path with score function g and let $T_n = T_n(X_1,\ldots,X_n)$ be statistics such that

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

for a function h with $Ph = 0$ and $Ph^2 < \infty$. Then the sequence $\sqrt{n}\big(T_n - \psi(P)\big)$ converges in distribution to a normal measure with mean Pgh and variance Ph^2 under $P^n_{1/\sqrt{n}}$.

Consequently, if ψ is differentiable at P, then $\sqrt{n}\big(T_n - \psi(P_{1/\sqrt{n}})\big)$ converges in distribution to a normal measure with mean $Pg(h-\tilde\psi_P)$ and variance Ph^2. It follows that T_n is a regular estimator sequence if only if $h-\tilde\psi_P$ is orthogonal to the tangent set. In other words if and only if h is an influence function of ψ. □

4.2 Gaussian Representations

Let H be a subset of a Hilbert space with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. For each $n \in \mathbb{N}$ and $h \in H$, let $P_{n,h}$ be a probability measure on a measurable space $(\mathcal{X}_n, \mathcal{A}_n)$. Consider the problem of estimating a "parameter" $\kappa_n(h)$ given an "observation" X_n with law $P_{n,h}$.

For ease of notation, let $\{\Delta_h : h \in H\}$ be the "iso-Gaussian process" with zero mean and covariance function $\mathrm{E}\Delta_{h_1}\Delta_{h_2} = \langle h_1, h_2\rangle$. The sequence of experiments $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ is called *asymptotically (shift) normal* if

$$\log\frac{dP_{n,h}}{dP_{n,0}} = \Delta_{n,h} - \tfrac{1}{2}\|h\|^2,$$

for stochastic processes $\{\Delta_{n,h} : h \in H\}$ such that

$$\Delta_{n,h} \overset{0}{\rightsquigarrow} \Delta_h \qquad \text{marginally.}$$

Here $\overset{h}{\rightsquigarrow}$ denotes weak convergence under $P_{n,h}$. This terminology arises from the theory of limiting experiments due to Le Cam.

The sequence of parameters $\kappa_n(h)$ is assumed to belong to a Banach space $\mathbb{D}$. We assume that it is asymptotically differentiable in the sense that

$$r_n\big(\kappa_n(h) - \kappa_n(0)\big) \to \dot\kappa(h), \qquad \text{for every } h \in H,$$

for a continuous, linear map $\dot\kappa\colon \operatorname{lin} H \mapsto \mathbb{D}$ and certain linear maps $r_n\colon \mathbb{D} \mapsto \mathbb{D}$ ("norming operators"). Any maps $T_n\colon \mathcal{X}_n \mapsto \mathbb{D}$ are considered estimators of the parameter.

4.7 Example (I.i.d. observations). To cover the situation of Lectures 1–3, let $\mathcal{X}_n = \mathcal{X}^n$, $\mathcal{A}_n = \mathcal{A}^n$ and $P_{n,h} = P_{1/\sqrt{n},h}$ for a differentiable path with score h. Furthermore, let $\kappa_n(h) = \psi(P_{1/\sqrt{n},h})$. Then differentiability of ψ implies the asymptotic differentiability of κ_n relative to the norming rate $r_n = \sqrt{n}$, with derivative $\dot\kappa = \dot\psi_P$. The asymptotic normality of the experiments $(P^n_{1/\sqrt{n},h}\colon h \in \dot{\mathcal{P}}_P)$ follows from Lemma 1.10, where we may take $H = \dot{\mathcal{P}}_P$, contained in the Hilbert space $L_2(P)$. In all these definitions the measure P is fixed (and considered statistically "known"), and h is an unknown parameter, known to belong to the tangent set. □

4.8 Theorem (Gaussian Representation). *Suppose that $T_n\colon \mathcal{X}_n \mapsto \mathbb{D}$ are statistics with values in a Banach space $\mathbb{D}$ such that, for every $h \in H$ and tight probability measures L_h,*

$$\sqrt{n}\big(T_n - \kappa_n(h)\big) \overset{h}{\rightsquigarrow} L_h.$$

Assume that the parameters κ_n are asymptotically differentiable. Then for any orthonormal sequence $h_1, \ldots, h_m$ in $\overline{\operatorname{lin}} H$ there exists a measurable map $T\colon \mathbb{R}^m \times [0,1] \mapsto \mathbb{D}$ such that $T - \dot\kappa(h)$ is distributed as L_h if the law of T is calculated under the product of the normal measure with mean $\big(\langle h, h_1\rangle, \ldots, \langle h, h_m\rangle\big)$ and covariance the identity and the uniform measure on $[0,1]$.

Proof. By an easy calculation we see that the random variable $\Delta_{\Sigma a_i h_i} - \sum a_i \Delta_{h_i}$ has second moment zero. Hence the process $h \mapsto \Delta_{n,h}$ is linear, in an almost sure sense. From this we conclude that the sequence $\Delta_{n,\Sigma a_i h_i} - \sum a_i \Delta_{n,h_i}$ converges to zero in probability under $P_{n,0}$. Thus the sequence $h \mapsto \Delta_{n,h}$ is asymptotically linear.

Next define variables

$$Z_{n,h} = r_n\big(T_n - \kappa_n(h)\big),$$
$$\Lambda_n(h) = \log \frac{dP_{n,h}}{dP_{n,0}} = \Delta_{n,h} - \tfrac{1}{2}\|h\|^2, \qquad \text{for } h = \textstyle\sum a_i h_i.$$

By assumption, the sequence $Z_{n,0}$ and each sequence $\Delta_{n,h}$ converge in distribution in $\mathbb{D}$ and $\overline{\mathbb{R}}$, respectively. By Prohorov's theorem, there exists a subsequence of $\{n\}$ such that

$$\big(Z_{n',0}, \Delta_{n',h_1}, \ldots, \Delta_{n',h_k}\big) \overset{0}{\rightsquigarrow} (Z, \Delta_{h_1}, \ldots, \Delta_{h_k}),$$

in $\mathbb{D} \times \mathbb{R}^k$, where the random vector on the right can be defined on a suitable probability space and has marginal distributions L_0 and the standard normal distribution on $\mathbb{R}^k$, respectively. In view of the asymptotic linearity of the processes $h \mapsto \Delta_{n,h}$ and the asymptotic differentiability of the sequence of parameters we obtain, for every $h \in H$,

$$\big(Z_{n',h}, \Lambda_{n'}(h)\big) \overset{0}{\rightsquigarrow} \Big(Z - \dot\kappa(h), \Delta_h - \tfrac{1}{2}\|h\|^2\Big).$$

Next we can apply Le Cam's third lemma to see that there exist variables Z_h, defined on some probability space, such that $Z_{n',h} \overset{a}{\rightsquigarrow} Z_h$, where $\overset{a}{\rightsquigarrow}$ denotes weak convergence under $P_{n,h}$ for $h = \sum a_i h_i$, and Z_h is distributed according to

$$\mathrm{P}(Z_h \in B) = \mathrm{E}1_B\Big(Z - \sum a_i \dot\kappa(h_i)\Big)\, e^{\sum a_i \Delta_{h_i} - \frac{1}{2}\|a\|^2}. \tag{4.9}$$

By assumption the weak limit Z_h is distributed according to L_h.

We are ready to construct an appropriate randomized estimator T. For ease of notation let X_0 and U be independent variables with a standard normal distribution on $\mathbb{R}^k$ and the uniform distribution, respectively. Suppose that T is such that $\big(T(X_0,U), X_0\big)$ is distributed as $(Z, \Delta_{h_1}, \ldots, \Delta_{h_k})$. Then, if X is normally distributed with mean vector $\mu(h) = \big(\langle h,h_1\rangle, \ldots, \langle h, h_m\rangle\big)$ and covariance the identity and independent of U, we have

$$\begin{aligned}\mathrm{P}_h\big(T(X,U) \in B\big) &= \mathrm{E}_h 1_B\big(T(X,U)\big) = \mathrm{E}_0 1_B\big(T(X,U)\big) e^{\mu(h)^T X - \frac{1}{2}X^T X}\\ &= L_h(B),\end{aligned}$$

because (X,U) is distributed as (X_0,U) under $h = 0$ and hence $\big(T(X,U),X\big)$ is distributed as $(Z, \Delta_{h_1}, \ldots, \Delta_{h_k})$.

To conclude the proof it suffices to construct T as in the preceding paragraph. Because the second marginal distributions of the vectors $\big(T(X_0,U),X_0\big)$ and $(Z, \Delta_{h_1}, \ldots, \Delta_{h_k})$ are identical, it suffices to construct T such that the conditional distributions of the first marginal given the second marginal are identical. This is the case if for each x_0 the variable $T(x_0,U)$ is distributed according the conditional law of Z given $(\Delta_{h_1}, \ldots, \Delta_{h_k}) = x_0$. This is the problem of generating a variable with an arbitrary distribution on a Polish space from a uniform variable. It is well-known that this possible. One construction is to map the Polish space bimeasurably onto the real line, and next use the quantile transformation to construct the induced law. ∎

The preceding theorem is restricted to finite-dimensional models. As we remarked before an extension to infinite-dimensional Hilbert spaces H is possible, but maybe not very useful, because it is hard to analyse infinite-dimensional Gaussian experiments directly, without finite-dimensional approximation. For completeness we include an infinite-dimensional version.

4.10 Theorem. *Suppose that $T_n\colon \mathcal{X}_n \mapsto \mathbb{D}$ are statistics with values in a Banach space $\mathbb{D}$ such that, for every $h \in H$ and a tight probability measure L_h,*

$$\sqrt{n}\big(T_n - \kappa_n(h)\big) \overset{h}{\rightsquigarrow} L_h.$$

Assume that the parameters κ_n are asymptotically differentiable. Then for any orthonormal sequence $h_1, h_2, \ldots$, in $\overline{\mathrm{lin}}H$ there exists a measurable map $T\colon \mathbb{R}^\infty \times [0,1] \mapsto \mathbb{D}$ such that $T - \dot\kappa(h)$ is distributed as L_h if the law of T is calculated under the product of the normal measure with mean $\big(\langle h,h_1\rangle, \langle h,h_2\rangle, \ldots\big)$ and covariance the identity and the uniform measure on $[0,1]$.

The infinite-dimensional normal measure in the theorem is simply the distribution of a sequence $Z = (Z_1, Z_2, \ldots)$ of independent normal variables Z_i with means $\langle h, h_i \rangle$ and unit variances. Actually, the Gaussian experiment could be represented in many different forms, the present one probably being the simplest one. For instance, the theorem is also true if Z is replaced by $f(Z)$ for an arbitrary bimeasurable map f from $\mathbb{R}^\infty$ onto a measurable space (e.g. the unit interval).

The preceding theorems show that estimator sequences in the sequence of experiments $(P_{n,h}: h \in H)$ are asymptotically matched by an estimator in a Gaussian experiment. The next step is to analyse the Gaussian experiment. In our abstract set-up the "optimal" measure can be defined in terms of the adjoint $\dot\kappa^*: \mathbb{D}^* \mapsto \overline{\text{lin}}\, H$ of the asymptotic derivative of the parameters κ_n, which maps the dual space of $\mathbb{D}$ into the closed linear span of H. This is determined by the equation $\langle \dot\kappa^* b^*, h\rangle = b^* \dot\kappa(h)$.

The optimal Gaussian measure can be uniquely determined by its marginal distributions: its induced laws under continuous, linear maps $d^*: \mathbb{D} \mapsto \mathbb{R}$. It can be represented as the distribution of a Borel measurable random element G in $\mathbb{D}$ such that d^*G is $N\big(0, \|\dot\kappa^* d^*\|^2\big)$ distributed, for any element d^* from the dual space $\mathbb{D}^*$. In the case that the Banach space $\mathbb{D}$ is infinite-dimensional such a measure does not necessarily exist, but the the theorem below shows that it does exist when we need it: if there exist good estimator sequences.

4.11 Example. Consider the case of i.i.d. observations as considered in Lecture 2, with $\dot\kappa = \dot\psi_P$, $H = \dot\psi_P$ equipped with the $L_2(P)$-norm, and ψ taking values in $\mathbb{D} = \mathbb{R}^k$. Then κ is representable as a vector-valued inner product $\dot\psi_P(h) = Ph\tilde\psi_P$, $\mathbb{D}^* = \mathbb{R}^k$, and its adjoint is the map $\dot\kappa^* a = a^T \tilde\psi_P$, because

$$P(\dot\kappa^* a)h = \langle a, \dot\kappa h\rangle = a^T Ph\tilde\psi_P = P(a^T\tilde\psi_P)h.$$

It follows that the optimal limit measure is the distribution of a vector G such that $a^T G$ is normally distributed with mean 0 and variance $\|a^T\tilde\psi_P\|^2 = P(a^T\tilde\psi_P)^2$. This agrees with the optimal normal measure found in Lecture 2. □

In our present set-up we call a sequence of estimators T_n *regular* with respect to the norming operators r_n if

$$r_n\big(T_n - \kappa_n(h)\big) \overset{h}{\rightsquigarrow} L, \qquad \text{for every } h \in H,$$

for a fixed, tight, Borel probability measure L on $\mathbb{D}$.

4.12 Theorem (Convolution). *Assume that the parameters κ_n are asymptotically differentiable.*

(i) If there exists a sequence of regular estimators for κ_n, then there exists a tight, Borel measurable variable G in $\mathbb{D}$ such that

$$d^*G \sim N\big(0, \|\dot\kappa^* d^*\|^2\big), \qquad \text{for every } d^* \in \mathbb{D}^*.$$

(ii) The limit law L of every regular sequence of estimators can be represented as the distribution of a sum $G + W$ of independent, tight, Borel measurable variables in $\mathbb{D}$ with G as distributed in (i).

Proof. (a). Assume that H is a finite-dimensional, linear space and let $h_1, \ldots, h_m$ be an orthonormal base. Then the assumptions of Theorem 4.8 are satisfied and we obtain that L is the distribution of $T - \dot\kappa(h)$ under every h. As shown in the proof of this theorem (see (4.9)), this means that, for every $a \in \mathbb{R}^k$,

$$L(B) = \mathrm{E}1_B\Big(Z - \sum a_i \dot\kappa(h_i)\Big)\, e^{\sum a_i \Delta_{h_i} - \frac{1}{2}\|a\|^2}.$$

We average this equation over a with respect to a $N_k(0, \lambda^{-1} I)$ weight function. Straightforward calculations yield

$$L(B) = \int \mathrm{E}1_B\left(Z - \frac{\sum \Delta_{h_i}\dot\kappa(h_i)}{1+\lambda} - \frac{\sum a_i \dot\kappa(h_i)}{(1+\lambda)^{1/2}}\right) c_\lambda(\Delta)\, dN_k\big(0, I\big)(a),$$

where $c_\lambda(\Delta) = (1+\lambda^{-1})^{k/2} \exp\big(\frac{1}{2}(1+\lambda)^{-1}\sum \Delta_{h_i}^2\big)$. Conclude that L can be written as the law of the sum $G_\lambda + W_\lambda$ of independent random elements G_λ and W_λ, where $G_\lambda = -\sum A_i \dot\kappa(h_i)/(1+\lambda)^{1/2}$ for a $N_k(0, I)$-distributed vector $(A_1, \ldots, A_k)$ and W_λ is distributed according to

$$\mathrm{P}(W_\lambda \in B) = \mathrm{E}1_B\left(Z - \frac{\sum \Delta_{h_i}\dot\kappa(h_i)}{1+\lambda}\right) c_\lambda(\Delta).$$

As $\lambda \downarrow 0$, we have $G_\lambda \rightsquigarrow G = \sum A_i \dot\kappa(h_i)$. The variable $d^*G = \sum A_i d^*\dot\kappa(h_i)$ is normally distributed with zero mean and variance

$$\mathrm{E}d^*G_\lambda^2 = \sum \big(d^*\dot\kappa(h_i)\big)^2 = \|\dot\kappa^* d^*\|^2.$$

By the converse part of Prohorov's theorem, the variables $\{G_\lambda : 0 < \lambda < 1\}$ are uniformly tight. Combined with the tightness of L it follows that there exists, for every $\varepsilon > 0$, a compact K such that

$$1 - \varepsilon < L(K) = \int \mathrm{P}(W_\lambda + g)\, dP^{G_\lambda}(g), \qquad \text{and} \qquad \mathrm{P}(G_\lambda \in K) > 1 - \varepsilon.$$

This implies that for every λ there exists $g_\lambda \in K$ such that $\mathrm{P}(W_\lambda + g_\lambda \in K) > 1 - 2\varepsilon$ and hence $\mathrm{P}(W_\lambda \in K - K) > 1 - 2\varepsilon)$. We conclude that set of the variables $\{W_\lambda : 0 < \lambda < 1\}$ is uniformly tight.

If $W_{\lambda_m} \rightsquigarrow W$ for a sequence $\lambda_m \downarrow 0$, then $(G_{\lambda_m}, W_{\lambda_m}) \rightsquigarrow (G, W)$, where G and W are independent and $G + W$ is distributed according to L. This concludes the proof of the theorem for finite-dimensional H.

(b) Let H be arbitrary. For any finite orthonormal set $h_1, \ldots, h_k$, the previous argument yields tight independent processes G_k and W_k such that $G_k + W_k$ is distributed according to L and G_k is zero-mean Gaussian with

$$\mathrm{E}d^*G_k^2 = \sum_{i=1}^k \langle \dot\kappa^* d^*, h_i\rangle^2.$$

The set of all variables G_k and W_k so obtained is uniformly tight. Indeed, by tightness of L, there exists for any given $\varepsilon > 0$ a compact set K such that $L(K) = \int \mathrm{P}(G_k \in K - x)\, dP^{W_k}(x) > 1 - \varepsilon$. Thus there exists x_0 with $\mathrm{P}(G_k \in K - x_0) > 1 - \varepsilon$. By symmetry, $\mathrm{P}(G_k \in x_0 - K) > 1 - \varepsilon$, whence $\mathrm{P}\big(G_k \in \frac{1}{2}(K - K)\big) > 1 - 2\varepsilon$. Next,

the uniform tightness of L and the collection G_k imply the uniform tightness of the collection W_k.

Direct the finite-dimensional subspaces of H by inclusion, and construct variables (G_k, W_k) for every subspace. Every weak limit point (G, W) of the net of laws $(G_{k'}, W_{k'})$ satisfies the requirements of the theorem. ∎

In the following minimax theorem we show that the maximum risk

$$\sup_h \mathrm{E}_{h*}\ell\big(r_n(T_n - \kappa_n(h))\big)$$

of an estimator sequence can never asymptotically fall below $\mathrm{E}\ell(G)$. A little (asymptotic) measurability is the only requirement on T_n, but measurability can be restrictive, so we shall be careful about it. Let $\mathbb{D}'$ be a given subspace of $\mathbb{D}^*$ that separates points of $\mathbb{D}$, and let $\tau(\mathbb{D}')$ be the weak topology induced on $\mathbb{D}$ by the maps $b'\colon \mathbb{D} \mapsto \mathbb{R}$ when b' ranges over $\mathbb{D}'$.

4.13 Definition. A map $\ell\colon \mathbb{D} \mapsto \mathbb{R}$ is called $\tau(\mathbb{D}')$-*subconvex* if for every $c > 0$ the set $\big\{y\colon \ell(y) \le c\big\}$ is $\tau(\mathbb{D}')$-closed, convex, and symmetric.

4.14 Theorem (Minimax theorem). *Assume that the parameters κ_n are asymptotically differentiable. Suppose a tight, Borel measurable Gaussian element G as in (i) of the statement of the convolution theorem exists. Then for every estimator sequence T_n such that $d'T_n\colon \mathcal{X}_n \mapsto \mathbb{R}$ is measurable for every $d' \in \mathbb{D}'$ and every $\tau(\mathbb{D}')$-subconvex function $\ell\colon \mathbb{D} \mapsto \mathbb{R}$,*

$$\sup_{I \subset H} \liminf_{n\to\infty} \sup_{h\in I} \mathrm{E}_{h*}\ell\Big(r_n\big(T_n - \kappa_n(h)\big)\Big) \ge \mathrm{E}\ell(G).$$

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

Proof. In a general sense the proof is based on an analysis of the minimax risk in the Gaussian representation provided by Theorem 4.8. The main work is to force our estimator sequence to have limit laws, so that the theorem becomes applicable. This is achieved by compactification of the range space of T_n, so that limit laws exist at least along subsequences and with limits concentrating on the compactification, by Prohorov's theorem. Because it will be necessary to extend the loss function to the compactification, the compactification must be chosen dependent on the loss function. Therefore the proof proceeds in several steps, building more complicated loss functions from simple ones.

(a). Assume first that the loss function can be written in the special form $\ell(y) = \sum_{i=1}^r 1_{K_i^c}\big(d'_{i,1}y, \ldots, d'_{i,p_i}y\big)$ for compact, convex, symmetric subsets $K_i \subset \mathbb{R}^{p_i}$ and arbitrary elements $d'_{i,j}$ of $\mathbb{D}'$. Fix an arbitrary orthonormal set $h_1, \ldots, h_k$ in H, and set

$$Z^i_{n,a} = \big(d'_{i,1}, \ldots, d_{i,p_i}\big) \circ r_n\big(T_n - \kappa_n(\textstyle\sum a_i h_i)\big), \qquad 1 \le i \le r.$$

Considered as maps into the one-point compactification of $\mathbb{R}^{p_i}$, the sequences $Z^i_{n,a}$ are certainly asymptotically tight. The sequences are asymptotically measurable by assumption.

Direct the finite subsets of H by inclusion. There exists a subnet $\{n_I : I \subset H, \text{finite}\}$ such that the left side of the statement of the theorem equals

$$\text{minimax risk} = \lim_I \sup_{h \in I} \mathrm{E}_{h*} \ell\Big(r_n\big(T_n - \kappa_n(h)\big)\Big).$$

By the same arguments as in the proof of the convolution theorem there is a further subnet $\{n'\} \subset \{n_I\}$ such that $Z^i_{n',a} \overset{a}{\rightsquigarrow} Z^i_a$ in the one-point compactifications, for every $a \in \mathbb{R}^k$ and every i. Here the limiting processes satisfy, for each i,

$$\int \mathcal{L}(Z^i_a)\, dN_k(0, \lambda^{-1} I) \sim G^i_\lambda + W^i_\lambda, \tag{4.15}$$

for independent elements G^i_λ and W^i_λ such that

$$G^i_\lambda = \big(d'_{i,1}, \ldots, d_{i,p_i}\big) \circ G_\lambda = \big(d'_{i,1}, \ldots, d_{i,p_i}\big) \circ \frac{\sum A_i \dot\kappa(h_i)}{(1+\lambda)^{1/2}},$$

for a $N_k(0, I)$-distributed vector $(A_1, \ldots, A_k)$. By the portmanteau theorem,

$$\text{minimax risk} \geq \liminf_{n'} \sum_{i=1}^r \mathrm{P}_{a*}(Z^i_{n',a} \notin K_i) \geq \sum_{i=1}^r \mathrm{P}(Z^i_a \notin K_i).$$

Since this is true for every a, the left side is also bounded below by the average of the right side, in particular the average under the $N_k(0, \lambda^{-1} I)$-distribution. In view of (4.15) we find that

$$\text{minimax risk} \geq \sum_{i=1}^r \mathrm{P}(G^i_\lambda + W^i_\lambda \notin K_i).$$

The right side becomes smaller if we replace the variables W^i_λ by 0. This follows by Anderson's lemma, according to which, given a mean zero Gaussian vector and a convex, symmetric set K, the probability $\mathrm{P}(G + a \in K)$ is maximized over a at $a = 0$, i.e. centering the Gaussian variable $G + a$ at zero. Thus the right side of the preceding display is bounded below by

$$\sum_{i=1}^r \mathrm{P}(G^i_\lambda \notin K_i) = \mathrm{E}\ell(G_\lambda).$$

We finish the proof for this special form of loss function by letting $\lambda \downarrow 0$ followed by taking the limit along finite-dimensional subspaces of H.

(b). The theorem is "closed" under taking monotone limits on ℓ: if the theorem holds for every function ℓ_r and $0 \leq \ell_r \leq \ell$ with $\ell_r \uparrow \ell$ almost surely under the law of G, then the theorem holds for ℓ. To see this, note that the minimax risk decreases by replacing ℓ by ℓ_r. Thus it is bounded below by $\mathrm{E}\ell_r(G)$ for every r, which increases to $\mathrm{E}\ell(G)$ as $r \to \infty$.

(c). An arbitrary subconvex ℓ can be approximated from below by a sequence of functions ℓ_r of the type as in (a). To see this, note first that

$$0 \leq 2^{-r} \sum_{i=1}^{2^{2r}} 1\big\{y : \ell(y) > i2^{-r}\big\} \uparrow \ell(y), \qquad \text{for every } \ y.$$

Each of the sets $\{y: \ell(y) > i/r\}$ is convex, $\tau(\mathbb{D}')$-closed, and symmetric. Thus, it suffices to approximate functions ℓ of the type 1_{C^c} for a convex, $\tau(\mathbb{D}')$-closed, and symmetric set C.

By the Hahn-Banach theorem, any such set C can be written

$$C = \bigcap_{b' \in \mathbb{D}'} \{y: |b'y| \le c_{b'}\}.$$

Thus the complement of C intersected with the support S of the limit variable G is the union of the sets $\{y \in S: |b'y| > c_{b'}\}$. These sets are relatively open in S and S is separable. Since a separable set is Lindelöf, the possibly uncountable union can be replaced by a countable subunion. Thus there exists a sequence d_i' in $\mathbb{D}'$ and numbers c_i such that $C^c \cap S = \cup_{i=1}^{\infty} \{y \in S: |d_i'y| > c_i\}$. This implies that

$$1_{C^c \cap S} = \sup_r 1_{K_r^c}(d_1'y, \ldots, d_r'y),$$

for the subsets of $\mathbb{R}^r$ defined by $K_r = \cap_{i=1}^r \{x \in \mathbb{R}^r : |x_i| \le c_i\}$. ∎

4.16 Example. For $\mathbb{D}' = \mathbb{D}^*$, the $\tau(\mathbb{D}')$-topology is the weak topology. Because convex subsets in a Banach space are weakly closed if and only if they are closed for the norm, a function which is subconvex relative to the norm is automatically $\tau(\mathbb{D}^*)$-subconvex. The theorem is applicable to the combination of such loss functions and estimator sequences T_n that are weakly measurable: d^*T_n should be a measurable map in $\mathbb{R}$ for every $d^* \in \mathbb{D}^*$.

This will typically be the case if the Banach space is separable, when estimators will usually be required to be Borel measurable. □

4.17 Example (Skorohod space). The Skorohod space $D[a, b]$, for a given interval $[a, b] \subset \overline{\mathbb{R}}$, is a Banach space if equipped with the uniform norm. The dual space consists of maps of the form

$$d^*(z) = \int z(u)\, d\mu(u) + \sum_{i=1}^{\infty} \alpha_i \big(z(u_i) - z(u_i-)\big),$$

for a finite signed measure μ on $[a, b]$, an arbitrary sequence u_i in $(a, b]$, and a sequence α_i with $\sum |\alpha_i| < \infty$. (This is an extension of the representation theorem for the dual space of the space of continuous functions on a compact due to Riesz, obtained in [36], pages 81–85.) Each such d^* is the pointwise limit of a sequence of linear combinations of coordinate projections. Thus, the σ-field generated by the dual space equals the σ-field generated by the coordinate projections.

It follows that an estimator sequence is $D[a, b]^*$-measurable if and only if it is a stochastic process. Since "$\tau(D[a, b]^*)$-subconvex" is identical to "subconvex with respect to the norm", the minimax theorem is valid for any sequence of stochastic processes T_n and subconvex loss function ℓ.

Examples of subconvex loss functions include

$$z \mapsto \ell_0\big(\|z\|_\infty\big),$$
$$z \mapsto \int |z|^p(t)\, d\mu(t),$$

for a nondecreasing, left-continuous function $\ell_0: \mathbb{R} \mapsto \mathbb{R}$, a finite Borel measure μ, and $p \ge 1$. □

4.18 Example (Bounded functions). On the space $\ell^\infty(\mathcal{F})$, functions of the type

$$z \mapsto \ell_0\Big(\Big\|\frac{z}{q}\Big\|_{\mathcal{F}}\Big),$$

for a nondecreasing, left-continuous function $\ell_0\colon \mathbb{R} \mapsto \mathbb{R}$ and an arbitrary map $q\colon \mathcal{F} \mapsto \mathbb{R}$ are subconvex with respect to the linear space spanned by the coordinate projections $z \mapsto z(f)$. Indeed, for any c there exists d such that

$$\Big\{z\colon \ell_0\Big(\Big\|\frac{z}{q}\Big\|_{\mathcal{F}}\Big) \le c\Big\} = \Big\{z\colon \Big\|\frac{z}{q}\Big\|_{\mathcal{F}} \le d\Big\} = \bigcap_{f\in\mathcal{F}} \Big\{z\colon \big|z(f)\big| \le d\,q(f)\Big\}.$$

Thus, the minimax theorem is valid for any estimator sequence T_n that is coordinatewise measurable and any loss function of this type.

For general loss functions that are subconvex with respect to the norm, the preceding minimax theorem applies only under strong measurability conditions on the estimator sequences. It is of interest that these measurability conditions are satisfied by sequences T_n such that $T_n(f)$ is measurable for every f and such that the sequence $r_n\big(T_n - \kappa_n(0)\big)$ is asymptotically tight under $P_{n,0}$. Indeed, such sequences are asymptotically $\tau(\ell^\infty(\mathcal{F})^*)$-measurable. It follows that, given any subconvex loss function, the minimax theorem may be used to designate optimal estimator sequences among the asymptotically tight sequences. □

Finally, consider the testing problem. The Gaussian representation theorem given previously was meant to be applied to the estimation problem, but we can easily transform it into a theorem on tests by taking $\kappa_n(h) \equiv 0$.

4.19 Theorem (Gaussian Representation). *Let $\phi_n\colon \mathcal{X}_n \mapsto [0,1]$ be arbitrary statistics such that, for every $h \in H$ and some function $\pi\colon H \mapsto \mathbb{R}$,*

$$P_{n,h}\phi_n \to \pi(h).$$

Then for any orthonormal sequence $h_1, \ldots, h_m$ in $\overline{\text{lin}}\, H$ there exists a measurable map $\phi\colon \mathbb{R}^m \mapsto [0,1]$ such that $P_h\phi = \pi(h)$ for every $h \in H$, where P_h is the normal measure with mean $\big(\langle h, h_1\rangle, \ldots, \langle h, h_m\rangle\big)$ and covariance the identity.

Proof. Because the unit interval is compact we can extract a subsequence of ϕ_n that converges in distribution under $P_{n,h}$ to a limit law L_h. By contiguity arguments, using Le Cam's third lemma as in the proofs of the preceding theorems, we can even find a subsequence that works for all h in the linear span of $h_1, \ldots, h_m$. We next apply Theorem 4.8 to the corresponding subsequence of $\kappa_n \equiv 0$ and $T_n = n^{-1/2}\phi_n$ to find that there exist a randomized estimator T with values in $[0,1]$ that has law L_h under the product of P_h and the uniform measure. Then $\phi(x) = \mathrm{E}T(x,U)$ has the desired properties. ∎

The message of the theorem is that every limiting power function is necessary the power function of a test in the limiting Gaussian experiment. The assumption that there exists a limiting power function is very weak, because by the compactness of the unit interval we can always construct subsequences along which a limit exists. An analysis of tests in the Gaussian experiment yields concrete bounds on, for instance, the power of level α tests. Compare Theorem 2.12 in Lecture 2.

4.20 Open Problem. The theorems as presented in this lecture apply to many time series models. However, the semiparametric theory for such models, e.g. discretely observed diffusion processes, appears to be largely undeveloped.

Notes

This lecture is based on [36], [38] and Chapter 3.11 of [41]. It is strongly motivated by ideas of Le Cam, in particular from his papers [15], [16] and [17], and earlier results by [20] and [22] and [23].

Lecture 5
Empirical Processes and Consistency of Z-Estimators

In this lecture and the next lecture we discuss empirical processes. Our main focus is the application of empirical processes to the derivation of asymptotic properties of estimators in semiparametric models. In this first lecture we discuss entropy numbers, Glivenko-Cantelli classes and their application to proving consistency of M- and Z-estimators.

5.1 Empirical Measures and Entropy Numbers

Given i.i.d. random variables $X_1, \ldots, X_n$ with law P on a measurable space $(\mathcal{X}, \mathcal{A})$ and a measurable function $f: \mathcal{X} \mapsto \mathbb{R}$ we let

$$\begin{aligned}
\mathbb{P}_n f &= \frac{1}{n}\sum_{i=1}^n f(X_i), \\
Pf &= \int f\, dP, \\
\mathbb{G}_n f &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \big(f(X_i) - Pf\big) = \sqrt{n}(\mathbb{P}_n - P)f, \\
\|f\|_{P,r} &= \big(P|f|^r\big)^{1/r}.
\end{aligned}$$

Given a class $\mathcal{F}$ of measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$ we view $\mathbb{P}_n$ as a map $f \mapsto \mathbb{P}_n f$ on $\mathcal{F}$. Of course, we can also think of $\mathbb{P}_n$ as the discrete uniform random measure on the points $X_1, \ldots, X_n$. We denote by F a measurable *envelope function* of the class $\mathcal{F}$: a function $F: \mathcal{X} \mapsto \mathbb{R}$ such that $\big|f(x)\big| \le F(x)$ for every $x \in \mathcal{X}$ and $f \in \mathcal{F}$. For a function $z: \mathcal{F} \mapsto \mathbb{R}$ the norm $\|z\|_{\mathcal{F}}$ is the supremum norm: $\|z\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \big|z(f)\big|$.

The law of large numbers asserts that $\mathbb{P}_n f \to Pf$ almost surely if Pf exists, and the central limit theorem asserts that $\mathbb{G}_n f$ is asymptotically normal if $Pf^2 < \infty$. An important aim in empirical process theory is to make these statements uniform in f ranging over a class $\mathcal{F}$, in an appropriate sense. We shall also be concerned with the behaviour of $\mathbb{P}_n \hat{f}_n$ and $\mathbb{G}_n \hat{f}_n$ for $\hat{f}_n$ a "random function", which is related to uniformity.

Uniformity over a class of functions depends on the size of a class. An appropriate measure of size are entropy numbers, which come in two types: with or without bracketing. Given two measurable functions $l, u: \mathcal{X} \mapsto \mathbb{R}$, the *bracket* $[l, u]$ is the collection of all functions $f: \mathcal{X} \mapsto \mathbb{R}$ such that $l \leq f \leq u$. Let $\|\cdot\|$ be a norm on a collection of functions. An *ε-bracket* is a bracket $[l, u]$ such that $\|u - l\| < \varepsilon$. Here it is required that both l and u are of finite norm.

5.1 Definition. The *bracketing number* $N_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)$ is the smallest number of ε-brackets needed to cover $\mathcal{F}$.

5.2 Definition. The *covering number* $N\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)$ is the smallest number of balls of radius ε needed to cover $\mathcal{F}$.

The logarithms of bracketing or covering numbers are called *entropies*. An ε-bracket $[l, u]$ is contained in a ball of radius $\varepsilon/2$ around the midpoint $\frac{1}{2}(l + u)$ of the bracket. It follows that $N\big(\varepsilon/2, \mathcal{F}, \|\cdot\|\big) \leq N_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|\big)$ and hence bracketing numbers are bigger than covering numbers (the factor 2 is of no importance in the following). On the other hand, brackets give pointwise control over functions, whereas for many norms knowing that some function is in a ball, even a very small one, still leaves irregular behaviour on a set of small measure open. Such small sets are important when the function is applied to random variables X_i. This observation explains that typically conditions using covering numbers use many different norms simultaneously, whereas conditions using bracketing numbers use the "true" law P only.

The best results using covering numbers are in terms of *random covering numbers*. For simplicity, we state the results in terms of the bigger uniform covering numbers.

5.3 Definition. The *L_r-uniform covering numbers* relative to the envelope function F are the numbers $\sup_Q N\big(\varepsilon \|F\|_{Q,r}, \mathcal{F}, \|\cdot\|_{Q,r}\big)$, where the supremum is taken over all discrete probability measures Q on $(X, \mathcal{A})$ with $\|F\|_{Q,r} > 0$.

A class $\mathcal{F}$ is, by definition, totally bounded if and only if $N\big(\varepsilon, \mathcal{F}, \|\cdot\|\big) < \infty$ for every $\varepsilon > 0$. (Then its completion is compact.) This will be necessary for the desired uniform law of large numbers or central limit theorem to hold, but it is by far not enough. A more precise measure of the size of a class $\mathcal{F}$ is the rate at which the covering or bracketing numbers increase as $\varepsilon \downarrow 0$.

5.2 Glivenko-Cantelli Classes

The Glivenko-Cantelli theorem is the uniform version of the law of large numbers. The classical Glivenko-Cantelli theorem concerns the uniformity in the convergence of the empirical cumulative distribution function of real-valued random variables. The abstract version is named after this.

5.4 Definition. A collection $\mathcal{F}$ of measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$ is P-Glivenko-Cantelli if $\|\mathbb{P}_n - P\|_{\mathcal{F}} \to 0$ almost surely.

We note that the random distance $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ need not be measurable. By "almost sure" convergence $Z_n \to Z$ of a sequence of possibly unmeasurable maps with values in a metric space, we shall understand that there exist measurable maps Δ_n on the underlying probability space such that $d(Z_n, Z) \leq \Delta_n$ for each n and $\Delta_n \to 0$ almost surely.

Two basic theorems on Glivenko-Cantelli classes are as follows.

5.5 Theorem. *If $N_{[]}\big(\varepsilon, \mathcal{F}, \|\cdot\|_{P,1}\big) < \infty$ for every $\varepsilon > 0$, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

5.6 Theorem. *If $\sup_Q N\big(\varepsilon \|F\|_{Q,1}, \mathcal{F}, \|\cdot\|_{Q,1}\big) < \infty$ for every $\varepsilon > 0$, $PF < \infty$ and $\mathcal{F}$ is suitably measurable, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

The condition that $\mathcal{F}$ be "suitably measurable" will recur, but what is suitable will depend on the situation. In the present case it may be taken to mean that the suprema

$$\left\| \frac{1}{n} \sum_{i=1}^{n} e_i f(X_i) \right\|_{\mathcal{F}}$$

are measurable, for every fixed vector $(e_1, \ldots, e_n) \in \{-1, 1\}^n$, and every $n \in \mathbb{N}$. A simple sufficient condition for this is that the supremum be equal to the same supremum but then computed over a countable class $\mathcal{F}$, e.g. a subclass $\mathcal{G} \subset \mathcal{F}$.

The suitable measurability is necessary because the proof of the theorem is based on a symmetrization and conditioning device, requiring an application of Fubini's theorem. The second, uniform entropy theorem is much harder to prove than the bracketing Glivenko-Cantelli theorem, which can be modelled after the proof of the classical Glivenko-Cantelli theorem.

The condition of the first theorem implies that $PF < \infty$: if we cover $\mathcal{F}$ with finitely many brackets, of for instance size 1, and next take the supremum of the absolute values of all upper and lower bracketing functions, we obtain an integrable envelope. Thus the difference between the two theorems resides solely in the use of bracketing or covering numbers. The stronger bracketing numbers may be replaced by the weaker covering numbers, but only at the cost of using *uniform* covering numbers.

Upper bounds on the covering or bracketing numbers of many classes of functions are known from the classical references on these subjects (1950/60s), from more recent work in approximation theory, and from the combinatorial theory employed by Vapnik and Chervonenkis.

Statistical problems, in particular in semiparametric modelling, generate many new classes of functions, sometimes of a complicated nature, for which such estimates are not known. Then we must either derive new estimates or can use stability theorems that allow the construction of new Glivenko-Cantelli classes from known Glivenko-Cantelli classes. The following theorem is in this spirit and can save much work.

For ease of terminology we call a collection of measurable functions $f: \mathcal{X} \mapsto \mathbb{R}^k$ Glivenko-Cantelli if each of the k collections of coordinate functions is Glivenko-Cantelli.

5.7 Theorem. *If $\mathcal{F}$ is a Glivenko-Cantelli class of functions $f: \mathcal{X} \mapsto \mathbb{R}^k$ with integrable envelope and $\phi: \mathbb{R}^k \mapsto \mathbb{R}$ is continuous, then the class of functions $\phi \circ f: \mathcal{X} \mapsto \mathbb{R}$ is Glivenko-Cantelli provided that it has an integrable envelope.*

5.3 Consistency of M- and Z-estimators

Glivenko-Cantelli classes are useful to carry out proofs that M- or Z-estimators are consistent. These are estimators defined to be a point of maximum or a zero of a given stochastic process.

To remain within the theme of empirical processes, we restrict ourselves to criterion functions that are averages over the observations. For every θ in a metric space Θ let $m_\theta: \mathcal{X} \mapsto \mathbb{R}$ be a measurable function. An M-estimator $\hat{\theta}_n$ is a point of maximum of the map $\theta \mapsto \mathbb{P}_n m_\theta$. The aim is to show that this converges in probability to a point of maximum θ_0 of the map $\theta \mapsto P m_\theta$. The following theorem states a stronger result.

5.8 Theorem. *Suppose that the class of functions $\{m_\theta: \theta \in \Theta\}$ is P-Glivenko-Cantelli and that there exists a point $\theta_0 \in \Theta$ such that $\inf_{\theta: d(\theta, \theta_0) > \delta} P m_\theta < P m_{\theta_0}$ for every $\delta > 0$. Then $\mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0}$ implies that $d(\hat{\theta}_n, \theta_0) \to 0$ almost surely.*

Proof. By the property of $\hat{\theta}_n$, we have $\mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} = P m_{\theta_0} - o(1)$, almost surely. Hence

$$\begin{aligned} P m_{\theta_0} - P m_{\hat{\theta}_n} &\leq \mathbb{P}_n m_{\hat{\theta}_n} - P m_{\hat{\theta}_n} + o(1) \\ &\leq \sup_\theta |\mathbb{P}_n m_\theta - P m_\theta| + o(1) \to 0, \end{aligned}$$

almost surely. By assumption there exists for every $\delta > 0$ a number $\eta > 0$ such that $P m_\theta < P m_{\theta_0} - \eta$ for every θ with $d(\theta, \theta_0) > \delta$. Thus, the event $\{d(\hat{\theta}_n, \theta_0) \geq \delta\}$ is contained in the event $\{P m_{\hat{\theta}_n} < P m_{\theta_0} - \eta\}$. The latter sequence of events decreases to a zero event, in view of the preceding display. ∎

This theorem is good enough for most purposes, but can be improved in two important ways:

- As is clear from the proof, the double-sided convergence given by the Glivenko-Cantelli property is used only to ensure a one-sided convergence, corresponding to the fact that we maximize a criterion function. However, we like the simple Glivenko-Cantelli condition over a more complicated one-sided condition. The tricks that we present below often blur the difference.

- If θ is far from θ_0, then usually Pm_θ will be far from Pm_{θ_0}. The closeness of the random criterion $\mathbb{P}m_\theta$ to the limit Pm_θ need therefore not be uniform in θ as it is required by the Glivenko-Cantelli property. We shall make use of this in Lecture 8 when discussing rates of convergence. (It appears that any type of relaxation of the Glivenko-Cantelli condition to make this point precise, automatically results in a stronger statement concerning a rate of convergence.)

Thus we shall not formulate any refinements here. Our interest will go in another direction: application to semiparametric estimation problems. Before discussing a concrete example, it is instructive to compare the present theorem to the one obtained by Wald in the 1940s. (Wald had maximum likelihood estimators in mind, but his proof applies equallly well to general M-estimators.) Wald's main conditions were compactness of the parameter set (or the possibility of compactification) and local domination. Taking the preceding remarks into account the present theorem contains Wald's theorem, in view of the following lemma.

5.9 Lemma. *Let Θ be a compact metric space, let the map $\theta \mapsto m_\theta(x)$ be continuous for every $x \in \mathcal{X}$ and suppose that every θ has a neighbourhood B such that $\sup_{\theta\in B}|m_\theta|$ is dominated by an integrable function. Then the class $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli and $\inf_{\theta:d(\theta,\theta_0)>\delta} Pm_\theta < Pm_{\theta_0}$ for every $\delta > 0$ if and only if $\theta \mapsto Pm_\theta$ possesses a unique global maximum at θ_0.*

Proof. The compactness of Θ and the local domination of the functions m_θ imply that the class $\{m_\theta : \theta \in \Theta\}$ possesses an integrable envelope function. The dominated convergence and the assumed continuity of the maps $\theta \mapsto m_\theta(x)$ imply that the map $\theta \mapsto Pm_\theta$ is continuous. Thus it attains its maximum on the compact set $\{\theta \in \Theta : d(\theta, \theta_0) \geq \delta\}$ for every given $\delta > 0$, and this is smaller than its value at θ_0, by the assumption that θ_0 is a unique maximum.

To complete the proof we show that the $L_1(P)$-bracketing numbers of the class $\{m_\theta : \theta \in \Theta\}$ are finite. If B_m is a decreasing sequence of neighbourhoods of a fixed θ such that $\cap_m B_m = \{\theta\}$ and u_m and l_m are defined as the supremum and infimum of the functions m_θ with $\theta \in B_m$, then $u_m - l_m \to m_\theta - m_\theta = 0$ as $m \to \infty$, by the continuity of the functions $\theta \mapsto m_\theta$. By the dominated convergence theorem $P(u_m - l_m) \to 0$. We conclude that for every $\varepsilon > 0$ and $\theta \in \Theta$ there exists a neighbourhood B such that $P(u_B - l_B) < \varepsilon$, for u_B and l_B the supremum and infimum of the functions m_θ with $\theta \in B$. The collection of neighbourhoods B obtained this way by varying θ over Θ has a finite subcollection that covers Θ, by the compactness of Θ. The corresponding brackets $[l_B, u_B]$ cover the class $\{m_\theta : \in \Theta\}$. ∎

The preceding theorem reduces the consistency proof of an M-estimator to verification of the good behaviour of the limit criterion function $\theta \mapsto Pm_\theta$ and a Glivenko-Cantelli property. The same methods apply to Z-estimators.

For every θ in a set $\Theta \subset \mathbb{R}^k$ let $\psi_\theta : \mathcal{X} \mapsto \mathbb{R}^k$ be a measurable, vector-valued function. A Z-estimator $\hat{\theta}_n$ is a zero of the map $\theta \mapsto \mathbb{P}_n\psi_\theta$. The aim is to show that this converges in probability to a zero θ_0 of the map $\theta \mapsto P\psi_\theta$.

5.10 Theorem. *Suppose that the class of functions* $\{\psi_\theta : \theta \in \Theta\}$ *is P-Glivenko-Cantelli and that there exists a point* $\theta_0 \in \Theta$ *such that* $\inf_{\theta: d(\theta,\theta_0)>\delta} \|P\psi_\theta\| > 0 = \|P\psi_{\theta_0}\|$ *for every* $\delta > 0$. *Then* $\mathbb{P}_n \psi_{\hat\theta_n} = 0$ *implies that* $d(\hat\theta_n, \theta_0) \to 0$ *almost surely.*

Proof. By the Glivenko-Cantelli property $\|P\psi_{\hat\theta}\| = \|\mathbb{P}_n\psi_{\hat\theta}\| + o(1) = o(1)$, almost surely as $n \to \infty$, by the property of $\hat\theta$. Thus it is impossible that $d(\hat\theta, \theta_0) > \delta$ infinitely often, for any $\delta > 0$. ∎

Notwithstanding beautiful and simple results as the preceding theorems, it remains an unfortunate fact that consistency proofs are not easily forced into a single mould. Because consistency concerns the behaviour of estimators on the global model, a differential analysis, such as possible for normality proofs, is impossible. (Unless one is satisfied with statements as: there exists some sequence of local maxima that converges to a true value, without worrying about the selection of such a sequence or the behaviour of an arbitrary sequence of maxima. We are not.) Proving consistency remains somewhat of an art, and is sometimes the hardest part of the analysis of a maximum likelihood estimator. This is true in particular for semiparametric maximum likelihood estimators, because semiparametric likelihoods may be ill-behaved. In the following three sections we discuss some useful tricks, each time illustrated by an example of interest.

5.3.1 Trick 1: Errors-in-variables

Consider the errors-in-variables models $p_{\theta,\eta}(x,y) = \int \phi(x-z)\phi\big(y - f_\theta(z)\big)\, d\eta(z)$, where ϕ is the standard normal density. The regression function f_θ is assumed known up to a parameter $\theta \in \Theta \subset \mathbb{R}^k$. We wish to prove that the maximum likelihood estimator $(\hat\theta, \hat\eta)$ defined as the maximizer of $\prod_{i=1}^n p_{\theta,\eta}(X_i, Y_i)$ over all $\theta \in \Theta$ and probability distributions η on some interval $\mathcal{Z} \subset \mathbb{R}$ is consistent.

To simplify we assume that Θ and $\mathcal{Z}$ are compact. In the case that the natural parameter space for z is the real line, we could achieve this by extending the model to all probability distributions on the extended real line $\overline{\mathbb{R}}$, defining $\phi(x-z)\phi\big(y-f_\theta(z)\big)$ to be zero for $z = \pm\infty$. Furthermore, we assume that $(\theta, z) \mapsto f_\theta(z)$ is continuous on $\Theta \times \mathcal{Z}$.

The set of all probability measures on $\mathcal{Z}$ is compact under the weak topology. Furthermore, the map $(\theta, \eta) \mapsto p_{\theta,\eta}(x,y)$ is continuous for every (x,y). To analyse the maximum likelihood estimator we might apply the preceding theorem with the functions $m_{\theta,\eta} = \log p_{\theta,\eta}$. These would form a Glivenko-Cantelli class by the preceding lemma, except for the fact that we need to find an integrable envelope function. These functions are bounded above, but their unboundedness from below could prevent this from being true. Because we are interested in a point of maximum, unboundness from below should not cause problems. We could see this by improving the preceding theorem, along the lines of the remarks following its proof. A simpler approach is to apply the theorem not with the functions $\log p_{\theta,\eta}$, but with the functions

$$m_{\theta,\eta} = \log\Big(\frac{p_{\theta,\eta} + p_{\theta_0,\eta_0}}{2}\Big).$$

It is not true that the maximum likelihood estimator $(\hat\theta, \hat\eta)$ maximizes $\mathbb{P}_n m_{\theta,\eta}$ for this choice of $m_{\theta,\eta}$, but it is true that

$$\begin{aligned}\mathbb{P}_n m_{\hat\theta,\hat\eta} = \mathbb{P}_n \log\Big(\frac{p_{\hat\theta,\hat\eta} + p_{\theta_0,\eta_0}}{2}\Big) &\geq \mathbb{P}_n \tfrac{1}{2}\Big(\log p_{\hat\theta,\hat\eta} + \log p_{\theta_0,\eta_0}\Big)\\ &\geq \mathbb{P}_n \log p_{\theta_0,\eta_0}.\end{aligned}$$

In the first inequality we use the concavity of the logarithm, and in the second the definition of $(\hat\theta, \hat\eta)$. Thus $\mathbb{P}m_{\hat\theta,\hat\eta} \geq \mathbb{P}m_{\theta_0,\eta_0}$ and this is good enough for the application of the theorem, because we also have that $Pm_{\theta,\eta} < Pm_{\theta_0,\eta_0}$ unless the densities $\frac{1}{2}(p_{\theta,\eta} + p_{\theta_0,\eta_0})$ and p_{θ_0,η_0} define the same measure. Equivalently, unless $p_{\theta,\eta}$ and p_{θ_0,η_0} define the same probability measure.

The last requirement concerns the identifiability of the parameter (θ_0, η_0). This depends on the nature of the functions f_θ and is a nontrivial matter. For the case of linear functions f_θ it was settled in the 1970s.

The functions $p_{\theta,\eta}$ and hence the functions $m_{\theta,\eta}$ are uniformly bounded above. Furthermore, the functions $m_{\theta,\eta}$ are bounded below by the function $\log p_{\theta_0,\eta_0} - \log 2$. Hence the class of functions $m_{\theta,\eta}$ has a P_{θ_0,η_0}-integrable envelope if $P_{\theta_0,\eta_0} \log p_{\theta_0,\eta_0} > -\infty$. By Jensen's inequality

$$\begin{aligned}P_{\theta_0,\eta_0}(-\log p_{\theta_0,\eta_0}) &\leq \int P_{\theta_0,\eta_0}(-\log)\big(\phi(x-z)\phi\big(y - f_{\theta_0}(z)\big)\, d\eta_0(z)\\ &\lesssim \mathrm{E}_{\theta_0,\eta_0}\big(X^2 + Y^2 + Z^2 + f_{\theta_0}(Z)^2\big).\end{aligned}$$

The right side is finite under reasonable assumptions on η_0.

5.3.2 Trick 2: Cox model

In many semiparametric models "likelihoods" are defined through point masses. A Wald-type proof of consistency is then ruled out by the lack of continuity of the likelihood relative to a useful topology. A proof of consistency may then proceed by an intermediate step using "likelihood equations", but still relying on the Glivenko-Cantelli theorem at several points. We illustrate this for the Cox model, as described in Lecture 3, Example 3.13. Other models have been treated by the same method, albeit that the exact arguments usually are more complicated. We make the same assumptions as in Lecture 3. In particular, C is smaller than some τ with probability one and satisfies $\mathrm{P}(C = \tau) > 0$ and $\mathrm{P}(T > \tau) > 0$.

The density of an observation in the Cox model takes the form

$$\Big(e^{\theta z}\lambda(y)e^{-e^{\theta z}\Lambda(y)}\big(1 - F_{C|Z}(y-|\, z)\big)\Big)^{\delta}\Big(e^{-e^{\theta z}\Lambda(y)} f_{C|Z}(y|\, z)\Big)^{1-\delta} p_Z(z).$$

To define a maximum likelihood estimator for (θ, Λ), we of course drop the terms involving the distribution of (C, Z), which is assumed not to depend on the parameter of interest. Unfortunately, the supremum of

$$\prod_{i=1}^n \Big(e^{\theta Z_i}\lambda(Y_i)e^{-e^{\theta Z_i}\Lambda(Y_i)}\Big)^{\Delta_i}\Big(e^{-e^{\theta Z_i}\Lambda(Y_i)}\Big)^{1-\Delta_i}$$

over all parameters θ and hazard functions λ is infinite. We can approximate this supremum by choosing hazard functions that have very high, but very thin peaks

around the values Y_i with $\Delta_i = 1$. By making the peaks sufficiently thin we can ensure that the values $\Lambda(Y_i)$ are arbitrarily close to zero and hence the value of the preceding display will be determined by the factor $\prod_{i=1}^n \lambda(Y_i)$.

Thus we cannot define a maximum likelihood estimator in this way. The way out is to *define* the likelihood instead by

$$\prod_{i=1}^n \Big(e^{\theta Z_i} \Lambda\{Y_i\} e^{-e^{\theta Z_i}\Lambda(Y_i)} \Big)^{\Delta_i} \Big(e^{-e^{\theta Z_i}\Lambda(Y_i)} \Big)^{1-\Delta_i}$$

Next we maximize over all $\theta \in \Theta$ and nondecreasing, cadlag functions $\Lambda\colon [0,\infty) \mapsto \mathbb{R}$ with $\Lambda(0) = 0$. (This is a bit bigger than the set of cumulative hazard functions, defined as finite measures of the type $d\Lambda = dF/(1-F-)$ for cumulative distributions F, which are restricted to having jumps of size less than 1, but asymptotically this will not make a difference.) Maximizing relative to Λ entails maximizing the jumps $\Lambda\{Y_i\}$ at points Y_i with $\Delta_i = 1$, meanwhile minimizing the cumulative masses $\Lambda(Y_i)$ at Y_i such that $\Delta_i = 0$. The best choice is among the discrete distributions Λ that jump at the points Y_i with $\Delta_i = 1$ only. This observation reduces the maximization problem to a finite-dimensional one (finding the jump sizes), and the compactness of the unit simplex implies that a solution exists, also jointly in θ and Λ.

What we have called "likelihood" does not have the continuity property we would require for a Wald type proof. Also the parameter space for Λ is not a-priori compact. We get around this problem by using likelihood equations. For a bounded function h we can define by $d\hat\Lambda_t = (1+th)\,d\hat\Lambda$ a perturbation of $\hat\Lambda$, defined for at least every t in a neighbourhood of 0. The likelihood evaluated at $(\hat\theta, \hat\Lambda_t)$ viewed as a function of t must be maximal at $t = 0$. Differentiating at $t = 0$ we obtain the stationary equation

$$\mathbb{P}_n B_{\hat\theta_n, \hat\Lambda_n} h = 0,$$

where $B_{\theta,\Lambda}$ is (the version of) the score operator given in Example 3.13. We can rewrite this equation as

$$\mathbb{P}_n \delta h(y) = \mathbb{P}_n e^{\hat\theta_n z} \int_{[0,y]} h\, d\hat\Lambda_n = \int \mathbb{P}_n e^{\hat\theta_n z} h(s) 1_{s\le y}\, d\hat\Lambda_n(s).$$

In this notation $\mathbb{P}_n$ is the empirical measure of the observations $X_i = (Y_i, \Delta, Z_i)$, and we write $\mathbb{P}_n f(x)$ instead of $\mathbb{P}_n f$ for clarity (we hope). Inverting the preceding display (i.e. replacing h by $h/\hat M_{n,0}$), we find

$$\hat\Lambda h_n = \mathbb{P}_n \frac{\delta h(y)}{\hat M_{n,0}(y)}, \qquad \hat M_{n,0}(s) = \mathbb{P}_n e^{\hat\theta z} 1_{s\le y}.$$

If we knew that $\hat\theta_n$ were consistent, then we could use this representation directly to prove the consistency of $\hat\Lambda_n$. The Cox model, as usual, is much simpler here than other models. In other situations we find a recursive expression for $\hat\Lambda_n$ with both $\hat\Lambda_n$ and $\hat\theta_n$ appearing on the right side, but the argument may proceed in the same way.

The Wald argument is based on comparing the value of the likelihood at the maximum likelihood estimator and at the true value of the parameter. In the present

case this causes a problem, because the likelihood at the maximum likelihood estimator, a random discrete distribution, and at the true parameter are different in character. This is solved by comparing the likelihood at the maximum likelihood estimator and at the random parameter $(\theta_0, \tilde\Lambda_n)$ for $\tilde\Lambda_n$ defined by

$$\tilde\Lambda_n h = \mathbb{P}_n \frac{\delta h(y)}{M_0(y)}, \qquad M_0(s) = P_0 e^{\theta_0 z} 1_{s \le y}.$$

The function $\tilde\Lambda_n$ is similar in structure to $\hat\Lambda_n$, but is also similar to Λ_0: applying the same algebra as previously to the equation $P_0 B_0 h = 0$ we see that

$$\Lambda_0 h = P_0 \frac{\delta h(y)}{M_0(y)}.$$

Under our assumptions $M_0(s) \ge M_0(\tau)$ is bounded away from zero. Therefore, the functions $(y, \delta) \mapsto \delta h(y)/M_0(y)$ form a Glivenko-Cantelli class if h ranges over a Glivenko-Cantelli class and hence $\tilde\Lambda_n h \to P_0 \delta h(y)/M_0(y) = \Lambda_0 h$, uniformly in h ranging over a Glivenko-Cantelli class.

The log likelihood evaluated at $(\hat\theta, \hat\Lambda)$ is bigger than the log likelihood evaluated at $(\theta_0, \tilde\Lambda)$. The point masses $\Lambda\{Y_i\}$ in the likelihood when evaluated at $\hat\Lambda$ and $\tilde\Lambda$ can be reexpressed in the functions $\hat M_{n,0}$ and M_0. Specifically we have that $\hat\Lambda/\tilde\Lambda\{Y_i\} = M_0/\hat M_{n,0}(Y_i)$. This yields the equation

$$(\hat\theta - \theta_0)\mathbb{P}_n z\delta - \mathbb{P}_n\Big(e^{\hat\theta z}\hat\Lambda(y) - e^{\theta_0 z}\tilde\Lambda(y)\Big) + \mathbb{P}_n \delta \log \frac{M_0}{\hat M_{n,0}}(y) \ge 0. \tag{5.11}$$

In the next paragraphs we prove that this implies that for almost all ω in the underlying probability space there exists $\theta_\infty \in \Theta$ such that along a subsequence $(\hat\theta_n, \hat\Lambda_n) \to (\theta_\infty, \Lambda_\infty)$ and

$$(\theta_\infty - \theta_0)P_0 z\delta - P_0\Big(e^{\theta_\infty z}\Lambda_\infty(y) - e^{\theta_0 z}\Lambda_0(y)\Big) + P_0\delta \log \frac{M_0}{M_{\infty,0}}(y) \ge 0, \tag{5.12}$$

for

$$M_{\infty,0}(s) = P_0 e^{\theta_\infty z} 1_{s \le y}, \qquad \Lambda_\infty h = P_0 \frac{\delta h(y)}{M_{\infty,0}(y)}.$$

The topology on Λ can be taken equal to the uniform norm on $[0, \tau]$. Noting that $M_0/M_{\infty,0} = d\Lambda_\infty/d\Lambda_0$, we recognize (5.12) as the Kullback-Leibler divergence $P_0 \log(p_{\theta_\infty,\Lambda_\infty}/p_{\theta_0,\Lambda_0})$, which is strictly negative by the identifiability of (θ_0, Λ_0), unless $(\theta_\infty, \Lambda_\infty) = (\theta_0, \Lambda_0)$. This would finish the proof that $(\hat\theta_n, \hat\Lambda_n) \to (\theta_0, \Lambda_0)$ almost surely.

To deduce (5.12) from (5.11) we note first that the functions $\hat M_{n,0}$ are bounded below by $\hat M_{n,0}(\tau) = \mathbb{P}_n e^{\hat\theta z} 1_{y=\tau}$, which is asymptotically bounded away from zero under our assumptions. Therefore the functions $(\delta, y) \mapsto \delta h(y)/\hat M_{n,0}(y)$ are contained in a Glivenko-Cantelli class, almost surely, if h ranges over a Glivenko-Cantelli class. It follows that $\hat\Lambda h = P_0 \delta h(y)/\hat M_{n,0}(y) + o(1)$, almost surely, uniformly in h running through a Glivenko-Cantelli class.

Second, we note that $\mathbb{P}_n e^{\hat\theta z} 1_{y=\tau}\hat\Lambda(\tau) \le \mathbb{P}_n e^{\hat\theta z}\hat\Lambda(y) = \mathbb{P}_n \delta$, by the likelihood equation with $h = 1$ and hence $\hat\Lambda(\tau)$ is uniformly bounded above, eventually, almost surely.

By the compactness of Θ, the sequence $\hat\theta_n$ converges to a limit θ_∞, at least along subsequences. Then $\hat M_{n,0}(s) \to M_{\infty,0}(s)$, uniformly in s, almost surely, and hence $\hat\Lambda_n h \to P_0\delta h(y)/\hat M_{\infty,0}(y) = \Lambda_\infty h$ almost surely, uniformly in h running through a Glivenko-Cantelli class with integrable envelope, still along a subsequence. It now suffices to take limits in (5.11). This is done in two steps. We first replace $\mathbb{P}_n$ by P_0 adding a $o(1)$-term, which is permitted, because the classes of functions $(z,\delta) \mapsto z\delta$, $(y,z) \mapsto e^{\theta z}\Lambda(y)$ and $(y,\delta) \mapsto \delta \log M_0/M(y)$ with $\theta \in \Theta$, Λ ranging over a uniformly bounded set of monotone, cadlag functions and M ranging over monotone, cadlag functions that are bounded away from zero, is Glivenko-Cantelli. The second step is to replace $\hat\theta_n$, $\hat\Lambda_n$, $\tilde\Lambda_n$ and $\hat M_{n,0}$ by their limits, which is justified by the dominated convergence theorem.

5.3.3 Trick 3: Mixture models

Our first trick already showed that for proving consistency of a maximum likelihood estimator, it may be useful to apply a general result for M-estimators not to the log density, but to a slightly modified function. In models that depend linearly on a parameter belonging to a convex set, there is an even better choice.

Given a kernel $p(x|\,z)$ indexed by $z \in \mathcal{Z}$ and a probability distribution η on $\mathcal{Z}$, let $p_\eta(x) = \int p(x|\,z)\,d\eta(z)$. Consider proving the consistency of the maximum likelihood estimator $\hat\eta$, which maximizes $\eta \mapsto \prod_{i=1}^n p_\eta(X_i)$ over the set of all probability measures.

We can use the linearity of this model, by starting from the observation that the likelihood is bigger at $\hat\eta$ than at $\eta_t = t\eta + (1-t)\hat\eta$ for every η and $t \in [0,1]$. Differentiating the inequality $\mathbb{P}_n \log p_{\hat\eta}/p_{\eta_t} \geq 0$ from the right at $t = 0$ we obtain

$$\mathbb{P}_n \frac{p_\eta}{p_{\hat\eta}} \leq 1.$$

We might try and use this equation for a consistency proof, but the quotients $p_\eta/p_{\hat\eta}$ may lack integrability, and it is useful to make a second step. Let $L\colon [0,\infty) \mapsto \mathbb{R}$ be a nondecreasing function such that $t \mapsto L(1/t)$ is convex. Then

$$\mathbb{P}_n L\Big(\frac{p_{\hat\eta}}{p_\eta}\Big) = \mathbb{P}_n L\Big(\frac{1}{p_\eta/p_{\hat\eta}}\Big) \geq L\Big(\frac{1}{\mathbb{P}_n p_\eta/p_{\hat\eta}}\Big) \geq L(1) = \mathbb{P}_n L\Big(\frac{p_{\eta_0}}{p_{\eta_0}}\Big).$$

Thus we may use Theorem 5.8 with the choice $m_\eta = L(p_\eta/p_{\eta_0})$. The choice

$$L(t) = \frac{t^\alpha - 1}{t^\alpha + 1}, \qquad \alpha \in (0,1],$$

is attractive, because then $L(t) = -L(1/t)$ is strictly concave. By Jensen's inequality

$$P_{\eta_0} L\Big(\frac{p_\eta}{p_{\eta_0}}\Big) \leq L\Big(P_{\eta_0}\frac{p_\eta}{p_{\eta_0}}\Big) \leq L(1).$$

Unless $p_\eta = p_{\eta_0}$ almost surely under P_{η_0}, the first inequality will be strict and hence the left side will be strictly less than the right side.

If the set of functions $x \mapsto p(x|\,z)$ where z ranges over $\mathcal{Z}$ is Glivenko-Cantelli, then so is its convex hull, the set of all functions p_η. The one-element class consisting of the function $1/p_{h_0}$ is Glivenko-Cantelli and hence so is the class of all functions

p_η/p_{η_0} and the class of functions $L(p_\eta/p_{\eta_0})$ when η ranges over all probability distributions on $(\mathcal{Z},\mathcal{C})$, by two applications of Theorem 5.7.

We now obtain the consistency of $\hat\eta$ for η_0 if we can verify that $P_{\eta_0}L\big(p_\eta/p_{\eta_0}\big)$ is strictly bounded away from its maximal value $L(1)$ if η varies over the complement of a ball of radius δ around η_0. This, of course, depends on the metric we choose for the set of mixing distributions. If we choose a metric for which this set is compact, then it suffices to verify that the map $\eta \mapsto p_\eta(x)$ is continuous for P_{η_0}-almost every x, because then so is the map $\eta \mapsto P_{\eta_0}L\big(p_\eta/p_{\eta_0}\big)$, by the dominated convergence theorem. In many examples the weak topology is appropriate, possibly after first compactifying $\mathcal{Z}$. A semi-metric that always works is the induced Hellinger metric, because, for $\alpha = 1/2$,

$$P_{\eta_0}L\Big(\frac{p_\eta}{p_{\eta_0}}\Big) \le -\tfrac{1}{2}h^2(p_\eta, p_{\eta_0}).$$

5.4 Nuisance Parameters

We close this lecture by noting that the preceding theorems have easy extensions to M- and Z-estimators defined in the presence of nuisance parameters. In the case of M-estimators we might be given measurable functions $m_{\theta,\eta}\colon \mathcal{X} \mapsto \mathbb{R}$ indexed by a parameter of interest θ and a nuisance parameter η. Given an initial estimator $\hat\eta$ for η, we consider $\hat\theta$ maximizing $\theta \mapsto \mathbb{P}_n m_{\theta,\hat\eta}$. More generally, given an "estimator" $\hat\eta(\theta)$ for η that may depend on θ, we consider $\hat\theta$ maximizing $\theta \mapsto \mathbb{P}_n m_{\theta,\hat\eta(\theta)}$. (The latter criterion is sometimes called a *profile criterion* function.) Both cases are covered if we allow a general random criterion function

$$\hat m_{n,\theta}(x) = \hat m_{n,\theta}(x; X_1, \ldots, X_n).$$

We shall assume that asymptotically the randomness disappears: $\hat m_{n,\theta} \to m_\theta$ for deterministic, measurable functions m_θ.

5.13 Theorem. *Suppose that there exists a Glivenko-Cantelli class $\mathcal{F}$ of functions with integrable envelope such that $P^n\big(\{\hat m_{n,\theta}\colon \theta \in \Theta\} \subset \mathcal{F}\big) \to 1$, suppose that $\sup_{\theta\in\Theta} |\hat m_{n,\theta} - m_\theta|(x) \xrightarrow{\mathrm{P}} 0$ for all x, and that there exists a point $\theta_0 \in \Theta$ such that $\inf_{\theta:d(\theta,\theta_0)>\delta} Pm_\theta < Pm_{\theta_0}$ for every $\delta > 0$. Then $\mathbb{P}_n m_{n,\hat\theta_n} \ge \mathbb{P}_n \hat m_{n,\theta_0}$ implies that $d(\hat\theta_n, \theta_0) \to 0$ in probability.*

Proof. For any random sequence $\tilde\theta$ and every x, the sequence $|m_{n,\tilde\theta} - m_{\tilde\theta}|(x)$ is bounded by $2F(x) < \infty$, for F an envelope function of the class, and converges in probability to zero. This implies that it converges to zero in mean and hence, by Fubini's theorem and the dominated convergence theorem, $\mathrm{E}P|m_{n,\tilde\theta} - m_{\tilde\theta}| \to 0$. Consequently, the sequence $P(m_{n,\tilde\theta} - m_{\tilde\theta})$ converges to zero in probability.

Combining this with the Glivenko-Cantelli assumption we obtain that the sequence $|\mathbb{P}_n \hat m_{n,\tilde\theta_n} - Pm_{\tilde\theta_n}|$ converges to zero in probability.

The remainder of the proof is similar to the proof of Theorem 5.10. ∎

Notes

The proofs of most results on empirical processes given in this lecture and the following ones can be found in the book [41]. This work also contains a reasonable number of references to the huge literature on empirical processes. We do not refer to this here, apart from mentioning that the Saint-Flours notes by Dudley [9] were a major step in developing the abstract theory of empirical processes. Applications of empirical processes to the analysis of M-estimators and Z-estimators were pioneered by Pollard. See [29], [30].

Trick 1 I learned from [4], trick 2 (applied here to the Cox model for the first time) from Susan Murphy (see [24]), and trick 3 from [27].

Lecture 6
Empirical Processes and Normality of Z-Estimators

In this lecture we continue the discussion of empirical processes, now concentrating on the central limit theorem and uniformity in convergence in distribution, and its applications to deriving the asymptotic distribution of Z-estimators.

6.1 Weak Convergence in Metric Spaces

Let $(\Omega_n, \mathcal{U}_n, \mathrm{P}_n)$ be a sequence of probability spaces and, for each n, let $X_n \colon \Omega_n \mapsto \mathbb{D}$ be an arbitrary map from Ω_n into a metric space $\mathbb{D}$.

6.1 Definition. The sequence X_n converges in distribution to a Borel measure L on $\mathbb{D}$ if and only if $\mathrm{E}^* f(X_n) \to \int f \, dL$ for every bounded, continuous function $f \colon \mathbb{D} \mapsto \mathbb{R}$.

Here the asterisk * denotes outer expectation, and is necessary because we have not assumed that the maps X_n are Borel measurable. It is defined as

$$\mathrm{E}^* f(X) = \inf\Big\{ \mathrm{E}U \colon\ U \colon \Omega \mapsto \mathbb{R}, \text{measurable}, U \geq f(X), \mathrm{E}U \text{ exists} \Big\}.$$

If X is a Borel measurable map in $\mathbb{D}$, defined on some probability space, with law L, then we also write $X_n \rightsquigarrow X$ instead of $X_n \rightsquigarrow L$. The limit is always assumed to be Borel measurable. Even though the X_n and X are ordinary maps, we also refer to them as "random elements", as they are defined on a probability space and hence induce randomness on $\mathbb{D}$.

In the following, we do not stress the measurability issues. However, we write stars, when necessary, as a reminder that there are measurability issues that need to be taken care of. Although Ω_n may depend on n, we do not let this show up in the notation for E^* and P^*.

Next consider convergence in probability and almost surely.

6.2 Definition. An arbitrary sequence of maps $X_n \colon \Omega_n \mapsto \mathbb{D}$ *converges in probability* to X if $\mathrm{P}^*\big(d(X_n, X) > \varepsilon\big) \to 0$ for all $\varepsilon > 0$. This is denoted by $X_n \xrightarrow{\mathrm{P}} X$.

6.3 Definition. An arbitrary sequence of maps $X_n: \Omega_n \mapsto \mathbb{D}$ *converges almost surely* to X if there exists a sequence of (measurable) random variables Δ_n such that $d(X_n, X) \leq \Delta_n$ and $\Delta_n \stackrel{\text{as}}{\to} 0$ This is denoted by $X_n \stackrel{\text{as}*}{\to} X$.

These definitions also do not require the X_n to be Borel measurable. In the definition of "convergence of probability" we added a star, for "outer probability". Similar to outer expectation, we define *outer probability* by

$$\mathrm{P}^*(X \in B) = \inf\Big\{\mathrm{P}(A): A \in \mathcal{A}, A \supset X^{-1}(B)\Big\}.$$

The definition of "almost sure convergence" is unpleasantly complicated. This cannot be avoided easily, because, even for Borel measurable maps X_n and X, the distance $d(X_n, X)$ need not be a random variable.

Most of the well-known properties and relationships of these modes of convergence remain valid under the generalized definitions. We collect the most important ones in the following theorem.

6.4 Theorem. *For arbitrary maps $X_n, Y_n: \Omega_n \mapsto \mathbb{D}$ and every random element X with values in $\mathbb{D}$,*

(i) $X_n \stackrel{\mathrm{P}}{\to} X$ *implies* $X_n \rightsquigarrow X$;
(ii) $X_n \stackrel{\mathrm{P}}{\to} c$ *for a constant c if and only if* $X_n \rightsquigarrow c$;
(iii) *if $X_n \rightsquigarrow X$, then $\phi(X_n) \rightsquigarrow \phi(X)$ for every map $\phi: \mathbb{D} \mapsto \mathbb{E}$ that is continuous at every point of a set $\mathbb{D}_0$ such that $\mathrm{P}(X \in \mathbb{D}_0) = 1$ and such that $\phi(X)$ is Borel measurable;*
(iv) *if $X_n \rightsquigarrow X$ and $d(X_n, Y_n) \stackrel{\mathrm{P}}{\to} 0$, then $Y_n \rightsquigarrow X$;*
(v) *if $X_n \rightsquigarrow X$ and $Y_n \stackrel{\mathrm{P}}{\to} c$ for a constant c, then $(X_n, Y_n) \rightsquigarrow (X, c)$;*
(vi) *if $X_n \stackrel{\mathrm{P}}{\to} X$ and $Y_n \stackrel{\mathrm{P}}{\to} Y$, then $(X_n, Y_n) \stackrel{\mathrm{P}}{\to} (X, Y)$.*

The metric spaces we are mostly interested in are, besides the Euclidean spaces, spaces of bounded functions equipped with the uniform norm. Given an arbitrary set T let $\ell^\infty(T)$ be the collection of all bounded functions $z: T \mapsto \mathbb{R}$. This is a Banach space under the uniform norm

$$\|z\|_T = \sup_{t \in T} \big|z(t)\big|.$$

Most of the random elements X with values in $\ell^\infty(T)$ of interest to us are *stochastic processes* in that their coordinate values $X_t = \pi_t \circ X$, for $\pi: \ell^\infty(T) \mapsto \mathbb{R}$ the coordinate projection $z \mapsto z(t)$, are random variables. However, many of them are not Borel measurable in $\ell^\infty(T)$ and hence the preceding extensions of the usual definitions are useful. Earlier extensions based on the σ-field generated by the closed balls, initiated by Dudley and expounded by Pollard, are special cases of the present approach, which is due to Hoffmann-Jørgensen.

We use the space $\ell^\infty(T)$ for defining "uniform weak convergence" of stochastic processes, such as the empirical processes. The next theorem gives a characterization of weak convergence in this space by finite approximation. It is required that, for any $\varepsilon > 0$, the index set T can be partitioned into finitely many sets $T_1, \ldots, T_k$ such that (asymptotically) the variation of the sample paths $t \mapsto X_{n,t}$ is less than ε on every one of the sets T_i, with large probability. Then the behaviour of the process can

be described, within a small error margin, by the behaviour of the *marginal vectors* $(X_{n,t_1}, \ldots, X_{n,t_k})$ for arbitrary fixed points $t_i \in T_i$. If these marginals converge, then the processes converge.

6.5 Theorem. *A sequence of arbitrary maps $X_n: \Omega_n \mapsto \ell^\infty(T)$ converges weakly to a tight random element if and only if both of the following conditions hold:*

(i) the sequence $(X_{n,t_1}, \ldots, X_{n,t_k})$ converges in distribution in $\mathbb{R}^k$ for every finite set of points $t_1, \ldots, t_k$ in T;

(ii) for every $\varepsilon, \eta > 0$ there exists a partition of T into finitely many sets $T_1, \ldots, T_k$ such that

$$\limsup_{n\to\infty} \mathrm{P}^* \left(\sup_i \sup_{s,t\in T_i} \left| X_{n,s} - X_{n,t} \right| \geq \varepsilon \right) \leq \eta.$$

Proof. We only give the proof of the more constructive part, the sufficiency of (i)-(ii). For each natural number m, partition T into sets $T_1^m, \ldots, T_{k_m}^m$ as in (ii) corresponding to $\varepsilon = \eta = 2^{-m}$. Since the probabilities in (ii) decrease if the partition is refined, we can assume without loss of generality that the partitions are successive refinements as m increases. For fixed m define a semimetric ρ_m on T by $\rho_m(s,t) = 0$ when s and t belong to the same partioning set T_j^m, and by $\rho_m(s,t) = 1$ otherwise. Every ρ_m-ball of radius $0 < \varepsilon < 1$ coincides with a partitioning set. In particular, T is totally bounded for ρ_m, and the ρ_m-diameter of a set T_j^m is zero. By the nesting of the partitions, $\rho_1 \leq \rho_2 \leq \cdots$. Define $\rho(s,t) = \sum_{m=1}^\infty 2^{-m} \rho_m(s,t)$. Then ρ is a semimetric such that the ρ-diameter of T_j^m is smaller than $\sum_{k>m} 2^{-k} = 2^{-m}$, and hence T is totally bounded for ρ. Let T_0 be the countable ρ-dense subset constructed by choosing an arbitrary point t_j^m from every T_j^m.

By assumption (i) and Kolmogorov's consistency theorem we can construct a stochastic process $\{X_t: t \in T_0\}$ on some probability space such that $(X_{n,t_1}, \ldots, X_{n,t_k}) \rightsquigarrow (X_{t_1}, \ldots, X_{t_k})$ for every finite set of points $t_1, \ldots, t_k$ in T_0. By the portmanteau lemma and assumption (ii), for every finite set $S \subset T_0$,

$$\mathrm{P}\left(\sup_j \sup_{\substack{s,t\in T_j^m \\ s,t\in S}} |X_s - X_t| > 2^{-m} \right) \leq 2^{-m}.$$

By the monotone convergence theorem this remains true if S is replaced by T_0. If $\rho(s,t) < 2^{-m}$, then $\rho_m(s,t) < 1$ and hence s and t belong to the same partitioning set T_j^m. Consequently, the event in the preceding display with $S = T_0$ contains the event in the following display, and

$$\mathrm{P}\left(\sup_{\substack{\rho(s,t)<2^{-m} \\ s,t\in T_0}} |X_s - X_t| > 2^{-m} \right) \leq 2^{-m}.$$

This sums to a finite number over $m \in \mathbb{N}$. Hence, by the Borel-Cantelli lemma, for almost all ω, $\left| X_s(\omega) - X_t(\omega) \right| \leq 2^{-m}$ for all $\rho(s,t) < 2^{-m}$ and all sufficiently large m. This implies that almost all sample paths of $\{X_t: t \in T_0\}$ are contained in $UC(T_0, \rho)$. Extend the process by continuity to a process $\{X_t: t \in T\}$ with almost all sample paths in $UC(T, \rho)$.

Define $\pi_m: T \mapsto T$ as the map that maps every partioning set T_j^m onto the point $t_j^m \in T_j^m$. Then, by the uniform continuity of X, and the fact that the ρ-diameter

of T_j^m is smaller than 2^{-m}, $X \circ \pi_m \rightsquigarrow X$ in $\ell^\infty(T)$ as $m \to \infty$ (even almost surely). The processes $\{X_n \circ \pi_m(t): t \in T\}$ are essentially k_m-dimensional vectors. By (i), $X_n \circ \pi_m \rightsquigarrow X \circ \pi_m$ in $\ell^\infty(T)$ as $n \to \infty$, for every fixed m. Consequently, for every Lipschitz function $f: \ell^\infty(T) \mapsto [0,1]$, $\mathrm{E}^* f(X_n \circ \pi_m) \to \mathrm{E} f(X)$ as $n \to \infty$, followed by $m \to \infty$. Conclude that, for every $\varepsilon > 0$,

$$\begin{aligned}\big|\mathrm{E}^* f(X_n) - \mathrm{E} f(X)\big| &\le \big|\mathrm{E}^* f(X_n) - \mathrm{E}^* f(X_n \circ \pi_m)\big| + o(1)\\ &\le \|f\|_{\mathrm{lip}}\varepsilon + \mathrm{P}^*\Big(\|X_n - X_n \circ \pi_m\|_T > \varepsilon\Big) + o(1).\end{aligned}$$

For $\varepsilon = 2^{-m}$ this is bounded by $\|f\|_{\mathrm{lip}} 2^{-m} + 2^{-m} + o(1)$, by the construction of the partitions. The proof is complete. ∎

In the course of the proof of the preceding theorem a semimetric ρ is constructed such that the weak limit X has uniformly ρ-continuous sample paths, and such that (T, ρ) is totally bounded. This is surprising: even though we are discussing stochastic processes with values in the very large space $\ell^\infty(T)$, the limit is concentrated on a much smaller space of continuous functions. Actually, this is a consequence of imposing the condition (ii), which can be shown to be equivalent to asymptotic tightness. (A sequence X_n is called *asymptotically tight* if for every $\varepsilon > 0$ there exists a compact set $K \subset \mathbb{D}$ such that $\liminf \mathrm{P}(d(X_n, K) < \eta) \ge 1 - \varepsilon$ for every $\eta > 0$.) It can be shown, more generally, that every tight random element X in $\ell^\infty(T)$ necessarily concentrates on $UC(T, \rho)$ for some semimetric ρ (depending on X) that makes T totally bounded.

In view of this connection between the partitioning condition (ii), continuity and tightness, we sometimes refer to this condition as the condition of *asymptotic tightness* or *asymptotic equi-continuity*. One consequence of this is that a tight random element X is completely determined by its values on a countable set (taken dense in (T, ρ)), and hence its distribution is determined by the distributions of all its finite-dimensional projections.

The existence of a semi-metric that induces continuity will enable us to use empirical process theory in the analysis of Z-estimators. Thus we record the existence of the semimetric for later reference. We also note that, for a Gaussian limit process, this can always be taken equal to the "intrinsic" standard deviation semimetric. This will help a good deal to make our results on Z-estimators more concrete.

6.6 Lemma. *Under the conditions (i)–(ii) of the preceding theorem there exists a semimetric ρ on T for which T is totally bounded, and such that the weak limit of the sequence X_n can be constructed to have almost all sample paths in $UC(T, \rho)$. Furthermore, if the weak limit X is zero-mean Gaussian, then this semimetric can be taken equal to $\rho(s,t) = \mathrm{sd}(X_s - X_t)$.*

Proof. A semimetric ρ is constructed explicitly in the proof of the preceding theorem. It suffices to prove the statement concerning Gaussian limits X.

Let ρ be the semimetric obtained in the proof of the theorem and let ρ_2 be the standard deviation semimetric. Since every uniformly ρ-continuous function has a unique continuous extension to the ρ-completion of T, which is compact, it is no loss of generality to assume that T is ρ-compact. Furthermore, assume that *every* sample path of X is ρ-continuous.

An arbitrary sequence t_n in T has a ρ-converging subsequence $t_{n'} \to t$. By the ρ-continuity of the sample paths, $X_{t_{n'}} \to X_t$ almost surely. Since every X_t is Gaussian, this implies convergence of means and variances, whence $\rho_2(t_{n'}, t)^2 = \mathrm{E}(X_{t_{n'}} - X_t)^2 \to 0$. Thus $t_{n'} \to t$ also for ρ_2 and hence T is ρ_2-compact.

Suppose that a sample path $t \mapsto X_t(\omega)$ is not ρ_2-continuous. Then there exists an $\varepsilon > 0$ and a $t \in T$ such that $\rho_2(t_n, t) \to 0$, but $\left|X_{t_n}(\omega) - X_t(\omega)\right| \geq \varepsilon$ for every n. By the ρ-compactness and continuity, there exists a subsequence such that $\rho(t_{n'}, s) \to 0$ and $X_{t_{n'}}(\omega) \to X_s(\omega)$ for some s. By the argument of the preceding paragraph, $\rho_2(t_{n'}, s) \to 0$, so that $\rho_2(s,t) = 0$ and $\left|X_s(\omega) - X_t(\omega)\right| \geq \varepsilon$. Conclude that the path $t \mapsto X_t(\omega)$ can only fail to be ρ_2-continuous for ω for which there exist $s, t \in T$ with $\rho_2(s,t) = 0$, but $X_s(\omega) \neq X_t(\omega)$. Let N be the set of ω for which there do exist such s, t. Take a countable, ρ-dense subset A of $\{(s,t) \in T \times T \colon \rho_2(s,t) = 0\}$. Since $t \mapsto X_t(\omega)$ is ρ-continuous, N is also the set of all ω such that there exist $(s,t) \in A$ with $X_s(\omega) \neq X_t(\omega)$. From the definition of ρ_2, it is clear that for every fixed (s,t), the set of ω such that $X_s(\omega) \neq X_t(\omega)$ is a nullset. Conclude that N is a null set. Hence, almost all paths of X are ρ_2-continuous. ∎

6.2 Donsker Classes

Given a random sample $X_1, \ldots, X_n$ from a probability distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$, let again $\mathbb{G}_n$ be the empirical process $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf)$ indexed by a given class $\mathcal{F}$ of measurable functions. Under the assumptions that the class possesses a finite envelope F and that $\|P\|_{\mathcal{F}}$ is finite (in particular if $PF < \infty$) the sample paths $f \mapsto \mathbb{G}_n f$ are contained in the space $\ell^\infty(\mathcal{F})$.

6.7 Definition. A class $\mathcal{F}$ of functions is *P-Donsker* if the sequence of empirical processes converges in distribution to a tight limit process in the space $\ell^\infty(\mathcal{F})$.

The convergence of the process in $\ell^\infty(\mathcal{F})$ implies the convergence of the marginals $\big(\mathbb{G}_n f_1, \ldots, \mathbb{G}_n f_k\big)$ for given any finite set elements $f_i \in \mathcal{F}$, by the continuous mapping theorem. This is possible only if $Pf_i^2 < \infty$ for every i and then the limit distribution is multivariate normal with mean zero and covariances $P(f_i - Pf_i)(f_j - Pf_j)$ by the multivariate central limit theorem. Thus if $\mathcal{F}$ is Donsker, then $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ for a tight Gaussian random element in $\ell^\infty(\mathcal{F})$ with mean zero and covariance function

$$\mathrm{E}\mathbb{G}_P f \, \mathbb{G}_P g = Pfg - PfPg.$$

This is known as a *P-Brownian bridge*. In view of the results of the preceding section this is also determined by:

- $\mathbb{G}$ is a Gaussian process;
- $\mathrm{E}\mathbb{G}f = 0$, $\mathrm{cov}(\mathbb{G}f, \mathbb{G}g) = Pfg - PfPg$;
- the sample paths of $\mathbb{G}$ can be constructed to be uniformly continuous relative to the semimetric $\rho(f,g) = \mathrm{sd}(\mathbb{G}f - \mathbb{G}g)$.
- $\mathcal{F}$ is totally bounded under ρ.

The $L_2(P)$-metric is slightly stronger than the metric ρ, because

$$\rho^2(f,g) = P\big((f-Pf)-(g-Pg)\big)^2 \le P(f-g)^2.$$

Thus the sample paths are also uniformly continuous relative to the $L_2(P)$-semimetric. It is not hard to see that $\mathcal{F}$ will also be totally bounded relative to the $L_2(P)$-semimetric as soon as $\|P\|_{\mathcal{F}} < \infty$. Thus there is not much loss in replacing ρ by the $L_2(P)$-metric and for this reason we shall work with the simpler $L_2(P)$-metric from now on.

Just as for the Glivenko-Cantelli theorem, there are two basic theorems that imply that a class of functions is Donsker, using bracketing or covering numbers. It is required that the numbers

$$N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big) \quad \text{or} \quad \sup_Q N\big(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_2(Q)\big)$$

do not grow too fast as $\varepsilon \downarrow 0$. The rate of growth is elegantly measured through the *bracketing integral* and the *uniform entropy integral* defined as

$$J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) = \int_0^\delta \sqrt{\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\, d\varepsilon,$$

$$J\big(\delta, \mathcal{F}, L_2\big) = \int_0^\delta \sqrt{\log \sup_Q N\big(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)\big)}\, d\varepsilon.$$

The convergence of these integrals depends only on the size of the integrands as $\varepsilon \downarrow 0$. Because $\int_0^1 \varepsilon^{-r}\, d\varepsilon$ converges for $r < 1$ and diverges for $r \ge 1$, convergence of the integrals roughly requires that the entropies grow at slower order than $(1/\varepsilon)^2$.

6.8 Theorem (Donsker theorem). *Every class $\mathcal{F}$ of measurable functions with $J_{[]}\big(1, \mathcal{F}, L_2(P)\big) < \infty$ is P-Donsker.*

6.9 Theorem (Donsker theorem). *Every suitably measurable class $\mathcal{F}$ of measurable functions with $J(1, \mathcal{F}, L_2) < \infty$ and $P^*F^2 < \infty$ is P-Donsker.*

The condition that the class $\mathcal{F}$ be "suitably measurable" is satisfied in most examples, but cannot be omitted. We do not give a general definition here, but note that it suffices that there exists a countable collection $\mathcal{G}$ of functions such that each f is the pointwise limit of a sequence g_m in $\mathcal{G}$. We shall call a class with this property *separable*.

As remarked in the preceding lecture, many estimates of the bracketing or uniform entropy are available in the literature and can be used to derive concrete Donsker classes. Alternatively, new Donsker classes can be constructed out of known Donsker classes. The following theorem is in this spirit and will be useful.

For ease of terminology we call a collection of measurable functions $f\colon \mathcal{X} \mapsto \mathbb{R}^k$ Donsker if each of the k collections of coordinate functions is Donsker.

6.10 Theorem. *If $\mathcal{F}$ is a Donsker class of functions $f: \mathcal{X} \mapsto \mathbb{R}^k$ with square-integrable envelope, and $\phi: \mathbb{R}^k \mapsto \mathbb{R}$ is Lipschitz, then the class of functions $\phi \circ f: \mathcal{X} \mapsto \mathbb{R}$ is Donsker provided that it has a square-integrable envelope.*

Our result on Z-estimators should cover the classical results, which are obtained by Taylor expansions. This concerns classes of functions $\psi_\theta: \mathcal{X} \mapsto \mathbb{R}^k$, where θ ranges over a bounded subset of $\mathbb{R}^k$ and the dependence $\theta \mapsto \psi_\theta$ is "smooth". The following lemma gives a bound on the entropy of such a class, which shows that these classes are very easily Donsker.

6.11 Lemma (Parametric class). *Let $\mathcal{F} = \{f_\theta: \theta \in \Theta\}$ be a collection of measurable functions indexed by a bounded subset $\Theta \subset \mathbb{R}^d$. Suppose that there exists a measurable function m such that*

$$\big|f_{\theta_1}(x) - f_{\theta_2}(x)\big| \le m(x)\|\theta_1 - \theta_2\|, \qquad \text{every } \theta_1, \theta_2.$$

If $P|m|^r < \infty$, then there exists a constant K, depending on Θ and d only, such that the bracketing numbers satisfy

$$N_{[]}\big(\varepsilon\|m\|_{P,r}, \mathcal{F}, L_r(P)\big) \le K\Big(\frac{\operatorname{diam}\Theta}{\varepsilon}\Big)^d, \qquad \text{every } 0 < \varepsilon < \operatorname{diam}\Theta.$$

Proof. We use brackets of the type $[f_\theta - \varepsilon m, f_\theta + \varepsilon m]$ for θ ranging over a suitably chosen subset of Θ. These brackets have $L_r(P)$-size $2\varepsilon\|m\|_{P,r}$. If θ ranges over a grid of meshwidth ε over Θ, then the brackets cover $\mathcal{F}$, since, by the Lipschitz condition, $f_{\theta_1} - \varepsilon m \le f_{\theta_2} \le f_{\theta_1} + \varepsilon m$ if $\|\theta_1 - \theta_2\| \le \varepsilon$. Thus, we need as many brackets as we need balls of radius $\varepsilon/2$ to cover Θ.

The size of Θ in every fixed dimension is at most $\operatorname{diam}\Theta$. We can cover Θ with fewer than $(\operatorname{diam}\Theta/\varepsilon)^d$ cubes of size ε. The circumscribed balls have radius a multiple of ε and also cover Θ. If we replace the centers of these balls by their projections into Θ, then the balls of twice the radius still cover Θ. ∎

6.3 Maximal Inequalities

We do not include the proofs of the two Donsker theorems here, but we do include the basic *maximal inequalities*, on which the proofs rest. These are bounds on the distribution of the supremum variables $\|\mathbb{G}_n\|_{\mathcal{F}}$. For our main purpose inequalities on the L_1-norm of these variables are sufficient. We use these inequalities in the next section to treat empirical processes indexed by random functions. Actually, the Theorem 6.15 obtained there can easily be turned into a proof of the Donsker theorems.

6.12 Lemma. *For any class $\mathcal{F}$ of measurable functions $f: \mathcal{X} \mapsto \mathbb{R}$ such that $Pf^2 < \delta^2$ for every f, we have, with $a(\delta) = \delta/\sqrt{\operatorname{Log} N_{[]}(\delta, \mathcal{F}, L_2(P))}$,*

$$\mathrm{E}_P^*\|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big) + \sqrt{n}P^*F\big\{F > \sqrt{n}a(\delta)\big\}.$$

6.13 Corollary. *For any class $\mathcal{F}$ of measurable functions with envelope function F,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\|F\|_{P,2}, \mathcal{F}, L_2(P)\big).$$

Proof. Since $\mathcal{F}$ is contained in the single bracket $[-F, F]$, the bracketing number $N_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)$ can be taken equal to 1 for $\delta = 2\|F\|_{P,2}$. Then the constant $a(\delta)$ as defined in the preceding lemma reduces to a multiple of $\|F\|_{P,2}$, and $\sqrt{n}P^*F\big\{F > \sqrt{n}a(\delta)\big\}$ is bounded above by a multiple of $\|F\|_{P,2}$, by Markov's inequality. ∎

6.14 Lemma. *For any suitably measurable class $\mathcal{F}$ of measurable functions $f\colon \mathcal{X} \mapsto \mathbb{R}$, we have, with $\theta_n^2 = \sup_{f\in\mathcal{F}} \mathbb{P}_n f^2 / \mathbb{P}_n F^2$,*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \mathrm{E}\big[J(\theta_n, \mathcal{F}, L_2)\|F\|_{\mathbb{P}_n,2}\big] \lesssim J(1, \mathcal{F}, L_2)\|F\|_{P,2}.$$

6.4 Random Functions

In Lecture 10 we shall use the preceding theorems directly to ensure that certain stochastic processes appearing in the asymptotic analysis of Z-estimators converge in distribution. However, our main use for Donsker classes in these lectures is indirect: they give a tool to show study averages of "random functions". Here by "random functions" we mean measurable functions $x \mapsto \hat{f}_n(x; X_1, \ldots, X_n)$ that, for every fixed x, are functions of the observations. We write $\hat{f}_n$ for $\hat{f}_n(\cdot; X_1, \ldots, X_n)$ and use the notations $\mathbb{P}_n\hat{f}_n$ and $P\hat{f}_n$ as abbreviations for the expectations of the functions $x \mapsto \hat{f}_n(x; X_1, \ldots, X_n)$ with $X_1, \ldots, X_n$ fixed. Thus

$$\mathbb{G}_n\hat{f}_n = \frac{1}{\sqrt{n}}\Big(\sum_{i=1}^n \hat{f}_n(X_i; X_1, \ldots, X_n) - P\hat{f}_n\Big),$$

$$P\hat{f}_n = \int \hat{f}_n(x; X_1, \ldots, X_n)\, dP(x).$$

Note that $\mathbb{G}_n\hat{f}_n$ is not centered at mean zero, although it could be considered centered in a wide sense.

Obviously, the central limit theorem does not apply to a sequence of the form $\mathbb{G}_n\hat{f}_n$. However, if the functions $\hat{f}_n$ are sufficiently stable, then its result is still true.

6.15 Theorem. *If there exists a P-Donsker class $\mathcal{F}$ such that $P^n(\hat{f}_n \in \mathcal{F}) \to 1$ and $P(\hat{f}_n - f_0)^2 \to 0$ in probability, for some $f_0 \in L_2(P)$, then $\mathbb{G}_n(\hat{f}_n - f_0) \to 0$ in probability.*

Proof. Assume without of loss of generality that f_0 is contained in $\mathcal{F}$. Define a function $g: \ell^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ by $g(z, f) = z(f) - z(f_0)$. The set $\mathcal{F}$ is a semimetric space relative to the $L_2(P)$-metric. The function g is continuous with respect to the product semimetric at every point (z, f) such that $f \mapsto z(f)$ is continuous. Indeed, if $(z_n, f_n) \to (z, f)$ in the space $\ell^\infty(\mathcal{F}) \times \mathcal{F}$, then $z_n \to z$ uniformly and hence $z_n(f_n) = z(f_n) + o(1) \to z(f)$ if z is continuous at f.

By assumption, $\hat{f}_n \xrightarrow{\mathrm{P}} f_0$ as maps in the metric space $\mathcal{F}$. Since $\mathcal{F}$ is Donsker, $\mathbb{G}_n \rightsquigarrow \mathbb{G}_P$ in the space $\ell^\infty(\mathcal{F})$, and it follows that $(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow (\mathbb{G}_P, f_0)$ in the space $\ell^\infty(\mathcal{F}) \times \mathcal{F}$. By Lemma 6.6, almost all sample paths of $\mathbb{G}_P$ are continuous on $\mathcal{F}$. Thus the function g is continuous at almost every point $(\mathbb{G}_P, f_0)$. By the continuous mapping theorem, $\mathbb{G}_n(\hat{f}_n - f_0) = g(\mathbb{G}_n, \hat{f}_n) \rightsquigarrow g(\mathbb{G}_P, f_0) = 0$. The lemma follows, since convergence in distribution and convergence in probability are the same for a degenerate limit. ∎

Employing a fixed Donsker class in the preceding lemma gives a useful, relatively simple condition for getting rid of randomness in the function $\hat{f}_n$. The lemma covers many examples. However, other methods may give better results. Sometimes it is possible to study $\mathbb{G}_n(\hat{f}_n - f_0)$ by direct methods, such as computing means and variances. In other situations it is good to know that what is really needed is not that the functions $\hat{f}_n$ remain within a fixed class, as $n \to \infty$, but that the complexity of the set of functions $\hat{f}_n$ does not increase too much with n. We can make this precise through a formulation using entropy conditions. On the one hand this gives more flexibility. On the other hand, nice results such as Theorem 6.10, which allow a calculus to create new Donsker classes, become unavailable.

In the next theorem we require that the realizations of the random functions $\hat{f}_n$ belong to classes $\mathcal{F}_n$ that may change with n. We assume that these classes possess envelope functions F_n that satisfy the Lindeberg condition

$$PF_n^2 = O(1),$$
$$PF_n^2\{F_n > \varepsilon\sqrt{n}\} \to 0, \qquad \text{every } \varepsilon > 0.$$

Then the result of the preceding theorem remains true provided the entropy integrals of the classes behave well.

6.16 Theorem. *Let $\mathcal{F}_n$ be classes of measurable functions such that $P^n(\hat{f}_n \in \mathcal{F}_n) \to 1$ and such that either*
(i) $J_{[]}\big(\delta_n, \mathcal{F}_n, L_2(P)\big) \to 0$ or
(ii) $J(\delta_n, \mathcal{F}_n, L_2) \to 0$, for every $\delta_n \downarrow 0$,
and with envelope functions that satisfy the Lindeberg condition. In the case of (ii) also assume that the classes are suitably measurable. If $P(\hat{f}_n - f_0)^2 \to 0$ in probability for some $f_0 \in L_2(P)$, then $\mathbb{G}_n(\hat{f}_n - f_0) \to 0$.

Proof. Without loss of generality assume that $f_0 = 0$. Otherwise, replace $\mathcal{F}_n$ by $\mathcal{F}_n - f_0$ and $\hat{f}_n$ by $\hat{f}_n - f_0$.

First assume that (i) holds. Let $\mathcal{G}_n(\delta)$ be the set of functions $\{f \in \mathcal{F}_n : Pf^2 \leq \delta^2\}$ By assumption we have that $P^n\big(\hat{f}_n \in \mathcal{G}_n(\delta)\big) \to 1$ as $n \to \infty$, for every $\delta > 0$. On the event $\{\hat{f}_n \in \mathcal{G}_n(\delta)\}$ we have $|\mathbb{G}_n \hat{f}_n| \leq \sup_{g \in \mathcal{G}_n(\delta)} |\mathbb{G}_n g|$. By Lemma 6.12

$$\mathrm{E}^* \sup_{g \in \mathcal{G}_n(\delta)} |\mathbb{G}_n g| \lesssim J_{[]}\big(\delta, \mathcal{G}_n(\delta), L_2(P)\big) + \frac{PF_n^2 1\big\{F_n > a_n(\delta)\sqrt{n}\big\}}{a_n(\delta)},$$

where $a_n(\delta)$ is the number given in Lemma 6.12 evaluated for the class of functions $\mathcal{G}_n(\delta)$. The first term on the right increases if we replace $\mathcal{G}_n(\delta)$ by $\mathcal{F}_n$ and hence converges to zero as $\delta \to 0$. Since $J_{[]}\big(\delta_n, \mathcal{F}_n, L_2(P)\big) \to 0$ for every $\delta_n \downarrow 0$, we must have that $J_{[]}\big(\delta, \mathcal{F}_n, L_2(P)\big) = O(1)$ for every $\delta > 0$ and hence

$$\delta\sqrt{\log N_{[]}\big(\delta, \mathcal{F}_n, L_2(P)\big)} \leq J_{[]}\big(\delta, \mathcal{F}_n, L_2(P)\big) = O(1).$$

Therefore, $a_n(\delta)$ is bounded away from zero, for every fixed δ as $n \to \infty$. Conclude that $PF_n^2 1\big\{F_n > a_n(\delta)\sqrt{n}\big\} \to 0$ as $n \to \infty$ followed by $\delta \to 0$. The proof under (i) is complete.

Next assume that (ii) holds. The class $\mathcal{G}_n(\delta)$, defined as before, has envelope function $1 + F_n$ and hence, by Lemma 6.14,

$$\mathrm{E}^* \sup_{g \in \mathcal{G}_n(\delta)} |\mathbb{G}_n g| \lesssim \mathrm{E}^*\Big[J\big(\theta_n(\delta), \mathcal{G}_n(\delta), L_2\big)\sqrt{\mathbb{P}_n(1+F_n)^2}\Big], \tag{6.17}$$

for J the uniform entropy integral of $\mathcal{G}_n(\delta)$ relative to the envelope function $1 + F_n$ and

$$\theta_n^2(\delta) = \frac{\|\mathbb{P}_n f^2\|_{\mathcal{G}_n(\delta)}}{\mathbb{P}_n(1+F_n)^2} \leq \|\mathbb{P}_n f^2\|_{\mathcal{G}_n(\delta)} \wedge 1.$$

The covering numbers of $\mathcal{G}_n(\delta)$ are bounded by by the covering numbers of $\mathcal{F}_n$ and hence the uniform entropy integral of $\mathcal{G}_n(\delta)$ is bounded by the uniform entropy integral of $\mathcal{F}_n$ if we compute them relative to the same envelope function. If for $\mathcal{F}_n$ we replace the envelope $1 + F_n$ by the natural envelop F_n, then the uniform entropy integral increases. Thus we can further bound the right side of (6.17) by

$$\Big[\mathrm{E}^* J^2\big(\theta_n(\delta), \mathcal{F}_n, L_2\big)\mathrm{E}^*\mathbb{P}_n(1+F_n)^2\Big]^{1/2}$$
$$\lesssim \Big[J^2(1, \mathcal{F}_n)\mathrm{P}^*\big(\theta_n(\delta) \geq \varepsilon\big) + J^2(\varepsilon, \mathcal{F}_n, L_2)\Big]^{1/2}\big(P(1+F_n)^2\big)^{1/2}.$$

We conclude that the theorem is proved if we can show that $\theta_n(\delta) \to 0$ in probability as $n \to \infty$, followed by $\delta \to 0$.

Fix $\eta > 0$. The class of functions $\mathcal{H}_n(\delta, \eta) = \big\{f^2 1_{F_n \leq \eta\sqrt{n}} : f \in \mathcal{G}_n(\delta)\big\}$ has envelope function $\eta\sqrt{n}F_n$. Hence by Lemma 6.14

$$\mathrm{E}^*\|\mathbb{G}_n\|_{\mathcal{H}_n(\delta,\eta)} \lesssim J\big(1, \mathcal{H}_n(\delta, \eta), L_2\big)\|\eta\sqrt{n}F_n\|_{P,2}. \tag{6.18}$$

Because

$$Q(f^2 1_{F_n \leq \eta\sqrt{n}} - g^2 1_{F_n \leq \eta\sqrt{n}})^2 \leq Q(f-g)^2(2\eta\sqrt{n})^2,$$

we have

$$N\big(\varepsilon\|\eta\sqrt{n}F_n\|_{Q,2}, \mathcal{H}_n(\delta, \eta), L_2(Q)\big) \leq N\big(\tfrac{1}{2}\varepsilon\|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q)\big).$$

Inserting this in the right side of (6.18) we see that the left side of (6.18) is bounded by $J(1, \mathcal{F}_n, L_2)\eta\sqrt{n}\|F_n\|_{P,2}$. We conclude that $\mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{H}_n(\delta,\eta)} \to 0$ as $n \to \infty$ followed by $\eta \to 0$.

For any fixed $\eta > 0$ the class of functions $\overline{\mathcal{H}}_n(\delta, \eta) = \{f^2 1_{F_n > \eta\sqrt{n}} : f \in \mathcal{G}_n(\delta)\}$ satisfies

$$\mathrm{E}^*\|\mathbb{P}_n - P\|_{\overline{\mathcal{H}}_n(\delta,\eta)} \le 2PF_n^2 1_{F_n > \eta\sqrt{n}} \to 0.$$

Combined with the result of the preceding paragraph this yields $\mathrm{E}^*\|\mathbb{P}_n - P\|_{\mathcal{G}_n(\delta)^2} \to 0$, as $n \to \infty$, for every $\delta > 0$. Because also $\|P\|_{\mathcal{G}_n(\delta)^2} \le \delta^2$ by the definition of the class $\mathcal{G}_n(\delta)$, we conclude that $\|\mathbb{P}_n\|_{\mathcal{G}_n(\delta)^2} \to 0$ as $n \to \infty$ followed by $\delta > 0$. This concludes the proof. ∎

6.5 Asymptotic Normality of Z-Estimators

In the preceding lecture we showed that a Z-estimator $\hat{\theta}$, defined as a zero of a random criterion function $\theta \mapsto \mathbb{P}_n\psi_\theta$, is typically consistent for a zero of the limiting criterion function $\theta \mapsto P\psi_\theta$. The asymptotic distribution of the difference $\hat{\theta} - \theta$ depends on the fluctuations of the random criterion function $\mathbb{P}_n\psi_\theta$ around its limit $P\psi_\theta$. Empirical processes are what we need to study such fluctuations.

We start with a simple theorem. For every θ in a set $\Theta \subset \mathbb{R}^k$ let $\psi_\theta : \mathcal{X} \mapsto \mathbb{R}^k$ be a measurable, vector-valued function.

6.19 Theorem. *Suppose that the class of functions $\{\psi_\theta : \theta \in \Theta\}$ is P-Donsker, that the map $\theta \mapsto P\psi_\theta$ is differentiable at θ_0 with nonsingular derivative V, and that the map $\theta \mapsto \psi_\theta$ is continuous in $L_2(P)$ at θ_0. Then any $\hat{\theta}_n$ such that $\mathbb{P}_n\psi_{\hat{\theta}_n} = 0$ and such that $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ for a zero θ_0 of $\theta \mapsto P\psi_\theta$ satisfies*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

Proof. The consistency of $\hat{\theta}_n$ and the Donsker condition on the functions ψ_θ imply that

$$\mathbb{G}_n\psi_{\hat{\theta}_n} - \mathbb{G}_n\psi_{\theta_0} \xrightarrow{\mathrm{P}} 0. \tag{6.20}$$

By the definitions of $\hat{\theta}_n$ and θ_0, we can rewrite $\mathbb{G}_n\psi_{\hat{\theta}_n}$ as $\sqrt{n}P(\psi_{\theta_0} - \psi_{\hat{\theta}_n}) + o_P(1)$. Combining this with the Delta-method and the differentiability of the map $\theta \mapsto P\psi_\theta$, we find that

$$\sqrt{n}V_{\theta_0}(\theta_0 - \hat{\theta}_n) + \sqrt{n}\, o_P\big(\|\hat{\theta}_n - \theta_0\|\big) = \mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

In particular, by the invertibility of the matrix V_{θ_0},

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\| \le \|V_{\theta_0}^{-1}\|\sqrt{n}\big\|V_{\theta_0}(\hat{\theta}_n - \theta_0)\big\| = O_P(1) + o_P\big(\sqrt{n}\|\hat{\theta}_n - \theta_0\|\big).$$

This implies that $\hat\theta_n$ is $\sqrt{n}$-consistent: the left side is bounded in probability. Inserting this in the previous display, we obtain that $\sqrt{n}V_{\theta_0}(\hat\theta_n - \theta_0) = -\mathbb{G}_n\psi_{\theta_0} + o_P(1)$. We conclude the proof by taking the inverse $V_{\theta_0}^{-1}$ left and right. Since matrix multiplication is a continuous map, the inverse of the remainder term still converges to zero in probability. ∎

This theorem as stated covers most (or all?) of the popular examples of Z-estimators, the condition that the functions ψ_θ form a Donsker class being not at all very restrictive. The Donsker class condition is used to ensure (6.20) and can be relaxed to

$$\mathbb{G}_n(\psi_{\hat\theta_n} - \psi_{\theta_0}) = o_P\big(1 + \sqrt{n}\|\hat\theta_n - \theta_0\|\big)$$

without changing the remainder of the proof. Of course, this or (6.20) does not really require that the class $\{\psi_\theta\colon \|\theta - \theta_0\| < \delta\}$ is Donsker for any fixed δ, but concerns a limiting property of these classes as $\delta \to 0$. Potentially, the Donsker condition could be relaxed to a condition that directly involves entropy numbers.

Such a relaxation does not appear to be worth the trouble in the situation of the preceding theorem, but is potentially of use in situations with nuisance parameters or criterion functions that change with n.

6.6 Nuisance parameters

An important method of estimation for semiparametric models, but also in general, is Z-estimation in the presence of nuisance parameters. We are given measurable functions $\psi_{\theta,\eta}\colon \mathcal{X} \mapsto \mathbb{R}^k$ indexed by a parameter of interest $\theta \in \mathbb{R}^k$ and a nuisance parameter η belonging to some metric space. Given an initial estimator $\hat\eta$ for η, we consider the (near) solution $\hat\theta$ of the equation $\mathbb{P}_n\psi_{\theta,\hat\eta} = 0$.

6.21 Theorem. *Suppose that the class of functions $\big\{\psi_{\theta,\eta}\colon \|\theta - \theta_0\| < \delta, d(\eta,\eta_0) < \delta\big\}$ is Donsker for some $\delta > 0$, that the maps $\theta \mapsto P\psi_{\theta,\eta}$ are differentiable at θ_0, uniformly in η in a neighbourhood of η_0 with nonsingular derivative matrices $V_{\theta_0,\eta}$ such that $V_{\theta_0,\eta} \to V_{\theta_0,\eta_0}$, and assume that the map $(\theta,\eta) \mapsto \psi_{\theta,\eta}$ is continuous in $L_2(P)$ at (θ_0,η_0). If $\sqrt{n}\,\mathbb{P}_n\psi_{\hat\theta_n,\hat\eta_n} = o_P(1)$ and $(\hat\theta_n,\hat\eta_n) \xrightarrow{\mathrm{P}} (\theta_0,\eta_0)$ for a point (θ_0,η_0) satisfying $P\psi_{\theta_0,\eta_0} = 0$, then*

$$\sqrt{n}(\hat\theta_n - \theta_0) = -V_{\theta_0,\eta_0}^{-1}\sqrt{n}\,P\psi_{\theta_0,\hat\eta_n} - V_{\theta_0,\eta_0}^{-1}\mathbb{G}_n\psi_{\theta_0,\eta_0}(X_i) + o_P\big(1 + \sqrt{n}\|P\psi_{\theta_0,\hat\eta_n}\|\big).$$

Proof. The proof closely follows the proof of the theorem without nuisance parameters. The consistency of $(\hat\theta_n,\hat\eta_n)$ and the Donsker condition imply that

$$\mathbb{G}_n\psi_{\hat\theta_n,\hat\eta_n} - \mathbb{G}_n\psi_{\theta_0,\eta_0} \xrightarrow{\mathrm{P}} 0. \tag{6.22}$$

Because $(\hat\theta_n, \hat\eta_n)$ and (θ_0, η_0) are zeros of the random criterion function and its limit, we can rewrite this as

$$\begin{aligned} -\mathbb{G}_n \psi_{\hat\theta_0,\hat\eta_0} &= \sqrt{n} P(\psi_{\hat\theta_n,\hat\eta_n} - \psi_{\theta_0,\eta_0}) + o_P(1) \\ &= \sqrt{n}(P(\psi_{\hat\theta_n,\hat\eta_n} - \psi_{\theta_0,\hat\eta_n}) + \sqrt{n} P\psi_{\theta_0,\hat\eta_n} + o_P(1). \end{aligned} \tag{6.23}$$

By the uniform differentiability of the map $\theta \mapsto P\psi_\theta$ and the uniform nonsingularity of its derivative, we find that there exists $c > 0$ such that for all (θ, η) in a sufficiently small neighbourhood of (θ_0, η_0)

$$\big\| P(\psi_{\theta,\eta} - \psi_{\theta_0,\eta_0}) \big\| \geq c \|\theta - \theta_0\|.$$

Combined with the preceding display this shows that with probability tending to one,

$$c\|\hat\theta - \theta_0\| \leq \big\| \mathbb{G}_n \psi_{\hat\theta_0,\hat\eta_0} \big\| + \sqrt{n} \| P\psi_{\theta_0,\hat\eta_n} \| = O_P\big(1 + \sqrt{n} \| P\psi_{\theta_0,\hat\eta_n} \|\big).$$

We now linearize the first term on the far right of (6.23) in $\hat\theta - \theta_0$ and finish the proof as before. ∎

Under the conditions of this theorem, the limiting distribution of the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ depends on the estimator $\hat\eta_n$ through the "drift" term $\sqrt{n} P\psi_{\theta_0,\hat\eta_n}$. In general, this gives a contribution to the limiting distribution, and $\hat\eta_n$ must be chosen with care. If $\hat\eta_n$ is $\sqrt{n}$-consistent and the map $\eta \mapsto P\psi_{\theta_0,\eta}$ is differentiable, then the drift term can be analyzed using the Delta-method.

It may happen that the drift term is zero. If the parameters θ and η are "orthogonal" in this sense, then the auxiliary estimators $\hat\eta_n$ may converge at an arbitrarily slow rate and affect the limit distribution of $\hat\theta_n$ only through their limiting value η_0. In semiparametric situations it is quite common to set up the estimating equations such that the drift term gives a zero contribution. Then the advantage of using a random value $\hat\eta_n$ over a fixed value could be a gain in efficiency: we choose $\hat\eta_n$ to converge to a value η_0 such that the asymptotic covariance matrix

$$V_{\theta_0,\eta_0}^{-1} P\psi_{\theta_0,\eta_0} \psi_{\theta_0,\eta_0}^T V_{\theta_0,\eta_0}^T$$

is "small".

This theorem and discussion is valid whether (θ, η) completely parametrizes a model, or not. In the first case, we would write a true distribution P_{θ_0,η_0} rather than as P. The asymptotic covariance matrix in the preceding display would then be at least equal to the inverse of the efficient information matrix. It would be equal to this if $\psi_{\theta,\eta}$ is proportional to the the efficient score function for θ.

6.24 Example (Regression). Let a typical observation be a pair $X = (Y, Z)$ whose distribution is described structurally by the equation $Y = f_\theta(Z) + e$ for (Z, e) having a distribution η such that $\mathrm{E}_\eta(e|\, Z) = 0$.

Consider the estimation equation defined by

$$\psi_{\theta,\eta}(x) = \big(y - f_\theta(z)\big) w_{\theta,\eta}(z),$$

for given weight functions $w_{\theta,\eta}$. We have

$$P_{\theta_0,\eta_0}\psi_{\theta_0,\eta} = \mathrm{E}_{\theta_0,\eta_0}\big[\mathrm{E}_{\theta_0,\eta_0}\big(Y - f_{\theta_0}(Z)\big)\big|\, Z\big]w_{\theta_0,\eta}(Z) = 0.$$

Thus the drift term in the preceding theorem vanishes. To obtain an efficient estimator we must choose the weight function equal to $w_{\theta,\eta}(z) = \dot{g}_\theta(z)/\mathrm{E}_\eta(e^2|\, Z = z)$ and use estimators for η such that $w_{\theta,\hat\eta}$ is consistent for this weight function, but (almost) any choice of the weight function will work to obtain an asymptotically normal estimator. One explanation for the fact that these estimating equations are unbiased is that the functions belong to the orthocomplement of the tangent set. □

In a number of models, such as the regression model in the preceding example, setting up good estimating equations is easy. In general, calculation of the tangent set of a model, or rather its orthocomplement can be of help. First, if some function $\psi_{\theta,\eta}$ is orthogonal to the tangent set due to the nuisance parameters, its mean $P_{\theta,\eta}\psi_{\theta,\hat\eta}$ should be fairly insensitive to the estimator $\hat\eta$, because by definition a nuisance score gives the change in the underlying distribution if perturbing the nuisance parameter. One attempt to make this idea formal is to write

$$P_{\theta,\eta}\psi_{\theta,\hat\eta} = (P_{\theta,\eta} - P_{\theta,\hat\eta})(\psi_{\theta,\hat\eta} - \psi_{\theta,\eta}) - P_{\theta,\eta}\Big[\frac{p_{\theta,\hat\eta} - p_{\theta,h}}{p_{\theta,\eta}} - B_{\theta,\eta}h\Big]\psi_{\theta,\eta}, \tag{6.25}$$

where $B_{\theta,\eta}h$ can be any η-score if $\psi_{\theta,\eta}$ is orthogonal to the nuisance tangent space. If $B_{\theta,\eta}h$ can approximate $(p_{\theta,\hat\eta} - p_{\theta,h})/p_{\theta,\eta}$, then we might hope that the right side of the display is of the order $O_P(d(\hat\eta,\eta)^2)$, for the metric d giving the approximation. Then the drift term will give no contribution to the limit distribution if $d(\hat\eta,\eta) = o_P(n^{-1/4})$. This informal argument can be useful, but it should not be concluded that a $n^{-1/4}$-rate for the nuisance parameter is "minimal" in some sense. Special properties of the model, as in the regression example, may make the drift term zero for any $\hat\eta$. The point is that $P\psi_{\theta,\hat\eta}$ is an integrated quantity and it is far to crude to analyse it by a Taylor expansion, replacing the integrand by its absolute value after subtracting the beginning of the expansion.

Nevertheless, we can formalize the expansion, for instance, as follows. Given some semiparametric model $\mathcal{P} = \{P_{\theta,\eta}\colon \theta \in \Theta, \eta \in H\}$ with H a metric space, suppose that, for some nonnegative numbers α, β, γ,

$$\begin{aligned} P_{\theta,\eta}\|\psi_{\theta,\hat\eta} - \psi_{\theta,\eta}\|^2 &= O_P\big(d(\hat\eta,\eta)^{2\alpha}\big) \\ \inf_{g\in\mathrm{lin}\ \dot{\mathcal{P}}_{P_{\theta,\eta}}} P_{\theta,\eta}\Big[\frac{p_{\theta,\hat\eta} - p_{\theta,\eta}}{p_{\theta,\eta}} - g\Big]^2 &= O_P\big(d(\hat\eta,\eta)^{2\beta+2\gamma}\big) \\ P_{\theta,\eta}\Big[\frac{p_{\theta,\hat\eta} - p_{\theta,\eta}}{p_{\theta,\eta}}\Big]^2 &= O_P\big(d(\hat\eta,\eta)^{2\beta}\big). \end{aligned}$$

Then $P_{\theta,\eta}\psi_{\theta,\hat\eta} = O_P\big(d(\hat\eta,\eta)^\delta\big)$ for $\delta = (a \vee \gamma) + \beta$.

If the underlying measure $P = P_{\theta,\eta}$ belongs to a semiparametric model, then it is worth while to adapt the conditions of Theorem 6.21 somewhat and to use the differentiability of the model in θ. This leads to the following theorem, which we shall apply in the next lectures to construct efficient estimators or analyse the maximum likelihood estimator. We now make the disappearance of the bias term part of the conditions.

Let $\psi_{\theta,\eta}: \mathcal{X} \mapsto \mathbb{R}$ be measurable functions and and let $\hat{\eta}_n$ be estimators such that

$$P_{\hat{\theta}_n,\eta}\psi_{\hat{\theta}_n,\hat{\eta}_n} = o_P\big(n^{-1/2} + \|\hat{\theta}_n - \theta\|\big), \tag{6.26}$$

$$P_{\theta,\eta}\big\|\psi_{\hat{\theta}_n,\hat{\eta}_n} - \psi_{\theta,\eta}\big\|^2 \stackrel{\mathrm{P}}{\to} 0, \quad P_{\hat{\theta}_n,\eta}\big\|\psi_{\hat{\theta}_n,\hat{\eta}_n}\big\|^2 = O_P(1). \tag{6.27}$$

The second condition (6.27) merely requires that the "plug-in" estimator $\psi_{\theta,\hat{\eta}_n}$ is a consistent estimator for the "true" estimating function $\psi_{\theta,\eta}$. If $P_{\theta,\eta}\psi_{\theta,\eta} = 0$, as we shall require, then the first condition (6.26) can be understood as requiring that the "bias" of the plug-in estimator, due to estimating the nuisance parameter, converges to zero faster than $1/\sqrt{n}$. Note that the derivative of $\theta \mapsto P_{\theta,\eta}\psi_{\theta,\hat{\eta}}$ should converge to the derivative of $\theta \mapsto P_{\theta,\eta}\psi_{\theta,\eta}$, which is zero, and hence, informally the condition (6.26) must be equivalent to

$$\sqrt{n}P_{\theta,\eta}\psi_{\theta,\hat{\eta}_n} \stackrel{\mathrm{P}}{\to} 0, \tag{6.28).}$$

6.29 Theorem. *Suppose that the model $\{P_{\theta,\eta}: \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ,η). Let the matrix $P_{\theta,\eta}\psi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$ be nonsingular. Assume that (6.26) and (6.27) hold. Furthermore, suppose that there exists a Donsker class with square-integrable envelope function that contains every function $\psi_{\hat{\theta}_n,\hat{\eta}_n}$ with probability tending to 1. Then a zero $\hat{\theta}_n$ of $\theta \mapsto \mathbb{P}_n\psi_{\theta,\hat{\eta}_n}$ that is consistent for θ satisfies that $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal with mean zero and covariance matrix*

$$(P_{\theta,\eta}\psi_{\theta,\eta}\dot{\ell}_{\theta,\eta}^T)^{-1}P\psi_{\theta,\eta}\psi_{\theta,\eta}^T(P_{\theta,\eta}\dot{\ell}_{\theta,\eta}\psi_{\theta,\eta}^T)^{-1}.$$

Proof. Let $G_n(\theta',\eta') = \sqrt{n}(\mathbb{P}_n - P_{\theta,\eta})\psi_{\theta',\eta'}$ be the empirical process indexed by the functions $\psi_{\theta',\eta'}$. By the assumption that the functions $\psi_{\hat{\theta},\hat{\eta}}$ are contained in a Donsker class, together with (6.27),

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

(Cf. Theorem 6.15.) By the defining relationship of $\hat{\theta}_n$ and the "no-bias" condition (6.26), this is equivalent to

$$\sqrt{n}(P_{\hat{\theta}_n,\eta} - P_{\theta,\eta})\psi_{\hat{\theta}_n,\hat{\eta}_n} = G_n(\theta,\eta) + o_P\big(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|\big).$$

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

$$\begin{aligned}
&\sqrt{n}\int \psi_{\hat{\theta}_n,\hat{\eta}_n}(p_{\hat{\theta}_n,\eta}^{1/2} + p_{\theta,\eta}^{1/2})\big[(p_{\hat{\theta}_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2}) - \tfrac{1}{2}(\hat{\theta}_n - \theta)^T\dot{\ell}_{\theta,\eta}\,p_{\theta,\eta}^{1/2}\big]d\mu \\
&\quad + \int \psi_{\hat{\theta}_n,\hat{\eta}_n}(p_{\hat{\theta}_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2})\,\tfrac{1}{2}\dot{\ell}_{\theta,\eta}^T\,p_{\theta,\eta}^{1/2}\,d\mu\,\sqrt{n}(\hat{\theta}_n - \theta) \\
&\quad - \int (\psi_{\hat{\theta}_n,\hat{\eta}_n} - \psi_{\theta,\eta})\,\dot{\ell}_{\theta,\eta}^T\,p_{\theta,\eta}\,d\mu\,\sqrt{n}(\hat{\theta}_n - \theta).
\end{aligned}$$

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 (6.27). 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

$$m_n^2 \int \|\psi_{\hat\theta_n,\hat\eta_n}\|\, p_{\theta,\eta}^{1/2} |p_{\hat\theta_n,\eta}^{1/2} - p_{\theta,\eta}^{1/2}|\, d\mu^2 \\ + \int \|\psi_{\hat\theta_n,\hat\eta_n}\|^2 (p_{\hat\theta_n,\eta} + p_{\theta,\eta})\, d\mu \int_{\|\dot\ell_{\theta,\eta}\|>m_n} \|\dot\ell_{\theta,\eta}\|^2\, p_{\theta,\eta}\, d\mu.$$

In view of (6.27), 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 $m_n\|\hat\theta_n - \theta\| \xrightarrow{\mathrm{P}} 0$. (Such a sequence exists. If $Z_n \xrightarrow{\mathrm{P}} 0$, then there exists a sequence $\varepsilon_n \downarrow 0$ such that $\mathrm{P}\bigl(|Z_n| > \varepsilon_n\bigr) \to 0$. Then $\varepsilon_n^{-1/2} Z_n \xrightarrow{\mathrm{P}} 0$.) In view of the last part of (6.27), the second term converges to zero in probability for every $m_n \to \infty$. This concludes the proof of the theorem. ∎

In the preceding theorems we have assumed that the realizations of the functions $\psi_{\hat\theta,\hat\eta}$ are contained in a fixed Donsker class, with high probability. This condition is overly strong. As we pointed out in Lecture 6, what is needed is that the entropy of these collections of realizations is asymptotically stable and not too big. Hence the condition can be replaced by the condition that there exist classes $\mathcal{F}_n$ of functions satisfying the conditions of Theorem 6.16 such that $\psi_{\hat\theta,\hat\eta}$ is contained in $\mathcal{F}_n$ with probability tending to one. One further extension is to permit $\psi_{\theta,\eta}$ to change with n itself.

Notes

See the notes to Lecture 5. The general topic of Section 6.4 is taken from [42], but the main result here is new.

Lecture 7
Efficient Score and One-step Estimators

In this lecture we consider the construction of efficient estimators in semiparametric models using the efficient score equation or the related one-step method. We apply it to the linear errors-in-variables model and the symmetric location model.

7.1 Efficient Score Estimators

The most important method to estimate 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 neighbourhood 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*

$$\sum_{i=1}^n \tilde{\ell}_{\theta,\hat{\eta}_n}(X_i) = 0. \tag{7.1}$$

Here we use (a version of) the efficient score function instead of the ordinary score function, and we substitute an estimator $\hat{\eta}_n$ for the unknown nuisance parameter. Alternatively, it may be more workable to find an "estimator" $\hat{\eta}_n(\theta)$ for η acting as if θ is known already and next solve θ from the "profile efficient score equations"

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

A solution $\hat{\theta}_n$ also satisfies the efficient score equation (7.1) if we set $\hat{\eta}_n = \hat{\eta}_n(\hat{\theta}_n)$. This choice of $\hat{\eta}_n$ may beat the purpose of finding an estimator $\hat{\theta}_n$, but this remark does indicate that to prove something about $\hat{\theta}_n$ it is not necessary to consider the profile efficient score equation. Hence we concentrate on solutions of (7.1).

We can derive the asymptotic normality of $\hat{\theta}_n$ from Theorem 6.29. Here $P_{\theta,\eta}\tilde{\psi}_{\theta,\eta}\dot{\ell}_{\theta,\eta}^T$ is the efficient information matrix $\tilde{I}_{\theta,\eta}$ and hence the asymptotic covariance matrix in this theorem reduces to $\tilde{I}_{\theta,\eta}^{-1}$.

7.2 Theorem. *Suppose that conditions of Theorem 6.29 are satisfied with* $\psi_{\theta,\eta} = \tilde{\ell}_{\theta,\eta}$. *Then a consistent sequence of zeros* $\hat{\theta}_n$ *of* $\theta \mapsto \mathbb{P}_n \tilde{\ell}_{\theta,\hat{\eta}_n}$ *is asymptotically efficient for* $\psi(P_{\theta,\eta}) = \theta$ *at* (θ, η).

We remark again that the condition that the functions $\tilde{\ell}_{\theta,\eta}$ are contained in a fixed Donsker class can be relaxed along the lines of Theorem 6.16.

7.2 One-step Estimators

Theorem 7.2 applies to many examples, but its conditions are not the minimal ones to ensure existence of asymptotically efficient estimators. There are may ways in which its conditions can be relaxed, all leading to estimators that are less natural but have better properties, in theory. We shall immediately go to the most extreme modification, which can be shown to work whenever there is anything that works.

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 mesh width $n^{-1/2}$, which will simplify the constructions and proof. Then the one-step estimator is defined as

$$\hat{\theta}_n = \tilde{\theta}_n - \Big(\sum_{i=1}^n \hat{\ell}_{n,\tilde{\theta}_n,i} \hat{\ell}^T_{n,\tilde{\theta}_n,i}(X_i)\Big)^{-1} \sum_{i=1}^n \hat{\ell}_{n,\tilde{\theta}_n,i}(X_i),$$

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

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

Thus, for X_i belonging to the first half of the sample, we use an estimator $\hat{\ell}_{n,\theta,i}$ 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 $\hat{\ell}_{n,\theta,i}$ is always independent of X_i, simultaneously for X_i running through each of the two halves of the sample. The trick is not recommended in practice.

The conditions of the preceding theorem can now be relaxed in two ways: we can drop the Donsker condition and we need an analogue of the "no-bias" condition (6.26) only for deterministic sequences θ_n. We assume that, for every deterministic sequence $\theta_n = \theta + O(n^{-1/2})$,

$$\sqrt{n} P_{\theta_n,\eta} \hat{\ell}_{n,\theta_n} \xrightarrow{\mathrm{P}} 0, \qquad P_{\theta_n,\eta} \big\| \hat{\ell}_{n,\theta_n} - \tilde{\ell}_{\theta_n,\eta} \big\|^2 \xrightarrow{\mathrm{P}} 0. \tag{7.3}$$

$$\int \Big\| \tilde{\ell}_{\theta_n,\eta} dP^{1/2}_{\theta_n,\eta} - \tilde{\ell}_{\theta,\eta} dP^{1/2}_{\theta,\eta} \Big\|^2 \to 0. \tag{7.4}$$

7.5 Theorem. *Suppose that the model $\{P_{\theta,\eta}: \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ,η), let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be non-singular. Assume that (7.3) and (7.4) 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 $\sum \hat{\ell}_{n,\theta_n,i}(X_i)$ is a sum of conditionally independent terms, given the second half of the sample. Thus,

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

Both expressions converge to zero in probability by assumption (7.3). 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

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

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

Combining the preceding display with the result of Lemma 7.6, we find that

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

In view of the discretised nature of $\tilde{\theta}_n$, this remains true if the deterministic sequence θ_n is replaced by $\tilde{\theta}_n$. This follows, because, for a given M, on the event $\{\|\sqrt{n}\tilde{\theta}_n - \theta\| \le M\}$ the estimator $\tilde{\theta}_n$ can take on only finitely many values, with the total number of different values being bounded independent of n. Thus an expression of the type $G_n(\tilde{\theta}_n)$ can be bounded above by $\sup_{\theta_n} \mathcal{G}_n(\theta_n)$ for the supremum ranging over a finite number of points. If each of the sequences $G_n(\theta_n)$ converges to zero in probability, then $G_n(\tilde{\theta}_n)$ converges to zero in probability on the event $\{\|\sqrt{n}\tilde{\theta}_n - \theta\| \le M\}$. Finally, by the assumed $\sqrt{n}$-consistently of $\tilde{\theta}_n$, we can fix M such that the probability of this event is arbitrarily close to 1.

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

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

By (7.3), 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 (7.4) and the law of large numbers, we see that

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

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

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

7.6 Lemma. *Suppose that the model $\{P_{\theta,\eta}: \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ,η), let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular, and assume that (7.4) holds. Then, for any $\theta_n = \theta + O(n^{-1/2})$,*

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

Proof. By the definition of the efficient score function as an orthogonal projection, $P_{\theta,\eta}\tilde{\ell}_{\theta,\eta}\dot{\ell}^T_{\theta,\eta} = \tilde{I}_{\theta,\eta}$. We shall use this identity several times in the following proof.

The lemma follows from adding the two assertions

$$\begin{aligned}
&\sqrt{n}\mathbb{P}_n\left(\tilde{\ell}_{\theta_n,\eta}\left(1 - \frac{p^{1/2}_{\theta_n,\eta}}{p^{1/2}_{\theta,\eta}}\right)\right) + \tfrac{1}{2}\tilde{I}_{\theta,\eta}\sqrt{n}(\theta_n - \theta) \xrightarrow{\mathrm{P}} 0 \\
&\sqrt{n}\mathbb{P}_n\left(\tilde{\ell}_{\theta_n,\eta}\frac{p^{1/2}_{\theta_n,\eta}}{p^{1/2}_{\theta,\eta}} - \tilde{\ell}_{\theta,\eta}\right) + \tfrac{1}{2}\tilde{I}_{\theta,\eta}\sqrt{n}(\theta_n - \theta) \xrightarrow{\mathrm{P}} 0.
\end{aligned} \tag{7.7}$$

For the second assertion we note that the variance of the variable on the left side under (θ,η) converges to zero by (7.4). Furthermore, the mean of this variable is equal to

$$\sqrt{n}\int \tilde{\ell}_{\theta_n,\eta}p^{1/2}_{\theta_n,\eta}p^{1/2}_{\theta,\eta}\,d\mu = \sqrt{n}\int \tilde{\ell}_{\theta_n,\eta}p^{1/2}_{\theta_n,\eta}(p^{1/2}_{\theta,\eta} - p^{1/2}_{\theta_n,\eta})\,d\mu.$$

This is asymptotically equivalent to $-\frac{1}{2}\sqrt{n}\tilde{I}_{\theta,\eta}(\theta_n - \theta)$ by (7.4), the differentiability of the model and the continuity of the inner product.

We prove the first assertion in (7.7) also by computing moments, but this time under the measures obtained by letting $X_1,\ldots,X_n$ be an i.i.d. sample from the probability measure with density $q_n = c_n p^{1/2}_{\theta_n,\eta}p^{1/2}_{\theta,\eta}$, where c_n is the norming constant. By the differentiability of the model we have

$$\begin{aligned}
c_n^{-1} &= \int p^{1/2}_{\theta_n,\eta}p^{1/2}_{\theta,\eta}\,d\mu = 1 - \tfrac{1}{2}\int (p^{1/2}_{\theta_n,\eta} - p^{1/2}_{\theta,\eta})^2\,d\mu \\
&= 1 - \tfrac{1}{2}(\theta_n - \theta)^T I_{\theta,\eta}(\theta_n - \theta) + o(n^{-1}).
\end{aligned}$$

From an expansion of the log likelihood ratio of the n-fold product measure Q_n^n corresponding to q_n and the n-fold product $P^n_{\theta,\eta}$, we see that these product measures are contiguous. Thus it suffices to prove convergence in probability to zero under Q_n^n. We have

$$\begin{aligned}
&Q_n^n\left|\sqrt{n}\mathbb{P}_n\left(\tilde{\ell}_{\theta_n,\eta}\left(1 - \frac{p^{1/2}_{\theta_n,\eta}}{p^{1/2}_{\theta,\eta}}\right)\right) + \sqrt{n}\tfrac{1}{2}\mathbb{P}_n\tilde{\ell}_{\theta_n,\eta}\dot{\ell}^T_{\theta,\eta}(\theta_n - \theta)\right| \\
&\quad\le c_n\int |\tilde{\ell}_{\theta_n,\eta}p^{1/2}_{\theta_n,\eta}|\,\sqrt{n}\left|(p^{1/2}_{\theta_n,\eta} - p^{1/2}_{\theta,\eta}) - \tfrac{1}{2}(\theta_n - \theta)^T\dot{\ell}_{\theta,\eta}p^{1/2}_{\theta,\eta}\right|\,d\mu \to 0,
\end{aligned}$$

by the differentiability of the model, (7.4) and the fact that $c_n \to 1$. Finally, it suffices to show that the sequence $\mathbb{P}_n \tilde{\ell}_{\theta_n,\eta} \dot{\ell}_{\theta,\eta}^T$ converges in probability to $\tilde{I}_{\theta,\eta}$ under Q_n^n. For this we first note that

$$\mathrm{E}_{Q_n} \mathbb{P}_n \tilde{\ell}_{\theta_n,\eta} \dot{\ell}_{\theta,\eta}^T = c_n \int \tilde{\ell}_{\theta_n,\eta} p_{\theta_n,\eta}^{1/2} \dot{\ell}_{\theta,\eta}^T p_{\theta,\eta}^{1/2} \, d\mu \to \tilde{I}_{\theta,\eta},$$

$$\mathrm{var}_{Q_n} \mathbb{P}_n \tilde{\ell}_{\theta_n,\eta} 1_{\|\tilde{\ell}_{\theta,\eta}\| \le M} \dot{\ell}_{\theta,\eta}^T 1_{\|\dot{\ell}_{\theta,\eta}\| \le M} \le c_n M^2 \frac{1}{n} \int \|\tilde{\ell}_{\theta_n,\eta}\| p_{\theta_n,\eta}^{1/2} \, \|\dot{\ell}_{\theta,\eta}\| p_{\theta,\eta}^{1/2} \, d\mu \to 0,$$

for every fixed M. We also have that

$$\mathrm{E}_{Q_n} \mathbb{P}_n \tilde{\ell}_{\theta_n,\eta} 1_{\|\tilde{\ell}_{\theta,\eta}\| > M} \dot{\ell}_{\theta,\eta}^T 1_{\|\dot{\ell}_{\theta,\eta}\| > M} \to 0,$$

as $n \to \infty$, followed by $M \to \infty$. The proof is complete upon combining the last two displays. ∎

The theorems reduce the problem of efficient estimation of θ to estimation of the efficient score function. At first sight we have made the problem harder. The estimator of the efficient score function must satisfy a "no-bias" and a consistency condition. The consistency is usually easy to arrange, but the no-bias condition, such as (7.3), is connected to the structure and the size of the model, as the bias must converge to zero at a rate faster than $1/\sqrt{n}$. It may happen that the bias is identically zero and then we only need to produce a consistent estimator of the efficient score function. In general, we can at best hope that the bias is a second order term, just as in our discussion of general estimating equations in Lecture 6.

The good news is that if an efficient estimator sequence exists, then it can always be constructed by the one-step method. In that sense the no-bias condition is necessary.

7.8 Theorem. *Suppose that the model $\{P_{\theta,\eta}: \theta \in \Theta\}$ is differentiable in quadratic mean with respect to θ at (θ, η), let the efficient information matrix $\tilde{I}_{\theta,\eta}$ be nonsingular, and assume that (7.4) holds. Then the existence of an asymptotically efficient sequence of estimators of $\psi(P_{\theta,\eta}) = \theta$ implies the existence of a sequence of estimators $\hat{\ell}_{n,\theta}$ satisfying (7.3).*

Proof. An efficient estimator sequence T_n must be asymptotically linear in the efficient influence function. By Lemma 7.6 and the continuity of $\theta \mapsto \tilde{I}_{\theta,\eta}$ this implies that, for every $\theta_n = \theta + O(n^{-1/2})$,

$$\sqrt{n}(T_n - \theta_n) = \mathbb{G}_n \tilde{\psi}_{\theta_n,\eta} + o_P(1),$$

where $\tilde{\psi}_{\theta,\eta} = I_{\theta,\eta}^{-1} \tilde{\ell}_{\theta,\eta}$. For simplicity we assume that this expansion is actually true in the stronger sense that, for every $\theta_n = \theta + O(n^{-1/2})$,

$$\mathrm{E}_{\theta_n,\eta} \big[\sqrt{n}(T_n - \theta_n) - \mathbb{G}_n \tilde{\psi}_{\theta_n,\eta}\big]^2 \to 0.$$

The general case can be handled by a truncation argument, which turns convergence in probability in convergence in second mean. (See [14].) Furthermore, to simplify notation we assume that T_n is permutation symmetric in its arguments.

In view of Hájek's projection lemma (which gives the orthogonal projection onto the space of all sums $\sum_{i=1}^n f(X_i)$), our assumption implies that

$$\mathrm{E}_{\theta_n,\eta}\Big[\sum_{i=1}^n \mathrm{E}_{\theta_n,\eta}\big(\sqrt{n}(T_n-\theta_n)\big|\, X_i\big) - \mathrm{E}_{\theta_n,\eta}\sqrt{n}(T_n-\theta_n) - \mathbb{G}_n\tilde\psi_{\theta_n,\eta}\Big]^2 \to 0,$$

which can be rewritten as

$$\mathrm{E}_{\theta_n,\eta}\Big[\mathrm{E}_{\theta_n,\eta}\big(n(T_n-\theta_n)\big|\, X_1\big) - \mathrm{E}_{\theta_n,\eta}n(T_n-\theta_n) - \tilde\psi_{\theta_n,\eta}(X_1)\Big]^2 \to 0.$$

Rather than estimate $\tilde\psi_{\theta,\eta}$ we can therefore "estimate" the function $x \mapsto \mathrm{E}_{\theta,\eta}\big(n(T_n-\theta)\big|\, X_1 = x\big)$ and its expectation. Given k_n independent copies $Y_{j1},\ldots,Y_{jn}$ of the sample $X_1,\ldots,X_n$, define

$$J_n(x) = \frac{1}{k_n}\sum_{j=1}^{k_n} n\big(T_n(x,Y_{j2},\ldots,Y_{jn}) - T_n(Y_{j1},\ldots,Y_{jn})\big).$$

Then $\mathrm{E}_{\theta_n,\eta}\big(J_n(X_1)\big|\, X_1\big)$ is identical to $\mathrm{E}_{\theta_n,\eta}\big(n(T_n-\theta_n)\big|\, X_1\big) - \mathrm{E}_{\theta_n,\eta}n(T_n-\theta_n)$ and hence

$$\begin{aligned}&\mathrm{E}_{\theta_n,\eta}\Big[J_n(X_1) - \mathrm{E}_{\theta_n,\eta}\big(n(T_n-\theta_n)\big|\, X_1\big) + \mathrm{E}_{\theta_n,\eta}n(T_n-\theta_n)\Big]^2\\ &\quad= \frac{1}{k_n}\mathrm{E}_{\theta_n,\eta}n^2\big(T_n(X_1,Y_{j2},\ldots,Y_{jn}) - T_n(Y_{j1},\ldots,Y_{jn})\big)^2 \lesssim \frac{n}{k_n},\end{aligned}$$

because $n\mathrm{E}_{\theta_n,\eta}(T_n-\theta_n)^2$ is bounded. This converges to zero for e.g. $k_n = n^2$. Then the estimator J_n is based on $m_n = k_n n = n^3$ observations. We define an estimator based on m observations, for every $m \in \mathbb{N}$, by $\tilde J_m = J_{\lfloor m^{1/3}\rfloor}$. A sequence $\tilde\theta_m = \theta + O(m^{-1/2})$ yields a sequence $\theta_n = \theta + O(n^{-3/2})$ on our original scale and hence is covered by the previous calculations. We conclude that, for every $\theta_n = \theta + O(n^{-1/2})$,

$$\mathrm{E}_{\theta_n,\eta}\int (\tilde J_n - \tilde\psi_{\theta_n,\eta})^2(x)\, p_{\theta_n,\eta}(x)\, d\mu(x) \to 0.$$

Thus the sequence $\tilde J_n$ is consistent as desired. To find a sequence of estimators that is both consistent and has small bias, we replace $\tilde J_n$ by, with $m = \lfloor n/2\rfloor$,

$$\begin{aligned}\hat\ell_{n,\theta}(x) &= \tilde J_{m_n}(x)\\ &\quad + T_{n-m_n}(X_{m_n+1},\ldots,X_n) - \theta - \frac{1}{n-m_n}\sum_{i=m_n+1}^n \tilde J_{m_n}(X_i).\end{aligned}$$

By assumption this is equivalent to

$$\begin{aligned}&\tilde J_{m_n}(x) + \frac{1}{n-m_n}\sum_{i=m_n+1}^n (\tilde\psi_{\theta_n,\eta} - \tilde J_{m_n})(X_i) + o_P(n^{-1/2})\\ &\tilde J_{m_n}(x) - \int \tilde J_{m_n} p_{\theta_n,\eta}\, d\mu + o_P(n^{-1/2}),\end{aligned}$$

by comparing conditional means and variances given $X_1,\ldots,X_{m_n}$, and where the $o_P(n^{-1/2})$-term does not depend on x. Thus the estimator $\hat\ell_{n,\theta}$ is both consistent and has small bias. ∎

7.3 Symmetric location

Suppose that we observe a random sample from a density $\eta(x-\theta)$ that is symmetric about θ. In Example 2.19 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 7.5 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 the techniques of Lectures 4 and 5.

Second, it suffices to construct estimators $\hat{\ell}_{n,\theta}$ that satisfy (7.3). 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

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

Since this function is skew-symmetric about the point θ, the bias condition in (7.3) is satisfied, with a bias of zero. Since 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 (7.3) reduces to consistency of $\hat{k}_n$ for the function g'/g in that

$$\int \left(\hat{k}_n - \frac{g'}{g}\right)^2(s)\, g(s)\, ds \xrightarrow{\mathrm{P}} 0. \tag{7.9}$$

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

$$\begin{aligned} \hat{g}_n(s) &= \frac{1}{\sigma_n}\sum_{i=1}^{n}\omega\Big(\frac{s-T_i}{\sigma_n}\Big), \\ \hat{k}_n(s) &= \frac{\hat{g}_n'}{\hat{g}_n}(s)1_{\hat{B}_n}(s), \\ \hat{B}_n &= \big\{s\colon |\hat{g}_n'(s)| \le \alpha_n, \hat{g}_n(s) \ge \beta_n, s \ge \gamma_n\big\}. \end{aligned} \tag{7.10}$$

Then (7.3) 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 consists of the usual manipulations of kernel estimators. (See [42], page 398, for a precise statement, or one of the many papers on this model.)

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 will be theoretically efficient under some regularity conditions.

7.11 Open Problem. It may be verified that the preceding construction generalize to higher dimensions. The problem of estimating θ from a sample of observations from a density $\eta(x-\theta)$ on $\mathbb{R}^d$ such that η has finite Fisher information and $\eta(x) = \eta(-x)$ is adaptive for any $d \geq 1$. Theoretically, one can estimate θ as well knowing η as not knowing η. However, in practice this appears to be nonsense. If $d = 10$, for instance, it cannot make sense to try and estimate η nonparametrically from $n = 1000$ observations and the preceding construction will presumably yield bad estimators. The problem is to develop a theory for this phenomenon, maybe using minimax bounds. Note that the problem of estimating θ for $d = 10$ is by itself not difficult. For instance, we could use an M-estimator and this will be asymptotically normal in the usual way and the asymptotics will be reliable for $n \geq 30$. See [32] and [31] for further questions regarding the asymptotic information bounds.

7.4 Errors-in-Variables

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

$$X = Z + e$$
$$Y = \alpha + \beta Z + f,$$

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 (where 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

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

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 (6.27)). One possibility for $\hat{\eta}$ is the maximum likelihood estimator, which was shown to be consistent in Lecture 5 in the case that Σ is known. This proof can be extended to the case that Σ is unknown.

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.

The submodel $t \mapsto P_{\theta,t\eta_1+(1-t)\eta}$ is well-defined for every $0 \le t \le 1$ and every $\eta_1 \in H$. Its score function is the function

$$p_{\theta,\eta_1}/p_{\theta,\eta} - 1$$

As is clear from the factorization theorem or direct calculation, it is a function 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, for every η_1,

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

Consequently, $b = 0$ almost surely by the completeness of $\psi_\theta(X,Y)$. We conclude that the closure of the linear span of the nuisance tangent space contains all measurable, square-integrable functions of $\psi_\theta(x,y)$.

The efficient score function can be written in the form

$$\tilde\ell_{\theta,\eta}(x,y) = Q_\theta(x,y) + P_\theta(x,y)\mathrm{E}\big(Z \,|\, \psi_\theta(X,Y)\big)$$

for polynomials Q_θ and P_θ of orders 2 and 1, respectively. The main work is now to show that the class of all functions of this type, when η ranges over a large class of distributions, is Donsker. Because we already know that $\hat\eta$ is consistent for the weak topology, it is enough to show this for η ranging over a weak neighbourhood of the true mixing distribution. The following lemma is the main part of the verification.

7.12 Lemma. *For every $0 < \alpha \le 1$ and every probability distribution η_0 on $\mathbb{R}$ and compact $K \subset (0,\infty)$, there exists an open neighbourhood 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

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

for every $V \ge 1/\alpha$, 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.

Proof. We only give a sketch of the main steps. See Lemma 7.3 in [25] for the details. First consider the functions

$$t \mapsto g_{c,\eta}(t) = \frac{\int z e^{zt^2} e^{-cz^2}\, d\eta(z)}{\int e^{zt} e^{-cz^2}\, d\eta(z)}.$$

By some clever applications of Jensen's and other inequalities it can be proved that there exists a weak neighbourhood U of η_0 such that, for $\eta \in U$ and $c \in K$,

$$\big|g_{c,\eta}(t)\big| \le C(1+|t|), \qquad \big|g'_{c,\eta}(t)\big| \le C(1+|t|)^2.$$

The classical bounds of Kolmogorov give estimates on the covering numbers $N(\varepsilon, \mathcal{F}_j, \|\cdot\|)$ of the classes $\mathcal{F}_j$ of all functions $f: [j, j+1] \mapsto \mathbb{R}$ such that $\|f\|_\infty \vee \|f'\|_\infty \le M_j$ for some constant M_j. In the present situation we apply these boundeds with $\mathcal{F}_j$ the restrictions of the functions $g_{c,\eta}$ to the intervals $[j, j+1]$ and with $M_j = (1+|j|)^2$.

Given an ε_j-net $f_{j,1}, \ldots, f_{j,N_j}$ over $\mathcal{F}_j$ we can construct brackets for the functions $f: \mathbb{R} \mapsto \mathbb{R}$ by first forming brackets $[f_{j,i} - \varepsilon_j, f_{j,i} + \varepsilon_j]$ on each interval $[j, j+1)$ and next glueing these brackets together in every possible combination. Naturally, we choose ε_j big enough so that for all but finitely many intervals we need to use only one bracket, because otherwise the number of brackets would be infinite. We can optimize the numbers ε_j and M_j such that resulting brackets on $\mathbb{R}$ are ε-brackets relative to the $L_2(Q)$-norm and such that they are almost a minimal set of ε-brackets, for Q the measure constructed below.

For fixed (a, b) the functions $f_{a,b,c,\eta}$ are essentially the functions $g_{c,\eta}$, because

$$f_{a,b,c,\eta}(x, y) = (a_0 + a_1 x + a_2 y) g_{c,\eta}(b_0 + b_1 x + b_2 y).$$

A bracket $[l, u]$ for the functions $g_{c,\eta}$ yields a bracket for the functions $f_{a,b,c,\eta}$ of the form

$$\Big[(a_0 + a_1 x + a_2 y)^+ l(b_0 + b_1 x + b_2 y) - (a_0 + a_1 x + a_2 y)^- u(b_0 + b_1 x + b_2 y),$$
$$(a_0 + a_1 x + a_2 y)^+ u(b_0 + b_1 x + b_2 y) - (a_0 + a_1 x + a_2 y)^+ l(b_0 + b_1 x + b_2 y)\Big].$$

Its size in $L_2(P)$ is equal to the size of $[l, u]$ in $L_2(Q)$ for the measure Q defined by

$$Q(B) = \int 1_B(b_0 + b_1 x + b_2 y)(a_0 + a_1 x + a_2 y)^2 \, dP(x, y).$$

Thus we can construct the desired brackets for the functions $f_{a,b,c,\eta}$ as c and η vary, for any fixed value (a, b).

For fixed (x, y) the dependence $(a, b) \mapsto f_{a,b,c,\eta}(x, y)$ is Lipschitz of order $\beta = \alpha/2$ with Lipschitz constant $h(x, y) = (1 + |x| + |y|)^{2+2\beta}$. Now construct brackets over the class of all $f_{a,b,c,\eta}$ by first choosing an $\varepsilon^{1/\beta}/\|h\|_{P,2}$-net over the set of all (a, b), next for every (a_i, b_i) in this net choose a minimal number of brackets $l, u]$ over the class of all $f_{a_i,b_i,c,\eta}$ and finally form the brackets $\big[l - \varepsilon h/\|h\|_{P,2}, u + \varepsilon h/\|h\|_{P,2}\big]$. Because we need only of the order $(1/\varepsilon)^{6/\beta}$ points (a_i, b_i) this last step hardly increases the entropy. ∎

Notes

Theorem 7.6 is due to [14]. The semiparametric one-step method has a long history, starting with special constructions in the symmetric location model.

Lecture 8
Rates of Convergence

In this lecture we apply maximal inequalities for empirical processes to obtain rates of convergence of minimum contrast estimators, in particular in semiparametric models. These rates are of interest by themselves, but will also be needed to prove the asymptotic normality of semiparametric likelihood estimators in certain models.

8.1 A General Result

The set-up is the same as the one in Lecture 5 on consistency. Let Θ be a metric space and for each $\theta \in \Theta$, let $m_\theta \colon \mathcal{X} \mapsto \mathbb{R}$ be a measurable function. Suppose that we are interested in the maximizer $\hat{\theta}$ of $\theta \mapsto \mathbb{P}_n m_\theta$. We may expect that this converges in probability to the maximizer θ_0 of the limiting criterion function $\theta \mapsto P m_\theta$. It is useful to picture the random criterion function $\mathbb{P}_n m_\theta$ as the sum of its limit and the scaled empirical process

$$\mathbb{P}_n m_\theta = P m_\theta + \frac{1}{\sqrt{n}} \mathbb{G}_n m_\theta .$$

Because $P m_\theta$ is maximal at θ_0 we could picture the function $\theta \mapsto P m_\theta$ as an inverse parabola with its top at θ_0. Without the second, random term on the right, the estimator $\hat{\theta}$ would always choose the top of the parabola, but the fluctuations may pull the maximum of $\mathbb{P}_n m_\theta$ away from θ_0. It is the size of the fluctuations that determines how far. If $P m_\theta \approx -d(\theta, \theta_0)^2$ and $\sup\{\mathbb{G}_n m_\theta \colon d(\theta, \theta_0) \leq \delta\} \approx \phi_n(\delta)$, then $d(\hat{\theta}, \theta_0)$ will probably be approximately equal to the value δ that balances the positive and negative parts of

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

In other words, we expect that $d(\hat{\theta}, \theta_0) \approx \delta_n$ for $\phi_n(\delta_n) \approx \sqrt{n}\delta_n^2$. The following theorem makes this precise.

As for the consistency results, we do not need $\hat{\theta}$ to maximize $\mathbb{P}_n m_\theta$. We only need that $\mathbb{P}_n m_{\hat{\theta}} \geq \mathbb{P}_n m_{\theta_0}$.

8.1 Theorem. *Suppose that, for all sufficiently small $d(\theta,\theta_0)$ all sufficiently small $\delta > 0$ and a function ϕ_n such that $\phi_n(c\delta) \le c^\alpha \phi_n(\delta)$ for all $c > 1$, and some $\alpha < 2$,*

$$P(m_\theta - m_{\theta_0}) \le -d^2(\theta,\theta_0),$$
$$\mathrm{E}^* \sup_{d(\theta,\theta_0)<\delta} \bigl|\mathbb{G}_n(m_\theta - m_{\theta_0})\bigr| \le \phi_n(\delta).$$

Then $\mathbb{P}_n m_{\hat\theta} \ge \mathbb{P}_n m_{\theta_0}$ and $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$ together imply that $d(\hat\theta_n,\theta_0) = O_P(\delta_n)$ for every δ_n satisfying $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$.

Proof. For each n, the parameter space (minus the point θ_0) can be partitioned into the "shells" $S_{j,n} = \{\theta\colon 2^{j-1}\delta_n < d(\theta,\theta_0) \le 2^j\delta_n\}$ with j ranging over the integers. If $d(\hat\theta_n,\theta_0)$ is larger than $2^M\delta_n$ for a given integer M, then $\hat\theta_n$ is in one of the shells $S_{j,n}$ with $j \ge M$. In that case the supremum of the map $\theta \mapsto \mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0}$ over this shell is nonnegative by the property of $\hat\theta_n$. Conclude that, for every $\eta > 0$,

$$\mathrm{P}^*\Bigl(d(\hat\theta_n,\theta_0) > 2^M\delta_n\Bigr) \le \sum_{\substack{j\ge M\\ 2^j\delta_n\le\eta}} \mathrm{P}^*\Bigl(\sup_{\theta\in S_{j,n}}\bigl(\mathbb{P}_n m_\theta - \mathbb{P}_n m_{\theta_0}\bigr) \ge 0\Bigr) + \mathrm{P}^*\bigl(2d(\hat\theta_n,\theta_0) \ge \eta\bigr).$$

Because the sequence $\hat\theta_n$ is consistent for θ_0, the second probability on the right converges to 0 as $n \to \infty$ for every $\eta > 0$. Choose $\eta > 0$ small enough that the first condition of the theorem holds for every $d(\theta,\theta_0) \le \eta$ and the second for every $\delta \le \eta$. Then for every j involved in the sum, we have, for every $\theta \in S_{j,n}$,

$$Pm_\theta - Pm_{\theta_0} \le -d^2(\theta,\theta_0) \lesssim -2^{2j-2}\delta_n^2.$$

Therefore, the series may be bounded by

$$\sum_{\substack{j\ge M\\ 2^j\delta_n\le\eta}} \mathrm{P}^*\Bigl(\bigl\|\mathbb{G}_n(m_\theta - m_{\theta_0})\bigr\|_{S_{j,n}} \ge \sqrt{n}2^{2j-2}\delta_n^2\Bigr) \lesssim \sum_{j\ge M} \frac{\phi_n(2^j\delta_n)}{\sqrt{n}\delta_n^2 2^{2j}} \lesssim \sum_{j\ge M} 2^{j\alpha-2j},$$

by Markov's inequality, the definition of δ_n, and the fact that $\phi_n(c\delta) \le c^\alpha\phi_n(\delta)$ for every $c > 1$. This expression converges to zero for every $M = M_n \to \infty$. ∎

The first condition of the theorem can be expected to hold if θ_0 is a point of maximum of $\theta \mapsto Pm_\theta$ and this function is twice differentiable. More generally, we can see it as simply defining the type of metric that we can work with. For instance, if m_θ is a log likelihood under parameter θ and $P = P_{\theta_0}$, then $P_{\theta_0}(m_\theta - m_{\theta_0})$ is the Kullback-Leibler divergence and we can either use this directly (inspection of the proof of the theorem, shows that it is not really necessary that d is a metric), or a metric whose square dominates this, such as the Hellinger distance. It is well known that for any pair of probability densities p and q,

$$P\log(q/p) \le -h^2(P,Q) = -\int\bigl(\sqrt{p} - \sqrt{q}\bigr)^2\,d\mu. \tag{8.2}$$

Thus the Hellinger distance is a natural distance when considering rates of convergence of maximum likelihood estimators.

The latter observation also points out a severe limitation of the theorem: the choice of metrics with which it works is limited. For instance, in a semiparametric model with parameter (θ, η) we might wish to prove that the maximum likelihood estimator, or some other contrast estimator, possesses a $\sqrt{n}$-rate of convergence. This will very rarely follow with the help of the preceding theorem, because the theorem will give a rate for the joint estimator $(\hat{\theta}, \hat{\eta})$, rather than for $\hat{\theta}$ only. The joint rate will typically be determined by the rate of $\hat{\eta}$ and this will typically be slower than $\sqrt{n}$.

Even a natural rate on the nuisance parameter $\hat{\eta}$ may not be derivable from the theorem, if "natural" refers to a particular, natural distance, which does not combine well with the distance imposed by the theorem. As a consequence, unfortunately, the applicability of the theorem to semiparametric models is limited.

The second condition of the theorem requires a maximal inequality for the modulus of the empirical process. Here the inequalities of Lecture 6 may work if the size of the functions $m_\theta - m_{\theta_0}$ is comparable to the size of the envelope function of the class of all such functions with $d(\theta, \theta_0) < \delta$. This is not always the case. The following maximal inequalities directly take the size of the functions $m_\theta - m_{\theta_0}$ into account.

8.3 Lemma. *Let $\mathcal{F}$ be a class of measurable functions with $\|f\|_\infty \leq M$, and $Pf^2 < \delta^2$ for every $f \in \mathcal{F}$. Then*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)\Big(1 + \frac{M J_{[]}\big(\delta, \mathcal{F}, L_2(P)\big)}{\delta^2 \sqrt{n}}\Big).$$

The preceding lemma is sufficient for many examples. However, sometimes the assumption that the class is uniformly bounded is restrictive. This can be remedied by computing the size of the brackets relative to a larger norm. Specifically, consider

$$\|f\|_{P,B} = \sqrt{2P(e^{|f|} - 1 - |f|)}.$$

The subscript "B" is for Bernstein, as this "norm" is essential in an exponential inequality for sums due to Bernstein, which plays a major role in the proofs of maximal inequalities. Actually, the quantity $\|\cdot\|_{P,B}$ is not a norm, as it does not satisfy the triangle inequality. However, we can use $\|\cdot\|_{P,B}$ as a measure of the size of a function and hence as a measure of the size of a bracket $[l, u]$ by applying it to the function $u - l$. We can define an entropy integral relative to it accordingly.

8.4 Lemma. *Let $\mathcal{F}$ be a class of measurable functions with $Pf^2 < \delta^2$ for every $f \in \mathcal{F}$. Then*

$$\mathrm{E}_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr)\Bigl(1 + \frac{J_{[]}\bigl(\delta, \mathcal{F}, \|\cdot\|_{P,B}\bigr)}{\delta^2\sqrt{n}}\Bigr).$$

8.2 Nuisance Parameters

In this section we consider the same problem of finding an upper bound on the rate of convergence of a minimum contrast estimator $\hat{\theta}$, but now in the presence of an estimated nuisance parameter. Using the "wrong", estimated contrast function should bring the rate of convergence down, but only if the estimation of the nuisance parameter is the harder part of the problem. The following theorem implements this idea.

The theorem is of interest not only because it takes care of problems with nuisance parameters of the type as considered before, but also of certain penalized minimum contrast estimators, in which the smoothing parameter of the penalty can be thought of as an estimated nuisance parameter.

Consider "estimators" $\hat{\theta}_n$ contained in a metric space Θ_n satisfying, for a given "estimators" $\hat{\eta}_n$ contained in a metric space H_n,

$$\mathbb{P}_n m_{\hat{\theta}_n, \hat{\eta}_n} \geq \mathbb{P}_n m_{\theta_0, \hat{\eta}_n}$$

for given measurable functions $x \mapsto m_{\theta,\eta}(x)$. This is valid, for example, for $\hat{\theta}_n$ equal to the maximizer of the function $\theta \mapsto \mathbb{P}_n m_{\theta,\hat{\eta}_n}$ over Θ_n, if this set contains θ_0.

Assume that the following conditions are satisfied for every $\theta \in \Theta_n$, every $\eta \in H_n$ and every $\delta > 0$.

$$P\bigl(m_{\theta,\eta} - m_{\theta_0,\eta}\bigr) \lesssim -d_\eta^2(\theta, \theta_0) + d^2(\eta, \eta_0), \tag{8.5}$$

$$\mathrm{E}^* \sup_{\substack{d_\eta(\theta,\theta_0)<\delta, d(\eta,\eta_0)<\delta \\ \theta\in\Theta_n, \eta\in H_n}} \bigl|\mathbb{G}_n(m_{\theta,\eta} - m_{\theta_0,\eta})\bigr| \lesssim \phi_n(\delta). \tag{8.6}$$

Here $d_\eta^2(\theta, \theta_0)$ may be thought of as the square of a distance, but the following theorem is true for arbitrary functions $\theta \mapsto d_\eta^2(\theta, \theta_0)$. Usually d_η does not depend on η, but in this form the following theorem is flexible enough to apply to penalized minimum contrast estimators, where the smoothing parameter can be included in η.

8.7 Theorem. *Suppose that (8.6) is valid, for all sufficiently small $\delta > 0$ and a function ϕ_n such that $\phi_n(c\delta) \le c^\alpha \phi_n(\delta)$ for all $c > 1$, and some $\alpha < 2$, and for sets $\Theta_n \times H_n$ that contain $(\hat\theta, \hat\eta)$ with probability tending to 1. Then $d_{\hat\eta}(\hat\theta, \theta_0) = O_P^*\big(\delta_n + d(\hat\eta, \eta_0)\big)$ for any sequence of positive numbers δ_n such that $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ for every n.*

Proof. For each $n \in \mathbb{N}$, $j \in \mathbb{Z}$ and $M > 0$ define a set

$$S_{n,j,M} = \Big\{(\theta,\eta) \in \Theta_n \times H_n \colon 2^{j-1}\delta_n < d_\eta(\theta,\theta_0) \le 2^j\delta_n, d(\eta,\eta_0) \le 2^{-M} d_\eta(\theta,\theta_0)\Big\}.$$

Then the intersection of the events $\hat\theta \in \Theta_n$, $\hat\eta \in H_n$ and $d_{\hat\eta}(\hat\theta,\theta_0) \ge 2^M\big(\delta_n + d(\hat\theta,\theta_0)\big)$ is contained in the union of the events $\big\{(\hat\theta,\hat\eta) \in S_{n,j,M}\big\}$ over $j \ge M$. By the definition of $\hat\theta$, the variable $\sup_{(\theta,\eta)\in S_{n,j,M}} \mathbb{P}_n(m_{\theta,\eta} - m_{\theta_0,\eta})$ is nonnegative on the event $\big\{(\hat\theta,\hat\eta) \in S_{n,j,M}\big\}$. Conclude that, for every $\delta > 0$,

$$\begin{aligned}\mathrm{P}^*\Big(d_{\hat\eta}(\hat\theta,\theta_0) \ge 2^M\big(\delta_n + d(\hat\eta,\eta_0)\big), \hat\theta \in \Theta_n, \hat\eta \in H_n\Big)& \\ \le \sum_{j \ge M} \mathrm{P}^*\Big(\sup_{(\theta,\eta)\in S_{j,n,M}} \mathbb{P}_n\big(m_{\theta,\eta} - m_{\theta_0,\eta}\big) \ge 0\Big).&\end{aligned}$$

For every j involved in the sum, we have, for every $(\theta,\eta) \in S_{j,n,M}$ and every sufficiently large M,

$$\begin{aligned}P\big(m_{\theta,\eta} - m_{\theta_0,\eta}\big) &\lesssim -d_\eta^2(\theta,\theta_0) + d^2(\eta,\eta_0)\\ &\lesssim -(1 - 2^{-2M})\, d_\eta^2(\theta,\theta_0) \lesssim -2^{2j-2}\delta_n^2.\end{aligned}$$

We now finish the proof as the proof of Theorem 8.1. ∎

For $d_\eta = d$ not depending on η condition (8.5) is implied by the conditions

$$\begin{aligned}P\big(m_{\theta_0,\eta} - m_{\theta_0,\eta_0}\big) &\gtrsim -d^2(\eta,\eta_0),\\ P\big(m_{\theta,\eta} - m_{\theta_0,\eta_0}\big) &\lesssim -d^2(\theta,\theta_0).\end{aligned}$$

These two conditions are the natural requirement that the criterion function $(\theta,\eta) \mapsto Pm_{\theta,\eta}$ behaves quadratically (relative to a distance) around the point of maximum (θ_0,η_0).

8.3 Cox Regression with Current Status Data

Let us apply the Theorem 8.1 to one example, which illustrates the potential and difficulties, and for which we shall need the rate of convergence in the next lecture as input to proving asymptotic normality of the maximum likelihood estimator. It is again the Cox model, but this time with a type of censoring that changes everything.

Suppose that we observe a random sample from the distribution of $X = (C, \Delta, Z)$, where $\Delta = 1\{T \leq C\}$, 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

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

We define this as the likelihood for one observation x and are interested in the estimator $(\hat\theta_n, \hat\Lambda_n)$ obtained by maximizing the full likelihood. Here we restrict the parameter θ to a compact $\Theta \subset \mathbb{R}^k$ 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" (the end point of the distribution of C).

We make the following assumptions. The observation time C possesses a Lebesgue density which 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 $\mathrm{E}\,\mathrm{cov}(Z|\,C) > 0$. The true parameter θ_0 is an inner point of the parameter set and the efficient information for θ is positive. (We make the latter condition concrete in the next lecture.)

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

Actually, we shall show that $\hat\theta_n$ also possesses a rate of convergence of at least $n^{-1/3}$. However, in the next lecture we shall see that the true rate is $n^{-1/2}$. It is a good illustration of what cannot be achieved with the preceding rate theorem.

Remembering Trick 1 of Lecture 5 we apply Theorem 8.1 not with m_θ equal to the log likelihood (as would be the straightforward thing to do), but with the functions

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

where the 0 denote the "true" parameter (θ_0, Λ_0). 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, which is of some help.

In Lemma 8.9 below we explicitly bound the bracketing numbers of the class of functions $m_{\theta,\Lambda}$, from which we infer that these are finite. Therefore, the class of functions $m_{\theta,\Lambda}$ forms a Glivenko-Cantelli class. The parameter set Θ is compact by assumption and the parameter set for Λ is compact for the weak topology, also partly because of our assumptions. If the parameter (θ_0, Λ_0) were identifiable, we could conclude by Theorem 5.8 that $(\hat\theta_n, \Lambda_n)$ is consistent. However, under our assumptions the parameter is not fully identifiable: the parameter Λ_0 is identifiable only on the interval (σ, τ). We can still conclude that $\hat\theta \xrightarrow{\mathrm{P}} \theta_0$ and that $\hat\Lambda(t) \xrightarrow{\mathrm{P}} \Lambda_0(t)$ for every $\sigma < t < \tau$. (The convergence of $\hat\Lambda$ at the points σ and τ does not appear to be guaranteed.)

By (8.2) the Kullback-Leibler divergence $P_0(m_{\theta,\Lambda} - m_0)$ is dominated by the square Hellinger distance between $(p_{\theta,\Lambda} + p_0)/2$ and p_0, and this in turn is equivalent to the square Hellinger distance between $p_{\theta,\Lambda}$ and p_0. By a lucky coincidence this

distance translates easily in natural distances on θ and Λ. By Lemma 8.10 below, we have

$$P_0(m_{\theta,\Lambda} - m_0) \lesssim -\|\theta - \theta_0\|^2 - \|\Lambda - \Lambda_0\|_2^2.$$

Thus we can take minus the right side as the square distance in Theorem 8.1. We only need to bound the modulus of the empirical process for this distance. By Lemma 8.9 below, the bracketing entropy of the class of functions $m_{\theta,\Lambda}$ is of the order $(1/\varepsilon)$. By Lemma 8.3 we can choose the function ϕ_n in Theorem 8.1 equal to

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

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

We finish with the technical work in the form of two lemmas.

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

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

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 $\Lambda_1 \le \Lambda \le \Lambda_2$ for Λ leads, by substitution, readily to a bracket for $m_{\theta,\Lambda}$. Furthermore, since this dependence is Lipschitz, there exists a constant D such that

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

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. It is well known that the set of all monotone functions $\Lambda\colon \mathbb{R} \mapsto [0, M]$ possesses a bracketing entropy of the order $1/\varepsilon$. Therefore, 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. The partial derivative $\partial/\partial\theta\, m_{\theta,\Lambda}(x)$ is uniformly bounded in (θ, Λ, x). Therefore, if $m_{\theta,\Lambda}$ is contained in a bracket $[l, u]$, then $m_{\theta',\Lambda}$ is contained in the bracket $[l - \varepsilon, u + \varepsilon]$ for every θ' with $\|\theta' - \theta\| \lesssim \varepsilon$. If the bracket $[l, u]$ is of size ε, then the bracket $[l - \varepsilon, u + \varepsilon]$ is of size 2ε. It follows that we can construct a set of brackets for the functions $m_{\theta,\Lambda}$ by first selecting an ε-net $\theta_1, \ldots, \theta_p$ over Θ, then apply the procedure of the first paragraph to find brackets for the functions $m_{\theta_i,\Lambda}$ for each i, and finally enlarging this bracket. The total number of brackets will be of the order $(1/\varepsilon)^k \exp c(1/\varepsilon)$. ∎

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

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

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

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

Since 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

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

Let $g(t)$ be the function $\exp(-e^{\theta^T z}\Lambda(c))$ 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

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

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.

The remainder of the proof is best understood in terms of semiparametric information. We adopt the notation of the information calculations given in the next lecture. Since 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 B_{θ_0,Λ_0} is the score operator for the model, which we derive in the next lecture. 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,

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

by the Cauchy-Schwarz inequality. Since the efficient information $\tilde{I}_0$ is positive-definite by assumption, 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 8.11 ahead. ∎

8.11 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*

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

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 ineqalities

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

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

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

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

Notes

Rates of convergence have been a hot topic in the 1990s. Here we have only said enough in order to be able to treat the Cox model with current status censoring in Lecture 9. The papers [5] and [6] are important contributions and contain good references. Another source of references is the book [41], which also gives an overview.

Lecture 9
Maximum and Profile Likelihood

In this lecture we study likelihood methods for semiparametric models. This concerns both ordinary likelihoods indexed by infinite-dimensional parameters and empirical likelihoods.

9.1 Examples

"Likelihood" is the key unifying element in classical statistics and hence it is worth while to seek a theory of likelihood for semiparametric models. This will be the subject of our last two lectures. Unfortunately, what we shall have to say is not completely satisfying. As known today likelihood theory for semiparametric models falls short of the beautiful and simple theory for parametric models.

A first problem is that it is not obvious what we should define to be the "likelihood" of a given semiparametric model, in general. It is obvious that the likelihood has something to do with a density of the observations, viewed as function of the parameter. Apart from the fact that we also need to choose particular versions of these densities, we encounter the further, major problem that many semiparametric models are not dominated, or are defined in terms of densities that maximize to infinity.

The good news is that given a concrete example it is usually not difficult to choose a "likelihood", albeit that often other, slightly different choices would be just as reasonable. Sometimes a likelihood can be taken equal to a density with respect to a dominating measure, for other models we use an "empirical likelihood", but 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, some authors use the phrase "nonparametric maximum likelihood estimators" (NPMLE). We prefer to speak simply of "maximum likelihood estimators", accepting the risk of being charged that nothing new is happening here. (In fact, it would be nice if nothing new needed to happen.) After all, in each of the models we are thinking of there is only one likelihood. We shall not give an abstract definition

of "likelihood", but shall describe "likelihoods that work" for a number of examples to set the stage. We denote the likelihood for the parameter P given one observation x by $\mathrm{lik}(P)(x)$ or $\mathrm{lik}(\theta,\eta)$ if $P = P_{\theta,\eta}$.

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.

9.1 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 would be restricted to distributions with a continuous Lebesgue density p, then we still could not 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. □

9.2 Open Problem. Suppose we use $p \mapsto \prod_{i=1}^{n} p(X_i)$ as a likelihood, restricted to a Hölder ball of densities $p\colon [0,1] \mapsto \mathbb{R}$, e.g. all densities which are twice continuously differentiable with second derivative bounded by 1 and which are themselves bounded by some fixed number. Is it true that $\int h(x)\, \hat{p}(x)\, dx$ is an asymptotically efficient estimator for $\psi(P) = \int h(x)\, p(x)\, dx$ for every reasonable function h?

9.3 Example (Cox model). We already discussed the problem of finding a likelihood for the Cox model in Lecture 5. There we settled on using the function

$$\mathrm{lik}(\theta, \Lambda)(y, \delta, x) = \Big(e^{\theta z} \Lambda\{y\} e^{-e^{\theta z}\Lambda(y)}\Big)^{\delta} \Big(e^{-e^{\theta z}\Lambda(y)}\Big)^{1-\delta}.$$

We also agreed to maximize this over all θ and over all nondecreasing, cadlag functions Λ with $\Lambda(0) = 0$. This is close, but not quite an empirical likelihood. Furthermore, we have enlarged the parameter set slightly, by not restricting the jumps of Λ to be at most 1. At the end of this lecture, when discussing profile likelihood, we reveal the reason for the latter. □

9.4 Example (Mixtures). Mixture models usually are based on well-behaved parametric families of densities, and then lead to well-behaved likelihoods equal to the ordinary density. Thus for a given kernel $p_\theta(\cdot|\, z)$ and $p_{\theta,\eta}$ the corresponding mixture density we simply set

$$\mathrm{lik}(\theta,\eta)(x) = p_{\theta,\eta}(x).$$

Surprisingly little is known about the behaviour of such likelihoods. For example, it is known for only a handful of examples that the θ-component of the maximum likelihood estimator $(\hat\theta,\hat\eta)$ is asymptotically efficient for estimating θ, as we would certainly expect. □

9.5 Open Problem. Just to show how little is known. Suppose that $X_1,\ldots,X_n$ are sampled from a normal location mixture $p_\eta(x) = \int \phi(x-z)\,dz$ and let $\hat\eta$ be the maximum likelihood estimator for η. Then $\int z\,d\hat\eta(z) = \overline{X}_n$ (as can be ascertained by manipulation of likelihood equations) and hence $\int z\,d\hat\eta(z)$ is asymptotically efficient for estimating the mean of η, if this exists. Is the analogous statement true for the higher moments $\int z^k\,d\hat\eta(z)$?

9.6 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

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

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 $\theta V + \eta(W)$, where 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 could restrict the maximization to a smaller set of η, which we could 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 $\hat{\lambda}_n$ will lead to smooth maximizers $\hat{\eta}$, while for small values the maximizer will be more like the unrestricted maximum likelihood estimator. Intermediate values are best, and are often chosen by a data-dependent scheme, such as cross validation. □

9.2 Asymptotic Normality

There are two ways of proving of asymptotic normality of the maximum likelihood estimator in parametric models: one based on maximization and one based on the likelihood equations. We like the first proof better, but it appears to be hard to generalize it to general semiparametric models, with its different types of likelihoods and possibly hard to estimate nuisance parameters. The proof based on the likelihood equations is easier to adapt to semiparametric models. If we are interested in the behaviour of the maximum likelihood estimator for θ in a semiparametric model with parameter (θ, η), then we have two possibilities. The first is to set up a system of likelihood equations for both parameters θ and η and infer the joint asymptotic normality of the maximum likelihood estimators. We shall discuss this method in the last lecture, Lecture 10.

The second possibility is to treat η as a nuisance parameter in the likelihood equation for θ. In fact, if $\hat{\theta}$ would satisfy the efficient score equation discussed in Lecture 7, then we have already proved its asymptotic normality and efficiency, under some conditions.

Sometimes the analysis is this easy, but not in general. Perhaps 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 becasue, by definition, the efficient score function is a projection (and L_2-approximation), and nothing guarantees that this projection is the derivative of the log likelihood along some submodel. If there exists a "least favourable" path $t \mapsto \eta_t(\hat{\theta}, \hat{\eta})$ such that $\eta_0(\hat{\theta}, \hat{\eta}) = \hat{\eta}$, and, for every x,

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

then the maximum likelihood estimator satisfies the efficient score equation; if not, then this is not clear. The existence of an exact least favourable 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.

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 7.2 goes through for every consistent estimator sequence $\hat{\theta}$ such that $\sqrt{n}\mathbb{P}_n \tilde{\ell}_{\hat{\theta},\hat{\eta}} = o_P(1)$. Second, this theorem is based on the more general Theorem 6.29, which yields asymptotic normality of estimators satisfying a more general estimating equation $\mathbb{P}_n \psi_{\theta,\hat{\eta}} \approx 0$, and actually uses the special property of

an efficient score function only to reduce the asymptotic variance to the inverse efficient influence function. As long as we can show that the maximum likelihood estimator $\hat\theta$ satisfies an equation $\mathbb{P}_n\psi_{\hat\theta,\hat\eta} \approx 0$ for functions $\psi_{\theta,\eta}$ that, if evaluated at the true parameter (θ,η), give the efficient score function, then we still can conclude that $\hat\theta$ is asymptotically efficient.

This motivates us to introduce approximately least favourable subprovided models.

9.7 Definition. An *approximately least favourable subprovided models* is a collection of maps $t \mapsto \eta_t(\theta,\eta)$ from a neighbourhood of $0 \in \mathbb{R}^k$ to the parameter set for η with $\eta_0(\theta,\eta) = \eta$ (for every (θ,η)) such that

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

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 favourable 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

$$t \mapsto \mathbb{P}_n \log \mathrm{lik}\big(\theta + t, \eta_t(\hat\theta,\hat\eta)\big)$$

is maximal at $t = 0$ and hence $(\hat\theta,\hat\eta)$ satisfies the stationary equation $\mathbb{P}_n\psi_{\hat\theta,\hat\eta} = 0$. Now Theorem 6.29 yields the asymptotic efficiency of $\hat\theta_n$. The main assumptions are that the entropies of the classes of realizations of the functions $\psi_{\hat\theta,\hat\eta}$ are stable and not too big, and the no-bias and consistency conditions (6.26) and (6.27).

Two obvious questions arise:

- Does an approximately least favourable submodel always exist?
- If it exists can it be chosen to satisfy the "regularity" conditions, such as (6.26)?

We discussed the nature of (6.26) in Lecture 6 and have nothing to add to it. We do not have a satisfying answer to the first question either. In many examples such submodels exist, but we have already mentioned some examples where the question of asymptotic normality of the maximum likelihood estimator is still open. To give some insight in the difficulties we discuss one example in some detail below. More in general, we note that we can often use our insight in the calculus of scores developed in the preceding lectures. Assume, for instance, that the information operator $B_0^*B_0$, evaluated at the true parameter (θ_0,Λ_0) is continuously invertible. Then the efficient score function is given by

$$\tilde\ell_0 = \dot\ell_0 - B_0(B_0^*B_0)^{-1}B_0^*\dot\ell_0.$$

A score function $B_{\theta,\eta}h$ would presumably arise from some path $t \mapsto \eta_t(\eta)(h)$ in the H-space. Then a potential least favourable path is given by

$$\eta_t(\theta,\eta) = \eta_t(\eta)(-h_0), \qquad h_0 = (B_0^*B_0)^{-1}B_0^*\dot\ell_0.$$

This, of course, is only possible if h_0 is a valid direction for perturbation of η in the H-space. It may be necessary to recenter h_0 first, and we may have to ascertain that h_0 is a nice function, e.g. bounded, or continuous, to make the path well-defined.

9.8 Example (Cox model). A convenient approximately least favourable submodel in the Cox model is defined by

$$d\Lambda_t(\theta, \Lambda) = \big(1 - th_0\big)\, d\Lambda,$$

where $h_0 = L_{1,\theta_0}/L_{0,\theta_0}$ is the least favourable direction in the Λ-space at the true parameter (θ_0, Λ_0). (See Example 3.13.) This is a valid cumulative hazard function, at least for $t \approx 0$, if h_0 is a bounded function, and this is true for instance if Z ranges over a bounded interval.

Substituting this submodel in Cox likelihood and differentiating with respect to t gives

$$\psi_{\theta,\Lambda}(x) = \frac{\partial}{\partial t}_{t=0} \operatorname{lik}\big(\theta + t, \Lambda_t(\theta, \lambda)\big)(x) = \dot\ell_{\theta,\Lambda} - B_{\theta,\Lambda} h_0(x).$$

This is not the efficient score function at every choice (θ, Λ), but it is the efficient score function for $(\theta, \Lambda) = (\theta_0, \Lambda_0)$, which is enough. The regularity conditions of Theorem 6.29 can be verified. Let us restrict ourselves to the most interesting one, the no-bias condition (6.28). We have

$$\begin{aligned}
P_{\theta_0,\Lambda_0}\psi_{\theta_0,\hat\Lambda} &= P_{\theta_0,\Lambda_0}(\dot\ell_{\theta_0,\hat\Lambda} - B_{\theta_0,\hat\Lambda} h_0) \\
&= P_{\theta_0,\Lambda_0}\big[(\dot\ell_{\theta_0,\hat\Lambda} - B_{\theta_0,\hat\Lambda} h_0) - (\dot\ell_{\theta_0,\Lambda_0} - B_{\theta_0,\Lambda_0} h_0)\big] \\
&= -P_{\theta_0,\Lambda_0}\big[z e^{\theta_0 z}(\hat\Lambda - \Lambda_0)(y) - e^{\theta_0 z}\int_{[0,y]} h_0\, d(\hat\Lambda - \Lambda_0)\big] \\
&= -\int (L_{1,\theta_0} - L_{0,\theta_0} h_0)\, d(\hat\Lambda - \Lambda_0).
\end{aligned}$$

The right side vanishes by the definition of the least favourable direction h_0. Therefore, the "no bias" condition is satisfied in the strongest possible sense. (We could have inferred this immediately from the linearity of the score functions in Λ (even though the likelihood is not linear in Λ)). Again, the Cox model is as nice as it can be; in other cases we do find a remainder term, and need to establish some rate of convergence. □

9.3 Cox Regression with Current Status Data

We take up the example for which we computed rates of convergence in Lecture 8. Thus we observe a random sample from the density

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

We define this density as the likelihood for one observation $x = (c, \delta, z)$. We make the same assumptions as in Lecture 8, but add the assumption that the function h_{θ_0,Λ_0} given by (9.9) ahead has a version which is differentiable with a bounded derivative on $[\sigma, \tau]$.

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

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

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

$$h_{\theta,\Lambda}(c) = \Lambda(c)\frac{\mathrm{E}_{\theta,\Lambda}\big(ZQ^2_{\theta,\Lambda}(X)|\, C = c\big)}{\mathrm{E}_{\theta,\Lambda}\big(Q^2_{\theta,\Lambda}(X)|\, C = c\big)}. \tag{9.9}$$

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 neighbourhood 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 vanishes 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 favourable submodel.

Our second guess on a least favourable submodel is to use $\Lambda_t(\theta, \Lambda) = \Lambda - th_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1} \circ \Lambda$. This alleviates the problem of different supports of Λ and its perturbation. Indeed, if the function $h_{\theta_0,\Lambda_0} \circ \Lambda_0^{-1}$ is Lipschitz, then for any $a \le b$ and C the Lipschitz constant,

$$\Lambda_t(\theta,\Lambda)(a) - \Lambda_t(\theta,\Lambda)(b) \le \big(\Lambda(a) - \Lambda(b)\big)(1 - tC).$$

Hence the function $\Lambda_t(\theta, \Lambda)$ is nondecreasing for sufficiently small $|t|$.

However, it is not clear that the range of $\Lambda_t(\theta,\Lambda)$ is inside $[0,M]$, whereas we have decided to maximize only over functions with range inside this interval. (It would have been better at this point to drop that restriction, to maximize over all nondecreasing functions, and next to prove that the maximizers remain uniformly bounded with high probability. However, we imposed the restriction to $[0,M]$ precisely, because we do not know if the last is true. Now we have to pay for it.) This motivates a third guess of a least favourable submodel. We take it to be, with ϕ a suitably chosen function,

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

If ϕ is Lipschitz, then $\Lambda_t(\theta,\Lambda)$ is nondecreasing, by the same argument as before. If $y - t\phi(y)h_{\theta_0,\Lambda_0}\circ\Lambda_0^{-1}(y)$ is contained in $[0,M]$ for all y in the range of Λ, then $\Lambda_t(\theta,\Lambda)$ takes its values in $[0,M]$. We achieve this if $0 \le \phi(y) \le c\big(y \wedge (M-y)\big)$ for every $0 \le y \le M$, Under our assumptions we can choose ϕ in such a way that this holds and, moreover, ϕ is the identity on the range $[\Lambda_0(s),\Lambda(\tau)]$ of Λ_0 (which is strictly contained in $[0,M]$).

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

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

When 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 favourable. To prove the asymptotic efficiency of $\hat\theta_n$ it suffices to verify the conditions of Theorem 6.29.

To verify the no-bias condition (6.28) we can use the decomposition (6.25) in combination with the inequalities

$$\begin{aligned}
|p_{\theta_0,\Lambda} - p_{\theta_0,\Lambda_0}|(x) &\lesssim |\Lambda-\Lambda_0|(c),\\
\big|\psi_{\theta_0,\Lambda} - \psi_{\theta_0,\Lambda_0}\big|(x) &\lesssim |\Lambda-\Lambda_0|(c),\\
\big|p_{\theta_0,\Lambda} - p_{\theta_0,\Lambda_0} - B_{\theta_0,\Lambda_0}(\Lambda-\Lambda_0)p_{\theta_0,\Lambda_0}\big|(x) &\lesssim |\Lambda-\Lambda_0|^2(c).
\end{aligned}$$

For every fixed x, the expressions on the left depend on Λ only through the scalar $\Lambda(y)$. For this reason these inequalities follow from ordinary Taylor expansions and uniform bounds on the first and second derivatives. By writing the bias as in (6.25), we now easily obtain that

$$|P_{\theta_0,\Lambda_0}\psi_{\theta_0,\hat\Lambda}| \lesssim \int_\sigma^\tau |\hat\Lambda-\Lambda_0|^2(c)\,dc.$$

The right side was shown to be of the order $O_P(n^{-2/3})$ in Lecture 8, and this is better than the $o_P(n^{-1/2})$ that is needed for asymptotic efficiency of $\hat\theta$.

The functions $\psi_{\theta,\Lambda}$ can be written in the form

$$\psi_{\theta,\Lambda}(x) = \psi(z, e^{\theta^T z}, \Lambda(c), \delta),$$

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 when θ and Λ range freely. Hence the functions $x\mapsto\Lambda(c)Q_{\theta,\Lambda}(x)$ form a Donsker class, by Theorem 6.10.

9.10 Open Problem. Find the limit distribution (if any) of the sequence $n^{1/3}(\hat{\Lambda} - \Lambda)(t)$.

9.4 Profile Likelihood

Given a partitioned parameter (θ, η) and a likelihood $\mathrm{lik}(\theta, \eta)(x)$ the *profile likelihood* for θ is defined as the function

$$\theta \mapsto \mathrm{proflik}(\theta) := \sup_{\eta} \prod_{i=1}^{n} \mathrm{lik}(\theta, \eta)(X_i).$$

The supremum is taken over all possible values of η, or given a sieve all values in the sieve at "time" n. It is rarely possible to compute a profile likelihood explicitly, but its numerical evaluation is often feasible.

The profile likelihood can be used as a computational device, because its point of maximum 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.

However, the importance of the profile likelihood goes far beyond computational issues. Profile likelihood functions can be used in the same way as (ordinary) likelihood functions of parametric models. Besides defining the maximum likelihood estimator $\hat{\theta}$, the curvature of the log profile likelihood at $\hat{\theta}$ can be used as an estimate of minus the inverse of the asymptotic covariance matrix of $\hat{\theta}$. Furthermore, the quotient $\mathrm{proflik}(\hat{\theta})/\mathrm{proflik}(\theta_0)$ between the maximum value and the value at a fixed point θ_0 is the likelihood ratio statistic for testing the (composite) null hypothesis $H_0{:}\,\theta = \theta_0$. In this section we study these quantities more closely.

It is well known that for parametric models with log likelihood $\ell_\theta(x) = \log \mathrm{lik}(\theta)$ the likelihood ratio statistic $2n\mathbb{P}_n(\ell_{\hat{\theta}} - \ell_{\theta_0})$ is under some assumptions and under the null hypothesis $H_0{:}\,\theta = \theta_0$ asymptotically chisquared distributed with degrees of freedom equal to the dimension of the parameter. Furthermore, it is well known that the *observed information* $-\mathbb{P}_n \ddot{\ell}_{\hat{\theta}}$ is, under some conditions, a consistent estimator of the Fisher information $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T = -P_\theta \ddot{\ell}_\theta$. Under some (more) conditions we can prove analogous results for semiparametric models, but with the profile likelihood function for θ replacing the ordinary likelihood.

At the basis of these results is an asymptotic expansion of the (profile) likelihood function as follows. For any random sequence $\tilde{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$,

$$\begin{aligned} \log \mathrm{proflik}_n(\tilde{\theta}_n) &= \log \mathrm{proflik}_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^T \sum_{i=1}^{n} \tilde{\ell}_{\theta_0,\eta_0}(X_i) \\ &\quad - \tfrac{1}{2} n (\tilde{\theta}_n - \theta_0)^T \tilde{I}_{\theta_0,\eta_0} (\tilde{\theta}_n - \theta_0) + o_{P_{\theta_0,\eta_0}}\left(\sqrt{n}\|\tilde{\theta}_n - \theta_0\| + 1\right)^2. \end{aligned} \tag{9.11}$$

If the maximum likelihood estimator is asymptotically efficient, then it possesses the asymptotic expansion

$$\sqrt{n}(\hat\theta_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde I^{-1}_{\theta_0,\eta_0} \tilde\ell_{\theta_0,\eta_0}(X_i) + o_{P_{\theta_0,\eta_0}}(1). \tag{9.12}$$

Taking this into account we see that the parabolic approximation to the log profile likelihood given by equation (9.11) is centered, to the first order, at $\hat\theta_n$. In other words, it is possible to expand the log profile likelihood function around $\hat\theta_n$, in the form

$$\begin{aligned} \log \text{proflik}_n(\tilde\theta_n) &= \log \text{proflik}_n(\hat\theta_n) - \tfrac12 n(\tilde\theta_n - \hat\theta_n)^T \tilde I_{\theta_0,\eta_0}(\tilde\theta_n - \hat\theta_n) \\ &\quad + o_{P_{\theta_0,\eta_0}}\bigl(\sqrt{n}\|\tilde\theta_n - \theta_0\| + 1\bigr)^2. \end{aligned} \tag{9.13}$$

Actually (9.12)–(9.13) are a consequence of (9.11), as we prove below. The expansion (9.11) is firmly believed to be true in some generality. We shall not give precise conditions for its validity here, but note that such conditions have been given in terms of the existence of approximately least favourable paths, much in the spirit of our treatment of maximum likelihood estimators earlier in this lecture.

The asymptotic expansions (9.11) and (9.13) justify using a semiparametric profile likelihood as an ordinary likelihood, at least asymptotically. In particular, we present three corollaries. We assume that the true parameter θ_0 is interior to the parameter set.

9.14 Corollary. *If (9.11) holds, $\tilde I_{\theta_0,\eta_0}$ is invertible, and $\hat\theta_n$ is consistent, then (9.12)–(9.13) hold. In particular, the maximum likelihood estimator $\hat\theta$ is asymptotically efficient at (θ_0, η_0).*

9.15 Corollary. *If (9.11) holds, $\tilde I_{\theta_0,\eta_0}$ is invertible, and $\hat\theta_n$ is consistent, then under the null hypothesis $H_0\colon \theta = \theta_0$, then the sequence $2\log\bigl(\text{proflik}_n(\hat\theta_n)/\,\text{proflik}_n(\theta_0)\bigr)$ is asymptotically chi-squared distributed with d degrees of freedom.*

9.16 Corollary. *If (9.11) holds and $\hat\theta_n$ is consistent, then, for all sequences $v_n \xrightarrow{\text{P}} v \in \mathbb{R}^d$ and $h_n \xrightarrow{\text{P}} 0$ such that $(\sqrt{n}h_n)^{-1} = O_P(1)$,*

$$-2\frac{\log \text{proflik}_n(\hat\theta_n + h_n v_n) - \log \text{proflik}_n(\hat\theta_n)}{nh_n^2} \xrightarrow{\text{P}} v^T \tilde I_0 v.$$

Proofs. The second and third corollaries are immediate consequences of (9.11)–(9.13). Relation (9.13) follows from (9.11)–(9.12) and some algebra. We shall derive (9.12) from (9.11). Set $\Delta_n = n^{-1/2}\sum_{i=1}^n \tilde\ell_{\theta_0,\eta_0}(X_i)$ and $\hat h = \sqrt{n}(\hat\theta - \theta_0)$.

Applying (9.11) with the choices $\tilde\theta = \hat\theta$ and $\tilde\theta = \theta_0 + n^{-1/2}\tilde I^{-1}_{\theta_0,\eta_0}\Delta_n$, we find

$$\begin{aligned} &\log \text{proflik}_n(\hat\theta) \\ &\quad = \log \text{proflik}_n(\theta_0) + \hat h^T \Delta_n - \tfrac12 \hat h^T \tilde I_{\theta_0,\eta_0} \hat h + o_P\bigl(\|\hat h\| + 1\bigr)^2, \\ &\log \text{proflik}_n(\theta_0 + n^{-1/2}\tilde I^{-1}_{\theta_0,\eta_0}\Delta_n) \\ &\quad = \log \text{proflik}_n(\theta_0) + \Delta_n^T \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n - \tfrac12 \Delta_n^T \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n + o_P(1). \end{aligned}$$

By the definition of $\hat\theta$, the expression on the left (and hence on the right) in the first equation is larger than the expression on the left in the second equation. It follows that

$$\hat h^T\Delta_n - \tfrac12\hat h^T\tilde I_{\theta_0,\eta_0}\hat h - \tfrac12\Delta_n^T\tilde I^{-1}_{\theta_0,\eta_0}\Delta_n \geq -o_P\big(\|\hat h\|+1\big)^2.$$

The left side of this inequality is equal to

$$-\tfrac12(\hat h - \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n)^T\tilde I_0(\hat h - \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n) \leq -c\|\hat h - \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n\|^2,$$

for a positive constant c, by the nonsingularity of $\tilde I_{\theta_0,\eta_0}$. Conclude that

$$\|\hat h - \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n\| = o_P\big(\|\hat h\|+1\big).$$

This implies first that $\|\hat h\| = O_P(1)$, and next, by reinsertion, that $\|\hat h - \tilde I^{-1}_{\theta_0,\eta_0}\Delta_n\| = o_P(1)$. This completes the proof of (9.12). ∎

9.17 Example (Cox model). Consider again the Cox model of Example 3.13. In Lecture 5 we noted that the second component of the maximum likelihood estimator $(\hat\theta,\hat\Lambda)$, relative to the likelihood chosen there, will be a step function with steps only at the Y_i such that $\Delta_i = 0$. It follows that the profile likelihood function takes the form

$$\theta \mapsto \sup_{\lambda_1,\ldots,\lambda_n\geq 0}\prod_{i=1}^n\Big(e^{\theta Z_i}\lambda_i\Big)^{\Delta_i}e^{-e^{\theta Z_i}\sum_{j:Y_j\leq Y_i}\lambda_j\Delta_j}.$$

In this (very special) case the supremum can be explicitly computed. Finding the maximizers over $(\lambda_1,\ldots,\lambda_n)$ is equivalent to maximizing

$$\sum_{i=1}^n\Delta_i\log\lambda_i - \sum_{i=1}^n\sum_{j:Y_j\leq Y_i}e^{\theta Z_i}\lambda_j\Delta_j.$$

Interchanging the sums and next taking the partial derivative relative to λ_k for a k such that $\Delta_k = 1$, yields the stationary equation

$$\frac{1}{\lambda_k} = \sum_{i:Y_i\geq Y_k}e^{\theta Z_i}.$$

Upon inserting this in the likelihood we find the profile likelihood for θ

$$\theta\mapsto\prod_{i=1}^n\Big(\frac{e^{\theta Z_i}}{\sum_{j:Y_j\geq Y_i}e^{\theta Z_j}}\Big)^{\Delta_i}e^{-\sum_{i=1}^n\Delta_i}.$$

This expression is known as the *Cox partial likelihood*. Cox's original motivation for this criterion function is that the terms in the product are the conditional probabilities that the ith subject dies at time Y_i given that one of the subjects at risk dies at that time.

The Cox partial log likelihood is a sum over the observations, but the terms in the sum are dependent. Direct study of such a sum therefore is nontrivial at first sight. Initially the Cox partial likelihood estimator was studied along the classical lines: characterizing $\hat\theta$ as the solution of the derivative of the partial likelihood and next using Taylor series arguments on this partial score equation. The difficulty is

then to show that the partial score and its derivative are asymptotically normal or satisfy a law of large numbers. Later it turned out that martingale arguments can both justify this derivation and facilitate the calculation of means and variances. Elegant as this arguments are, they are restricted to a special case such as the Cox model. In the final lecture we shall show how the asymptotic normality of the Cox estimators can be derived within a framework that applies to general semiparametric models. Alternatively, the asymptotic normality of $\hat{\theta}$ follows along the lines of the present lecture. □

Notes

The treatment of the Cox model with current status data follows [12], who also presents a general set-up. Our definition of approximately least favourable submodels is based on [40] and [26]. The latter paper discusses the profile likelihood function and summarizes other work on the likelihood ratio statistic and the observed information. For an analysis of the sieved or penalized logistic regression model see [13] and [21].

Lecture 10
Infinite-dimensional Z-Estimators

In this lecture we consider infinite-dimensional systems of estimating equations and show that solutions are asymptotically normal if the system is appropriately differentiable, extending the results on finite-dimensional Z-estimators to infinite dimensions. Next we show that this method can be applied to proving asymptotic normality of maximum likelihood estimators in semiparametric models, with as example, again, the Cox model.

10.1 General Result

A system of estimating equations for a parameter must be of the same dimension as the parameter. For an infinite-dimensional parameter we need infinitely many estimating equations. It turns out that such a system can be analyzed much in the same way as a finite-dimensional system, provided that we substitute functional analysis for multivariate calculus. 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. Whereas in the finite-dimensional situation the use of empirical processes can be avoided through higher order Taylor expansions, now empirical processes appear indispensable. But we do not mind that, of course, having established already all the tools we need.

For each θ in a subset Θ of a Banach space and each h in an arbitrary set $\mathcal{H}$, let $\psi_{\theta,h}\colon \mathcal{X} \mapsto \mathbb{R}$ be a measurable function. Denote by $\psi_\theta(x)$ the vector-valued function $\{\psi_{\theta,h}(x)\colon h \in \mathcal{H}\}$ and let $\mathbb{P}_n\psi_\theta$ and $P\psi_\theta$ be the corresponding vector-valued empirical and "true" means. We are interested in zeros $\hat\theta$ of the map $\theta \mapsto \mathbb{P}_n\psi_\theta$. Equivalently, in random elements $\hat\theta$ with values in Θ such that

$$\mathbb{P}_n\psi_{\theta,h} = 0, \qquad \text{every } h \in \mathcal{H}.$$

We expect that the sequence $\hat\theta_n$ converges in probability to a zero of the map $\theta \mapsto P\psi_\theta$. In applications where $\hat\theta$ is a maximum likelihood or another contrast estimator, we usually already know this from applying a standard method to the contrast function. It may also be possible to establish consistency from the fact that $\hat\theta$ is a zero only. In any case, the consistency issue does not yield structurally different questions from before and we omit further discussion.

We assume that the maps $h \mapsto \psi_\theta(x)$ and $h \mapsto P\psi_{\theta,h}$ are uniformly bounded, so that the maps $\theta \mapsto \mathbb{P}_n\psi_\theta$ and $\theta \mapsto P\psi_\theta$ map Θ into $\ell^\infty(\mathcal{H})$. This may seem a bit special, but even when considering maps $\theta \mapsto \psi_\theta(x)$ with values in a general Banach space, we can always force this in the present form by choosing the right index set $\mathcal{H}$. The advantage of the present special form is that we set the theorems immediately within the context of empirical processes.

The following theorem establishes the asymptotic normality of $\sqrt{n}(\hat\theta - \theta)$ and should be compared to Theorem 6.19. Recall that *Fréchet differentiability* is ordinary differentiability. Thus the map $\theta \mapsto P\psi_\theta$ is Fréchet differentiable at θ_0 if there exists a continuous, linear map $V\colon \operatorname{lin}\Theta \mapsto \ell^\infty(\mathcal{H})$ such that, as $\theta \mapsto \theta_0$,

$$\big\|P\psi_\theta - P\psi_{\theta_0} - V(\theta - \theta_0)\big\|_{\mathcal{H}} = o\big(\|\theta - \theta_0\|\big).$$

In our setting we do not assume that the domain of the map $\theta \mapsto P\psi_\theta$ contains θ_0 as an interior point, but allow Θ to be arbitrary. The sequence $\theta \mapsto \theta_0$ in the preceding display is restricted to Θ.

10.1 Theorem. *Suppose that the class of functions $\{\psi_{\theta,h}\colon \theta \in \Theta, h \in \mathcal{H}\}$ is P-Donsker, that the map $\theta \mapsto P\psi_\theta$ is Fréchet differentiable at θ_0 with derivative $V\colon \operatorname{lin}\Theta \mapsto \ell^\infty(\mathcal{H})$ that is one-to-one and has a continuous inverse $V^{-1}\colon \mathrm{R}(V) \mapsto \operatorname{lin}\Theta$. Furthermore, assume that the maps $\theta \mapsto \psi_{\theta,h}$ are continuous in $L_2(P)$ at θ_0, uniformly in $h \in \mathcal{H}$. Then any zero $\hat\theta_n$ of $\theta \mapsto \mathbb{P}_n\psi_\theta$ that converges in probability to a zero θ_0 of $\theta \mapsto P\psi_\theta$ satisfies*

$$V\sqrt{n}(\hat\theta - \theta_0) = \mathbb{G}_n\psi_{\theta_0} + o_P(1).$$

Proof. The first step is to prove that $\mathbb{G}_n(\psi_{\hat\theta_n} - \psi_{\theta_0}) \xrightarrow{\mathrm{P}} 0$ in $\ell^\infty(\mathcal{H})$. Equip the set $\mathcal{H} \times \Theta$ with the semi-metric

$$\rho\big((h,\theta),(h',\theta')\big) = \sqrt{P(\psi_{\theta,h} - \psi_{\theta',h'})^2},$$

and define a map $\phi\colon \ell^\infty(\mathcal{H}\times\Theta) \times \Theta \mapsto \ell^\infty(\mathcal{H})$ by $\phi(z,\theta) = z(\cdot,\theta) - z(\cdot,\theta_0)$. By assumption we have that $\rho\big((h,\theta),(h,\theta_0)\big) \to 0$, uniformly in $h \in \mathcal{H}$, as $\theta \to \theta_0$. Thus if $z \in \ell^\infty(\mathcal{H}\times\Theta)$ is ρ-uniformly continuous, then $|z(h,\theta) - z(h,\theta_0)| \to 0$, uniformly in $h \in \mathcal{H}$, if $\theta \to \theta_0$. Consequently, for such z and for $(z_n,\theta_n) \mapsto (z,\theta_0)$ an arbitrary sequence in $\ell^\infty(\mathcal{H}\times\Theta)\times\Theta$,

$$\begin{aligned}\big\|\phi(z_n,\theta_n) - \phi(z,\theta_0)\big\|_{\mathcal{H}} &= \big\|z_n(h,\theta_n) - z_n(h,\theta_0)\|_{\mathcal{H}} \\ &\le 2\|z_n - z\|_{\mathcal{H}\times\Theta} + \big\|z(h,\theta_n) - z(h,\theta_0)\big\|_{\mathcal{H}} \to 0.\end{aligned}$$

We conclude that the map ϕ is continuous at every point (z,θ_0) such z is ρ-uniformly continuous at θ_0. Almost all sample paths of a Brownian bridge are uniformly continuous relative to the $L_2(P)$-norm and therefore almost all sample paths $(\theta,h) \mapsto Z(\theta,h)$ of the process $Z(\theta,\eta) = \mathbb{G}\psi_{\theta,h}$ are uniformly continuous relative to ρ. By assumption we have that $Z_n \rightsquigarrow Z$ and that $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$. Hence $(Z_n,\hat\theta_n) \rightsquigarrow (Z,\theta_0)$ and by the continuous mapping theorem we conclude that $\phi(Z_n,\hat\theta_n) \rightsquigarrow \phi(Z,\theta_0) = 0$. This is equivalent to the claim that $\mathbb{G}_n(\psi_{\hat\theta_n} - \psi_{\theta_0}) \xrightarrow{\mathrm{P}} 0$ in $\ell^\infty(\mathcal{H})$.

Using the fact that $\hat\theta$ and θ_0 are zeros we can rewrite the claim as

$$-\sqrt{n}P(\psi_{\hat\theta_n} - \psi_{\theta_0}) = \mathbb{G}_n \psi_{\theta_0} + o_P(1).$$

The right side converges in distribution in $\ell^\infty(\mathcal{H})$, by the Donsker assumption. Hence its norm is $O_P(1)$. The left side can be written as

$$-\sqrt{n}\big(V(\hat\theta_n - \theta_0) + o_P(\|\hat\theta_n - \theta_0\|)\big)$$

by the assumption of Fréchet differentiability. Because V has a continuous inverse on its range, there exists a constant $c > 0$ such that $\|V(\theta-\theta_0)\| \geq c\|\theta-\theta_0\|$ for every $\theta \in \Theta$. We use this and the preceding displays to conclude that $\sqrt{n}\|\hat\theta_n-\theta_0\| = O_P(1)$. Next we insert this in the preceding display to see that the display is equivalent to $-V\sqrt{n}(\hat\theta_n - \theta_0) + o_P(1)$. ∎

We can invert the assertion of the preceding theorem to see that $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically distributed as $V^{-1}\mathbb{G}\psi_{\theta_0}$ provided we use the correct (continuous) extension of the inverse operator V^{-1} to a domain that contains the support of the Brownian bridge $\mathbb{G}\psi_{\theta_0}$. Because continuous, linear transformations of Gaussian processes are Gaussian we obtain that the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal. In many situations, though, the limit distribution is easier found by performing the inversion of the relation for a finite n. We shall see an example of this in the following treatment of maximum likelihood estimators.

An important condition in the theorem is the continuous invertibility of the derivative V. Since 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. Since 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 V^{-1} more readily continuous, but makes the differentiability of the map $\theta \mapsto P\psi_\theta$ and the Donsker condition more stringent.

10.2 Maximum Likelihood

Consider a semiparametric model, indexed by a finite-dimensional parameter θ of interest and a nuisance parameter η, assumed to be contained in some Banach space. We wish to apply the preceding theorem to derive the asymptotic distribution of the pair $(\hat\theta, \hat\eta)$ of maximum likelihood estimators. (Thus θ of the theorem becomes (θ, η) in this section.) This approach gives an alternative to the one of Lecture 9 based on the efficient score equation. A limitation of the present 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 be parametrized such that this is the case, while it is certainly not always the case for the natural parametrization. An advantage is that we obtain the joint asymptotic distribution of $\hat\theta$ and $\hat\eta$.

The system of estimating equations that we are looking for will consist 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

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

for $\operatorname{lik}(\theta, \eta)(x)$ 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 $\log \operatorname{lik}(\theta, \eta)(x)$ 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 will take 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 will be 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,

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

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).

Our total set of estimating equations may be thought of as indexed by the set $\{1, \ldots, k\} \cup \mathcal{H}$. We can summarize the estimating equations in a random map $\Psi_n \colon \mathbb{R}^k \times H \mapsto \mathbb{R}^k \times \ell^\infty(\mathcal{H})$ given 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,

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

Under some conditions, Theorem 10.1 gives the asymptotic distribution of the sequence $\sqrt{n}(\hat\theta-\theta_0,\hat\eta-\eta_0)$ as a function of the derivative $\dot\Psi_0$ of Ψ at (θ_0,η_0) and the limit process of $\sqrt{n}(\Psi_n-\Psi)(\theta_0,\eta_0)$, a pair of a Gaussian vector and a Brownian bridge process.

We would like to make this limit process more concrete and ascertain that the maximum likelihood estimator is asymptotically efficient. Then we need to relate the derivative of Ψ to the score and information operators of the model. 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 $h\colon \mathcal{Z}\mapsto\mathbb{R}$, 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 second. 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

$$\begin{aligned}
\dot\Psi_{11}(\theta-\theta_0)&=-P_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}\dot\ell^T_{\theta_0,\eta_0}(\theta-\theta_0),\\
\dot\Psi_{12}(\eta-\eta_0)&=-\int B^*_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}\,d(\eta-\eta_0),\\
\dot\Psi_{21}(\theta-\theta_0)h&=-P_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}h)\dot\ell^T_{\theta_0,\eta_0}\,(\theta-\theta_0),\\
\dot\Psi_{22}(\eta-\eta_0)h&=-\int B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}h\,d(\eta-\eta_0).
\end{aligned}\tag{10.2}$$

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

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

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

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

On comparing the preceding pair of displays, we obtain the last line of (10.2). These arguments are purely heuristic, and this form of the derivative must be established for every example. For instance, within the context of Theorem 10.1, we may need to apply $\dot\Psi_0$ to η that are not absolutely continuous with respect to η_0. Then the validity of (10.2) already 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 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 when 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

$$\dot V(\eta - \eta_0)h = -\int (B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0} + K)h\, d(\eta - \eta_0),$$

where the operator K is defined as

$$Kh = -\big(P_{\theta_0,\eta_0}(B_{\theta_0,\eta_0}h)\dot\ell^T_{\theta_0,\eta_0}\big)I^{-1}_{\theta_0,\eta_0}B^*_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0}.$$

The operator $\dot V\colon \operatorname{lin} H \mapsto \ell^\infty(\mathcal{H})$ is certainly continuously invertible if there exists a positive number ϵ such that, for all $\eta \in \operatorname{lin}\mathcal{H}$

$$\sup_{h\in\mathcal{H}}\big|\dot V(\eta - \eta_0)h\big| \geq \varepsilon\|\eta - \eta_0\|.$$

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 $\sup_{h\in\mathcal{H}}\big|(\eta - \eta_0)h\big|$. Then the display is certainly satisfied if, for some $\varepsilon > 0$,

$$\Big\{(B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0} + K)h\colon h \in \mathcal{H}\Big\} \supset \epsilon\mathcal{H}.$$

This condition has a nice interpretation if $\mathcal{H}$ is equal to the unit ball of a Banach space $\mathbb{D}$ of functions. Then the preceding display is implied by the operator $B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0} + K\colon \mathbb{D} \mapsto \mathbb{D}$ being continuously invertible. The first part of this operator is the information operator for the nuisance parameter. Typically, this would be continuously invertible if the nuisance parameter is regularly estimable at a $\sqrt{n}$-rate (relatively to the norm used) when 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, i.e. the parameters θ and η are not locally confounded.

10.3 Lemma. *Let $\mathcal{H}$ be the unit ball in a Banach space $\mathbb{D}$ contained in $\ell^\infty(\mathcal{Z})$. If $\tilde I_{\theta_0,\eta_0}$ is nonsingular, $B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0}\colon \mathbb{D} \mapsto \mathbb{D}$ is continuous, onto and continuously invertible and $B^*_{\theta_0,\eta_0}\dot\ell_{\theta_0,\eta_0} \in \mathbb{D}$, then $B^*_{\theta_0,\eta_0}B_{\theta_0,\eta_0} + K\colon \mathbb{D} \mapsto \mathbb{D}$ is continuous, 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 10.4 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 $h \in \mathbb{D}$. By definition $Kh = a_0^TB_o^*\dot\ell_0$ for for $a_0 = -I_0^{-1}P_0(B_0h)\dot\ell_0$. By assumption there exists a path $t \mapsto \eta_t$ with score function $\overline{B}_0h = B_0h - P_0B_0h$ at $t = 0$. Then the submodel indexed by $t \mapsto (\theta_0 + ta_0, \eta_t)$ has score function $a_0^T\dot\ell_0 + \overline{B}_0h$ at $t = 0$, and information

$$a_0^TI_0a_0 + P_0(\overline{B}_0h)^2 + 2a_0^TP_0\dot\ell_0(B_0h) = P_0(\overline{B}_0h)^2 + a_0^TI_0a_0.$$

Since 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 would be 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 $K: \mathbb{D} \mapsto \mathbb{D}$ 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.

10.4 Lemma. *Let $\mathbb{D}$ be a Banach space, let the operator $A: \mathbb{D} \mapsto \mathbb{D}$ be continuous, onto and continuously invertible and let $K: \mathbb{D} \mapsto \mathbb{D}$ 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 (10.2), the assertion of Theorem 10.1 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 equation (3.12), 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.

10.5 Example (Cox model). We come back to the Cox model one more time. We recall that the scores and the information operator are given by

$$\begin{aligned} \dot{\ell}_{\theta,\Lambda}(x) &= \delta z - z e^{\theta z}\Lambda(y) \\ B_{\theta,\Lambda}h(x) &= \delta h(y) - e^{\theta z}\int_{[0,y]} h\, d\Lambda \\ B^*_{\theta,\Lambda}B_{\theta,\Lambda}h(y) &= h(y)\,\mathrm{E}_{\theta,\Lambda}1_{Y \geq y}e^{\theta Z} \\ B^*_{\theta,\Lambda}\dot{\ell}_{\theta,\Lambda} &= \mathrm{E}_{\theta,\Lambda}1_{Y \geq y}Z e^{\theta Z}. \end{aligned}$$

As in the preceding discussion we set up estimating equations $\mathbb{P}_n \dot{\ell}_{\theta,\Lambda} = 0$ and $\mathbb{P}_n B_{\theta,\Lambda} h = 0$. Here we let h range over the unit ball of the space $\mathrm{BV}[0,\tau]$ of functions $h\colon [0,\tau] \mapsto \mathbb{R}$ of bounded variation (with norm the supremum of the uniform norm and the variation norm). The expectations of these equations are given by the maps $\Psi_1(\theta,\Lambda) = P_0 \dot{\ell}_{\theta,\Lambda}$ and $\Psi_2(\theta,\Lambda)h = P_0 B_{\theta,\Lambda} h$.

We can now directly verify the validity of formula (10.2). for the derivative of the map $\Psi = (\Psi_1, \Psi_2)$ The map Ψ is already linear in Λ. With $G_0(y|\, Z)$, the distribution function of Y given Z, it can be written as

$$\Psi_1(\theta,\Lambda) = \mathrm{E} Z e^{\theta_0 Z} \int \overline{G}_0(y|\, Z)\, d\Lambda_0(y) - \mathrm{E} Z e^{\theta Z} \int \Lambda(y)\, dG_0(y|\, Z),$$

$$\Psi_2(\theta,\Lambda)h = \mathrm{E} e^{\theta_0 Z} \int h(y) \overline{G}_0(y|\, Z)\, d\Lambda_0(y) - \mathrm{E} e^{\theta Z} \iint_{[0,y]} h\, d\Lambda\, dG_0(y|\, Z).$$

The map $\Psi\colon \mathbb{R} \times \ell^\infty(\mathcal{H}) \mapsto \mathbb{R} \times \ell^\infty(\mathcal{H})$ 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 neighbourhood of (θ_0, Λ_0). Several applications of Fubini's theorem show that indeed the derivative takes the form (10.2).

The operator $B_0^* B_0$, initially introduced as acting on $L_2(\Lambda)$, can also be viewed as an operator of the space $\mathrm{BV}[0,\tau]$ into itself. 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 $[0,\tau]$, which is part of our assumptions. In Lecture 3 we already computed the efficient information and noted its positivity (under the assumption that Z is not almost surely equal to a function of $h(Y)$). Thus, we can conclude that the map $\dot{\Psi}_0$ is continuously invertible by Lemma 10.3.

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 Theorem 6.10.

The other conditions of Theorem 10.1 are satisfied too. We finish our lectures with the conclusion that the maximum likelihood estimator in the Cox model, alias the partial likelihood estimator, is asymptotically efficient.

We are not the first to conclude this, but we still feel that this is a worthy conclusion of the lectures, remembering that the present approach also applies to other models. □

Notes

This lecture has its roots in [39].

References

[1] Begun, J.M., Hall, W.J., Huang, W.M. and Wellner, J.A., (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Annals of Statistics* **11**, 432–452.

[2] Bickel, P.J., (1982). On adaptive estimation. *Annals of Statistics* **10**, 647–671.

[3] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A., (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

[4] Birgé, L. and Massart, P., (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields* **97**, 113–150.

[5] Birgé, L. and Massart, P., (1994). Minimum contrast estimators on sieves. *preprint*,.

[6] Birgé, L. and Massart, P., (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli* **4**, 329–375.

[7] Cox, D.R., (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society b* **34**, 186–220.

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

[9] Dudley, R.M., (1984). A course on empirical processes. École d'été de Probabilités de Saint-Flour XII-1982. *Lecture Notes in Mathematics* **1097**, 1–142. Springer.

[10] Hájek, J., (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 323–330.

[11] Hájek, J., (1972). Local asymptotic minimax and admissibility in estimation. Proc. Sixth Berkeley Symp. Math. Statist. Prob. **1**, 175–194, (eds: L.M. LeCam, J. Neyman, and E. Scott).

[12] Huang, J., (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics* **24**, 540–568.

[13] Chen, H., (1996). Asymptotically efficient estimation in semiparametric generalized linear models. *Annals of Statistics* **23**, 1102–1129.

[14] Klaassen, C.A.J., (1987). Consistent estimation of the influence function of locally asymptotically linear estimates. *Annals of Statistics* **15**, 1548-1562.

[15] Le Cam, L., (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics* **3**, 37–98.

[16] Le Cam, L., (1969). *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal.

[17] Le Cam, L., (1972). Limits of experiments. Proc. Sixth Berkeley Symp. Math. Statist. Probab. **1**, 245–261, (eds: L.M. LeCam, J. Neyman, and E. Scott). University of California Press, Berkeley.

[18] Le Cam, L., (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

[19] Koshevnik, Yu.A. and Levit, B.Ya., (1976). On a nonparametric analogue of the information matrix. *Theory of Probability and its Applications* **21**, 738–753.

[20] Levit, B.Ya., (1978). Infinite-dimensional informational lower bounds. *Theory of Probability and its Applications* **23**, 388–394.

[21] Mammen, E. and van de Geer, S.A., (1997). Penalized quasi-likelihood estimation in partial linear models. *Annals of Statistics* **25**, 1014–1037.

[22] Millar, P.W., (1983). The minimax principle in asymptotic statistical theory. École d'été de Probabilités de Saint-Flour XI-1981. *Lecture Notes in Mathematics* **976**, 67-267. Springer.

[23] Millar, P.W., (1985). Nonparametric applications of an infinite dimensional convolution theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68**, 545–556.

[24] Murphy, S.A., Rossini, T.J. and van der Vaart, A.W., (1997). MLE in the proportional odds model. *Journal of the American Statistical Association* **92**, 968–986.

[25] Murphy, S.A. and van der Vaart, A.W., (1996). Likelihood ratio inference in the errors-in-variables model. *Journal of Multivariate Analysis* **59**, 81–108.

[26] Murphy, S.A. and van der Vaart, A.W., (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, to appear.

[27] Pfanzagl, J., (1988). Consistency of maximum likelihood estimators for certain nonparametric families, in particular mixtures. *Journal of Statistical Planning and Inference* **19**, 137–158.

[28] Pfanzagl, J. and Wefelmeyer, W., (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statistics* **13**. Springer Verlag, New York.

[29] Pollard, D., (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.

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

[31] Ritov, Y. and Bickel, P.J., (1990). Achieving information bounds in non and semi-parametric models. *Annals of Statistics* **18**, 925–938.

[32] Robins, J.M. and Ritov, Y., (1992). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16**, 285–319.

[33] Rudin, W., (1974). *Functional Analysis*. McGraw Hill.

[34] Stein, C., (1956). Efficient nonparametric testing and estimation. Proceedings Third Berkeley Symposium Math. Statist. Probability **1**, 267–284. University of California, Berkeley.

[35] Taupin, M.-.L., (1998). Estimation in the nonlinear errors-in-variables model. *C. R. Acad. Sci. Paris Sr. I Math.* **326 - 7**, 885–890.

[36] van der Vaart, A.W., (1988). *Statistical Estimation in Large Parameter Spaces*. *CWI Tracts* **44**. Centrum voor Wiskunde en Informatica, Amsterdam.

[37] van der Vaart, A.W., (1991). On differentiable functionals. *Annals of Statistics* **19**, 178–204.

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

[39] van der Vaart, A.W., (1994). Maximum likelihood estimation with partially censored observations. *Annals of Statistics* **22**, 1896–1916.

[40] van der Vaart, A.W., (1996). Efficient estimation in semiparametric models. *Annals of Statistics* **24**, 862–878.

[41] van der Vaart, A.W. and Wellner, J.A., (1996). *Weak Convergence and Empirical Processes*. Springer, New York.

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