EFFICIENT ESTIMATION FOR RESTRICTED SEMIPARAMETRIC MODELS

by

Yuly Koshevnik
Department of Statistical Science
Southern Methodist University

Technical Report No. SMU/DS/TR-269

Research Sponsored by
NSF Grant DMS–9311477

May 1994

Abstract[1].

Asymptotically efficient estimates are studied for semiparametric models under restrictions. Independent identically distributed (i.i.d.) observations are drawn from a distribution P, satisfying the equation $\Psi(P) = 0$. A transform Ψ is possibly nonlinear, takes its values in a suitable Banach space, and is defined on a sufficiently large family $\mathcal{Q}$ of probabilities. The asymptotically efficient estimate of a real-valued functional $\Phi(P)$ is obtained using a general procedure based upon initial estimates for both $\Phi(P)$ and $\Psi(P)$, which are assumed to be asymptotically well behaved under the assumption $P \in \mathcal{Q}$. The geometric ideas, often appearing in nonparametric and semiparametric estimation problems, are applied for both the construction of the desired estimates by means of the adjustment procedure and justification of their asymptotic behavior.

[1] *AMS* 1980 *subject classfications.* 62G20, 62G05, 60F05.
Keywords: *semiparametric estimation, asymptotically efficient estimates, restricted estimation problems, lower bounds of risks, uniform weak convergence.*

Chapter 1

Introduction.

This paper presents a unified approach to the construction of asymptotically efficient estimates for a real–valued functional (or coparameter) $\Phi(P)$ of the unknown true distribution P. The desired estimator $\{\hat{\Phi}^n = \hat{\Phi}(X^n) : n \geq 1\}$ is based on the sample of i.i.d. random variables

$$X^n = (X_1, X_2, \ldots, X_n).$$

This sample may be thought of as n i.i.d. observations of a random variable X, taking values in a measure space $(\mathbf{X}, \mathcal{A})$. The unknown distribution P belongs to a family $\mathcal{P}$ of probabilities on this measure space.

A restricted estimation problem is described as follows. Suppose a wider family $\mathcal{Q}$ of distributions on $(\mathbf{X}, \mathcal{A})$ is given and for a rich enough class $\mathbf{\Lambda}$ of functionals Λ, defined on $\mathcal{Q}$, the asymptotically efficient estimator–sequence $\{\Lambda^n(X^n)\}$ can be constructed. The family $\mathcal{P}$ can be represented as a subfamily of $\mathcal{Q}$, satisfying the system of equations:

$$\Psi_a(P) = 0.$$

The subscript a here numbers the equations, and Ψ_a belongs to the class $\mathbf{\Lambda}$, for every a. The restrictions on P provide additional prior information, so that the estimate $\tilde{\Phi}$ of $\Phi(P)$, derived for the unrestricted family $\mathcal{Q}$, may be improved. There is a more convenient way to describe this problem. We will denote any map from the family $\mathcal{Q}$ into an arbitrary linear space $\mathbf{B}$ as a coparameter, while the parameter will be thought of as a subscript distinguishing distributions from a a finite–dimensional family or subfamily in $\mathcal{Q}$. In other words, a parameter is defined by means of a map (that is called a parametrization) from a subset $U \subset \mathbf{R}^m$ into a family $\mathcal{Q}$ of distributions. As far as $\mathbf{B}$ is concerned, it will be always a Banach space (infinite–dimensional, perhaps), and $\Psi(P)$ will be referred to as a $\mathbf{B}$–valued coparameter on $\mathcal{Q}$. The family $\mathcal{P}$ is then defined by the equation $\Psi(P) = 0$. The main question is as follows: *What should we do to construct the asymptotically efficient estimator–sequence for $\Phi(P)$ given that $P \in \mathcal{P}$?*

Several examples will show what has already been done in this area. The general results formulated here will allow us to develop a unified approach to the desired general construction. This approach may be thought of as an infinite–dimensional version of the well known one–step approximation to the maximum likelihood estimate (MLE). Some of the results in this direction can be found in Millar (1983), Pfanzagl (1982), Koshevnik and Levit (1976), Koshevnik (1985, 1984a, 1984c), Bickel and Ritov (1990), Bickel, Ritov, and Wellner (1991), Gill (1989), and Schick (1986, 1987).

Before we go on, let us consider a simple example from Koshevnik and Levit (1976), also mentioned by Pfanzagl (1982). It suggests a useful approach for this paper and explains what should be done generally.

Example. Suppose that the family $\mathcal{Q}$ is sufficiently large and the coparameters $\Phi(P) \in \mathbf{R}$ and $\Psi(P) \in \mathbf{R}^m$ are expressed as the expectations: $\Phi(P) = \int \phi(x)\,dP(x)$ and $\Psi_j(P) = \int \psi_j(x)\,dP(x)$, where $j = 1, 2 \ldots, m$. For the sake of convenience, we also assume that all of these functions, ϕ and $\psi(\cdot, j)$ are bounded. The restricted problem is related to a subfamily $\mathcal{P} \subset \mathcal{Q}$ such that $\Psi_j(P) = 0,$ for all j. If the unknown true value $\Phi(P)$ is initially estimated by means of its empirical version, then the restricted problem admits some improvement, at least asymptotically. The modified estimate may be constructed as a simple adjustment to $\Phi(P^n) = \tilde{\Phi}^n$, where P^n is the empirical *CDF*.

Consider all possible estimates of $\Phi(P)$ that can be represented as a linear combination

$$\hat{\Phi} = n^{-1} \sum_{i=1}^{n} \left[\phi(X_i) - \sum_{j=1}^{m} \psi_j(X_i,)\beta_j \right]$$

of the empirical versions of $\Phi(P)$ and all $(\Psi_j(P) : 1 \leq j \leq m)$. The coefficients $\{\beta_j : 1 \leq j \leq m\}$ should be chosen to minimize the asymptotic variance of $\hat{\Phi}$. Of course, the optimal choice depends on P. However, after the empirical estimate $\beta(P^n)$ is substituted for $\beta(P)$, the asymptotic efficiency can be usually achieved. These heuristic explanations have been justified by the results of Koshevnik and Levit (1976), showing that the asymptotic bounds of risks can be achieved for these "natural" estimates.

The system of equations, describing the choice of coefficients $\beta = (\beta_1, \ldots, \beta_m)$, is simple enough:

$$\sum_{k=1}^{m} \langle \psi_j, \psi_k \rangle_P \, \beta_k = \langle \psi_j, \phi \rangle_P \, .$$

Here, for two square integrable functions, say $\xi, \eta \in \mathbf{L}^2(P)$, their inner product $\int \xi(x)\,\eta(x)\,dP(x)$ is denoted as $\langle \xi, \eta \rangle_P$.

Remark. In fact, this system involves a projection of the coparameter of interest onto the subspace spanned over restrictions. The similarity with the least square estimates for a linear regression model is obvious here. ■

We assume generally that the coparameters $\Phi(P)$ and $\Psi(P) \in \mathbf{B}$ admit asymptotically efficient estimates, say $\tilde{\Phi}$ and $\tilde{\Psi} \in \mathbf{B}$, respectively, when the condition $P \in \mathcal{Q}$ holds. Later, for an infinite–dimensional space $\mathbf{B}$, even the requirement $\tilde{\Psi} \in \mathbf{B}$ will be relaxed, and the estimates will no longer be considered as $\mathbf{B}$–valued random variables. Instead, they may take their values in a certain extension $\mathbf{D} \supset \mathbf{B}$, as we will show.

Let $\mathbf{B}^*$ denote the dual space to $\mathbf{B}$, that is a set of all linear and continuous functionals β. It will be convenient to denote the value that β takes at $b \in \mathbf{B}$ as $\langle b, \beta \rangle = \beta(b)$. The system of equations $\langle \Psi(P), \beta \rangle = 0$, with β running over the dual space $\mathbf{B}^*$, is simply the equivalent form of the equation $\Psi(P) = 0$. We will develop this tool to make it work in much more complicated situations.

Both the main results and their applications extensively use the geometric approach to estimation problems, as it has been exploited in Koshevnik and Levit (1976), and then developed by Pfanzagl (1982), Begun, Hall, Huang, and Wellner (1983), as well as many other authors. The necessary explanations are given throughout the paper. Chapter 2 recalls some important notions and definitions and exhibits the information inequalities, describing the lower bounds of asymptotic risks. Chapter 3 contains the weak convergence theorems for the proposed estimates. It turns out, under suitable assumptions, that the asymptotic normality holds uniformly in $P \in V$, where V is a small neighborhood of an unknown true distribution. Chapter 4 provides detailed explanations for some examples. Although they are either relatively simple or well known, the attractiveness of the proposed approach is more evident, when the complicated matters are avoided in the examples. The proofs and useful technical details on a uniform weak convergence are presented in Chapter 5.

Chapter 2

Geometric approach to efficient nonparametric estimation.

This section develops the geometric tools to answer two major questions arising in the nonparametric theory. For estimates of a real valued coparameter $\Phi(P)$ from i.i.d. observations with a common distribution $P \in \mathcal{P}$, these questions are formulated as follows.

1. How can one describe the lower bounds of risk for any estimator–sequence $\hat{\Phi} = \{\hat{\Phi}^n : n \geq 1\}$?

2. How can one construct the asymptotically efficient estimator–sequence, achieving the described bound in the limit, as the sample size $n \to \infty$?

The geometric language, applied to restricted estimation problems for a real–valued coparameter $\Phi(P)$, obtains lower bounds of risks quite simply. It turns out, however, that the desired asymptotic efficiency requires some additional conditions on the estimates of both $\Phi(P)$ and $\Psi(P)$, in terms of their asymptotic performance for a family $\mathcal{Q} \supset \mathcal{P}$.

The notions and definitions used to formulate the key ideas are reviewed in the next subsection. (References to Begun, et al. (1983), Beran (1980), Bickel (1982), van der Vaart (1991), Pfanzagl (1982), Koshevnik and Levit (1976, 1980), and Koshevnik, (1984c) are relevant, as well as many others.) Although parametric families are not given careful consideration here, several definitions are essentially based on the limiting properties of the likelihood ratio and its square root for parametric subfamilies, or finite–dimensional subsets surrounding any fixed given distribution P.

2.1 Geometric notions and differentiable coparameters.

2.1.1 Smooth parametric subfamilies and paths.

Let $U = (-\delta, \delta)$ be an open interval surrounding $0 \in \mathbf{R}$, and $\gamma : u \mapsto P_u$ from U into $\mathcal{Q}$ be a given map. For any measures, say P and Q on $(\mathbf{X}, \mathcal{A})$, let $Q = Q^c + Q^s$ denote the decomposition of Q into the sum of an absolutely continuous (with respect to P) component Q^c and a singular part Q^s. The total variation of any measure R is denoted as $\| R \|$. For measures Q and P, let dQ/dP denote the Radon–Nikodym derivative of the measure Q^c with respect to P.

Definition. Let $\{P_u : u \in U\}$ be a subfamily in $\mathcal{Q}$, indexed by $U \subset \mathbf{R}$, such that $0 \in U$ and $P_0 = P$, where P is a fixed given probability distribution belonging to $\mathcal{Q}$. This subfamily will be called a *smooth path, passing through* P, if the following assumptions hold:

1. The total variation $\| P_u \| = o(| u |^2)$, as $u \to 0$.
2. The map $\Gamma : u \mapsto [dP_u/dP]^{1/2}$, from U into the space $\mathbf{L}^2(P)$, is Fréchet–differentiable at $u = 0$, i.e., there exists $h \in \mathbf{L}^2(P)$, such that $\lim_{u \to 0} u^{-2} \int \left[(dP_u \,/\, dP)^{1/2}(x) - (1 + u\,/\,2\,h(x))\right]^2 dP(x) = 0$. ■

The function $h \in \mathbf{L}^2(P)$, appearing here, will be called a *tangent vector* both to $\mathcal{Q}$ and to the path $\{P_u : u \in U\}$ at $P = P_0$. The minimal closed linear subspace $\mathbf{T}(P, \mathcal{Q}) \subset \mathbf{L}^2(P)$, containing tangent vectors at P to every possible smooth path $\{P_u : u \in U\}$, passing through $P = P_0$, will be called a *tangent space* to $\mathcal{Q}$ at P.

2.1.2 Coparametrization.

The estimation in nonparametric or semiparametric models often presumes an infinite–dimensional parameter, so that parametric subfamilies of any finite dimension can surround P inside $\mathcal{Q}$. The following definition introduces differentiable coparameters.

Definition. The map $\Phi : \mathcal{Q} \to \mathbf{B}$, from the family $\mathcal{Q}$ of probabilities into a Banach space $\mathbf{B}$, is called a *differentiable coparametrization* at P (and the value $\Phi(P)$ is called a differentiable coparameter), if there exists a linear and continuous operator $\mathrm{D}_P\,\Phi : \mathbf{T}(P, \mathcal{Q}) \to \mathbf{B}$, a derivative of Φ at P, such that whatever smooth path $\{P_u : u \in U\} \subset \mathcal{Q}$ through $P = P_0$ is chosen, the composition $u \mapsto P_u \mapsto \Phi(P_u)$ is differentiable at the point zero, and the derivative of this composition may be calculated via the chain rule, i.e.:

$$\frac{d}{du}\,\Phi(P_u)\,|_{u=0} = \mathrm{D}_P\,\Phi\,[h].$$

Here $h \in \mathbf{T}(P, \mathcal{Q})$ is a tangent vector to a chosen path at P. ■

The definition introduces a coparametrization, differentiable at P. If it holds at every point $P \in \mathcal{Q}$, it is called simply differentiable. This paper considers differentiable coparametrizations only, so the word "differentiable" will often be omitted. Linear functionals, arising as derivatives of real–valued coparameters, can be represented by more common P–square integrable functions. That is why the following notion of slope or score is introduced. It is known as the influence function in robustness theory.

2.1.3 Canonic gradients.

Consider a real–valued coparameter $\Phi(P)$. Its derivative $\mathrm{D}_P\,\Phi$ is a linear continuous functional on the space $\mathbf{T}(P,\mathcal{Q})$. If we assume that every element of this space can be approximated by tangent vectors (rather than by their linear combinations), corresponding to smooth paths through P, then the value $\mathrm{D}_P\,\Phi(P)\,[h]$ is uniquely defined at each $h \in \mathbf{T}(P,\mathcal{Q})$. Therefore, due to the fact that a closed linear subspace of the Hilbert space $\mathbf{L}^2(P)$ will itself be Hilbert (with respect to the same inner product), there may exist more than one function $\phi \in \mathbf{L}^2(P)$, representing the functional $\mathrm{D}_P\,\Psi$ in terms of the common inner product in the space $\mathbf{L}^2(P)$, i.e. such that $\mathrm{D}_P\,\Phi\,[h] = \int \phi(x)\cdot h(x)\,dP(x)$. First, the fact that each $h \in \mathbf{T}(P,\mathcal{Q})$ is orthogonal to the function $\mathbf{1}$ (its value $\mathbf{1}(x)$ is identically equal to 1), implies that there exist more than one function ϕ, representing the linear functional. Nevertheless, the "canonic" version of this function, or so called "canonic gradient" exists and is uniquely defined as a function $\phi_P = \phi(P,\cdot)$ belonging to $\mathbf{T}(P,\mathcal{Q})$ and representing the derivative of a coparameter $\Phi(P)$ at $P \in \mathcal{Q}$.

Remark: The notion of a canonic gradient was introduced in Koshevnik and Levit (1976). Both its existence and uniqueness are implied by the fact that a functional $\mathrm{D}_P\,\Phi$ is linear and continuous on the Hilbert space $\mathbf{T}(P,\mathcal{Q})$, and hence it admits the above described representation due to the Riesz - Fischer theorem. Canonic gradients have been used and discussed by Koshevnik and Levit (1980), Pfanzagl (1982), Millar (1983), Begun, et al. (1983), Wellner (1985), van der Vaart (1989, 1991).

2.1.4 Canonic kernels.

A similar definition, introducing the canonic kernel of a Banach space valued coparameter can be given. It provides a convenient representation for the linear operator $\mathrm{D}_P\,\Phi : \mathbf{T}(P,\mathcal{Q}) \to \mathbf{B}$. For any differentiable coparameter $\Phi(P) \in \mathbf{B}$ and any functional β from the dual space $\mathbf{B}^*$, consider a real valued coparameter $\langle\,\Phi(P),\beta\,\rangle$, and let $K_P\,(\cdot;\beta)$ denote its canonic gradient.

Definition. A function $K_P : \mathbf{X}\times\mathbf{B}^* \to \mathbf{R}$ is called a canonic kernel representing the derivative $\mathrm{D}_P\,\Phi$ of Φ at P. ■

Remark. The following characterization can be easily given for the canonic kernel.

1. As a function of its first argument, $K_P\,(\,\cdot,\beta) \in \mathbf{T}(P,\mathcal{Q})$, for every $\beta \in \mathbf{B}^*$.

2. As a function of its second argument, K_P is linear and a continuous map from $\mathbf{B}^*$ into $\mathbf{T}(P, \mathcal{Q})$.

The definition of canonic kernel implies both of the properties, provided the canonic gradients exist and Ψ is a differentiable coparameter.

Brownian bridges.

The limiting properties of asymptotically efficient estimates of any coparameter can be described in terms of the family of abstract Brownian bridges.

Definition. Let $W_P = \{W_P[h] : h \in \mathbf{L}^2(P)\}$ be a set of Gaussian zero mean random variables. We call it a Brownian bridge (with respect to P) or a P–Brownian bridge, if the variance–covariance structure is given by

$$\mathbf{E}\,(W_P[h_1]W_P[h_2]) = \langle h_1, h_2 \rangle_P - \int h_1\,dP \cdot \int h_2\,dP.$$

Remark. It can be seen that in the particular case when $\mathbf{X}$ is the unit interval and P is the uniform distribution on $\mathbf{X}$, the common Brownian bridge, say $w(\cdot)$, will satisfy this definition, provided a collection of random variables $(W[h] : h \in \mathbf{L}^2[0,1])$ is defined by means of the stochastic integral

$$W_P[h] = \int h(x)\,dw(x).$$

Recall that w is a zero mean Gaussian process on the unit interval, having continuous sample paths and covariance function $\mathbf{E}\,[w(t) \cdot w(s)] = \min\,(t,s) - t \cdot s$, and the Wiener integral is used. This Brownian bridge $w(\cdot)$ appears as a limiting process for the empirical cumulative distribution function (*CDF*), based on the sample from the uniform distribution on the unit interval. Meanwhile, its abstract analogue W_P plays the similar role under general circumstances.

2.2 Information inequalities.

The lower bounds of risks for estimation of a quite general infinite–dimensional coparameter $\Psi(P) \in \mathbf{B}$ can be formulated in terms of the canonic kernels and the family of Brownian bridges. Suppose the canonic kernel for Ψ at $P \in \mathcal{Q}$ is equal to $\{K_P(\cdot, \beta) \in \mathbf{T}(P, \mathcal{Q}) : \beta \in \mathbf{B}^*\}$. Consider the cylindric Gaussian measure $\mathbf{M}_P$ on the space $\mathbf{B}$, having zero mean and covariance structure

$$\int \langle b, \beta_1 \rangle \cdot \langle b, \beta_2 \rangle\, d\mathbf{M}_P(b) = \int K_P(x, \beta_1) \cdot K_P(x, \beta_2)\, dP(x).$$

Due to the fact that for every canonic gradient

$$\int K_P(x, \beta)\, dP(x) = 0$$

and using the adjoint operators

$$\mathrm{D}_P^* \Phi : \mathbf{B}^* \to \mathbf{T}(P, \mathcal{Q}),$$

we can define the symmetric bilinear forms, say $\{\mathbf{J}_P : P \in \mathcal{Q}\}$ on the dual space $\mathbf{B}^*$: $\mathbf{J}_P(\beta_1, \beta_2) = \langle K_P(\cdot, \beta_1), K_P(\cdot, \beta_2)\rangle_P$. Then $\mathbf{M}_P$ is a cylindric Gaussian measure with a zero mean and covariance $\mathbf{J}_P$. The key assumption we need to add, is that each of these cylindric Gaussian measures on $\mathbf{B}$ may be extended to a Borel probability on $\mathbf{B}$. Another assumption that we make, the so–called "extensiveness property" of subsets $U \subset \mathcal{Q}$, that will be called neighborhoods in $\mathcal{Q}$. The class $\mathcal{U}$ of sets, which we will consider later, will contain the extensive sets only. They may be considered as neighborhoods with respect to a suitable topology $\mathcal{T}$ on $\mathcal{Q}$. As a matter of fact, the assumptions on $\mathcal{T}$ are not too restrictive. For instance, the main results of Koshevnik and Levit (1976), Pfanzagl (1982), and Koshevnik (1984c) remain valid for various topologies on the given family of probabilities, i.e. are "topology free", provided $\mathcal{T}$ satisfies sufficiently mild requirements.

The extensiveness property can be roughly interpreted as approximability of every linear combination of tangent vectors by a sequence of tangent vectors, corresponding to smooth paths in $\mathcal{Q}$ through $P = P_0$. More precisely, for each $P \in \mathcal{Q}$, every natural number m, any collection $\{h_1, h_2, \ldots, h_m\} \subset \mathbf{T}(P, \mathcal{Q})$ and every $\epsilon > 0$, there exists a smooth m-parameter subfamily $\{P_{u,\epsilon} = P_u : u \in U\}$, indexed by a neighborhood $U \equiv U_\epsilon \subset \mathbf{R}^m$, with the tangent vectors $(h_{j,\epsilon} = g_j : j = 1, \ldots, m)$ approximating the given set of tangent vectors with the error less than ϵ, i.e.

$$\max_{1 \le j \le m} \left[\int (g_j(x) - h_j(x))^2 \, dP(x) \right]^{1/2} < \epsilon.$$

This condition from Koshevnik and Levit (1980), generally means that any finite–dimensional subspace in $\mathbf{T}(P, \mathcal{Q})$ can be approximated by a tangent space to a suitable smooth m-dimensional parametric subfamily through P.

Information inequalities. The lower bounds of risks, derived by Koshevnik and Levit (1980) for a quite general restricted problem, are described by the inequality:

$$\liminf_{n\to\infty} \sup_{P \in U} \mathbf{E}_P \, L \left[n^{1/2} \left(\tilde{\Phi}^n - \Phi(P) \right) \right] \ge \sup_{P \in U} \int L(b) \, d\mathbf{M}_P(b).$$

Here L is any real–valued, symmetric about zero, non–negative, semiconvex, and low semicontinuous loss function on the decision space (real line $\mathbf{R}$, or finite–dimensional vector space $\mathbf{R}^m$ or a certain Banach space $\mathbf{B}$). In other words, once the decision space $\mathbf{B}$ is chosen, the function L is assumed to be non–negative, $L(0) = 0$, and all the sets

$$\{ b \in \mathbf{B} : L(b) \le a \}$$

are convex and closed, for every positive a.

Remark: Note that each of the Gaussian measures $\{\mathrm{M}_P : P \in \mathcal{Q}\}$, introduced at the beginning of this section, is simply the distribution $\mathcal{L}(B_P)$ of $\mathbf{B}$-valued random element B_P, such that for every $\beta \in \mathbf{B}^*$ the equality $\langle B_P, \beta \rangle = \int K_P(x, \beta)\, dW_P(x)$, is valid.

2.3 Asymptotic efficiency and uniform weak convergence.

It is clear that once the lower bounds are described in terms of a family $\{B_P : P \in \mathcal{Q}\}$ of $\mathbf{B}$–valued Gaussian random variables and there exists a sequence of $\mathbf{B}$-valued estimates, say $\{\hat{\Psi}^n : n \geq 1\}$, such that weak convergence of $n^{1/2}\left[\hat{\Psi} - \Psi(P)\right]$ to random elements B_P holds, as $n \to \infty$, the sequence may happen to be asymptotically efficient. This occurs if, for instance, the weak convergence holds uniformly in $P \in U$, where $U \subset \mathcal{Q}$ denotes a "small neighborhood" of a distribution P. Under these assumptions, the asymptotic bounds are automatically achieved for all continuous and bounded loss functions ($L \in \mathbf{CB}$) simultaneously, by the same estimator–sequence.

Remark. Later we will be required to replace the space $\mathbf{B}$ by a richer one, say $\mathbf{D} \supset \mathbf{B}$. In the meanwhile, a sequence of $\mathbf{D}$ - valued random elements $B_P^n = n^{1/2}\left[\tilde{\Psi} - \Psi(P)\right]$ will converge weakly to $\mathbf{B}$–valued Gaussian random elements, as $n \to \infty$, and uniformity holds in $P \in U$ for a wide enough class $\mathcal{U}$ of "neighborhoods" $U \in \mathcal{Q}$. ■

The loss functions considered here will be from $\mathbf{CB}$. As far as more general loss functions are concerned, a sequence of functions from $\mathbf{CB}$, increasingly converging to the given one, as Millar (1983) indicated, usually can be constructed. The standard techniques may then apply to derive the achievability of the asymptotic bounds (given by the information inequality), even for a wider class of loss functions. Sometimes, the estimates themselves should be either trimmed or subjected to another kind of a "fine tuning", but these techniques seem to be less important here.

Therefore we will focus our attempts on the uniform weak convergence and show how it can be established. On the other hand, Bickel (1984) notes the importance of uniform weak convergence. It is quite natural to interpret asymptotic efficiency as the weak convergence to the Gaussian random elements B_P (the same as in the Information Inequality) uniformly in $P \in U$. Here U runs over a class $\mathcal{U}$ of subsets in $\mathcal{Q}$. Some assumptions on $\mathcal{U}$ will be formulated later. Typically the subsets U are chosen as suitable neighborhoods of $\mathcal{Q}$. They can be described in terms of coparameters of interest and restrictions in each concrete problem.

2.3.1 How are information bounds affected by restrictions?

The question in the title admits an equivalent formulation. *How do the canonic kernels and gradients change due to restrictions?* The following lemma helps to provide the general answer.

Lemma. Let $\mathbf{H}$ be a Hilbert space, $\mathbf{H}_1$ its closed linear subspace, $\mathbf{B}$ a Banach space. Suppose a linear and continuous operator $A : \mathbf{H} \to \mathbf{B}$ is given, and let $A_1 : \mathbf{H}_1 \to \mathbf{B}$ denote its restriction on the subspace $\mathbf{H}_1$. Then the adjoint operator A_1^* to A_1 can be expressed as

$$A_1^*\beta = \Pi(A^*\beta)$$

for every $\beta \in \mathbf{B}^*$, where Π is the orthogonal projector from $\mathbf{H}$ onto $\mathbf{H}_1$.

Proof. It is obvious from the definitions. ■

In restricted problems, when the real-valued coparameter Φ is to be estimated, first for $P \in \mathcal{Q}$, and then using additional information $P \in \mathcal{P}$, the relations between tangent spaces $\mathbf{T}(P, \mathcal{Q})$ and $\mathbf{T}(P, \mathcal{P})$ are easy to describe. Namely, suppose $\mathcal{P}$ is defined in terms of a coparameter $\Psi(P) \in \mathbf{B}$, so that $\mathcal{P} \equiv \{P \in \mathcal{Q} : \Psi(P) = 0\}$. Then the tangent space to $\mathcal{P}$ coincides with the orthogonal complement to the image of the adjoint operator $(\mathrm{D}_P\Psi)^*$. Equivalently, the tangent space $\mathbf{T}(P, \mathcal{P})$ is the orthogonal complement in $\mathbf{T}(P, \mathcal{Q})$ to the collection of all canonic gradients $\{K(\cdot, \beta) : \beta \in \mathbf{B}^*\}$. We use the following notation for orthogonal complementation

$$\mathbf{T}(P, \mathcal{Q}) \ominus (\mathrm{D}_P\Psi)^*(\mathbf{B}^*) = \mathbf{T}(P, \mathcal{P}).$$

Hence in the previous Lemma, setting $\mathbf{H}$ and $\mathbf{H}_1$ equal, respectively, to the tangent spaces to $\mathcal{Q}$ and $\mathcal{P}$ at the same $P \in \mathcal{P}$, the canonic gradient for the restricted problem is obtained as a result of the orthogonal projection of the initial gradient, onto the space, "complementary to restrictions".

2.3.2 Some heuristic techniques of efficient estimation.

Considering restricted estimation problems, one may find the following device attractive. The above construction under a finite set of affine restrictions may be extended and applied to more general situations. Suppose the restrictions are defined by means of a $\mathbf{B}$-valued coparameter $\Psi(P)$. For finite-dimensional restrictions, with $\mathbf{B} = \mathbf{R}^m$, and $\Psi \equiv (\Psi_1, \ldots, \Psi_m)$, it is easy to represent the above considered estimator as

$$\hat{\hat{\Phi}} = \tilde{\Phi} - \langle \Psi, \beta \rangle,$$

with $\beta \in (\mathbf{R}^m)^*$. The space $\mathbf{B} = \mathbf{R}^m$ coincides with its dual, but it is convenient to treat both of them differently, taking into account that generally, under infinite-dimensional restrictions, the spaces $\mathbf{B}$ and $\mathbf{B}^*$ will no longer be the same. The equation we have already discussed for $\beta \in \mathbf{R}^m$, can be written as

$$(\mathrm{D}_P\,\Psi)\,(\mathrm{D}_P\,\Psi)^*\,[\beta] = (D_P\,\Psi)\,[\phi_P].$$

Though there is a difference between the infinite-dimensional and finite-dimensional restrictions that may cause additional difficulties, we will attempt to apply the same construction in the general case. The weak convergence results for a transformed estimator–sequence $\hat{\Phi}$ with the "ideal " (or optimal) functional $\beta \in \mathbf{B}^*$ replaced by its approximation, will be proved later as well as the desired asymptotic efficiency.

We need to mention two "exceptions" in this regard. The estimates in both of these cases are very well known and do not actually require the geometric approach. Their limiting behavior can also be derived easily. However, the knowledge of these examples will help us to realize the efficient estimates considered in Section 4.

2.3.3 Example: Symmetry about zero.

Suppose the distribution P of a real valued sample $(X_i : i = 1, \ldots, n)$ is known to be symmetric about zero and continuous. In other words, the cumulative distribution function F satisfies the equations:

$$F(-t) \equiv 1 - F(t),$$

for every $t \in \mathbf{R}$. Then for the family $\mathcal{P}$ of all symmetric about zero distributions from $\mathcal{Q}$, where $\mathcal{Q}$ is the family of all continuous probability distributions on $\mathbf{R}$, the efficient estimate of a real valued functional $\Phi(P)$ may be obtained by substituting the symmetrized empirical CDF

$$\hat{F}(t) \equiv \frac{1}{2}\left[\tilde{F}(t) + 1 - \tilde{F}(-t)\right],$$

instead of the usual empirical CDF $\tilde{F}(\cdot)$. Hence

$$\hat{\Phi} = \frac{1}{2n}\sum_{i=1}^{n}\left[\phi(X_i) + \phi(-X_i)\right]$$

will be the asymptotically efficient estimate of $\Phi(P) = \int \phi(x)\, dP(x)$ in this case (Koshevnik and Levit (1980), Koshevnik (1984c) and Pfanzagl (1982)).

2.3.4 Example: Independent components.

Another occasionally easy construction may be found when the bivariate random observation X has independent components. In terms of a joint CDF $F(\cdot\,,\cdot)$ of the components $(X(1), X(2))$, the restriction can be written as

$$F(t,s) \equiv F(t,\infty) \cdot F(\infty, s).$$

The efficient estimation of the functional $\Phi(P) = \int \phi(x)\, dP(x)$ can be found by means of the direct substitution of the product of marginal empirical CDFs into Φ, i.e. $\hat{\Phi} = n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi(X_i(1), X_j(2))$. See also Pfanzagl (1982), and Koshevnik (1982, 1984c).

2.3.5 What can be done generally?

Generally, we can hardly hope to avoid all sophisticated procedures, based on several approximation theorems. That is a motivation for the general approach developed here. Let us consider an arbitrary restricted estimation problem. Suppose that for a real valued coparameter $\Phi(P)$ and all real valued coparameters $\{\langle\Psi(P),\beta\rangle : \beta \in \mathbf{B}^*\}$ there can be constructed estimates, say $\tilde{\Phi}$ and $\{\langle\tilde{\Psi},\beta\rangle : \beta \in \mathbf{B}^*\}$, in other words, there exists an estimate $\left(\tilde{\Phi}(P), \tilde{\Psi}(P)\right) \in \mathbf{R} \oplus \mathbf{D}$, that is asymptotically well behaved.

Remark. In the case of a multivariate sample, with restrictions expressed via the CDF F, $\mathbf{B}$ and $\mathbf{D}$ are usually chosen as the spaces of all continuous and all bounded Borel measurable functions respectively, defined on the extended Euclidean space $[-\infty,\infty]^m$. Uniformity in the Central Limit Theorem for the empirical *CDF* is derived in Koshevnik (1982, 1984c). ■

We assume that the estimates $\tilde{\Phi}$ and $\tilde{\Psi}$ satisfy the usual joint asymptotic normality conditions. Equivalently, we assume weak convergence of the joint distributions

$$n^{1/2}\left[\left(\tilde{\Phi}^n - \Phi(P)\right), \left(\tilde{\Psi}^n - \Psi(P)\right)\right] \xrightarrow{\mathcal{D}} \left[\int \phi(P,x)\,dW_P(x),\ B_P\right],$$

as $n \to \infty$. If this convergence holds uniformly in $P \in U$, for a rich enough class $\mathcal{U}$ of "neighborhoods" $U \subset \mathcal{Q}$, then it will be inherited by neighborhoods $V = U \cap \mathcal{P}$ in the restricted family.

The above equations define the ideal functionals $\beta = \beta(P)$. Their estimates, say $\beta^n = \beta^n(X^n)$ will be used, just as in the case of a finite set of affine restrictions. It is unclear, what is to be done when the solution $\beta(P)$ fails to exist. Some possible answers were proposed by Koshevnik (1985). Other suggestions, applicable to more general cases, were put forth in Koshevnik (1981, 1984a,b). This paper extends the above mentioned results to cover more complicated situations. The new theorems may apply to both the nonparametric and semiparametric versions of mixture models and bias selected models. They can be also thought of as an alternative to methods proposed by Bickel and Ritov (1990), Schick (1986, 1987), Bickel, Ritov, and Wellner (1991), because a restricted family of probabilities is considered here in terms of its codimension rather than its dimension. Meanwhile, often the theorems work even for infinite–dimensional (or nonparametric) families. These applications are considered in detail in a paper in preparation. Examples shown in the previous section are simple and illustrate how the main results apply to more common and well developed models.

Chapter 3

Uniform weak convergence theorems.

The uniform weak convergence results for the proposed estimates of real–valued coparameter $\Phi(P)$ are formulated in this section. Their proofs, based on theorems from Pollard (1989) and Koshevnik (1982, 1984a), are outlined in the last section. The restriction $\Psi(P) = 0$ is invoked to improve the initial estimator–sequence $\tilde{\Phi}$. The main idea is similar to the standard tool in unbiased estimation. The minimum variance estimate is constructed there as a sum of the initial one and a suitably chosen unbiased estimate of zero. In the case under consideration, a linear functional $\langle \tilde{\Psi}(P) , \beta \rangle$, for each $\beta \in \mathbf{B}^*$, will estimate zero asymptotically unbiasedly, provided the true distribution P belongs to the subfamily $\mathcal{P}$.

It will be convenient to slightly simplify the notation. For a given pair of families $\mathcal{Q}$ and $\mathcal{P}$ of probabilities on $(\mathbf{X}, \mathcal{A})$, the canonic gradients of real–valued coparameters $\Phi(P)$ and $\langle \Psi(P) , \beta \rangle$ are denoted as $\phi(P, \cdot) \in \mathbf{T}(P, \mathcal{Q})$ and $\psi(P, \cdot\,; \beta) \in \mathbf{T}(P, \mathcal{Q})$ respectively. The similar canonic gradients, calculated under the assumption $P \in \mathcal{P}$, will be denoted as $\phi_1 (P, \cdot) \in \mathbf{T}(P, \mathcal{P})$ and $\psi_1(P, \cdot\,; \beta) \in \mathbf{T}(P, \mathcal{P})$. The relations between ϕ and ϕ_1 may be also expressed in terms of the family $\{\mathbf{\Pi}_P : P \in \mathcal{P}\}$ of orthogonal projectors, each $\mathbf{\Pi}_P$ mapping the space $\mathbf{L}^2(P)$ onto its subspace

$$\mathbf{T}(P, \mathcal{P}) = \mathbf{T}(P, \mathcal{Q}) \ominus \{\psi(P, \cdot, \beta) : \beta \in \mathbf{B}^*\},$$

orthogonal to canonic gradients of coparameters determining the restrictions.

Remark. It should be noticed that if the projector $\mathbf{\Pi}_P$ onto $\mathbf{T}(P, \mathcal{P})$ is known, and for instance, does not depend on P (just as in the examples with distributions symmetric about zero or bivariate observations with independent components), general constructions are no longer needed. The useful idea (unless serious technical difficulties accompany its implementation) is to estimate the projector $\mathbf{\Pi}_P$ and then use its estimate, in the same manner as in the case of finite–dimensional affine restrictions. In fact, we will have to use this approach cautiously. The technical details are needed to guarantee that the asymptotic bounds of risk can be achieved in a quite general situation.

3.1 Regularity assumptions.

We recall and clarify the main assumptions. It may happen, for instance, that the space $\mathbf{B}$, containing all the values of a coparameter $\Psi(P)$, is not large enough. In particular, the estimates $\tilde{\Psi}$ often may fail to be random elements of the space $\mathbf{B}$, where Ψ takes its values. That is the reason why we need to introduce another space, $\mathbf{D} \supset \mathbf{B}$, such that all the estimates $\tilde{\Psi}$ are random elements of $\mathbf{D}$.

3.1.1 Assumptions concerning Banach spaces.

First, we assume that $\mathbf{D} \subset \mathbf{B}^{**}$, or equivalently, that every $b \in \mathbf{D}$ may be interpreted as a functional on $\mathbf{B}^*$, bounded on norm bounded sets in $\mathbf{B}^*$. The same notation $\langle b, \beta \rangle$ is used for both $b \in \mathbf{B}$ and $b \in \mathbf{D}$. In the second case it relates to $b(\beta)$ rather than $\beta(b)$ in the first one.

The unit ball S^* of the space $\mathbf{B}^*$ can be equipped with the topology of $*$–weak convergence and is known to be compact. (This topology is generated by all functionals $\{\beta \mapsto \langle b, \beta \rangle : b \in \mathbf{B}\}$.) Due to the fact that $\mathbf{B}$ is separable, S^* will be metrizable compact. Any element $b \in \mathbf{B}$ can be identified with the linear functional on $\mathbf{B}^*$, continuous on the compact S^*. The elements of $\mathbf{D}$ can be identified with bounded linear functionals on S^*, their continuity may be lost because S^* is equipped with a topology, different from its original one (induced by the norm of $\mathbf{B}^*$). This construction eliminates common difficulties, appearing when abstract Banach spaces (like $\mathbf{B}$ or $\mathbf{D}$) are involved to prove weak convergence of empirical processes, treated as random elements of a suitable Banach space. See also Pollard (1989).

3.1.2 Neighborhoods.

We now describe the sets $U \subset \mathcal{Q}$, appropriate as neighborhoods. The intersections $V = U \cap \mathcal{P}$, when U runs over a class $\mathcal{U}$, will be treated as a neighborhood in $\mathcal{P}$ and will inherit these important properties of U. We assume that the neighborhoods (in both families) have the extensiveness property and the corresponding sets $\{\Psi(P) : P \in U\}$ are precompact (or totally bounded) in $\mathbf{B}$. The "typical" neighborhood is supposed to satisfy another natural requirement, namely, the tangent space $\mathbf{T}(P, U)$ should coincide with a tangent space to the whole family $\mathcal{Q}$ at P. The same assumption is imposed on V and $\mathcal{P}$, respectively. Some additional assumptions on U are typically motivated by the behavior of canonic gradients. For instance, we need to assume the canonic gradient $\phi(P, \cdot)$ is uniformly square integrable with respect to $P \in U$, i.e.

$$\lim_{a \to \infty} \sup_{P \in U} \int \mathbf{1}\{|\phi(P, x)| > a\} = 0.$$

3.1.3 The limiting Gaussian measures.

The assumption introduced here concerns a family of Gaussian probabilities (rather than cylindric measures) $\{\mathbf{M}_P : P \in U\}$ corresponding to $\mathbf{B}$–valued random elements $\{B_P : P \in \mathcal{P}\}$. Random elements B_P of the space $\mathbf{B}$ are interpreted, with attention being focused on finite–dimensional distributions only, as stochastic integrals $\langle B_P \, , \beta \rangle \, = \, \int \psi(P, x; \beta) \, dW_P \, (x)$. Here W_P is the P-Brownian bridge indexed by $\mathbf{L}^2(P)$, and the integrand is the canonic kernel of Ψ at $P \in \mathcal{Q}$, while $\beta \in \mathbf{B}^*$ is an arbitrary linear functional on $\mathbf{B}$. This family of Gaussian measures on $\mathbf{B}$ is also supposed to be locally uniformly tight, i.e. when P runs over a small enough neighborhood U,

$$\forall \epsilon \, > \, 0 \, \exists \text{ a compact set } K \subset \mathbf{B} \text{ such that } \mathrm{M}_P \, (K) \, \geq \, 1 \, - \, \epsilon \, , \, \forall P \in U.$$

This assumption restricts the class $\mathcal{U}$ of possible neighborhoods.

Remark. The uniform tightness assumption for the family $\mathcal{M}(U) \, = \, \{ \, \mathrm{M}_P \, : \, P \in U \, \}$ of limiting probabilities on $\mathbf{B}$ is to ensure that the estimates of $\Phi(P)$ constructed by means of β^n "adapt" themselves to the unknown value of $\beta(P)$. In other words, if the value $\beta(P)$ were known, we would have adjusted the initial estimate, otherwise we have to replace this true value by its estimate. Several examples in Section 4 will illustrate this principle.

3.2 Auxiliary results on uniform weak convergence.

The necessary and sufficient conditions of uniform weak convergence, formulated in Koshevnik (1982, 1984c), are reviewed here. They will be invoked later to show that the asymptotic efficiency of the proposed estimates $\hat{\Phi}$ is implied by the limiting behavior of the initial estimator–sequences $\tilde{\Psi}$ and $\tilde{\Phi}$.

3.2.1 Weak convergence in nonseparable Banach spaces.

Let $\mathbf{D}$ be a Banach space and $\mathbf{B}$ its separable subspace. (If $\mathbf{D}$ is separable, $\mathbf{B}$ and $\mathbf{D}$ may be chosen equal, this construction being needed to avoid some difficulties appearing otherwise.) Let $\mathcal{D}$ denote a σ–field of subsets $A \subset \mathbf{D}$, satisfying two assumptions. These assumptions make sense due to the following definition (see Pollard (1989) for details) of the weak convergence $\mathrm{M}^n \Longrightarrow \mathrm{M}$ that covers cases of either separable or nonseparable Banach spaces $\mathbf{D}$.

1. The σ–field of all sets $\{A \cap \mathbf{B} \, : \, A \in \mathcal{D}\}$, called a trace of $\mathcal{D}$ on $\mathbf{B}$, coincides with the Borel σ–field $\mathcal{B}$ of subsets $\mathbf{B}$.

2. Any open ball (with respect to the norm topology) in $\mathbf{D}$ is $\mathcal{D}$–measurable.

Definition. Suppose that for every $\mathbf{CB}$ and $\mathcal{D}$-measurable function G on $\mathbf{D}$, the convergence of integrals

$$\int G(b) \, dM^n \, (b) \, \rightarrow \, \int G(b) \, dM \, (b)$$

holds, as $n \rightarrow \infty$. Then we say that probabilities $\{ \, M^n \, \}$ weakly converge to M. ■

The space of functions G appearing in this definition, will be also denoted as $\mathbf{CB}$.

Definition: Uniform weak convergence. Let both M^n and M be indexed by a subscript, say $v \in V$. Suppose that the weak convergence assumption holds uniformly in $v \in V$, i.e.

$$\lim_{n\to\infty} \sup_{v\in V} \left| \int G(b)\, d\mathbf{M}_v^n(b) - \int G(b)\, d\mathbf{M}_v(b) \right| = 0, \; \forall G \in \mathbf{CB}(D).$$

Then the weak convergence is called uniform (in $v \in V$). ■

Asymptotic uniform tightness. Suppose that weak convergence of all "scalar" distributions

$$\mathcal{L}\left(\langle B_v^n, \beta \rangle\right) \Longrightarrow \mathcal{L}\left(\langle B_v, \beta \rangle\right),$$

as $n \to \infty$, holds uniformly in $v \in V$. If a limiting family is uniformly tight, the necessary and sufficient condition for uniform weak convergence is given by the Proposition (proved in Koshevnik (1982)).

Proposition. The following condition is necessary and sufficient for the uniform weak convergence $B_v^n \xrightarrow{\mathcal{D}} B_v$, provided the scalar distributions converge uniformly.

For every $\epsilon > 0$, there exist a compact set $K \subset \mathbf{B}$ and a number n_0 such that

$$\sup_{v\in V} \mathbf{Prob}\,\{ d(B_v^n, K) \geq \epsilon \} < \epsilon, \quad \text{as soon as } n > n_0. \blacksquare$$

Here, for every set $K \subset \mathbf{D}$ and every element $b \in \mathbf{D}$, the distance $d(b, K)$ is defined as $\inf_{d\in K} \| b - d \|$. Simple examples from (Koshevnik (1982, 1984c) show that the uniform tightness of the limiting family is important and cannot be removed entirely.

3.3 Main results.

The following theorems can now be formulated.

Asymptotically efficient estimation under firm assumptions.

The asymptotic distribution result for the proposed estimator–sequence is formulated under assumptions that may seem to be too strict. They will be relaxed later. These assumptions hold in both Examples 2.5.1 and 2.5.2, due to the fact that $\beta(P)$ there can be chosen "distribution free". Moreover, they hold for "well behaved" coparameters, though Theorem 2 obtains the same asymptotic distribution for a wider class of coparameters.

Theorem 1. Let $\tilde{\Psi}$ be a **D**–valued estimator–sequence for a **B**–valued coparameter $\Psi(P)$ and $\tilde{\Phi}$ be an estimator–sequence for $\Phi(P) \in \mathbf{R}$. Suppose a neighborhood U in $\mathcal{Q}$ satisfies the following set of assumptions.

1. Weak convergence

$$n^{1/2}\left[\left(\tilde{\Psi} - \Psi(P)\right), \left(\tilde{\Phi} - \Phi(P)\right)\right] \xrightarrow{\mathcal{D}} \left[\int \phi(P,x)\, dW_P(x)\,,\, B_P\right]$$

   holds uniformly in $P \in U$.

2. The limiting family $\{ \mathrm{M}_P = \mathcal{L}(B_P) : P \in U \}$ is uniformly tight.

3. The family of functionals $\{ \beta(P) \,:\, P \in U \}$ has uniformly bounded norm, i.e.

$$\sup_{P \in U} \| \beta(P) \| < \infty.$$

4. There exists a sequence $\{ \beta^n = \beta^n(X^n) \}$ of estimates for $\beta(P)$ such that

   (a) for every $b \in \mathbf{B}$ the convergence $\langle b\,, \beta^n \rangle \to \langle b\,, \beta(P) \rangle$ in P-probability holds uniformly in $P \in U$;

   (b) for every $\epsilon > 0$ there exists a positive number M and $n_0 = n_0(\epsilon)$ such that

$$P\,\{ \| \beta^n \| > M \} < \epsilon \text{ for all } P \in U \text{ and for every } n \geq n_0.$$

Then for the estimator–sequence

$$\hat{\Phi} = \tilde{\Phi} - \langle \tilde{\Psi}\,, \beta^n \rangle$$

weak convergence

$$n^{1/2}\left[\hat{\Phi}^n - \Phi(P)\right] \xrightarrow{\mathcal{D}} \int \phi_1(P,x)\, dW_P(x)$$

holds uniformly in $P \in U \cap \mathcal{P}$, as $n \to \infty$.

3.3.1 The relaxed regularity assumptions.

Various examples, exploiting trimmed statistics, smoothed empirical distributions and even nonparametric density estimates, to construct the asymptotically efficient estimator for a real valued functional, are known due to Beran (1974), Sacks (1975), Stone (1975). These successful attempts have essentially inspired the development of semiparametric theory, outlined by Bickel (1982) and Wellner (1985). A slightly different explanation for some of these results is given below. Possible extensions will be discussed later.

Theorem 2. Suppose that the estimator–sequences $\tilde{\Psi} \in \mathbf{D}$, $\tilde{\Phi}$ for coparameters $\Psi(P) \in \mathbf{D}$, $\Phi(P) \in \mathbf{R}$ and a neighborhood $U \subset \mathcal{Q}$ are chosen such that the Assumptions (1) and (2) of Theorem 1 are satisfied. Further assume

3′. There exists a sequence $\{\beta_m(P) \in \mathbf{B}^*\}$, such that for every m the sets $\{\beta_m(P) : P \in U\}$ are uniformly bounded with respect to the norm in $\mathbf{B}^*$.

4′. For every m an estimator–sequence $\beta_m^n = \beta_m^n(X^n)$ for $\beta_m(P)$ exists, the assumptions 4a, 4b of Theorem 1 being satisfied uniformly in $P \in U$.

5. The functions $\xi_m(P,\cdot) = \phi(P,\cdot) - \psi(P,\cdot;\beta_m(P))$ converge to $\phi_1(P,\cdot)$ with respect to the $\mathbf{L}^2(P)$ norm, uniformly in $P \in U$.

Then a sequence $m(n) \uparrow \infty$ can be chosen, such that for the estimator–sequence $\hat{\Phi} = \tilde{\Phi} - \langle \tilde{\Psi}, \beta_m^n \rangle$ weak convergence $n^{1/2}\left[\hat{\Phi}^n - \Phi(P)\right] \xrightarrow{\mathcal{D}} \int \phi_1(P,\cdot)\, dW_P(\cdot)$ holds uniformly in $P \in U \cap \mathcal{P}$, as $n \to \infty$.

3.3.2 Approximation theorem.

The extended version of this result was used in Koshevnik (1982) in order to prove that for the suitably smoothed empirical CDF, uniform weak convergence to Brownian Bridge will remain valid.

Theorem 3. Let Φ and $\{\Phi_m : m \geq 1\}$ be real valued differentiable coparameters. Suppose the following conditions are satisfied uniformly in $P \in U$.

1. $\Phi(P) = \lim_{m\to\infty} \Phi_m(P)$.

2. The canonic gradients $\phi_m(P,\cdot) \in \mathbf{T}(P,\mathcal{P})$ converge to $\phi(P,\cdot) \in \mathbf{T}(P,\mathcal{P})$, as $m \to \infty$. (This convergence holds with respect to $\mathbf{L}^2(P)$ norm and is uniform in $P \in U$, i.e.

$$\lim_{m\to\infty} \sup_{P\in V} \int \left[\phi_m(P,x) - \phi(P,x)\right]^2 dP(x) = 0.)$$

3. For every m there exists the asymptotically efficient estimator–sequence Φ_m^n for $\Phi_m(P)$ such that

$$n^{1/2}\left[\Phi_m^n - \Phi_m(P)\right] \xrightarrow{\mathcal{D}} \int \phi_m(P,x)\, dW_P(x).$$

Then there exists a sequence $m = m(n) \uparrow \infty$, such that $n^{1/2}\left[\Phi_m^n - \Phi(P)\right] \xrightarrow{\mathcal{D}} \int \phi(P,x)\, dW_P(x)$ uniformly in $P \in V$, as $n \to \infty$.

Chapter 4

Examples.

The examples collected in this section illustrate how the methods can be applied to concrete models. Fortunately, tedious calculations for several models have been already done before and the references will be made to the corresponding papers.

4.1 Symmetric distributions with unknown center.

This example has been among the most popular illustrations of how the adaptive estimate of a location coparameter θ can be constructed under the assumption that the observations are distributed symmetrically about its value. Beran (1974), Sacks (1975), Stone (1975) provided three different versions of the desired estimate, and then the possiblity of the successful adaptive estimation is explained in many papers, e.g. Begun, et al. (1983), Bickel (1982). Wellner (1985) proposed a general construction that was refined by Schick (1986, 1987).

This example will be considered here, too. Some of the results, obtained by Stone (1975), may be used to illustrate the Theorems 1 - 3. The estimate of a distribution of a random variable $X - \theta$ symmetric about zero, which is done in Koshevnik (1984a, c), will be discussed further. The real valued coparameter in this case is represented as a functional $\Phi(F) = \int \phi(x)\, dG(x) \equiv \int \phi(x - \theta)\, dF(x)$ of the *CDF* $F(x) = G(x - \theta)$.

4.1.1 Estimation of center.

Assume that real valued observations $(X_1, \ldots, X_n)$ have a common CDF $F(x) = G(x - \theta)$, where θ is unknown and G satisfies a symmetry condition, just as in the Example 2.5.1. Estimates for the coparameter of interest θ, constructed by Stone (1975), can be more easily described step by step as follows.

1. Suppose that θ is known. Then replacing every observation X_i by $X_i - \theta$ and using the construction from the Example 2.5.1, the estimate of $G(t)$ at every point $t \in \mathbf{R}$ can be constructed.

2. If G is known and satisfies the Rao - Cramér regularity conditions, and $\tilde{\theta}$ is any $n^{1/2}$-consistent estimate of θ, then the well known Newton - Raphson method will provide the improved estimate $\hat{\theta}$, asymptotically efficient for θ under given fixed G.

3. When G is symmetric about zero and $\tilde{\theta}$ is any $n^{1/2}$-consistent estimate of θ, the symmetrization of the empirical *CDF* based on $\{ X_i - \tilde{\theta} \}$ will give the asymptotically efficient estimate of G, even if it is considered as an infinite dimensional coparameter, as in Koshevnik (1984a, c).

4. It can be shown from Koshevnik (1982, 1984c), that the constructed estimate of G can be smoothed and still retain its asymptotic properties. Requiring firmer regularity assumptions (on G and its density) and using the Theorems 2 and 3, we may find that the procedure, proposed by Stone (1975), will work and provide the asymptotically efficient estimate for θ.

5. Let $J(F) = \int [(d/dx)(\log(f))(x)]^2 \, dF(x)$ and $\eta(x) = \mathrm{D}_F\theta(x) = -[(d/dx)\log(f)](x)/J(F)$ denote the Fisher information at F and the semiparametric efficient score for the functional $\theta = \theta(F)$, respectively. Then, Theorem 2 shows that weak convergence for the estimate, constructed by Stone (1975),

$$n^{1/2}\left[\hat{\theta} - \theta\right] \xrightarrow{\mathcal{D}} \mathrm{N}[0, (J(F))^{-1}]$$

holds uniformly in $F \in U$, provided the asymptotic expansion

$$\hat{\theta} = \theta + 1/n \sum_{i=1}^{n} \eta(X_i) + o(n^{-1/2})$$

is valid uniformly in $F \in U$. This uniformity can be achieved by means of the additional differentability assumptions on the density function.

Similar arguments explain the asymptotic efficiency for alternative constructions from Beran (1974) and Sacks (1975).

4.1.2 Estimation of the error distribution.

To some extent, estimation of $G(t)$ at the fixed point $t \in \mathbf{R}$, as well as the entire CDF G (treated as an infinite dimensional coparameter) may be even more attractive. Let us consider the same model as before, using the "signal + noise" representation:

$$X = \theta + Y,$$

where Y is distributed according to a distribution G symmetric about zero on the real line, and $\theta = \theta(F)$ is the unknown center of F. The following result can be derived from Theorems 1 and 2 above.

Proposition. Let $\tilde{\theta}$ be any $n^{1/2}$-consistent estimate for θ and $\hat{G}(t)$ be a symmetrized empirical *CDF* based on the observations shifted by $\tilde{\theta}$, i.e. $\{X_i - \tilde{\theta} : 1 \leq i \leq n\}$. Assume also that the unknown true distribution G belongs to a family $\mathcal{G}$ of symmetric distributions having densities $g(x) = dG(x)/dx$ and such that $\lim_{a \to 0} \sup_{|h| \leq a} 1/h\, [G(x + h) - G(x) - g(x)h] = 0$ holds uniformly in $G \in U$. Then the estimate $\hat{G}(t)$ is asymptotically efficient at every $t \in \mathbf{R}$. It follows that the integrals $\int \zeta(x)\, d\hat{G}(x)$ will be asymptotically efficient estimates of $\int \zeta(x)\, dG(x)$ for every bounded and measurable function ζ. Theorem 3 suggests that this class of functions ζ can be extended.

Remark. The assumptions of the uniformity in the first order Taylor expansion for $G(\cdot)$ impose some additional requirements on the topology $\mathcal{T}$ on the family $\mathcal{F}$ of probabilities F, symmetric about unknown center. These additional requirements as indicated in Koshevnik and Levit (1980) and Pfanzagl (1982), do not affect the tangent spaces.

Proof. The Proposition follows immediately from two facts.

1. Given $\theta = \theta_0$ we can construct the symmetrized empirical CDF $\hat{G}(\cdot, \theta_0)$ based on the observations $\{X_i - \theta_0 : 1 \leq i \leq n\}$. Convergence of the estimate of $\Phi(\hat{G})$ given $\theta = \theta_0$ is uniform with respect to θ_0.

2. Convergence in probability
$$\left[n^{1/2} \left(G(t - \tilde{\theta}) - G(t - \theta)\right) - \left(G(-t - \tilde{\theta}) - G(-t - \theta)\right)\right] \to 0$$
holds uniformly in $F \in U$, by assumption.

4.2 A two sample problem with unknown location.

Consider the problem when observations $\{X_i : 1 \leq i \leq n\}$ and $\{Y_j : 1 \leq j \leq m\}$ are independent, all X_i distributed according to F, all Y_j distributed according to H, while F and H are continuous and satisfy the following system of equations:
$$F(t) \equiv H(t - \theta),$$
with unknown real coparameter $\theta = \theta(F, H)$. Just as before, both estimation of θ and F (or H) can be considered. Similar results may be obtained for estimates of θ from Beran (1974) and Sacks (1975). As far as F or H is concerned, the following "infinite dimensional reparametrization" is recommended.

Let G and γ be a distribution function and a real number, such that
$$F(t + \gamma) = G(t) = H(t - \gamma)$$
and hence G is a common *CDF* of $X - \gamma$ and $Y + \gamma$, while $\theta = 2\gamma$. Then, having found any $n^{1/2}$ - consistent estimate $\tilde{\gamma}$ for γ (or, equivalently, $\tilde{\theta} = 2\tilde{\gamma}$ for $\theta = 2\gamma$) and calculated the empirical CDF

$\hat{G}$ from all "pseudo-observations" $\{X_i - \tilde{\gamma} : 1 \leq i \leq n\}$ and $\{Y_j + \tilde{\gamma} : 1 \leq j \leq m\}$ we obtain the asymtotically efficient estimate for $G(t)$ and for functionals $\int \zeta(x)\, dG(x)$, just as in the previosly considered case of symmetry with unknown center. The same regularity assumptions on the CDF G are required for asymptotic efficiency in this case, too.

4.3 Several transformation models.

Suppose the observations $\{X_i : 1 \leq i \leq n\}$ admit the representation $X_i = \theta + H_b(Y_i)$, where b is a finite dimensional vector, indexing elements of a transformation group $\mathbf{H} = \{H_b(\cdot) = H(b, \cdot) : b \in B\}$, and θ is a real number, while unobservable random variables $\{Y_i : 1 \leq i \leq n\}$ are i.i.d. with a common distribution G, symmetric about zero. H_b will be assumed odd, i.e. $H_b(-x) = -H_b(x)$, for all $x \in \mathbf{R}$ and $b \in B$. In order to ensure identifiability of G and b, suppose that there is a coparameter, say $\beta : \mathcal{F} \to B$ defined for any distribution of X, while the distribution G is to have the value $\beta(G) = 0$. In other words, $b = \beta(F)$. The asymptotically efficient estimates are constructed for a given affine coparameter of G, such as $\Phi(G) = \int \phi(x)\, dG(x)$. Let us assume that $H_b(0) = 0$ for every b, and H_b is strictly increasing.

Remark. If b is a scale parameter, so that $X = a + bY$, we may use the median and upper quartile to suitably define a and b. The symmetry about zero does not affect the results concerning estimates of coparameters of G, while estimation of b should be done according to the definition of β. ∎

Proposition. Suppose that $n^{1/2}$–consistent estimates $\tilde{a}$ and $\tilde{b}$ for $a \in \mathbf{R}$ and $b = \beta(F) \in B$ respectively are given. For every observation X_i define $Y_i\left(\tilde{a}, \tilde{b}\right) = \left(H_{\tilde{b}}\right)^{-1}(X_i - \tilde{a})$ and use them as if they were i.i.d. observations from G. Let $\hat{G}$ denote the symmetrized empirical CDF based upon these random variables. Suppose G admits the first order Taylor approximation uniformly in $F \in U$ for a chosen neighborhood $U \subset \mathcal{F}$ in the family of all underlying distributions F. Then

1. for any $t \in \mathbf{R}$ the estimate $\hat{G}(t)$ is asymptotically efficient for $G(t)$;
2. for any function ϕ uniformly square integrable with respect to $F \in U$, i.e. such that

$$\lim_{A \to \infty} \sup_{F \in U} \int \mathbf{1}\{|\phi(x)| > A\}\, \phi^2(x)\, dF(x) = 0,$$

the estimate $\hat{\Phi} = \int \phi(x)\, d\hat{G}(x)$ will be asymptotically efficient for $\Phi(F) = \int \phi(x)\, dG(x)$.

Proof. The first part can be derived from Theorem 1 and the first order Taylor expansion uniformly in $F \in U$. The second essentially invokes Theorems 2 and 3. Let us start with the estimate of a. It enables us to eliminate the effect of the unknown center, just as before. Hence, everything is reduced to the symmetric distribution case. Instead of G we may consider the distribution K of the random variable $Z = |Y|$, so that $K(t) = 0$ for $t \leq 0$ and $K(t) = G(t) - G(-t) = 2\,G(t) - 1$ for $t > 0$. The symmetrized empirical CDF $\tilde{G}$ is no longer asymptotically efficient for G because the assumption that $\beta(G) = 0$ is not satisfied. However, due to the assumption that H_b is an odd function, the value $\beta(F)$ of the transfomation parameter is defined in terms of K or equivalently, via the CDF $\bar{G}$ of a random variable $|X - a|$. Once again, the uniformity in the first–order Taylor expansion implies that estimates based on $\{X_i - \tilde{a} : 1 \leq i \leq n\}$ and similar statistics calculated from the unobservable random variables $\{X_i - a : 1 \leq i \leq n\}$ have the same asymptotic behavior. That is, the supremum over $t \in \mathbf{R}$ of their difference is $o(n^{-1/2})$ uniformly in $F \in U$. In terms of the CDF $\bar{G}$ the restrictions $\beta(\bar{G}) = 0$ are handled in the same manner as we discussed before. (We assume that the family $\{H_b : b \in B\}$ forms a group, and the value b, corresponding to the identity of this group is equal to 0.) Hence, combining the two constructions, we obtain the desired asymptotically efficient estimation after the following steps.

1. Using a $n^{1/2}$–consistent estimate $\tilde{a}$, a symmetrized empirical *CDF* $\tilde{\bar{G}}$ is constructed.

2. The corresponding estimate for $\bar{G}$ is modified by means of restrictions $b(G) = 0$. ■

Remark. As a matter of fact, there is a hope that even the assumption that H_b is odd, can be sometimes relaxed. For instance, if

$$H_b(x) = \begin{cases} x & \text{for } x > 0 \\ b\,x\text{otherwise}, & \end{cases}$$

with unknown $b > 0$, then similar arguments work well and will lead to efficient estimates for a symmetric error distribution, provided the $n^{1/2}$–consistent estimates for a and b are found.

4.4 Censored data.

Consider the censored data case. A typical observation can be represented as a pair (X, δ) where $X = \min(T, C) = T \wedge C$ is a non–negative random variable, while $\delta = 1\{T \leq C\}$, is the indicator of the event $\{T \leq C\}$ and takes values 0 and 1 only. The random variables T and C are independent with unknown continuous CDFs F and G, respectively.

The coparameter of interest typically involves F only, so we let it be equal $\int \phi(x)\,dF(x)$ for some function ϕ. Theorem 3 will allow us to extend the class of possible functions ϕ, starting with the indicator functions of some intervals, up to functions that can be approximated with respect to $\mathbf{L}^2(P)$ norm, by a finite linear combination ϕ_m of bounded indicator functions, vanishing outside of a bounded interval, say $[0, b_m]$. If the approximation holds uniformly in $P \in U$, where P denotes a distribution of a pair (X, δ) and U is a neighborhood of P, then the coparameters $\Phi(P) = \int \phi(x)dF(x)$ can be estimated for a wider class of functions ϕ than indicators of intervals $\{t : a \leq t \leq b\}$.

In order to describe this problem as a restricted one, let us introduce the following functions.

$$K_j(x) = P\{X \leq x, \delta = j\}$$

with $j = 0$ or 1. In terms of the CDFs F and G, these two functions can be expressed, due to Millar (1983) and references given there, as $K_0(x) = \int_0^\infty F(c \wedge x)\, dG(c)$, and the similar expression for K_1 can be easily written, due to the symmetry between F and G. Dividing each of the functions K_j by the corresponding probability $\pi_j = P\{\delta = j\}$ for $j = 0$ and 1, we obtain the conditional CDFs, say $H_j(x) = P\{X \leq x \,|\delta = j\}$. The probabilities π_1 and $\pi_0 = 1 - \pi_1$ can be expressed via F and G, for instance, $\pi_1 = \int_0^\infty F(c)\, dG(c)$.

Hence, it seems worthwhile to use another infinite–dimensional parametrization via the conditional distributions H_0, H_1 and probability π_1, expressed in terms of the distribution of a typical observation (X, δ), so that $K_0(x) = (1 - \pi_1) \cdot H_0(x)$ and $K_1(x) = \pi_1 \cdot H_1(x)$. If two *CDFs* H_0, H_1 are given, the joint distribution of (X, δ) can be uniquely defined and vice versa, the joint distribution of (X, δ) uniquely defines H_0, H_1 and $\pi_1 = K(\infty, 1)$.

In terms of the conditional CDFs, F can be expressed as

$$F(x) = 1 - \exp\left(\int_0^\infty \frac{dH_1(t)}{1 - H(t-)}\right),$$

where $H = H_0 + H_1$ is the sum of the conditional CDFs. In terms of the distribution of censored data, the restrictions are expressed as follows: let F be the above defined function, and let G be defined similarly, that is

$$G(x) = 1 - \exp\left(\int_0^\infty \frac{dH_0(t)}{1 - H(t-)}\right).$$

Then functions K_0 and K_1, uniquely determined by any distribution on the product of the half–line $[0, \infty)$ and a set $\{0; 1\}$, must satisfy the above given system of equations, in terms of the CDFs H_0 and H_1 and a number π_1.

It is now easy to realize that under these restrictions nothing else is needed to estimate F and G, using the given expressions, unless there is some relation between them. For instance, having known G, we may improve the estimate for F.

Without any additional information on F and G, the well known modification of the Kaplan - Meier estimator for F can be obtained. The version expressed via integrals is asymptotically equivalent to the product limit estimator, initially proposed by Kaplan and Meier. See also van der Vaart (1991), Millar (1983), and Gill (1989) for additional details and references.

4.5 Other applications.

Starting with initially given estimates for a coparameter of interest $\Phi(P)$ and infinite–dimensional coparameter $\Psi(P)$ of the true distribution $P \in \mathcal{Q}$, the adjusted estimates can be constructed for a restricted problem, i.e. when $\Phi(P)$ is estimated under the assumption $\Psi(P) = 0$. The initial estimates are assumed to be asymptotically efficient for a larger family $\mathcal{Q}$ and asymptotically normal uniformly in $P \in U$, where U is a neighborhood in $\mathcal{Q}$. Then, under reasonable regularity assumptions, the asymptotically efficient estimates for $\Phi(P)$ can be found via the common adjustment procedure. It can be thought of as an infinite dimensional version of the one step approximation to the maximum likelihood estimate.

Further applications to more complex models, such as a bias selected model, considered by Vardi (1985), Gill, Vardi, and Wellner (1988), Pollard (1989), mixture models appearing in van der Vaart and Wellner (1990), Wellner (1985), van der Vaart (1991), models with censored data (see, e.g. Begun, et al. (1983), van der Vaart (1989)) and some other models admitting adaptive estimation (Bickel (1982), Bickel, Ritov, and Wellner (1991) and Wellner (1985)), will be considered in a subsequent paper. The uniform weak convergence for many interesting infinite dimensional coparameters is proved, under suitable assumptions, for these models as well. There are a few possible applications of the represented weak convergence results. One of the most attractive ideas is to justify the approximate confidence intervals considered in the bootstrap theory.

Chapter 5

Proofs.

5.1 Auxiliary lemmas.

The most important part of Theorem 1 is proved here.

Lemma 1. Suppose that the assumptions of Theorem 1 are satisfied uniformly in $P \in U$. Then the convergence $\langle B_P^n , (\beta^n - \beta(P)) \rangle \to 0$ in probability holds uniformly in $P \in U \cap \mathcal{P}$.

Proof of Lemma 1. Fix $\epsilon > 0$. Choose a compact set $K \subset \mathbf{B}$ and $n_0 = n_0(\epsilon)$, such that

$$\mathbf{Prob}\{ d(B_P^n , K) > \epsilon \} < \epsilon ,$$

for every $P \in U$ and every $n \geq n_0$. The sequence β^n is uniformly consistent and uniformly bounded in probability, therefore, using the standard arguments, we obtain that convergence $\langle b , (\beta^n - \beta(P)) \rangle \to 0$ holds uniformly in $b \in K$. If $\sup_{b \in K} |\langle b, (\beta^n - \beta(P)) \rangle| < \epsilon$ with probability at least $1 - \epsilon$, for all $n \geq n_1(\epsilon)$, and $d(B_P^n , K) < \epsilon$ with probability at least $1 - \epsilon$, for all $n \geq n_0(\epsilon)$, then we obtain that the inequality $|\langle B_P^n , (\beta^n - \beta(P)) \rangle| < 2\epsilon$ holds eventually with a probability at least $1 - 2\epsilon$, *Q.E.D.*

5.1.1 Uniform weak convergence of linear functionals.

Lemma 2. Suppose that $\mathbf{D}$–valued random variables B_P^n converge weakly to $\mathbf{B}$–valued random variables B_P uniformly in $P \in U$ as $n \to \infty$. Suppose also that the functionals $\beta(P)$ are uniformly bounded, that is $\sup_{P \in U} \|\beta(P)\| < \infty$. Then $\langle B_P^n , \beta(P) \rangle$ converges to $\langle B_P , \beta(P) \rangle$ uniformly in $P \in U$, as $n \to \infty$.■

This Lemma proves the easiest part of Theorem 1 and it will apply to the case where instead of $\langle B_P^n , \beta(P) \rangle$ the random variables $n^{1/2} \left(\hat{\Phi} - \Phi(P) \right) = n^{1/2} \tilde{\Phi} - \langle B_P^n , \beta(P) \rangle$ are considered. The spaces $\mathbf{B}$ and $\mathbf{D}$ must be replaced by the direct sums $\mathbf{B} \oplus \mathbf{R}$ and $\mathbf{D} \oplus \mathbf{R}$ respectively, and the result will be exploited for the functionals

$$\int \phi(P,x)\, dW_P(x) - \langle B_P^n, \beta(P) \rangle.$$

The proof is obvious here. In fact, a generalized result was proved in Koshevnik (1984c). It can be easily established, using standard arguments of weak convergence theory.

5.1.2 Proposition.

Suppose that the first two assumptions of Theorem 1 hold for **D**–valued random elements B_v^n and **B**–valued random elements B_v. These two assumptions are recalled below (1 and 2).

1. $B_v^n \xrightarrow{\mathcal{D}} B_v$ uniformly in $v \in V$, as $n \to \infty$.
2. The family $\{ \mathcal{L} \, [B_v] \, : \, v \in V \}$ of distributions is uniformly tight.
3. Assume also that the image of **B** under every γ_v belongs **B**.

Then the weak convergence $\gamma_v(B_v^n) \xrightarrow{\mathcal{D}} \gamma_v(B_v)$ holds uniformly in $v \in V$.

5.1.3 Proof of Theorem 1.

The combination of the above Lemmas will prove the theorem. Due to Lemma 1, consistency and uniform boundedness of β^n will reduce the study of the asymptotic distribution of the random variables

$$n^{1/2} \left[\tilde{\Phi} - \langle \tilde{\Psi}, \beta^n \rangle - \Phi(P) \right]$$

to the case of

$$n^{1/2} \left[\tilde{\Phi} - \langle \tilde{\Psi} , \beta(P) \rangle \right] - \Phi(P).$$

If $\Psi(P) = 0$ for $P \in \mathcal{P}$, then the limiting behavior of these random variables is the same as for

$$n^{1/2} \left[\tilde{\Phi} - \langle B_P^n , \beta(P) \rangle \right],$$

and due to Lemma 2, we obtain the required convergence, also uniformly in $P \in U$, if $U \subset \mathcal{P}$ and satisfies the assumptions of the Theorem.

5.1.4 Proof of Theorem 2.

Now we return to Theorem 2. The assumptions there and Theorem 1 will give, for every m, the sequence of estimates $\{\hat{\Phi}^n_m = \tilde{\Phi}^n - \langle \tilde{\Psi}, \beta^n_m \rangle\}$, such that weak convergence (as $n \to \infty$)

$$n^{1/2}\left[\hat{\Phi}^n_m - \Phi(P)\right] \xrightarrow{\mathcal{D}} \int \phi(x)\,dW_P(x) - \langle B_P, \beta_m(P)\rangle$$

holds uniformly in $P \in U$, for $U \subset \mathcal{P}$ and U satisfies the assumptions of the theorem. On the other hand, the variances converge to zero, i.e.

$$\lim_{m\to\infty}\int [\phi_1(P,x) - \phi(P,x) + \psi(P,x,\beta_m)]^2\,dP(x) = 0$$

uniformly in $P \in U$, by assumption. Hence, similar convergence also holds for the values of the corresponding functionals and $\langle B_P, \beta_m(P)\rangle$ has limit as $m \to \infty$. This convergence in mean implies the convergence in probability, the Chebyshev inequality provides uniform bounds for variances of these Gaussian random variables, hence we obtain that convergence in probability to their limits holds uniformly in $P \in U$. Once again, using uniform tightness of the limiting family $\{\mathrm{M}_P : P \in U\}$ of probabilities on $\mathbf{B}$, we obtain that the random variables $\int \phi(P,x)\,dW_P(x) - \langle B_P, \beta_m(P)\rangle$ converge in distribution to the limit formulated in the theorem.

Hence, for every m the sequence of estimates is constructed, and its limit approaches $\int \phi_1(P,x)\,dW_P(x)$, as $m \to \infty$. To complete the proof, we can use the "diagonal process" to choose a subsequence $m_n \uparrow \infty$, such that $n \int [K(P,x,\beta_m(P)) - \xi(P,x)]^2 \to \infty$ uniformly in $P \in U$, and it is this sequence that will provide the desired convergence for $\hat{\Phi} = \tilde{\Phi} - \langle \tilde{\Psi}, \beta^n_m \rangle$.

The proof of Theorem 3 is based on arguments similar to those used for Theorem 2.

Acknowledgements.

I thank all my colleagues from Moscow University, Harvard University and Southern Methodist University for their patient attention, useful suggestions and helpful advice. My special thanks to Marvin Zelen, Bill Schucany and Wayne Woodward. I very much appreciate fruitful discussions with Dick Dudley, Dorota Dabrowska, Georgia Thompson, Andrzej Kozek and Sam Efromovich. I am grateful to Nan M. Laird for extending the hospitality of the Harvard School of Public Health.

References.

Begun, J. M., Hall, W. J., Huang, W. M., Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.* **10** 432-452.

Beran, R. J. (1974). Asymptotically efficient adaptive rank estimates in location models. *Ann. Statist.*, **2**, 63-74.

Beran, R. J. (1980). Asymptotic lower bounds for risk in robust estimation. *Ann. Statist.* **8** 1252-1264.

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

Bickel, P. J. (1984). Book Review. *Ann. Statist.* **12** 786-791.

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

Bickel, P. J. , Ritov, Y., Wellner, J. A. (1991). Efficient estimation of linear functionals of a probability measure P with known marginal distributions. *Ann. Statist.* **19** 1316-1346.

Gill, R. D. (1989). Non- and semiparametric maximum likelihood estimators and the von Mises method. *Scand. J. Statist.* **16** 97-128.

Gill, R. D., Vardi, Y., Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16** 1069-1112.

Koshevnik, Y., Levit, B. (1976). On a non-parametric analogue for the information matrix. *Theory Probab. Appl.* **21** 738-753.

Koshevnik, Y., Levit, B. (1980). Risk bounds in estimation of symmetrical distributions. *J. Sov. Math.* **21** 65-75 *(Translated 1983)*

Koshevnik, Y. (1981). Nonparametric estimates and regular families of distribution functions. *3rd Intern. Conf. Probab. Theor. Math. Statist.* Vilnius **1** 247-249. (*In Russian.*)

Koshevnik, Y. (1982). *Limit Theorems of Nonparametric Statistics.* Sov. Inst. of Scientific and Technical Information. (*In Russian.*)

Koshevnik, Y. (1984a). Limit theorems for nonparametric estimates of some symmetric distribution functions. *Methods of Data Analysis, Estimation, and Choice:* Sov. Inst. Syst. Res. **11** 71-80. (*In Russian.*)

Koshevnik, Y. (1984b). Asymptotic properties of nonparametric estimators of the characteristic function. *J. Sov. Math.* **33** 758-767. *(Translated 1986.)*

Koshevnik, Y. (1984c). On some limit properties of nonparametric estimates of a distribution function. *Theory Probab. Appl.* **29** 807-813.

Koshevnik, Y. (1985). On the conditional estimation of smooth functionals. *4th Intern. Conf. Probab. Theor. Math. Statist.* Vilnius **2** 67-69. (*In Russian.*)

Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. *Lect. Notes in Math.* **976** 75-265.

Pfanzagl, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lect. Notes in Statist.* **13.** Springer, New York.

Pollard, D. (1982). *Empirical Processes: Theory and Applications. CBMS Regional Conferences in Probability and Statistics.*

Sacks, J. (1975). An asymptotically efficient sequence of estimators of a location parameter. *Ann. Statist.* **3** 285-298.

Stone, C. J. (1975). Adaptive maximum likelihood estimators of a location parameter. *Ann. Statist.* **3** 267-284.

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

Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inf.* **16** 89-105.

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

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

van der Vaart, A. W., Wellner, J. A. (1990). Existence and consistency of maximum likelihood in upgraded mixture models. *Preprint.*

Vardi, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **10** 178-203.

Wellner, J. A. (1985). Semiparametric models: Progress and problems. *Bull. Inst. Internat. Statist.* **51** 23.1-23.20.

Department of Statistical Science
Southern Methodist University
Dallas, Texas 75275-0332