Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Electron J Stat. Author manuscript; available in PMC 2018 April 04.

Published in final edited form as:

Electron J Stat. 2009 ; 3: 1305–1321. doi:10.1214/09-EJS479.

Semiparametric Minimax Rates

James Robins,
Department of Biostatistics and Epidemiology, School of Public Health, Harvard University

Eric Tchetgen Tchetgen,
Department of Biostatistics and Epidemiology, School of Public Health, Harvard University

Lingling Li, and
Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care, Boston, MA, 02215

Aad van der Vaart
Department of Mathematics, Vrije Universiteit, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Abstract

We consider the minimax rate of testing (or estimation) of non-linear functionals defined on semiparametric models. Existing methods appear not capable of determining a lower bound on the minimax rate of testing (or estimation) for certain functionals of interest. In particular, if the semiparametric model is indexed by several infinite-dimensional parameters. To cover these examples we extend the approach of [1], which is based on comparing a "true distribution" to a convex mixture of perturbed distributions to a comparison of two convex mixtures. The first mixture is obtained by perturbing a first parameter of the model, and the second by perturbing in addition a second parameter. We apply the new result to two examples of semiparametric functionals:the estimation of a mean response when response data are missing at random, and the estimation of an expected conditional covariance functional.

AMS 2000 subject classifications

Primary 62G05; 62G20; 62G20; 62F25

Keywords and phrases

Nonlinear functional; nonparametric estimation; Hellinger distance

1. Introduction

Let $X_1, X_2, \ldots, X_n$ be a random sample from a density p relative to a measure ν on a sample space $(\mathcal{X}, \mathcal{A})$. It is known that p belongs to a collection $\mathcal{P}$ of densities, and we wish to estimate the value $\chi(p)$ of a functional $\chi:\mathcal{P} \to \mathbb{R}$. In this setting the mimimax rate of estimation of $\chi(p)$ relative to squared error loss can be defined as the root of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\inf_{T_n} \sup_{p \in \mathcal{P}} \mathrm{E}_p |T_n - \chi(p)|^2,$$

where the infimum is taken over all estimators $T_n = T_n(X_1, \ldots, X_n)$. Determination of a minimax rate in a particular problem often consists of proving a "lower bound", showing that the mean square error of no estimator tends to zero faster than some rate ε_n^2, combined with the explicit construction of an estimator with mean square error ε_n^2.

The lower bound is often proved by a testing argument, which tries to separate two subsets of the set $\{P^n: p \in \mathcal{P}\}$ of possible distributions of the observation $(X_1, \ldots, X_n)$. Even though testing is a statistically easier problem than estimation under quadratic loss, the corresponding minimax rates are often of the same order. The testing argument can be formulated as follows. If P_n and Q_n are in the convex hull of the sets $\{P^n: p \in \mathcal{P}, \mathcal{X}(p) \leq 0\}$ and $\{P^n: p \in \mathcal{P}, \mathcal{X}(p) \geq \varepsilon_n\}$ and there exist no sequence of tests of P_n versus Q_n with both error probabilities tending to zero, then the minimax rate is not faster than a multiple of ε_n. Here existence of a sequence of tests with errors tending to zero (a perfect sequence of tests) is determined by the asymptotic separation of the sequences P_n and Q_n and can be described, for instance, in terms of the *Hellinger affinity*

$$\rho(P_n, Q_n) = \int \sqrt{dP_n}\sqrt{dQ_n}.$$

If $\rho(P_n, Q_n)$ is bounded away from zero as $n \to \infty$, then no perfect sequence of tests exists (see e.g. Section 14.5 in [2]).

One difficulty in applying this simple argument is that the relevant (least favorable) two sequences of measures P_n and Q_n need not be product measures, but can be arbitrary convex combinations of product measures. In particular, it appears that for nonlinear functionals at least one of the two sequences must be a true mixture. This complicates the computation of the affinity $\rho(P_n, Q_n)$ considerably. [1] derived an elegant nice lower bound on the affinity when P_n is a product measure and Q_n a convex mixture of product measures, and used it to determine the testing rate for functionals of the type $\int f(p)\, d\nu$, for a given smooth function $f: \mathbb{R} \to \mathbb{R}$, the function $f(x) = x^2$ being the crucial example.

In this paper we are interested in structured models $\mathcal{P}$ that are indexed by several subparameters and where the functional is defined in terms of the subparameters. It appears that testing a product versus a mixture is often not least favorable in this situation, but testing two mixtures is. Thus we extend the bound of [1] to the case that both P_n and Q_n are mixtures. In our examples P_n is equal to a convex mixture obtained by perturbing a first parameter of the model, and Q_n is obtained by perturbing in addition a second parameter. We also refine the bound in other, less essential directions.

The main general results of the paper are given in Section 2. In Section 3 we apply these results to two examples of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2. Main result

For $k \in \mathbb{N}$ let $\mathcal{X} = \cup_{j=1}^{k} \mathcal{X}_j$ be a measurable partition of the sample space. Given a vector $\lambda = (\lambda_1, \ldots, \lambda_k)$ in some product measurable space $\Lambda = \Lambda_1 \times \cdots \times \Lambda_k$ let P_λ and Q_λ be probability measures on $\mathcal{X}$ such that

1. $P_\lambda(\mathcal{X}_j) = Q_\lambda(\mathcal{X}_j) = p_j$ for every $\lambda \in \Lambda$, for some probability vector $(p_1, \ldots, p_k)$.
2. The restrictions of P_λ and Q_λ to $\mathcal{X}_j$ depend on the jth coordinate λ_j of $\lambda = (\lambda_1, \ldots, \lambda_k)$ only.

For p_λ and q_λ densities of the measures P_λ and Q_λ that are jointly measurable in the parameter λ and the observation, and π a probability measure on Λ, define $p = \int p_\lambda \, d\pi(\lambda)$ and $q = \int q_\lambda \, d\pi(\lambda)$, and set

$$a = \max_j \sup_\lambda \int_{\chi_j} \frac{(p_\lambda - p)^2}{p_\lambda} \frac{d\nu}{p_j},$$

$$b = \max_j \sup_\lambda \int_{\chi_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda} \frac{d\nu}{p_j},$$

$$d = \max_j \sup_\lambda \int_{\chi_j} \frac{(q - p)^2}{p_\lambda} \frac{d\nu}{p_j}.$$

Theorem 2.1

If $np_j(1 \vee a \vee b) \le A$ for all j and $\underline{B} \le p_\lambda \le \bar{B}$ for positive constants A, $\underline{B}$, $\bar{B}$, then there exists a constant C that depends only on A, $\underline{B}$, $\bar{B}$ such that, for any product probability measure $\pi = \pi_1 \otimes \cdots \otimes \pi_k$,

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda)\right) \ge 1 - Cn^2(\max_j p_j)(b^2 + ab) - Cnd.$$

Proof

The numbers a, b and d are the maxima over j of the numbers a, b and d defined in Lemma 2.2, but with the measures P_λ and Q_λ replaced there by the measures P_{j,λ_j} and Q_{j,λ_j} given in (2.1). Define a number c similarly as

$$\max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{p^2}{p_\lambda} \frac{d\nu}{p_j}.$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Under the assumptions of the theorem c is bounded above by $\bar{B}^2/\underline{B}$.

By applying Lemma 2.1 and next Lemma 2.2 we see that the left side is at least

$$\mathrm{E}\prod_{j=1}^{k}\left(1-\frac{1}{4}\sum_{r=2}^{N_j}\binom{N_j}{r}b^r-\frac{1}{2}N_j^2\sum_{r=1}^{N_j-1}\binom{N_j-1}{r}a^r b-\frac{1}{2}N_j^2 c^{N_j-1}d\right)$$
$$\geq 1-\mathrm{E}\sum_{j=1}^{k}\left(1-\frac{1}{4}\sum_{r=2}^{N_j}\binom{N_j}{r}b^r-\frac{1}{2}N_j^2\sum_{r=1}^{N_j-1}\binom{N_j-1}{r}a^r b-\frac{1}{2}N_j^2 c^{N_j-1}d\right),$$

because $\prod_{j=1}^{k}(1-a_j)\geq 1-\sum_{j=1}^{k}a_j$ for any nonnegative numbers $a_1, \ldots, a_k$. The expected values on the binomial variables N_j can be evaluated explicitly, using the identities, for N a binomial variable with parameters n and p,

$$\mathrm{E}\sum_{r=2}^{N}\binom{N}{r}b^r=\mathrm{E}((1+b)^N-1-Nb)=(1+bp)^n-1-npb,$$

$$\mathrm{E}N^2c^{N-1}=np(cp+1-p)^{n-2}(cnp+1-p),$$

$$\mathrm{E}N^2\sum_{r=1}^{N-1}\binom{N-1}{r}a^r=\mathrm{E}N^2((1+a)^{N-1}-1)=np(1+ap)^{n-2}(1+nap+np-p)-np(1-p)-n^2p^2.$$

Under the assumption that $np(1\vee a\vee b\vee c)\lesssim 1$, the right sides of these expressions can be seen to be bounded by multiples of $(npb)^2$, np and $(np)^2a$, respectively. We substitute these bounds in the first display of the proof, and use the equality $\Sigma_j p_j=1$ to complete the proof.

Remark 2.1

If $\min p_j\sim\max_j p_j\sim 1/n^{1+\varepsilon}$ for some $\varepsilon>0$, which arises for equiprobable partitions in $k\sim n^{1+\varepsilon}$ sets, then there exists a number n_0 such that $\mathrm{P}(\max_j N_j>n_0)\to 0$. (Indeed, the probability is bounded by $k(n\max_j p_j)^{n_0+1}$.) Under this slightly stronger assumption the computations need only address $N_j\leq n_0$ and hence can be simplified.

The proof of Theorem 2.1 is based on two lemmas. The first lemma factorizes the affinity into the affinities of the restrictions to the partitioning sets, which are next lower bounded using the second lemma. The reduction to the partioning sets is useful, because it reduces the n-fold products to lower order products for which the second lemma is accurate.

Define probability measures P_{λ_j} and Q_{λ_j} on $\mathscr{X}_j$ by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$dP_{j,\lambda_j} = \frac{1_{\chi_j} dP_\lambda}{p_j}, \qquad dQ_{j,\lambda_j} = \frac{1_{\chi_j} dQ_\lambda}{p_j}. \quad (2.1)$$

Lemma 2.1

For any product probability measure $\pi = \pi_1 \otimes \cdots \otimes \pi_k$ on Λ and every $n \in \mathbb{N}$,

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda)\right) = \mathrm{E} \prod_{j=1}^k \rho_j(N_j),$$

where $(N_1, \ldots, N_k)$ is multinomially distributed on n trials with success probability vector $(p_1, \ldots, p_k)$ and $\rho_j : \{0, \ldots, n\} \to [0, 1]$ is defined by $\rho_j(0) = 1$ and

$$\rho_j(m) = \rho\left(\int P_{j,\lambda_j}^m d\pi_j(\lambda), \int Q_{j,\lambda_j}^m d\pi_j(\lambda)\right), \qquad m \geq 1.$$

Proof

Set $\bar{P}_n := \int P_\lambda^n d\pi(\lambda)$ and consider this as the distribution of the vector $(X_1, \ldots, X_n)$. Then, for p_λ and q_λ densities of P_λ and Q_λ relative to some dominating measure, the left side of the lemma can be written as

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda)\right) = \mathrm{E}_{\bar{P}_n} \sqrt{\frac{\int \prod_{j=1}^k \prod_{i: X_i \in \chi_j} q_\lambda(X_i) d\pi(\lambda)}{\int \prod_{j=1}^k \prod_{i: X_i \in \chi_j} p_\lambda(X_i) d\pi(\lambda)}}.$$

Because by assumption on each partitioning set $\mathcal{X}_j$ the measures Q_λ and P_λ depend on λ_j only, the expressions $\prod_{i: X_i \in \mathcal{X}_j} q_\lambda(X_i)$ and $\prod_{i: X_i \in \mathcal{X}_j} p_\lambda(X_i)$ depend on λ only through λ_j. In fact, within the quotient on the right side of the preceding display, they can be replaced by $\prod_{i: X_i \in \mathcal{X}_j} q_{j,\lambda_j}(X_i)$ and $\prod_{i: X_i \in \mathcal{X}_j} p_{j,\lambda_j}(X_i)$ for q_{j,λ_j} and p_{j,λ_j} densities of the measures Q_{j,λ_j} and P_{j,λ_j}. Because π is a product measure, we can next use Fubini's theorem and rewrite the resulting expression as

$$\mathrm{E}_{\bar{P}_n} \sqrt{\frac{\prod_{j=1}^k \int \prod_{i: X_i \in \chi_j} q_{j,\lambda_j}(X_i) d\pi_j(\lambda_j)}{\prod_{j=1}^k \int \prod_{i: X_i \in \chi_j} p_{j,\lambda_j}(X_i) d\pi_j(\lambda_j)}}.$$

Here the two products over j can be pulled out of the square root and replaced by a single product preceding it. A product over an empty set (if there is no $X_i \in \mathcal{X}_j$) is interpreted as 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Define variables $I_1, \ldots, I_n$, I_n that indicate the partitioning sets that contain the observations: $I_i = j$ if $X_i \in \mathcal{X}_j$ for every i and j, and let $N_j = (\#1 \leq i \leq n\colon I_i = j)$ be the number of X_i falling in $\mathcal{X}_j$.

The measure $\bar{P}_n$ arises as the distribution of $(X_1, \ldots, X_n)$ if this vector is generated in two steps. First λ is chosen from π and next given this λ the variables $X_1, \ldots, X_n$ are generated independently from P_λ. Then given λ the vector $(N_1, \ldots, N_k)$ is multinomially distributed on n trials and probability vector $(P_\lambda(\mathcal{X}_1), \ldots, P_\lambda(\mathcal{X}_k))$. Because the latter vector is independent of λ and equal to $(p_1, \ldots, p_k)$ by assumption, the vector $(N_1, \ldots, N_k)$ is stochastically independent of λ and hence also unconditionally, under $\bar{P}_n$, multinomially distributed with parameters n and $(p_1, \ldots, p_k)$. Similarly, given λ the variables $I_1, \ldots, I_n$ are independent and the event $I_i = j$ has probability $P_\lambda(\mathcal{X}_j)$, which is independent of λ by assumption. It follows that the random elements $(I_1, \ldots, I_n)$ and λ are stochastically independent under $\bar{P}_n$.

The conditional distribution of $X_1, \ldots, X_n$ given λ and $I_1, \ldots, I_n$ can be described as: for each partitioning set $\mathcal{X}_j$ generate N_j variables independently from P_λ restricted and renormalized to $\mathcal{X}_j$, i.e. from the measure P_{j,λ_j}; do so independently across the partitioning sets; and attach correct labels $\{1, \ldots, n\}$ consistent with $I_1, \ldots, I_n$ to the n realizations obtained. The conditional distribution under $\bar{P}_n$ of $X_1, \ldots, X_n$ given I_n is the mixture of this distribution relative to the conditional distribution of λ given $(I_1, \ldots, I_n)$, which was seen to be the unconditional distribution, π. Thus we obtain a sample from the conditional distribution under $\bar{P}_n$ of $(X_1, \ldots, X_n)$ given $(I_1, \ldots, I_n)$ by generating for each partitioning set $\mathcal{X}_j$ a set of N_j variables from the measure $\int P_{j,\lambda_j}^{N_j} d\pi_j(\lambda_j)$, independently across the partitioning sets, and next attaching labels consistent with $I_1, \ldots, I_n$.

Now rewrite the right side of the last display by conditioning on $I_1, \ldots, I_n$ as

$$\mathrm{E}_{\bar{P}_n} \mathrm{E}_{\bar{P}_n} \left[\prod_{j=1}^{k} \sqrt{\frac{\int \prod_{i:I_i=j} q_{j,\lambda_j}(X_i) d\pi_j(\lambda_j)}{\int \prod_{i:I_i=j} p_{j,\lambda_j}(X_i) d\pi_j(\lambda_j)}} \middle| I_1, \ldots, I_n \right].$$

The product over j can be pulled out of the conditional expectation by the conditional independence across the partitioning sets. The resulting expression can be seen to be of the form as claimed in the lemma.

The second lemma does not use the partitioning structure, but is valid for mixtures of products of arbitrary measures on a measurable space. For λ in a measurable space Λ let P_λ and Q_λ be probability measures on a given sample space $(\mathcal{X}, \mathcal{A})$, with densities p_λ and q_λ relative to a given dominating measure ν, which are jointly measurable. For a given (arbitrary) density p define functions $\ell_\lambda = q_\lambda - p_\lambda$ and $\kappa_\lambda = p_\lambda - p$, and set

$$a = \sup_{\lambda \in \Lambda} \int \frac{\kappa_\lambda^2}{p_\lambda} d\nu,$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$b = \sup_{\lambda \in \Lambda} \int \frac{\ell_\lambda^2}{p_\lambda} d\nu,$$

$$c = \sup_{\lambda \in \lambda} \int \frac{p^2}{p_\lambda} d\nu,$$

$$d = \sup_{\lambda \in \Lambda} \int \frac{\left(\int \ell_\mu d\pi(\mu)\right)^2}{p_\lambda} d\nu.$$

Lemma 2.2

For any probability measure π on Λ and every $n \in \mathbb{N}$,

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda)\right) \geq 1 - \frac{1}{4}\sum_{r=2}^{n}\binom{n}{r} b^r - \frac{1}{2} n^2 \sum_{r=1}^{n-1}\binom{n-1}{r} a^r b - \frac{1}{2} n^2 c^{n-1} d.$$

Proof

Consider the measure $\bar{P}_n = \int P_\lambda^n d\pi(\lambda)$, which has density $\bar{p}_n(\vec{x}_n) = \int \prod_{i=1}^n p_\lambda(x_i) d\pi(\lambda)$ relative to ν^n, as the distribution of $(X_1, \ldots, X_n)$. Using the inequality $\mathrm{E}\sqrt{1+Y} \geq 1 - \mathrm{E}Y^2/8$, valid for any random variable Y with $1 + Y \geq 0$ and $\mathrm{E}\, Y = 0$ (see for example [1], we see that

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda)\right) = \mathrm{E}_{\bar{P}_n} \sqrt{1 + \frac{\int [\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)] d\pi(\lambda)}{\bar{p}_n(X_1, \ldots, X_n)}} \quad (2.2)$$

$$\geq 1 - \frac{1}{8}\mathrm{E}_{\bar{P}_n} \frac{\int [\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)] d\pi(\lambda)^2}{\bar{p}_n(X_1, \ldots, X_n)^2}.$$

It suffices to upper bound the expected value on the right side. To this end we expand the difference $\prod_{i=1}^n q_\lambda(X_i) - \prod_{i=1}^n p_\lambda(X_i)$ as $\Sigma_{|I| \geq 1} \prod_{i \in I^c} p_\lambda(X_i) \prod_{i \in I} \ell_\lambda(X_i)$, where the sum ranges over all nonempty subsets $I \subset \{1, \ldots, n\}$. We split this sum in two parts, consisting of the terms indexed by subsets of size 1 and the subsets that contain at least 2 elements, and separate the square of the sum of these two parts by the inequality $(A + B)^2 \leq 2A^2 + 2B^2$.

If $n = 1$, then there are no subsets with at least two elements and the second part is empty. Otherwise the sum over subsets with at least two elements contributes two times

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\int \frac{\int \sum_{|I|\geq 2} \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I} \ell_\lambda(x_i) d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) d\pi(\lambda)} d\nu^n(\vec{x}_n)$$
$$\leq \int\int \left(\sum_{|I|\geq 2} \prod_{i\in I^c} \sqrt{p_\lambda}(x_i) \prod_{i\in I} \frac{\ell_\lambda}{\sqrt{p_\lambda}}(x_i) \right)^2 d\pi(\lambda) d\nu^n(\vec{x}_n)$$
$$= \sum_{|I|\geq 2} \int\int \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I} \frac{\ell_\lambda^2}{p_\lambda}(x_i) d\pi(\lambda) d\nu^n(\vec{x}_n).$$

To derive the first inequality we use the inequality $(\mathrm{E}\,U)^2/\mathrm{E}\,V \leq \mathrm{E}(U^2/V)$, valid for any random variables U and $V \geq 0$, which can be derived from Cauchy-Schwarz' or Jensen's inequality. The last step follows by writing the square of the sum as a double sum and noting that all off-diagonal terms vanish, as they contain at least one term $\ell_\lambda(x_i)$ and $\int \ell_\lambda \, d\nu = 0$. The order of integration in the right side can be exchanged, and next the integral relative to ν^n can be factorized, where the integrals $\int p_\lambda \, d\nu$ are equal to 1. This yields the contribution $2 \sum_{|I|\geq 2} b^{|I|}$ to the bound on the expectation in (2.2).

The sum over sets with exactly one element contributes two times

$$\int \frac{\int \sum_{j=1}^n \prod_{i\neq j} p_\lambda(x_i) \ell_\lambda(x_j) d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) d\pi(\lambda)} d\nu^n(\vec{x}_n). \quad (2.3)$$

Here we expand

$$\prod_{i\neq j} p_\lambda(x_i) - \prod_{i\neq j} p(x_i) = \prod_{i\neq j} p_\lambda(x_i) - \prod_{i\neq j} (p_\lambda - \kappa_\lambda)(x_i) = -\sum_{|I|\geq 1, j\notin I} \prod_{i\in I^c} p_\lambda(x_i) \prod_{i\in I} (-\kappa_\lambda)(x_i),$$

where the sum is over all nonempty subsets $I \subset \{1, \ldots, n\}$ that do not contain j. Replacement of $\prod_{i\neq j} p_\lambda(x_i)$ by $\prod_{i\neq j} p(x_i)$ changes (2.3) into

$$\int \frac{\int \sum_{j=1}^n \prod_{i\neq j} p(x_i) \ell_\lambda(x_j) d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) d\pi(\lambda)} d\nu^n(\vec{x}_n) \leq n \sum_{j=1}^n \int \frac{\prod_{i\neq j} p^2(x_i) \int \ell_\lambda(x_j) d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) d\pi(\lambda)} d\nu^n(\vec{x}_n)$$
$$\leq n \sum_{j=1}^n \int\int \prod_{i\neq j} \frac{p^2}{p_\mu}(x_i) \frac{\int \ell_\lambda d\pi(\lambda)^2}{p_\mu}(x_j) d\pi(\mu) d\nu^n(\vec{x}_n).$$

In the last step we use that $1/\mathrm{E}\,V \leq \mathrm{E}(1/V)$ for any positive random variable V. The integral with respect to ν^n in the right side can be factorized, and the expression bounded by $n^2 c^{n-1} d$. Four this this must be added to the bound on the expectation in (2.2).

Finally the remainder after substituting $\prod_{i\neq j} p(x_i)$ for $\prod_{i\neq j} p_\lambda(x_i)$ in (2.3) contributes

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\int \frac{\int \sum_{j=1}^{n} \sum_{|I| \geq 1, j \notin I} \prod_{i \in I^c} p_\lambda(x_i) \prod_{i \in I} (-\kappa_\lambda)(x_i) \ell_\lambda(x_j) d\pi(\lambda)^2}{\int \prod_i p_\lambda(x_i) d\pi(\lambda)} d\nu^n(\vec{x}_n)$$

$$\leq \int \int \left(\sum_{j=1}^{n} \sum_{|I| \geq 1, j \notin I} \prod_{i \in I^c} \sqrt{p_\lambda}(x_i) \prod_{i \in I} \frac{-\kappa_\lambda}{\sqrt{p_\lambda}}(x_i) \frac{\ell_\lambda}{\sqrt{P_\lambda}}(x_j) \right)^2 d\pi(\lambda) d\nu^n(\vec{x}_n)$$

$$\leq n \sum_{j=1}^{n} \int \int \left(\sum_{|I| \geq 1, j \notin I} \prod_{i \in I^c} \sqrt{p_\lambda}(x_i) \prod_{i \in I} \frac{-\kappa_\lambda}{\sqrt{p_\lambda}}(x_i) \right)^2 \frac{\ell_\lambda^2}{p_\lambda}(x_j) d\pi(\lambda) d\nu^n(\vec{x}_n)$$

$$= n \sum_{j=1}^{n} \sum_{|I| \geq 1, j \notin I} \int \int \prod_{i \in I^c} p_\lambda(x_i) \prod_{i \in I} \frac{\kappa_\lambda^2}{p_\lambda}(x_i) \frac{\ell_\lambda^2}{p_\lambda}(x_j) d\pi(\lambda) d\nu^n(\vec{x}_n).$$

We exchange the order of integration and factorize the integral with respect to ν^n to bound this by $n^2 \Sigma_{|I| \geq 1, j \notin I} a^{|I|} b$.

3. Applications

3.1. Estimating the mean response in missing data models

Suppose that a typical observation is distributed as $X = (Y A, A, Z)$ for Y and A taking values in the two-point set $\{0, 1\}$ and conditionally independent given Z. We think of Y as a response variable, which is observed only if the indicator A takes the value 1, and are interested in estimating the mean response $\mathrm{E}\, Y$. The covariate Z is chosen such that it contains all information on the dependence between response and missingness indicator ("missing at random"). We assume that Z takes its values in $\mathcal{Z} = [0, 1]^d$.

The model can be parameterized by the marginal density f of Z relative to Lebesgue measure measure ν on $\mathcal{Z}$, and the probabilities $b(z) = \mathrm{P}(Y = 1 | Z = z)$ and $a(z)^{-1} = \mathrm{P}(A = 1 | Z = z)$. Alternatively, the model can be parameterized by the function $g = f/a$, which is the conditional density of Z given $A = 1$ up to the norming factor $\mathrm{P}(A = 1)$. Under this latter parametrization which we adopt henceforth, the density p of an observation X is described by the triple (a, b, g) and the functional of interest is expressed as $\mathcal{X}(p) = \int abg \, d\nu$.

Define $C_M^\alpha[0, 1]^d$ as M times the unit ball of the Hölder space of α-smooth functions on $[0, 1]^d$. For given positive constants $\alpha, \beta, \gamma, \phi$ and $\underline{M}, M$, we consider the models

- $$\mathcal{P}_1 = \{(a, b, g) : a \in C_M^\alpha[0, 1]^d, b \in C_M^\beta[0, 1]^d, g = 1/2, a, b \geq \underline{M}\}.$$
- $$\mathcal{P}_2 = \{(a, b, g) : a \in C_M^\alpha[0, 1]^d, b \in C_M^\beta[0, 1]^d, g \in C^\gamma[0, 1]^d, a, b \geq \underline{M}\}.$$

If $(\alpha + \beta)/2 \geq d/4$, then a $\sqrt{n}$-rate is attainable over $\mathcal{P}_2$ (see [3]), and a standard "two-point" proof can show that this rate cannot be improved. Here we are interested in the case $(\alpha +$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\beta)/2 < d/4$, when the rate becomes slower than $1/\sqrt{n}$. The paper [3] constructs an estimator that attains the rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ uniformly over $\mathcal{P}_2$ if

$$\frac{\gamma}{2\gamma+d} > \left(\frac{\alpha \vee \beta}{d}\right)\left(\frac{d-2\alpha-2\beta}{d+2\alpha+2\beta}\right) := \gamma(\alpha,\beta). \quad (3.1)$$

We shall show that this result is optimal by showing that the minimax rate over the smaller model $\mathcal{P}_1$ is not faster than $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.

In the case that $\alpha = \beta$ these results can be proved using the method of [1], but in general we need a construction as in Section 2 with P_λ based on a perturbation of the smoothest parameter of the pair (a, b) and Q_λ constructed by perturbing in addition the coarsest of the two parameters.

Theorem 3.1—If $(\alpha + \beta)/2 < d/4$ the minimax rate over $\mathcal{P}_1$ for estimating $\int abg\, d\nu$ is at least $n^{-2\alpha-2\beta/(2\alpha+2\beta+d)}$.

Proof: Let $H: \mathbb{R}^d \to \mathbb{R}$ be a C^∞ function supported on the cube $[0, 1/2]^d$ with $\int H\, d\nu = 0$ and $\int H^2\, d\nu = 1$. Let k be the integer closest to $n^{2d/(2\alpha+2\beta+d)}$ and let $\mathcal{X}_1, \ldots, \mathcal{X}_k$ be translates of the cube $k^{-1/d}[0, 1/2]^d$ that are disjoint and contained in $[0, 1]^d$. For $z_1, \ldots, z_k$ the bottom left corner of these cubes and $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda = \{-1, 1\}^k$, let

$$a_\lambda(z) = 2 + k^{-\alpha/d} \sum_{i=1}^{k} \lambda_i H((z - z_i)k^{1/d}),$$

$$b_\lambda(z) = 1/2 + k^{-\beta/d} \sum_{i=1}^{k} \lambda_i H((z - z_i)k^{1/d}).$$

These functions can be seen to be contained in $C^\alpha[0, 1]^d$ and $C^\beta[0, 1]^d$ with norms that are uniformly bounded in k. We choose a uniform prior π on λ, so that $\lambda_1, \ldots, \lambda_k$ are i.i.d. Rademacher variables.

We partition the sample space $\{0, 1\}\times\{0, 1\}\times\mathcal{X}$ into the sets $\{0, 1\}\times\{0, 1\}\times\mathcal{X}_j$ and the remaining set.

We parameterize the model by the triple (a, b, g). The likelihood can then be written as

$$(a-1)^{1-A(Z)}\left(b^{Y(Z)}(1-b)^{1-Y(Z)}\right)^{A}.$$

Because $\int H\, d\nu = 0$ the values of the functional $\int abg\, d\nu$ at the parameter values $(a_\lambda, 1/1, 1/2)$ and $(2, b_\lambda, 1/2)$ are both equal to 1/2, whereas the value at $(a_\lambda, b_\lambda, 1/2)$ is equal to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\int a_\lambda b_\lambda \frac{d\nu}{2} = \frac{1}{2} + \left(\frac{1}{k}\right)^{\alpha/d+\beta/d} \int \left(\sum_{i=1}^{k} H((z-z_i)k^{1/d}) \right)^2 \frac{d\nu}{2} = \frac{1}{2} + \frac{1}{2}\left(\frac{1}{k}\right)^{\alpha/d+\beta/d}.$$

Thus the minimax rate is not faster than $(1/k)^{\alpha/d+\beta/d}$ for $k = k_n$ such that the convex mixtures of the products of the perturbations do not separate completely as $n \to \infty$. We choose the mixtures differently in the cases $\alpha \le \beta$ and $\alpha \ge \beta$.

$\alpha \le \beta$. We define p_λ by the parameter $(a_\lambda, 1/2, 1/2)$ and q_λ by the parameter $(a_\lambda, b\lambda, 1/2)$. Because $\int a_\lambda \, d\pi(\lambda) = 2$ and $\int b_\lambda \, d\pi(\lambda) = 1/2$, we have

$$p(X) := \int p_\lambda(X) d\pi(\lambda) = (b^Y(Z)(1-b)^{1-Y}(Z))^A,$$

$$(p_\lambda - p)(X) = (1-A)(a_\lambda - 2)(Z),$$

$$(q_\lambda - p_\lambda)(X) = A(b_\lambda - 1/2)^Y (1/2 - b_\lambda)^{1-Y},$$

$$(q-p)(X) := \int (q_\lambda - p_\lambda)(X) d\pi(\lambda) = 0.$$

Therefore, it follows that the number d in Theorem 2.1 vanishes, while the numbers a and b are of the orders $k^{-2\alpha/d}$ and $k^{-2\beta/d}$ times

$$\max_j \int_{\mathcal{Z}_j} \left(\sum_{i=1}^{k} \lambda_i H((z-z_i)k^{1/d}) \right)^2 \frac{d\nu}{1/k} \sim 1,$$

respectively. Theorem 2.1 shows that

$$\rho\left(\int P_\lambda^n d\pi(\lambda), \int Q_\lambda^n d\pi(\lambda) \right) \ge 1 - C' n^2 \frac{1}{k}\left(k^{-4\beta/d} + k^{-2\alpha/d} k^{-2\beta/d} \right).$$

For $k \sim n^{2d/(2\alpha+2\beta+d)}$ the right side is bounded away from 0. Substitution of this number in the magnitude of separation $(1/k)^{\alpha/d+\beta/d}$ leads to the rate as claimed in the theorem.

$\alpha \ge \beta$. We define p_λ by the parameter $(2, b_\lambda, 1/2)$ and q_λ by the parameter $(a_\lambda, b_\lambda, 1/2)$. The computations are very similar to the ones in the case $\alpha \le \beta$.

3.2. Estimating an expected conditional covariance

Suppose that we observe n independent and identically distributed copies of $X = (Y, A, Z)$, where as in the previous section, Y and A are dichotomous, and Z takes its values in $\mathcal{Z} = [0,$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$1]^d$ with joint density given by f. Let $b(z) = P(Y = 1|Z = z)$ and $a(z) = P(A = 1|Z = z)$. We note that

$$\begin{aligned} b(Z) &= P(Y = 1|A = 1, Z)a(Z) + \{1 - a(Z)\}P(Y = 1|A = 0, Z) \\ &= \{P(Y = 1|A = 1, Z) - P(Y = 1|A = 0, Z)\}a(Z) + P(Y = 1|A = 0, Z) \\ &= \{P(Y = 1|A = 0, Z) - P(Y = 1|A = 1, Z)\}\{1 - a(Z)\} + P(Y = 1|A = 1, Z) \end{aligned}$$

so that by combining the last two equations above, we can write

$$P(Y = 1|A, Z) = \Delta(Z)\{A - a(Z)\} + b(Z)$$

where $\Delta(Z) = P(Y = 1|A = 1, Z) - P(Y = 1|A = 0, Z)$. This allows us to parametrize the density p of an observation by (Δ, a, b, f). The functional $\chi(p)$ is given by expected conditional covariance

$$E_f\{cov_{\Delta, p, b}(Y, A|Z)\} = E_{\Delta, p, b, f}(Y\,A) - \int abf d\nu \quad (3.2)$$

We consider the models

- $$\mathcal{B}_1 = \left\{(\Delta, a, b, f) : \Delta \text{ is unrestricted}, a \in C_M^{\alpha}[0, 1]^d, b \in C_M^{\beta}[0, 1]^d, f = 1, a, b \geq M\right\}$$
- $$\mathcal{B}_2 = \left\{(\Delta, a, b, f) : \Delta \text{ is unrestricted}, a \in C_M^{\alpha}[0, 1]^d, b \in C_M^{\beta}[0, 1]^d, f \in C_M^{\gamma}[0, 1]^d, a, b \geq M\right\}$$

We are mainly interested in the case $(\alpha + \beta)/2 < d/4$ when the rate of estimation of $\chi(p)$ becomes slower than $1/\sqrt{n}$. The paper [3] constructs and estimator that attains the rate $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$ uniformely over $\mathcal{B}_2$ if equation 3.1 of the previous section holds. We will show that this rate is optimal by showing that the minimax rate over the smaller model $\mathcal{B}_1$ is not faster than $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.

The first term of the difference on the right side of equation (3.2) can be estimated by the sample average $n^{-1}\sum_{i=1}^{n} Y_i A_i$ at rate $n^{-1/2}$. It follows that $\chi(p)$ can be estimated at the maximum of $n^{-1/2}$ and the rate of estimation of $\int abf d\nu$. In other words, to establish that the minimax rate for estimating $\chi(p)$ over $\mathcal{B}_1$ is $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$, we shall show that the minimax rate for estimating $\int abf d\nu$ over $\mathcal{B}_1$ is $n^{-(2\alpha+2\beta)/(2\alpha+2\beta+d)}$.

Theorem 3.2—If $(\alpha + \beta)/2 < d/4$ the minimax rate over $\mathcal{B}_1$ for estimating $\int abf d\nu$ is at least $n^{-2(\alpha+\beta)/(2\alpha+2\beta+d)}$.

Proof: Under the parametrization (Δ, a, b, f), the density of an observation X is given by

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$([\Delta(Z)\{A-a(Z)\}+b(Z)]a(Z))^{YA}\times([1-\Delta(Z)\{A-a(Z)\}-b(Z)]a(Z))^{(1-Y)A}$$
$$\times([\Delta(Z)\{A-a(Z)\}+b(Z)]\{1-a(Z)\})^{Y(1-A)}\times[\{1-\Delta(Z)\{A-a(Z)\}-b(Z)\}\{1-a(Z)\}]^{(1-Y)(1-A)}\times f(Z)$$

Suppose $\alpha < \beta$ and set

$$a_\lambda(z) = 1/2 + \delta a_\lambda(z) = 1/2 + k^{-\alpha/d}\sum_{i=1}^{k}\lambda_i H\left((z-z_i)k^{1/d}\right)$$

$$b_\lambda(z) = 1/2 + \delta b_\lambda(z) = 1/2 + k^{-\beta/d}\sum_{i=1}^{k}\lambda_i H\left((z-z_i)k^{1/d}\right)$$

$$\Delta_\lambda(Z) = \frac{-\delta b_\lambda(Z)}{1/2-\delta a_\lambda(Z)}$$

then at the parameters values (0, a_λ, 1/2, 1), $\int ab\,f\,d\upsilon = 1/4$ with a corresponding likelihood $p_\lambda = \{a_\lambda(Z)\}^A\times[\{1-a_\lambda(Z)\}]^{(1-A)}$, whereas at parameter values (Δ_λ, a_λ, b_λ, 1), $\int ab\,f\,d\upsilon = 1/4 + n^{-2(\alpha+\beta)/(d+2(\alpha+\beta))}$ and the likelihood is given by

$$q_\lambda(X) = \{a_\lambda(Z)/2\}^{YA}\times\{a_\lambda(Z)/2\}^{(1-Y)A}\times([1/2+\delta b_\lambda(Z)])^{Y(1-A)}([1/2-a_\lambda(Z)-\delta b_\lambda(Z)])^{(1-Y)(1-A)}$$

so that

$$(q_\lambda - p_\lambda)(X) = (1-A)\times\delta b_\lambda(Z)^Y\times\{-\delta b_\lambda(Z)\}^{(1-Y)}$$

And we conclude that $(q-p)(X) = \int(q_\lambda - p_\lambda)(X)\,d\pi(\lambda) = 0$. Furthermore

$$(p_\lambda - p)(X) = \{\delta a_\lambda(Z)\}^A\times[-\delta a_\lambda(Z)]^{(1-A)}$$

so that

$$\max_j\sup_\lambda\int_{\mathcal{X}_j}\frac{(q_\lambda-p_\lambda)^2}{p_\lambda}\frac{1}{P(\mathcal{X}_j)}d\upsilon = k^{-2\beta/d}\max_j\sup_\lambda\int_{\mathcal{X}_j}\frac{\left(\sum_{i=1}^{k}\lambda_i H((z-z_i)k^{1/d})\right)^2}{p_\lambda}\frac{1}{P(\mathcal{X}_j)}d\upsilon \lesssim k^{-2\beta/d},$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\sup_{\lambda}\int_{\mathcal{X}_j}\frac{\left(\int(q_\lambda-p_\lambda)d\pi(\lambda)\right)^2}{p_\lambda}\frac{1}{\mathrm{P}(\mathcal{X}_j)}d\upsilon=0$$

and

$$\max_j\sup_{\lambda}\int_{\mathcal{X}_j}\frac{(q_\lambda-p)^2}{p_\lambda}\frac{1}{\mathrm{P}(\mathcal{X}_j)}d\upsilon=k^{-2\alpha/d}\max_j\sup_{\lambda}\int_{\mathcal{X}_j}\frac{\left(\sum_{i=1}^{k}\lambda_i H((z-z_i)k^{1/d})\right)^2}{p_\lambda}\frac{1}{\mathrm{P}(\mathcal{X}_j)}d\upsilon\lesssim k^{-2\alpha/d}$$

Therefore, it follows that the number d of Theorem 2.1 vanishes, while the numbers a and b are of order $k^{-2\alpha/d}$ and $k^{-2\beta/d}$ respectively. Theorem 2.1. shows that

$$\rho\left(\int P_\lambda^n d\pi(\lambda),\int Q_\lambda^n d\pi(\lambda)\right)\geq 1-C''n^2\frac{1}{k}\left(k^{-4\beta/d}+k^{-2\beta/d}k^{-2\alpha/d}\right)$$

which gives the desired result for the choice of $k\sim n^{2d/(2\alpha+2\beta+d)}$.

Next, suppose $\alpha>\beta$, set $a_\lambda(Z)$ and $b_\lambda(Z)$ as above, and let

$$\Delta_\lambda(Z)=\frac{-\delta a_\lambda(Z)b_\lambda(Z)}{(1/2-\delta a_\lambda(Z))a(Z)}$$

then at the parameters values $(0, 1/2, b_\lambda, 1)$, $\int ab\,f\,d\upsilon = 1/4$ with corresponding likelihood

$$p_\lambda(X)=[(b_\lambda(Z))]^Y\times[(1-b_\lambda(Z))]^{(1-Y)}$$

whereas at parameter values $(0, p_\lambda, b_\lambda, 1)$, $\int ab\,f\,d\upsilon = 1/4 + n^{-2(\alpha+\beta)/(d+2(\alpha+\beta))}$ with corresponding likelihood given by

$$q_\lambda(X)=[b_\lambda(Z)/2]^Y\times([a_\lambda(Z)-b_\lambda(Z)/2])^{(1-Y)A}\times[1/2-\delta a_\lambda(Z)-b_\lambda(Z)/2]^{(1-Y)(1-A)}$$

so that

$$(q_\lambda-p_\lambda)(X)=(1-Y)\times\delta a(Z)^A\times[-\delta a_\lambda(Z)]^{(1-A)}$$

and we conclude that $(q-p)(X)=\int(q_\lambda-p_\lambda)(X)\,d\pi(\lambda)=0$. Furthermore

$$(p_\lambda-p)(X)=\delta b_\lambda(Z)^Y\times[-\delta b_\lambda(Z)]^{(1-Y)}$$

so that

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(q_\lambda - p_\lambda)^2}{p_\lambda} \frac{1}{\mathrm{P}(\mathcal{X}_j)} d\upsilon \lesssim k^{-2\alpha/d},$$

$$\sup_\lambda \int_{\mathcal{X}_j} \frac{(\int (q_\lambda - p_\lambda) d\pi(\lambda))^2}{p_\lambda} \frac{1}{\mathrm{P}(\mathcal{X}_j)} d\upsilon = 0$$

and

$$\max_j \sup_\lambda \int_{\mathcal{X}_j} \frac{(p_\lambda - p)^2}{p_\lambda} \frac{1}{\mathrm{P}(\mathcal{X}_j)} d\upsilon \lesssim k^{-2\beta/d}$$

which yields the desired result by arguments similar to the previous case.

References

1. Birge L, Massart P. Estimation of integral functionals of a density. Annals of statistics. 1995; 23:11–29.
2. van der Vaart, A. Asymptotic statistics. Cambridge University Press; 1998.
3. Robins J, Li L, Tchetgen E, van der Vaart A. Higher order inuence functions and minimax estimation of nonlinear functionals. IMS Lecture Notes-Monograph Series Probability and Statistics Models:Essays in Honor of David A. Freedman. 2008; 2:335–421.