Minimax rate of convergence for an estimator of the functional component in a semiparametric multivariate partially linear model

Michael Levine

Purdue University E-mail: mlevins@purdue.edu

Abstract:

A multivariate semiparametric partial linear model for both fixed and random design cases is considered. Earlier, in Brown, Levine and Wang (2014), the model has been analyzed using a difference sequence approach. In particular, the functional component has been estimated using a multivariate Nadaraya-Watson kernel smoother of the residuals of the linear fit. Moreover, this functional component estimator has been shown to be rate optimal if the Lipschitz smoothness index exceeds half the dimensionality of the functional component domain. In the current manuscript, we take this research further and show that, for both fixed and random designs, the rate achieved is the minimax rate under both risk at a point and the L_2 risk. The result is achieved by proving lower bounds on both pointwise risk and the L_2 risk of possible estimators of the functional component.

Key words and phrases: Multivariate semiparametric partial linear model, minimax rate of convergence, functional component, lower bound, Fano's lemma, Varshamov-Gilbert bound.

1. Introduction

Semiparametric models have a long history in statistics and have received considerable attention in the last several decades. The main reason they are of considerable interest is that, quite often, the relationships between the response and predictors are very heterogeneous in the same model. Some of the relationships are clearly linear whereas the detailed information about others is hard to come by. In many situations, a small subset of variables is presumed to have an unknown relationship with the response that is modeled nonparametrically while the rest are assumed to have a linear relationship with it. As an example, Engle, Granger, Rice and Weiss (1986) studied the nonlinear relationship between temperature and electricity usage where other related factors, such as income and price, are parameterized linearly.

The model we consider in this paper is a semiparametric partial linear multivariate model

$$Y_{\mathbf{i}} = a + X_{\mathbf{i}}^{'}\beta + f(U_{\mathbf{i}}) + \varepsilon_{\mathbf{i}} \tag{1.1}$$

where $X_{\mathbf{i}} \in \mathbb{R}^p$ and $U_{\mathbf{i}} \in \mathbb{R}^q$, β is an unknown $p \times 1$ vector of parameters, a is an unknown intercept term, $f(\cdot)$ is an unknown function and $\varepsilon_{\mathbf{i}}$ are independent and identically distributed random variables with mean 0 and constant variance σ^2. We consider two cases with respect to U: a random design case whereby $U_{\mathbf{i}}$ is a q-dimensional random variable and a fixed design case with $U_{\mathbf{i}}$ being a q-dimensional vector where each coordinate is defined on an equispaced grid on $[0, 1]$. In the fixed design case the errors are independent of $X_{\mathbf{i}}$ while in the random design case they are independent of $(X_{\mathbf{i}}^{'}, U_{\mathbf{i}})$. To obtain meaningful results, the function f is assumed to belong in the Lipschitz ball $\Lambda_\alpha(M)$ where α is the Lipschitz exponent. Of particular interest is the fact that, to be coherent, in the fixed design case when $q > 1$ the model (1.1) must have multivariate indices. The version with $q = 1$ was earlier considered in Wang, Brown and Cai (2011) while Brown, Levine and Wang (2014) considered the case of $q > 1$ in detail. The latter defined two conceptually similar difference based estimators of the parametric component for fixed and random design cases, respectively, and showed $\sqrt{n}$ asymptotic normality of both of these estimators. Moreover, it was also established that, in order for the estimator of the parametric component to be efficient, the order of the difference sequence must go to infinity. Brown, Levine and Wang (2014) also obtained a uniform over a Lipschitz ball $\Lambda^\alpha(M)$ convergence result for an estimator of the functional component, establishing the rate of convergence $n^{-2\alpha/(2\alpha+q)}$.

To the best of our knowledge, the optimal in the minimax sense rate of convergence for estimators of the nonparametric component of multivariate partial linear model (1.1) has not been established. Since results of Brown, Levine and

Wang (2014) amount to establishing the upper bound of that rate, the remaining task is to establish the lower bound. In this manuscript, we are doing just that for both fixed and random designs as well as for the two different functional distances. The first distance considered is the difference at a given fixed point and the second is that $L_2[0,1]^q$ distance. A number of different techniques are used to obtain these results.

Before proceeding, it is probably useful to recap quickly how the functional component estimator is constructed. The detailed discussion is available in Brown, Levine and Wang (2014). We only describe what happens in the fixed design case. We begin with (normalized) "diagonal" differences of observations $Y_{\mathbf{i}}$. As in Cai, Levine and Wang (2009) and Munk, Bissantz, Wagner and Freitag (2005), we select first a set of q-dimensional indices $J = \{(0,\dots,0),(1,\dots,1),\dots,(\gamma,\dots,\gamma)\}$. Some specialized notation is needed first to describe resulting differences. For any vector $u \in \mathbb{R}^q$, a real number v and a set $A \subset \mathbb{R}^q$, we define the the affine transformation of the set A is the set $B = u + vA = \{y \in \mathbb{R}^q : y = u + va, a \in A \subset \mathbb{R}^q\}$; then, we introduce a set R that consists of all indices $\mathbf{i} = (i_1,\dots,i_q)$ such that $R + J \equiv \{(\mathbf{i}+j)|\mathbf{i} \in R, j \in J\} \subset \{1,\dots,m\}^q$. Let a subset of $R+J$ corresponding to a specific $\mathbf{i} \in R$ be $\mathbf{i}+J$. In order to define a difference of observations of order γ, we define first a sequence of real numbers $\{d_j\}$ such that $\sum_{j=0}^{\gamma} d_j = 0$, and $\sum_{j=0}^{\gamma} d_j^2 = 1$ and $\sum_{j=0}^{\gamma} d_j j^k = 0$ for any power $k = 1,\dots,\gamma$. Moreover, denote $c_k = \sum_{i=0}^{\gamma-k} d_i d_{i+k}$. Then the difference of order γ "centered" around the point $Y_{\mathbf{i}}$, $\mathbf{i} \in R$ is defined as

$$D_{\mathbf{i}} = \sum_{j \in J} d_j Y_{\mathbf{i}+J} \tag{1.2}$$

Note that this particular choice of the set J makes numbering of difference coefficients d_j very convenient; since each q-dimensional index j consists of only identical scalars, that particular scalar can be thought of as a scalar index of d; thus, $\sum_{j\in J} d_j$ is the same as $\sum_{j=0}^{\gamma} d_j$ whenever needed. Now, let $Z_{\mathbf{i}} = \sum_{j\in J} d_j X_{\mathbf{i}+J}$, $\delta_{\mathbf{i}} = \sum_{j\in J} d_j f(U_{\mathbf{i}+J})$, and $\omega_{\mathbf{i}} = \sum_{j\in J} d_j \varepsilon_{\mathbf{i}+J}$, for any $\mathbf{i} \in R$. Then, by differencing the original model (2.1), one obtains $D_{\mathbf{i}} = Z_{\mathbf{i}}'\beta + \delta_{\mathbf{i}} + \omega_{\mathbf{i}}$ for all $\mathbf{i} \in R$. The ordinary least squares solution for β can then be written as

$$\hat{\beta} = argmin \sum_{\mathbf{i} \in R} (D_{\mathbf{i}} - Z_{\mathbf{i}}^{'}\beta)^2$$

In Brown, Levine and Wang (2014), the estimator of the nonparametric component f has been constructed in several stages. First, the vector coefficient β has been estimated as described above. Next, the intercept a has been estimated using the natural estimator $\hat{a} = \frac{1}{n}\sum_{\mathbf{i}\le n}(Y_{\mathbf{i}} - X_{\mathbf{i}}^{'}\hat{\beta})$. Finally, the multivariate Nadaraya-Watson kernel smoother has been applied to the residuals $r_{\mathbf{i}} = Y_{\mathbf{i}} - \hat{a} - X_{\mathbf{i}}^{'}\hat{\beta}$ from that fit to estimate the unknown function f. To construct the kernel smoother, one can, for example, select a univariate kernel function $K(U^l)$ for a specific coordinate U^l, $l = 1, \ldots, q$ such that $\int K(U^l)\, dU^l = 1$ and that has $\lfloor \alpha \rfloor$ vanishing moments. Next, one would usually chose an asymptotically optimal bandwidth $h = n^{-1/(2\alpha+q)}$ (see, for example, J. Fan and I. Gijbels (1995)), and define the univariate rescaled kernel as $K_h(U^l) = h^{-1}K(h^{-1}U^l)$. The q-dimensional product kernel is, then $K_h(U) = h^{-q}\prod_{l=1}^{q} K(h^{-1}U^l)$. Armed with this framework, the Nadaraya-Watson kernel weights can be defined as $W_{\mathbf{i},h}(U - U_{\mathbf{i}}) = \frac{K_h(U-U_{\mathbf{i}})}{\sum_{\mathbf{i}\le n} K_h(U-U_{\mathbf{i}})}$ Finally, the resulting kernel estimator of the function $f(U)$ can then be defined as

$$\hat{f}(U) = \sum_{\mathbf{i}\le n} W_{\mathbf{i},h}(U - U_{\mathbf{i}}) r_{\mathbf{i}}$$

Note that in the univariate case, Wang, Brown and Cai (2011) used the Gasser-Müller kernel to obtain the estimator of the functional component; for the multivariate case, Nadaraya-Watson estimator seems to be a better choice because it can be generalized easier to the multivariate case.

The next two sections present detailed results for the fixed and random design cases, respectively.

2. Optimal rates of convergence for the deterministic design case We consider the following semiparametric model

$$Y_{\mathbf{i}} = a + X_{\mathbf{i}}^{'}\beta + f(U_{\mathbf{i}}) + \varepsilon_{\mathbf{i}} \tag{2.1}$$

where $X_{\mathbf{i}} \in \mathbb{R}^p$, $U_{\mathbf{i}} \in S = [0,1]^q \subset \mathbb{R}^q$, $\varepsilon_{\mathbf{i}}$ are iid zero mean random variables with variance σ^2 and finite absolute moment of the order $\delta + 2$ for some small

$\delta > 0$, that is, $E\left|\varepsilon_{\mathbf{i}}\right|^{\delta+2} < \infty$. As noticed earlier in Brown, Levine and Wang (2014), the model (2.1) must have multidimensional indices $\mathbf{i} = (i_1, \ldots, i_q)^{'}$ to be coherent. Throughout this work, we will use bold font $\mathbf{i}$ to refer to multivariate indices and regular font to refer to coordinates of a multivariate index. For some positive integer m, we can take $i_k = 0, 1, \ldots, m$ for $k = 1, \ldots, q$; thus, the total sample size is $n = m^q$. Note that this assumption implies that $m = o(n)$ as $n \to \infty$. Due to the use of multivariate indices, one can also say that $\varepsilon_{\mathbf{i}}$ form an independent random field with the marginal density function $h(x)$ where x is a generic argument. We will say that two indices $\mathbf{i}^1 = (i_1^1, \ldots, i_q^1) \leq \mathbf{i}_2 = (i_1^2, \ldots, i_q^2)$ if $i_k^1 \leq i_k^2$ for any $k = 1, \ldots, q$; the relationship between $\mathbf{i}^1$ and $\mathbf{i}^2$ is that of partial ordering. Also, for a multivariate index $\mathbf{i}$, we denote $|\mathbf{i}| = |i_1| + \ldots + |i_q|$. In this section, we assume that $U_{\mathbf{i}}$ follows a fixed equispaced design: $U_{\mathbf{i}} = (u_{i_1}, \ldots, u_{i_q})^{'} \in \mathbb{R}^q$ where each coordinate is $u_{i_k} = \frac{i_k}{m}$. In the model (2.1), β is an unknown p-dimensional vector of parameters and a is an unknown intercept term. We assume that $X_{\mathbf{i}}$'s are independent random vectors that are also independent of $\varepsilon_{\mathbf{i}}$; moreover, we denote the non-singular covariance matrix of X Σ_X. For convenience, we also denote $N = \{1, \ldots, m\}^q$. This model requires an identifiability condition to be satisfied; more specifically, $\int_{[0,1]^q} f(u)du = 0$. The version of (2.1) with $q = 1$ has been considered earlier in Wang, Brown and Cai (2010). The case of $q = 1$ is quite different in that it only requires univariate indices for the model to be coherent. As a reminder, we consider functions f belonging to the Lipschitz ball $\Lambda^\alpha(M)$ for some positive constant M that is defined as follows. For a q-dimensional index $j = (j_1, \ldots, j_q)$, we define $j(l) = \{j : |j| = j_1 + \ldots + j_q = l\}$. Then, for any function $f : \mathbb{R}^q \to R$, $\frac{D^{j(l)} f}{\partial u_1^{j_1} \ldots \partial u_q^{j_q}}$ is defined for all j such that $|j| = l$. Then, the Lipschitz ball $\Lambda^\alpha(M)$ consists of all functions $f(u) : [0,1]^q \to R$ such that $|D^{j(l)} f(u)| \leq M$ for $l = 0, 1, \ldots, \lfloor\alpha\rfloor$ and $|D^{j(\lfloor\alpha\rfloor)} f(v) - D^{j(\lfloor\alpha\rfloor)} f(w)| \leq M||v - w||^{\alpha^{'}}$ with $\alpha^{'} = \alpha - \lfloor\alpha\rfloor$. Here and in the future, $||\cdot||$ stands for the regular l_2 norm in $\mathbb{R}^q$. Brown, Levine and Wang (2014) established a uniform upper bound on the risk at a point for the difference based estimator $\hat{f}$ of the nonparametric component f. More specifically, they proved that, for any Lipschitz indicator $\alpha > 0$ and any $U_0 \in [0,1]^q$, the estimator $\hat{f}$ satisfies

$$\sup_{f \in \Lambda^\alpha(M)} \mathbb{E}[(\hat{f}(U_0) - f(U_0))^2] \leq C n^{-2\alpha/(2\alpha+q)}$$

for a constant $C > 0$. The following result also establishes the lower bound on the risk of that estimator, therefore proving that $n^{-\alpha/(2\alpha+q)}$ is the minimax rate of convergence. Since this result requires that the difference-based estimator of β be asymptotically normal, the assumptions from Theorem (2.1) are still needed. However, they are not sufficient to obtain the lower bound. One extra assumption still has to be included in the statement of the Theorem (2.2) to obtain the lower bound. Also, the lower bound proof in this context requires the use of the so-called Varshamov-Gilbert Lemma that has been known for a long time in information theory (see, e.g. Gilbert (1952) as well as Ibragimov and Hasminskii (1977)). For convenience, we give the full text of this important result.

Lemma 2.1. *Choose a positive integer $m \geq 8$. Let Ω be a set of all binary sequences of length m; clearly, the cardinality of this set is 2^m. There exists a subset $\{\omega^{(0)}, \ldots, \omega^{(J)}\}$ of Ω such that $\omega^{(0)} = (0, \ldots, 0)$, and $\rho((\omega^{(j)}, \omega^{(k)}) \geq \frac{m}{8}$ for any $0 \leq j \leq k \leq J$; moreover, $J \geq 2^{m/8}$.*

The Varshamov-Gilbert lower bound essentially implies that one can select a "large" subset of binary sequences of length m from set Ω in such a way that the Hamming distance between any two sequences from this subset can be guaranteed to be no less than $\frac{m}{8}$. In this context, "large" means that the the size of such a subset can be guaranteed to be no less than $2^{m/8}$.

Theorem 2.2. *Let $\varepsilon_{\mathbf{i}}$ be independent identically distributed random variables with zero mean and finite variance σ^2; moreover, we assume that, for some small $\delta > 0$, $\mathbb{E}\varepsilon_{\mathbf{i}}^{2+\delta} < \infty$. We also assume that the marginal density function $h(x)$ of the field $\varepsilon_{\mathbf{i}}$ must have a bounded variation over the real line. Next, a difference sequence used d_j is of order $\gamma \geq \lfloor \alpha \rfloor$ and such that $\sum_{j=0}^{\gamma} d_j = 0$, $\sum_{j=0}^{\gamma} d_j^2 = 1$, $\sum_{j=0}^{\gamma} d_j j^k = 0$ for $k = 1, \ldots, \gamma$. Finally, we assume that the marginal density function of observations Y_i $p(Y_{\mathbf{i}}) = \int p_\varepsilon(Y_{\mathbf{i}} - X_{\mathbf{i}}'\beta)\, dX_{\mathbf{i}}$ satisfies the following assumption: there exists $p_* > 0$, $v_0 > 0$ such that*

$$\int p(U) \log \frac{p(U)}{p(U+V)}\, dU \leq p_* v^2 \text{ for all } ||V|| < v_0 \tag{2.2}$$

Then, the convergence rate $n^{-\alpha/(2\alpha+q)}$ is optimal

1. *on* $(\Lambda_\alpha(M), d_0)$ *where* d_0 *is the distance at a fixed point* $U_0 \in [0,1]^q$ *and*

2. *on* $(\Lambda_\alpha(M), ||\cdot||_2)$ *where* $||\cdot||_2$ *is an* L_2 *distance on* $[0,1]^q$.

Remark 2.3. *Note that the condition* (2.2) *is the one that is new here. It is of a fairly general nature; in particular, a standard Gaussian density satisfies this condition. For more information on this condition, see Tsybakov (2009).*

Proof. 1. To obtain the minimax convergence rate for the first case, we will use a two-point argument going all the way back to Ibragimov and Has'minskii (1977) and Has'minskii (1978). First, we need to define an appropriate "test" function. We will use the optimal bandwidth $h_n = n^{-\frac{1}{2\alpha+q}}$ in each dimension $j = 1, \ldots, q$ and a bandwidth matrix $H_n = \text{diag}\{h_n\} = h_n I_{q\times q}$ for the estimator of function $f(U)$. Next, consider a non-negative function $K \in \Lambda^\alpha\left(\frac{1}{2}\right)$ such that $K(U) > 0$ if and only if $||U|| \leq \frac{1}{2}$. As an example of such function, we can select $K(U) = a \exp\left(-\frac{1}{1-||U||^2}\right) I\left(||U|| \leq \frac{1}{2}\right)$ for a sufficiently small $a > 0$. Note that this function reaches its maximum when $||U|| = 0$ and that this maximum is $\exp(-1)$. A pair of functions that we will need for our problem are $f_{0,n}(U) \equiv 0$ and $f_{1,n}(U) = M h_n^\alpha K(H_n^{-1}(U - U_0)) = M h_n^\alpha K(h_n^{-1}(U - U_0))$. The following three conditions must be checked for us to conclude that the lower bound is achieved at the convergence rate $n^{-\frac{\alpha}{2\alpha+q}}$.

- The first condition is that $f_{j,n} \in \Lambda^\alpha(M)$ for $j = 0, 1$ and sufficiently large n. For $f_{0,n} \equiv 0$ it is clearly satisfied immediately. Next, for any $l = 0, 1, \ldots, \lfloor\alpha\rfloor$, $|D^{j(l)} f_{1,n}(U)| = M h_n^{\alpha-l} |D^{j(l)} K(h_n^{-1}(U - U_0))| \leq M$ for sufficiently large n since the function $K \in \Lambda^\alpha\left(\frac{1}{2}\right)$. Finally, since $|D^{(j(\lfloor\alpha\rfloor))} f_{1,n}(V) - D^{(j(\lfloor\alpha\rfloor))} f_{1,n}(W)| \leq h_n^{\alpha-\lfloor\alpha\rfloor} |D^{(j(\lfloor\alpha\rfloor))} K(h_n^{-1}(V - U_0)) - D^{(j(\lfloor\alpha\rfloor))} K(h_n^{-1}(W - U_0))| \leq M ||V - W||^{\alpha'}$ for all sufficiently large n, we can say that $f_{1,n} \in \Lambda^\alpha(M)$

- Next, we need to check that, for sufficiently large n, the distance between the two "test" functions $d(f_{1,n}(U_0), f_{0,n}(U_0)) \geq n^{-\alpha/(2\alpha+q)}$. Indeed, $|f_{1,n}(U_0)| = |h_n^\alpha K(0)| \geq 2An^{-\frac{\alpha}{2\alpha+q}}$ for e.g. $A = \frac{M}{4}\exp(-1)$. Therefore, the distance between the two "points" in the Lipschitz ball $\Lambda^\alpha(M)$ that we selected is, indeed, of the right order $n^{-\frac{\alpha}{2\alpha+q}}$.

- Finally, let us define two product densities $m_{0,n}$ and $m_{1,n}$ that are densities of observations generated by (2.1) when the functional component is equal to $f_{0,n}$ and $f_{1,n}$, respectively. Using assumption (2.2), the Kullback distance between the two is

$$
\begin{aligned}
&K(m_{0,n}, m_{1,n}) = \int \cdots \int \log \prod_{\mathbf{i} \le n} \frac{p(Y_{\mathbf{i}})}{p(Y_{\mathbf{i}} - f_{1,n}(U_{\mathbf{i}}))} \prod_{\mathbf{i} \le n} [p(Y_{\mathbf{i}})\, dY_{\mathbf{i}}] \\
&= \sum_{\mathbf{i} \le n} \int p(Y) \log \frac{p(Y)}{p(Y - f_{1,n}(U_{\mathbf{i}}))}\, dY \le \sum_{\mathbf{i} \le n} p^* f_{1,n}^2(U_{\mathbf{i}}) \\
&= M p_* h_n^{2\alpha} \exp(-2) \sum_{\mathbf{i} \le n} I\left(||U_i - U_0|| \le \frac{h_n}{2} \right) \\
&\le p_* h_n^{2\alpha} \exp(-2) \max((M/2) n h_n^q, 1) \\
&\le (M/2) p_* h_n^{2\alpha} \exp(-2) n h_n^q = (M/2) p_* \exp(-2) n h_n^{2\alpha+q} \le C
\end{aligned}
$$

for sufficiently large n where $C = (M/2)p_* \exp(-2)$. This establishes the optimality of the rate $n^{-\alpha/(2\alpha+q)}$.

2. Next, we establish the minimax rate of convergence for the $L_2[0,1]^q$ risk. First, recall the earlier result that, for any $\alpha > 0$,

$$
\sup_{f \in \Lambda^\alpha(M)} \mathbb{E} \left[\int_{[0,1]^q} (\hat{f}(U) - f(U))^2\, dU \right] \le C n^{-2\alpha/(2\alpha+q)}
$$

We will argue that the rate $n^{-\alpha/(2\alpha+q)}$ is also the minimax rate under the $L_2[0,1]^q$ loss. In what follows, we denote $\lceil x \rceil$ the smallest integer that is larger than $x \in \mathbb{R}$. First, let us define $m = \lceil c_0 n^{\frac{q}{2\alpha+q}} \rceil$ where $c_0 > 0$ is some real number. As a second step, we choose the bandwidth $h_n = m^{-1/q}$. Our next purpose is to define a partition of $[0,1]^q$ into a set of disjoint subsets and define a sequence of functions that take non-zero values on just one of these subsets. Such a multivariate partition with $\mathbb{R}^q$-valued partition points (vectors) $U_k = (u_k^1, \ldots, u_k^q)'$ can be defined by selecting $u_k^j = \frac{k-\frac{1}{2}}{m}$, for $k = 1, \ldots, m$. Now, denote $\Delta_k = \left\{ [\frac{k-1}{m}, \frac{k}{m}), \ldots, [\frac{k-1}{m}, \frac{k}{m}) \right\}' \in \mathbb{R}^q$; note that the entire $[0,1]^q = \cup_k \Delta_k$ and that $\Delta_k \cap \Delta_{k'} = \emptyset$ if $k \ne k'$, that is Δ_ks are disjoint. The next step consists of selecting hypotheses based on a function $K(U) : [0,1]^q \to \mathbb{R}$ such that $K(U) \in \Lambda^\alpha\left(\frac{1}{2}\right)$ and $K(U) > 0$ if and only if $||U|| < \frac{1}{2}$. As before, we select the function $K(U) = \exp\left(-\frac{1}{1-||U||^2}\right) I\left(||U|| \le \frac{1}{2}\right)$.

Also, denote $||K||_2$ the $L_2[0,1]^q$ norm of the function K. To simplify the notation, let the diagonal bandwidth matrix be $H = \text{diag}\; h_n = h_n I_{q\times q}$. Now, we can define a set of m functions $\Phi_k(U) = Mh_n^{\alpha}K(H^{-1}(U-U_k))$, for $k=1,\ldots,m$. Finally, denote the set of all binary sequences of length m $\Omega = \{\omega = (\omega_1,\ldots,\omega_m), \omega_i \in (0,1)\} = \{0,1\}^m$. Then, the "test functions" $f_{j,n}$, $j=0,\ldots,J$ will be selected from the set of functions

$$\mathbb{E} = \{f_\omega(u) = \sum_{k=1}^{m} \omega_k \Phi_k(u), \omega \in \Omega\} \tag{2.3}$$

.

The following three conditions now must be verified to ensure that $n^{-\alpha/(2\alpha+q)}$ is, indeed, the minimax rate.

(a) First, we need to show that the $L_2[0,1]^q$ distance between any two of the "test functions" is bounded below by the multiple of $n^{-\alpha/(2\alpha+q)}$. For any two functions $f_\omega, f_{\omega'} \in \mathbb{E}$, the $L_2[0,1]^q$-distance between them is

$$d(f_\omega, f_{\omega'}) = \sqrt{\int_{[0,1]^q} \left[f_\omega(U) - f_{\omega'}(U)\right]^2 dU}$$

$$= \sqrt{\int_{[0,1]^q} \left[\sum_{k=1}^{m}(\omega_k - \omega_k^{'})\Phi_k(U)\right]^2 dU}$$

$$= \sqrt{\sum_{k=1}^{m}(\omega_k - \omega_k^{'})^2 \int_{\Delta_k} \Phi_k^2(U)\, dU} = Mh_n^{\alpha+\frac{q}{2}}||K||_2\sqrt{\rho(\omega,\omega')}$$

where $\rho(\omega,\omega^{'}) = \sum_{k=1}^{q} I(\omega_k \neq \omega_k^{'})$ is the Hamming distance between ω and $\omega^{'}$. Now we need to use the Lemma(2.1). In our context, it suffices to choose the ω and $\omega^{'}$ such that $\sqrt{\rho(\omega,\omega')} \asymp h_n^{-q/2}$ which is equivalent to $\rho(\omega,\omega^{'}) \asymp m$. In other words, to show that the rate $n^{-\alpha/(2\alpha+q)}$ is, indeed, the minimax rate of convergence, we need to use an *infinite* number of "testing hypotheses" J. It is now easy to verify that, for a sufficiently large n,

$$d(f_\omega, f_{\omega'}) \geq M h_n^{\alpha+\frac{q}{2}} ||K||_2 \sqrt{\frac{m}{16}} \tag{2.4}$$
$$= \frac{M}{4}||K||_2 h_n^\alpha = \frac{M}{4}||K||_2 n^{-\frac{\alpha}{2\alpha+q}}$$

and so the rate is correct

(b) Clearly, each $\Phi_k(U) \in \Lambda^\alpha(M)$; since each $\omega_k \leq 1$ and the functions $\Phi_k(U)$ have disjoint supports for different k, $f_\omega \in \Lambda^\alpha(M)$.

(c) Finally, we also need to verify that the average Kullback-Leibler distance between the null hypothesis and others is bounded from above as follows: $\frac{1}{J}\sum_{j=1}^J K(f_{0,n}, f_{j,n}) \leq \alpha \log J$. Indeed, proceeding as before in the case of pointwise risk, one can find that

$$K(f_{0,n}, f_{j,n}) \leq p_* \sum_{i\leq n} f_{j,n}^2(U_i) \leq p_* \sum_{k=1}^m \sum_{i:U_i \in \Delta_K} \Phi_k^2(U_i)$$
$$\leq p_* M^2 K_{\max}^2 h_n^{2\alpha} \sum_{k=1}^m \{\# : U_i \in \Delta_k\}$$
$$\leq p_* M^2 K_{\max}^2 n h_n^{2\alpha} \leq p_* M^2 K_{\max}^2 c_0^{(-2\alpha+q)/q} m$$

for a sufficiently large n. Since Varshamov-Gilbert result suggests that $m \leq 8\log M/\log 2$, the claim is, indeed, true if we select $c_0 = \left(\frac{8p_* L^2 K_{\max}^2}{\alpha \log 2}\right)^{\frac{q}{2\alpha+q}}$.

□

3. Optimal rates of convergence for the random design case

In the same way as in Wang, Brown and Cai (2011), our next step is to obtain minimax convergence rates in the case of random design. For convenience purposes, we restate the assumptions of that case. Our model is again

$$Y_i = a + X_i^{'}\beta + f(U_i) + \varepsilon_i \tag{3.1}$$

for $i = 1, \ldots, n$; we also assume that U_i are random variables on $[0,1]^q$ and that $(X_i^{'}, U_i) \in \mathbb{R}^p \times \mathbb{R}^q$ are independent with an unknown joint density $g(x,u)$. The random errors ε_i are independent identically distributed with mean zero, variance σ^2 and are independent of (X_i', U_i). Moreover, we assume that the conditional covariance matrix $\Sigma_* = \mathbb{E}[(X_1 - \mathbb{E}(X_1|U_1))(X_1 - \mathbb{E}(X_1|U_1))']$ is nonsingular. As in any linear regression model, $\beta \in \mathbb{R}^p$ is a vector of coefficients. For any U with the marginal distribution $g(u)$, we also need to assume that $\mathbb{E}(f(U)) \equiv \int f(u)g(u)\,du = 0$ to ensure identifiability of the model (3.1). Finally, an individual coordinate of the vector X_i is denoted X_i^l, for $l = 1, \ldots, p$ and an individual coordinate of the random vector U is denoted U^r, for $r = 1, \ldots, q$. Note that, unlike the fixed design case, the indices i here are univariate.

As a first step, we obtain least squares estimates of the coefficient vector $\hat{\beta}$ and the intercept $\hat{a}$. Note that, unlike in the fixed design case, the nearest neighbor type approach has to be used because it is impossible to arrange the points U_i in a meaningful order while keeping them in a small neighborhood of the point U where the function has to be estimated. In other words, only the points U_i such that the Euclidean norm $||U_i - U||^2 \le \varepsilon$ for some small $\varepsilon > 0$ are considered. Let the number of these points be $\gamma_i(\varepsilon)$; clearly, this number depends on the choice of ε as well as on the marginal distribution of U_i. Then, a difference "centered" on the point U_i will be $\delta_i = \sum_{t=1}^{\gamma_i(\varepsilon)} d_t f(U_{i+t})$. Note that, as opposed to the fixed design case, the difference sequence considered here is of a *variable* order that depends on the value of the marginal density function $g(U_i)$ at which the function f is to be estimated as well as the "tuning" parameter ε. From this point on, the estimation of β proceeds exactly as in the fixed design case.

The asymptotic normality and efficiency of the estimator $\hat{\beta}$ in the random design case were established in the Theorem 3.5 of Brown, Levine and Wang (2014). To estimate the function $f(U)$, we apply a multivariate kernel smoother to the residuals $r_i = Y_i - \hat{a} - X_i'\hat{\beta}$ in the same way as it was done in the fixed design case. As in the fixed design case, we use again the multivariate version of the Nadaraya-Watson estimator that uses a product kernel. The Nadaraya-Watson estimator of $f(U)$ is, then

$$\hat{f}_n(U) = \sum_{i=1}^{n} W_{i,h}(U - U_i) r_i$$

where the weights $W_{i,h}(U - U_i)$ are the multivariate Nadaraya-Watson weights. We stress the dependence of this estimator on the sample size n by using it as a subscript. To make the notation shorter, we will also use $||\cdot||_2$ to denote the $L_2[0,1]^q$ - norm and $||\cdot||_2^2$ the squared norm in the same space. As a first step, we need to establish the analogue of Theorem 2.4 from Brown, Levine and Wang (2014) in the case of random design. Since the proof of that result is almost analogous to Theorem 2.4, we omit it and only state the final result.

Theorem 3.4. *Let the marginal density function of U_i $g(u)$ be bounded everywhere on $\mathbb{R}^q$. Also, let the function $f(U)\in\Lambda^\alpha(M_f)$ and $h(U)\equiv\mathbb{E}(X|U)\in\Lambda^\rho(M_h)$. Define the difference based estimator of β as described above with $\varepsilon\to 0$ as $n\to\infty$; the "nearest neighbor distance" ε is selected in such a way that $o(n)\varepsilon^{2(\rho+\alpha)}\to 0$ when $n\to\infty$. Then, for any Lipschitz indicator $\alpha>0$ and any $U_0\in[0,1]^q$, the estimator $\hat{f}_n$ satisfies*

$$\sup_{f\in\Lambda^\alpha(M)}\mathbb{E}[(\hat{f}_n(U_0)-f(U_0))^2]\le Cn^{-2\alpha/(2\alpha+q)}$$

for a constant $C>0$. Also, for any $\alpha>0$,

$$\sup_{f\in\Lambda^\alpha(M)}\mathbb{E}\left[\int_{[0,1]^q}(\hat{f}_n(U)-f(U))^2\,dU\right]\le Cn^{-2\alpha/(2\alpha+q)}$$

Theorem (3.4) establishes upper bounds on the rate of convergence for the distance at a point and the $L_2[0,1]^q$ distance. In order to obtain the optimality of this convergence rate, we need to match these upper bounds with lower bounds. To characterize these lower bounds, we need a result from the information theory generally known as Fano's Lemma (see e.g. Fano (1952)). To make the exposition easier to follow, we describe this result in full. First, let $P_0,P_1,\dots,P_M$ be probability measures on some measurable probability space. Suppose we have a test ψ that can differentiate between $\{0,1,\dots,M\}$ based on a given outcome. Based on such a test, define the average probability of error and the minimum average probability of error by

$$\bar{p}_{e,M}(\psi)=\frac{1}{M+1}\sum_{j=0}^{M}P_j(\psi\ne j)$$

and

$$\bar{p}_{e,M}=\inf_{\psi}\bar{p}_{e,M}(\psi),$$

respectively. Also, define the average probability measure $\bar{P} = \frac{1}{M+1}\sum_{j=0}^{M} P_j$. For any $0 \le x \le 1$, we define the function $g(x) = x\log M + \mathcal{H}(x)$ where $\mathcal{H}(x) = -x\log x - (1-x)\log(1-x)$. Now we are ready to introduce Fano's Lemma.

Lemma 3.5. *Let $P_0, P_1, \dots, P_M$ be a set of probability measures for some $M \ge 1$. Then, $\bar{p}_{e,M} \le \frac{M}{M+1}$ and, moreover,*

$$g(\bar{p}_{e,M}) \ge \log(M+1) - \frac{1}{M+1}\sum_{j=0}^{M} K(P_j, \bar{P})$$

Now, we are ready to state and prove the main result of this section.

Theorem 3.6. *1. Let T_n be an arbitrary estimator of the function f. For any Lipschitz indicator $\alpha > 0$ and any $U_0 \in [0,1]^q$, the following holds:*

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Lambda^\alpha(M)} \mathbb{E}_f \left[n^{\frac{2\alpha}{2\alpha+q}} (T_n(U_0) - f(U_0))^2 \right] \ge c_1$$

where c_1 is a constant that does not depend on n. In other words, $n^{-\alpha/(2\alpha+q)}$ is an optimal (minimax) rate of convergence when estimating the function $f(U)$ at a given fixed point U_0.

2. Again, let T_n be an arbitrary estimator of the function f. For any Lipschitz indicator $\alpha > 0$,

$$\liminf_{n\to\infty} \inf_{T_n} \sup_{f\in\Lambda^\alpha(M)} \mathbb{E}_f \left[\int_{[0,1]^q} n^{\frac{2\alpha}{2\alpha+q}} (T_n(U) - f(U))^2 \, dU \right] \ge c_2$$

where c_2 is a constant that doesn't depend on n.

Proof. 1. As a first step, we will consider the case of estimation at a point $U_0 \in [0,1]^q$. The proposed minimax rate is $\psi_n = n^{-\alpha/(2\alpha+q)}$. The subscript n is used to stress its dependence on the sample size n. We also define two test functions $f_{0,n}(U) \equiv 0$ and $f_{1,n}(U) = Mh_n^\alpha K(h_n^{-1}(U-U_0))$ that are exactly the same as those used in the proof of the first part of the theorem (2.2). Recall that the distance at a point between these functions is $|f_{1n}(U_0)| \ge 2A\psi_n$ where the convergence rate $\psi_n = n^{-\alpha/(2\alpha+q})$ and $A = \frac{M}{4}e^{-1}$. Denote $\mathbb{E}_{U_1,\dots,U_n}$ the conditional expectation with respect to the joint distribution of $U_1, \dots, U_n$. Also, denote the "test" function Ψ that, for a given set of data, selects either the first or the second function $f_{0,n}$ or $f_{1,n}$. Then, using Chebyshev's inequality, we have

$$\sup_{f\in\Lambda_\alpha(M)} \mathbb{E}[\psi_n[T_n(U_0)-f(U_0)]]^2 \geq A^2 \max_{f\in f_0,f_{1n}} P(|T_n(U_0)-f(U_0)| \geq A\psi_n)$$

$$\geq \frac{A^2}{2}\sum_{j=0}^{1}\mathbb{E}_{U_1,\ldots,U_n}\left[P(|T_n(U_0)-f(U_0 \geq A\psi_n|U_1,\ldots,U_n)\right]$$

$$= A^2\mathbb{E}_{U_1,\ldots,U_n}\left[\frac{1}{2}\sum_{j=0}^{1}P\left(|T_n(U_0)-f(U_0)| \geq A\psi_n|U_1,\ldots,U_n\right)\right]$$

$$A^2\mathbb{E}_{U_1,\ldots,U_n}\left[\inf_{\Psi}\frac{1}{2}\sum_{j=0}^{1}P((\Psi\neq j|U_1,\ldots,U_n)\right]$$

where the last inequality follows from the triangle inequality and the fact that the distance between the two functions at the point U_0 is greater than or equal to $2A\psi_n$. For fixed $U_1,\ldots,U_n$ we have the distance between the two product densities $m_{0,n}$ and $m_{1,n}$ associated with functions $f_{0,n}$ and $f_{1,n}$ $K(P_0,P_1) \leq C$ for the same finite C that was obtained in the proof of theorem (2.2). Thus, the lower bound becomes $\bar{p}_{e,1} = \inf_\Psi P((\Psi\neq j|U_1,\ldots,U_n) \geq \max\left(\frac{1}{4}\exp(-C),\frac{1-\sqrt{C/2}}{2}\right)$ according to the Theorem 2.2 of Tsybakov (2009) which finishes our proof.

2. Next, we need to consider the $L_2[0,1]^q$ case. Here, yet again, we will need not just two, but $J+1$ "test" functions with $J\to\infty$ as $n\to\infty$. The needed "test" functions are defined as f_{jn}, $j=0,\ldots,J$ where f_{jn} are again selected from the function set (2.3) as before in the case of fixed design. Earlier, we proved that the L_2 distance between any two functions from this set is

$$||f_{jn}-f_{kn}||_2 \geq \frac{M}{4}||K||_2 n^{-\alpha/(2\alpha+q)}$$

for any $0\leq j,k\leq J$. The basic idea we are going to use is to bound the $L_2[0,1]^q$ risk from below by the average probability of error rather than the minimax probability of error. This approach uses the Fano's Lemma (3.5) that we described earlier. Again, using Chebyshev's inequality, we obtain

$$\begin{aligned}
&\sup_{f\in\Lambda^{\alpha}(M)} \mathbb{E}\left[n^{2\alpha/(2\alpha+q)}||\hat{f}-f||_2^2\right]\\
&\ge A^2 \max_{f\in\{f_{0n},\dots,f_{Jn}\}} \mathbb{P}_f\left(||\hat{f}_n-f||_2\ge A\psi_n\right)\\
&\ge A^2\frac{1}{J+1}\sum_{j=0}^{J}\mathbb{E}_{U_1,\dots,U_n}\left[\mathbb{P}_j\left(||\hat{f}_n-f||_2\ge A\psi_n|U_1,\dots,U_n\right)\right]\\
&= A^2\mathbb{E}_{U_1,\dots,U_n}\left[\frac{1}{J+1}\sum_{j=0}^{J}\mathbb{P}_j\left(||\hat{f}_n-f||_2\ge A\psi_n|U_1,\dots,U_n\right)\right]
\end{aligned}$$

Next, let Ψ be a test that based on the data, makes a choice between the $J+1$ hypotheses considered. Then, the above result means that

$$\begin{aligned}
&\sup_{f\in\Lambda^{\alpha}(M)} \mathbb{E}\left[n^{2\alpha/(2\alpha+q)}||\hat{f}-f||_2^2\right]\\
&\ge A^2\mathbb{E}_{U_1,\dots,U_n}\left[\inf_{\Psi}\frac{1}{J+1}\sum_{j=0}^{J}\mathbb{P}_j(\Psi\ne j|U_1,\dots,U_n)\right]
\end{aligned}$$

While proving the theorem (2.2), we showed that $\frac{1}{J}\sum_{j=1}^{J}K(f_{0,n},f_{j,n})\le \alpha\log J$. A direct consequence of Fano's lemma (for details see, for example, Corollary 2.6 in Tsybakov (2009) is that the minimum average probability of error is bounded from below as

$$\inf_{\Psi}\frac{1}{J+1}\sum_{j=0}^{J}\mathbb{P}_j(\Psi\ne j|U_1,\dots,U_n)\ge\frac{\log(J+1)-\log 2}{\log J}-\alpha \qquad (3.2)$$

The right hand side of the last inequality does not depend on $U_1,\dots,U_n$ and so the desired result is obtained.

□

4. Acknowledgements

Michael Levine's research has been partially supported by the NSF grant DMS # 1208994

References

Brown, L.D., Levine,M., and Wang, L. (2014) A semiparametric multivariate partially linear model: a difference approach. A Technical Report at http://www.stat.purdue.edu/research/technical_reports/2014-tr.html

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *Journal of the American Statistical Association*, **81(394)**, 310-320.

Ibragimov, I.A. and Has'minskii, R.Z. (1977) On the estimation of an infinite-dimensional parameter in Gaussian white noise. *Soviet Mathematics. Doklady*, **18**, 1307-1309

Has'minskii, R.Z. (1978) A lower bound on the risks of nonparametric estimates of densities in the uniform metric *Theory of Probability and Its Applications*, **23**, 794-798

Fan, J., and Gijbels, I. (1995) Local polynomial modelling and its applications: Monographs on Statistics and Applied Probability, Chapman& Hall, London

Fano, R.M. (1952) Class Notes for Transmission of Information. Course 6.574, MIT, Cambridge, Massachusetts

Gilbert (1952) A comparison of signalling alphabets. *Bell System Technical Journal*, **31**, 504-522

Tsybakov, A.B. (2009) Introduction to Nonparametric Estimation, Springer

Wang, L., Brown, L.D. and Cai, Tony T. (2011). A difference based approach to the semiparametric partial linear model. *Electronic Journal of Statistics*, **5**, 619-641.