Vitesses de convergence et estimation adaptative sous les hypothèses single- et multi-index

Nora Serdyukova

Departamento de Estadística, Facultad de Ciencias Físicas y Matemáticas
Universidad de Concepción
Avda. Esteban Iturra s/n - Barrio Universitario, Concepción, Región VIII, CHILE.
E-mail : Nora.Serdyukova@gmail.com

Résumé. On présente les résultats d'estimation adaptative d'une fonction multidimensionnelle sous les hypothèses structurelles. Dans un premier temps, on suppose que la fonction à estimer possède la structure « single-index ». On propose une procédure s'appuyant sur l'idée de la méthode de Lepski, qui s'adapte simultanément à l'indice inconnu ainsi qu'à la régularité de la fonction de lien. Pour des pertes ponctuelles on montre la borne supérieure du risque maximal au cas où la fonction de lien appartient à l'échelle des espaces de Hölder. D'après la borne inférieure obtenue pour le risque minimax l'estimateur construit est un estimateur adaptatif optimal sur l'ensemble de classes considérées. Ensuit, on parle du modèle de « multi-index » anisotrope pour lequel on présente une borne inférieure pour le risque minimax sur l'échelle d'espaces de Hölder anisotropes.

Mots-clés. Estimation adaptative, Borne inférieure, Vitesse minimax, Modèle de single-index, multi-index, Adaptation structurelle.

Abstract. New results about adaptive estimating a multivariate function under structural constraints are presented. First, one supposes that the function to be estimated possesses the structure "single-index". A procedure developing the idea of Lepski's method, that adapts simultaneously to the unknown index vector and the regularity of the link function is proposed. For pointwise losses an upper bound for the maximal risk when the link function belongs to a collection of the Hölder classes is obtained. The lower bound on the minimax risk demonstrates that the proposed estimator is rate optimal adaptive over the considered classes of functions. Next, one discusses the anisotropic multi-index model for which lower bounds on the minimax risk over the classes of structured anisotropic functions are provided.

Keywords. Adaptive estimation, Lower bounds, Minimax rate, Multi-index model, Single-index, Structural adaptation.

Introduction. Consider a function $F : \mathbb{R}^d \to \mathbb{R}$ which has the single-index (ridge) structure : there exist a link function $f : \mathbb{R} \to \mathbb{R}$ and an index vector $\theta \in \mathbb{S}^{d-1}$ such that

$$F(x) = f(\theta^\top x). \tag{1}$$

In statistical modeling this type of constraint (single-index model) appears in semiparametric estimation [see, for instance, Horowitz (1998)] as a natural relaxation of the generalized linear models [see McCullagh and Nelder (1989)] and implies that both the link function and the index vector are unknown.

During the last two decades, numerous techniques for estimating in the single-index model have been advanced. Among them : M-estimation [see Delecroix and Hristache (1999), Delecroix et al. (2006), Xia and Li (1999)] ; the so-called "direct", average derivative (gradient) based, methods [see Härdle and Stoker (1989), Hristache et al. (2001a,b)] ; iterative methods, see Xia and Härdle (2006) among others. This list is far from being complete and just gives an overview of research directions. Roughly speaking, the existing methodologies essentially consist of two steps : first, one estimates the index θ , possibly with the use of some preliminary nonparametric estimator of f , as for instance in M-estimation ; second, one uses a plug-in estimator to obtain the final estimator of F .

There are at least two issues related : the calibration of the initial value of θ in the iterative methods and the selection of the smoothing parameters (bandwidths or cut-off indices) related to the nonparametric estimating the link function, e.g. to its unknown regularity. Obviously, a suboptimal choice may dramatically affect the overall estimation quality so it is one of central points in semiparametric modeling [see the discussion in Carroll et al. (1997) and Xia and Härdle (2006)]. Moreover, for example, in M-estimation one has to choose a bandwidth twice, for a pilot nonparametric estimator (like the Nadaraya-Watson estimator) of f when estimating θ by $\widehat{\theta}$ maximizing the corresponding criterion and the next time in order to build a final estimator of F . As pointed out in Delecroix et al. (2006), the bandwidth optimal for estimating θ is not necessarily even satisfactory for the use in the final estimator. Therefore, some adaptive to the unknown direction of θ , selection of the smoothing parameters is strongly desirable.

Yet another aspect of the problem. To the best of our knowledge, the overwhelming majority of the methods require the link function f to be k -times differentiable, see Delecroix et al. (2006) for the range of the values of k .

Estimation in the single-index regression. Consider a regression model :

$$Y_i = F(X_i) + \varepsilon_i, \qquad i = 1, \dots, n, \tag{2}$$

where X_i are independent random vectors in $\mathbb{R}^d$ with common density g w.r.t. the Lebesgue measure. The noise $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. centered symmetric random variables satisfying with $\Omega \in]0,1]$ and $\omega > 0$ the tail condition $\int_x^\infty p(y)\mathrm{d}y \leq \Upsilon \exp\{-\Omega x^\omega\} \forall x \geq 0$. The sequences $\{\varepsilon_i\}_{i=1}^n$ and $\{X_i\}_{i=1}^n$ are assumed to be independent.

To judge the quality of estimation we use the "pointwise" risk defined as follows :

$$\mathcal{R}_{r,t}^{(n)}(\widehat{F}, F) = \Big(\mathbb{E}_F |\widehat{F}(t) - F(t)|^r\Big)^{1/r}, \quad t \in [-1/2, 1/2]^d.$$

Here $\widehat{F}(\cdot)$ is an $\{(X_i, Y_i)\}_{i=1}^n$ -measurable function and $\mathbb{E}_F$ denotes the mathematical expectation with respect to $\mathbb{P}_F$, the joint distribution of the sequence $\{(X_i, Y_i)\}_{i=1}^n$.

We will present an adaptive procedure that develops the idea of pointwise adaptation introduced in Lepski (1990) and Kerkyacharian et al. (2001). Due to the lack of space we provide here only some of obtained results, see Lepski and Serdyukova (2014) for more details. Particularly, here we describe the estimation procedure and the results on the adaptive rate of convergence over a collection of classes of functions possessing the single index structure defined by

$$\mathbb{F}_d(\beta,L) \overset{\text{def}}{=} \Big\{F:\mathbb{R}^d\to\mathbb{R} \mid F(x)=f\Big(\theta^\top x\Big),\ f\in\mathbb{H}(\beta,L),\ \theta\in\mathbb{S}^{d-1}\Big\}, \tag{3}$$

where $\mathbb{H}(\beta,L)$, $\beta>0$, $L>0$, denotes a standard Hölder class of univariate functions. In what follows we assume that the design density g is bounded away from zero on some compact larger than the estimation interval. In addition, we suppose that the link function possesses some minimal smoothness, that is that f is uniformly bounded and continuous. However, this information is not required for the estimation procedure and appears only in proofs of the theoretical results. On the other hand, all the results are correct, if f is discontinuous, but its uniform upper bound is available.

Kernel estimators. Let $\mathcal{K}:\mathbb{R}\to\mathbb{R}$ be a function satisfying the following assumption.

Assumption 1. **(1)** $\mathrm{supp}(\mathcal{K})\subseteq[-1/2,1/2]$, $\int\mathcal{K}=1$, $\mathcal{K}$ *is symmetric;*

(2) *there exists* $Q>0$ *such that* $\Big|\mathcal{K}(u)-\mathcal{K}(v)\Big|\le Q|u-v|$, $\forall u,v\in\mathbb{R}$.

Let $d=2$. For any $(\theta,h)\in\mathbb{S}^1\times[h_{\min},1]$, define the matrix

$$E_{(\theta,h)}=\begin{pmatrix} h^{-1}\theta_1 & h^{-1}\theta_2\\ -\theta_2 & \theta_1\end{pmatrix},\qquad \det\Big(E_{(\theta,h)}\Big)=h^{-1},$$

and consider kernel estimators with $K(u,v)=\mathcal{K}(u)\mathcal{K}(v)$ so that

$$\widehat{F}_{(\theta,h)}(\cdot)=\det\Big(E_{(\theta,h)}\Big)n^{-1}\sum_{i=1}^n Y_iK\Big(E_{(\theta,h)}(X_i-\cdot)\Big)g(X_i)^{-1}.$$

For any $\theta,\nu\in\mathbb{S}^1$ and any $h\in[h_{\min},1]$, define

$$\overline{E}_{(\theta,h)(\nu,h)}=\begin{pmatrix} \frac{(\theta_1+\nu_1)}{2h(1+|\nu^\top\theta|)} & \frac{(\theta_2+\nu_2)}{2h(1+|\nu^\top\theta|)}\\ -\frac{(\theta_2+\nu_2)}{2(1+|\nu^\top\theta|)} & \frac{(\theta_1+\nu_1)}{2(1+|\nu^\top\theta|)}\end{pmatrix},$$

where

$$E_{(\theta,h)(\nu,h)}=\begin{cases}\overline{E}_{(\theta,h)(\nu,h)}, & \nu^\top\theta\ge 0,\\ \overline{E}_{(-\theta,h)(\nu,h)}, & \nu^\top\theta<0;\end{cases}\qquad \frac{1}{4h}\le\det\Big(E_{(\theta,h)(\nu,h)}\Big)\le\frac{1}{2h}.$$

A kernel estimator associated with the matrix $E_{(\theta,h)(\nu,h)}$ is defined by

$$\widehat{F}_{(\theta,h)(\nu,h)}(\cdot)=\det\Big(E_{(\theta,h)(\nu,h)}\Big)n^{-1}\sum_{i=1}^n Y_iK(E_{(\theta,h)(\nu,h)}(X_i-\cdot)\Big)g(X_i)^{-1}.$$

Procedure. Define $\mathrm{TH}(\eta) = 2\Big[\|\mathcal{K}\|_\infty^2\sqrt{\ln(n)} + \widehat{F}_\infty C_1(n) + C_2(n)\Big](\eta n)^{-1/2}, \quad \eta \in (0,1]$, where $\widehat{F}_\infty = 2\sup_{v\in[-5/2,5/2]^2}\Big|\widehat{F}(v)\Big| + 2C_5(n)$ and $\widehat{F}(v)$ is an auxiliary kernel estimator allowing estimating without knowledge of the uniform upper bound on the regression function. The quantities $C_1(n)$, $C_2(n)$ and $C_5(n)$ are given in the paper.

Set $\mathcal{H}_n = \{h_k = 2^{-k},\ k \in \mathbb{N}^0\} \cap [2^{-1}h_{\min}, 1]$ and let for any $\theta \in \mathbb{S}^1$ and $h \in \mathcal{H}_n$

$$\begin{aligned} R_t^{(1)}(\theta,h) &= \sup_{\eta\in\mathcal{H}_n:\ \eta\le h}\ [\sup_{\nu\in\mathbb{S}^1}|\widehat{F}_{(\theta,\eta)(\nu,\eta)}(t) - \widehat{F}_{(\nu,\eta)}(t)| - \mathrm{TH}(\eta)]_+; \\ R_t^{(2)}(h) &= \sup_{\eta\in\mathcal{H}_n:\ \eta\le h}\ [\sup_{\theta\in\mathbb{S}^1}|\widehat{F}_{(\theta,h)}(t) - \widehat{F}_{(\theta,\eta)}(t)| - \mathrm{TH}(\eta)]_+. \end{aligned}$$

Subsequently, define $(\widehat{\theta},\widehat{h})$ as a solution of the following minimization problem :

$$R_t^{(1)}(\widehat{\theta},\widehat{h}) + R_t^{(2)}(\widehat{h}) + \mathrm{TH}(\widehat{h}) = \inf_{(\theta,h)\in\mathbb{S}^1\times\mathcal{H}_n}[R_t^{(1)}(\theta,h) + R_t^{(2)}(h) + \mathrm{TH}(h)].$$

Then our final estimator is $\widehat{F}(t) = \widehat{F}_{(\widehat{\theta},\widehat{h})}(t)$.

Theorem 1. *Let $b>0$ be fixed; and let the kernel $\mathcal{K}$ additionally satisfy $\int z^j\mathcal{K}(z)\mathrm{d}z = 0,\ \ \forall j = 1,\ldots,\lfloor b\rfloor$. Then, for any $\beta \le b,\ L>0,\ r\ge 1$ and $t\in[-1/2,1/2]^2$, we have*

$$\sup_{F\in\mathbb{F}_2(\beta,L)} \mathcal{R}_{r,t}^{(n)}\Big(\widehat{F}_{(\widehat{\theta},\widehat{h})}, F\Big) \le \varkappa_1\psi_n(\beta,L),$$

where $\psi_n(\beta,L) = L^{1/(2\beta+1)}\left[n^{-1}\ln(n)\right]^{\beta/(2\beta+1)}$ and $\varkappa_1$ is independent of n.

Under additional assumptions on the densities of the noise and design we have the following lower bound result.

Theorem 2. *For any $t\in[-1/2,1/2]^d$, $d\ge 2$, $\beta, L>0$, and any n large enough,*

$$\inf_{\widetilde{F}}\ \sup_{F\in\mathbb{F}_d(\beta,L)} \mathcal{R}_{r,t}^{(n)}\Big(\widetilde{F}, F\Big) \ge \varkappa_2\psi_n(\beta,L),$$

where the infimum is over all possible estimators. Here $\varkappa_2$ is a numerical constant independent of n and L, and $\psi_n(\beta,L)$ is defined in Theorem 1.

We see that the logarithmic term appearing in the lower bound is a sort of “payment” for the complicated structure but not for the pointwise adaptation. A natural question is : If we consider more involved structured functions, should we “pay” only the ln ? In the second part of the present note we introduce a class of anisotropic multi-index functions and give the best obtainable rate of convergence in estimating of the such functions.

Estimation in the multi-index model. Let $D_j^l g$ denote the lth order partial derivative of $g : \mathbb{R}^m \to \mathbb{R}$ with respect to the variable z_j ; and let $\lfloor \beta_k \rfloor$ be the largest integer strictly less than β_k .

Definition 1. *Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, $\beta_k > 0,\ k = 1, \ldots, m$ and $L > 0$. A function $g : \mathbb{R}^m \to \mathbb{R}$ belongs to the anisotropic Hölder class $\mathbb{H}_m(\boldsymbol{\beta}, L)$ if g has continuous partial derivatives of all orders $l \leq \lfloor \beta_k \rfloor$, $k = 1, \ldots, m$, and for all $k = 1, \ldots, m$*

$$\begin{aligned}
&\|D_k^l f\|_\infty \leq L \qquad \forall l \leq \lfloor \beta_k \rfloor \\
&\Big| D_k^{\lfloor \beta_k \rfloor} g(z_1, \ldots, z_{k-1}, z_k + \tau, z_{k+1} \ldots, z_m) - D_k^{\lfloor \beta_k \rfloor} g(z_1, \ldots, z_k, \ldots, z_m) \Big| \leq L \tau^{\beta_k - \lfloor \beta_k \rfloor} \\
&\forall z \in \mathbb{R}^m, \tau \in \mathbb{R}.
\end{aligned}$$

Let $\theta_k \in \mathbb{S}^{d-1}, k = 1, \ldots, m, m \leq d$, be linearly independent unit-length vectors. Define the class of anisotropic multi-index functions as follows :

$$\mathbb{F}_{m,d}(\boldsymbol{\beta}, L) \stackrel{\text{def}}{=} \Big\{ F : \mathbb{R}^d \to \mathbb{R} \;\Big|\; F(x) = f\Big(\theta_1^\top x, \ldots, \theta_m^\top x\Big),\ \theta_k \in \mathbb{S}^{d-1},\ f \in \mathbb{H}_m(\boldsymbol{\beta}, L), m \leq d \Big\}.$$

When $\boldsymbol{\beta} = (\beta, \ldots, \beta)$, we will write $\mathbb{F}_{m,d}(\beta, L)$ (the isotropic case).

Let $m_d = \begin{cases} d/2, & \text{if } d \text{ is even,} \\ \lfloor d/2 \rfloor, & \text{if } d \text{ is odd,} \end{cases}$ and let $\mathbb{I}(d) = \begin{cases} 1, & \text{if } d \text{ is odd and } m > m_d, \\ 0, & \text{otherwise.} \end{cases}$

It is interesting that in the accordance with Theorem 2 the best obtainable rate depends on whether $m \leq m_d$ or not. In order to formulate our lower bound result in a unified way we need the following

Assumption 2. *Let $d \geq m > m_d + \mathbb{I}(d)$. Suppose that there exist m_d directions in which the corresponding harmonic mean of smoothness indexes strictly dominates : $\sum_{(k)=1}^{m_d} \beta_{(k)}^{-1} > \sum_{(k)=m_d+1+\mathbb{I}(d)}^{m} \beta_{(k)}^{-1}$, where $(k) \in \{1, \ldots, m\}$ are the permutated indexes from Definition 1.*

Note that the number of terms on the right-hand side of the inequality is at most m_d with the equality only if $m = d$. Therefore, in the isotropic case Assumption 2 is trivially fulfilled for any $m < d$ and fails for $m = d$. Consequently, the result of the theorem below is not applicable if $m = d$ and the function is isotropic. It is not surprising because the isotropic case with $m = d$ can be reduced to estimating of a d-variate function with no structure constraint and the minimax rate in this case is $n^{-\beta/(2\beta+d)}$. On the contrary, if the anisotropy presents, Assumption 2 can be viewed as an intolerance level of the function to the rotation of the coordinate axes.

Denote by $\mathbb{F}_{m,d}^{\text{anis}}(\boldsymbol{\beta}, L) = \mathbb{F}_{m,d}(\boldsymbol{\beta}, L) \backslash \mathbb{F}_{m,d}(\beta, L)$ the class of multi-index functions with purely anisotropic link functions. Under the conditions of the preceding theorem we have

Theorem 3. *For any $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)$, $\beta_k > 0,\ k = 1, \ldots, m,\ L > 0,\ r \geq 1$, $t \in [-1/2, 1/2]^d$, $d \geq 2$, and any n sufficiently large*

(1) *in the isotropic case if $m < d$,*

$$\inf_{\widetilde{F}} \sup_{F \in \mathbb{F}_{m,d}(\beta, L)} \mathcal{R}_{r,t}^{(n)}\Big(\widetilde{F}, F\Big) \geq \varkappa L^{m/(2\beta+m)} \Big[n^{-1} \ln(n)\Big]^{\beta/(2\beta+m)};$$

(2) *in the anisotropic case if either* $m \le m_d$ *or* $d \ge m > m_d + \mathbb{I}(d)$ *and Assumption 2 holds,*

$$\inf_{\widetilde{F}} \sup_{F \in \mathbb{F}^{anis}_{m,d}(\beta,L)} \mathcal{R}^{(n)}_{r,t}\Big(\widetilde{F}, F\Big) \ge \varkappa L^{1/(2\gamma+1)} \Big[n^{-1}\ln(n)\Big]^{\gamma/(2\gamma+1)}, \ \gamma^{-1} = \sum_{k=1}^{m} \beta_k^{-1},$$

where infimum is taken over all estimators and $\varkappa$ *is a constant independent of* n *and* L.

It is worth mentioning that all the obtained rates of convergence agree with the prominent Stone's dimensionality reduction principle [see Stone (1985)], particularly, in the isotropic case $\beta_k = \beta \ \forall k$ we observe in the rate the “effective dimension” m.

Références

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models. *J. Amer. Statist. Assoc.* **92 :438** 477–489.

Delecroix, M. and Hristache, M. (1999). M-estimateurs semi-paramétriques dans les modèles à direction révélatrice unique. *Bull. Belg. Math. Soc. Simon Stevin* **6 :2** 161–185.

Delecroix, M. Hristache, M. and Patilea, V. (2006). On semiparametric M-estimation in single-index regression. *J. Statist. Plann. Inference* **136 :3** 730–769.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84 :408** 986–995.

Horowitz J. L. (1998). *Semiparametric and Nonparametric Methods in Econometrics.*

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Ann. Statist.* **29 :3** 595–623.

Hristache, M., Juditsky, A., Polzehl, J. and Spokoiny, V. (2001). Structure adaptive approach for dimension reduction. *Ann. Statist.* **29 :6**, 1537–1566.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multi–index denoising. *Probab. Theory Related Fields* **121**, 137–170.

Lepskii, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35 :3** 454–466.

Lepski, O. and Serdyukova, N. (2014). Adaptive estimation under single-index constraint in a regression model. *Ann. Statist.* **42 :1** 1–28.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.*

Serdyukova, N. (2014). On the best obtainable rates of convergence in estimation at a given point under multi-index constraint. Manuscript.

Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13 :2** 689–705.

Xia, Y. and Li, W. K. (1999). On single-index coefficient regression models. *J. Amer. Statist. Assoc.* **94 :448** 1275–1285.

Xia, Y. and Härdle, W. (2006). Semi-parametric estimation of partially linear single-index models. *J. Multivariate Anal.* **97 :5** 1162–1184.