Estimation of the effective dimension reduction subspace

Joint work with A. Iouditsky and V. Spokoiny

Arnak DALALYAN
LPMA, Université Paris 6

Nonparametric Regression with Additive Noise

- We observe $(x_1, Y_1, \ldots, x_n, Y_n)$ with

$$\boxed{Y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n,}$$

- $x_i \in \mathbb{R}^d$ are called explanatory variables,
- $f : \mathbb{R}^d \to \mathbb{R}$ is the unknown regression function,
- ε_i are i.i.d. centered random variables with finite variance σ^2.

In our theoretical results, we will assume that x_i are deterministic and that ε_i are gaussian.

Statistical problems

- **Non-parametric inference :** under no particular assumption on f,
  - predict the value of the response Y at a new point $x_0 \in \mathbb{R}^d$,
  - find a function $\hat{f}$ that fits well the values of f at the observed design points x_i.
- **Semi-parametric inference :** under some structural assumption on f, estimate the structural parameters. For example,
  - in the partial linear model $f(x) = \theta^T x^{(1)} + g(x^{(2)})$, $x = (x^{(1)}, x^{(2)}) \in \mathbb{R}^{d_1+d_2}$ the structural parameter is $\theta \in \mathbb{R}^{d_1}$,
  - in the single-index model $f(x) = g(\theta^T x)$, $\theta \in \mathbb{S}_{d-1}$, the structural parameter is θ.

Motivation : curse of dimensionality
Local-linear smoothing

- A well-known method for estimating f non parametrically is based on local-linear smoothing.
- For a kernel $K(\cdot)$ and a bandwidth $h > 0$, the local-linear estimator is defined by

$$\begin{bmatrix} \hat{f}_n(x_i) \\ \widehat{\nabla f}_n(x_i) \end{bmatrix} = \min_{(a,b)\in\mathbb{R}^{d+1}} \sum_{j=1}^{n} \{Y_j - a - b^T(x_j - x_i)\}^2 w_{ij}$$

$$= \left\{ \sum_{j=1}^{n} \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix} \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix}^T w_{ij} \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix} w_{ij} \ ,$$

  where $x_{ij} = x_j - x_i$ and $w_{ij} = K(|x_{ij}/h|)$.
- If the design $\{x_i\}$ is “regular”, it holds

$$\inf_{K,h} \sup_{f:\|\nabla^2 f\|_\infty \le R} \mathbf{E}[\|\hat{f}_n - f\|_2^2] \underset{n\to\infty}{\asymp} n^{-4/(4+d)}.$$

Curse of dimensionality

- The risk of the local-linear estimator with “ideal” kernel and bandwidth is of order $n^{-4/(4+d)}$.
- No estimator do better! The rate $n^{-4/(4+d)}$ is minimax on the Sobolev ball $\Sigma(2, R)$.

This rate is too slow when d is large.

- For functions f smoother than C^2, better rate can be attained using local-polynomial smoothing instead of local-linear one.
- The computation of the local-polynomial estimator of degree $\ell \geq 2$ may be highly time-consuming; since it requires the inversion of a $d^\ell \times d^\ell$ matrix.

One way of dealing with the curse of dimensionality

- Leave the fully non-parametric model in favor of a semiparametric model where the function f is assumed to have some “structure”.
- This assumption is helpful, even if the structure is unknown, since rough estimates of the unknown function may lead to a good estimator of the “structure”.
- Using the estimated structure, we can reduce the dimensionality and significantly improve the quality of estimators of the function f.

Effective dimension-reduction subspace (EDRS)

Let us introduce the following structural assumption:

$$\boxed{f(x) = g(\Theta^T x), \qquad \forall x \in \mathbb{R}^d,} \tag{1}$$

- $g : \mathbb{R}^m \to \mathbb{R}$ for some $m \leq d$,
- Θ is a $d \times m$ matrix such that $\Theta^T\Theta = I_m$,
- furthermore, Θ is "the smallest" matrix satisfying (1): for every orthogonal matrix $\bar\Theta$ of size $d \times m'$ such that $f(x) = g(\bar\Theta^T x)$, $\forall\, x \in \mathbb{R}^d$, it holds

$$Span(\Theta) \subset Span(\bar\Theta).$$

Statement of the problem

We postulate that the data $(x_1, Y_1, \dots, x_n, Y_n)$ obeys the model

$$Y_i = f(x_i) + \varepsilon_i = g(\Theta^T x_i) + \varepsilon_i, \quad i = 1, \dots, n \qquad (2)$$

where Θ is a $d \times m$ matrix with orthonormal columns.

- The function g as well as the matrix Θ are unknown.
- We are interested in the inference on Θ.

We say that $\mathcal{S} = Span(\Theta)$ is the **index space** or the **effective dimension-reduction subspace**.

Identifiability issues

- On the one hand, consistent estimation of Θ is impossible since Θ is not uniquely defined by f ! In fact, for every orthognal matrix $U \in \mathbb{R}^m \otimes \mathbb{R}^m$, we have
$$f(x) = \tilde{g}(\tilde{\Theta}^T x)$$
with $\tilde{g}(\cdot) = g(U^T \cdot)$ and $\tilde{\Theta} = \Theta U^T$.
- On the other hand, the orthogonal projector $\Pi^* = \Theta\Theta^T$ onto $\mathcal{S}$ is uniquely defined by f and, consequently, can be consistently estimated.
- In what follows we assume that the structural dimension $m = Tr(\Pi^*)$ is known.

Sliced inverse regression of Li (JASA, 1991)
Main idea

- Assume that (x_i, ε_i) are iid $\mathcal{N}_{d+1}(0, I_{d+1})$, then for every set A, the conditional expectation $\mathbf{E}(x_i | Y_i \in A)$ lies in the EDRS $\mathcal{S}$.

$$\begin{aligned}
\mathbf{E}(x_i | Y_i) &= \mathbf{E}(\Pi_{\mathcal{S}} x_i | Y_i) + \mathbf{E}(\Pi_{\mathcal{S}^\perp} x_i | Y_i)\\
&= \underbrace{\Pi_{\mathcal{S}} \mathbf{E}(x_i | Y_i)}_{\in \mathcal{S}} + \mathbf{E}[\underbrace{\mathbf{E}(\Pi_{\mathcal{S}^\perp} x_i | \Theta^T x_i, \varepsilon_i)}_{=0} | Y_i]\\
&= \Pi_{\mathcal{S}} \mathbf{E}(x_i | Y_i).
\end{aligned}$$

- If $x_i \sim \mathcal{N}_d(\mu, \Sigma)$ and $x_i \perp\!\!\!\perp \varepsilon_i$, then

$$\Sigma^{-1} \mathbf{E}(x_i - \mu | Y_i) \in \mathcal{S}.$$

This feature holds true for elliptically contoured distributions.

Sliced inverse regression

The method

For a fixed $h > 0$,

1. $\forall k \in \mathbb{Z}$ estimate the vector $\mathbf{E}(x|Y \in [kh,(k+1)h[)$ by

$$\hat{\beta}_k = \sum_{i=1}^{n} \frac{x_i}{n_k} \mathbb{1}_{[kh,(k+1)h[}(Y_i), \quad n_k = \sum_{i=1}^{n} \mathbb{1}_{[kh,(k+1)h[}(Y_i).$$

2. conduct a PCA on $\{\hat{\beta}_k\}_{k\in\mathbb{Z}}$: compute the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d$ and the eigenvectors $v_1, \ldots, v_d$ of the matrix

$$\mathcal{B}_n = \frac{1}{n} \sum_{k\in\mathbb{Z}} \hat{\beta}_k \hat{\beta}_k^T n_k.$$

3. Then set $\hat{\mathcal{S}} = Span\{v_1, \ldots, v_m\}$.

Slice inverse regression

- Advantages :
  - easy to implement, the algorithm is speedy even for large d,
  - nice theoretical features obtained by simple arguments.
- Limitations :
  - strong probabilistic assumption on the design,
  - no guarantee that all the directions of $\mathcal{S}$ are recovered.

MAVE method of Xia et al. (JRSS, 2002)

- Idea : choose $\hat{\Theta}_n$ by minimizing

$$PL_n(\Theta) = \min_{\{a_j,b_j\}_j} \sum_{i=1}^{n} \sum_{j=1}^{n} \{Y_i - a_j - b_j^T \Theta^T (x_i - x_j)\}^2 w_{ij}$$

  where the weights w_{ij} vanish when x_i is far from x_j.
- Iterative method :
  $\{w_{ij}^{(0)}\} \rightsquigarrow \hat{\Theta}^{(1)} \rightarrowtail \{w_{ij}^{(1)}\} \rightsquigarrow \ldots \rightsquigarrow \hat{\Theta}^{(K)}$
- Advantage: good empirical performance.
- Limitations :
  - theoretical properties are poorly studied,
  - non-convex optimization,
  - does not classify the directions.

SA-PCA method of HRISTACHE ET AL. (AoS 2001)
Heuristic ideas

- $\mathcal{S}$ coincides with $Span\{\nabla f(x_1), \ldots, \nabla f(x_n)\}$.
- Let $\{\psi_\ell, \ell \leq L\} \subset \mathbb{R}^n$ be such that

$$\begin{cases} \frac{1}{n}\sum_{i=1}^n \psi_\ell(x_i)^2 = 1 \\ \text{Rank}\{\psi_\ell, \ell \leq L\} = n \end{cases}$$

then

$$\mathcal{S} = Span\{(\beta_\ell)_{\ell \leq L}\} \text{ where } \beta_\ell = \frac{1}{n}\sum_{i=1}^n \nabla f(x_i)\psi_\ell(x_i).$$

- estimation of β_ℓ is easier than that of $\nabla f(x_i)$.
- few ψ_ℓ suffice for capturing the structure.

SA-PCA method
Heuristic ideas

- Thus, on the one hand, a “good” estimator of ∇f may be very helpful for recovering the structure.
- On the other hand, the knowledge of the structure leads to a significant improvement in the estimation of ∇f. In fact, if an “oracle” gives us $\mathcal{S}$, we may consider

$$\begin{bmatrix} \hat{f}(x_i) \\ \widehat{\nabla f}(x_i) \end{bmatrix} = \left\{ \sum_{j=1}^{n} \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix} \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix}^T w_{ij}^* \right\}^{-1} \sum_{j=1}^{n} Y_j \begin{bmatrix} 1 \\ x_{ij} \end{bmatrix} w_{ij}^* ,$$

with the “ideal” weights $w_{ij}^* = K(|\Pi^* x_{ij}|/h)$, where

- Π^* stands for the orthogonal projector onto $\mathcal{S}$,
- $h > 0$ is a bandwidth,
- K is a function $\in C^2$, > 0, vanishing outside $[-1, 1]$.

SA-PCA method

The algorithm: First of all, choose $\{\psi_\ell, \ell \leq L\}$ and standardize the design.

1. Set $k = 1$, $\rho_1 = 1$, $\hat{\Pi}_1 = I$ and choose h_1 conveniently.
2. Estimate $\nabla f(x_i)$ for every i by local linear smoothing with
$$w_{ij}^{(k)} = K\Big(\frac{|\hat{\Pi}_k x_{ij}|}{\rho_k h_k} + \frac{|(I - \hat{\Pi}_k)x_{ij}|}{h_k}\Big).$$
3. Compute $\hat{\beta}_{\ell,k} = n^{-1}\sum_i \widehat{\nabla f_k}(x_i)\psi_\ell(x_i)$.
4. Determine $\hat{\Pi}_{k+1}$ by a PCA on $\hat{\beta}_{\ell,k}$,
5. Set $\rho_{k+1} = a_\rho \rho_k$, $h_{k+1} = a_h h_k$ and increment k.
6. Terminate if $\rho_k < n^{-1/(3\vee m)}$, otherwise return to step 2.

Structural adaptation via maximum minimization

<u>The idea:</u> in the algorithm SA-PCA, modify the way of extracting the structural information from $(\hat{\beta}_\ell)_{\ell \leq L}$.

<u>The reason:</u> the risk of SA-PCA is proportional to L.

<u>Our proposal:</u>

- PCA is equivalent to the optimization problem

$$\textit{minimize} \quad \sum_\ell \hat{\beta}_{\ell,k}^T (I - \Pi) \hat{\beta}_{\ell,k}$$

  over the set of all projectors Π of rank $\leq m$.
- We replace this optimization by:

$$\textit{minimize} \quad \max_\ell \hat{\beta}_{\ell,k}^T (I - \Pi) \hat{\beta}_{\ell,k}$$

  over the set of all symmetric matrices Π such that $0 \leq \Pi \leq I$ and $tr(\Pi) \leq m$.

Assumptions

(A1) There exists $C_g > 0$ such that $|\nabla g(x)| \leq C_g$ and

$$|g(x) - g(x') - (x - x')^T \nabla g(x)| \leq C_g |x - x'|^2$$

for all $x, x' \in \mathbb{R}^m$.

(A2) Let $\mathcal{B}^* = \big\{\bar\beta = \sum_{\ell=1}^L c_\ell \beta_\ell : \sum_{\ell=1}^L |c_\ell| \leq 1\big\}$. There exist $\bar\beta_1, \ldots, \bar\beta_m \in \mathcal{B}^*$ and $\mu_1, \ldots, \mu_m \in \mathbb{R}_+$ such that

$$\Pi^* \leq \sum_{k=1}^{m^*} \mu_k \bar\beta_k \bar\beta_k^T.$$

(A3) Technical assumption on the design.

(A4) The errors are gaussian.

The vectors (functions) ψ_ℓ satisfy $\max_{i,\ell} |\psi_\ell(x_i)| \leq \bar\psi$.

Main result

If assumptions (A1)-(A4) are satisfied, then there exists a constant $C > 0$ such that $\forall z \in]0, 2\sqrt{\log(nL)}]$ and for n large enough, it holds

$$\boldsymbol{P}\left(\|\widehat{\Pi}_n - \Pi^*\|_2 > \frac{C\log(nL)}{n^{\frac{2}{3\vee m}}} + \frac{Cz\sigma}{\sqrt{n}}\right) \leq Lze^{-\frac{z^2-1}{2}} + \frac{6\log n}{n}.$$

- For $m \leq 4$, we get the optimal rate $1/\sqrt{n}$.
- For $m > 4$, the rate is probably sub-optimal. It can be improved by using local polynomial smoothing of degree > 1 with stronger smoothness assumptions on g.

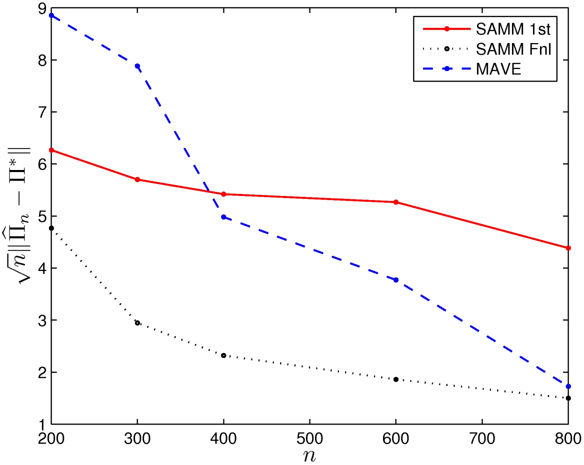
Example 1 : "single-index model"

We set

$$\begin{aligned} d &= 5, \\ f(x) &= g(\theta^T x), \\ g(t) &= 4|t|^{1/2}\sin^2(\pi t), \\ \theta &= (a, 2a, 0, 0, 0). \end{aligned}$$

Further, we choose $x_i^{(j)}$ i.i.d. uniformly distributed on $[-1, 1]$ and ε_i i.i.d. $0.5\mathcal{N}(0, 1)$ independent of x.

Example 1 : “single-index model”

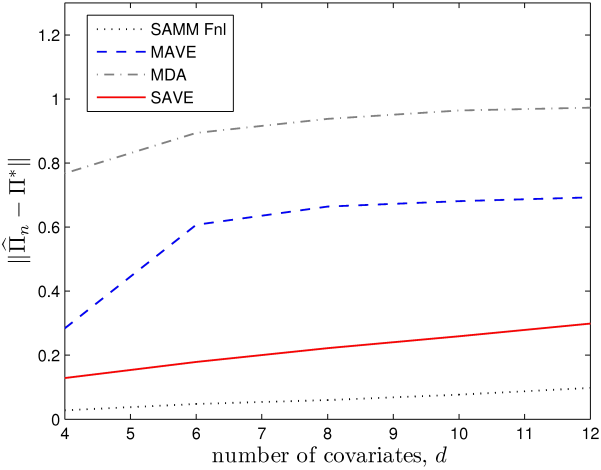


Example 2 : “double-index model”

We set

$$
\begin{aligned}
n &= 300, \\
g(x) &= (x_1 - x_2^3)(x_1^3 + x_2), \\
\theta_1 &= (1, 0, \ldots, 0), \\
\theta_2 &= (0, 1, \ldots, 0), \\
x_i^{(j)} &\overset{iid}{\sim} \mathcal{U}([-40, 40]), \\
\varepsilon_i &\overset{iid}{\sim} 0.1\mathcal{N}(0, 1).
\end{aligned}
$$

Example 2 : "double-index model"



Conclusion and outlook

- The SAMM method provides a consistent estimator of $\mathcal{S}$ under very mild identifiability assumptions.
- In almost all simulations we did, SAMM is much better than SIR.
- SAMM is comparable to MAVE, but
  - SAMM seems to deal better with the bias than MAVE,
  - SAMM has the advantage of classifying the directions.
- Extension of SAMM to the case of unknown m is a challenging problem.
- Consistent estimation of m under realistic assumptions is an interesting open problem.