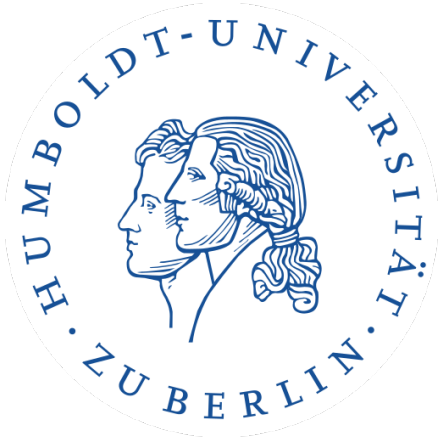
Statistical Inference in Sparse High-Dimensional Nonparametric Models

"Meeting in Mathematical Statistics"
CIRM, Luminy, Marseille, France
December 18-22, 2017

Karl Gregory, Enno Mammen, Martin Wahl

Universität Heidelberg



Humboldt-Universität zu Berlin

Research question:

Consider a model that contains

- **a nonparametric component of interest** and
- **a high-dimensional nonparametric nuissance component**.

Research question:

Consider a model that contains

- **a nonparametric component of interest** and
- **a high-dimensional nonparametric nuissance component**.

We give conditions under which the nonparametric component of interest can be estimated with the same asymptotic accuracy regardless of if the high-dimensional nuissance component is known or not known.

Research question:

Consider a model that contains

- **a nonparametric component of interest** and
- **a high-dimensional nonparametric nuissance component**.

We give conditions under which the nonparametric component of interest can be estimated with the same asymptotic accuracy regardless of if the high-dimensional nuissance component is known or not known.

Then the nonparametric estimator of the nonparametric component of interest has the same asymptotic distribution as a well studied nonparametric estimator in a model with only one nonparametric component.

Research question:

Consider a model that contains

- **a nonparametric component of interest** and
- **a high-dimensional nonparametric nuissance component**.

We give conditions under which the nonparametric component of interest can be estimated with the same asymptotic accuracy regardless of if the high-dimensional nuissance component is known or not known.

Then the nonparametric estimator of the nonparametric component of interest has the same asymptotic distribution as a well studied nonparametric estimator in a model with only one nonparametric component.

This allows the implementation of methods for statistical inference on the nonparametric function of interest.

Motivating example:

$Y = c + \sum_{j=1}^{q} f_j(X_j) + \varepsilon; \quad E[f_j(X_j)] = 0; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q$ large, $|\{j : f_j \neq 0\}|$ smaller

"Sparse high-dimensional additive model"

Motivating example:

$Y = c + \sum_{j=1}^{q} f_j(X_j) + \varepsilon; \quad E[f_j(X_j)] = 0; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q$ large, $|\{j : f_j \neq 0\}|$ smaller

"Sparse high-dimensional additive model"

Much work on estimation: Ingster and Lepski (03), Ravikumar et al. (08), Meier et al. (09), Koltschinskii and Yuan (10), Huang et al. (10), Raskutti et al. (11), Gayraud and Ingster (12), Suzuki (12), Dalalyan et al. (12).

- Good estimation results under sparsity
- Can use group Lasso to select/estimate components

Motivating example:

$$Y = c + \sum_{j=1}^{q} f_j(X_j) + \varepsilon; \quad E[f_j(X_j)] = 0; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q \text{ large, } |\{j : f_j \neq 0\}| \text{ smaller}$$

"Sparse high-dimensional additive model"

Much work on estimation: Ingster and Lepski (03), Ravikumar et al. (08), Meier et al. (09), Koltschinskii and Yuan (10), Huang et al. (10), Raskutti et al. (11), Gayraud and Ingster (12), Suzuki (12), Dalalyan et al. (12).

- Good estimation results under sparsity
- Can use group Lasso to select/estimate components

Little work on inference:

- Lasso estimators have a complicated distribution.
- However, in the *parametric setting*, where

$$Y = \sum_{j=1}^{q} \beta_j X_j + \varepsilon; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q \text{ large}, \quad |\{j : \beta_j \neq 0\}| \text{ small};$$

  Zhang and Zhang (2014), van de Geer et al. (2014) and Javanard and Montanari (2014) propose the "desparsified/debiased Lasso", which, under sparsity conditions, produces asymptotically normal estimators $\hat{\beta}_1; \ldots; \hat{\beta}_q$.

Motivating example:

$$Y = c + \sum_{j=1}^{q} f_j(X_j) + \varepsilon; \quad E[f_j(X_j)] = 0; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q \text{ large, } |\{j : f_j \neq 0\}| \text{ smaller}$$

"Sparse high-dimensional additive model"

Much work on estimation: Ingster and Lepski (03), Ravikumar et al. (08), Meier et al. (09), Koltschinskii and Yuan (10), Huang et al. (10), Raskutti et al. (11), Gayraud and Ingster (12), Suzuki (12), Dalalyan et al. (12).

- Good estimation results under sparsity
- Can use group Lasso to select/estimate components

Little work on inference:

- Lasso estimators have a complicated distribution.
- However, in the *parametric setting*, where

$$Y = \sum_{j=1}^{q} \beta_j X_j + \varepsilon; \quad \varepsilon \sim \text{Normal}(0; \sigma^2); \quad q \text{ large}, \quad |\{j : \beta_j \neq 0\}| \text{ small};$$

  Zhang and Zhang (2014), van de Geer et al. (2014) and Javanard and Montanari (2014) propose the "desparsified/debiased Lasso", which, under sparsity conditions, produces asymptotically normal estimators $\hat{\beta}_1; \ldots; \hat{\beta}_q$.

We propose a nonparametric "debiased" Lasso to enable inference in the additive model. We describe a two-step procedure based on resmoothing the "debiased" Lasso for optimal estimation.

Table of Contents

Debiasing: general model

General setting: $Y = f_1 + f_{-1} + "$

with

- $" \sim \mathrm{N}(0;\ ^2 I_n)$,

Debiasing: general model

General setting: $Y = f_1 + f_{-1} + "$

with

- $" \sim \mathrm{N}(0;\ ^2 I_n)$,
- f_1 (random) element of $\mathbb{R}^n$, component of interest
  (e.g. $f_1 = \left(f_1(X_1^i)\right)_{i=1}^n$ in the additive model),
- f_{-1} (random) element of $\mathbb{R}^n$, high-dimensional nonparametric nuissance component
  (e.g. $f_{-1} = f_2 + \cdots + f_d$ in the additive model with $f_j = \left(f_j(X_j^i)\right)_{i=1}^n$),

Debiasing: general model

General setting: $Y = f_1 + f_{-1} + "$

with

- $" \sim \mathrm{N}(0;\ ^2 I_n)$,
- f_1 (random) element of $\mathbb{R}^n$, component of interest
  (e.g. $f_1 = \left(f_1(X_1^i)\right)_{i=1}^n$ in the additive model),
- f_{-1} (random) element of $\mathbb{R}^n$, high-dimensional nonparametric nuissance component
  (e.g. $f_{-1} = f_2 + \cdots + f_d$ in the additive model with $f_j = \left(f_j(X_j^i)\right)_{i=1}^n$),
- $V_1; V_{-1} \subset \mathbb{R}^n$ approximating linear subspaces for f_1 or f_{-1}, respectively,
- $\hat{\Pi}_1 : \mathbb{R}^n \to V_1$, $\hat{\Pi}_{-1} : \mathbb{R}^n \to V_{-1}$ linear maps (e.g. in additive model: $\hat{\Pi}_1$ projection onto V_1, $\hat{\Pi}_{-1}$ "LASSO"-projection onto V_{-1})

Debiasing: general model

General setting: $Y = f_1 + f_{-1} + "$

with

- $" \sim \mathrm{N}(0;\ ^2 I_n)$,
- f_1 (random) element of $\mathbb{R}^n$, component of interest
  (e.g. $f_1 = \left(f_1(X_1^i)\right)_{i=1}^n$ in the additive model),
- f_{-1} (random) element of $\mathbb{R}^n$, high-dimensional nonparametric nuissance component
  (e.g. $f_{-1} = f_2 + \cdots + f_d$ in the additive model with $f_j = \left(f_j(X_j^i)\right)_{i=1}^n$),
- $V_1; V_{-1} \subset \mathbb{R}^n$ approximating linear subspaces for f_1 or f_{-1}, respectively,
- $\hat{\Pi}_1 : \mathbb{R}^n \to V_1$, $\hat{\Pi}_{-1} : \mathbb{R}^n \to V_{-1}$ linear maps (e.g. in additive model: $\hat{\Pi}_1$ projection onto V_1, $\hat{\Pi}_{-1}$ "LASSO"-projection onto V_{-1})

Aim: do asymptotically/approximately as well as

$$\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T Y^{(oracle)}; \text{ where } Y^{(oracle)} = Y - f_{-1} = " + f_1;$$

Aim:

do asymptotically/approximately as well as

$$\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T Y^{(oracle)}; \text{ where } Y^{(oracle)} = Y - f_{-1} = " + f_1:$$

Debiased estimator:

$$\hat{f}_1 = A(Y - \hat{f}_{-1}^{(init)})$$

with

- $\hat{f}_{-1}^{(init)}$ available initial estimator of f_{-1} (e.g. group LASSO-estimator in additive model),
- $A = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)$:

Aim:

do asymptotically/approximately as well as

$$\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T Y^{(oracle)}; \text{ where } Y^{(oracle)} = Y - f_{-1} = " + f_1:$$

Debiased estimator:

$$\hat{f}_1 = A(Y - \hat{f}_{-1}^{(init)})$$

with

- $\hat{f}_{-1}^{(init)}$ available initial estimator of f_{-1} (e.g. group LASSO-estimator in additive model),
- $A = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)$:

Motivation: Because of

$$\hat{f}_1 - f_1 = A(Y - \hat{f}_{-1}^{(init)}) - f_1 = (A - I)f_1 + A(f_{-1} - \hat{f}_{-1}^{(init)}) + A";$$

$\hat{f}_1$ is a bias corrected version of $\tilde{f}_1 = AY$. Note that for $\tilde{f}_1$ we have

$$\tilde{f}_1 - f_1 = (A - I)f_1 + Af_{-1} + A":$$

We have the following theorem for the comparison of $\hat{f}_1 = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)(Y - \hat{f}_{-1}^{(init)})$ and $\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T (Y - f_{-1})$:

We have the following theorem for the comparison of $\hat{f}_1 = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)(Y - \hat{f}_{-1}^{(init)})$ and $\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T (Y - f_{-1})$:

Theorem 1.

Make the Assumptions (A1)–(A6) (will be introduced in a second). Then it holds with probability $\geq 1 - 6$

$$\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty} \quad \leq \quad 1 + C_2(1 - \quad _1)^{-1} \times \quad \sqrt{\frac{2(\log(2n) + 2\log(1/\quad)}{n}} + (C_1 + C_2) \quad _2 + \quad _1 + \quad _1 \quad _2 \quad .$$

Here

- $\|z\|_{n;\infty} = \max_{1 \leq i \leq n} |z_i|$ empirical sup norm,

We have the following theorem for the comparison of $\hat{f}_1 = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)(Y - \hat{f}_{-1}^{(init)})$ and $\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T (Y - f_{-1})$:

Theorem 1.

Make the Assumptions (A1)–(A6) (will be introduced in a second). Then it holds with probability $\geq 1 - 6$

$$\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty} \quad \leq \quad 1 + C_2(1 - \ _1)^{-1} \times \quad \sqrt{\frac{2(\log(2n) + 2\log(1=\ )}{n}} + (C_1 + C_2)\ _2 + \ _1 + \ _1\ _2 \quad :$$

Here

- $\|z\|_{n;\infty} = \max_{1 \leq i \leq n} |z_i|$ empirical sup norm,
- $C_1; C_2$ slowly growing constant (in additive regression: polynomials in log terms and sparsity of additive components and dependence structure of $X_2; \ldots; X_d$),
- $_1; \ _2$ small constants (in additive regression: slowly growing constant times nonparametric rate),

We have the following theorem for the comparison of
$\hat{f}_1 = (I - \hat{\Pi}_1^T \hat{\Pi}_{-1}^T)^{-1} \hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T)(Y - \hat{f}_{-1}^{(init)})$ and $\hat{f}_1^{(oracle)} = \hat{\Pi}_1^T (Y - f_{-1})$:

Theorem 1.

Make the Assumptions (A1)–(A6) (will be introduced in a second). Then it holds with probability $\geq 1 - 6$

$$\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty} \leq \left(1 + C_2(1 - \ _1)^{-1} \times \sqrt{\frac{2(\log(2n) + 2\log(1/\ ))}{n}} + (C_1 + C_2)\ _2 + \ _1 + \ _1\ _2\right) :$$

Here

- $\|z\|_{n;\infty} = \max_{1 \leq i \leq n} |z_i|$ empirical sup norm,
- $C_1; C_2$ slowly growing constant (in additive regression: polynomials in log terms and sparsity of additive components and dependence structure of $X_2; \ldots; X_d$),
- $\ _1; \ _2$ small constants (in additive regression: slowly growing constant times nonparametric rate),
- $\ _1; \ _2$ bias terms, approximation error of V_1 or V_{-1} , respectively,
- $0 < \ _1 < 1$ constant.

Interpretation of the bound of the Theorem 1:

$$\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty} \quad \leq \quad \left(1 + C_2(1 - \ _1)^{-1}\right)$$
$$\times \left(\sqrt{\frac{2(\log(2n) + 2\log(1= \ )}{n}} + (C_1 + C_2) \ _2 + \ _1 + \ _1 \ _2 \right)$$

Up to small terms, $\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty}$ is only bounded by bias terms.

Interpretation of the bound of the Theorem 1:

$$\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty} \leq \left(1 + C_2(1 - \ _1)^{-1}\right) \times \left(\sqrt{\frac{2(\log(2n) + 2\log(1= \ )}{n}} + (C_1 + C_2)\ _2 + \ _1 + \ _1\ _2\right)$$

Up to small terms, $\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_{n;\infty}$ is only bounded by bias terms.

Two applications of Theorem 1:

1. In case of undersmoothing, the same asymptotic theory applies for $\hat{f}_1$ as for $\hat{f}_1^{(\text{oracle})}$. E.g. one gets the same asymptotic distributions for pointwise inference and for uniform bounds of $\hat{f}_1$ as for $\hat{f}_1^{(\text{oracle})}$.

Second application of Theorem 1: Resmoothing.

Consider application of a second smoothing step $\mathcal{S}$ after a first step with undersmoothed $\hat{f}_1$.

Resmoothig estimator $\hat{f}_1^{(resmooth)} = \mathcal{S}\hat{f}_1$ with smoothing operator $\mathcal{S} : \mathbb{R}^n \to \mathcal{C}[I]$.

Second application of Theorem 1: Resmoothing.

Consider application of a second smoothing step $\mathcal{S}$ after a first step with undersmoothed $\hat{f}_1$.

Resmoothig estimator $\hat{f}_1^{(resmooth)} = \mathcal{S}\hat{f}_1$ with smoothing operator $\mathcal{S} : \mathbb{R}^n \to \mathcal{C}[I]$.

E.g. Nadaraya-Watson smoothing in the additive model:

$$\mathcal{S}g(\cdot) = \frac{\sum_{i=1}^n K_h(X_1^i - \cdot)g_i}{\sum_{i=1}^n K_h(X_1^i - \cdot)} = \frac{\sum_{i=1}^n K_h(X_1^i - \cdot)g(X_1^i)}{\sum_{i=1}^n K_h(X_1^i - \cdot)}$$

For the smoothing operator $\mathcal{S}$ we make the following two assumptions:

$$\|\mathcal{S}g\|_\infty \leq C\|g\|_{n;\infty}$$

for some $C > 0$

and

$$\Delta = \|\mathcal{S}Y^{(\text{oracle})} - \mathcal{S}\hat{f}_1^{(\text{oracle})}\|_\infty \text{ small}$$

with $Y^{(\text{oracle})} = Y - f_{-1}$, i.e. $\mathcal{S}Y^{(\text{oracle})}$ smoothing estimator with smoothing $\mathcal{S}$ in oracle model.

Then

$$\|\mathcal{S}\hat{f}_1 - \mathcal{S}Y^{(\text{oracle})}\|_\infty \leq C\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_\infty + \Delta;$$

where the first term can be bounded with the help of Theorem 1.
Δ is typically small if the amount of smoothing in $\hat{f}_1$ is small compared to the smoothing in $\mathcal{S}$:

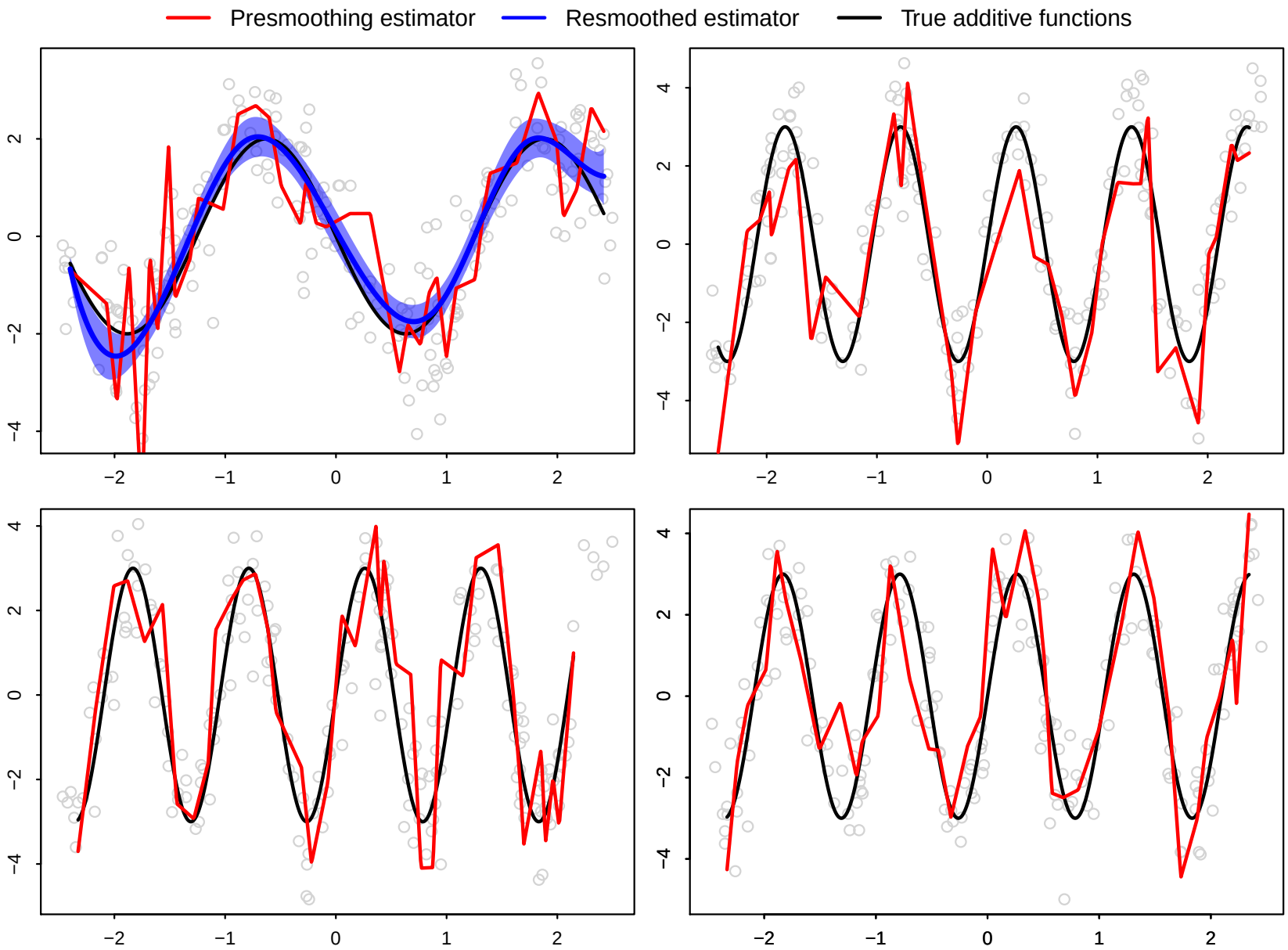
$$\text{Smoothing} \circ \text{Undersmoothing} \approx \text{Smoothing};$$

This can be used

- for the construction of estimators of f_1 with good performance,
- for the development of an optimality theory:

  For a smoothing estimator in the oracle model (where one knows f_{-1}) one can construct an estimator in the full model (where one does not know f_{-1}) with nearly the same operation characteristics.

  Using this argument one can show (under mild conditions) that (sharp) asymptotic minimax theorems valid in the oracle model carry over to be valid in the full model (e.g. in the additive model one has the same minimax results for the first component regardless of whether the nuissance components $f_2, \ldots, f_d$ are known or not known.



Plots 2–4: undersmoothed $\hat{f}_2^{(init)}$,..., $\hat{f}_4^{(init)}$, f_2,...,f_4; Plot 1: $\hat{f}_1$, $\mathcal{S}\hat{f}_1$, f_1.

The assumptions of Theorem 1

For some constants $C_1; C_2; \ _1; \ _2; \ _1; \ _2 > 0$ and $0 < \ ; \ _1 < 1$ we make the following assumptions.

(A1) **(∞-norm preservation of $\hat{\Pi}_1$)**

$$\|\hat{\Pi}_1 y\|_{n;\infty} \leq C_1 \|y\|_{n;\infty}$$

for all $y \in \mathbb{R}^n$ with probability $\geq 1 - \ $.

The assumptions of Theorem 1

For some constants $C_1; C_2; \ _1; \ _2; \ _1; \ _2 > 0$ and $0 < \ ; \ _1 < 1$ we make the following assumptions.

(A1) **(∞-norm preservation of $\hat{\Pi}_1$)**

$$\|\hat{\Pi}_1 y\|_{n;\infty} \leq C_1 \|y\|_{n;\infty}$$

for all $y \in \mathbb{R}^n$ with probability $\geq 1 - \ $.

(A2) **(Smoothing property)**

$$\|\hat{\Pi}_{-1}\hat{\Pi}_1 e_i\|_{n;2} \leq \frac{C_2}{n}$$

for all $i = 1; :::; n$ with probability $\geq 1 - \ $, where e_i is the ith standard basis vector and $\|\cdot\|_{n;2}$ is the Euclidean norm of $\mathbb{R}^n$.

The assumptions of Theorem 1

For some constants $C_1; C_2; \ _1; \ _2; \ _1; \ _2 > 0$ and $0 < \ ; \ _1 < 1$ we make the following assumptions.

(A1) **(∞-norm preservation of $\hat{\Pi}_1$)**

$$\|\hat{\Pi}_1 y\|_{n;\infty} \leq C_1 \|y\|_{n;\infty}$$

for all $y \in \mathbb{R}^n$ with probability $\geq 1 - \ $.

(A2) **(Smoothing property)**

$$\|\hat{\Pi}_{-1} \hat{\Pi}_1 e_i\|_{n;2} \leq \frac{C_2}{n}$$

for all $i = 1; :::; n$ with probability $\geq 1 - \ $, where e_i is the ith standard basis vector and $\|\cdot\|_{n;2}$ is the Euclidean norm of $\mathbb{R}^n$.

(A3) **(Empirical minimal angle assumption)**

$$\|\hat{\Pi}_{-1} \hat{\Pi}_1 y\|_{n;2} \leq \ _1 \|y\|_{n;2}$$

for all $y \in \mathbb{R}^n$ with probability $\geq 1 - \ $.

The assumptions of Theorem 1, continued.

(A4) **(Bias conditions)**

$$\|f_1 - g_1^*\|_{n;\infty} \leq \quad _1 \text{ for some } g_1^* \in V_1$$

and

$$\|f_{-1} - g_{-1}^*\|_{n;\infty} \leq \quad _2 \text{ for some } g_{-1}^* \in V_{-1}$$

with probability $\geq 1 - \quad$.

The assumptions of Theorem 1, continued.

(A4) **(Bias conditions)**

$$\|f_1 - g_1^*\|_{n;\infty} \leq \quad_1 \text{ for some } g_1^* \in V_1$$

and

$$\|f_{-1} - g_{-1}^*\|_{n;\infty} \leq \quad_2 \text{ for some } g_{-1}^* \in V_{-1}$$

with probability $\geq 1 - \quad$.

(A5) **(Approximate orthogonality assumption)**

$$\|\hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T) g_{-1}\|_{n;\infty} \leq \quad_1 \text{pen}(g_{-1})$$

for all $g_1 \in V_1$ with probability $\geq 1 - \quad$. Here pen: $V_{-1} \to \mathbb{R}_+$ (e.g. in additive models we will choose $\text{pen}(g_{-1}) = \|g_2\|_{n;2} + \cdots + \|g_d\|_{n;2}$.

The assumptions of Theorem 1, continued.

(A4) **(Bias conditions)**

$$\|f_1 - g_1^*\|_{n;\infty} \leq \quad_1 \text{ for some } g_1^* \in V_1$$

and

$$\|f_{-1} - g_{-1}^*\|_{n;\infty} \leq \quad_2 \text{ for some } g_{-1}^* \in V_{-1}$$

with probability $\geq 1 - \quad$.

(A5) **(Approximate orthogonality assumption)**

$$\|\hat{\Pi}_1^T (I - \hat{\Pi}_{-1}^T) g_{-1}\|_{n;\infty} \leq \quad_1 \text{pen}(g_{-1})$$

for all $g_1 \in V_1$ with probability $\geq 1 - \quad$. Here pen: $V_{-1} \to \mathbb{R}_+$ (e.g. in additive models we will choose $\text{pen}(g_{-1}) = \|g_2\|_{n;2} + \cdots + \|g_d\|_{n;2}$.

(A6) **(Condition on the initial estimator)** $\text{pen}\,(\hat{f}_{-1}^{(init)} - g_1^*) \leq \quad_2$ with probability $\geq 1 - \quad$.

Table of Contents

The debiased estimator in sparse high-dimensional additive models

We now want to apply Theorem 1 to

The sparse high-dimensional additive model

$$Y^i = \sum_{j=1}^{q} f_j(X_j^i) + \ ^i; \quad ^i \sim \text{Normal}(0; \ ^2); \quad i = 1; : : : ; n; \quad q \text{ large}, \quad |\{j : f_j \neq 0\}| \ll n$$

Now, f_1 function of interest,

$f_2; : : : ; f_q$ nuissance components.

We suppose that X_j^i take values in $[0; 1]$.

Our choice of the function spaces $V_1, \ldots, V_q$

- V_j = piecewise polynomials in $x_j \in [0,1]$ of maximal degree t_j defined on

$$\text{Interval } I_{jk} = \left(\frac{k}{m_j}, \frac{k+1}{m_j} \right], \quad k = 0, \ldots, m_j - 1$$

  for $j = 1, \ldots, q$.
- Wlog: $m_2 = \cdots = m_q$, $t_2 = \cdots = t_q$.

Our choice of the function spaces $V_1, \ldots, V_q$

- V_j = piecewise polynomials in $x_j \in [0,1]$ of maximal degree t_j defined on

$$\text{Interval } I_{jk} = \left(\frac{k}{m_j}, \frac{k+1}{m_j} \right], \quad k = 0, \ldots, m_j - 1$$

  for $j = 1, \ldots, q$.
- Wlog: $m_2 = \cdots = m_q$, $t_2 = \cdots = t_q$.
- $V_2, \ldots, V_q$ centered so that $\mathbb{E} g_j(X_j) = 0$ for all $g_j \in V_j$, $j = 2, \ldots, q$ and

$$V_{-1} = \sum_{j=2}^{q} V_j.$$

- Bases for $V_1, \ldots, V_q$ can be constructed from the Legendre polynomials

$$b_{j,k(t_j+1)+1}, \ldots, b_{j,k(t_j+1)+t_j+1}$$

  as orthonormal basis which are zero outside the interval I_{jk}

To reconstruct the desparsified Lasso estimator in the additive model context, we will need Lasso estimators of $f_1; : : : ; f_q$ as well as a Lasso version of the projection of the V_1 basis functions onto V_{-1}.

We first define the nonparametric Lasso estimator of f by

$$\left(\hat{f}_1^L; : : : ; \hat{f}_q^L\right) = \underset{g_j \in V_j}{\operatorname{argmin}} \left\{ \left\| Y - \sum_{j=1}^{q} g_j \right\|_n^2 + 2 \sum_{j=1}^{q} \|g_j\|_n \right\};$$

where > 0 is some tuning parameter.

We set

$$\hat{f}_{-1}^{(init)} = \hat{f}_2^L + \cdots + \hat{f}_q^L:$$

For the Lasso version of the projection of the V_1 basis functions onto V_{-1} we put for $k = 1;:::;d_1$

$$\hat{\Pi}^L_{-1} b_{1k} = \sum_{j=2}^{q} (\hat{\Pi}^L_{-1} b_{1k})_j \in V_{-1}$$

with

$$\left((\hat{\Pi}^L_{-1} b_{1k})_2;:::;(\hat{\Pi}^L_{-1} b_{1k})_q\right) = \underset{g_j \in V_j}{\operatorname{argmin}} \left\{ \left\| b_{1k} - \sum_{j=2}^{q} g_j \right\|_n^2 + 2 \sum_{j=2}^{q} \|g_j\|_n \right\};$$

where > 0 is some tuning parameter. Moreover, we define $\hat{\Pi}_{-1}$ as the linear extension of $\hat{\Pi}^L_{-1}$ to V_1.

Some quantities for stating the result for additive models

Dimensions

- $q =$ total # of functions, $s_0 = \#\{\text{nonzero } f_j\}$, $s_1 =$ sparsity of projection Π_{-1}

Some quantities for stating the result for additive models

Dimensions

- $q =$ total # of functions, $s_0 = \#\{\text{nonzero } f_j\}$, $s_1 =$ sparsity of projection Π_{-1}
- $d_1 = \dim(V_1)$, $d_2 = \dim(V_2) = \cdots = \dim(V_q)$, $d = \max_j d_j$

Some quantities for stating the result for additive models

Dimensions

- $q =$ total # of functions, $s_0 = \#\{\text{nonzero } f_j\}$, $s_1 =$ sparsity of projection Π_{-1}
- $d_1 = \dim(V_1)$, $d_2 = \dim(V_2) = \cdots = \dim(V_q)$, $d = \max_j d_j$

Geometric quantities

- ϕ, ψ are theoretical compatibility constants
- ρ_0 the minimal angle between V_1 and V_{-1}
- $\delta = \frac{C}{} \sqrt{\frac{s_1 d(x+\log d+\log q)}{n}}$, governs diff. betwn empirical $\|\cdot\|_n$ and true norms $\|\cdot\|_2$

Some quantities for stating the result for additive models

Dimensions

- q = total # of functions, $s_0 = \#\{\text{nonzero } f_j\}$, s_1 = sparsity of projection Π_{-1}
- $d_1 = \dim(V_1)$, $d_2 = \dim(V_2) = \cdots = \dim(V_q)$, $d = \max_j d_j$

Geometric quantities

- ϕ, ψ are theoretical compatibility constants
- ρ_0 the minimal angle between V_1 and V_{-1}
- $\delta = \frac{C}{\sqrt{\frac{s_1 d(x+\log d+\log q)}{n}}}$, governs diff. betwn empirical $\|\cdot\|_n$ and true norms $\|\cdot\|_2$

Approximation quantities

- r_1, r_2 are smoothnesses such that there exists $g_j^* \in V_j$ for which

$$\|f_1 - g_1^*\|_\infty \leq C_0 d_1^{-r_1} \quad \text{and} \quad \|f_j - g_j^*\|_\infty \leq C_0 d_2^{-r_2}, \quad j = 2, \ldots, q$$

Some quantities for stating the result for additive models

Dimensions

- $q =$ total # of functions, $s_0 = \#\{\text{nonzero } f_j\}$, $s_1 =$ sparsity of projection Π_{-1}
- $d_1 = \dim(V_1)$, $d_2 = \dim(V_2) = \cdots = \dim(V_q)$, $d = \max_j d_j$

Geometric quantities

- ϕ, ψ are theoretical compatibility constants
- ρ_0 the minimal angle between V_1 and V_{-1}
- $\delta = \frac{C}{\sqrt{\frac{s_1 d(x+\log d+\log q)}{n}}}$, governs diff. betwn empirical $\|\cdot\|_n$ and true norms $\|\cdot\|_2$

Approximation quantities

- r_1, r_2 are smoothnesses such that there exists $g_j^* \in V_j$ for which

$$\|f_1 - g_1^*\|_\infty \leq C_0 d_1^{-r_1} \quad \text{and} \quad \|f_j - g_j^*\|_\infty \leq C_0 d_2^{-r_2}, \quad j = 2, \ldots, q$$

Lasso tuning parameters

- $\lambda = 2\sigma\sqrt{\frac{d}{n}} + 2\sigma\sqrt{\frac{2x+2\log q}{n}}$

  $\eta = C\sqrt{\frac{d(x+\log d_1+\log q)}{n}} + \frac{\sqrt{s_1} d(x+\log d_1+\log q)}{n}$, $x > 1$

Theorem 2

If Assumptions (B1)–(B5) hold and

$$\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1\eta^2}{\phi^2} \leq (1-\rho_0)^2/C$$

as well as

$$\max(s_0, s_1)\left(\frac{d}{\sqrt{n}} + \sqrt{\frac{d(x+\log q)}{n}} + \frac{d(x+\log q)}{n}\right) \leq \phi^2/C,$$

Theorem 2

If Assumptions (B1)–(B5) hold and

$$\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1\eta^2}{\phi^2} \leq (1-\rho_0)^2/C$$

as well as

$$\max(s_0, s_1)\left(\frac{d}{\sqrt{n}} + \sqrt{\frac{d(x+\log q)}{n}} + \frac{d(x+\log q)}{n}\right) \leq \phi^2/C,$$

then

$$\mathbb{P}\left(\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_\infty \geq C(\Delta_1+\Delta_2+\Delta_3)\right) \leq 4\exp(-x) + \exp(-y),$$

Theorem 2

If Assumptions (B1)–(B5) hold and

$$\frac{s_1\delta}{\psi^2} + \frac{s_1\sqrt{d_1}\eta}{\psi\phi} + \frac{s_1 d_1\eta^2}{\phi^2} \leq (1-\rho_0)^2/C$$

as well as

$$\max(s_0, s_1)\left(\frac{d}{\sqrt{n}} + \sqrt{\frac{d(x+\log q)}{n}} + \frac{d(x+\log q)}{n}\right) \leq \phi^2/C,$$

then

$$\mathbb{P}\left(\|\hat{f}_1 - \hat{f}_1^{(\text{oracle})}\|_\infty \geq C(\Delta_1+\Delta_2+\Delta_3)\right) \leq 4\exp(-x) + \exp(-y),$$

where

$$\Delta_1 = \frac{1}{\psi(1-\rho_0)}\left(s_1 d_1^{-r_1} + s_1 s_0 d_2^{-r_2}\right)$$

$$\Delta_2 = \frac{1}{\psi(1-\rho_0)}\left((\eta/\lambda)\sqrt{s_1 d_1}\left(d_1^{-r_1} + s_0 d_2^{-r_2}\right)^2 + s_0\sqrt{s_1}\sqrt{d_1}\lambda\eta\right)$$

$$\Delta_3 = \frac{1}{\psi(1-\rho_0)}\sqrt{\frac{s_1(\log d_1 + y)}{n}}.$$

An asymptotic interpretation of Theorem 2

Set $x = y = \log q$ and suppose

$$\log\log q = o(\log n), \quad s_0 = O(n^{\ 0}), \quad s_1 = O(n^{\ 1}), \quad n \to \infty$$

for $0 \leq \gamma_0 < 1/2$ and $0 \leq \gamma_1 \leq 1/4$. Then

$$\Delta_1 + \Delta_2 + \Delta_3 = o(n^{-\ })$$

An asymptotic interpretation of Theorem 2

Set $x = y = \log q$ and suppose

$$\log\log q = o(\log n), \quad s_0 = O(n^{\ 0}), \quad s_1 = O(n^{\ 1}), \quad n \to \infty$$

for $0 \leq \gamma_0 < 1/2$ and $0 \leq \gamma_1 \leq 1/4$. Then

$$\Delta_1 + \Delta_2 + \Delta_3 = o(n^{-\ })$$

if

$$\begin{aligned}
1 + \frac{1}{r_2} \quad \gamma_0 + \quad \frac{1}{2} + \frac{1}{2r_1} + \frac{1}{r_2} \quad \gamma_1 \quad &< \quad 1 - \quad 1 + \frac{1}{2r_1} + \frac{1}{r_2} \quad \beta, \\
2(\gamma_0 \vee \gamma_1) + \frac{2}{r_1}\gamma_1 \quad &< \quad 1 - \frac{2}{r_1}\beta, \\
\frac{2}{r_2}(\gamma_0 \wedge \gamma_1) + \quad 2 + \frac{2}{r_2} \quad (\gamma_0 \vee \gamma_1) \quad &< \quad 1 - \frac{2}{r_2}\beta.
\end{aligned}$$

An asymptotic interpretation of Theorem 2

Set $x = y = \log q$ and suppose

$$\log\log q = o(\log n), \quad s_0 = O(n^{\ 0}), \quad s_1 = O(n^{\ 1}), \quad n \to \infty$$

for $0 \leq \gamma_0 < 1/2$ and $0 \leq \gamma_1 \leq 1/4$. Then

$$\Delta_1 + \Delta_2 + \Delta_3 = o(n^{-\ })$$

if

$$\begin{aligned}
1 + \frac{1}{r_2} \quad \gamma_0 + \quad \frac{1}{2} + \frac{1}{2r_1} + \frac{1}{r_2} \quad \gamma_1 \quad &< \quad 1 - \quad 1 + \frac{1}{2r_1} + \frac{1}{r_2} \quad \beta, \\
2(\gamma_0 \vee \gamma_1) + \frac{2}{r_1}\gamma_1 \quad &< \quad 1 - \frac{2}{r_1}\beta, \\
\frac{2}{r_2}(\gamma_0 \wedge \gamma_1) + \quad 2 + \frac{2}{r_2} \quad (\gamma_0 \vee \gamma_1) \quad &< \quad 1 - \frac{2}{r_2}\beta.
\end{aligned}$$

The optimal rate $\beta = r_1/(2r_1 + 1)$ is achievable if

- $r_2 \geq 2r_1/(2r_1 + 1)$ and $r_1 > 1/2$, i.e. RHSs positive,
- and when $\gamma_0, \gamma_1 \geq 0$ are small enough

Table of Contents

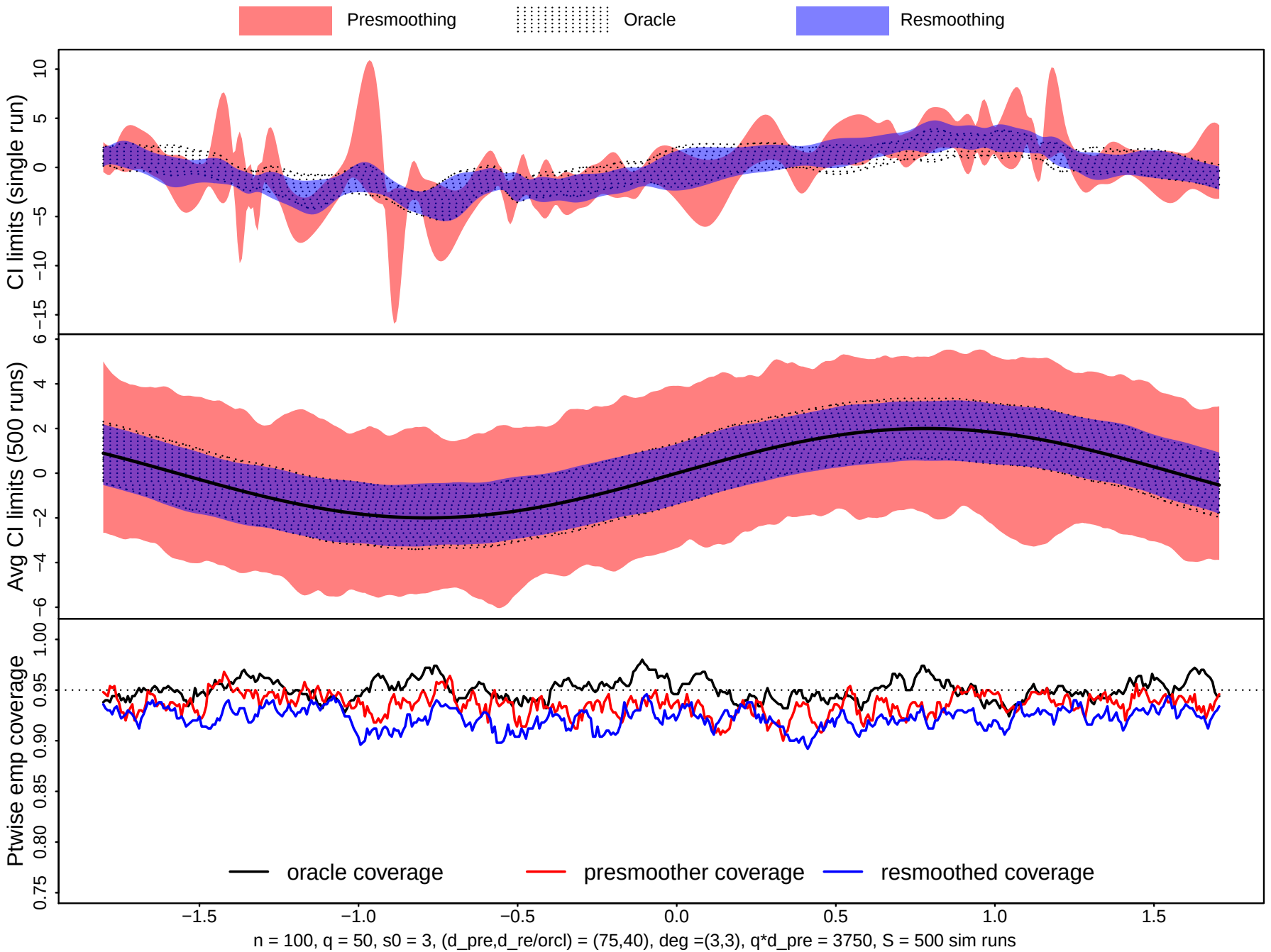
Performance on simulated data

Model

$$Y = \sum_{j=1}^{q} \underbrace{(1=j) f(X_j) \mathbf{1}(j \leq s_0)}_{s_0 \text{ ``active'' covariates}} + ''$$

- f: sine, line, expo, quad
- $X_j \sim \text{Unif}(-2{:}5; 2{:}5)$, correlation $0{:}9$ within s_0-size groups
- Choose ; with crossvalidation
- $n = 100; 500; 1000$, $p = 50; 200$ ($s_0 = 3; 10$).

Evaluate empirical coverage of 95% *pointwise CIs*


Presmoothing
Oracle
Resmoothing
CI limits (single run)
Avg CI limits (500 runs)
Ptwise emp coverage
oracle coverage
presmoother coverage
resmoothed coverage
n = 100, q = 50, s0 = 3, (d_pre,d_re/orcl) = (75,40), deg =(3,3), q*d_pre = 3750, S = 500 sim runs

			sine			line			expo			quad		
		$x =$	-1.5	0	1	-1.5	0	1	-1.5	0	1	-1.5	0	1
$n = 100$	$p = 50,\ s_0 = 3$	orcl	0.95	0.95	0.94	0.94	0.95	0.95	0.93	0.91	0.93	0.94	0.94	0.94
	$(d_{\text{pre}} = 75)$	pre-s	0.92	0.94	0.95	0.97	0.97	0.98	0.83	0.94	0.95	0.94	0.97	0.96
	$(d_{\text{re/orcl}} = 40)$	re-s	0.92	0.92	0.92	0.93	0.98	0.97	0.78	0.93	0.88	0.93	0.93	0.92
	$p = 200,\ s_0 = 10$	orcl	0.94	0.95	0.95	0.95	0.96	0.96	0.93	0.93	0.93	0.94	0.94	0.93
	$(d_{\text{pre}} = 75)$	pre-s	0.92	0.88	0.92	0.97	0.98	0.98	0.70	0.88	0.85	0.88	0.92	0.90
	$(d_{\text{re/orcl}} = 57)$	re-s	0.90	0.90	0.89	0.93	0.98	0.97	0.67	0.88	0.79	0.87	0.87	0.87
$n = 500$	$p = 50,\ s_0 = 3$	orcl	0.97	0.94	0.95	0.93	0.94	0.93	0.94	0.94	0.95	0.93	0.95	0.95
	$(d_{\text{pre}} = 200)$	pre-s	0.94	0.94	0.91	0.97	0.97	0.97	0.92	0.96	0.97	0.94	0.95	0.93
	$(d_{\text{re/orcl}} = 100)$	re-s	0.94	0.92	0.92	0.96	0.97	0.96	0.90	0.94	0.94	0.92	0.93	0.94
	$p = 200,\ s_0 = 10$	orcl	0.95	0.96	0.94	0.95	0.95	0.93	0.94	0.94	0.95	0.94	0.94	0.93
	$(d_{\text{pre}} = 200)$	pre-s	0.90	0.89	0.94	0.98	1.00	0.99	0.80	0.97	0.99	0.93	0.99	0.96
	$(d_{\text{re/orcl}} = 129)$	re-s	0.88	0.90	0.94	0.98	0.99	0.99	0.75	0.97	0.94	0.93	0.99	0.97
$n = 1000$	$p = 50,\ s_0 = 3$	orcl	0.96	0.94	0.96	0.96	0.94	0.95	0.94	0.94	0.94	0.95	0.95	0.94
	$(d_{\text{pre}} = 300)$	pre-s	0.93	0.92	0.93	0.97	0.96	0.96	0.90	0.94	0.93	0.91	0.94	0.94
	$(d_{\text{re/orcl}} = 147)$	re-s	0.94	0.90	0.93	0.95	0.94	0.96	0.90	0.93	0.93	0.94	0.92	0.94
	$p = 200,\ s_0 = 10$	orcl	0.95	0.95	0.93	0.93	0.96	0.93	0.94	0.95	0.93	0.94	0.94	0.93
	$(d_{\text{pre}} = 300)$	pre-s	0.88	0.89	0.93	0.99	0.98	0.99	0.78	0.97	0.99	0.92	0.97	0.94
	$(d_{\text{re/orcl}} = 162)$	re-s	0.88	0.90	0.93	0.97	0.99	0.98	0.76	0.98	0.96	0.92	0.93	0.92

Table: Coverage of confidence intervals based on oracle, presmoothing, and resmoothed estimators at points $x = -1.5, 0, 1$ for the sine, line, expo and quad functions for $n = 100, 500, 1000$ and $q = 50, 200$ over 500 simulation runs. Dimension d_{pre} used in presmoothing and $d_{\text{re/orcl}}$ for the oracle and the resmoothed estimator shown.

Summary

- We consider models that contain **a nonparametric component of interest** and **a high-dimensional nonparamtric nuissance component**.
- We discuss **debiased LASSO-estimation** for such models.
- We give conditions under which the nonparametric component of interest can be estimated with the same asymptotic accuracy regardless of if the high-dimensional nuissance component is known or not known.
- This holds for **undersmoothed orthogonal series estimators** and, under weak conditions, it can be achieved **for a large class of other nonparametric estimators**.

Summary

- We consider models that contain **a nonparametric component of interest** and **a high-dimensional nonparamtric nuissance component**.
- We discuss **debiased LASSO-estimation** for such models.
- We give conditions under which the nonparametric component of interest can be estimated with the same asymptotic accuracy regardless of if the high-dimensional nuissance component is known or not known.
- This holds for **undersmoothed orthogonal series estimators** and, under weak conditions, it can be achieved **for a large class of other nonparametric estimators**.
- This allows an **optimality theory** for such models.
- We verified the assumptions for additive nonparametric models. In particular, for **additive models** this implies that an additive function can be estimated with the same asymptotic accuracy regardless of if the other functions are known or not known.

Assumption (B1)

Suppose that for $j = 1;:::;q$, X_j takes values in $[0;1]$ and has a density p_j with respect to the Lebesgue measure on $[0;1]$ which satisfies $c_1 \leq p_j \leq 1{=}c_1$ for some constant $c_1 > 0$. Moreover, suppose that for $j = 2:::;q$, $(X_1;X_j)$ has a density p_{1j} with respect to the Lebesgue measure on $[0;1]^2$ which is bounded from above by $1{=}c_1$.

Assumption (B2) introduces a geometric quantity $\ _0$ which governs the degree of collinearity between the spaces V_1 and V_{-1}. The closer $\ _0$ is to 1, the harder it is to distinguish the effects of X_1 from those of $X_2;:::;X_q$.

Assumption (B2). **[Minimal angle assumption]**

Suppose that there is a constant $0 \leq \ _0 < 1$ such that for all $g_1 \in V_1$,

$$\|\Pi_{-1} g_1\| \leq \ _0 \|g_1\|;$$

where $\Pi_{-1} : L^2(\mathbb{P}^X) \to V_{-1}$ is the orthogonal projection from $L^2(\mathbb{P}^X)$ to V_{-1} given by

$$\Pi_{-1} f = \underset{g \in V_{-1}}{\operatorname{argmin}} \|f - g\|^2:$$

Note that $\ _0$ can also be defined as the minimal angle between V_1 and V_{-1}.

Assumption (B3). **[Bias conditions]**

Suppose that there exist some $r_1; r_2 > 0$ and a subset $J_0 \subseteq \{1; : : : ; q\}$ with $1 \in J_0$ and $|J_0| \leq s_0$ such that for each $j \in J_0$ there is a $g_j^* \in V_j$ satisfying

$$\|f_1 - g_1^*\|_\infty \leq C_0 d_1^{-r_1}$$

if $j = 1$ and

$$\|f_j - g_j^*\|_\infty \leq C_0 d_2^{-r_2}$$

otherwise for some constant $C_0 > 0$. Moreover, setting

$$g^* = \sum_{j \in J_0} g_j^*;$$

suppose that

$$\|f - g^*\|_\infty \leq C_0 \left(d_1^{-r_1} + s_0 d_2^{-r_2}\right):$$

Assumption (B4) states that the projection of each basis function of V_1 onto the space V_{-1} may be approximated sufficiently well by its projection onto a subspace of V_{-1} of s_1 or fewer additive components.

Assumption (B4)

For each $k = 1;:::;d_1$, suppose that there is a subset $J_k \subseteq \{2;:::;q\}$ with $|J_k| \leq s_1$, such that there is a decomposition

$$\Pi_{J_k} b_{1k} - \Pi_{-1} b_{1k} = \sum_{j=2}^{q} v_j$$

with $v_j \in V_j$ satisfying

$$\sum_{j=2}^{q} \|v_j\| \leq C_1 \sqrt{s_1} \sqrt{\frac{d}{n}}$$

for some constant $C_1 > 0$. Finally, suppose that $d \leq n$ and

$$\geq \sqrt{\frac{d}{n}}:$$

Assumption (B5) [Theoretical compatibility conditions]

Suppose that there is a real number $0 < \leq 1$ such that

$$\sum_{j \in J_0} \|g_j\|^2 \leq \Big\| \sum_{j=1} g_j \Big\|^2 = {}^2$$

for all $(g_1; : : : ; g_q) \in (V_1; : : : ; V_q)$ satisfying

$$\sum_{j=1} \|g_j\| \leq 8\sqrt{3} \sum_{j \in J_0} \|g_j\| :$$

Moreover, for $k = 1; : : : ; d_1$, suppose that

$$\sum_{j \in J_k} \|g_j\|^2 \leq \Big\| \sum_{j=2} g_j \Big\|^2 = {}^2$$

for all $(g_2; : : : ; g_q) \in (V_2; : : : ; V_q)$ satisfying

$$\sum_{j=2} \|g_j\| \leq 8\sqrt{3} \sum_{j \in J_k} \|g_j\| :$$