Joint estimation of the conditional mean and the conditional variance in high-dimensions

Joint work with M. Hebiri, K. Meziani, J. Salmon

"Estimation et traitement statistique en grande dimension"
May 16, 2013
Paris, FRANCE

Arnak S. Dalalyan

ENSAE / CREST / GENES

I. Problem presentation

Heteroscedastic Regression

Observations : finite collection of pairs $\{(\boldsymbol{x}_t, y_t); t = 1, \ldots, T\}$

- $\boldsymbol{x}_t \in \mathbb{R}^d$ multidimensional feature vector ;
- $y_t \in \mathbb{R}$ real valued label.

Prediction : for a new feature $\boldsymbol{x}_{T+1}$, predict y_{T+1}.

- Quadratic loss : $\ell[y_{T+1}, \mathrm{b}(\boldsymbol{x}_{T+1})] = (y_{T+1} - \mathrm{b}(\boldsymbol{x}_{T+1}))^2$.
- Bayes predictor : $\mathrm{b}^* = \arg\min_{\mathrm{b}} \mathbf{E}\big\{\ell[y_{T+1}, \mathrm{b}(\boldsymbol{x}_{T+1})]\big\}$

$$\mathrm{b}^*(\boldsymbol{x}) = \mathbf{E}[y_{T+1} | \boldsymbol{x}_{T+1} = \boldsymbol{x}].$$

- Given $\boldsymbol{x}_{T+1} = \boldsymbol{x}$, the average loss of the Bayes predictor :

$\mathrm{s}^{*2}(\boldsymbol{x}) = \mathbf{E}\big\{\ell[y_{T+1}, \mathrm{b}^*(\boldsymbol{x}_{T+1})] \,\big|\, \boldsymbol{x}_{T+1} = \boldsymbol{x}\big\} = \mathbf{Var}[y_{T+1} | \boldsymbol{x}_{T+1} = \boldsymbol{x}]$.

The goal is to estimate the functions b^* and s^*.

Problem reformulation

Observations : finite collection $(\boldsymbol{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ obeying

$$y_t = \mathrm{b}^*(\boldsymbol{x}_t) + \mathrm{s}^*(\boldsymbol{x}_t)\,\xi_t, \qquad t \in \mathcal{T} = \{1, \ldots, T\},$$

where $\mathrm{b}^* : \mathbb{R}^d \to \mathbb{R}$ and $\mathrm{s}^* : \mathbb{R}^d \to \mathbb{R}_+$ such that

$$\text{Conditional mean : } \ \mathbf{E}[y_t|\boldsymbol{x}_t] = \mathrm{b}^*(\boldsymbol{x}_t).$$
$$\text{Conditional variance : } \ \mathbf{Var}[y_t|\boldsymbol{x}_t] = \mathrm{s}^{*2}(\boldsymbol{x}_t).$$

Therefore, ξ_t's are such that $\mathbf{E}[\xi_t|\boldsymbol{x}_t] = 0$ and $\mathbf{Var}[\xi_t|\boldsymbol{x}_t] = 1$. They are often assumed Gaussian $\mathcal{N}(0, 1)$ for simplicity.

Goal : to jointly estimate the functions b^* and s^* by a computationally tractable procedure with strong theoretical guarantees.

Sparsity Assumption

- In these settings, estimating b^* and s^* under no further assumption is an ill-posed problem.
- Sparsity scenario : b^* and s^* belong to some low dimensional spaces.

Example : Homoscedastic regression

$$\forall \boldsymbol{x}, \quad \mathrm{b}^*(\boldsymbol{x}) = \sum_{j=1}^{p} \mathrm{f}_j(\boldsymbol{x})\beta_j^* = [\mathrm{f}_1(\boldsymbol{x}), \ldots, \mathrm{f}_p(\boldsymbol{x})]\boldsymbol{\beta}^*, \quad \text{and} \quad \mathrm{s}^*(\boldsymbol{x}) \equiv \sigma^*$$

$\hookrightarrow$ Dictionary $\{\mathrm{f}_1, \ldots, \mathrm{f}_p\}$ of functions from $\mathbb{R}^d$ to $\mathbb{R}$

$\hookrightarrow$ Unknown vector $(\boldsymbol{\beta}^*, \sigma^*) \in \mathbb{R}^p \times \mathbb{R}$, sparse vector $\boldsymbol{\beta}^*$

$\hookrightarrow$ Sparsity index : $p^* = |\boldsymbol{\beta}^*|_0 := \sum_{j=1}^{p} \mathbb{1}(\beta_j^* \neq 0)$ with $p^* \ll p$

Some remarks

- Because of its nonparametric nature, this problem is hard even for small values of the dimension d.
- The literature on estimating s^* is very scarce as compared to the literature on estimating b^*.
- Estimators of s^* may be used for constructing confidence intervals for the predictions.
- The case of time-inhomogeneous observations is included in the previous set-up. Indeed, if

$$\mathsf{b}_t^*(\boldsymbol{x}) = \mathbf{E}[y_t|\boldsymbol{x}_t = \boldsymbol{x}]$$

  depends on "time" t, one can include the time as a feature $\bar{\boldsymbol{x}}_t = (\boldsymbol{x}_t, t)$ and set $\bar{\mathsf{b}}^*(\bar{\boldsymbol{x}}_t) = \mathsf{b}_t^*(\boldsymbol{x}_t)$.

II. Previous work

Homoscedastic regression

The model is

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^*\boldsymbol{\xi}$$

Observations :	$\boldsymbol{Y} = [y_1, \ldots, y_T]^\top \in \mathbb{R}^T$
Noise :	$\boldsymbol{\xi} = [\xi_1, \ldots, \xi_T]^\top \in \mathbb{R}^T$
Design Matrix :	$\mathbf{X} = x_{t,j}$ with $x_{t,j} = [\mathrm{f}_j(\boldsymbol{x}_t)] \in \mathbb{R}$
Coefficients :	$\boldsymbol{\beta}^* = \left[\beta_1^*, \ldots, \beta_p^*\right]^\top \in \mathbb{R}^p$
Standard deviation :	$\mathrm{s}^*(\boldsymbol{x}_t) \equiv \sigma^* \in \mathbb{R}_*^+$

Recall that the sparsity assumption postulates that $|\boldsymbol{\beta}^*|_0 = p^* \ll p$.

Most popular methods : Lasso and Dantzig selector

- The LASSO of Tibshirani (1996) is defined as

$$\widehat{\beta}^{\text{Lasso}} = \arg\min_{\beta\in\mathbb{R}^p} \left(\frac{|\mathbf{Y} - \mathbf{X}\beta|_2^2}{2\sigma^{*2}} + \lambda \sum_{j=1}^{p} |\mathbf{X}_j|_2 |\beta_j| \right)$$

- The Dantzig selector of Candès and Tao (2007) is

$$\widehat{\beta}^{\text{DS}} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{j=1}^{p} |\mathbf{X}_j|_2 |\beta_j| : \max_{j=1,\cdots,p,} \frac{|\mathbf{X}_j^\top(\mathbf{Y} - \mathbf{X}\beta)|}{|\mathbf{X}_j|_2} \le \lambda \right\}$$

For a tuning parameter satisfying $\lambda \propto 1/\sigma^*$, if $\mathbf{X}$ is "nice", sharp oracle inequalities are available, *e.g.*, Bickel *et al.* (2009).

$$\mathbf{E}\left(\frac{1}{T}|\mathbf{X}(\widehat{\beta}^{\bullet} - \beta)|_2^2\right) \le \mathrm{C}\,\frac{p^*\log(p)}{T}.$$

To correctly tune the parameter λ, the knowledge of σ^* is necessary.

Joint estimation of β^* and σ^* (1/2)

- Scaled Lasso, Städler *et al.* (2010),

$$(\widehat{\beta}^{\text{ScL}}, \widehat{\sigma}^{\text{ScL}}) = \arg\min_{\boldsymbol{\beta},\sigma} \Big(T \log(\sigma) + \frac{|\mathbf{Y} - \mathbf{X}\beta|_2^2}{2\sigma^2} + \frac{\lambda}{\sigma} \sum_{j=1}^{p} |\mathbf{X}_j|_2 |\beta_j| \Big).$$

This can be recast in a convex problem (do $\rho := \frac{1}{\sigma}$ and $\phi := \frac{\beta}{\sigma}$) :

$$\arg\min_{\boldsymbol{\phi},\rho} \Big(-T \log(\rho) + \frac{|\rho \mathbf{Y} - \mathbf{X}\phi|_2^2}{2} + \lambda \sum_{j=1}^{p} |\mathbf{X}_j|_2 |\phi_j| \Big).$$

- Scaled DS version proposed by Dalalyan & Chen (2012) : sharp analysis and computational advantages.

Joint estimation of β^* and σ^* (2/2)

- Square-Root Lasso Antoniadis (2010), Belloni *et al.* (2011), Sun & Zhang (2012),

$$\widehat{\beta}^{\text{SqR-Lasso}} = \arg\min_{\beta} \Big(\big|\boldsymbol{Y} - \mathbf{X}\beta\big|_2 + \lambda \sum\nolimits_{j=1}^{p} |\boldsymbol{X}_j|_2 |\beta_j| \Big)$$

$$\widehat{\sigma}^{\text{SqR-Lasso}} = T^{-1/2} \big|\boldsymbol{Y} - \mathbf{X}\widehat{\beta}^{\text{SqR-Lasso}}\big|_2 .$$

- Self Tuning Instrumental Variables (STIV) Gautier & Tsybakov (2011), $(\widehat{\beta}^{\text{STIV}}, \widehat{\sigma}^{\text{STIV}})$ minimizes

$$\sigma + \lambda \sum\nolimits_{j=1}^{p} |\boldsymbol{X}_j|_2 |\beta_j|$$

subject to the constraints

$$\big|\boldsymbol{Y} - \mathbf{X}\beta\big|_2 \leq \sigma; \qquad \forall j = 1, \cdots, p,\ |\boldsymbol{X}_j^\top (\boldsymbol{Y} - \mathbf{X}\beta)| \leq \widetilde{\lambda} |\boldsymbol{X}_j|_2 .$$

If the two tuning parameters coincide $\lambda = \widetilde{\lambda}$, STIV = SqR-Lasso.

Some remarks

- If the design matrix **X** satisfies some "nice" conditions (RIP, RE,...), the theoretical guarantees for the methods presented in this part are almost as strong as for the Lasso and the DS with known σ^*.
- All the estimators presented in this part are computable by solving a simple convex program. This can be done efficiently even for large dimensions p.
- Extensions to the matrix estimation under the rank-sparsity available Klopp (2012).

III. Main results

Functional transformation

- Re-parametrize by the inverse of the conditional standard deviation s^*

$$\mathrm{r}^*(\boldsymbol{x}) = \frac{1}{\mathrm{s}^*(\boldsymbol{x})} \qquad \text{and} \qquad \mathrm{f}^*(\boldsymbol{x}) = \frac{\mathrm{b}^*(\boldsymbol{x})}{\mathrm{s}^*(\boldsymbol{x})}.$$

- This leads to

$$\mathrm{r}^*(\boldsymbol{x}_t) \cdot y_t = \mathrm{f}^*(\boldsymbol{x}_t) + \xi_t, \qquad t = 1, \ldots, T,$$

  where $\mathrm{r}^* : \mathbb{R}^d \to \mathbb{R}$ is the inverse of conditional standard deviation (StD) and f^* is the conditional signal-to-noise ratio.

- We impose modeling assumptions on the pair $(\mathrm{r}^*, \mathrm{f}^*)$ rather than on $(\mathrm{s}^*, \mathrm{b}^*)$.

Main assumptions on b^* and s^*

Group Sparsity Assumption

For p given functions $\mathrm{f}_1, \ldots, \mathrm{f}_p$ mapping $\mathbb{R}^d$ into $\mathbb{R}$, there is a vector $\boldsymbol{\phi}^* \in \mathbb{R}^p$ such that

$$\mathrm{f}^*(\boldsymbol{x}) = \sum_{j=1}^{p} \phi_j^* \, \mathrm{f}_j(\boldsymbol{x}).$$

Furthermore, for a given partition $G_1, \ldots, G_K$ of $\{1, \ldots, p\}$, the vector $\boldsymbol{\phi}^*$ is group-sparse that is

$$\mathrm{Card}(\{k : |\boldsymbol{\phi}^*_{G_k}|_2 \neq 0\}) \ll K.$$

Low dimensional inverse StD assumption

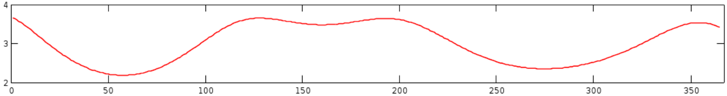
For q given functions $\mathrm{r}_1, \ldots, \mathrm{r}_q$ mapping $\mathbb{R}^d$ into $\mathbb{R}_+$, there is a vector $\boldsymbol{\alpha}^* \in \mathbb{R}^q$ such that

$$\mathrm{r}^*(\boldsymbol{x}) = \sum_{\ell=1}^{q} \alpha_\ell^* \mathrm{r}_\ell(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \in \mathbb{R}^d.$$

Motivations for these assumptions

- **Group sparsity assumption** is relevant in a sparse additive model, that is when
  - $\hookrightarrow$ $\mathrm{f}^*(\boldsymbol{x}) = \psi_1(x_1) + \ldots + \psi_d(x_d)$ s.t. $\psi_j \equiv 0$ for most j,
  - $\hookrightarrow$ projection on basis $\psi_j(x_j) \approx \sum_{\ell=1}^{K_j} \phi^*_{\ell,j} \mathrm{f}_\ell(x_j)$,
  - $\hookrightarrow$ group sparsity of $\boldsymbol{\phi} = (\phi_{\ell,j})$.

- **Low dimensionality of the inverse StD** occurs, for instance when the noise is block-wise homoscedastic or periodic.



Estimation for heteroscedastic regression

Observations : $(\boldsymbol{x}_t, y_t)_{t=1,\ldots,T}$ obeying

$$y_t = \mathrm{b}^*(\boldsymbol{x}_t) + \mathrm{s}^*(\boldsymbol{x}_t)\,\xi_t = \mathrm{r}^*(\boldsymbol{x}_t)^{-1}(\mathrm{f}^*(\boldsymbol{x}_t) + \xi_t).$$

Under our assumptions

$$\mathrm{f}^*(\boldsymbol{x}_t) = \sum\nolimits_{j=1}^{p} \phi_j^*\,\mathrm{f}_j(\boldsymbol{x}_t) = \mathbf{X}(t)\boldsymbol{\phi}^*,$$
$$\mathrm{r}^*(\boldsymbol{x}_t) = \sum\nolimits_{\ell=1}^{q} \alpha_\ell^*\,\mathrm{r}_\ell(\boldsymbol{x}_t) = \mathbf{R}(t)\boldsymbol{\alpha}^*.$$

Thus,

$$\begin{bmatrix} \mathrm{f}^*(\boldsymbol{x}_1) \\ \vdots \\ \mathrm{f}^*(\boldsymbol{x}_T) \end{bmatrix} = \mathbf{X}\boldsymbol{\phi}^* \qquad \text{and} \qquad \begin{bmatrix} \mathrm{r}^*(\boldsymbol{x}_1) \\ \vdots \\ \mathrm{r}^*(\boldsymbol{x}_T) \end{bmatrix} = \mathbf{R}\boldsymbol{\alpha}^*.$$

This leads to

$$\boxed{\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha}^* = \mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi}, \qquad \mathbf{D}_{\boldsymbol{Y}} = \mathrm{diag}(y_1,\ldots,y_T).}$$

Estimation for heteroscedastic regression

Scaled Heteroscedastic Lasso : $(\widehat{\phi}^{\mathrm{ScHeL}}, \widehat{\alpha}^{\mathrm{ScHeL}})$ **solution to**

$$\min_{(\phi,\alpha)} \Big\{ \underbrace{-\sum_{t=1}^{T} \log(\mathbf{R}(t)\alpha) + \frac{1}{2}|\mathbf{D}_{\mathbf{Y}}\mathbf{R}\alpha - \mathbf{X}\phi|_2^2}_{\text{Gaussian negative log-likelihood}} + \underbrace{\sum_{k=1}^{K} \lambda_k |\mathbf{X}_{G_k}\phi_{G_k}|_2}_{\text{sparsity promoting penalty}} \Big\}.$$

- The optimization problem is convex.
- ... but the gradient of the objective is not Lipschitz.

Estimation for heteroscedastic regression

Scaled Heteroscedastic Dantzig selector (ScHeDs) :

$(\widehat{\phi}^{\text{ScHeDs}}, \widehat{\alpha}^{\text{ScHeDs}})$ solution to

$$\min_{(\phi,\alpha)\in\mathbb{R}^{p+q}} \sum_{k=1}^{K} \lambda_k |\mathbf{X}_{G_k}\phi_{G_k}|_2 \qquad \text{s.t.}$$

$$\left|\mathbf{\Pi}_{G_k}\left(\operatorname{diag}(\boldsymbol{Y})\boldsymbol{R}\alpha - \mathbf{X}\phi\right)\right|_2 \le \lambda_k, \qquad \forall k \in \{1,\ldots,K\};$$

$$\sum_{t=1}^{T} \frac{\boldsymbol{R}_{t\ell}}{\boldsymbol{R}_{t,:}\alpha} \le \left(y_t \boldsymbol{R}_{t,:}\alpha - \boldsymbol{X}_{t,:}\phi\right) y_t \boldsymbol{R}_{t\ell}, \qquad \forall \ell \in \{1,\ldots,q\};$$

Theorem : ScHeDs can be solved by an SOCP. Furthermore, the feasible set of this problem is not empty and contains, in particular, the ScHeL.

Comments on the procedure

- Degrees of freedom :
  $\hookrightarrow$ Many tuning parameters in the procedure
  $\hookrightarrow$ Theory : $\lambda_k = \lambda_0\sqrt{r_k}$ with $\lambda_0 > 0$ and $r_k = \text{rank}(\mathbf{X}_{G_k})$
  $\hookrightarrow$ Most papers use $\lambda_k \propto \sqrt{|G_k|} \quad (k = 1, \ldots, K)$

- One can include additional constraints of boundedness of the conditional mean or conditional standard deviation without breaking convexity.

- Bias Correction, practical improvement :
  $\hookrightarrow$ Classical two-steps methods :
  i) our algorithm with $\lambda_k = \lambda_0\sqrt{r_k}$ $(k = 1, \ldots, K)$
  ii) Least squares on the selected variables ($\boldsymbol{\lambda} = 0$)

Comments on the implementation

Several off-the-shelves toolboxes (for instance in Matlab) exist to deal with SOCP

- Sedumi Sturm (1999) : popular interior point methods
  `http://sedumi.ie.lehigh.edu/`
  $\hookrightarrow$ highly accurate solution for moderately large datasets, *e.g.* $p, T \leq 2000$
- Tfocs Becker *et al.* (2011) : first-order proximal method
  `http://cvxr.com/tfocs/`
  $\hookrightarrow$ less accurate (but do we need high accuracy in a noisy setting ?)

  BUT can handle large scale datasets.

Finite sample risk bounds for the ScHeDs

Theorem

Consider the aforementioned heteroscedastic model with sub-Gaussian errors $\boldsymbol{\xi}$. Let K^* (resp. p^*) be the number of relevant groups (resp. corrdinates of ϕ^*). Let $\varepsilon \in (0,1)$ be a tolerance level and set

$$\lambda_k = 4\big(\sqrt{\operatorname{rank}(\mathbf{X}_{G_k})} + \sqrt{\log(K/\varepsilon)}\big).$$

Under some assumptions, with probability at least $1 - 2\varepsilon$,

$$\big|\mathbf{X}(\widehat{\phi} - \phi^*)\big|_2 \le D_{T,\varepsilon}^{3/2}\sqrt{q\log(2q/\varepsilon)} + D_{T,\varepsilon}\sqrt{p^* + K^*\log(K/\varepsilon)}.$$

$$\big|\mathbf{R}(\widehat{\alpha} - \alpha^*)\big|_2 \le D_{T,\varepsilon}^{3/2}\sqrt{q\log(2q/\varepsilon)} + D_{T,\varepsilon}\sqrt{p^* + K^*\log(K/\varepsilon)},$$

where $D_{T,\varepsilon} \propto (\max_t |\mathrm{f}^*(\boldsymbol{x}_t)| + \log(2T/\varepsilon))$.

IV. Numerical experiments

Homoscedastic noise

Data : 500 repetitions :

- Design matrix : $\mathbf{X} \in \mathbb{R}^{T\times p}$ i.i.d. entries $\mathcal{N}(0,1)$
- Noise vector : $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_T, \mathbf{I}_{T\times T})$ independent of $\mathbf{X}$; $\sigma_t \equiv \sigma^*$
- Regression vector : $\beta^0 = [\mathbf{1}_{p^*},\ \mathbf{0}_{p-p^*}]^\top$;
  $\hookrightarrow$ permutation of the entries of β^0 gives β^* ;
- Response vector : $\boldsymbol{Y} = \mathbf{X}\beta^* + \sigma^*\boldsymbol{\xi}$.

Setting : 8 different settings varying (T, p, p^*, σ^*)

Challenger : Square-root Lasso

Tuning parameter : universal choice for both $\lambda = \sqrt{2\log(p)}$ as good in most cases as Cross Validation, *cf.* Sun and Zhang (2012)

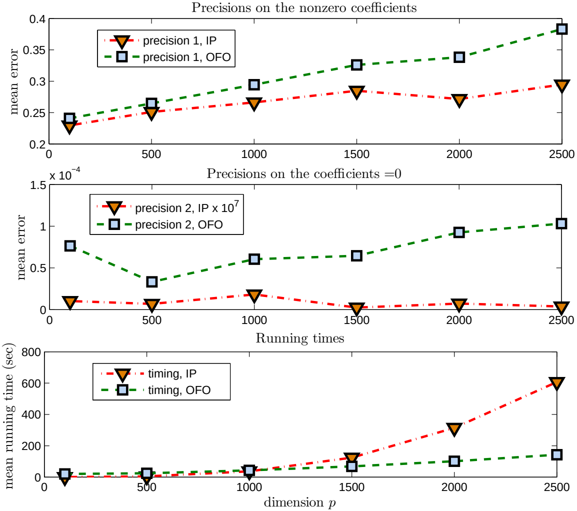


FIGURE: Comparing interior point (IP) *vs.* optimal first-order (OFO) method. Top & middle : MSE of $\widehat{\beta}^{\text{ScHeDs}}$. Bottom : running times.

Experiment with bias correction for the two methods :

ScHeDs (T, p, i^*, σ^*)	$\lvert\widehat{\beta}-\beta^*\rvert_2$ Ave	StD	$\lvert\widehat{p}-p^*\rvert$ Ave	StD	$10\lvert\widehat{\sigma}-\sigma^*\rvert$ Ave	StD
$(100, 100, 2, .5)$	**.06**	.03	**.00**	.00	**.29**	.21
$(100, 100, 5, .5)$	**.11**	.08	**.01**	.12	**.32**	.37
$(100, 100, 2, 1)$	**.13**	.07	**.03**	.16	**.57**	.46
$(100, 100, 5, 1)$	**.28**	.23	**.10**	.33	.77	.68
$(200, 100, 5, .5)$	.08	.02	**.00**	.00	**.23**	.16
$(200, 100, 5, 1)$	**.16**	.05	**.00**	.01	**.09**	.29
$(200, 500, 8, .5)$	**.09**	.03	**.00**	.00	**.22**	.16
$(200, 500, 8, 1)$	.21	.11	**.03**	.17	.48	.43

SqR Lasso	$\lvert\widehat{\beta}-\beta^*\rvert_2$		$\lvert\widehat{p}-p^*\rvert$		$10\lvert\widehat{\sigma}-\sigma^*\rvert$	
$(100, 100, 2, .5)$	.08	.06	.19	.44	.32	.23
$(100, 100, 5, .5)$	.12	.04	.18	.42	.33	.24
$(100, 100, 2, 1)$	.16	.10	.19	.44	.59	.48
$(100, 100, 5, 1)$	.25	.16	.21	.43	**.68**	.47
$(200, 100, 5, .5)$	**.09**	.03	.21	.45	.24	.17
$(200, 100, 5, 1)$	.18	.07	.21	.48	.48	.32
$(200, 500, 8, .5)$	.10	.03	.14	.38	.23	.17
$(200, 500, 8, .5)$	.21	.07	.18	.40	**.46**	.34

Real data : temperature in Paris

Data : daily temperature in Paris from 2003 to 2008 ;
$\hookrightarrow$ National Climatic Data Center (NCDC).

- Response variable y_t : the difference of temperature between two successive days.

- Covariates $\boldsymbol{x}_t = (t, \boldsymbol{u}_t)$: 17 dimensional vector (16+1)
  $\hookrightarrow$ time t ;
  $\hookrightarrow$ increments of temperature over the past 7 days ;
  $\hookrightarrow$ maximal intraday variation of temperature over the past 7 days ;
  $\hookrightarrow$ wind speed of the day before.

Construction of $\boldsymbol{R}$: $T \times 11$ matrix with columns r_ℓ.

$$\mathrm{r}_1(\boldsymbol{x}_t) = 1; \qquad \mathrm{r}_2(\boldsymbol{x}_t) = t; \qquad \mathrm{r}_3(\boldsymbol{x}_t) = 1/(t + 2 \times 365)^{\frac{1}{2}};$$
$$\mathrm{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell - 3)t/365); \qquad \ell = 4, \dots, 7;$$
$$\mathrm{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell - 7)t/365); \qquad \ell = 8, \dots, 11.$$

Construction of $\mathbf{X}$: $T \times 2176$ matrix with columns f_j. $\hookrightarrow$
Time-varying second-order polynomial in $\boldsymbol{u}_t$:

$$\mathrm{f}_j(t) = \psi_\ell(t) \times \chi_{m,m'}(\boldsymbol{u}_t);$$
$$|\{\mathrm{f}_j\}| = 16 \times 16 \times 17/2 = 2176.$$

Construction of groups : 136 groups of 16 functions

$$\mathcal{G}_{m,m'} = \{\psi_\ell(t) \times \chi_{m,m'}(\boldsymbol{u}_t) : \ell = 1, \dots, 16\}.$$

Results

Samples :

$\hookrightarrow$ Training set : temperatures from 2003 to 2007 (that is, 2172 values) ;

$\hookrightarrow$ Test set : temperatures from 2008 (that is, 366 values, leap year).

Conclusions of the study :

- Dimension reduction : from 2176 to 26 ;
- Sign estimation : 62% of right estimation ;
- Volatility estimation : the oscillation of the temperature during the period between May and July is significantly higher than in March, September and October ;

Summary

New procedures named ScHeL and ScHeDs :

- Suitable for fitting the heteroscedastic regression model.
- Simultaneous estimation of the mean and the variance functions.
- Takes into account group sparsity.
- Implemented using two different solvers :
  $\hookrightarrow$ primal-dual interior point method (highly accurate),
  $\hookrightarrow$ optimal first-order method (moderately accurate but with cheap iterations).
- Competitive with state-of-the art algorithms
  $\hookrightarrow$ applicable in a much more general framework.

Manuscript is available on arxiv, codes are available on request.


Thank You

References I

- A. Antoniadis, Comments on : ℓ_1-penalization for mixture regression models, TEST **19** (2010), no. 2, 257–258. MR 2677723
- S. R. Becker, E. J. Candès, and M. C. Grant, Templates for convex cone problems with applications to sparse signal recovery, Mathematical Programming Computation **3** (2011), no. 3, 165–218.
- A. Belloni, V. Chernozhukov, and L. Wang, Square-root Lasso : Pivotal recovery of sparse signals via conic programming, Biometrika **98** (2011), no. 4, 791–806.
- P. J. Bickel, Y. Ritov, and A. B. Tsybakov, Simultaneous analysis of Lasso and Dantzig selector, Ann. Statist. **37** (2009), no. 4, 1705–1732.
- E. J. Candès and T. Tao, The Dantzig selector : statistical estimation when p is much larger than n, Ann. Statist. **35** (2007), no. 6, 2392–2404.
- Arnak S. Dalalyan and Yin Chen, Fused sparsity and robust estimation for linear models with unknown variance, Advances in Neural Information Processing Systems 25 : NIPS, 2012, pp. 1268–1276.
- E. Gautier and A. Tsybakov, High-dimensional instrumental variables regression and confidence sets, September 2011.
- N. Städler, P. Bühlmann, and Sara s van de Geer, ℓ_1-penalization for mixture regression models, TEST **19** (2010), no. 2, 209–256.
- J. F. Sturm, Using sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones, Optimization Methods and Software **11–12** (1999), 625–653.
- T. Sun and C.-H. Zhang, Scaled sparse linear regression, Biometrika **99** (2012), no. 4, 879–898.
- R. Tibshirani, Regression shrinkage and selection via the Lasso, J. Roy. Statist. Soc. Ser. B **58** (1996), no. 1, 267–288.