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

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

SAPS IX, Le Mans, FRANCE

Arnak S. Dalalyan

ENSAE / CREST / GENES

I. Problem presentation

Continuous-time autoregression

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

$$\dot{y}_t = \mathrm{b}^*(\boldsymbol{x}_t) + \mathrm{s}^*(\boldsymbol{x}_t)\,\dot{W}_t, \qquad t \in \mathcal{T} = [0, T],$$

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

$$\text{Drift} : \ \mathrm{b}^*(\boldsymbol{x}_t) = \mathbf{E}\big[\dot{y}_t \big| \boldsymbol{x}_t\big].$$
$$\text{Volatility} : \ \mathrm{s}^{*2}(\boldsymbol{x}_t) = \mathbf{Var}\big[\dot{y}_t \big| \boldsymbol{x}_t\big].$$

$(W_t)_{t\geq 0}$ is a standard Wiener process (w.r.t a filtration $(\mathcal{F}_t)_{t\geq 0}$).

The goal is to estimate the function b^*.

Discrete time (auto-)regression

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

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

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

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

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.

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

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

$$\forall \boldsymbol{x}, \quad b^*(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(\boldsymbol{x})\beta_j^* = [f_1(\boldsymbol{x}), \ldots, f_p(\boldsymbol{x})]\boldsymbol{\beta}^*, \quad \text{and} \quad s^*(\boldsymbol{x}) \equiv \sigma^*$$

$\hookrightarrow$ Dictionary $\{f_1, \ldots, f_p\}$ 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 $p^* \ll p$

Main assumptions on b^* and s^*

Re-parametrize by the inverse of the conditional volatility s^*

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

Group Sparsity Assumption

For p given functions $f_1, \ldots, f_p$ mapping $\mathbb{R}^d$ into $\mathbb{R}$, there is a vector $\phi^* \in \mathbb{R}^p$ such that $f^*(\boldsymbol{x}) = \sum_{j=1}^p \phi_j^* f_j(\boldsymbol{x})$. Furthermore, for a given partition $G_1, \ldots, G_K$ of $\{1, \ldots, p\}$, the vector ϕ^* is group-sparse that is $\mathrm{Card}(\{k : |\phi^*_{G_k}|_2 \neq 0\}) \ll K$.

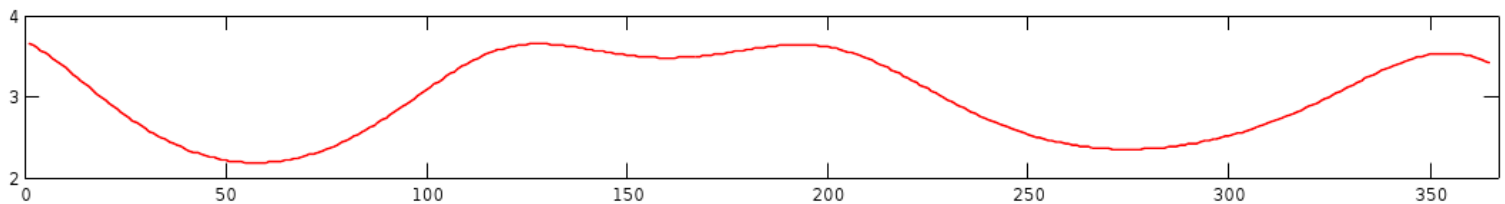
Low dimensional volatility assumption

For q given functions $r_1, \ldots, r_q$ mapping $\mathbb{R}^d$ into $\mathbb{R}_+$, there is a vector $\alpha^* \in \mathbb{R}^q$ such that $r^*(\boldsymbol{x}) = \sum_{\ell=1}^q \alpha_\ell^* r_\ell(\boldsymbol{x})$ for almost every $\boldsymbol{x} \in \mathbb{R}^d$.

Motivations for these assumptions

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

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



II. Relation to previous work

Homoscedastic regression

The model is

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

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 :	$\beta^* = \left[\beta_1^*, \ldots, \beta_p^*\right]^\top \in \mathbb{R}^p$
Standard deviation :	$\mathrm{s}^*(\boldsymbol{x}_t) \equiv \sigma^* \in \mathbb{R}_*^+$

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

Most popular methods : Lasso and Dantzig selector

- The LASSO of Tibshirani (1996) is defined as

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

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

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

For a tuning parameter satisfying $\lambda \propto 1/\sigma^*$, sharp oracle inequalities are available, *e.g.*, Bickel *et al.* (2009).

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

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

Joint estimation of β^* and σ^*

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

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

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

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

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

$$\widehat{\beta}^{\text{SqR-Lasso}} = \arg\min_{\beta} \left(\left| \boldsymbol{Y} - \mathbf{X}\beta \right|_2 + \lambda \sum_{j=1}^{p} |\boldsymbol{X}_{:,j}|_2 |\beta_j| \right)$$

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

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

III. Main results

Estimation for heteroscedastic regression

Continuous time

Observations : $(\boldsymbol{x}_t, y_t)_{t \le T}$ obeying $\dot{y}_t = \mathrm{b}^*(\boldsymbol{x}_t) + \mathrm{s}^*(\boldsymbol{x}_t)\, \dot{W}_t$, thus

$$\frac{\dot{y}_t}{\mathrm{s}^*(\boldsymbol{x}_t)} = \left(\sum_{j=1}^{p} \phi_j^* \, \mathrm{f}_j(\boldsymbol{x}_t) \right) + \dot{W}_t = \mathbf{X}(t)\phi^* + \dot{W}_t, \quad t \in \mathcal{T} = [0, T].$$

Scaled Heteroscedastic Lasso :

$$\widehat{\phi}^{\mathrm{ScHeL}} = \underset{\phi \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{2} \int_0^T \left(\frac{\dot{y}_t}{\mathrm{s}^*(\boldsymbol{x}_t)} - \mathbf{X}(t)\phi \right)^2 dt + \lambda \sum_{k=1}^{K} |\mathbf{X}_{G_k} \phi_{G_k}|_2 \right\},$$

with $|\mathbf{X}_G \phi_G|_2^2 = \int_0^T \left(\sum_{j \in G} \mathbf{X}_j(t) \phi_j \right)^2 dt = \int_0^T \left(\sum_{j \in G} \mathrm{f}_j(\boldsymbol{x}_t) \phi_j \right)^2 dt$.

Computation : the estimator $\widehat{\phi}^{\mathrm{ScHeL}}$ can be efficiently computed even for very large dimensions p by solving a second-order cone program.

Estimation for heteroscedastic regression

Discrete time

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

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

Under our assumptions

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

Thus, $[\mathsf{f}^*(\boldsymbol{x}_1), \dots, \mathsf{f}^*(\boldsymbol{x}_T)]^\top = \mathbf{X}\boldsymbol{\phi}^*$ and $[\mathsf{r}^*(\boldsymbol{x}_1), \dots, \mathsf{r}^*(\boldsymbol{x}_T)]^\top = \mathbf{R}\boldsymbol{\alpha}^*$. This leads to

$$\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha}^* = \mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi}, \qquad \mathbf{D}_{\boldsymbol{Y}} = \mathrm{diag}(y_t;\, t = 1, \dots, T).$$

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

$$\min_{\boldsymbol{\phi}\in\mathbb{R}^p} \left\{ \underbrace{-\sum_{t=1}^{T} \log(\mathbf{R}(t)\boldsymbol{\alpha}) + \frac{1}{2}\,|\mathbf{D}_{\boldsymbol{Y}}\mathbf{R}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\phi}|_2^2}_{\text{Gaussian log-likelihood}} + \underbrace{\lambda \sum_{k=1}^{K} |\mathbf{X}_{G_k}\boldsymbol{\phi}_{G_k}|_2}_{\text{sparsity promoting penalty}} \right\}.$$

Finite sample risk bounds for the ScHeL

Theorem Consider either continuous-time model or discrete-time with sub-Gaussian errors ξ. 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

$$\lambda_k = 4\big(\sqrt{|G_k|} + \sqrt{\log(K/\varepsilon)}\big).$$

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

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, proofs and codes available on request.

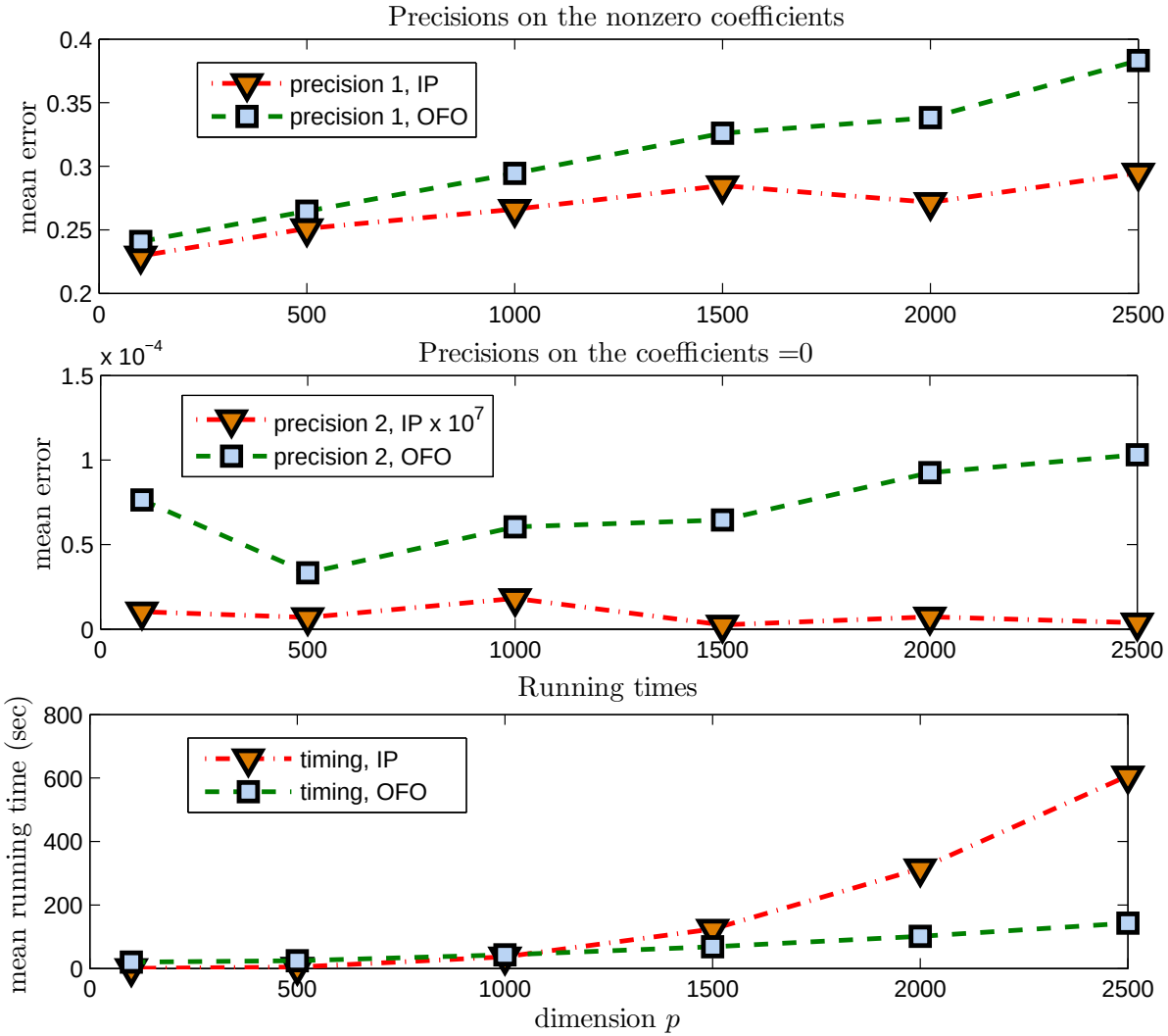


FIGURE: Comparing interior point (IP) *vs.* optimal first-order (OFO) method. Top & middle : MSE of $\widehat{\beta}^{\text{ScHeDs}}$. Bottom : running times.

References I

- A. Antoniadis, Comments on : ℓ_1-penalization for mixture regression models, TEST **19** (2010), no. 2, 257–258. MR 2677723
- 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.
- 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.


Thank You