Learning Heteroscedastic Models by Conic Programming under Group Sparsity

Joseph Salmon
Télécom ParisTech
http://josephsalmon.eu/

Joint work with: Arnak Dalalyan (ENSAE-CREST),
Mohamed Hebiri (Université Paris-Est),
Katia Meziani (Université Paris-Dauphine)

Outline

Heteroscedastic regression

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

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

- Conditional mean: $\mathsf{b}^* : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbf{E}[y_t|\boldsymbol{x}_t] = \mathsf{b}^*(\boldsymbol{x}_t)$
- Conditional variance: ${\mathsf{s}^*}^2 : \mathbb{R}^d \to \mathbb{R}_+$ such that $\mathbf{Var}[y_t|\boldsymbol{x}_t] = {\mathsf{s}^*}^2(\boldsymbol{x}_t)$
- Normalized errors: ξ_t such that $\mathbf{E}[\xi_t|\boldsymbol{x}_t] = 0$ and $\mathbf{Var}[\xi_t|\boldsymbol{x}_t] = 1$ (*e.g.* Gaussian for simplicity)

$\hookrightarrow$ Including the "time-dependent" mean and variance case: consider $[t; \boldsymbol{x}_t^\top]^\top$ instead of $\boldsymbol{x}_t$ as explanatory variable

Sparsity **Assumption**

- Estimating b^* and s^* is ill-posed
- sparsity senario: b^* and s^* belong to low dimensional spaces

Example: Homoscedastic regression

$$\forall \boldsymbol{x}, \quad \mathsf{b}^*(\boldsymbol{x}) = [\mathsf{f}_1(\boldsymbol{x}), \dots, \mathsf{f}_p(\boldsymbol{x})]\boldsymbol{\beta}^*, \qquad \text{and} \quad \mathsf{s}^*(\boldsymbol{x}) \equiv \sigma^*$$

$\hookrightarrow$ Dictionary $\{\mathsf{f}_1, \dots, \mathsf{f}_p\}$ 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: $i^* = |\boldsymbol{\beta}^*|_0 := \sum_{j=1}^p \mathbb{1}(\boldsymbol{\beta}_j^* \neq 0)$ with $i^* \ll T$

Homoscedastic case with known noise level

Regression formulation

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

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:	$\mathbf{X}_{t,j} = [\mathsf{f}_j(\boldsymbol{x}_t)] \in \mathbb{R}$
Coefficients:	$\boldsymbol{\beta}^* = \left[\beta_1^*, \ldots, \beta_p^*\right]^\top \in \mathbb{R}^p$
Standard deviation:	$\mathsf{s}^*(\boldsymbol{x}_t) \equiv \sigma^* \in \mathbb{R}_*^+$

<u>REM</u>:

- $\boldsymbol{Y}$ is observed
- $\mathbf{X}$ is known or chosen by the statistician
- $\boldsymbol{\beta}^*$ is to be recovered by $\hat{\beta}$

Pioneer methods: homoscedastic, σ^* known

LASSO Tibshirani (1996)

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\left(\frac{|\boldsymbol{Y}-\mathbf{X}\boldsymbol{\beta}|_2^2}{2T}+\lambda\sum_{j=1}^{p}|\boldsymbol{X}_{:,j}|_2|\beta_j|\right)$$

Dantzig-Selector Candès and Tao (2007)

$$\underset{\boldsymbol{\beta}|_2\in\mathbb{R}^p}{\arg\min}\left\{\sum_{j=1}^{p}|\boldsymbol{X}_{:,j}|_2|\beta_j| : \text{s.t.}\forall j=1,\cdots,p,\ \frac{|\boldsymbol{X}_{:,j}^\top(Y-\boldsymbol{X}\beta)|}{|\boldsymbol{X}_{:,j}|_2}\leq\lambda\right\}$$

Oracle inequalities (non-asymptotic bounds) available *e.g.* Bickel *et al.* (2009) for a tuning parameter satisfying $\lambda \propto \sigma^*$,
BUT knowledge of σ^* needed!

Homoscedastic case with unknown noise level

Matrix/vector formulation

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

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

REM:

- $\boldsymbol{Y}$ is observed,
- $\mathbf{X}$ is known or chosen by the statistician
- $\boldsymbol{\beta}^*$ and σ^* are to be recovered by $\hat{\beta}$ and $\hat{\sigma}$

Pioneering methods: homoscedastic, σ^* unknown

Scaled-Lasso, Städler *et al.* (2010)

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

↪ penalized (Gaussian, negative) log-likelihood minimization
↪ can be recast in a convex problem (do $\rho := \frac{1}{\sigma}$ and $\phi := \frac{\beta}{\sigma}$):

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

- **equivariant** estimator, *i.e.* if $\boldsymbol{Y} \leftarrow c\,\boldsymbol{Y}, \boldsymbol{\beta}^* \leftarrow c\boldsymbol{\beta}^*, \sigma^* \leftarrow c\sigma^*$, then $\hat{\boldsymbol{\beta}} \leftarrow c\hat{\boldsymbol{\beta}}$ and $\hat{\sigma} \leftarrow c\hat{\sigma}$
- Jointly convex problem but not a simple one (Linear Programming, etc.)

Pioneering methods: homoscedastic, σ^* unknown

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

$$\widehat{\beta}^{\text{SqR-Lasso}} = \arg\min_{\beta} \left(\frac{|\boldsymbol{Y} - \mathbf{X}\beta|_2}{2\sqrt{T}} + \lambda \sum_{j=1}^{p} |\boldsymbol{X}_{:,j}|_2 |\beta_j| \right)$$

$$\hat{\sigma}^* = \frac{1}{\sqrt{T}} |\boldsymbol{Y} - \mathbf{X}\widehat{\beta}^{\text{SqR-Lasso}}|_2$$

↪ equivalent to sequentially minimizing the following

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

- Can be solved by a **S**econd **O**rder **C**one **P**rogram (SOCP)
- Not easily extended to the heteroscedastic case
- Extension to matrix completion Klopp (2011)

Objectives

Extending previous works Dalalyan and Chen (2012) , propose a method for **jointly** estimating:

- the conditional mean function b^*
- the conditional volatility s^*

$\hookrightarrow$ for the **heteroscedastic** regression

$\hookrightarrow$ **without any knowledge** on the noise level

Problem re-formulation

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

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

Assumptions on the model (I)

Group Sparsity Assumption

For a given family $G_1, \ldots, G_K$ of disjoint subsets of $\{1, \ldots, p\}$, there is a vector $\boldsymbol{\phi}^* \in \mathbb{R}^p$ such that

$$[\mathsf{f}^*(\boldsymbol{x}_1), \ldots, \mathsf{f}^*(\boldsymbol{x}_T)]^\top = \mathbf{X}\boldsymbol{\phi}^*, \qquad \mathsf{Card}(\{k : |\boldsymbol{\phi}^*_{G_k}|_2 \neq 0\}) \ll K.$$

Sparse vector:

Group Sparse vector:

REM: Note that the groups have not necessarily the same size

Examples of application I

Group sparsity assumption (I)

- Sparse linear model with categorical data
  $\hookrightarrow$ linear regression with qualitative covariates
  $\hookrightarrow$ each covariate has several modalities

- Sparse additive model
  $\hookrightarrow$ $\mathsf{f}^*(\boldsymbol{x}) = \mathsf{f}_1^*(x_1) + \ldots + \mathsf{f}_d^*(x_d)$; $\mathsf{f}_j^* \equiv 0$ for most j
  $\hookrightarrow$ Projection on a basis:
  $\mathsf{f}_j^*(x) \approx \sum_{\ell=1}^{K_j} \phi_{\ell,j}\psi_\ell(x)$: group sparsity of $\boldsymbol{\phi} = (\phi_{\ell,j})$.

Assumptions on the model (II)

Low dimension volatility assumption

For q given functions $\mathsf{r}_1, \ldots, \mathsf{r}_q$ mapping $\mathbb{R}^d$ into $\mathbb{R}_+$, there is a vector $\boldsymbol{\alpha}^* \in \mathbb{R}^q$ such that $\mathsf{r}^*(\boldsymbol{x}) = \sum_{\ell=1}^q \alpha_\ell^* \mathsf{r}_\ell(\boldsymbol{x})$ for almost every $\boldsymbol{x} \in \mathbb{R}^d$, and $\mathcal{S}$ is the linear span of $\mathsf{r}_1, \ldots, \mathsf{r}_q$.

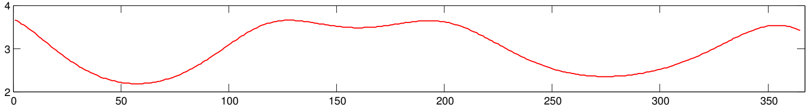
$$\left[\mathsf{r}^*(\boldsymbol{x}_1), \ldots, \mathsf{r}^*(\boldsymbol{x}_T)\right]^\top = \boldsymbol{R}\boldsymbol{\alpha}^*$$

<u>REM</u>: here and after $q \ll T$

Examples of application (II)

Low dimension volatility assumption

- Block-wise homoscedastic noise
  $\hookrightarrow$ r^* is well approximated by a piecewise constant function: time series modeling (smooth variations over time), image processing (neighboring pixels are often corrupted by noise levels of similar magnitude).
- Periodic noise-level
  $\hookrightarrow$ r^* belongs to the linear span of a few trigonometric functions: meteorology (seasonal variations), image processing (electronic disturbance of repeating nature, caused for instance by an electric motor).



Penalized $\log$-likelihood formulation

- penalized log-likelihood used for defining the group-Lasso
  - Tuning parameter: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}_+^K$

Introduce the $T \times q$ matrix $\boldsymbol{R} = (\mathsf{r}_\ell(\boldsymbol{x}_t))_{t,\ell}$
The cost function becomes $\mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha})$:

$$\mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = -\sum_{t=1}^{T} \log(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}) + \frac{1}{2}\sum_{t=1}^{T}\left(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\right)^2 + \sum_{k=1}^{K} \lambda_k \left|\boldsymbol{X}_{:,G_k}\boldsymbol{\phi}_{G_k}\right|_2$$

- REM: use penalty $\sum_{k=1}^{K} \lambda_k \left|\boldsymbol{X}_{:,G_k}\boldsymbol{\phi}_{G_k}\right|_2$ instead of $\sum_{k=1}^{K} \lambda_k \left|\boldsymbol{\phi}_{G_k}\right|_2$
  Simon and Tibshirani (2012)

Optimization considerations

- Minimization of PL can be seen as a log-det problem
  ↪ But higher computational complexity than **L**inear **P**rogramming (LP) and SOCP
- Reduce computation cost
  ↪ Dantzig Selector arguments;
  ↪ First-order conditions:

$$\forall k \in \{1, \dots, K\}, \qquad \frac{\partial}{\partial \phi_{G_k}} \mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = 0 \tag{1}$$

$$\forall \ell \in \{1, \dots, q\}, \qquad \frac{\partial}{\partial \alpha_\ell} \mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = 0 \tag{2}$$

First order conditions (1)

- $\forall k \in \{1, \ldots, K\}$, $\frac{\partial}{\partial \phi_{G_k}} \mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = 0$ implies:

$$-\mathbf{X}_{:,G_k}^{\top}(\text{diag}(\boldsymbol{Y})\boldsymbol{R\alpha} - \mathbf{X}\boldsymbol{\phi}) + \lambda_k \mathbf{X}_{:,G_k}^{\top} \frac{\mathbf{X}_{:,G_k}\boldsymbol{\phi}_{:,G_k}}{\left|\mathbf{X}_{:,G_k}\boldsymbol{\phi}_{:,G_k}\right|_2} = 0$$

$\hookrightarrow$ True if $\min_k |\boldsymbol{X}_{:,G_k}\boldsymbol{\phi}_{:,G_k}|_2 \neq 0$
$\hookrightarrow$ Difficult problem: non-linear part

- Equivalence with

$$\mathbf{\Pi}_{G_k}(\text{diag}(\boldsymbol{Y})\boldsymbol{R\alpha} - \mathbf{X}\boldsymbol{\phi}) = \lambda_k \mathbf{X}_{:,G_k}\boldsymbol{\phi}_{G_k} / |\mathbf{X}_{:,G_k}\boldsymbol{\phi}_{G_k}|_2$$

$\mathbf{\Pi}_{G_k} = \mathbf{X}_{:,G_k}(\mathbf{X}_{:,G_k}^{\top}\mathbf{X}_{:,G_k})^{+}\mathbf{X}_{:,G_k}^{\top}$: projector on $\mathrm{Span}(\mathbf{X}_{:,G_k})$

"Convexification" : $\left|\mathbf{\Pi}_{G_k}(\text{diag}(\boldsymbol{Y})\boldsymbol{R\alpha} - \mathbf{X}\boldsymbol{\phi})\right|_2 \leq \lambda_k$

First order conditions (2)

- $\forall \ell = 1, \ldots, q$, $\frac{\partial}{\partial \alpha_\ell} \mathsf{PL}(\boldsymbol{\phi}, \boldsymbol{\alpha}) = 0$ implies:
  $\exists\ \boldsymbol{\nu} \in \mathbb{R}_+^T$ such that

$$-\sum_{t=1}^{T} \frac{\boldsymbol{R}_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}} + \sum_{t=1}^{T} \left(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\right) y_t \boldsymbol{R}_{t\ell} - \boldsymbol{\nu}^\top \boldsymbol{R}_{:,\ell} = 0$$

  and $\nu_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} = 0$ for every t.

"Convexification" : $$\sum_{t=1}^{T} \frac{\boldsymbol{R}_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}} - \left(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\right) y_t \boldsymbol{R}_{t\ell} \leq 0$$

Relaxation

Scaled Heteroscedastic Dantzig selector (ScHeDs)

Minimizing with respect to $(\boldsymbol{\phi}, \boldsymbol{\alpha}) \in \mathbb{R}^p \times \mathbb{R}^q$ the problem

$$\min_{\phi,\alpha} \quad \sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\phi}_{G_k}\big|_2, \qquad s.t.$$

$$\Big|\mathbf{\Pi}_{G_k}\big(\mathrm{diag}(\boldsymbol{Y})\boldsymbol{R}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\phi}\big)\Big|_2 \leq \lambda_k, \qquad \forall k \in \{1, \ldots, K\};$$

$$\sum_{t=1}^{T} \frac{\boldsymbol{R}_{t\ell}}{\boldsymbol{R}_{t,:}\boldsymbol{\alpha}} - \big(y_t \boldsymbol{R}_{t,:}\boldsymbol{\alpha} - \boldsymbol{X}_{t,:}\boldsymbol{\phi}\big) y_t \boldsymbol{R}_{t\ell} \leq 0, \qquad \forall \ell \in \{1, \ldots, q\};$$

Theorem: ScHeDs can be solved by an SOCP.

REM: The feasible set of this problem is not empty and contains, in particular, all the minimizers of the penalized log-likelihood.

Homoscedastic case

Scaled Homoscedastic Dantzig selector (ScHeDs)

Minimizing with respect to $(\boldsymbol{\phi}, \rho) \in \mathbb{R}^p \times \mathbb{R}$ the problem

$$\min_{\phi,\rho} \quad \sum_{k=1}^{K} \lambda_k \big|\mathbf{X}_{:,G_k} \boldsymbol{\phi}_{G_k}\big|_2, \qquad s.t.$$

$$\Big|\mathbf{\Pi}_{G_k}\big(\mathrm{diag}(\boldsymbol{Y})\rho - \mathbf{X}\boldsymbol{\phi}\big)\Big|_2 \leq \lambda_k, \qquad\qquad \forall k \in \{1, \ldots, K\};$$

$$T - \rho\big(\boldsymbol{Y}\rho - \boldsymbol{X}\boldsymbol{\phi}\big)^\top \boldsymbol{Y} \leq 0,$$

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 $\lambda_k \propto \sqrt{|G_k|}\,(k = 1, \dots, K)$
- Bias correction, practical improvement:
  - $\hookrightarrow$ Classical two-steps methods:
    - i) our algorithm with $\lambda_k = \lambda_0 \sqrt{r_k}$ (k=1,. . . ,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 method
  `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 dimension,
  e.g. $p = 5000$ and $T = 3000$
  REM: early stopping could lead to better solutions than Sedumi

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: $\mathbb{R}^T \ni \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_T, \mathbf{I})$ independent of $\mathbf{X}$; $\sigma_t \equiv \sigma^*$
- Regression vector: $\boldsymbol{\beta}^0 = [\mathbf{1}_{i^*},\ \mathbf{0}_{p-i^*}]^\top$;

  $\hookrightarrow$ permutation of the entries of $\boldsymbol{\beta}^0$ gives $\boldsymbol{\beta}^*$;
- Response vector: $\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma^* \boldsymbol{\xi}$.

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

Experiment with bias correction for the two methods:

ScHeDs	$\|\widehat{\boldsymbol\beta}-\boldsymbol\beta^*\|_2$		$\|\widehat{i}-i^*\|$		$10\|\widehat\sigma-\sigma^*\|$	
$(\;T,\;\; p,\; i^*,\, \sigma^*)$	Ave	StD	Ave	StD	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

Square-root Lasso	$\|\widehat{\boldsymbol\beta}-\boldsymbol\beta^*\|_2$		$\|\widehat{i}-i^*\|$		$10\|\widehat\sigma-\sigma^*\|$	
$(\;T,\;\; p,\; i^*,\, \sigma^*)$	Ave	StD	Ave	StD	Ave	StD
$(100, 100, 2, .5)$	.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

Heteroscedastic (without blocks)

Data:

- Design matrix: $\mathbf{X} \in \mathbb{R}^{T \times p}$ i.i.d. entries $\mathcal{N}(0,1)$
- Noise vector: $\mathbb{R}^T \ni \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_T, \mathbf{I})$ independent of $\mathbf{X}$
- Variances: piecewise constant with blocks of length $T/10$
  1st block $\sigma_t \equiv 8\sigma^*$; 5th block $\sigma_t \equiv 4\sigma^*$;
  9th block $\sigma_t \equiv 5\sigma^*$; others 7 blocks have $\sigma_t \equiv \sigma^*$;
- $\boldsymbol{\beta}^* = (2, 3, 3, 3, 1.5, 1.5, 1.5, 0, 0, 0, 2, 2, 2, 0, \cdots, 0)^\top \in \mathbb{R}^p$
- Response vector: $y_t = \mathbf{X}_{t,:}\boldsymbol{\beta}^* + \sigma_t \boldsymbol{\xi}_t$.

Challenger: Square-root Lasso Belloni et al. (2011)
HRR (High dim. Heteroscedastic Regression) Daye et al. (2011)

Tuning parameters: "universal choice" $\lambda = \sqrt{2\log(p)}$;
$\boldsymbol{R}$: encodes blocks of size $T/20$ (*i.e.* $q = 20$)

Heteroscedastic noise

Prediction error $\frac{\|\mathbf{X}\hat{\beta}-\mathbf{X}\beta^*\|_2}{\sqrt{T}}$ (or $\|(\mathbf{X}\hat{\phi})./(\boldsymbol{R}\hat{\boldsymbol{\alpha}}) - \mathbf{X}\beta^*\|_2/\sqrt{T}$)

	Sqrt-Lasso	Sqrt-Lasso Deb.	Daye	ScHeDs	ScHeDs Deb.
T	$\sigma = 4,\ p = 500$				
100	6.37	5.92	**2.99**	5.61	6.17
200	6.26	4.48	**2.44**	4.89	3.75
500	3.75	**2.15**	2.36	2.33	2.33
T	$\sigma = 6,\ p = 500$				
100	7.67	7.67	**3.75**	6.44	5.43
200	6.82	6.32	**2.34**	4.54	3.21
500	5.73	3.92	8.24	2.98	**2.34**
T	$\sigma = 8,\ p = 500$				
100	7.55	7.55	**3.96**	6.69	6.16
200	6.68	6.46	**2.90**	4.62	4.68
500	6.53	5.23	10.21	3.91	**3.20**
T	$\sigma = 10,\ p = 500$				
100	7.53	7.53	**4.53**	5.99	7.63
200	6.84	6.84	**4.88**	5.92	4.95
500	6.55	5.31	5.21	3.94	**3.52**

Heteroscedastic noise

Prediction error $\frac{\|\mathbf{X}\hat{\beta}-\mathbf{X}\beta^*\|_2}{\sqrt{T}}$ (or $\|(\mathbf{X}\hat{\phi})./(\boldsymbol{R}\hat{\boldsymbol{\alpha}}) - \mathbf{X}\beta^*\|_2/\sqrt{T}$)

	Sqrt-Lasso	Sqrt-Lasso Deb.	Daye	ScHeDs	ScHeDs Deb.
T			$\sigma = 4,\ p = 200$		
100	6.00	5.18	**2.20**	5.53	5.80
200	6.05	5.53	**1.88**	4.90	4.74
500	4.08	**2.06**	2.26	2.55	2.21
T			$\sigma = 6,\ p = 200$		
100	7.77	7.77	6.96	**6.57**	7.14
200	6.75	6.17	**2.97**	5.02	3.63
500	5.08	2.78	3.80	2.77	**2.64**
T			$\sigma = 8,\ p = 200$		
100	7.28	7.28	9.35	6.38	**4.99**
200	6.94	6.94	5.96	4.61	**3.25**
500	5.46	5.10	4.95	3.59	**2.94**
T			$\sigma = 10,\ p = 200$		
100	6.01	6.91	**5.14**	5.30	9.15
200	7.14	7.14	11.11	5.52	**5.12**
500	6.53	6.43	6.07	4.21	**3.46**

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.

$$\mathsf{r}_1(\boldsymbol{x}_t) = 1; \qquad \mathsf{r}_2(\boldsymbol{x}_t) = t;$$

$$\mathsf{r}_3(\boldsymbol{x}_t) = 1/(t + 2 \times 365)^{\frac{1}{2}};$$

$$\mathsf{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell - 3)t/365); \qquad \ell = 4, \ldots, 7;$$

$$\mathsf{r}_\ell(\boldsymbol{x}_t) = 1 + \cos(2\pi(\ell - 7)t/365); \qquad \ell = 8, \ldots, 11.$$

Construction of $\mathbf{X}$: $t \times 2176$ matrix with columns f_j.

$$\chi_{m,m'}(\boldsymbol{u}_t) = u_t^m u_t^{m'}, \qquad \text{with } 1 \leq m \leq m'2 \text{ and } m + m' = 2;$$
$$\psi_1(t) = 1;$$
$$\psi_\ell(t) = t^{1/(\ell-1)}, \qquad \ell = 2, 3, 4;$$
$$\psi_\ell(t) = \cos(2\pi(\ell - 4)t/365); \qquad \ell = 5, \ldots, 10;$$
$$\psi_\ell(t) = \sin(2\pi(\ell - 10)t/365); \qquad \ell = 11, \ldots, 16.$$

$\hookrightarrow$ Time-varying second-order polynomial in $\boldsymbol{u}_t$:

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

Construction of groups: 136 groups of 16 functions

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

$\triangleright$ **This construction is arbitrary.**

Results

Samples:

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

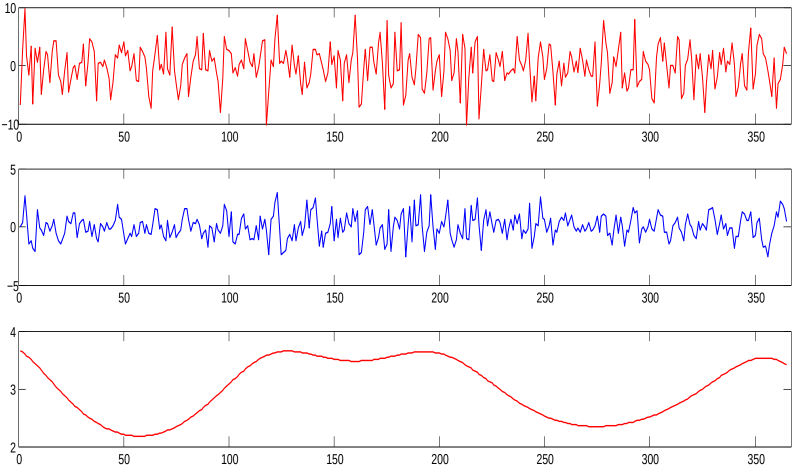
↪ Test set: temperatures from 2008 (that is, 366 values, leap year).

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;

Illustration

1) Increments observed in 2008;
2) Our prediction of these increments;
3) Noise level estimation.



Finite sample risk bound

Theorem

Under the **(GRE)** + assumptions on signal/noise ratio for any $\epsilon > 0$, w.p. $1 - \epsilon$, the ScHeDs estimator satisfies

$$|\boldsymbol{X}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^*)|_2 \precsim \left(\frac{1}{\kappa} \sqrt{i_{\boldsymbol{\phi}^*} + |\mathcal{K}^*| \log(\frac{K}{\epsilon})} + \sqrt{q \log(\frac{q}{\epsilon})} \right) D_{T,\delta}^{3/2}$$

$$\frac{|\boldsymbol{R}(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}^*)|_2}{|\boldsymbol{R}\boldsymbol{\alpha}^*|_\infty} \precsim \left(\frac{1}{\kappa} \sqrt{i_{\boldsymbol{\phi}^*} + |\mathcal{K}^*| \log(\frac{K}{\epsilon})} + \sqrt{q \log(\frac{q}{\epsilon})} \right)) D_{T,\delta}^{3/2}$$

with $D_{T,\delta} = \log(\frac{T}{\delta})$ and $i_{\boldsymbol{\phi}^*} = \sum_{k=1}^{K} \mathrm{rank}(\mathrm{X}_{:,\mathrm{G}_\mathrm{k}})$

REM:

- assumptions on the signal/noise ratio only needed for the theorem, not for the construction of the estimator.

Summary

New procedure named ScHeDs:

- Suitable for fitting the heteroscedastic regression model
- Simultaneous estimation of the mean and the variance functions;
- Takes into account group sparsity;
- Relaxation of first-order conditions for maximum penalized likelihood estimation
  - ↪ existence of a solution;
  - ↪ convex problem – second-order cone programming
- Competitive with state-of-the art algorithms
  - ↪ applicable in a much more general framework.

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, 2313–2351.

- A. S. Dalalyan and Y. Chen, *Fused sparsity and robust estimation for linear models with unknown variance*, NIPS, 2012, pp. 1268–1276.

References II

- J. Daye, J. Chen, and H. Li, *High-dimensional heteroscedastic regression with an application to eQTL data analysis*, Biometrics **68** (2012), no. 1, 316–326.
- O. Klopp, *High dimensional matrix estimation with unknown variance of the noise*, arXiv preprint arXiv:1112.3055 (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.
- N. Simon and R. Tibshirani, *Standardization and the Group Lasso penalty*, Stat. Sin. **22** (2012), no. 3, 983–1001 (English).
- 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. R. Stat. Soc. Ser. B Stat. Methodol. **58** (1996), no. 1, 267–288.

SOCP reformulation

$$
\begin{aligned}
\min \quad & \sum_{k=1}^{K} \lambda_k u_k \\
\text{subject to} \quad & \\
\forall k = 1, \cdots, K \quad & \left| \boldsymbol{X}_{:,G_k} \boldsymbol{\phi}_{G_k} \right|_2 \leq u_k, \\
\forall k = 1, \cdots, K, \quad & \left| \boldsymbol{\Pi}_{G_k} \big(\operatorname{diag}(\boldsymbol{Y}) \boldsymbol{R} \boldsymbol{\alpha} - \boldsymbol{X} \boldsymbol{\phi} \big) \right|_2 \leq \lambda_k, \\
& \boldsymbol{R}^\top \boldsymbol{v} \leq \boldsymbol{R}^\top \operatorname{diag}(\boldsymbol{Y}) (\operatorname{diag}(\boldsymbol{Y}) \boldsymbol{R} \boldsymbol{\alpha} - \boldsymbol{X} \boldsymbol{\phi}); \\
\forall t = 1, \cdots, T, \quad & \left| \left[v_t; \boldsymbol{R}_{t,:} \boldsymbol{\alpha}; \sqrt{2} \right] \right|_2 \leq v_t + \boldsymbol{R}_{t,:} \boldsymbol{\alpha};
\end{aligned}
$$

Assumption

Some notations:

$$\mathcal{K}^* = \left\{k : |\boldsymbol{\phi}^*_{G_k}|_1 \neq 0\right\},$$

$$J_{\phi^*} = \bigcup_{k\in\mathcal{K}^*} G_k, \qquad i_{\phi^*} = \sum_{k\in\mathcal{K}^*} |G_k|,$$

$$\Gamma(\mathcal{K}) = \left\{\boldsymbol{\delta} \in \mathbb{R}^p : \sum_{k\in\mathcal{K}^c} \lambda_k |\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}|_2 \leq \sum_{k\in\mathcal{K}} \lambda_k |\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}|_2\right\}.$$

Let $1 \leq b \leq K$ be a bound on the group sparsity: $|J_{\phi^*}| \leq b$

Group Restricted Eigenvalue Condition (GREC)

$$\exists\kappa, \forall\boldsymbol{\delta} \in \Gamma(\mathcal{K}) \setminus \{0\}, \text{s.t.} |\mathcal{K}| \leq \mathcal{K}^*, |\mathbf{X}\boldsymbol{\delta}|_2^2 \geq \kappa^2 T \sum_{k\in\mathcal{K}} |\mathbf{X}_{:,G_k}\boldsymbol{\delta}_{G_k}|_2^2$$

REM: extension of the RE Bickel *et al.* (2009)

Assumption signal/noise ratio

Define

$$C_1 = \min_{\ell=1,\dots,q} \frac{1}{T} \sum_{t\in\mathcal{T}} \frac{r_{t\ell}^2(\boldsymbol{X}_{t,:}\boldsymbol{\phi}^*)^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2} \ ,$$

$$C_2 = \max_{\ell=1,\dots,q} \frac{1}{T} \sum_{t\in\mathcal{T}} \frac{r_{t\ell}^2}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)^2} \ ,$$

$$C_3 = \min_{\ell=1,\dots,q} \frac{1}{T} \sum_{t\in\mathcal{T}} \frac{r_{t\ell}}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)}.$$

We denote $C_4 = (\sqrt{C_2} + \sqrt{2\,C_1})/C_3$ and

$$\max_{t=1,\cdots,T} \frac{(\boldsymbol{R}_{t,:}\hat{\boldsymbol{\alpha}})}{(\boldsymbol{R}_{t,:}\boldsymbol{\alpha}^*)} \leq \hat{D}_1$$

The constant in the oracle inequalities satisfies:

$$D_{T,\delta} = C_4\hat{D}_1(|\boldsymbol{X}\boldsymbol{\phi}^*|_\infty^2 + \log(\frac{T}{\delta}))$$