Heteroscedastic Regression 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)

Heteroscedastic regression

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

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

- 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 i.i.d such that $\mathbf{E}[\xi_t|\boldsymbol{x}_t] = 0$ and $\mathbf{Var}[\xi_t|\boldsymbol{x}_t] = 1$ (*e.g.* Gaussian for simplicity)

Sparsity **Assumption**

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

Example: Homoscedastic regression

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

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

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

$$|\boldsymbol{\beta}^*|_0 := \sum_{j=1}^p \mathbb{1}(\boldsymbol{\beta}_j^* \neq 0) \ll T$$

Homoscedastic case with known noise level

Regression 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
- $\boldsymbol{\beta}^*$ is to be recovered by $\hat{\beta}$

Pioneer methods: homoscedastic, σ^* known

LASSO Tibshirani (1996)

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

Dantzig-Selector Candès and Tao (2007)

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

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

Pioneering methods: homoscedastic, σ^*

Type of results proved Bickel *et al.* (2009)

$\boldsymbol{X}$ is s.t. $|\boldsymbol{X}_{:,j}|_2 = 1$ for all $j = 1, \ldots, p$; and $\boldsymbol{X}$ satisfies the RE condition (RIP type condition), for $\lambda = A\sigma^* \sqrt{\frac{\log p}{n}}$ $(A > 2\sqrt{2})$ then with probability $1 - p^{1-A^2/8}$

$$\|\boldsymbol{X}\hat{\beta} - \boldsymbol{X}\beta^*\|_2^2 \precsim \frac{(\sigma^*)^2}{\kappa^2}\|\beta^*\|_0 \log p$$

and

$$\|\hat{\beta} - \beta^*\|_1 \precsim \frac{\sigma^*}{\kappa^2}\|\beta^*\|_0 \sqrt{\frac{\log p}{n}}$$

Homoscedastic case with unknown noise level

Matrix/vector 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: $\mathsf{s}^*(\boldsymbol{x}_t) \equiv \sigma^* \in \mathbb{R}_*^+$

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)

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

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

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

- **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

Pioneering methods: homoscedastic, σ^* unknown

$\sqrt{\text{Lasso}}$ Antoniadis (2010) , Belloni *et al.* (2011) , Sun and Zhang (2012)

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

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

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

Objectives

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

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

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

$\hookrightarrow$ **without exact knowledge** of the noise level

Problem re-formulation

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

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

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

- 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
  - $\hookrightarrow$ Project on elementary functions (Fourier, Wavelet):

$$\mathsf{f}_j^*(x) \approx \sum_{\ell=1}^{K_j} \phi_{\ell,j}\psi_\ell(x)$$

then $\boldsymbol{\phi} = (\phi_{\ell,j})$ is group sparse

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 $\mathsf{r}_1, \ldots, \mathsf{r}_q$.

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

$\boldsymbol{R} = (\mathsf{r}_\ell(\boldsymbol{x}_t))_{t,\ell}$ is a $T \times q$ noise design matrix

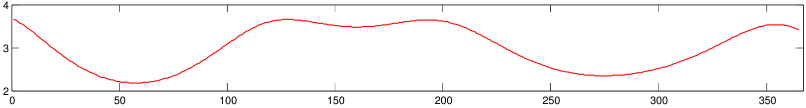
<u>REM</u>: here and after $q \ll T$

$$\boxed{\text{Reformulated model:} \quad \text{diag}(\boldsymbol{Y})\boldsymbol{R}\boldsymbol{\alpha}^* = \mathbf{X}\boldsymbol{\phi}^* + \boldsymbol{\xi}}$$

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)
- Periodic/seasonal noise-level
  $\hookrightarrow$ r^* belongs to the linear span of a few trigonometric functions: *e.g.*, meteorology (seasonal variations)



Penalized $\log$-likelihood formulation

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

- Remind $\boldsymbol{R} = (\mathsf{r}_\ell(\boldsymbol{x}_t))_{t,\ell}$ is the $T \times q$ noise design matrix
- Tuning parameter: $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}_+^K$
- Use $\sum_{k=1}^{K} \lambda_k \left|\boldsymbol{X}_{:,G_k}\boldsymbol{\phi}_{G_k}\right|_2$ instead of $\sum_{k=1}^{K} \lambda_k \left|\boldsymbol{\phi}_{G_k}\right|_2$ as in Simon and Tibshirani (2012) : **equivariance** w.r.t. invertible linear transformations of predictors within groups
- log-det problem not a **L**inear **P**rogramming (LP) or an SOCP

Relaxation of first order conditions (1)

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

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

$\hookrightarrow$ Difficult problem: non-linear part

- Equivalence with

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

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

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

Relaxation of 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.

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

Relaxation

Scaled Heteroscedastic Dantzig selector (ScHeDs)

Definition:

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

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

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

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.

Comments on the procedure

- Degrees of freedom:
  - ↪ Many tuning parameters in the procedure
  - ↪ Theory: $\lambda_k = \lambda_0 \sqrt{r_k}$ with $\lambda_0 > 0$ and $r_k = \text{rank}(\mathbf{X}_{:,G_k})$
  - ↪ Most papers use $\lambda_k \propto \sqrt{|G_k|}\,(k = 1, \ldots, K)$
- Bias correction, practical improvement:
  - ↪ 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 ($\boldsymbol{\lambda} = 0$)
- Implementation with SOCP solvers (Matlab):
  Sedumi Sturm (1999) : popular interior point method, highly accurate solution for small datasets, *e.g.* $p, T \leq 2000$
  Tfocs Becker *et al.* (2011) : first-order proximal method, less accurate BUT can handle larger dimensions, *e.g.* $p = 5000$ and $T = 3000$

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: $y_t = \mathbf{X}_{t,:}\boldsymbol{\beta}^* + \sigma_t \boldsymbol{\xi}_t$.

Compared with: 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	$\sigma = 4,\ p = 200$				
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**

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

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

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

with $D_{T,\delta} = \log(\frac{T}{\delta})$, $\mathcal{K}^* = \{k : |\boldsymbol{\phi}^*_{G_k}| \neq 0\}$, $i^* = \sum_{k \in \mathcal{K}^*} \mathrm{rank}(\mathrm{X}_{:,\mathrm{G_k}})$

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 heteroscedastic regression models
- Estimating both the mean and the variance functions
- Takes into account group sparsity
- Relaxation of 1st order conditions for penalized MLE
  - $\hookrightarrow$ existence of a solution
  - $\hookrightarrow$ convex problem – second-order cone programming
- Competitive with state-of-the-art algorithms
- More simulations + real data in the paper

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.
- 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

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

Assumption

Some notations:

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

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

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

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:

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