

**Learning Heteroscedastic Models  
by  
Conic Programming  
under  
Group Sparsity**

**Joseph Salmon**

**IMT**

**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  $(\mathbf{x}_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$  obeying

$$y_t = \mathbf{b}^*(\mathbf{x}_t) + \mathbf{s}^*(\mathbf{x}_t)\xi_t, \quad t = 1, \dots, T$$

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

## *Sparsity Assumption*

- ▶ Estimating  $\mathbf{b}^*$  and  $\mathbf{s}^*$  is ill-posed
- ▶ sparsity senario:  $\mathbf{b}^*$  and  $\mathbf{s}^*$  belong to low dimensional spaces

### Example: Homoscedastic regression

$$\forall \mathbf{x}, \quad \mathbf{b}^*(\mathbf{x}) = [f_1(\mathbf{x}), \dots, f_p(\mathbf{x})]\boldsymbol{\beta}^*, \quad \text{and} \quad \mathbf{s}^*(\mathbf{x}) \equiv \sigma^*$$

↔ Dictionary  $\{f_1, \dots, f_p\}$  of functions from  $\mathbb{R}^d$  to  $\mathbb{R}$

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

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

# Homoscedastic case with known noise level

## Regression formulation

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

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

### REM:

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

## Pioneer methods: homoscedastic, $\sigma^*$ known

### LASSO Tibshirani (1996)

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

### Dantzig-Selector Candès and Tao (2007)

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

Non-asymptotic guarantees available for both (e.g. [Bickel et al. \(2009\)](#)) for a tuning parameter satisfying  $\lambda \propto \sigma^*$ , **but** knowledge of  $\sigma^*$  needed!

## Non-asymptotic guarantees

### LASSO / Dantzig-Selector *Bickel et al. (2009)*

Under a normalizing hypothesis  $\|\mathbf{X}_{:,j}\|_2 = 1$ , and RE condition on  $\mathbf{X}$  (RIP type condition) then for any  $A > 2\sqrt{2}$ , and for  $\lambda = A\sigma^* \sqrt{\frac{\log p}{T}}$  then with probability at least  $1 - p^{1-A^2/8}$  one has

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \lesssim \frac{\lambda \|\boldsymbol{\beta}^*\|_0}{\kappa^2}$$

and  $\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \lesssim \frac{\lambda^2 \|\boldsymbol{\beta}^*\|_0 \sqrt{T}}{\kappa^2}$

where  $\kappa$  can be interpreted as a generalization of the conditioning number of  $\mathbf{X}$

# Homoscedastic case with unknown noise level

## Matrix/vector formulation

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

Observations:  $\mathbf{Y} = [y_1, \dots, y_T]^\top \in \mathbb{R}^T$

Noise:  $\boldsymbol{\xi} = [\xi_1, \dots, \xi_T]^\top \in \mathbb{R}^T$

Design Matrix:  $\mathbf{X}_{t,j} = [f_j(\mathbf{x}_t)] \in \mathbb{R}$

Coefficients:  $\boldsymbol{\beta}^* = [\beta_1^*, \dots, \beta_p^*]^\top \in \mathbb{R}^p$

Standard deviation:  $s^*(\mathbf{x}_t) \equiv \sigma^* \in \mathbb{R}_*^+$

### REM:

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

# Pioneering methods: homoscedastic, $\sigma^*$ unknown

Scaled-Lasso, Städler *et al.* (2010)

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

↔ penalized (Gaussian, negative) log-likelihood minimization

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

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

- ▶ **equivariant** estimator, i.e. if  $\mathbf{Y} \leftarrow c\mathbf{Y}$ ,  $\beta^* \leftarrow c\beta^*$ ,  $\sigma^* \leftarrow c\sigma^*$ , then  $\hat{\beta} \leftarrow c\hat{\beta}$  and  $\hat{\sigma} \leftarrow c\hat{\sigma}$
- ▶ **Jointly** convex problem



## Pioneering methods: homoscedastic, $\sigma^*$ unknown

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

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

- ▶ Can be solved by a **Second Order Cone Program (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  $s^*$

↪ for the **heteroscedastic** regression

↪ **without exact knowledge** of the noise level

## Problem re-formulation

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

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

# Assumptions on the model (I)

## Group Sparsity Assumption

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

$$[f^*(\mathbf{x}_1), \dots, f^*(\mathbf{x}_T)]^\top = \mathbf{X}\phi^*, \quad \text{Card}(\{k : |\phi_{G_k}^*|_2 \neq 0\}) \ll K.$$

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
  - ↪ linear regression with qualitative covariates
  - ↪ each covariate has several modalities
- ▶ Sparse additive model
  - ↪  $\mathbf{f}^*(\mathbf{x}) = \mathbf{f}_1^*(x_1) + \dots + \mathbf{f}_d^*(x_d)$  ;  $\mathbf{f}_j^* \equiv 0$  for most  $j$
  - ↪ Project on elementary functions (Fourier, Wavelet):

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

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

## Assumptions on the model (II)

### Low dimension volatility assumption

For  $q$  given functions  $r_1, \dots, r_q$  mapping  $\mathbb{R}^d$  into  $\mathbb{R}_+$ , there is a vector  $\alpha^* \in \mathbb{R}^q$  such that  $r^*(\mathbf{x}) = \sum_{\ell=1}^q \alpha_\ell^* r_\ell(\mathbf{x})$  for almost every  $\mathbf{x} \in \mathbb{R}^d$ , and  $\mathcal{S}$  is the linear span of  $r_1, \dots, r_q$ .

$$[r^*(\mathbf{x}_1), \dots, r^*(\mathbf{x}_T)]^\top = \mathbf{R}\alpha^*$$

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

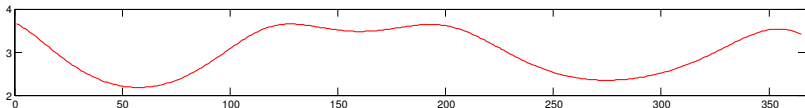
REM: here and after  $q \ll T$

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

## Examples of application (II)

### Low dimension volatility assumption

- ▶ Block-wise homoscedastic noise
  - ↪  $r^*$  is well approximated by a piecewise constant function: time series modeling (smooth variations over time)
- ▶ Periodic/seasonal noise-level
  - ↪  $r^*$  belongs to the linear span of a few trigonometric functions: e.g., meteorology (seasonal variations)



# Penalized log-likelihood formulation

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

- ▶ Remind  $\mathbf{R} = (r_\ell(\mathbf{x}_t))_{t,\ell}$  is the  $T \times q$  noise design matrix
- ▶ Tuning parameter:  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_K) \in \mathbb{R}_+^K$
- ▶ Use  $\sum_{k=1}^K \lambda_k |\mathbf{X}_{:,G_k} \phi_{G_k}|_2$  instead of  $\sum_{k=1}^K \lambda_k |\phi_{G_k}|_2$  as in [Simon and Tibshirani \(2012\)](#) : **equivariance** w.r.t. invertible linear transformations of predictors within groups
- ▶ log-det problem not a **Linear Programming** (LP) or an SOCP

# Relaxation of first order conditions (1)

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

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

↪ Difficult problem: non-linear part

- Equivalence with

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

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

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



## Relaxation of first order conditions (2)

►  $\forall \ell = 1, \dots, q$ ,  $\frac{\partial}{\partial \alpha_\ell} \text{PL}(\phi, \alpha) = 0$  implies:

$\exists \nu \in \mathbb{R}_+^T$  such that

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

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

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

# Relaxation

## Scaled Heteroscedastic Dantzig selector (ScHeDs)

Definition:

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

$$\left| \boldsymbol{\Pi}_{G_k} (\text{diag}(\mathbf{Y}) \mathbf{R} \boldsymbol{\alpha} - \mathbf{X} \boldsymbol{\phi}) \right|_2 \leq \lambda_k, \quad \forall k \in \{1, \dots, K\}$$

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

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  $\lambda_k \propto \sqrt{|G_k|}$  ( $k = 1, \dots, K$ )

► Bias correction, practical improvement:

↪ Classical two-steps methods:

i) our algorithm with  $\lambda_k = \lambda_0 \sqrt{r_k}$  ( $k=1, \dots, K$ )

ii) Least squares on the selected variables ( $\lambda = 0$ )

► 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  $\sigma_t \equiv \sigma^*$ ;
- ▶  $\boldsymbol{\beta}^* = (2, 3, 3, 3, 1.5, 1.5, 1.5, 0, 0, 0, 2, 2, 2, 0, \dots, 0)^\top \in \mathbb{R}^p$
- ▶ 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”  $\lambda = \sqrt{2 \log(p)}$ ;

$\mathbf{R}$ : 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  $\|(\mathbf{X}\hat{\phi}) ./ (\mathbf{R}\hat{\alpha}) - \mathbf{X}\beta^*\|_2 / \sqrt{T}$ )

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

# 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

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

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

## 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
  - ↪ existence of a solution
  - ↪ 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:  $\ell_1$ -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,  *$\ell_1$ -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

$$\min \sum_{k=1}^K \lambda_k u_k$$

subject to

$$\forall k = 1, \dots, K \quad |\mathbf{X}_{:,G_k} \boldsymbol{\phi}_{G_k}|_2 \leq u_k,$$

$$\forall k = 1, \dots, K, \quad \left| \boldsymbol{\Pi}_{G_k} (\text{diag}(\mathbf{Y}) \mathbf{R} \boldsymbol{\alpha} - \mathbf{X} \boldsymbol{\phi}) \right|_2 \leq \lambda_k,$$

$$\mathbf{R}^\top \mathbf{v} \leq \mathbf{R}^\top \text{diag}(\mathbf{Y}) (\text{diag}(\mathbf{Y}) \mathbf{R} \boldsymbol{\alpha} - \mathbf{X} \boldsymbol{\phi});$$

$$\forall t = 1, \dots, T, \quad |[v_t; \mathbf{R}_{t,:} \boldsymbol{\alpha}; \sqrt{2}]|_2 \leq v_t + \mathbf{R}_{t,:} \boldsymbol{\alpha};$$

# Assumption

Some notations:

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

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

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

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

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

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

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

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

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

The constant in the oracle inequalities satisfies:

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