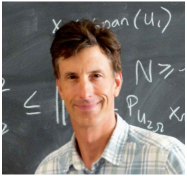
Linear Regression with Strongly Correlated Designs Using Ordered Weigthed ℓ_1 (OWL ) Regularization

Mário A. T. Figueiredo

Instituto de **T**elecomunicações
and
Instituto **S**uperior **T**écnico, Universidade de Lisboa
Portugal

Joint work with **Robert Nowak** (**U Wisconsin, USA**)



Introduction

Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}$

Introduction

Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}$

- Design matrix: $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p] \in \mathbb{R}^{n \times p}$;

Introduction

Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}$

- Design matrix: $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p] \in \mathbb{R}^{n\times p}$;
- Regression coefficients: $\mathbf{x} \in \mathbb{R}^p$;

Introduction

Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}$

- Design matrix: $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p] \in \mathbb{R}^{n\times p}$;
- Regression coefficients: $\mathbf{x} \in \mathbb{R}^p$;
- Noise (or random perturbations): $\mathbf{n} \in \mathbb{R}^n$;

Introduction

Linear regression: classical problem in statistics, machine learning, signal processing, with countless applications.

Observations: $\mathbf{y} = \mathbf{A}\,\mathbf{x} + \mathbf{n}$

- Design matrix: $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_p] \in \mathbb{R}^{n\times p}$;
- Regression coefficients: $\mathbf{x} \in \mathbb{R}^p$;
- Noise (or random perturbations): $\mathbf{n} \in \mathbb{R}^n$;
- **Goal**: estimate $\mathbf{x}$, from $\mathbf{y}$ and $\mathbf{A}$.

Regularization, Sparsity, and Variable Selection

Regularized linear regression (classical criteria):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\,\mathbf{x} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{x})$$

Regularization, Sparsity, and Variable Selection

Regularized linear regression (classical criteria):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\,\mathbf{x} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{x})$$

- $R(\mathbf{x}) = \|\mathbf{x}\|_2^2 \quad \Rightarrow \quad \widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{y}$;
  ridge regression, (Hoerl and Kennard, 1970)

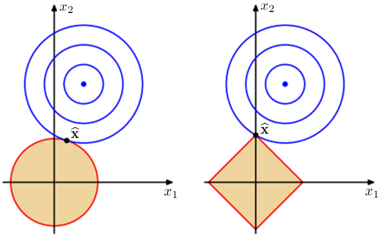
Regularization, Sparsity, and Variable Selection

Regularized linear regression (classical criteria):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\,\mathbf{x} - \mathbf{y}\|_2^2 + \lambda R(\mathbf{x})$$

- $R(\mathbf{x}) = \|\mathbf{x}\|_2^2 \quad \Rightarrow \quad \widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A} + \lambda \mathbf{I}\right)^{-1}\mathbf{A}^T\mathbf{y}$;
  ridge regression, (Hoerl and Kennard, 1970)

- $R(\mathbf{x}) = \|\mathbf{x}\|_1$;
  LASSO (Tibshirani, 1996), basis pursuit denoising (Chen et al., 1995)



Sparsity! (variable selection)

Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

Group sparsity: discard/keep entire *groups* of features (Bach et al., 2012)

- density inside each group
- sparsity with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

Group sparsity: discard/keep entire *groups* of features (Bach et al., 2012)

- density inside each group
- sparsity with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

Yields statistical gains if the assumption is correct (Huang and Zhang, 2010; Stojnic et al., 2009)

Group/Structured Sparsity

Promote certain sparsity patterns (usually groups)

Group sparsity: discard/keep entire *groups* of features (Bach et al., 2012)

- density inside each group
- sparsity with respect to the groups which are selected
- choice of groups: prior knowledge about the intended *sparsity patterns*

Yields statistical gains if the assumption is correct (Huang and Zhang, 2010; Stojnic et al., 2009)

Many applications:

- feature template selection (Martins et al., 2011)
- multi-task learning (Caruana, 1997; Obozinski et al., 2010)
- multiple kernel learning (Bach, 2008)
- learning the structure of graphical models (Schmidt and Murphy, 2010)

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task
- Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task
- Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof
- Group regularizers may solve this problem, but require *a priori* knowledge of group structure

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task
- Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof
- Group regularizers may solve this problem, but require *a priori* knowledge of group structure
- Alternatives (without predefined groups):
  - ◇ Elastic net (EN)
    (Zou and Hastie, 2005; De Mol et al., 2009)

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task
- Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof
- Group regularizers may solve this problem, but require *a priori* knowledge of group structure
- Alternatives (without predefined groups):
  - ◇ Elastic net (EN)
    (Zou and Hastie, 2005; De Mol et al., 2009)
  - ◇ Cluster lasso
    (Bühlmann et al., 2013)

Variable Selection and Grouping

- Goal: identify **all** the covariates (*e.g.*, genes, voxels,...) that are relevant in some problem/task
- Problem: with highly correlated covariates, LASSO may select an arbitrary subset thereof
- Group regularizers may solve this problem, but require *a priori* knowledge of group structure
- Alternatives (without predefined groups):
  - ◇ Elastic net (EN)
    (Zou and Hastie, 2005; De Mol et al., 2009)
  - ◇ Cluster lasso
    (Bühlmann et al., 2013)
  - ◇ Octagonal shrinkage and clustering algorithm for regression (OSCAR)
    (Bondell and Reich, 2007; Zhong and Kwok, 2012)

Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- Elastic net:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2$

Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- Elastic net:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2$

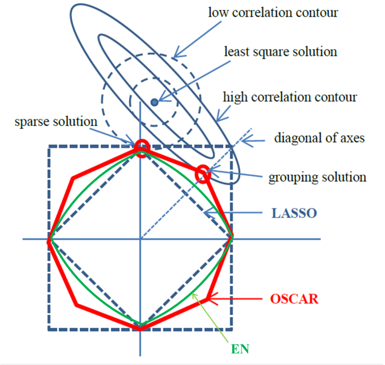
- OSCAR:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$

Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- Elastic net:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2$
- OSCAR:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$

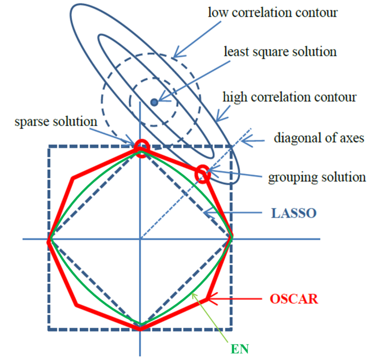


Elastic Net (EN) and OSCAR

Goal of **EN**: including groups of correlated variables.

Goal of **OSCAR**: grouping correlated variables.

- Elastic net:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \|\mathbf{x}\|_2^2$
- OSCAR:
  $R(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$



OSCAR is competitive with EN, LASSO, ridge, in terms of MSE;

OSCAR yields explicit variable grouping (Bondell and Reich, 2007)

Some OSCAR Results on Synthetic Data

Example		Med. MSE (Std. Err.)	MSE 10th perc.	MSE 90th perc.	Med. Df
1	Ridge	2.31 (0.18)	0.98	4.25	8
	Lasso	1.92 (0.16)	0.68	4.02	5
	Elastic Net	1.64 (0.13)	0.49	3.26	5
	Oscar	1.68 (0.13)	0.52	3.34	4
2	Ridge	2.94 (0.18)	1.36	4.63	8
	Lasso	2.72 (0.24)	0.98	5.50	5
	Elastic Net	2.59 (0.21)	0.95	5.45	6
	Oscar	2.51 (0.22)	0.96	5.06	5
3	Ridge	1.48 (0.17)	0.56	3.39	8
	Lasso	2.94 (0.21)	1.39	5.34	6
	Elastic Net	2.24 (0.17)	1.02	4.05	7
	Oscar	1.44 (0.19)	0.51	3.61	5
4	Ridge	27.4 (1.17)	21.2	36.3	40
	Lasso	45.4 (1.52)	32.0	56.4	21
	Elastic Net	34.4 (1.72)	24.0	45.3	25
	Oscar	25.9 (1.26)	19.1	38.1	15
5	Ridge	70.2 (3.05)	41.8	103.6	40
	Lasso	64.7 (3.03)	27.6	116.5	12
	Elastic Net	40.7 (3.40)	17.3	94.2	17
	Oscar	51.8 (2.92)	14.8	96.3	12

From (Bondell and Reich, 2007)

Generalizing OSCAR: The OWL

OSCAR: $$R_{\mathsf{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$$

Generalizing OSCAR: The OWL

OSCAR:
$$R_{\mathsf{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) = \lambda_1\|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$$
$$= \sum_{i=1}^{p} \big(\lambda_1 + \lambda_2(p-i)\big)\, |x|_{[i]},$$

where $|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]}$ (sorted entries of $|\mathbf{x}|$).

Generalizing OSCAR: The OWL

OSCAR: $$R_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\mathbf{x}) = \lambda_1 \|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$$
$$= \sum_{i=1}^{p} \big(\lambda_1 + \lambda_2(p-i)\big)\, |x|_{[i]},$$

where $|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]}$ (sorted entries of $|\mathbf{x}|$).

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]}$$

where $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$

Generalizing OSCAR: The OWL

OSCAR: $$R^{\lambda_1,\lambda_2}_{\text{OSCAR}}(\mathbf{x}) = \lambda_1\|\mathbf{x}\|_1 + \lambda_2 \sum_{i<j} \max\{|x_i|, |x_j|\}$$
$$= \sum_{i=1}^{p} \left(\lambda_1 + \lambda_2(p-i)\right) |x|_{[i]},$$

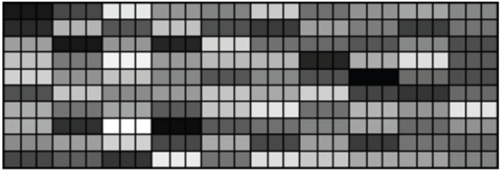
where $|x|_{[1]} \geq |x|_{[2]} \geq \cdots \geq |x|_{[p]}$ (sorted entries of $|\mathbf{x}|$).

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i\, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

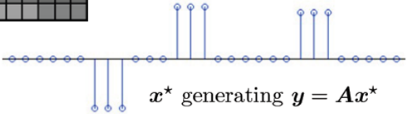
where $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $|\mathbf{x}|_{\downarrow} = \left[|x|_{[1]}, |x|_{[2]}, \ldots, |x|_{[p]}\right]^T$

Toy example

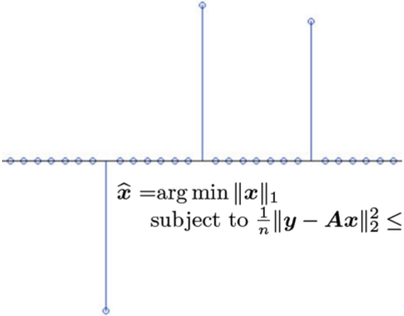


$\boldsymbol{A} \in \mathbb{R}^{10\times 30}$

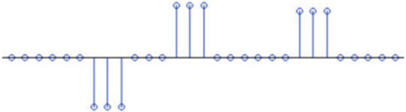
every column has 3 replicates



$\boldsymbol{x}^\star$ generating $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}^\star$



$$\widehat{\boldsymbol{x}} = \arg\min \|\boldsymbol{x}\|_1 \quad \text{subject to } \tfrac{1}{n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \varepsilon$$



$$\widehat{\boldsymbol{x}} = \arg\min \Omega_{\boldsymbol{w}}(\boldsymbol{x}) \quad \text{subject to } \tfrac{1}{n}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \varepsilon$$

The OWL Norm

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

The OWL Norm

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- Proposed independently by:
  - ◇ Bogdan et al. (2013), for false discovery rate (FDR) control in variable selection with weakly correlated covariates
  - ◇ Zeng and Figueiredo (2014), generalizing OSCAR, for variable grouping with strongly correlated covariates

The OWL Norm

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- Proposed independently by:
  - ◇ Bogdan et al. (2013), for false discovery rate (FDR) control in variable selection with weakly correlated covariates
  - ◇ Zeng and Figueiredo (2014), generalizing OSCAR, for variable grouping with strongly correlated covariates
- Remaining of the talk focuses on the **OWL**
  - ◇ Part I: covariate clustering analysis
  - ◇ Part II: statistical analysis

Some Properties of the OWL

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

Some Properties of the OWL

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.

Some Properties of the OWL

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.
- Relationship with ℓ_1

$$\bar{w} \, \|\mathbf{x}\|_1 \leq \Omega_{\mathbf{w}}(\mathbf{x}) \leq w_1 \, \|\mathbf{x}\|_1;$$

where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$.

Some Properties of the OWL

The **ordered weighted ℓ_1 (OWL)** norm

$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.
- Relationship with ℓ_1

$$\bar{w} \, \|\mathbf{x}\|_1 \leq \Omega_{\mathbf{w}}(\mathbf{x}) \leq w_1 \, \|\mathbf{x}\|_1;$$

  where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$.
- Obviously, $\Omega_{\mathbf{w}}(\mathbf{x}) \geq w_1 \, \|\mathbf{x}\|_\infty$ (equality if $w_2 = w_3 = \cdots = w_p = 0$).

Some Properties of the OWL

The **ordered weighted ℓ_1 (OWL)** norm

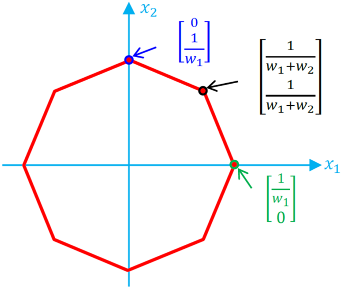
$$\Omega_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{p} w_i \, |x|_{[i]} = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$$

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \to \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.
- Relationship with ℓ_1

$$\bar{w} \, \|\mathbf{x}\|_1 \leq \Omega_{\mathbf{w}}(\mathbf{x}) \leq w_1 \, \|\mathbf{x}\|_1;$$

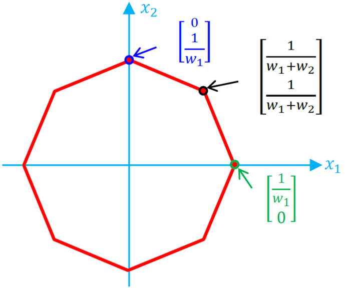
  where $\bar{w} = \frac{1}{p} \sum_{i=1}^{p} w_i$, with equalities if $w_1 = w_2 = \cdots = w_p$.
- Obviously, $\Omega_{\mathbf{w}}(\mathbf{x}) \geq w_1 \, \|\mathbf{x}\|_\infty$ (equality if $w_2 = w_3 = \cdots = w_p = 0$).
- Proximity operator ($O(p \log p)$), projection onto an OWL-ball ($O(p \log p)$), atomic formulation are all known (yesterday's poster).

Atoms

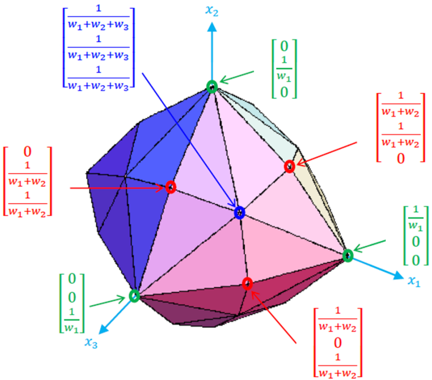


x_2
x_1
$\begin{bmatrix} 0 \\ \frac{1}{w_1} \end{bmatrix}$
$\begin{bmatrix} \frac{1}{w_1+w_2} \\ \frac{1}{w_1+w_2} \end{bmatrix}$
$\begin{bmatrix} \frac{1}{w_1} \\ 0 \end{bmatrix}$

Atoms



x_2
$\begin{bmatrix} 0 \\ \frac{1}{w_1} \end{bmatrix}$
$\begin{bmatrix} \frac{1}{w_1+w_2} \\ \frac{1}{w_1+w_2} \end{bmatrix}$
x_1
$\begin{bmatrix} \frac{1}{w_1} \\ 0 \end{bmatrix}$



$\begin{bmatrix} \frac{1}{w_1+w_2+w_3} \\ \frac{1}{w_1+w_2+w_3} \\ \frac{1}{w_1+w_2+w_3} \end{bmatrix}$
x_2
$\begin{bmatrix} 0 \\ \frac{1}{w_1} \\ 0 \end{bmatrix}$
$\begin{bmatrix} \frac{1}{w_1+w_2} \\ \frac{1}{w_1+w_2} \\ 0 \end{bmatrix}$
$\begin{bmatrix} 0 \\ \frac{1}{w_1+w_2} \\ \frac{1}{w_1+w_2} \end{bmatrix}$
$\begin{bmatrix} \frac{1}{w_1} \\ 0 \\ 0 \end{bmatrix}$
x_1
$\begin{bmatrix} 0 \\ 0 \\ \frac{1}{w_1} \end{bmatrix}$
x_3
$\begin{bmatrix} \frac{1}{w_1+w_2} \\ 0 \\ \frac{1}{w_1+w_2} \end{bmatrix}$

Part I: Clustering Analysis

Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. It is said that $\mathbf{x}$ majorizes $\mathbf{y}$, denoted $\mathbf{x} \succ \mathbf{y}$, if

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x_{[i]} \geq \sum_{i=1}^{j} y_{[i]}, \text{ for } j = 1, ..., p-1. \tag{1}$$

Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. It is said that $\mathbf{x}$ majorizes $\mathbf{y}$, denoted $\mathbf{x} \succ \mathbf{y}$, if

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x_{[i]} \geq \sum_{i=1}^{j} y_{[i]}, \;\; \text{for } j = 1, ..., p-1. \tag{1}$$

Examples: $(4,0,0,0) \succ (3,1,0,0) \succ (2,1,1,0) \succ (1,1,1,1)$

Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. It is said that $\mathbf{x}$ majorizes $\mathbf{y}$, denoted $\mathbf{x} \succ \mathbf{y}$, if

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x_{[i]} \geq \sum_{i=1}^{j} y_{[i]}, \;\; \text{for } j = 1, ..., p-1. \tag{1}$$

Examples: $(4,0,0,0) \succ (3,1,0,0) \succ (2,1,1,0) \succ (1,1,1,1)$

Definition (Schur-convexity (Marshall et al., 2011))

Let $\mathcal{A} \subseteq \mathbb{R}^P$; a function $f : \mathcal{A} \to \mathbb{R}$ is Schur-convex in $\mathcal{A}$ if,

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{A}, \;\; \mathbf{x} \succ \mathbf{y} \;\; \Rightarrow \;\; f(\mathbf{x}) \geq f(\mathbf{y}),$$

Majorization and Schur Convexity

Definition (Majorization (Marshall et al., 2011))

Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. It is said that $\mathbf{x}$ majorizes $\mathbf{y}$, denoted $\mathbf{x} \succ \mathbf{y}$, if

$$\sum_{i=1}^{p} x_i = \sum_{i=1}^{p} y_i \quad \text{and} \quad \sum_{i=1}^{j} x_{[i]} \geq \sum_{i=1}^{j} y_{[i]}, \quad \text{for } j = 1, ..., p-1. \qquad (1)$$

Examples: $(4,0,0,0) \succ (3,1,0,0) \succ (2,1,1,0) \succ (1,1,1,1)$

Definition (Schur-convexity (Marshall et al., 2011))

Let $\mathcal{A} \subseteq \mathbb{R}^P$; a function $f : \mathcal{A} \to \mathbb{R}$ is Schur-convex in $\mathcal{A}$ if,

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{A}, \;\; \mathbf{x} \succ \mathbf{y} \;\; \Rightarrow \;\; f(\mathbf{x}) \geq f(\mathbf{y}),$$

and strictly Schur-convex, if the second inequality is strict when $\mathbf{x}$ is not a permutation of $\mathbf{y}$.

Strong Schur Convexity

Definition (Pigou-Dalton transfer (Marshall et al., 2011))

Consider $\mathbf{x} \in \mathbb{R}_+^p$ and two components, x_i, x_j, s.t. $x_i > x_j$. We say that $\mathbf{y}$ ($\mathbf{y} \prec \mathbf{x}$) results from a Pigou-Dalton transfer of size $\varepsilon \in \big(0,\, (x_i - x_j)/2\big)$ if

$$y_i = x_i - \varepsilon, \;\; y_j = x_j + \varepsilon, \;\; y_k = x_k, \;\; \text{for } k \neq i, j.$$

Strong Schur Convexity

Definition (Pigou-Dalton transfer (Marshall et al., 2011))

Consider $\mathbf{x} \in \mathbb{R}_+^p$ and two components, x_i, x_j, s.t. $x_i > x_j$. We say that $\mathbf{y}$ ($\mathbf{y} \prec \mathbf{x}$) results from a Pigou-Dalton transfer of size $\varepsilon \in \big(0,\, (x_i - x_j)/2\big)$ if

$$y_i = x_i - \varepsilon, \;\; y_j = x_j + \varepsilon, \;\; y_k = x_k, \;\; \text{for } k \neq i, j.$$

The Pigou-Dalton transfer (a.k.a. Robin-Hood transfer) is used in the study of measures of economic inequality (Dalton, 1920; Pigou, 1912).

Strong Schur Convexity

Definition (Pigou-Dalton transfer (Marshall et al., 2011))

Consider $\mathbf{x} \in \mathbb{R}_+^p$ and two components, x_i, x_j, s.t. $x_i > x_j$. We say that $\mathbf{y}$ ($\mathbf{y} \prec \mathbf{x}$) results from a Pigou-Dalton transfer of size $\varepsilon \in \big(0, (x_i - x_j)/2\big)$ if

$$y_i = x_i - \varepsilon, \;\; y_j = x_j + \varepsilon, \;\; y_k = x_k, \;\; \text{for } k \neq i, j.$$

The Pigou-Dalton transfer (a.k.a. Robin-Hood transfer) is used in the study of measures of economic inequality (Dalton, 1920; Pigou, 1912).

Definition (Strong Schur convexity (Figueiredo and Nowak, 2014))

Function f is S-strongly Schur-convex if there exists a constant $S > 0$, s.t.

$$f(\mathbf{x}) - f(\mathbf{y}) \geq \varepsilon S,$$

whenever $\mathbf{y} \prec \mathbf{x}$ result from a Pigou-Dalton transfer of size ε applied to $\mathbf{x}$.

Strong Schur Convexity of $\Omega_{\mathbf{w}}$ and Exact Grouping

Lemma (Figueiredo and Nowak (2014))

Consider $\Omega_{\mathbf{w}}$, with $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, and let

$$\Delta = \min\{w_1 - w_2,\, w_2 - w_3, \ldots, w_{p-1} - w_p\}.$$

Then, $\Omega_{\mathbf{w}}$ is Δ-strongly Schur-convex.

Strong Schur Convexity of $\Omega_{\mathbf{w}}$ and Exact Grouping

Lemma (Figueiredo and Nowak (2014))

Consider $\Omega_{\mathbf{w}}$, with $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, and let

$$\Delta = \min\{w_1 - w_2,\, w_2 - w_3, \ldots, w_{p-1} - w_p\}.$$

Then, $\Omega_{\mathbf{w}}$ is Δ-strongly Schur-convex.

This lemma underlies the proof of the following theorem

Theorem (Exact grouping (Figueiredo and Nowak, 2014))

Let $\widehat{\mathbf{x}} \in \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Omega_{\mathbf{w}}(\mathbf{x})$; then,

(i) $\|\mathbf{a}_i - \mathbf{a}_j\|_2 < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = \widehat{x}_j$

(ii) $\|\mathbf{a}_i + \mathbf{a}_j\|_2 < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = -\widehat{x}_j$

Exact Grouping Corollaries

Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let $\widehat{\mathbf{x}} \in \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Omega_{\mathbf{w}}(\mathbf{x})$, *assume the columns of* $\mathbf{A}$ *have zero-mean* $\mathbf{1}^T\mathbf{a}_k = 0$ *and unit norm* $\|\mathbf{a}_k\|_2 = 1$, *and* $\rho_{ij} = \mathbf{a}_i^T\mathbf{a}_j$. *Then,*

(i) $\sqrt{2 - 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = \widehat{x}_j$

(ii) $\sqrt{2 + 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = -\widehat{x}_j$

Exact Grouping Corollaries

Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let $\widehat{\mathbf{x}} \in \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Omega_{\mathbf{w}}(\mathbf{x})$*, assume the columns of* $\mathbf{A}$ *have zero-mean* $\mathbf{1}^T\mathbf{a}_k = 0$ *and unit norm* $\|\mathbf{a}_k\|_2 = 1$*, and* $\rho_{ij} = \mathbf{a}_i^T\mathbf{a}_j$*. Then,*

(i) $\sqrt{2 - 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = \widehat{x}_j$

(ii) $\sqrt{2 + 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = -\widehat{x}_j$

- Recovers the theorem by Bondell and Reich (2007) for OSCAR ($\Delta = \lambda_2$), but under much weaker conditions.

Exact Grouping Corollaries

Corollary (Standardized Columns (Figueiredo and Nowak, 2014))

Let $\widehat{\mathbf{x}} \in \arg\min \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \Omega_{\mathbf{w}}(\mathbf{x})$, *assume the columns of* $\mathbf{A}$ *have zero-mean* $\mathbf{1}^T\mathbf{a}_k = 0$ *and unit norm* $\|\mathbf{a}_k\|_2 = 1$, *and* $\rho_{ij} = \mathbf{a}_i^T\mathbf{a}_j$. *Then,*

(i) $\sqrt{2 - 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = \widehat{x}_j$

(ii) $\sqrt{2 + 2\,\rho_{ij}} < \Delta/\|\mathbf{y}\|_2 \;\Rightarrow\; \widehat{x}_i = -\widehat{x}_j$

- Recovers the theorem by Bondell and Reich (2007) for OSCAR ($\Delta = \lambda_2$), but under much weaker conditions.
- Similar results can be proved for the absolute error loss.

Part II: Statistical Analysis

Statistical Bounds

Scenario and assumptions

Statistical Bounds

Scenario and assumptions

- $\mathbf{y} = \mathbf{A}\,\mathbf{x}^{\star} + \mathbf{n}$

Statistical Bounds

Scenario and assumptions

- $\mathbf{y} = \mathbf{A}\,\mathbf{x}^\star + \mathbf{n}$
- $\|\mathbf{x}^\star\|_1 \leq \sqrt{s}\,\|\mathbf{x}\|_2$ (*e.g.*, $\mathbf{x}^\star$ is s-sparse)

Statistical Bounds

Scenario and assumptions

- $\mathbf{y} = \mathbf{A}\,\mathbf{x}^\star + \mathbf{n}$
- $\|\mathbf{x}^\star\|_1 \leq \sqrt{s}\,\|\mathbf{x}\|_2$ (*e.g.*, $\mathbf{x}^\star$ is s-sparse)
- $\frac{1}{n}\|\mathbf{n}\|_1 \leq \varepsilon$ (no other assumptions on the noise)

Statistical Bounds

Scenario and assumptions

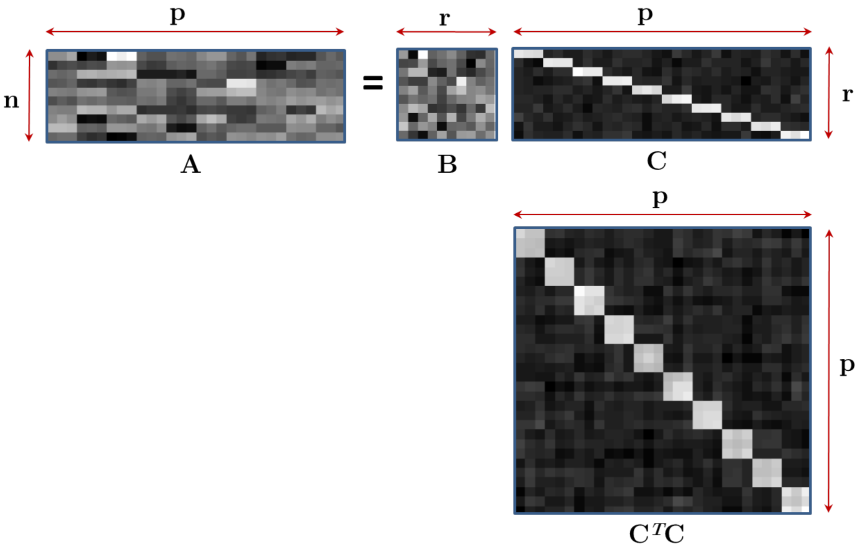
- $\mathbf{y} = \mathbf{A}\,\mathbf{x}^\star + \mathbf{n}$
- $\|\mathbf{x}^\star\|_1 \leq \sqrt{s}\,\|\mathbf{x}\|_2$ (*e.g.*, $\mathbf{x}^\star$ is s-sparse)
- $\frac{1}{n}\|\mathbf{n}\|_1 \leq \varepsilon$ (no other assumptions on the noise)
- Rows of $\mathbf{A} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(0, \mathbf{C}^T\mathbf{C})$

Scenario and assumptions

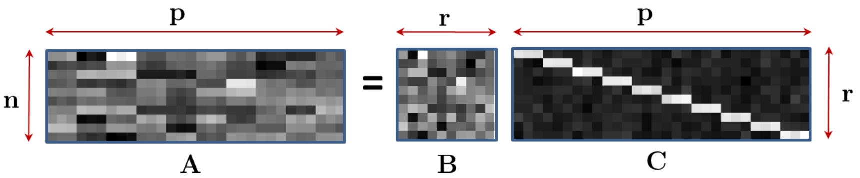
- $\mathbf{y} = \mathbf{A}\,\mathbf{x}^\star + \mathbf{n}$
- $\|\mathbf{x}^\star\|_1 \leq \sqrt{s}\,\|\mathbf{x}\|_2$ (*e.g.*, $\mathbf{x}^\star$ is s-sparse)
- $\frac{1}{n}\|\mathbf{n}\|_1 \leq \varepsilon$ (no other assumptions on the noise)
- Rows of $\mathbf{A} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(0, \mathbf{C}^T\mathbf{C})$
- ..equivalently, $\mathbf{A} = \mathbf{B}\mathbf{C}$, with rows of $\mathbf{B} \in \mathbb{R}^{n\times r}$ i.i.d. $\mathcal{N}(0, \mathbf{I})$, and $\mathbf{C} \in \mathbb{R}^{r\times p}$
- Illustration (exactly replicated columns):

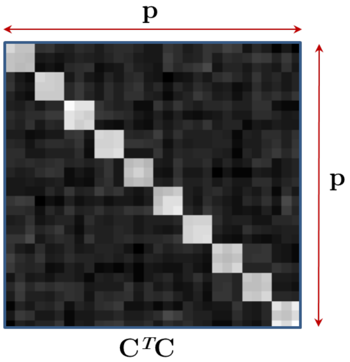
$\mathbf{A}$ $\mathbf{B}$ $\mathbf{C}$

Another Illustration: Highly Correlated Groups of Columns



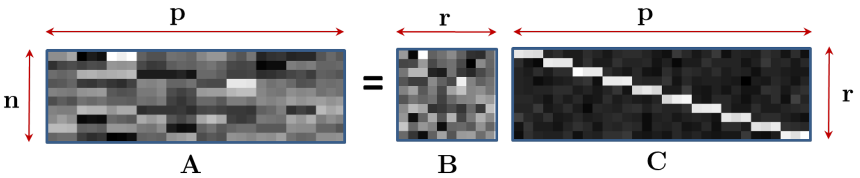
Another Illustration: Highly Correlated Groups of Columns

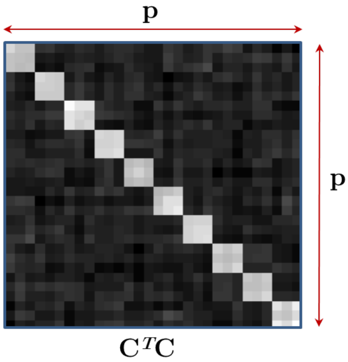




- Notice that $\text{rank}(\mathbf{A}) \leq r$

Another Illustration: Highly Correlated Groups of Columns





- Notice that $\text{rank}(\mathbf{A}) \leq r$
- Similar columns are contiguous only for visualization

Statistical Bound

Theorem (Figueiredo and Nowak (2014))

Let $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}^\star$, *and* ε *be as defined above, and* $\widehat{\mathbf{x}}$ *be a solution to one of the two following problems:*

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \;\; \textit{subject to} \;\; \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \;\leq\; \varepsilon^2$$

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \;\; \textit{subject to} \;\; \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_1 \;\leq\; \varepsilon.$$

Then (with $\gamma(\mathbf{C}) = \min\{\|\mathbf{C}\|_1, \|\mathbf{C}\|_2\}$*)*

$$\mathbb{E}\|\mathbf{C}(\widehat{\mathbf{x}}-\mathbf{x}^\star)\|_2 \leq \sqrt{8\pi}\left(\sqrt{32}\,\gamma(\mathbf{C})\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}}+\varepsilon\right),$$

Statistical Bound

Theorem (Figueiredo and Nowak (2014))

Let $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}^\star$, *and* ε *be as defined above, and* $\widehat{\mathbf{x}}$ *be a solution to one of the two following problems:*

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \quad \textit{subject to} \quad \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \;\leq\; \varepsilon^2$$

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \quad \textit{subject to} \quad \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_1 \;\leq\; \varepsilon.$$

Then (with $\gamma(\mathbf{C}) = \min\{\|\mathbf{C}\|_1, \|\mathbf{C}\|_2\}$*)*

$$\mathbb{E}\|\mathbf{C}(\widehat{\mathbf{x}}-\mathbf{x}^\star)\|_2 \leq \sqrt{8\pi}\left(\sqrt{32}\,\gamma(\mathbf{C})\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}}+\varepsilon\right),$$

- Proof based on techniques and tools by Vershynin (2014).

Statistical Bound

Theorem (Figueiredo and Nowak (2014))

Let $\mathbf{y}$, $\mathbf{A}$, $\mathbf{x}^\star$, *and* ε *be as defined above, and* $\widehat{\mathbf{x}}$ *be a solution to one of the two following problems:*

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \;\; \textit{subject to} \;\; \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 \;\leq\; \varepsilon^2$$

$$\min_{\mathbf{x}\in\mathbb{R}^p} \Omega_{\mathbf{w}}(\mathbf{x}) \;\; \textit{subject to} \;\; \frac{1}{n}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_1 \;\leq\; \varepsilon.$$

Then (with $\gamma(\mathbf{C}) = \min\{\|\mathbf{C}\|_1, \|\mathbf{C}\|_2\}$*)*

$$\mathbb{E}\|\mathbf{C}(\widehat{\mathbf{x}}-\mathbf{x}^\star)\|_2 \leq \sqrt{8\pi}\left(\sqrt{32}\,\gamma(\mathbf{C})\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \varepsilon\right),$$

- Proof based on techniques and tools by Vershynin (2014).
- Key step: extension of the *general $M^\star$ bound* for $\mathbf{A} = \mathbf{B}\mathbf{C}$.

Statistical Bound: Insight From a Special Case

- Columns of $\mathbf{A}$ are either identical or uncorrelated.

Statistical Bound: Insight From a Special Case

- Columns of $\mathbf{A}$ are either identical or uncorrelated.
- Let $\bar{\mathbf{x}}^\star$ have identical components, for identical columns of $\mathbf{A}$.

Statistical Bound: Insight From a Special Case

- Columns of $\mathbf{A}$ are either identical or uncorrelated.
- Let $\bar{\mathbf{x}}^\star$ have identical components, for identical columns of $\mathbf{A}$.
- In this case, the theorem claims that

$$\mathbb{E}\,\|\widehat{\mathbf{x}} - \bar{\mathbf{x}}^\star\|_2 \leq \sqrt{8\pi}\left(4\sqrt{2}\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \varepsilon\right).$$

Statistical Bound: Insight From a Special Case

- Columns of $\mathbf{A}$ are either identical or uncorrelated.
- Let $\bar{\mathbf{x}}^\star$ have identical components, for identical columns of $\mathbf{A}$.
- In this case, the theorem claims that

$$\mathbb{E}\,\|\widehat{\mathbf{x}} - \bar{\mathbf{x}}^\star\|_2 \leq \sqrt{8\pi}\left(4\sqrt{2}\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \varepsilon\right).$$

- *i.e.*, number of samples sufficient to achieve a given precision grows as

$$n \sim s\log p$$

as in bounds with stronger assumptions, *e.g.*, RIP or i.i.d. design (Candès et al., 2006; Candès and Tao, 2007; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)

Statistical Bound: Insight From a Special Case

- Columns of $\mathbf{A}$ are either identical or uncorrelated.
- Let $\bar{\mathbf{x}}^\star$ have identical components, for identical columns of $\mathbf{A}$.
- In this case, the theorem claims that

$$\mathbb{E}\,\|\widehat{\mathbf{x}} - \bar{\mathbf{x}}^\star\|_2 \leq \sqrt{8\pi}\left(4\sqrt{2}\,\|\mathbf{x}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \varepsilon\right).$$

- *i.e.*, number of samples sufficient to achieve a given precision grows as

$$n \sim s\log p$$

  as in bounds with stronger assumptions, *e.g.*, RIP or i.i.d. design (Candès et al., 2006; Candès and Tao, 2007; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)
- No price is paid for the colinearities in $\mathbf{A}$

Conclusions

- **OSCAR**: a regularizer that aims at identifying groups of correlated variables in linear regression.
- OSCAR is a particular case of the **OWL** norm.
- Exact clustering properties of OWL regularization
- Statistical sample complexity bounds for OWL regularization with correlated designs

Conclusions

- **OSCAR**: a regularizer that aims at identifying groups of correlated variables in linear regression.
- OSCAR is a particular case of the **OWL** norm.
- Exact clustering properties of OWL regularization
- Statistical sample complexity bounds for OWL regularization with correlated designs
- Ongoing work: how to select the weights?
- Ongoing work: other losses, *e.g.* logistic, hinge,...

Thank you.

References I

Bach, F. (2008). Consistency of the group Lasso and multiple kernel learning. *JMLR*, 9:1179–1225.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Structured sparsity through convex optimization. *Statistical Science*, 27:450–468.

Bogdan, J., Berg, E., Su, W., and Candes, E. (2013). Statistical estimation and testing via the ordered ℓ_1 norm. *arXiv preprint http://arxiv.org/pdf/1310.1969v1.pdf*.

Bondell, H. and Reich, B. (2007). Regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115–123.

Bühlmann, P., Rüttiman, P., van de Geer, S., and Zhang, C.-H. (2013). Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, pages 1835–1858.

Candès, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35:2313–2351.

Candès, E., Romberg, J., and Tao, T. (2006). *IEEE Transactions on Information Theory*, 52:489–509.

Caruana, R. (1997). Multitask learning. *Machine Learning*, 28(1):41–75.

Chen, S., Donoho, D., and Saunders, M. (1995). Atomic decomposition by basis pursuit. Technical report, Department of Statistics, Stanford University.

References II

Dalton, H. (1920). The measurement of the inequality of incomes. *The Economic Journal*, 30:348–361.

De Mol, C., De Vito, E., and Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity*, 25:201–230.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52:1289–1306.

Figueiredo, M. and Nowak, R. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. Technical report, available at `http://arxiv.org/abs/1409.4005`.

Haupt, J. and Nowak, R. (2006). Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52:4036–4048.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42:80–86.

Huang, J. and Zhang, T. (2010). *Annals of Statistics*, 38:1978–2004.

Marshall, A., Olkin, I., and Arnold, B. (2011). *Inequalities: Theory of Majorization and Its Applications*. Springer, New York.

Martins, A. F. T., Smith, N. A., Aguiar, P. M. Q., and Figueiredo, M. A. T. (2011). Structured Sparsity in Structured Prediction. In *Proc. of Empirical Methods for Natural Language Processing*.

References III

Obozinski, G., Taskar, B., and Jordan, M. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Pigou, A. (1912). *Wealth and Welfare*. Macmillan, London.

Schmidt, M. and Murphy, K. (2010). Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proc. of AISTATS*.

Stojnic, M., Parvaresh, F., and Hassibi, B. (2009). On the reconstruction of block-sparse signals with an optimal number of measurements. *Signal Processing, IEEE Transactions on*, 57(8):3075–3085.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B.*, pages 267–288.

Vershynin, R. (2014). Estimation in high dimensions: A geometric perspective. Technical report, available at `http://arxiv.org/abs/1405.5103`.

Zeng, X. and Figueiredo, M. (2014). Decreasing weighted sorted ℓ_1 regularization. *IEEE Signal Processing Letters*, 21:1240–1244.

Zhong, L. and Kwok, J. (2012). Efficient sparse modeling with automatic feature grouping. *IEEE Transactions on Neural Networks and Learning Systems*, 23:1436–1447.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320.