Dealing with **Correlated Variables** in **Supervised Learning**

Mário A. T. Figueiredo

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

Joint work with: Xiangrong Zeng (NUDT, China),
Robert Nowak (U Wisconsin, USA)

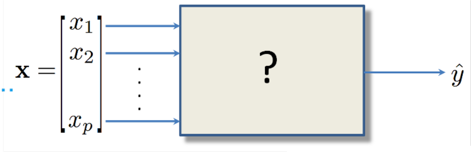
Outline

Outline

Regression

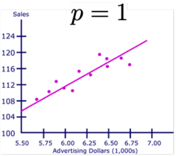
Predict a quantity, from several other quantities

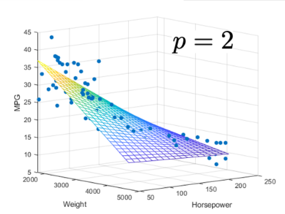
predictors,
explanatory variables, …



response

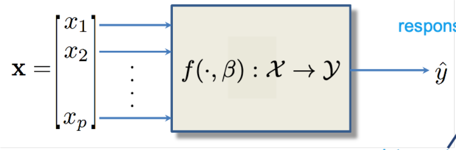
$$\mathbf{x} \in \mathcal{X} = \mathbb{R}^p$$





Linear Regression

Predicted response $=$ linear combination of predictors



predictors

response

change in response per unit change in predictor x_p

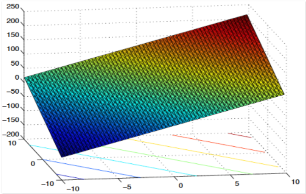
$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = f(\mathbf{x}, \boldsymbol{\beta})$$

intercept, offset (response w/ all predictors = 0)

change in response per unit change in predictor x_1

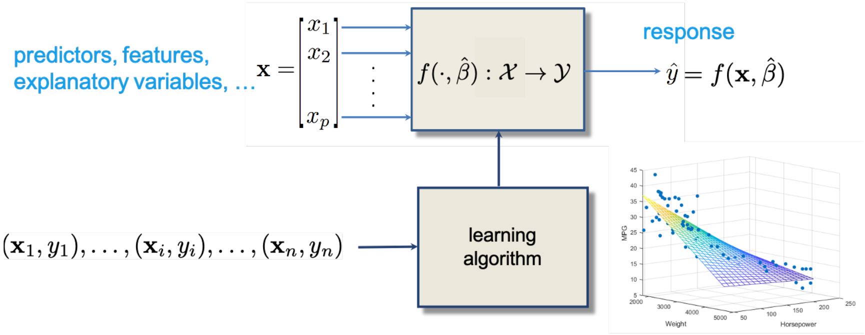
Example:

$$\hat{y} = 50 + 10\, x_1 + 7\, x_2$$



Linear Regression

Learn, from examples, to predict a quantity, from several other quantities



Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- n responses, $y_i \in \mathbb{R}$, modeled as

$$y_i = \ \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(i) + \varepsilon_i = \quad i = 1, ..., n$$

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- n responses, $y_i \in \mathbb{R}$, modeled as

$$y_i = \ \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(i) + \varepsilon_i = \quad i = 1, ..., n$$

- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$ (w.l.o.g. $\beta_0 = 0$)

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- n responses, $y_i \in \mathbb{R}$, modeled as

$$y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(i) + \varepsilon_i = \quad i = 1, ..., n$$

- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$ (w.l.o.g. $\beta_0 = 0$)
- Noise (often assumed Gaussian i.i.d.): $\boldsymbol{\varepsilon} \in \mathbb{R}^n$;

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- n responses, $y_i \in \mathbb{R}$, modeled as

$$y_i = \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(i) + \varepsilon_i = \quad i = 1, ..., n$$

- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$ (w.l.o.g. $\beta_0 = 0$)
- Noise (often assumed Gaussian i.i.d.): $\boldsymbol{\varepsilon} \in \mathbb{R}^n$;
- Vector notation: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \left[\mathbf{x}^{(1)} \cdots \mathbf{x}^{(n)}\right]^T \in \mathbb{R}^{n \times p}$;

Linear Regression

Linear regression: classical old problem (Galton, 1894; Pearson, 1896)

- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- n responses, $y_i \in \mathbb{R}$, modeled as

$$y_i = \; \beta_0 + \boldsymbol{\beta}^T \mathbf{x}(i) + \varepsilon_i = \quad i = 1, ..., n$$

- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$ (w.l.o.g. $\beta_0 = 0$)
- Noise (often assumed Gaussian i.i.d.): $\boldsymbol{\varepsilon} \in \mathbb{R}^n$;
- Vector notation: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X} = \left[\mathbf{x}^{(1)} \cdots \mathbf{x}^{(n)}\right]^T \in \mathbb{R}^{n \times p}$;
- **Goal**: estimate $\boldsymbol{\beta}$, from $\mathbf{y}$ and $\mathbf{X}$

Logistic Regression

Logistic regression: another classical problem, with many applications

Observations: $y_1, ..., y_n$, with $y_i \in \{0, 1\}$ is a sample of r.v. $Y_i|\mathbf{x}^{(i)}$

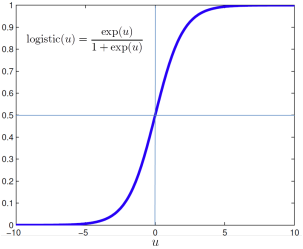
Logistic Regression

Logistic regression: another classical problem, with many applications

Observations: $y_1, ..., y_n$, with $y_i \in \{0, 1\}$ is a sample of r.v. $Y_i|\mathbf{x}^{(i)}$

$$\mathbb{E}(Y_i|\mathbf{x}^{(i)}) = \mathbb{P}(Y_i = 1|\mathbf{x}^{(i)})$$
$$= \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)})$$

$$\sigma(u) = \frac{e^u}{1 + e^u}$$



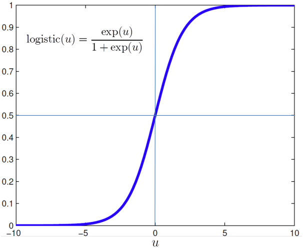
Logistic Regression

Logistic regression: another classical problem, with many applications

Observations: $y_1, ..., y_n$, with $y_i \in \{0, 1\}$ is a sample of r.v. $Y_i|\mathbf{x}^{(i)}$

$$\mathbb{E}(Y_i|\mathbf{x}^{(i)}) = \mathbb{P}(Y_i = 1|\mathbf{x}^{(i)}) = \sigma(\boldsymbol{\beta}^T \mathbf{x}^{(i)})$$

$$\sigma(u) = \frac{e^u}{1 + e^u}$$



- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$;

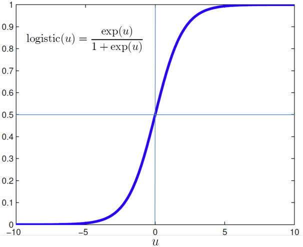
Logistic Regression

Logistic regression: another classical problem, with many applications

Observations: $y_1, ..., y_n$, with $y_i \in \{0, 1\}$ is a sample of r.v. $Y_i|\mathbf{x}^{(i)}$

$$\mathbb{E}(Y_i|\mathbf{x}^{(i)}) = \mathbb{P}(Y_i = 1|\mathbf{x}^{(i)}) = \sigma(\boldsymbol{\beta}^T\mathbf{x}^{(i)})$$

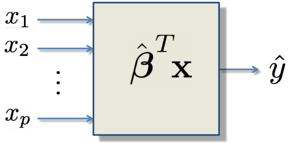
$$\sigma(u) = \frac{e^u}{1 + e^u}$$



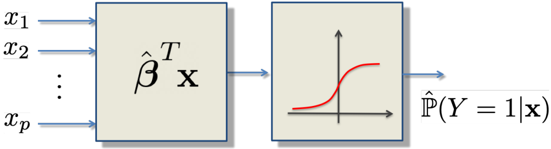
- n vectors of p features/variables: $\mathbf{x}^{(i)} \in \mathbb{R}^p$, for $i = 1, ..., n$.
- Regression coefficients: $\boldsymbol{\beta} \in \mathbb{R}^p$;
- **Goal**: estimate $\boldsymbol{\beta}$, from $\mathbf{y}$ and $\mathbf{x}^{(1)}, ..., \mathbf{x}^{(n)}$

Feature Selection

Linear regression:



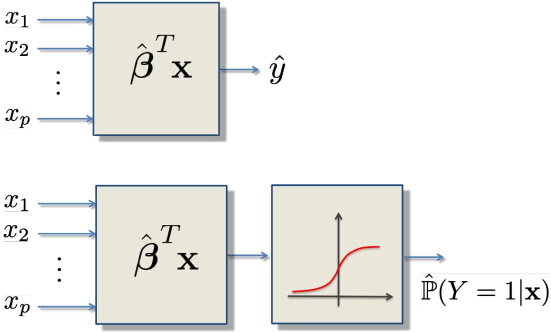
Logistic regression (binary):



Feature Selection

Linear regression:

Logistic regression (binary):

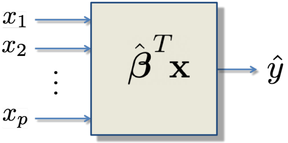


Questions:

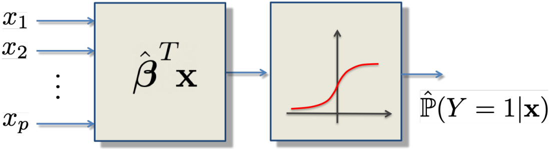
- Are some variables/features irrelevant? (should some $\hat{\beta}_j$ be zero?)

Feature Selection

Linear regression:



Logistic regression (binary):



Questions:

- Are some variables/features irrelevant? (should some $\hat{\beta}_j$ be zero?)
- Are there redundant features?

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

- Filters: ignore the subsequent learning algorithm

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

- Filters: ignore the subsequent learning algorithm
    - ◇ Supervised: use the desired responses (*e.g.* labels)

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

- Filters: ignore the subsequent learning algorithm
  - Supervised: use the desired responses (*e.g.* labels)
  - Unsupervised: use only the features

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

- Filters: ignore the subsequent learning algorithm
  - ◇ Supervised: use the desired responses (*e.g.* labels)
  - ◇ Unsupervised: use only the features
- Wrappers: make use of a learning algorithm to assess feature subsets

Feature Selection

Three main classes of feature selection methods
(Das, 1994; Escolano et al, 2009; Guyon et al., 2003):

- Filters: ignore the subsequent learning algorithm
  - Supervised: use the desired responses (*e.g.* labels)
  - Unsupervised: use only the features
- Wrappers: make use of a learning algorithm to assess feature subsets
- **Embedded**: in the learning algorithm

Embedded Feature/Variable Selection

Formulate the learning problem as a trade-off between

Embedded Feature/Variable Selection

Formulate the learning problem as a trade-off between

- accuracy (*i.e.*, fitting data well): minimize a loss

Embedded Feature/Variable Selection

Formulate the learning problem as a trade-off between

- accuracy (*i.e.*, fitting data well): minimize a loss
- "desirability" (*e.g.*, no irrelevant variables): minimizing a regularizer

Embedded Feature/Variable Selection

Formulate the learning problem as a trade-off between

- accuracy (*i.e.*, fitting data well): minimize a loss
- "desirability" (*e.g.*, no irrelevant variables): minimizing a regularizer
- Optimization (regularization) formulation:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \underbrace{\sum_{i=1}^{n} l(\boldsymbol{\beta}; y_i, \mathbf{x}(i))}_{L(\boldsymbol{\beta})} + R(\boldsymbol{\beta})$$

Embedded Feature/Variable Selection

Formulate the learning problem as a trade-off between

- accuracy (*i.e.*, fitting data well): minimize a loss
- "desirability" (*e.g.*, no irrelevant variables): minimizing a regularizer
- Optimization (regularization) formulation:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \underbrace{\sum_{i=1}^{n} l(\boldsymbol{\beta}; y_i, \mathbf{x}(i))}_{L(\boldsymbol{\beta})} + R(\boldsymbol{\beta})$$

- Often yields well-understood, solvable optimization problems, with analyzable solutions

Regularization and Sparsity in Linear Regression

Regularized linear regression criteria (classical choices):

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}), \quad \text{with } L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

Regularization and Sparsity in Linear Regression

Regularized linear regression criteria (classical choices):

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}), \quad \text{with } L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

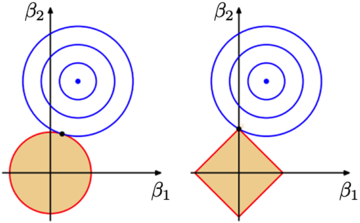
- The old classic: ridge regression (Wiener, 1949; Hoerl and Kennard, 1970):
  $R(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2^2 \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}.$

Regularization and Sparsity in Linear Regression

Regularized linear regression criteria (classical choices):

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}), \quad \text{with } L(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2$$

- The old classic: ridge regression (Wiener, 1949; Hoerl and Kennard, 1970):
  $R(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_2^2 \quad \Rightarrow \quad \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}$.
- The new classic: LASSO, $R(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$
  (Claerbout and Muir, 1973; Taylor et al., 1979; Levy and Fullagar, 1981; Chen et al., 1995; Williams, 1995; Tibshirani, 1996; Bühlmann and van de Geer, 2011):



Sparsity! (variable selection)

Regularization and Sparsity in Logistic Regression

Regularized logistic regression:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}),$$

with the logistic loss: $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\bigl(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}^{(i)})\bigr) - y_i\, \boldsymbol{\beta}^T \mathbf{x}^{(i)}$

Regularization and Sparsity in Logistic Regression

Regularized logistic regression:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}),$$

with the logistic loss: $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\big(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}^{(i)})\big) - y_i\, \boldsymbol{\beta}^T \mathbf{x}^{(i)}$

- The old classic: ridge logistic regression, $R(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_2^2$
  (Schaefer et al., 1984; Cessie and Houwelingen, 1992)

Regularization and Sparsity in Logistic Regression

Regularized logistic regression:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}),$$

with the logistic loss: $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\big(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}^{(i)})\big) - y_i\, \boldsymbol{\beta}^T \mathbf{x}^{(i)}$

- The old classic: ridge logistic regression, $R(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_2^2$
  (Schaefer et al., 1984; Cessie and Houwelingen, 1992)

- The new classic: sparse logistic regression, $R(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$
  (Tibshirani, 1996; Shevade and Keerthi, 2003; Krishnapuram et al., 2005)

Regularization and Sparsity in Logistic Regression

Regularized logistic regression:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + R(\boldsymbol{\beta}),$$

with the logistic loss: $L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log\big(1 + \exp(\boldsymbol{\beta}^T \mathbf{x}^{(i)})\big) - y_i\, \boldsymbol{\beta}^T \mathbf{x}^{(i)}$

- The old classic: ridge logistic regression, $R(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_2^2$
  (Schaefer et al., 1984; Cessie and Houwelingen, 1992)

- The new classic: sparse logistic regression, $R(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$
  (Tibshirani, 1996; Shevade and Keerthi, 2003; Krishnapuram et al., 2005)

Many other generalized linear models (GLM) can be used.
(Bühlmann and van de Geer, 2011)

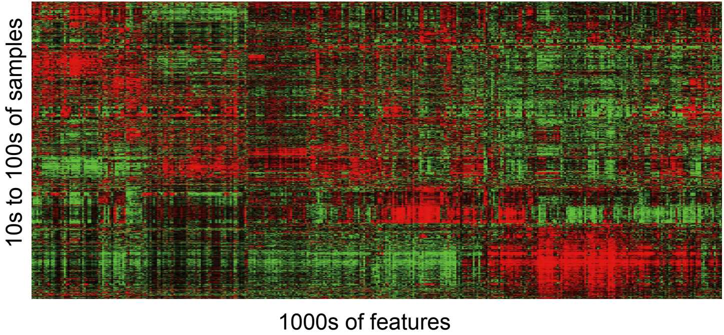
Outline

Group Sparsity with Unknown Groups

- **Goal**: identify **all** the relevant features/variables
  Why/when? If the variables have meaning (*e.g.*, genes, voxels,...)

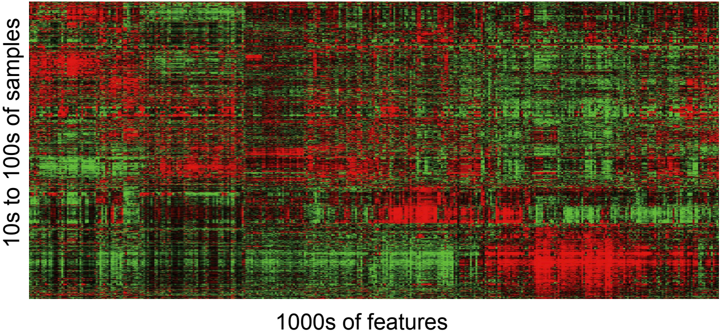
Group Sparsity with Unknown Groups

- **Goal**: identify **all** the relevant features/variables
  Why/when? If the variables have meaning (*e.g.*, genes, voxels,...)



Group Sparsity with Unknown Groups

- **Goal**: identify **all** the relevant features/variables
  Why/when? If the variables have meaning (*e.g.*, genes, voxels,...)



- Goal: not only good prediction, also identify all involved genes

LASSO with Highly-Correlated Features

Problem: with highly correlated variables, LASSO may select an arbitrary subset of variables; also, it is unstable (Bühlmann et al., 2013)

LASSO with Highly-Correlated Features

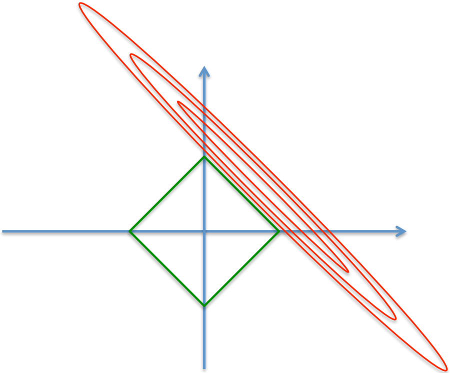
Problem: with highly correlated variables, LASSO may select an arbitrary subset of variables; also, it is unstable (Bühlmann et al., 2013)

Why?

LASSO with Highly-Correlated Features

Problem: with highly correlated variables, LASSO may select an arbitrary subset of variables; also, it is unstable (Bühlmann et al., 2013)

Why? $$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2 \;\text{ s.t. }\; \|\boldsymbol{\beta}\|_1 \leq \delta$$



Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known

Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known
- Approach 1: use covariance information ($\boldsymbol{\Sigma}$) to...

Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known
- Approach 1: use covariance information ($\boldsymbol{\Sigma}$) to...
  - ...infer and use groups (cluster LASSO) (Bühlmann et al., 2013)
    - ✓ *e.g.*, use group-LASSO or a representative variable per group

Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known
- Approach 1: use covariance information ($\boldsymbol{\Sigma}$) to...
  - ...infer and use groups (cluster LASSO) (Bühlmann et al., 2013)
    - ✓ *e.g.*, use group-LASSO or a representative variable per group
  - ...infer and use a correlation graph (Li et al, 2018)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j,k} |\Sigma_{j,k}| \big(\beta_j - \mathrm{sign}(\Sigma_{j,k})\beta_k\big)^2 + \lambda_3 \sum_{j,k} |\Sigma_{j,k}|^{\frac{1}{2}} \big|\beta_j - \mathrm{sign}(\Sigma_{j,k})\beta_k\big|$$

Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known
- Approach 1: use covariance information ($\boldsymbol{\Sigma}$) to...
  - ...infer and use groups (cluster LASSO) (Bühlmann et al., 2013)
    - ✓ *e.g.*, use group-LASSO or a representative variable per group
  - ...infer and use a correlation graph (Li et al, 2018)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j,k} |\Sigma_{j,k}| \big(\beta_j - \mathsf{sign}(\Sigma_{j,k})\beta_k\big)^2 + \lambda_3 \sum_{j,k} |\Sigma_{j,k}|^{\frac{1}{2}} \big|\beta_j - \mathsf{sign}(\Sigma_{j,k})\beta_k\big|$$

  - Several variants of this approach (Daye and Jeng, 2009; Hebiri and van de Geer, 2011; Kim and Xing, 2009; Sharma et al, 2013; She, 2010; Veríssimo et al, 2016)

Highly-Correlated Variables

- Group sparsity (group-LASSO) could help, **if** the groups **were** known
- Approach 1: use covariance information ($\boldsymbol{\Sigma}$) to...
  - ◇ ...infer and use groups (cluster LASSO) (Bühlmann et al., 2013)
    - ✓ *e.g.*, use group-LASSO or a representative variable per group
  - ◇ ...infer and use a correlation graph (Li et al, 2018)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\,\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{j,k} |\Sigma_{j,k}| \big(\beta_j - \mathsf{sign}(\Sigma_{j,k})\beta_k\big)^2 + \lambda_3 \sum_{j,k} |\Sigma_{j,k}|^{\frac{1}{2}} \big|\beta_j - \mathsf{sign}(\Sigma_{j,k})\beta_k\big|$$

  - ◇ Several variants of this approach (Daye and Jeng, 2009; Hebiri and van de Geer, 2011; Kim and Xing, 2009; Sharma et al, 2013; She, 2010; Veríssimo et al, 2016)
  - ◇ Key aspect: at least $O(n\,p^2)$; expensive in high-dimensional problems

Elastic Net (EN) and OSCAR

- Approach 2: regularizers that handle highly-correlated variables
  - Elastic net (EN) (Zou and Hastie, 2005; De Mol et al., 2009)
    Goal: include groups of correlated variables.

Elastic Net (EN) and OSCAR

- Approach 2: regularizers that handle highly-correlated variables
  - ◇ Elastic net (EN) (Zou and Hastie, 2005; De Mol et al., 2009)
    Goal: include groups of correlated variables.
  - ◇ Octagonal shrinkage and clustering algorithm for regression (OSCAR) (Bondell and Reich, 2007; Zhong and Kwok, 2012)
    Goal: exactly group sufficiently correlated variables

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

Elastic Net (EN) and OSCAR

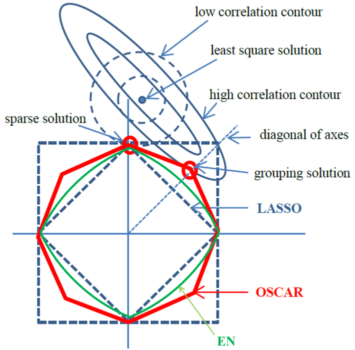
- Approach 2: regularizers that handle highly-correlated variables
  - ◇ Elastic net (EN) (Zou and Hastie, 2005; De Mol et al., 2009)
    Goal: include groups of correlated variables.
  - ◇ Octagonal shrinkage and clustering algorithm for regression (OSCAR) (Bondell and Reich, 2007; Zhong and Kwok, 2012)
    Goal: exactly group sufficiently correlated variables

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

- OSCAR:
  $R(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$

Elastic Net (EN) and OSCAR

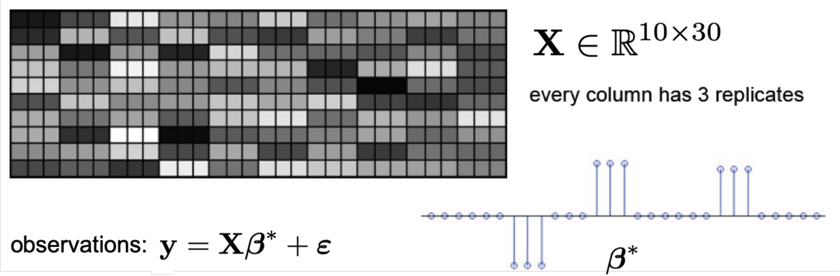
- Approach 2: regularizers that handle highly-correlated variables
  - ◇ Elastic net (EN) (Zou and Hastie, 2005; De Mol et al., 2009)
    Goal: include groups of correlated variables.
  - ◇ Octagonal shrinkage and clustering algorithm for regression (OSCAR) (Bondell and Reich, 2007; Zhong and Kwok, 2012)
    Goal: exactly group sufficiently correlated variables



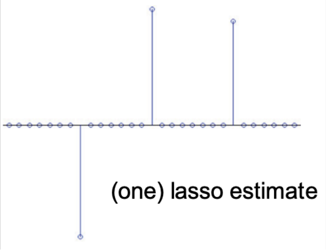
- Elastic net:
  $R(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \|\boldsymbol{\beta}\|_2^2$

- OSCAR:
  $R(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$

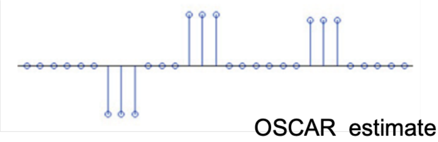
Toy example



$\mathbf{X} \in \mathbb{R}^{10\times 30}$
every column has 3 replicates
observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$
$\boldsymbol{\beta}^*$



(one) lasso estimate



OSCAR estimate

OSCAR on Synthetic Data (Bondell and Reich, 2007)

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

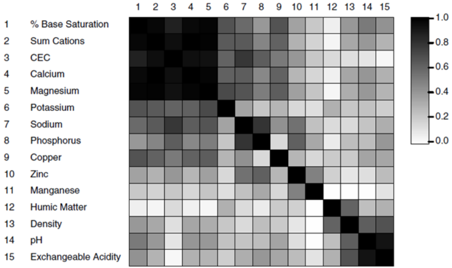
OSCAR on Synthetic Data (Bondell and Reich, 2007)

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

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

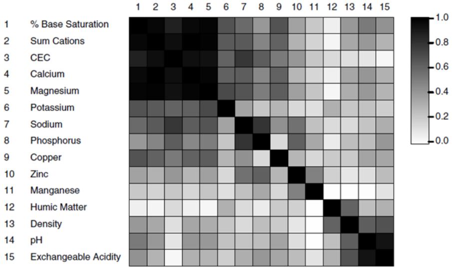
OSCAR yields explicit variable grouping (Bondell and Reich, 2007)

Real Data: Plant Diversity vs Soil Chemistry



(Bondell and Reich, 2007)

Real Data: Plant Diversity vs Soil Chemistry



(Bondell and Reich, 2007)

Estimated coefficients for the soil data example

Variable	OSCAR (5-fold CV)	OSCAR (GCV)	LASSO (5-fold CV)	LASSO (GCV)
% Base saturation	0	−0.073	0	0
Sum cations	−0.178	−0.174	0	0
CEC	−0.178	−0.174	−0.486	0
Calcium	−0.178	−0.174	0	−0.670
Magnesium	0	0	0	0
Potassium	−0.178	−0.174	−0.189	−0.250
Sodium	0	0	0	0
Phosphorus	0.091	0.119	0.067	0.223
Copper	0.237	0.274	0.240	0.400
Zinc	0	0	0	−0.129
Manganese	0.267	0.274	0.293	0.321
Humic matter	−0.541	−0.558	−0.563	−0.660
Density	0	0	0	0
pH	0.145	0.174	0.013	0.225
Exchangeable acidity	0	0	0	0

Generalizing OSCAR: The OWL

OSCAR: $$R_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$$

Generalizing OSCAR: The OWL

OSCAR:
$$R_{\text{OSCAR}}^{\lambda_1,\lambda_2}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$$
$$= \sum_{i=1}^{p} \underbrace{\big(\lambda_1 + \lambda_2(p-i)\big)}_{\text{linearly decreasing sequence}} \; |\beta|_{[i]},$$

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

Generalizing OSCAR: The OWL

OSCAR: $$R^{\lambda_1,\lambda_2}_{\text{OSCAR}}(\boldsymbol{\beta}) = \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$$
$$= \sum_{i=1}^{p} \underbrace{\big(\lambda_1 + \lambda_2(p-i)\big)}_{\text{linearly decreasing sequence}} \; |\beta|_{[i]},$$

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

Generalization: the **ordered weighted ℓ_1 (OWL)** norm (a.k.a. **SLOPE**) (**?**Zeng and F, 2014a)

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

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

Generalizing OSCAR: The OWL

OSCAR: $$R^{\lambda_1,\lambda_2}_{\text{OSCAR}}(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{i<j} \max\{|\beta_i|, |\beta_j|\}$$
$$= \sum_{i=1}^{p} \underbrace{\big(\lambda_1 + \lambda_2(p-i)\big)}_{\text{linearly decreasing sequence}} \; |\beta|_{[i]},$$

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

Generalization: the **ordered weighted ℓ_1** **(OWL)** norm (a.k.a. **SLOPE**) (**?**Zeng and F, 2014a)

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

where $w_1 \geq w_2 \geq \cdots \geq w_p \geq 0$ and $|\boldsymbol{\beta}|_{\downarrow} = \big[|\beta|_{[1]}, \, |\beta|_{[2]}, \ldots, |\beta|_{[p]}\big]^T$

Some Properties of the OWL

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

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

Some Properties of the OWL

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

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

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

Some Properties of the OWL

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

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

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.
- Relationship with ℓ_1 (with $\bar{w} = \frac{1}{p}\sum_{i=1}^{p} w_i$):

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

equalities if $w_1 = w_2 = \cdots = w_p$;

Some Properties of the OWL

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

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

- $\Omega_{\mathbf{w}} : \mathbb{R}^p \rightarrow \mathbb{R}_+$ is indeed a norm, iff $w_1 > 0$.
- Relationship with ℓ_1 (with $\bar{w} = \frac{1}{p}\sum_{i=1}^{p} w_i$):

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

  equalities if $w_1 = w_2 = \cdots = w_p$;

- Relationship with ℓ_∞:

$$w_1 \, \|\boldsymbol{\beta}\|_\infty \leq \Omega_{\mathbf{w}}(\boldsymbol{\beta}),$$

  with equality if $w_2 = w_3 = \cdots = w_p = 0$.

Real Data: Machine Translation

- OWL in machine translation (MT)
  (Clark, 2015)

Real Data: Machine Translation

- OWL in machine translation (MT) (Clark, 2015)

Condition	#NonZero	#Unique	Tune BLEU	Test BLEU
Zh→En				
Baseline (ℓ_2)	~10^6	~10^6	29.4	23.5
SparseFeats ($\ell_1 + \ell_2$)	75,952	55,746	36.0	22.8* (-0.7)
SparseFeats (**OWL**)	677,728	922	34.4	**24.1*** (+0.6)
Ar→En				
Baseline (ℓ_2)	~10^6	~10^6	42.2	47.7
SparseFeats ($\ell_1 + \ell_2$)	79,961	56,657	50.5	48.3* (+0.6)
SparseFeats (**OWL**)	824,563	608	48.6	**49.4*** (+1.7)
Cz→En				
Baseline (ℓ_2)	~10^6	~10^6	33.3	38.5
SparseFeats ($\ell_1 + \ell_2$)	115,210	77,732	36.9	38.3* (-0.2)
SparseFeats (**OWL**)	826,440	2,273	35.8	**38.7*** (+0.2)

- Massive parameter/weight sharing

Real Data: Machine Translation

- OWL in machine translation (MT) (Clark, 2015)

Condition	#NonZero	#Unique	Tune BLEU	Test BLEU
		Zh→En		
Baseline (ℓ_2)	~10^6	~10^6	29.4	23.5
SparseFeats (ℓ_1 + ℓ_2)	75,952	55,746	36.0	22.8* (-0.7)
SparseFeats (OWL)	677,728	922	34.4	**24.1*** (+0.6)
		Ar→En		
Baseline (ℓ_2)	~10^6	~10^6	42.2	47.7
SparseFeats (ℓ_1 + ℓ_2)	79,961	56,657	50.5	48.3* (+0.6)
SparseFeats (OWL)	824,563	608	48.6	**49.4*** (+1.7)
		Cz→En		
Baseline (ℓ_2)	~10^6	~10^6	33.3	38.5
SparseFeats (ℓ_1 + ℓ_2)	115,210	77,732	36.9	38.3* (-0.2)
SparseFeats (OWL)	826,440	2,273	35.8	**38.7*** (+0.2)

- Massive parameter/weight sharing
- OWL does adaptive weight sharing (cf. Nowlan and Hinton (1992))

Real Data: fMRI

- Neural decoding from fMRI data (Li et al, 2018)
  ($n = 90$, $p \sim [5, 9] \times 10^3$)

Real Data: fMRI

- Neural decoding from fMRI data (Li et al, 2018)
  ($n = 90$, $p \sim [5, 9] \times 10^3$)

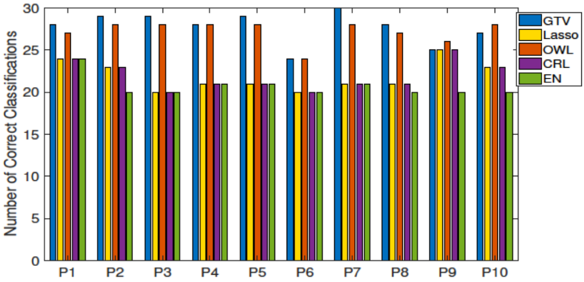


FIG 8. *Classification accuracy for five different methods applied to an fMRI study in which participants were asked to view images of faces and non-faces, described in Section 5. The horizontal axis represents ten different participants denoted from P1 to P10 and the vertical axis represents number of correct classifications for each method.*

- OWL performs competitively with GTV (Li et al, 2018), which uses correlation information and is much more costly.

Outline

Majorization

Key tools to understand OSCAR/OWL: majorization and Schur-convexity

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

Consider $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$; we say that $\boldsymbol{\gamma}$ majorizes $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta} \prec \boldsymbol{\gamma}$, if

$$\begin{aligned} \sum_{i=1}^{p} \beta_i &= \sum_{i=1}^{p} \gamma_i \\ \sum_{i=1}^{j} \beta_{[i]} &\leq \sum_{i=1}^{j} \gamma_{[i]}, \quad \text{for } j = 1, ..., p-1 \end{aligned}$$

Majorization

Key tools to understand OSCAR/OWL: majorization and Schur-convexity

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

Consider $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$; we say that $\boldsymbol{\gamma}$ majorizes $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta} \prec \boldsymbol{\gamma}$, if

$$\begin{aligned} \sum_{i=1}^{p} \beta_i &= \sum_{i=1}^{p} \gamma_i \\ \sum_{i=1}^{j} \beta_{[i]} &\leq \sum_{i=1}^{j} \gamma_{[i]}, \quad \text{for } j = 1, ..., p-1 \end{aligned}$$

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

Majorization

Key tools to understand OSCAR/OWL: majorization and Schur-convexity

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

Consider $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^p$; we say that $\boldsymbol{\gamma}$ majorizes $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta} \prec \boldsymbol{\gamma}$, if

$$\begin{aligned} \sum_{i=1}^{p} \beta_i &= \sum_{i=1}^{p} \gamma_i \\ \sum_{i=1}^{j} \beta_{[i]} &\leq \sum_{i=1}^{j} \gamma_{[i]}, \quad \text{for } j = 1, ..., p-1 \end{aligned}$$

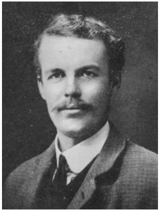
- Examples: $(1,1,1,1) \prec (2,1,1,0) \prec (3,1,0,0) \prec (4,0,0,0)$
- $\boldsymbol{\beta} \prec \boldsymbol{\gamma}$ and $\boldsymbol{\gamma} \prec \boldsymbol{\beta}$, iff $\boldsymbol{\beta}$ is a permutation of $\boldsymbol{\gamma}$

Majorization

- Majorization theory originated in the study of economic inequality

(Dalton, 1920)

(Lorenz, 1905)



(Pigou, 1912)

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma} \in \mathcal{A}}, \;\; \boldsymbol{\beta} \prec \boldsymbol{\gamma} \;\; \Rightarrow \;\; f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathcal{A}}, \;\; \boldsymbol{\beta} \prec \boldsymbol{\gamma} \;\; \Rightarrow \;\; f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Furthermore, f is strictly Schur-convex, if

$$(\boldsymbol{\beta} \prec \boldsymbol{\gamma}) \wedge (\boldsymbol{\beta} \not\succ \boldsymbol{\gamma}) \Rightarrow f(\boldsymbol{\beta}) < f(\boldsymbol{\gamma})$$

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathcal{A}}, \;\; \boldsymbol{\beta} \prec \boldsymbol{\gamma} \;\; \Rightarrow \;\; f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Furthermore, f is strictly Schur-convex, if

$$(\boldsymbol{\beta} \prec \boldsymbol{\gamma}) \wedge (\boldsymbol{\beta} \not\succ \boldsymbol{\gamma}) \Rightarrow f(\boldsymbol{\beta}) < f(\boldsymbol{\gamma})$$

- Schur-convex (SC) functions "favour" more uniform vectors

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathcal{A}}, \quad \boldsymbol{\beta} \prec \boldsymbol{\gamma} \quad \Rightarrow \quad f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Furthermore, f is strictly Schur-convex, if

$$(\boldsymbol{\beta} \prec \boldsymbol{\gamma}) \wedge (\boldsymbol{\beta} \not\succ \boldsymbol{\gamma}) \Rightarrow f(\boldsymbol{\beta}) < f(\boldsymbol{\gamma})$$

- Schur-convex (SC) functions "favour" more uniform vectors
- Symmetric (invariant under argument permutations) convex functions are SC

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma}\in\mathcal{A}}, \;\; \boldsymbol{\beta} \prec \boldsymbol{\gamma} \;\; \Rightarrow \;\; f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Furthermore, f is strictly Schur-convex, if

$$(\boldsymbol{\beta} \prec \boldsymbol{\gamma}) \wedge (\boldsymbol{\beta} \not\succ \boldsymbol{\gamma}) \Rightarrow f(\boldsymbol{\beta}) < f(\boldsymbol{\gamma})$$

- Schur-convex (SC) functions "favour" more uniform vectors
- Symmetric (invariant under argument permutations) convex functions are SC
- Example: ℓ_1 norm is SC, but not strictly

Majorization and Schur-Convexity

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

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

$$\forall_{\boldsymbol{\beta},\boldsymbol{\gamma} \in \mathcal{A}}, \;\; \boldsymbol{\beta} \prec \boldsymbol{\gamma} \;\; \Rightarrow \;\; f(\boldsymbol{\beta}) \leq f(\boldsymbol{\gamma}).$$

Furthermore, f is strictly Schur-convex, if

$$(\boldsymbol{\beta} \prec \boldsymbol{\gamma}) \wedge (\boldsymbol{\beta} \not\succ \boldsymbol{\gamma}) \Rightarrow f(\boldsymbol{\beta}) < f(\boldsymbol{\gamma})$$

- Schur-convex (SC) functions "favour" more uniform vectors
- Symmetric (invariant under argument permutations) convex functions are SC
- Example: ℓ_1 norm is SC, but not strictly
- Example: ℓ_2 norm and Shannon entropy are strictly SC

Pigou-Dalton Transfers

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

Let $\boldsymbol{\beta} \in \mathbb{R}_+^p$ and two components, β_i, β_j, s.t. $\beta_i > \beta_j$. A Pigou-Dalton transfer of size $\varepsilon \in \big(0, \, (\beta_i - \beta_j)/2\big)$ applied to $\boldsymbol{\beta}$ yields $\boldsymbol{\gamma} \prec \boldsymbol{\beta}$ according to

$$\gamma_i = \beta_i - \varepsilon, \quad \gamma_j = \beta_j + \varepsilon, \quad \gamma_k = \beta_k, \;\; \text{for } k \neq i, j.$$

Pigou-Dalton Transfers

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

Let $\boldsymbol{\beta} \in \mathbb{R}_+^p$ and two components, β_i, β_j, s.t. $\beta_i > \beta_j$. A Pigou-Dalton transfer of size $\varepsilon \in \big(0,\, (\beta_i - \beta_j)/2\big)$ applied to $\boldsymbol{\beta}$ yields $\boldsymbol{\gamma} \prec \boldsymbol{\beta}$ according to

$$\gamma_i = \beta_i - \varepsilon, \quad \gamma_j = \beta_j + \varepsilon, \quad \gamma_k = \beta_k, \ \text{for } k \neq i, j.$$

- Pigou-Dalton transfer: also known as the Robin-Hood transfer (Arnold, 1987)

Strong Schur-Convexity

Definition (Strong Schur convexity (F and Nowak, 2016)**)**

A function f is S-strongly Schur-convex if, for $S > 0$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\gamma}) \geq \varepsilon S,$$

when $\boldsymbol{\gamma}$ result from a Pigou-Dalton transfer of size ε applied to $\boldsymbol{\beta}$.

Strong Schur-Convexity

Definition (Strong Schur convexity (F and Nowak, 2016))

A function f is S-strongly Schur-convex if, for $S > 0$,

$$f(\boldsymbol{\beta}) - f(\boldsymbol{\gamma}) \geq \varepsilon S,$$

when $\boldsymbol{\gamma}$ result from a Pigou-Dalton transfer of size ε applied to $\boldsymbol{\beta}$.

- Strong SC $\Rightarrow$ strict SC
- Strong SC $\not\Leftarrow$ strict SC

Strong Schur Convexity of $\Omega_{\mathbf{w}}$

Lemma (F and Nowak (2016))

Consider $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^T|\boldsymbol{\beta}|_{\downarrow}$, *with* $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, *and let*

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

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

Strong Schur Convexity of $\Omega_{\mathbf{w}}$

Lemma (F and Nowak (2016))

Consider $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^T|\boldsymbol{\beta}|_{\downarrow}$, *with* $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, *and let*

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

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

- The ℓ_1 norm is not strongly (it is not even strictly) SC

Strong Schur Convexity of $\Omega_{\mathbf{w}}$

Lemma (F and Nowak (2016))

Consider $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^T |\boldsymbol{\beta}|_{\downarrow}$, *with* $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, *and let*

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

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

- The ℓ_1 norm is not strongly (it is not even strictly) SC
- The ℓ_2 norm and the EN regularizer are strictly, but not strongly, SC

Strong Schur Convexity of $\Omega_{\mathbf{w}}$

Lemma (F and Nowak (2016))

Consider $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \mathbf{w}^T |\boldsymbol{\beta}|_{\downarrow}$, *with* $w_1 \geq w_2 \geq \cdots \geq x_p \geq 0$, *and let*

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

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

- The ℓ_1 norm is not strongly (it is not even strictly) SC
- The ℓ_2 norm and the EN regularizer are strictly, but not strongly, SC
- This lemma is key to the proofs of the following theorems

Exact Grouping: Squared Error Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i$ *the* i*-th column of* $\mathbf{X}$

$$
\begin{array}{lll}
\textbf{(a)} & \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = \widehat{\beta}_j \\
\textbf{(b)} & \|\mathbf{x}_i + \mathbf{x}_j\|_2 < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = -\widehat{\beta}_j
\end{array}
$$

Exact Grouping: Squared Error Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i$ *the* i*-th column of* $\mathbf{X}$

$$\textbf{(a)} \quad \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \|\mathbf{x}_i + \mathbf{x}_j\|_2 < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = -\widehat{\beta}_j$$

Corollary

Let the columns of $\mathbf{X}$ *satisfy* $\mathbf{1}^T\mathbf{x}_k = 0$ *and* $\|\mathbf{x}_k\|_2 = 1$*, for* $k = 1, ..., p$*.*
Denote $\rho_{ij} = \mathbf{x}_i^T\mathbf{x}_j \in [-1, 1]$ *the sample correlation. Then,*

$$\textbf{(a)} \quad \sqrt{2 - 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \sqrt{2 + 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\|\mathbf{y}\|_2 \;\;\Rightarrow\;\; \widehat{\beta}_i = -\widehat{\beta}_j.$$

Significantly extends a theorem by Bondell and Reich (2007)

Exact Grouping: Absolute Error Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_1 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i$ *the* i*-th column of* $\mathbf{X}$

$$\textbf{(a)} \quad \|\mathbf{x}_i - \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \|\mathbf{x}_i + \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j$$

Exact Grouping: Absolute Error Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_1 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i$ *the* i*-th column of* $\mathbf{X}$

$$
\begin{array}{lll}
\textbf{(a)} & \|\mathbf{x}_i - \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} & \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j \\
\textbf{(b)} & \|\mathbf{x}_i + \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} & \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j
\end{array}
$$

Corollary

Let the columns of $\mathbf{A}$ *satisfy* $\mathbf{1}^T\mathbf{x}_k = 0$ *and* $\|\mathbf{x}_k\|_2 = 1$*, for* $k = 1, ..., p$*. Denote* $\rho_{ij} = \mathbf{x}_i^T\mathbf{x}_j \in [-1, 1]$ *the sample correlation. Then,*

$$
\begin{array}{lll}
\textbf{(a)} & \sqrt{2 - 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\sqrt{n} & \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j \\
\textbf{(b)} & \sqrt{2 + 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\sqrt{n} & \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j.
\end{array}
$$

Exact Grouping: Logistic Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i = [x_{(1),i}, ..., x_{(n),i}]^T$ *the* i*-th feature*

$$\textbf{(a)} \quad \|\mathbf{x}_i - \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \|\mathbf{x}_i + \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j$$

Exact Grouping: Logistic Loss

Theorem (F and Nowak (2016))

Let $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$ *and* $\mathbf{x}_i = [x_{(1),i}, ..., x_{(n),i}]^T$ *the* i*-th feature*

$$\textbf{(a)} \quad \|\mathbf{x}_i - \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \|\mathbf{x}_i + \mathbf{x}_j\|_1 < \Delta_{\mathbf{w}} \quad \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j$$

Corollary

Let $\mathbf{1}^T\mathbf{x}_k = 0$ *and* $\|\mathbf{x}_k\|_2 = 1$, *for* $k = 1, ..., p$. *Denote* $\rho_{ij} = \mathbf{x}_i^T\mathbf{x}_j \in [-1, 1]$ *the sample correlation. Then,*

$$\textbf{(a)} \quad \sqrt{2 - 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\sqrt{n} \quad \Rightarrow \quad \widehat{\beta}_i = \widehat{\beta}_j$$

$$\textbf{(b)} \quad \sqrt{2 + 2\,\rho_{ij}} < \Delta_{\mathbf{w}}/\sqrt{n} \quad \Rightarrow \quad \widehat{\beta}_i = -\widehat{\beta}_j.$$

Exact Grouping: Remarks

- Finding all pairs of sufficiently correlated features explicitly: $O(n\,p^2)$
- OWL costs $O(n\,p\log p)$ (shown later); much better, for large p
- OWL uses the responses $\mathbf{y}$, not just the covariates/variables

Statistical Bounds

Scenario and assumptions (F and Nowak, 2016)

Statistical Bounds

Scenario and assumptions (F and Nowak, 2016)

- Observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^\star$ is s-sparse

Statistical Bounds

Scenario and assumptions (F and Nowak, 2016)

- Observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^{\star}$ is s-sparse
- Noise: $\frac{1}{n}\|\boldsymbol{\varepsilon}\|_1 \leq \epsilon$

Statistical Bounds

Scenario and assumptions (F and Nowak, 2016)

- Observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^\star$ is s-sparse
- Noise: $\frac{1}{n}\|\boldsymbol{\varepsilon}\|_1 \leq \epsilon$
- Random design: the rows of $\mathbf{X} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(0, \mathbf{C}^T\mathbf{C})$

Statistical Bounds

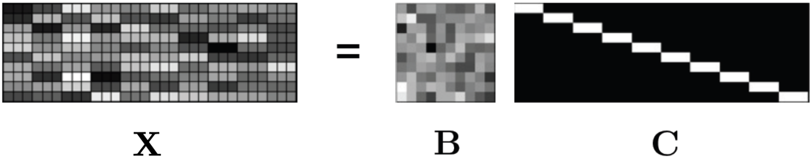
Scenario and assumptions (F and Nowak, 2016)

- Observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^\star$ is s-sparse
- Noise: $\frac{1}{n}\|\boldsymbol{\varepsilon}\|_1 \leq \epsilon$
- Random design: the rows of $\mathbf{X} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(0, \mathbf{C}^T\mathbf{C})$
- ...equivalently, $\mathbf{X} = \mathbf{B}\mathbf{C}$ where the rows of $\mathbf{B} \in \mathbb{R}^{n\times q}$ are i.i.d. $\mathcal{N}(0, \mathbf{I})$, and $\mathbf{C} \in \mathbb{R}^{q\times p}$

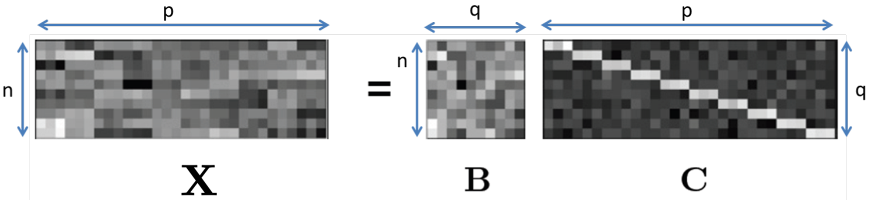
Statistical Bounds

Scenario and assumptions (F and Nowak, 2016)

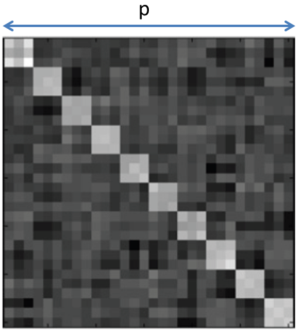
- Observations: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^\star$ is s-sparse
- Noise: $\frac{1}{n}\|\boldsymbol{\varepsilon}\|_1 \leq \epsilon$
- Random design: the rows of $\mathbf{X} \in \mathbb{R}^{n\times p}$ are i.i.d. $\mathcal{N}(0, \mathbf{C}^T\mathbf{C})$
- ...equivalently, $\mathbf{X} = \mathbf{B}\mathbf{C}$ where the rows of $\mathbf{B} \in \mathbb{R}^{n\times q}$ are i.i.d. $\mathcal{N}(0, \mathbf{I})$, and $\mathbf{C} \in \mathbb{R}^{q\times p}$
- Illustration:



Another Illustration: Highly Correlated Groups of Columns



$$\mathbf{C}^T\mathbf{C} =$$



Statistical Bound

Theorem (F and Nowak (2016))

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

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

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

Then

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^\star\|_{\mathbf{C}^T\mathbf{C}}\big] \leq \sqrt{8\pi}\left(\sqrt{32}\,\big(\min_{r=1,2}\|\mathbf{C}\|_r\big)\,\|\boldsymbol{\beta}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}}+\epsilon\right)$$

where $\|\mathbf{C}\|_r$ *is the matrix norm induced by the* ℓ_r *norm.*

Statistical Bound

Theorem (F and Nowak (2016))

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

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

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

Then

$$\mathbb{E}\big[\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_{\mathbf{C}^T\mathbf{C}}\big] \leq \sqrt{8\pi}\left(\sqrt{32}\,\big(\min_{r=1,2}\|\mathbf{C}\|_r\big)\,\|\boldsymbol{\beta}^\star\|_2\,\frac{w_1}{\bar{w}}\sqrt{\frac{s\log p}{n}} + \epsilon\right)$$

where $\|\mathbf{C}\|_r$ *is the matrix norm induced by the* ℓ_r *norm.*

The proof is based on work by Vershynin (2014)

Statistical Bound: Observations

- Particular case: $q = p$ and $\mathbf{C} = \mathbf{I}$ (as in compressive sensing)

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\|\boldsymbol{\beta}^\star\|_2 \sqrt{\frac{s \log p}{n}}\Big),$$

Statistical Bound: Observations

- Particular case: $q = p$ and $\mathbf{C} = \mathbf{I}$ (as in compressive sensing)

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\|\boldsymbol{\beta}^\star\|_2 \sqrt{\frac{s \log p}{n}}\Big),$$

recovering classical results that show that

$$n \sim s \log p$$

samples are needed to achieve a given precision
(Bühlmann and van de Geer, 2011; Candès et al., 2006; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)

Statistical Bound: Observations

- Particular case: $q = p$ and $\mathbf{C} = \mathbf{I}$ (as in compressive sensing)

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\|\boldsymbol{\beta}^\star\|_2 \sqrt{\frac{s \log p}{n}}\Big),$$

  recovering classical results that show that

$$n \sim s \log p$$

  samples are needed to achieve a given precision
  (Bühlmann and van de Geer, 2011; Candès et al., 2006; Donoho, 2006; Haupt and Nowak, 2006; Vershynin, 2014)

- Other analyses of OWL/SLOPE, not focusing on the highly-correlated case, by Hu and Lu (2019); Stucky and van de Geer (2017); Wang et al (2019)

Statistical Bound: Observations

- Particular case: correlated full-rank design matrix,

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\frac{\lambda_{\max}}{\lambda_{\min}} \|\boldsymbol{\beta}^\star\|_2 \sqrt{\frac{s \log p}{n}}\Big),$$

recovering results by Raskutti et al. (2010)

Statistical Bound: Observations

- Particular case: correlated full-rank design matrix,

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\frac{\lambda_{\max}}{\lambda_{\min}}\,\|\boldsymbol{\beta}^\star\|_2\,\sqrt{\frac{s\log p}{n}}\Big),$$

  recovering results by Raskutti et al. (2010)

- This result does not apply in the presence of identical columns

Statistical Bound: Observations

- Particular case: correlated full-rank design matrix,

$$\mathbb{E}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2 = O\Big(\frac{\lambda_{\max}}{\lambda_{\min}}\,\|\boldsymbol{\beta}^\star\|_2\,\sqrt{\frac{s\log p}{n}}\Big),$$

  recovering results by Raskutti et al. (2010)

- This result does not apply in the presence of identical columns

- Notice that these results also apply to ℓ_1

Statistical Bound: Observations

- Particular case: groups of replicated columns (here, q is the number of groups of replicated columns)

Statistical Bound: Observations

- Particular case: groups of replicated columns (here, q is the number of groups of replicated columns)

- Illustration with replicated columns, $q = 3$, $p = 4$, and

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad G_1 = \{1, 2\}, \;\; G_2 = \{3\}, \;\; G_3 = \{4\}$$

Statistical Bound: Observations

- Particular case: groups of replicated columns (here, q is the number of groups of replicated columns)

- Illustration with replicated columns, $q = 3$, $p = 4$, and

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad G_1 = \{1, 2\}, \;\; G_2 = \{3\}, \;\; G_3 = \{4\}$$

- Let $\bar{\boldsymbol{\beta}}^{\star}$ be the vector that has identical coefficients corresponding to identical columns (notice $\mathbf{X}\bar{\boldsymbol{\beta}}^{\star} = \mathbf{X}\boldsymbol{\beta}^{\star}$)

Statistical Bound: Observations

- Particular case: groups of replicated columns (here, q is the number of groups of replicated columns)
- Illustration with replicated columns, $q = 3$, $p = 4$, and

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad G_1 = \{1, 2\}, \;\; G_2 = \{3\}, \;\; G_3 = \{4\}$$

- Let $\bar{\boldsymbol{\beta}}^\star$ be the vector that has identical coefficients corresponding to identical columns (notice $\mathbf{X}\bar{\boldsymbol{\beta}}^\star = \mathbf{X}\boldsymbol{\beta}^\star$)
- In this case,

$$\mathbb{E}\,\|\widehat{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^\star\|_2 = O\Big(\|\boldsymbol{\beta}^\star\|_2\,\frac{w_1}{\bar{w}}\,\sqrt{\frac{s\log p}{n}} + \epsilon\Big)$$

i.e. no penalty for the presence of (nearly) replicated columns

Outline

Regularization Formulations

- Tikhonov regularization (**OWL-T**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \tau\, \Omega_{\mathbf{w}}(\boldsymbol{\beta}),$$

Regularization Formulations

- Tikhonov regularization (**OWL-T**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \tau \, \Omega_{\mathbf{w}}(\boldsymbol{\beta}),$$

- Ivanov regularization (**OWL-I**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \quad \text{s.t. } \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \varepsilon,$$

Regularization Formulations

- Tikhonov regularization (**OWL-T**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \tau\, \Omega_{\mathbf{w}}(\boldsymbol{\beta}),$$

- Ivanov regularization (**OWL-I**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \quad \text{s.t. } \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \varepsilon,$$

...under mild conditions, all equivalent.

Regularization Formulations

- Tikhonov regularization (**OWL-T**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \tau\, \Omega_{\mathbf{w}}(\boldsymbol{\beta}),$$

- Ivanov regularization (**OWL-I**)

$$\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}), \quad \text{s.t. } \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \varepsilon,$$

...under mild conditions, all equivalent.

Suggest different optimization methods:

- Proximal gradient, for OWL-T
- Projected gradient or Frank-Wolfe, for OWL-I

Proximal Gradient Algorithm (PGA)

The Tikhonov formulation: $\quad \min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

Proximal Gradient Algorithm (PGA)

The Tikhonov formulation: $\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- Proximal gradient algorithm: $\boldsymbol{\beta}_{t+1} = \text{prox}_{\tau\Omega_{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau\nabla L(\boldsymbol{\beta}_t)\Big)$

Proximal Gradient Algorithm (PGA)

The Tikhonov formulation: $\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- Proximal gradient algorithm: $\boldsymbol{\beta}_{t+1} = \text{prox}_{\tau\Omega_{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau\nabla L(\boldsymbol{\beta}_t)\Big)$
- Needs the proximity operator (Moreau, 1962) of $\Omega_{\mathbf{w}}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2}\,\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$$

(...next slide)

Proximal Gradient Algorithm (PGA)

The Tikhonov formulation: $\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- Proximal gradient algorithm: $\boldsymbol{\beta}_{t+1} = \operatorname{prox}_{\tau\Omega_{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau\nabla L(\boldsymbol{\beta}_t)\Big)$
- Needs the proximity operator (Moreau, 1962) of $\Omega_{\mathbf{w}}$

$$\operatorname{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2}\,\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$$

(...next slide)

- Needs the gradient of $L(\boldsymbol{\beta})$: simple for linear and logistic regression, and others losses

Proximal Gradient Algorithm (PGA)

The Tikhonov formulation: $\min_{\boldsymbol{\beta}} L(\boldsymbol{\beta}) + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- Proximal gradient algorithm: $\boldsymbol{\beta}_{t+1} = \text{prox}_{\tau\Omega_{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau\nabla L(\boldsymbol{\beta}_t)\Big)$
- Needs the proximity operator (Moreau, 1962) of $\Omega_{\mathbf{w}}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \tfrac{1}{2}\,\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$$

(...next slide)

- Needs the gradient of $L(\boldsymbol{\beta})$: simple for linear and logistic regression, and others losses
- Accelerated versions: FISTA (Beck and Teboulle, 2009); TwIST (Bioucas-Dias and F, 2007); SpaRSA (Wright, Nowak, and F, 2009);

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad\qquad |\mathbf{u}|_\downarrow$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\mathrm{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \qquad\qquad (|\mathbf{u}|_{\downarrow} - \mathbf{w})$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\mathrm{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad\qquad \mathrm{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad \text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad \text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\,(|\mathbf{u}|_\downarrow - \mathbf{w})\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad \mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\big(\, \text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \,\text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)\big)\Big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\big(\text{proj}_{\mathcal{K}_m}(|\mathbf{u}|_{\downarrow} - \mathbf{w})\big)\Big)$$

- $\text{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA): cost $O(p)$ (cf. isotonic regression (Barlow et al., 1972; Best and Chakravarti, 1990))

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\big(\text{proj}_{\mathcal{K}_m}(|\mathbf{u}|_{\downarrow} - \mathbf{w})\big)\Big)$$

- $\text{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA): cost $O(p)$ (cf. isotonic regression (Barlow et al., 1972; Best and Chakravarti, 1990))
- Computational cost: $O(p \log p)$, due to sorting; all else is $O(p)$.

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+} \big(\text{proj}_{\mathcal{K}_m} \left(|\mathbf{u}|_\downarrow - \mathbf{w} \right) \big) \Big)$$

- $\text{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA): cost $O(p)$ (cf. isotonic regression (Barlow et al., 1972; Best and Chakravarti, 1990))
- Computational cost: $O(p \log p)$, due to sorting; all else is $O(p)$.
- Compact expression for the algorithms by Bogdan et al. (2014); Zeng and F (2014b); Zhong and Kwok (2012)

Projected Gradient Algorithm

Ivanov formulation: $\min_{\boldsymbol{\beta} \in \mathcal{G}_\epsilon^{\mathbf{w}}} L(\boldsymbol{\beta}),$ where $\mathcal{G}_\epsilon^{\mathbf{w}} = \{\boldsymbol{\beta} : \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \epsilon\}$

- Projected gradient algorithm: $\boldsymbol{\beta}_{t+1} = \text{proj}_{\mathcal{G}_\epsilon^{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau_t \nabla L(\boldsymbol{\beta}_t)\Big)$

Projected Gradient Algorithm

Ivanov formulation: $\min_{\boldsymbol{\beta} \in \mathcal{G}_\epsilon^{\mathbf{w}}} L(\boldsymbol{\beta}), \quad$ where $\mathcal{G}_\epsilon^{\mathbf{w}} = \{\boldsymbol{\beta} : \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \epsilon\}$

- Projected gradient algorithm: $\boldsymbol{\beta}_{t+1} = \mathrm{proj}_{\mathcal{G}_\epsilon^{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau_t \nabla L(\boldsymbol{\beta}_t)\Big)$
- Gradient: simple for linear and logistic regressions, and other losses

Projected Gradient Algorithm

Ivanov formulation: $\min_{\boldsymbol{\beta} \in \mathcal{G}_\epsilon^{\mathbf{w}}} L(\boldsymbol{\beta}),$ where $\mathcal{G}_\epsilon^{\mathbf{w}} = \{\boldsymbol{\beta} : \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \epsilon\}$

- Projected gradient algorithm: $\boldsymbol{\beta}_{t+1} = \text{proj}_{\mathcal{G}_\epsilon^{\mathbf{w}}}\Big(\boldsymbol{\beta}_t - \tau_t \nabla L(\boldsymbol{\beta}_t)\Big)$
- Gradient: simple for linear and logistic regressions, and other losses
- Euclidean projection operator

$$\text{proj}_{\mathcal{G}_\varepsilon}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta} \in \mathcal{G}_\epsilon^{\mathbf{w}}} \|\mathbf{u} - \boldsymbol{\beta}\|_2^2$$

can also be computed with $O(p \log p)$ cost (Davis, 2015)

Atomic Norm Formulation

Atomic norms (Chandrasekaran, 2012; Jaggi, 2013)

Atomic Norm Formulation

Atomic norms (Chandrasekaran, 2012; Jaggi, 2013)

- A compact, centrally symmetric set $\mathcal{A} \subset \mathbb{R}^p$ induces an atomic norm:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t \geq 0 : \mathbf{x} \in t\, \mathsf{conv}(\mathcal{A})\}$$

...the Minkowski gauge of conv$(\mathcal{A})$ (Rockafellar, 1970)

Atomic Norm Formulation

Atomic norms (Chandrasekaran, 2012; Jaggi, 2013)

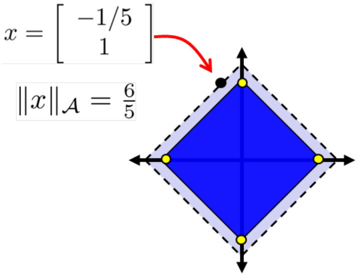
- A compact, centrally symmetric set $\mathcal{A} \subset \mathbb{R}^p$ induces an atomic norm:

$$\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t \geq 0 : \mathbf{x} \in t\,\text{conv}(\mathcal{A})\}$$

...the Minkowski gauge of conv$(\mathcal{A})$ (Rockafellar, 1970)

Example: the ℓ_1 norm as an atomic norm

- $\mathcal{A} = \left\{ \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix} \right\}$
- $\text{conv}(\mathcal{A}) = B_1(1)$ (ℓ_1 unit ball).
- $\|x\|_{\mathcal{A}} = \inf\{t > 0 : \; x \in t\; B_1(1)\}$
  $= \|x\|_1$



Atomic Formulation of the OWL Norm

OWL: $\Omega_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$. Atomic: $\|\mathbf{x}\|_{\mathcal{A}} = \inf \{t \geq 0 : \mathbf{x} \in t \, \mathrm{conv}(\mathcal{A})\}$

Proposition (Zeng and F (2015))

Let $\mathcal{B} = \left\{\mathbf{b}^{(1)}, ..., \mathbf{b}^{(i)}, ..., \mathbf{b}^{(p)}\right\}$, *where*

$$\mathbf{b}^{(i)} = \begin{bmatrix} \tau_i \\ \vdots \\ \tau_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}\tau_i\\ \vdots\\ \tau_i\end{matrix}}\right\} i \text{ entries} \\ \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} (p-i) \text{ entries} \end{matrix} \qquad \text{with } \tau_i = \frac{1}{w_1 + \cdots + w_i}$$

Atomic Formulation of the OWL Norm

OWL: $\Omega_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T |\mathbf{x}|_{\downarrow}$. Atomic: $\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t \geq 0 : \mathbf{x} \in t\, \text{conv}(\mathcal{A})\}$

Proposition (Zeng and F (2015))

Let $\mathcal{B} = \left\{\mathbf{b}^{(1)}, ..., \mathbf{b}^{(i)}, ..., \mathbf{b}^{(p)}\right\}$, *where*

$$\mathbf{b}^{(i)} = \begin{bmatrix} \tau_i \\ \vdots \\ \tau_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}\tau_i\\ \vdots\\ \tau_i\end{matrix}}\right\} i \text{ entries} \\ \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} (p-i) \text{ entries} \end{matrix} \qquad \text{with } \tau_i = \frac{1}{w_1 + \cdots + w_i}$$

and $\mathcal{A} = \{\mathbf{Q}\,\mathbf{b} : \mathbf{Q} \in \mathcal{P}_{\pm}, \mathbf{b} \in \mathcal{B}\}$.

$\mathcal{P}_{\pm}$ is the *hyperoctahedral group* (signed permutation matrices).

Atomic Formulation of the OWL Norm

OWL: $\Omega_{\mathbf{w}}(\mathbf{x}) = \mathbf{w}^T|\mathbf{x}|_\downarrow$. Atomic: $\|\mathbf{x}\|_{\mathcal{A}} = \inf\{t \geq 0 : \mathbf{x} \in t\,\text{conv}(\mathcal{A})\}$

Proposition (Zeng and F (2015))

Let $\mathcal{B} = \left\{\mathbf{b}^{(1)}, ..., \mathbf{b}^{(i)}, ..., \mathbf{b}^{(p)}\right\}$, *where*

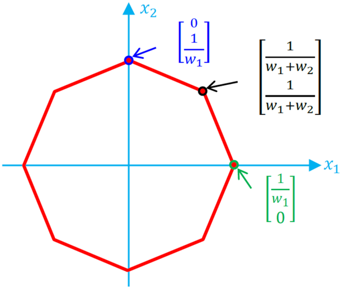
$$\mathbf{b}^{(i)} = \begin{bmatrix} \tau_i \\ \vdots \\ \tau_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}\tau_i\\ \vdots\\ \tau_i\end{matrix}}\right\} i \text{ entries} \\ \left.\vphantom{\begin{matrix}0\\ \vdots\\ 0\end{matrix}}\right\} (p-i) \text{ entries} \end{matrix} \qquad \text{with } \tau_i = \frac{1}{w_1 + \cdots + w_i}$$

and $\mathcal{A} = \{\mathbf{Q}\,\mathbf{b} : \ \mathbf{Q} \in \mathcal{P}_\pm, \mathbf{b} \in \mathcal{B}\}$. *Then, for any* $\mathbf{x} \in \mathbb{R}^p$, $\|\mathbf{x}\|_{\mathcal{A}} = \Omega_{\mathbf{w}}(\mathbf{x})$.

$\mathcal{P}_\pm$ is the *hyperoctahedral group* (signed permutation matrices).

Atomic Formulation of the OWL Norm

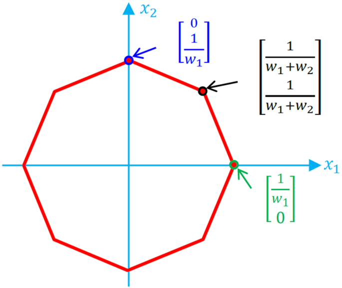
Illustration in $\mathbb{R}^2$ and $\mathbb{R}^3$

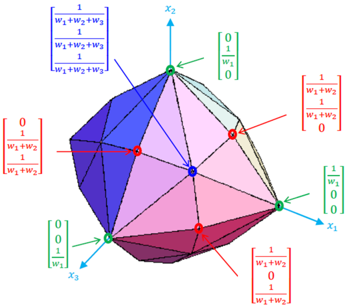


(Zeng and F, 2015)

Atomic Formulation of the OWL Norm

Illustration in $\mathbb{R}^2$ and $\mathbb{R}^3$





Cardinality: $|\mathcal{A}| = \sum_{i=1}^{p} \binom{n}{i} 2^i = 3^p - 1$

(Zeng and F, 2015)

Frank-Wolfe (Conditional Gradient) Algorithm

Ivanov formulation: $\min_{\boldsymbol{\beta} \in \mathcal{G}_\epsilon^{\mathbf{w}}} L(\boldsymbol{\beta}),$ where $\mathcal{G}_\epsilon^{\mathbf{w}} = \{\boldsymbol{\beta} : \Omega_{\mathbf{w}}(\boldsymbol{\beta}) \leq \epsilon\}$

- Frank-Wolfe algorithm:

$$\mathbf{s}_t = \arg\max_{\mathbf{s} \in \mathcal{G}_\epsilon^{\mathbf{w}}} \mathbf{s}^T \left(-\nabla L(\boldsymbol{\beta}_t)\right) \quad (1)$$

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \tau_k (\mathbf{s}_t - \boldsymbol{\beta}_t) \quad (2)$$

- Gradient: simple for linear and logistic regressions, and other losses

- Linear minimization oracle (Zeng and F, 2015)

$$\arg\max_{\mathbf{s} \in \mathcal{G}_\epsilon} \mathbf{s}^T \mathbf{u} = \epsilon \,\mathsf{sign}(\mathbf{u}) \odot \left(\mathbf{P}(|\mathbf{u}|)^T \arg\max_{\mathbf{b} \in \mathcal{B}} \mathbf{b}^T |\mathbf{u}|_{\downarrow}\right)$$

...also $O(p \log p)$, due to sorting

Outline

Group-OWL: GrOWL

- Matrix regression problem (m simultaneous linear regressions)

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + R(\mathbf{B})$$

Group-OWL: GrOWL

- Matrix regression problem (m simultaneous linear regressions)

$$\widehat{\mathbf{B}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + R(\mathbf{B})$$

- Group-LASSO: $$R(\mathbf{B}) = \|\mathbf{B}\|_{2,1} = \sum_{j=1}^{p} \|\boldsymbol{\beta}_{j:}\|_2$$

Group-OWL: GrOWL

- Matrix regression problem (m simultaneous linear regressions)

$$\widehat{\mathbf{B}} = \arg \min_{\mathbf{B} \in \mathbb{R}^{p \times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + R(\mathbf{B})$$

- Group-LASSO: $R(\mathbf{B}) = \|\mathbf{B}\|_{2,1} = \sum_{j=1}^{p} \|\boldsymbol{\beta}_{j:}\|_2$

- GrOWL (Oswal et al, 2016): $R(\mathbf{B}) = \Omega_{\mathbf{w}}(\mathbf{B}) = \sum_{j=1}^{p} w_j \|\boldsymbol{\beta}_{[j]:}\|_2$

$\|\boldsymbol{\beta}_{[1]:}\|_2 \geq \|\boldsymbol{\beta}_{[2]:}\|_2 \geq \cdots \geq \|\boldsymbol{\beta}_{[p]:}\|_2$

Group-OWL: GrOWL

- Matrix regression problem (m simultaneous linear regressions)

$$\widehat{\mathbf{B}} = \arg \min_{\mathbf{B} \in \mathbb{R}^{p \times m}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_F^2 + R(\mathbf{B})$$

- Group-LASSO: $R(\mathbf{B}) = \|\mathbf{B}\|_{2,1} = \sum_{j=1}^{p} \|\boldsymbol{\beta}_{j:}\|_2$

- GrOWL (Oswal et al, 2016): $R(\mathbf{B}) = \Omega_{\mathbf{w}}(\mathbf{B}) = \sum_{j=1}^{p} w_j \|\boldsymbol{\beta}_{[j]:}\|_2$

  $\|\boldsymbol{\beta}_{[1]:}\|_2 \geq \|\boldsymbol{\beta}_{[2]:}\|_2 \geq \cdots \geq \|\boldsymbol{\beta}_{[p]:}\|_2$

- Also a norm; proximity operator still efficiently computable; similar exact grouping guarantees

Group OWL: GrOWL

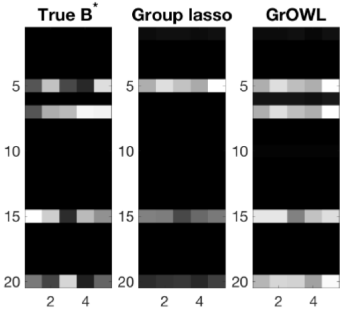


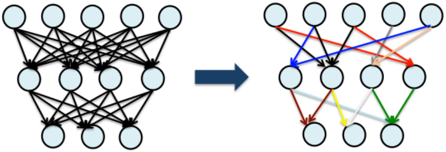
Figure 2. A comparison of group lasso (middle) and grOWL (right) optimization solutions with correlated columns in $\boldsymbol{X}$ showing that GrOWL selects relevant features (row 5 and 7) even if they happen to be strongly correlated and automatically cluster them by setting the corresponding coefficient rows to be equal

(Oswal et al, 2016)

Application to Deep Learning

With Laura Balzano and Haozhu Wang (U Michigan), and Dejiao Zhang (Amazon)

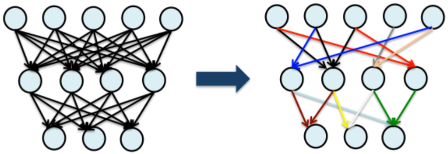
- Weigh sharing is good for DNN: can we learn it using OWL/GrOWL?
- OWL for adaptive weight sharing in deep networks (Zhang et al, 2018)

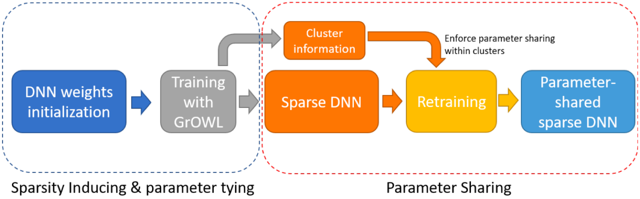


Application to Deep Learning

With Laura Balzano and Haozhu Wang (U Michigan), and Dejiao Zhang (Amazon)

- Weigh sharing is good for DNN: can we learn it using OWL/GrOWL?
- OWL for adaptive weight sharing in deep networks (Zhang et al, 2018)





Application to Deep Learning

Table 1: Sparsity, parameter sharing, and compression rate results on MNIST. Baseline model is trained with weight decay and we do not enforce parameter sharing for baseline model. We train each model for 5 times and report the average values together with their standard deviations.

Regularizer	**Sparsity**	**Parameter Sharing**	**Compression ratio**	**Accuracy**
none	$0.0 \pm 0\%$	1.0 ± 0	1.0 ± 0	$98.3 \pm 0.1\%$
weight decay	$0.0 \pm 0\%$	1.6 ± 0	1.6 ± 0	$98.4 \pm 0.0\%$
group-Lasso	$87.6 \pm 0.1\%$	1.9 ± 0.1	15.8 ± 1.0	$98.1 \pm 0.1\%$
group-Lasso+ℓ_2	$93.2 \pm 0.4\%$	1.6 ± 0.1	23.7 ± 2.1	$98.0 \pm 0.1\%$
GrOWL	$80.4 \pm 1.0\%$	3.2 ± 0.1	16.7 ± 1.3	$98.1 \pm 0.1\%$
GrOWL+ℓ_2	$83.6 \pm 0.5\%$	3.9 ± 0.1	24.1 ± 0.8	$98.1 \pm 0.1\%$

$$\text{Sparsity} = \frac{\#\text{ zeros}}{\#\text{ total}} \qquad \text{Sharing} = \frac{\#\text{ non-zeros}}{\#\text{ unique vaues}} \qquad \text{Compression} = \frac{\#\text{ total}}{\#\text{ unique vaues}}$$

(Zhang et al, 2018)

(Sparse) Subspace Clustering

- Grouping high-dimensional data points into distinct subspaces

(Sparse) Subspace Clustering

- Grouping high-dimensional data points into distinct subspaces
- Sparse subspace clustering (SSC) (Elhamifar et al, 2013)

(Sparse) Subspace Clustering

- Grouping high-dimensional data points into distinct subspaces
- Sparse subspace clustering (SSC) (Elhamifar et al, 2013)
  - ◇ N points in $\mathbb{R}^d$: $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$
  - ◇ Do sparse regression (use LASSO) of each point w.r.t. "all" the others

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{x}^{(n)} - \mathbf{X}_{\bar{n}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

where $\mathbf{X}_{\bar{n}} \in \mathbb{R}^{d\times(N-1)}$ contains the points other than $\mathbf{x}^{(n)}$

(Sparse) Subspace Clustering

- Grouping high-dimensional data points into distinct subspaces
- Sparse subspace clustering (SSC) (Elhamifar et al, 2013)
  - ◇ N points in $\mathbb{R}^d$: $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$
  - ◇ Do sparse regression (use LASSO) of each point w.r.t. "all" the others

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{x}^{(n)} - \mathbf{X}_{\bar{n}} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

    where $\mathbf{X}_{\bar{n}} \in \mathbb{R}^{d \times (N-1)}$ contains the points other than $\mathbf{x}^{(n)}$

  - ◇ Rationale: points on the same subspace are likely to be selected

(Sparse) Subspace Clustering

- Grouping high-dimensional data points into distinct subspaces
- Sparse subspace clustering (SSC) (Elhamifar et al, 2013)
  - ◇ N points in $\mathbb{R}^d$: $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$
  - ◇ Do sparse regression (use LASSO) of each point w.r.t. "all" the others

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{2} \|\mathbf{x}^{(n)} - \mathbf{X}_{\bar{n}} \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

where $\mathbf{X}_{\bar{n}} \in \mathbb{R}^{d \times (N-1)}$ contains the points other than $\mathbf{x}^{(n)}$

  - ◇ Rationale: points on the same subspace are likely to be selected
  - ◇ Selected points define an adjacency graph: apply clustering

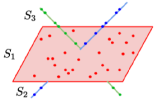
(Sparse) Subspace Clustering

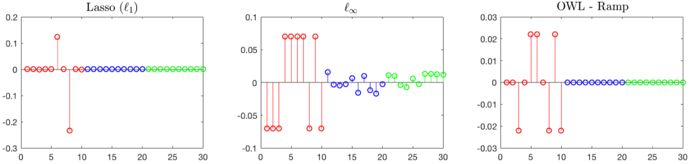
- Grouping high-dimensional data points into distinct subspaces
- Sparse subspace clustering (SSC) (Elhamifar et al, 2013)
  - ◇ N points in $\mathbb{R}^d$: $\{\mathbf{x}^{(1)}, ..., \mathbf{x}^{(N)}\}$
  - ◇ Do sparse regression (use LASSO) of each point w.r.t. "all" the others

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{x}^{(n)} - \mathbf{X}_{\bar{n}}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

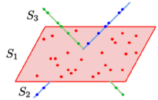
    where $\mathbf{X}_{\bar{n}} \in \mathbb{R}^{d\times(N-1)}$ contains the points other than $\mathbf{x}^{(n)}$
  - ◇ Rationale: points on the same subspace are likely to be selected
  - ◇ Selected points define an adjacency graph: apply clustering
- New approach: replace LASSO by OWL (Oswal et al, 2018)

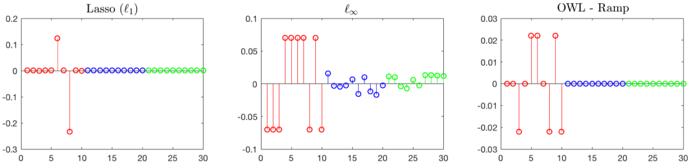
OWL Subspace Clustering: Illustration (Oswal et al, 2018)

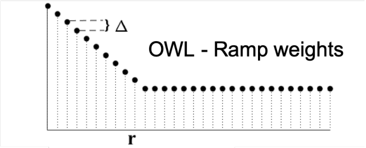




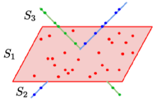
OWL Subspace Clustering: Illustration (Oswal et al, 2018)

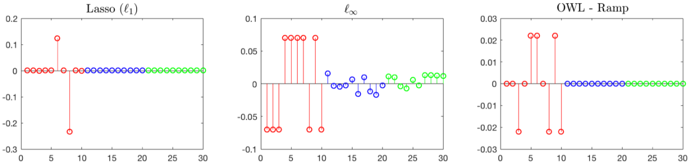






OWL Subspace Clustering: Illustration (Oswal et al, 2018)





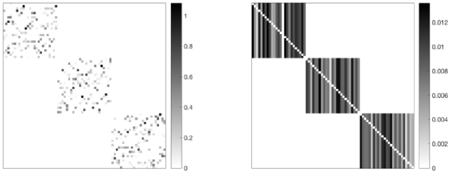
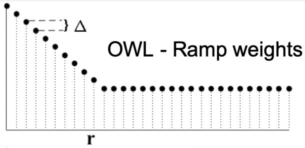


Fig. 3: Examples of coefficient matrices $|B| = [|\widehat{\beta}_1| \dots |\widehat{\beta}_N|]$ for exact ℓ_1 minimizations (left) and OWL optimizations (right) with the contiguous columns lying in three orthogonal subspaces each of dimension $d = 5$ in $\mathbb{R}^{15}$. The plots were generated using OWL-Ramp weights defined in Section III.



Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm

Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm
- Exact clustering sufficient conditions
- Statistical sample complexity bounds

Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm
- Exact clustering sufficient conditions
- Statistical sample complexity bounds
- The proximity operator of OWL (proximal gradient)
- The Euclidean projection onto OWL ball (projected gradient)
- Linear minimization oracle of the OWL ball (Frank-Wolfe)

Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm
- Exact clustering sufficient conditions
- Statistical sample complexity bounds
- The proximity operator of OWL (proximal gradient)
- The Euclidean projection onto OWL ball (projected gradient)
- Linear minimization oracle of the OWL ball (Frank-Wolfe)
- ...all with $O(p \log p)$ cost

Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm
- Exact clustering sufficient conditions
- Statistical sample complexity bounds
- The proximity operator of OWL (proximal gradient)
- The Euclidean projection onto OWL ball (projected gradient)
- Linear minimization oracle of the OWL ball (Frank-Wolfe)
- ...all with $O(p \log p)$ cost
- Extensions, applications, ...

Final Summary

- **OSCAR**: a regularizer able to identify groups of correlated variables
- OSCAR is a particular case of the **OWL** norm
- Exact clustering sufficient conditions
- Statistical sample complexity bounds
- The proximity operator of OWL (proximal gradient)
- The Euclidean projection onto OWL ball (projected gradient)
- Linear minimization oracle of the OWL ball (Frank-Wolfe)
- ...all with $O(p \log p)$ cost
- Extensions, applications, ...
- Key open question: how to choose $\mathbf{w}$?

Thank you.

References I

Arnold, B. (1987). *Majorization and the Lorenz Order: A Brief Introduction*, volume 43. Springer Verlag: Lecture Notes in Statistics, Berlin.

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

Bach, F. (2012). Consistency of the group-Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University.

Barlow, R., Bartholomew, D., Bremand, J., and Brunk, H. (1972). *Statistical inference under order restrictions; the theory and application of isotonic regression*. Wiley, New York.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2:183–202.

Best, M. and Chakravarti, N. (1990). Active set algorithms for isotonic regression; a unifying framework. *Mathematical Programming*, 47:425–439.

Bioucas-Dias, J. and Figueiredo, M. (2007). A new twist: two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16:2992–3004.

Bogdan, J., Berg, E., Su, W., and Candes, E. (2014). Statistical estimation and testing via the ordered ℓ_1 norm. *arXiv preprint http://arxiv.org/pdf/1310.1969v1.pdf*.

Bolstad, A., Veen, B. V., and Nowak, R. (2009). Space-time event sparse penalization for magnetoelectroencephalography. *NeuroImage*, 46:1066–1081.

References II

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.

Bühlmann, P., van de Geer, S., (2011). *Statistics for High-Dimensional Data,* Springer.

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.

Cessie, S. L. and Houwelingen, J. C. V. (1992). Ridge estimators in logistic regression. *Journal of the Royal Statistical Society; Series C*, 41:191–201.

Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12: 805–849.

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

Claerbout, J. and Muir, F. (1973). Robust modelling of erratic data. *Geophysics*, 38:826–844.

Clark, J. (2015). *Locally Non-Linear Learning via Feature Induction and Structured Regularization in Statistical Machine Translation*. PhD thesis, Language Technologies Institute, Carnegie Mellon University.

References III

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

Das, S. (1994). Filters, wrappers and a boosting-based hybrid for feature selection. *International Conference on Machine Learning*, pp. 74?-81.

Davis, D. (2015). An o(nlog(n)) algorithm for projecting onto the ordered weighted ℓ_1 norm ball. Technical report, arXiv:1505.00870.

Daye, Z. and Jeng, X. (2009). Shrinkage and model selection with correlated variables via weighted fusion. *Computational Statistics & Data Analysis*, 53(4):1284-?1298.

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.

Escolano, F., Suau, P., Bonev, B. (2009). *Information Theory in Computer Vision and Pattern Recognition*, Springer.

Ferreira, A., Figueiredo, M. (2012). Efficient feature selection filters for high dimensional data. *Pattern Recognition Letters*, 33: 1794–1804.

Figueiredo, M. and Nowak, R. (2016). Ordered weighted l1 regularized regression with strongly correlated covariates: Theoretical aspects. In *19th International Conference on Artificial Intelligence and Statistics*, Cadiz, Spain.

Galton, F. (1934). *Natural Inheritance*. Macmillan and Company, New York.

References IV

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hardy, G., Littlewood, J., and Pólya, G. (1934). *Inequalities*. Cambridge University Press.

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

Hebiri, M. and van de Geer, S. (2011). The smooth-lasso and other $\ell_1 + \ell_2$-penalized methods. *Electronic Journal of Statistics*, 5:1184?-1226.

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

Hu, H. and Lu, Y. (2019). Asymptotics and optimal designs of SLOPE for sparse linear regression. `https://arxiv.org/abs/1903.11582`

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

Jaggi, M. (2013). Revisiting Frank-Wolfe: projection-free sparse convex optimization, *Proceedings of the 30th International Conference on Machine Learning*, pp. 427–435.

Kazama, J. and Tsujii, J. (2003). Evaluation and extension of maximum entropy models with inequality constraints. *Proceedings of EMNLP*.

Kim, S. and Xing, E. (2009). Statistical estimation of correlated genome associations to a quantitative trait network. *PLoS Genetics*, 5:e1000587.

References V

Krishnapuram, B., Carin, L., Figueiredo, M., and Hartemink, A. (2005). Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27:957–968.

Levy, S. and Fullagar, P. (1981). Reconstruction of a sparse spike train from a portion of its spectrum and application to high-resolution deconvolution. *Geophysics*, 46:1235–1243.

Li, Y., Mark, B., Raskutti, G., and Willett, R. (2018). Graph-based regularization for regression problems with highly-correlated designs. Technical report, arXiv:1803.07658v2.

Lorenz, M. (1905). Methods of measuring concentration of wealth. *Journal of the American Statistical Association*, 9:209–219.

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

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

Moreau, J. (1962). Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899.

Negahban, S., Ravikumar, P., Wainwright, M., and Yu, B. (2012). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27:538–557.

Németh, A. and Németh, S. (2012). How to project on the monotone nonnegative cone using the pool adjacent violators type algorithms. `http://arxiv.org/pdf/1201.2343v2.pdf`.

References VI

Nowlan, S. and Hinton, G. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4: 473–493.

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.

Oswal, U., Cox, C., Ralph, M., Rogers, T., Nowak, R. (2016) Representational similarity learning with application to brain networks. *International Conference on Machine Learning.*

Oswal, U., Nowak, R. (2018). Scalable sparse subspace clustering via ordered weighted ℓ_1 regression *Annual Allerton Conference on Communication, Control, and Computing.*

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences.* 187: 253?318.

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

Raskutti, G., Wainwright, M., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259.

Rockafellar, R.T. (1970) *Convex Analysis*. Princeton University Press.

Schaefer, R., Roi, L., and Wolfe, R. (1984). A ridge logistic estimator. *Communications in Statistical Theory and Methods*, 13:99–113.

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

References VII

Sharma, D., Bondell, H. and Zhang, H. (2013). Consistent group identification and variable selection in regression with correlated predictors. *Journal of Computational and Graphical Statistics,* 22(2):319?-340.

She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4:1055-?1096.

Shevade, S. and Keerthi, S. (2003). A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19:2246–2253.

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

Stucky, B. and vande Geer, S. (2017). Sharp Oracle Inequalities for Square Root Regularization. *Journal of Machine Learning Research*: 18:1–29.

Taylor, H., Bank, S., and McCoy, J. (1979). Deconvolution with the ℓ_1 norm. *Geophysics*, 44:39–52.

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

Veríssimo, A., Oliveira, A. L., Sagot, M. F., and Vinga, S. (2016). DegreeCox - a network-based regularization method for survival analysis. *BMC Bioinformatics*. 17: 1310-1314.

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

References VIII

Virouleau, A., Gaiffas, S., Guilloux, A., Bogdan, M. (2017) High-dimensional robust regression and outliers detection with SLOPE. https://arxiv.org/abs/1712.02640

Wang, S., Weng, H., and Maleki, A. (2019). Does SLOPE outperform bridge regression? https://arxiv.org/abs/1909.09345

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Wiley, New York.

Williams, P. (1995). Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7:117–143.

Wright, S., Nowak, R., and Figueiredo, M. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57:2479–2493.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (B)*, 68(1):49.

Yurtsever, A., Tran-Dinh, Q., and Cevher, V. (2015). Universal primal-dual proximal-gradient methods. *arXiv preprint http://arxiv.org/pdf/1502.03123.pdf*.

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

Zeng, X. and Figueiredo, M. (2014b). Solving OSCAR regularization problems by fast approximate proximal splitting algorithms. *Digital Signal Processing*, 31:124–135.

Zeng, X. and Figueiredo, M. (2014a). The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms *arXiv preprint https://arxiv.org/abs/1409.4271*.

References IX

Zhang, D., Wang, H., Figueiredo, M., and Balzano, L. (2018). Learning to share: Simultaneous parameter tying and sparsification in deep learning. *International Conference on Learning Representations*.

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.

Elhamifar, E. and Vidal, R. (2013). Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765-?2781.

Rao, N., Recht, B., and Nowak, R. (2012) Tight measurement bounds for exact recovery of structured sparse signals. *AISTATS*.

Proximity Operator: Key Facts

Proximity operator of **OWL**: $\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- easy to show that, since $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \Omega_{\mathbf{w}}(|\mathbf{x}|)$,

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \text{prox}_{\Omega_{\mathbf{w}}}(|\mathbf{u}|)$$

Proximity Operator: Key Facts

Proximity operator of **OWL**: $\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- easy to show that, since $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \Omega_{\mathbf{w}}(|\mathbf{x}|)$,

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \text{prox}_{\Omega_{\mathbf{w}}}(|\mathbf{u}|)$$

- ...moreover, since $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \Omega_{\mathbf{w}}(\mathbf{P}\,\mathbf{x})$, for any permutation $\mathbf{P}$,

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \big(\mathbf{P}(|\mathbf{u}|)^T \,\text{prox}_{\Omega_{\mathbf{w}}}(|\mathbf{u}|_{\downarrow})\big)$$

Proximity Operator: Key Facts

Proximity operator of **OWL**: $\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{u} - \boldsymbol{\beta}\|_2^2 + \Omega_{\mathbf{w}}(\boldsymbol{\beta})$

- easy to show that, since $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \Omega_{\mathbf{w}}(|\mathbf{x}|)$,

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \text{prox}_{\Omega_{\mathbf{w}}}(|\mathbf{u}|)$$

- ...moreover, since $\Omega_{\mathbf{w}}(\boldsymbol{\beta}) = \Omega_{\mathbf{w}}(\mathbf{P}\,\mathbf{x})$, for any permutation $\mathbf{P}$,

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \big(\mathbf{P}(|\mathbf{u}|)^T \,\text{prox}_{\Omega_{\mathbf{w}}}(|\mathbf{u}|_{\downarrow})\big)$$

- Conclusion: we only need to compute $\text{prox}_{\Omega_{\mathbf{w}}}$ for arguments in $\mathcal{K}_{m+}$

$$\mathcal{B} \subset \mathcal{K}_{m+} = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p \geq 0\} \subset \mathbb{R}_+^p,$$

Proximity Operator: Key Facts

Proximity operator of **OWL**: $\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{u}-\boldsymbol{\beta}\|_2^2 + \Omega_\mathbf{w}(\boldsymbol{\beta})$

- easy to show that, since $\Omega_\mathbf{w}(\boldsymbol{\beta}) = \Omega_\mathbf{w}(|\mathbf{x}|)$,

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \text{prox}_{\Omega_\mathbf{w}}(|\mathbf{u}|)$$

- ...moreover, since $\Omega_\mathbf{w}(\boldsymbol{\beta}) = \Omega_\mathbf{w}(\mathbf{P}\,\mathbf{x})$, for any permutation $\mathbf{P}$,

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \, \text{prox}_{\Omega_\mathbf{w}}(|\mathbf{u}|_\downarrow)\Big)$$

- Conclusion: we only need to compute $\text{prox}_{\Omega_\mathbf{w}}$ for arguments in $\mathcal{K}_{m+}$

$$\mathcal{B} \subset \mathcal{K}_{m+} = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p \geq 0\} \subset \mathbb{R}^p_+,$$

- Finally, using the rearrangement inequality (Hardy et al., 1934)

Lemma

$$\mathbf{v} \in \mathcal{K}_{m+} \;\; \Rightarrow \;\; \text{prox}_{\Omega_\mathbf{w}}(\mathbf{v}) \in \mathcal{K}_{m+}$$

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathcal{K}_{m+}} \frac{1}{2}\|\mathbf{v} - \mathbf{x}\|_2^2 + \mathbf{w}^T\mathbf{x}$$

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \frac{1}{2}\|\mathbf{v}-\mathbf{x}\|_2^2 + \mathbf{w}^T\mathbf{x}$$

- Can be written as a projection onto $\mathcal{K}_{m+}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \|\mathbf{x}-(\mathbf{v}-\mathbf{w})\|_2^2 = \text{proj}_{\mathcal{K}_{m+}}\big(\mathbf{v}-\mathbf{w}\big)$$

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg \min_{\mathbf{x} \in \mathcal{K}_{m+}} \frac{1}{2} \|\mathbf{v} - \mathbf{x}\|_2^2 + \mathbf{w}^T \mathbf{x}$$

- Can be written as a projection onto $\mathcal{K}_{m+}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg \min_{\mathbf{x} \in \mathcal{K}_{m+}} \|\mathbf{x} - (\mathbf{v} - \mathbf{w})\|_2^2 = \text{proj}_{\mathcal{K}_{m+}}(\mathbf{v} - \mathbf{w})$$

- As shown by Németh and Németh (2012),

$$\text{proj}_{\mathcal{K}_{m+}}(\mathbf{z}) = \underbrace{\text{proj}_{\mathbb{R}_+^p}}\Big(\underbrace{\text{proj}_{\mathcal{K}_m}}(\mathbf{z})\Big)$$

where $\mathcal{K}_m = \{\mathbf{x} \in \mathbb{R}^p,\ x_1 \geq x_2 \cdots \geq x_p\}$ (the monotone cone).

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathcal{K}_{m+}} \frac{1}{2}\|\mathbf{v} - \mathbf{x}\|_2^2 + \mathbf{w}^T\mathbf{x}$$

- Can be written as a projection onto $\mathcal{K}_{m+}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x} \in \mathcal{K}_{m+}} \|\mathbf{x} - (\mathbf{v} - \mathbf{w})\|_2^2 = \text{proj}_{\mathcal{K}_{m+}}(\mathbf{v} - \mathbf{w})$$

- As shown by Németh and Németh (2012),

$$\text{proj}_{\mathcal{K}_{m+}}(\mathbf{z}) = \underbrace{\text{proj}_{\mathbb{R}^p_+}}_{\text{clipping}}\Big(\underbrace{\text{proj}_{\mathcal{K}_m}}(\mathbf{z})\Big)$$

where $\mathcal{K}_m = \{\mathbf{x} \in \mathbb{R}^p,\ x_1 \geq x_2 \cdots \geq x_p\}$ (the monotone cone).

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \frac{1}{2}\|\mathbf{v}-\mathbf{x}\|_2^2 + \mathbf{w}^T\mathbf{x}$$

- Can be written as a projection onto $\mathcal{K}_{m+}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \|\mathbf{x}-(\mathbf{v}-\mathbf{w})\|_2^2 = \text{proj}_{\mathcal{K}_{m+}}(\mathbf{v}-\mathbf{w})$$

- As shown by Németh and Németh (2012),

$$\text{proj}_{\mathcal{K}_{m+}}(\mathbf{z}) = \underbrace{\text{proj}_{\mathbb{R}^p_+}}_{\text{clipping}}\big(\underbrace{\text{proj}_{\mathcal{K}_m}}_{\text{PAVA}}(\mathbf{z})\big)$$

where $\mathcal{K}_m = \{\mathbf{x}\in\mathbb{R}^p,\ x_1 \geq x_2 \cdots \geq x_p\}$ (the monotone cone).

PAVA = *pool adjacent violators algorithm*, which has $O(p)$ cost (Barlow et al., 1972; Best and Chakravarti, 1990).

Proximity Operator: Derivation

- Let $\mathbf{v} \in \mathcal{K}_{m+}$ (of course, $\mathbf{w} \in \mathcal{K}_{m+}$), then

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \frac{1}{2}\|\mathbf{v}-\mathbf{x}\|_2^2 + \mathbf{w}^T\mathbf{x}$$

- Can be written as a projection onto $\mathcal{K}_{m+}$

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{v}) = \arg\min_{\mathbf{x}\in\mathcal{K}_{m+}} \|\mathbf{x}-(\mathbf{v}-\mathbf{w})\|_2^2 = \text{proj}_{\mathcal{K}_{m+}}(\mathbf{v}-\mathbf{w})$$

- As shown by Németh and Németh (2012),

$$\text{proj}_{\mathcal{K}_{m+}}(\mathbf{z}) = \underbrace{\text{proj}_{\mathbb{R}^p_+}}_{\text{clipping}}\Big(\underbrace{\text{proj}_{\mathcal{K}_m}}_{\text{PAVA}}(\mathbf{z})\Big)$$

where $\mathcal{K}_m = \{\mathbf{x}\in\mathbb{R}^p,\ x_1 \geq x_2 \cdots \geq x_p\}$ (the monotone cone).

PAVA = *pool adjacent violators algorithm*, which has $O(p)$ cost (Barlow et al., 1972; Best and Chakravarti, 1990).

- PAVA solves $\min_{\mathbf{x}} \|\mathbf{z}-\mathbf{x}\|_2^2$, subject to $x_1 \geq x_2 \geq \cdots \geq x_p$
monotone regression (Barlow et al., 1972; Best and Chakravarti, 1990).

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad\qquad\qquad\qquad |\mathbf{u}|_\downarrow$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \qquad\qquad (|\mathbf{u}|_\downarrow - \mathbf{w})$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \qquad\qquad \text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_{\downarrow} - \mathbf{w}\right)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_{\downarrow} = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \qquad \text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_{\downarrow} - \mathbf{w}\right)\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \qquad \text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\,(|\mathbf{u}|_\downarrow - \mathbf{w})\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_{\mathbf{w}}}(\mathbf{u}) = \qquad \mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\big(\text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)\big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \; \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\,(|\mathbf{u}|_\downarrow - \mathbf{w})\big)\Big)$$

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \; \text{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \,\text{proj}_{\mathbb{R}^p_+}\big(\,\text{proj}_{\mathcal{K}_m}\,(|\mathbf{u}|_\downarrow - \mathbf{w})\big)\Big)$$

- $\text{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA)

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\text{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \; \text{sign}(\mathbf{u}) \odot \left(\mathbf{P}(|\mathbf{u}|)^T \, \text{proj}_{\mathbb{R}^p_+}\left(\, \text{proj}_{\mathcal{K}_m}\left(|\mathbf{u}|_\downarrow - \mathbf{w}\right)\right)\right)$$

- $\text{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA)
- Computational cost: $O(p \log p)$, due to sorting; all else is $O(p)$.

Proximity Operator of OWL

- Monotone cone: $\mathcal{K}_m = \{\mathbf{z} \in \mathbb{R}^p : \; z_1 \geq z_2 \geq \cdots z_p\} \subset \mathbb{R}^p$
- Permutation matrix: $\mathbf{S} = \mathbf{P}(|\mathbf{u}|) \Leftrightarrow |\mathbf{u}|_\downarrow = \mathbf{S}|\mathbf{u}|$

Proposition (Zeng and F (2015))

$$\mathrm{prox}_{\Omega_\mathbf{w}}(\mathbf{u}) = \mathrm{sign}(\mathbf{u}) \odot \Big(\mathbf{P}(|\mathbf{u}|)^T \,\mathrm{proj}_{\mathbb{R}^p_+}\big(\mathrm{proj}_{\mathcal{K}_m}(|\mathbf{u}|_\downarrow - \mathbf{w})\big)\Big)$$

- $\mathrm{proj}_{\mathcal{K}_m}$: pool adjacent violators algorithm (PAVA)
- Computational cost: $O(p \log p)$, due to sorting; all else is $O(p)$.
- Compact expression for the algorithms by Bogdan et al. (2014); Zeng and F (2014b); Zhong and Kwok (2012)