OWL to the rescue of LASSO

IISc IBM day 2018

Joint Work

R. Sankaran and *Francis Bach*

AISTATS '17

Chiranjib Bhattacharyya

Professor, Department of Computer Science and Automation
Indian Institute of Science, Bangalore

March 7, 2018

Identifying groups of strongly correlated variables through Smoothed Ordered Weighted L_1-norms

LASSO: The method of choice for feature selection

Ordered Weighted L1(OWL) norm and Submodular penalties

$\Omega_{\mathcal{S}}$: Smoothed OWL (SOWL)

LASSO: The method of choice for feature selection

Linear Regression in High dimension

Question?

Let $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$ and

$$y = {w^*}^\top \mathbf{x} + \epsilon \;\; \epsilon \sim N(0, \sigma^2)$$

From $D = \{(\mathbf{x}_i, y_i) | i \in [n], \mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\}$ can we find w^*

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix}, Y \in \mathbb{R}^n$$

$X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^n$

Linear Regression in High dimension

Least Squares Linear Regression

$X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^d$:

$$\boxed{w_{LS} = \text{argmin}_{w \in \mathbb{R}^d} \frac{1}{2} \|Xw - Y\|_2^2}$$

Assumptions

- Labels centered : $\sum_{j=1}^n y_j = 0$.
- Features normalized : $x_i \in \mathbb{R}^d, \|x_i\|_2 = 1, x_i^\top 1_d = 0$.

Linear Regression in High dimension

Least Squares Linear Regression

$w_{LS} = \left(X^\top X\right)^{-1} X^\top Y$ and $E(w_{LS}) = w^*$

Variance of Predictive error: $\frac{1}{n}E(\|X(w_{LS} - w^*)\|^2) = \sigma^2 \frac{d}{n}$

Linear Regression in High dimension

$$\boxed{rank(X) = d}$$

- w_{LS} is unique
- Poor predictive performance, d is close to n

Linear Regression in High dimension

$rank(X) = d$

- w_{LS} is unique
- Poor predictive performance, d is close to n

$rank(X) < d$

- $d > n$
- w_{LS} is not unique.

Regularized Linear Regression

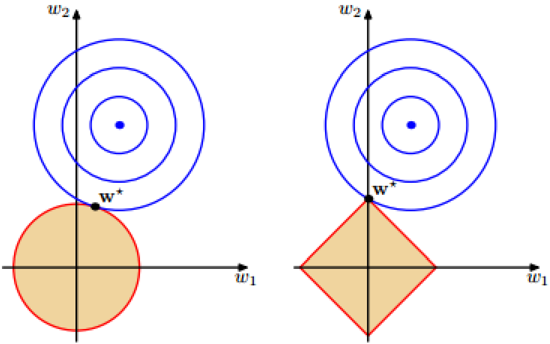
Regularized Linear Regression $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^d, \Omega : \mathbb{R}^d \to \mathbb{R}$:

$$\boxed{\min_{w \in \mathbb{R}^d} \frac{1}{2}\|Xw - y\|_2^2 \quad \text{s.t. } \Omega(w) \le t}$$

Regularizer: $\Omega(w)$

- non-negative
- Convex function, typically a norm.
- Possibly non-differentiable.

Lasso regression[Tibshirani, 1994]



Ridge: $\Omega(w) = \|w\|_2^2$

- Does not promote sparsity
- Closed form solution

Lasso: $\Omega(w) = \|w\|_1$

- Encourages sparse solutions.
- Solve convex optimization problem

Regularized Linear Regression

Regularized Linear Regression $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^d, \Omega : \mathbb{R}^d \to \mathbb{R}$:

$$\boxed{\min_{w \in \mathbb{R}^d} \frac{1}{2}\|Xw - y\|_2^2 + \lambda \Omega(w)}$$

- Equivalent to the constraint version
- Unconstrained

Lasso: Properties at a glance

Computational

- Proximal methods : IST, FISTA, Chambolle-Pock [Chambolle and Pock, 2011, Beck and Teboulle, 2009].
- Convergence rate : $O(1/T^2)$ in T iterations.
- Assumption: availability of *proximal operator* of Ω (Easy for ℓ_1).

Lasso: Properties at a glance

Computational

- Proximal methods : IST, FISTA, Chambolle-Pock [Chambolle and Pock, 2011, Beck and Teboulle, 2009].
- Convergence rate : $O(1/T^2)$ in T iterations.
- Assumption: availability of *proximal operator* of Ω (Easy for ℓ_1).

Statistical properties[Wainwright, 2009]

- Support recovery (Will it recover the true support ?).
- Sample complexity (How many samples needed ?).
- Prediction error (What is the expected error in prediction ?).

Lasso Model Recovery[Wainwright, 2009, Theorem 1]

Setup

$$y = Xw^* + \epsilon, \epsilon_i \sim \mathcal{N}(0, \sigma^2), X \in \mathbb{R}^{n \times d}$$

Support of w^* be $S = \{j | w_j^* \neq 0 \; j \in [d]\}$

Lasso

$$\min_{w \in \mathbb{R}^d} \frac{1}{2} \|Xw - y\|_2^2 + \lambda \|w\|_1$$

Lasso Model Recovery[Wainwright, 2009, Theorem 1]

Conditions[1].

$$\left\| X_{S^c}^\top X_S \left(X_S^\top X_S \right)^{-1} \right\|_\infty \leq 1 - \gamma, \text{Incoherence with } \gamma \in (0,1],$$

$$\Lambda_{\min}\left(\frac{1}{n} X_S^\top X_S \right) \geq C_{\min}$$

$$\lambda > \lambda_0 \triangleq \frac{2}{\gamma}\sqrt{\frac{2\sigma^2 \log d}{n}}$$

[1]Define $\|M\|_\infty = \max_i \sum_j |M_{ij}|$

Lasso Model Recovery[Wainwright, 2009, Theorem 1]

Conditions[1].

$$\left\| X_{S^c}^\top X_S \left(X_S^\top X_S \right)^{-1} \right\|_\infty \leq 1 - \gamma, \text{Incoherence with } \gamma \in (0,1],$$

$$\Lambda_{\min} \left(\frac{1}{n} X_S^\top X_S \right) \geq C_{\min}$$

$$\lambda > \lambda_0 \triangleq \frac{2}{\gamma} \sqrt{\frac{2\sigma^2 \log d}{n}}$$

W.h.p, the following holds:

$$\| \hat{w}_S - w_S^* \|_\infty \leq \lambda \left(\left\| \left(X_S^\top X_S / n \right)^{-1} \right\|_\infty + 4\sigma / \sqrt{C_{\min}} \right)$$

[1] Define $\|M\|_\infty = \max_i \sum_j |M_{ij}|$

Lasso Model Recovery: Special cases

Case: $X_{S^c}^\top X_S = 0$.

- The incoherence condition trivially holds, and $\gamma = 1$.
- The threshold λ_0 is lesser $\Rightarrow$ The recovery error is lesser.

Case: $X_S^\top X_S = I$.

- $C_{\min} = 1/n$, the largest possible for a given n.
- Larger $C_{\min} \Rightarrow$ lesser recovery error.

Lasso Model Recovery: Special cases

Case: $X_{S^c}^\top X_S = 0$.

- The incoherence condition trivially holds, and $\gamma = 1$.
- The threshold λ_0 is lesser $\Rightarrow$ The recovery error is lesser.

Case: $X_S^\top X_S = I$.

- $C_{\min} = 1/n$, the largest possible for a given n.
- Larger $C_{\min} \Rightarrow$ lesser recovery error.

When does Lasso work well?

- Lasso prefers low correlation between support and non-support columns.
- Low correlation of columns within support lead to better recovery.

Lasso Model Recovery: Implications

Setting: Strongly correlated columns in X.

- Correlation between feature i and feature j

$$\rho_{ij} \approx x_i^\top x_j$$

- Large correlation between X_S and $X_{S^c} \Rightarrow \gamma$ is small.
- Large correlation within $X_S \Rightarrow C_{\min}$ is small.

Lasso Model Recovery: Implications

Setting: Strongly correlated columns in X.

- Correlation between feature i and feature j

$$\rho_{ij} \approx x_i^\top x_j$$

- Large correlation between X_S and $X_{S^c} \Rightarrow \gamma$ is small.
- Large correlation within $X_S \Rightarrow C_{\min}$ is small.
- The r.h.s. of the bound is large. (loose bound).
- Hence w.h.p., lasso fails in **model recovery**.

In other words:

- Lasso solutions differ with the solver used.
- Solution is not unique typically.
- The prediction error may not be as worse though [Hebiri and Lederer, 2013].

Lasso Model Recovery: Implications

Setting: Strongly correlated columns in X.

- Correlation between feature i and feature j

$$\rho_{ij} \approx x_i^\top x_j$$

- Large correlation between X_S and $X_{S^c} \Rightarrow \gamma$ is small.
- Large correlation within $X_S \Rightarrow C_{\min}$ is small.
- The r.h.s. of the bound is large. (loose bound).
- Hence w.h.p., lasso fails in **model recovery**.

In other words:

- Lasso solutions differ with the solver used.
- Solution is not unique typically.
- The prediction error may not be as worse though [Hebiri and Lederer, 2013].

Requirements.

- Need consistent estimates independent of the solver.
- Preferably select all the correlated variables as a group.

Illustration: Lasso under correlation[Zeng and Figueiredo, 2015]

Setting: strongly correlated features.

- $\{1, \ldots, d\} = \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_k$, $\mathcal{G}_m \cap \mathcal{G}_l = \emptyset, \forall l \neq m$
- $\rho_{ij} \equiv |x_i^\top x_j|$ very high (≈ 1) for pairs $i, j \in \mathcal{G}_m$.

Illustration: Lasso under correlation[Zeng and Figueiredo, 2015]

Setting: strongly correlated features.

- $\{1, \ldots, d\} = \mathcal{G}_1 \cup \cdots \cup \mathcal{G}_k$, $\mathcal{G}_m \cap \mathcal{G}_l = \emptyset, \forall l \neq m$
- $\rho_{ij} \equiv |x_i^\top x_j|$ very high (≈ 1) for pairs $i, j \in \mathcal{G}_m$.

Toy Example

- $d = 40$, $k = 4$.
- $\mathcal{G}_1 = [1:10]$, $\mathcal{G}_2 = [11:20]$, $\mathcal{G}_3 = [21:30]$, $\mathcal{G}_4 = [31:40]$.

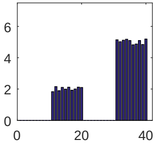


Figure: Original signal

Lasso: $\Omega(w) = \|w\|_1$.

- Sparse recovery.
- Arbitrarily selects the variables within a group.

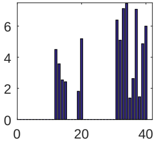


Figure: Recovered signal.

Possible solutions: 2-stage procedures

Cluster Group Lasso [Buhlmann et al., 2013]

- Identify strongly correlated groups $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$
  - Canonical Correlation.
- Group selection. $\Omega(w) = \sum_{j=1}^{k} \alpha_j \|w_{\mathcal{G}_j}\|_2$.
- Select all or no variables from each group.

Possible solutions: 2-stage procedures

Cluster Group Lasso [Buhlmann et al., 2013]

- Identify strongly correlated groups $\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_k\}$
  - Canonical Correlation.
- Group selection. $\Omega(w) = \sum_{j=1}^{k} \alpha_j \|w_{\mathcal{G}_j}\|_2$.
- Select all or no variables from each group.

Goal: Learn $\mathcal{G}$ and w simultaneously ?

Ordered Weighted ℓ_1 (OWL) norms

OSCAR[2] [Bondell and Reich, 2008]

$$\Omega_{\mathcal{O}}(w) = \sum_{i=1}^{d} c_i |w|_{(i)}$$
$$c_i = c_0 + (d-i)\mu,$$
$$c_0, \mu, c_d > 0.$$

[2]**Notation:** $|w|_{(i)}$: i^{th} largest in $|w|$.

Ordered Weighted ℓ_1 (OWL) norms

OSCAR[2] [Bondell and Reich, 2008]

$$\Omega_{\mathcal{O}}(w) = \sum_{i=1}^{d} c_i |w|_{(i)}$$
$$c_i = c_0 + (d-i)\mu,$$
$$c_0, \mu, c_d > 0.$$

OWL [Figueiredo and Nowak, 2016]:

- $c_1 \geq \cdots \geq c_d \geq 0.$

[2] **Notation:** $|w|_{(i)}$: i^{th} largest in $|w|$.

Ordered Weighted ℓ_1 (OWL) norms

OSCAR[2] [Bondell and Reich, 2008]

$$\Omega_{\mathcal{O}}(w) = \sum_{i=1}^{d} c_i |w|_{(i)}$$
$$c_i = c_0 + (d-i)\mu,$$
$$c_0, \mu, c_d > 0.$$

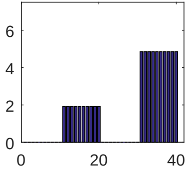


Figure: Recovered: OWL

OWL [Figueiredo and Nowak, 2016]:

- $c_1 \geq \cdots \geq c_d \geq 0.$

[2]**Notation:** $|w|_{(i)}$: i^{th} largest in $|w|$.

Oscar: Sparsity Illustrated

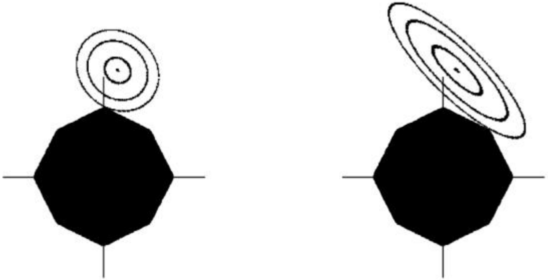


Figure: Examples of solutions:[Bondell and Reich, 2008]

- Solutions encouraged towards vertices.
- Encourages blockwise constant solutions.
- See [Bondell and Reich, 2008].

OWL-Properties

Grouping covariates [Bondell and Reich, 2008, Theorem 1], [Figueiredo and Nowak, 2016, Theorem 1]

$$\boxed{|w_i| = |w_j|, \text{if } \lambda \geq \lambda_{ij}^0.}$$

OWL-Properties

Grouping covariates [Bondell and Reich, 2008, Theorem 1], [Figueiredo and Nowak, 2016, Theorem 1]

$$\boxed{|w_i| = |w_j|, \text{if } \lambda \geq \lambda_{ij}^0.}$$

- $\lambda_{ij}^0 \propto \sqrt{1 - \rho_{ij}^2}$.
- Strongly correlated pairs grouped early in the regularization path.
- Groups: $\mathcal{G}_j = \{i \; ||w_i| = \alpha_j\}$.

OWL-Issues

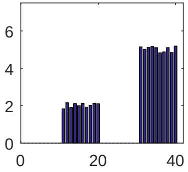


Figure: True model

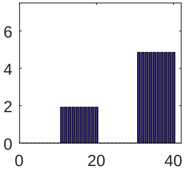


Figure: Recovered: OWL

- **Bias for piecewise constant $\hat{w}$**
  - Easily understood through the norm balls.
  - Requires more samples to consistent estimation.

OWL-Issues

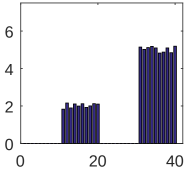


Figure: True model

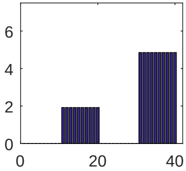


Figure: Recovered: OWL

- **Bias for piecewise constant** $\hat{w}$
  - Easily understood through the norm balls.
  - Requires more samples to consistent estimation.
- **Lack of interpretations for choosing** c

Ordered Weighted L1(OWL) norm and Submodular penalties

Preliminaries: Penalties on the Support

Goal

Encourage w to have desired support structure.

$$\mathrm{supp}(w) = \{i | w_i \neq 0\}$$

Preliminaries: Penalties on the Support

Goal

Encourage w to have desired support structure.

$$\mathrm{supp}(w) = \{i | w_i \neq 0\}$$

Idea

Penalty on support [Obozinski and Bach, 2012]:

$$\mathrm{pen}(w) = F(\mathrm{supp}(w)) + \|w\|_p^p, p \in [1, \infty].$$

Preliminaries: Penalties on the Support

Goal

Encourage w to have desired support structure.

$\mathrm{supp}(w) = \{i | w_i \neq 0\}$

Idea

Penalty on support [Obozinski and Bach, 2012]:

$\mathrm{pen}(w) = F(\mathrm{supp}(w)) + \|w\|_p^p, p \in [1, \infty]$.

Relaxation(pen(w)): $\Omega_p^F(w)$

Tightest positively homogenous, convex lower bound.

Preliminaries: Penalties on the Support

Goal

Encourage w to have desired support structure.

$\mathrm{supp}(w) = \{i | w_i \neq 0\}$

Idea

Penalty on support [Obozinski and Bach, 2012]:

$\mathrm{pen}(w) = F(\mathrm{supp}(w)) + \|w\|_p^p, p \in [1, \infty]$.

Relaxation(pen(w)): $\Omega_p^F(w)$

Tightest positively homogenous, convex lower bound.

Example: $F(\mathrm{supp}(w)) = |\mathrm{supp}(w)|$

Preliminaries: Penalties on the Support

Goal

Encourage w to have desired support structure.

$$\text{supp}(w) = \{i | w_i \neq 0\}$$

Idea

Penalty on support [Obozinski and Bach, 2012]:

$$\text{pen}(w) = F(\text{supp}(w)) + \|w\|_p^p, p \in [1, \infty].$$

Relaxation(pen(w)): $\Omega_p^F(w)$

Tightest positively homogenous, convex lower bound.

Example: $F(\text{supp}(w)) = |\text{supp}(w)| \Rightarrow \Omega_p^F(w) = \|w\|_1$. **Familiar!**

Message: The cardinality function always relaxes to the ℓ_1 norm.

Nondecreasing Submodular Penalties on Cardinality

Assumptions. Denote $F \in \mathcal{F}$, if $F : A \subseteq \{1, \ldots, d\} \to \mathbb{R}$ is:

1. **Submodular** [Bach, 2011].
   - $\forall A \subseteq B, F(A \cup \{k\}) - F(A) \geq F(B \cup \{k\}) - F(B)$.
   - Lovász extension: $f : \mathbb{R}^d \to \mathbb{R}$. (Convex extension of F to $\mathbb{R}^d$).
2. **Cardinality based**.
   - $F(A) = g(|A|)$ (Invariant to permutations).
3. **Non Decreasing**.
   - $g(0) = 0, g(x) \geq g(x-1)$.

Implication: $F \in \mathcal{F} \Rightarrow F$ completely specified through g.
Example: Let $V = \{1, \ldots, d\}$, define $F(A) = |A||V \setminus A|$.
Then $f(w) = \sum_{i<j} |w_i - w_j|$.

Ω_∞^F and Lovász extension

Result: Case $p = \infty$ [Bach, 2010]:

$$\boxed{\Omega_\infty^F(w) = f(|w|)}$$

- The ℓ_∞ relaxation coincides with the Lovász extension in the positive orthant.
- To work with Ω_∞^F, may use existing results of submodular function minimization.
- Ω_p^F not known in closed form for $p < \infty$.

Equivalence of OWL and Lovász extensions: Statement

Proposition [Sankaran et al., 2017]: $F \in \mathcal{F}, \Omega_\infty^F(w) \Leftrightarrow \Omega_\mathcal{O}(w)$

1. Given $F(A) = f(|A|)$, $\Omega_\infty^F(w) = \Omega_\mathcal{O}(w)$, with $c_i = f(i) - f(i-1)$.
2. Given $c_1 \geq \ldots c_d \geq 0$, $\Omega_\mathcal{O}(w) = \Omega_\infty^F(w)$ with $f(i) = c_1 + \cdots + c_i$.

Interpretations

- Gives alternate interpretations for OWL.

Equivalence of OWL and Lovász extensions: Statement

Proposition [Sankaran et al., 2017]: $F \in \mathcal{F}, \Omega_\infty^F(w) \Leftrightarrow \Omega_\mathcal{O}(w)$

1. Given $F(A) = f(|A|)$, $\Omega_\infty^F(w) = \Omega_\mathcal{O}(w)$, with $c_i = f(i) - f(i-1)$.
2. Given $c_1 \geq \dots c_d \geq 0$, $\Omega_\mathcal{O}(w) = \Omega_\infty^F(w)$ with $f(i) = c_1 + \dots + c_i$.

Interpretations

- Gives alternate interpretations for OWL.
- $\Omega_\infty^\mathcal{F}$ has undesired extreme points [Bach, 2011].
  - Explains piecewise constant solutions of OWL.

Equivalence of OWL and Lovász extensions: Statement

Proposition [Sankaran et al., 2017]: $F \in \mathcal{F}, \Omega_\infty^F(w) \Leftrightarrow \Omega_{\mathcal{O}}(w)$

1. Given $F(A) = f(|A|)$, $\Omega_\infty^F(w) = \Omega_{\mathcal{O}}(w)$, with $c_i = f(i) - f(i-1)$.
2. Given $c_1 \geq \ldots c_d \geq 0$, $\Omega_{\mathcal{O}}(w) = \Omega_\infty^F(w)$ with $f(i) = c_1 + \cdots + c_i$.

Interpretations

- Gives alternate interpretations for OWL.
- $\Omega_\infty^{\mathcal{F}}$ has undesired extreme points [Bach, 2011].
  - Explains piecewise constant solutions of OWL.
- Motivates $\Omega_p^F(w)$ for $p < \infty$.

$\Omega_{\mathcal{S}}$: Smoothed OWL (SOWL)

SOWL: Definition

Smoothed OWL

$$\boxed{\Omega_{\mathcal{S}}(w) := \Omega_2^F(w).}$$

SOWL: Definition

Smoothed OWL

$$\boxed{\Omega_{\mathcal{S}}(w) := \Omega_2^F(w).}$$

Variational form for Ω_2^F [Obozinski and Bach, 2012].

$$\boxed{\Omega_{\mathcal{S}}(w) = \min_{\eta \in \mathbb{R}_+^d} \frac{1}{2}\left(\sum_{i=1}^{d} \frac{w_i^2}{\eta_i} + f(\eta)\right).}$$

SOWL: Definition

Smoothed OWL

$$\boxed{\Omega_{\mathcal{S}}(w) := \Omega_2^F(w).}$$

Variational form for Ω_2^F [Obozinski and Bach, 2012].

$$\boxed{\Omega_{\mathcal{S}}(w) = \min_{\eta \in \mathbb{R}_+^d} \frac{1}{2} \left(\sum_{i=1}^{d} \frac{w_i^2}{\eta_i} + f(\eta) \right).}$$

Use OWL equivalance: $f(|\eta|) = \Omega_\infty^F(\eta) = \sum_{i=1}^d c_i |\eta|_{(i)}$,

$$\boxed{\Omega_{\mathcal{S}}(w) = \min_{\eta \in \mathbb{R}_+^d} \underbrace{\frac{1}{2} \sum_{i=1}^{d} \left(\frac{w_i^2}{\eta_i} + c_i \eta_{(i)} \right)}_{\Psi(w,\eta)}.} \qquad \text{(SOWL)}$$

OWL vs SOWL

Case: $c = 1_d$.

- $\Omega_{\mathcal{S}}(w) = \|w\|_1$.
- $\Omega_{\mathcal{O}}(w) = \|w\|_1$.

Case: $c = [1, \underbrace{0, \ldots, 0}_{d-1}]^\top$.

- $\Omega_{\mathcal{S}}(w) = \|w\|_2$.
- $\Omega_{\mathcal{O}}(w) = \|w\|_\infty$.

OWL vs SOWL

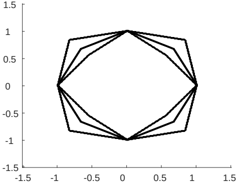
Case: $c = 1_d$.

- $\Omega_{\mathcal{S}}(w) = \|w\|_1$.
- $\Omega_{\mathcal{O}}(w) = \|w\|_1$.

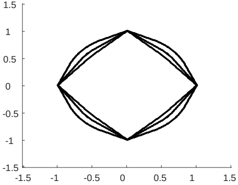
Case: $c = [1, \underbrace{0, \ldots, 0}_{d-1}]^\top$.

- $\Omega_{\mathcal{S}}(w) = \|w\|_2$.
- $\Omega_{\mathcal{O}}(w) = \|w\|_\infty$.

Norm Balls



: OWL



: SOWL

Figure: Norm balls for OWL, SOWL, for different values of c

Group Lasso and $\Omega_{\mathcal{S}}$

SOWL objective (eliminating η):

$$\boxed{\Omega_{\mathcal{S}}(w) = \sum_{j=1}^{k} \left(\|w_{\mathcal{G}_j}\| \sqrt{\sum_{i \in \mathcal{G}_j} c_i} \right) .}$$

Denote η_w: denote the optimal η, given w.

Group Lasso and $\Omega_{\mathcal{S}}$

SOWL objective (eliminating η):

$$\Omega_{\mathcal{S}}(w) = \sum_{j=1}^{k} \left(\|w_{\mathcal{G}_j}\| \sqrt{\sum_{i \in \mathcal{G}_j} c_i} \right).$$

Denote η_w: denote the optimal η, given w.
Key differences:

- Groups defined through $\eta_w = [\underbrace{\delta_1, \ldots, \delta_1}_{\mathcal{G}_1}, \ldots, \underbrace{\delta_k, \ldots, \delta_k}_{\mathcal{G}_k}]$.
- Influenced by the choice of c.

Open Questions:

1. Does $\Omega_{\mathcal{S}}$ promotes grouping of correlated variables as $\Omega_{\mathcal{O}}$?
   - Are there any benefits over $\Omega_{\mathcal{O}}$?

Open Questions:

1. Does $\Omega_{\mathcal{S}}$ promotes grouping of correlated variables as $\Omega_{\mathcal{O}}$?
    - Are there any benefits over $\Omega_{\mathcal{O}}$?
2. Is using $\Omega_{\mathcal{S}}$ computationally feasible ?

Open Questions:

1. Does $\Omega_{\mathcal{S}}$ promotes grouping of correlated variables as $\Omega_{\mathcal{O}}$?
   - Are there any benefits over $\Omega_{\mathcal{O}}$?
2. Is using $\Omega_{\mathcal{S}}$ computationally feasible ?
3. Theoretical properties of $\Omega_{\mathcal{S}}$ vs Group Lasso?

Grouping property $\Omega_{\mathcal{S}}$: Statement

Learning Problem: LS-SOWL

$$\min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}^d_+} \underbrace{\frac{1}{2n}\|Xw - y\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\left(\frac{w_i^2}{\eta_i} + c_i \eta_{(i)}\right)}_{\Gamma^{(\lambda)}(w,\eta)}.$$

Grouping property $\Omega_{\mathcal{S}}$: Statement

Learning Problem: LS-SOWL

$$\min_{w\in\mathbb{R}^d,\eta\in\mathbb{R}^d_+} \underbrace{\frac{1}{2n}\|Xw-y\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\left(\frac{w_i^2}{\eta_i} + c_i\eta_{(i)}\right)}_{\Gamma^{(\lambda)}(w,\eta)}.$$

Theorem: [Sankaran et al., 2017]

Define the following:

- $\left(\hat{w}^{(\lambda)}, \hat{\eta}^{(\lambda)}\right) = \text{argmin}_{w,\eta}\Gamma^{(\lambda)}(w,\eta)$.
- $\rho_{ij} = x_i^\top x_j$.
- $\tilde{c} = \min_i c_i - c_{i+1}$.

Grouping property $\Omega_{\mathcal{S}}$: Statement

Learning Problem: LS-SOWL

$$\min_{w \in \mathbb{R}^d, \eta \in \mathbb{R}^d_+} \underbrace{\frac{1}{2n}\|Xw - y\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{d}\left(\frac{w_i^2}{\eta_i} + c_i \eta_{(i)}\right)}_{\Gamma^{(\lambda)}(w,\eta)}.$$

Theorem: [Sankaran et al., 2017]

Define the following:

- $\left(\hat{w}^{(\lambda)}, \hat{\eta}^{(\lambda)}\right) = \text{argmin}_{w,\eta}\Gamma^{(\lambda)}(w,\eta)$.
- $\rho_{ij} = x_i^\top x_j$.
- $\tilde{c} = \min_i c_i - c_{i+1}$.

There exists $0 \leq \lambda^0 \leq \frac{\|y\|_2}{\sqrt{\tilde{c}}}(4 - 4\rho_{ij}^2)^{\frac{1}{4}}$, such that $\forall \lambda > \lambda^0$, $\hat{\eta}_i^{(\lambda)} = \hat{\eta}_j^{(\lambda)}$

Grouping property $\Omega_{\mathcal{S}}$: Interpretation

1. Variables η_i, η_j grouped if $\rho_{ij} \approx 1$ (Even for small λ).
   - Similar to Figueiredo and Nowak [2016, Theorem 1], which is for absolute values of w.

Grouping property $\Omega_{\mathcal{S}}$: Interpretation

1. Variables η_i, η_j grouped if $\rho_{ij} \approx 1$ (Even for small λ).
   - Similar to Figueiredo and Nowak [2016, Theorem 1], which is for absolute values of w.
2. SOWL differentiates grouping variable η and model variable w.

Grouping property $\Omega_{\mathcal{S}}$: Interpretation

1. Variables η_i, η_j grouped if $\rho_{ij} \approx 1$ (Even for small λ).
   - Similar to Figueiredo and Nowak [2016, Theorem 1], which is for absolute values of w.
2. SOWL differentiates grouping variable η and model variable w.
3. Allows model variance within group.
   - $\hat{w}_i^{(\lambda)} \neq \hat{w}_j^{(\lambda)}$ as long as c has distinct values.

Illustration: Group Discovery using SOWL

Aim: Illustrate group discovery of SOWL.

- Consider $z \in \mathbb{R}^d$, Compute $\text{prox}_{\lambda\Omega_{\mathcal{S}}}(z)$.
- Study the regularization path.

Illustration: Group Discovery using SOWL

Aim: Illustrate group discovery of SOWL.

- Consider $z \in \mathbb{R}^d$, Compute $\text{prox}_{\lambda\Omega_{\mathcal{S}}}(z)$.
- Study the regularization path.

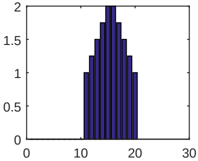


Figure: Original signal

Illustration: Group Discovery using SOWL

Aim: Illustrate group discovery of SOWL.

- Consider $z \in \mathbb{R}^d$, Compute $\text{prox}_{\lambda\Omega_{\mathcal{S}}}(z)$.
- Study the regularization path.

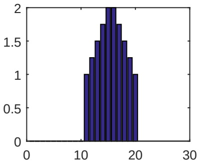


Figure: Original signal

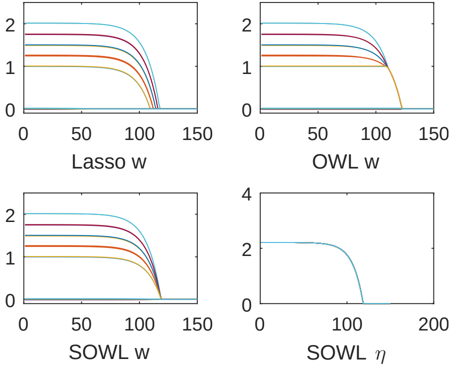


Figure: x-axis : λ, y-axis: $\hat{w}$.

- Early group discovery.
- Model variation.

Proximal Methods: A brief overview

Proximal operator

$$\boxed{\mathrm{prox}_{\Omega}(z) = \mathrm{argmin}_w \frac{1}{2}\|w - z\|_2^2 + \Omega(w)}$$

- Easy to evaluate for many simple norms.
- $\mathrm{prox}_{\lambda \ell_1}(z) = sign(z)\left(|z| - \lambda\right)_+$.
- Generalization of Projected Gradient Descent

FISTA [Beck and Teboulle, 2009]

Initialization

- $t^{(1)} = 1,\ \tilde{w}^{(1)} = x^{(1)} = 0.$

Steps : $k > 1$

- $w^{(k)} = \text{prox}_{\Omega}\left(\tilde{w}^{(k-1)} - \frac{1}{L}\nabla f(\tilde{w}^{(k-1)})\right).$
- $t^{(k)} = \left(1 + \sqrt{1 + 4\left(t^{(k-1)}\right)^2}\right)/2.$
- $\tilde{w}^{(k)} = w^{(k)} + \left(\frac{t^{(k-1)}-1}{t^{(k)}}\right)\left(w^{(k)} - w^{(k-1)}\right).$

Guarantee

- Convergence rate $O(1/T^2)$.
- No additional assumptions than IST.
- Known to be optimal for this class of minimization problems.

Computing $\text{prox}_{\Omega_{\mathcal{S}}}$

Problem:

$$\text{prox}_{\lambda\Omega}(z) = \text{argmin}_w \frac{1}{2}\|w - z\|_2^2 + \lambda\Omega(w).$$

$$w^{(\lambda)} = \text{prox}_{\lambda\Omega_{\mathcal{S}}}(z), \eta_w^{(\lambda)} = \text{argmin}_\eta \Psi(w^{(\lambda)}, \eta).$$

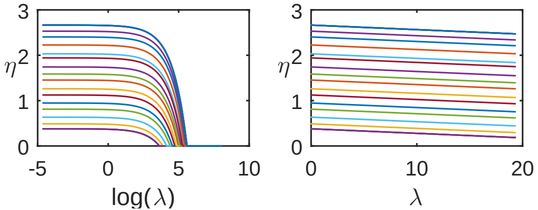
Computing $\text{prox}_{\Omega_{\mathcal{S}}}$

Problem:

$$\text{prox}_{\lambda\Omega}(z) = \text{argmin}_w \frac{1}{2}\|w - z\|_2^2 + \lambda\Omega(w).$$

$$w^{(\lambda)} = \text{prox}_{\lambda\Omega_{\mathcal{S}}}(z), \eta_w^{(\lambda)} = \text{argmin}_\eta \Psi(w^{(\lambda)}, \eta).$$

Key idea: Ordering of $\eta_w^{(\lambda)}$ remains same for all λ.



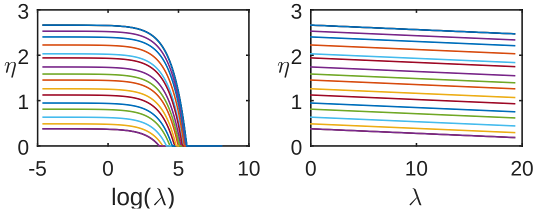
Computing $\text{prox}_{\Omega_{\mathcal{S}}}$

Problem:

$$\text{prox}_{\lambda\Omega}(z) = \text{argmin}_w \frac{1}{2}\|w - z\|_2^2 + \lambda\Omega(w).$$

$$w^{(\lambda)} = \text{prox}_{\lambda\Omega_{\mathcal{S}}}(z), \eta_w^{(\lambda)} = \text{argmin}_\eta \Psi(w^{(\lambda)}, \eta).$$

Key idea: Ordering of $\eta_w^{(\lambda)}$ remains same for all λ.



- $\eta_w^{(\lambda)} = (\eta_z - \lambda)_+$.
- Same complexity as computing the norm $\Omega_{\mathcal{S}}(O(d \log d))$.
- True for all cardinality based F.

Random Design

Problem setting: LS-SOWL.

- True model: $y = Xw^* + \varepsilon$, $\left(X^\top\right)_i \sim \mathcal{N}(\mu, \Sigma)$, $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Notation: $\mathcal{J} = \{i | w_i^* \neq 0\}, \eta_{w^*} = [\underbrace{\delta_1^*, \ldots, \delta_1^*}_{G_1}, \ldots, \underbrace{\delta_k^*, \ldots, \delta_k^*}_{G_k}]$.

[3] D: Diagonal matrix, defined using groups $\mathcal{G}_1, \ldots, \mathcal{G}_k$, and w^*

Random Design

Problem setting: LS-SOWL.

- True model: $y = Xw^* + \varepsilon$, $\left(X^\top\right)_i \sim \mathcal{N}(\mu, \Sigma)$, $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Notation: $\mathcal{J} = \{i | w_i^* \neq 0\}, \eta_{w^*} = [\underbrace{\delta_1^*, \ldots, \delta_1^*}_{G_1}, \ldots, \underbrace{\delta_k^*, \ldots, \delta_k^*}_{G_k}]$.

Assumptions: [3]

- $\Sigma_{\mathcal{J},\mathcal{J}}$ is invertible, $\lambda \to 0$, and $\lambda\sqrt{n} \to \infty$.

[3] D: Diagonal matrix, defined using groups $\mathcal{G}_1, \ldots, \mathcal{G}_k$, and w^*

Random Design

Problem setting: LS-SOWL.

- True model: $y = Xw^* + \varepsilon$, $\left(X^\top\right)_i \sim \mathcal{N}(\mu, \Sigma)$, $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Notation: $\mathcal{J} = \{i | w_i^* \neq 0\}, \eta_{w^*} = [\underbrace{\delta_1^*, \ldots, \delta_1^*}_{G_1}, \ldots, \underbrace{\delta_k^*, \ldots, \delta_k^*}_{G_k}]$.

Assumptions: [3]

- $\Sigma_{\mathcal{J},\mathcal{J}}$ is invertible, $\lambda \to 0$, and $\lambda\sqrt{n} \to \infty$.

Irrepresentability conditions

1. $\delta_k^* = 0$ if $|\mathcal{J}^c| \neq \emptyset$.
2. $\dfrac{\left\| \Sigma_{\mathcal{J}^c,\mathcal{J}} \left(\Sigma_{\mathcal{J},\mathcal{J}}\right)^{-1} \mathrm{D} w_{\mathcal{J}}^* \right\|_2}{\beta} < 1$.

[3] D: Diagonal matrix, defined using groups $\mathcal{G}_1, \ldots, \mathcal{G}_k$, and w^*

Random Design

Problem setting: LS-SOWL.

- True model: $y = Xw^* + \varepsilon$, $\left(X^\top\right)_i \sim \mathcal{N}(\mu, \Sigma)$, $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Notation: $\mathcal{J} = \{i | w_i^* \neq 0\}, \eta_{w^*} = [\underbrace{\delta_1^*, \dots, \delta_1^*}_{G_1}, \dots, \underbrace{\delta_k^*, \dots, \delta_k^*}_{G_k}]$.

Assumptions: [3]

- $\Sigma_{\mathcal{J},\mathcal{J}}$ is invertible, $\lambda \to 0$, and $\lambda\sqrt{n} \to \infty$.

Irrepresentability conditions

1. $\delta_k^* = 0$ if $|\mathcal{J}^c| \neq \emptyset$.
2. $\frac{\left\| \Sigma_{\mathcal{J}^c,\mathcal{J}} \left(\Sigma_{\mathcal{J},\mathcal{J}}\right)^{-1} \mathrm{D} w_{\mathcal{J}}^* \right\|_2}{\beta} < 1$.

$\longrightarrow$

Result[Sankaran et al., 2017]

1. $\hat{w} \to_p w^*$.
2. $\mathbb{P}(\hat{\mathcal{J}} = \mathcal{J}) \to 1$.

[3] D: Diagonal matrix, defined using groups $\mathcal{G}_1, \dots, \mathcal{G}_k$, and w^*

Random Design

Problem setting: LS-SOWL.

- True model: $y = Xw^* + \varepsilon$, $\left(X^\top\right)_i \sim \mathcal{N}(\mu, \Sigma)$, $\varepsilon_i \in \mathcal{N}(0, \sigma^2)$.
- Notation: $\mathcal{J} = \{i | w_i^* \neq 0\}, \eta_{w^*} = [\underbrace{\delta_1^*, \ldots, \delta_1^*}_{G_1}, \ldots, \underbrace{\delta_k^*, \ldots, \delta_k^*}_{G_k}]$.

Assumptions: [3]

- $\Sigma_{\mathcal{J},\mathcal{J}}$ is invertible, $\lambda \to 0$, and $\lambda\sqrt{n} \to \infty$.

Irrepresentability conditions

1. $\delta_k^* = 0$ if $|\mathcal{J}^c| \neq \emptyset$.
2. $\dfrac{\left\| \Sigma_{\mathcal{J}^c,\mathcal{J}} \left(\Sigma_{\mathcal{J},\mathcal{J}}\right)^{-1} \mathrm{D} w_{\mathcal{J}}^* \right\|_2}{\beta} < 1$.

$\longrightarrow$

Result[Sankaran et al., 2017]

1. $\hat{w} \to_p w^*$.
2. $\mathbb{P}(\hat{\mathcal{J}} = \mathcal{J}) \to 1$.

- Similar to Group Lasso [Bach, 2008, Theorem 2].
- Learns the weights, without explicit groups information.

[3] D: Diagonal matrix, defined using groups $\mathcal{G}_1, \ldots, \mathcal{G}_k$, and w^*

Quantitative Simulation: Predictive Accuracy

Aim: Learn $\hat{w}$ using LS-SOWL, evaluate prediction error.

[4] The experiments followed the setup of Bondell and Reich [2008]

Quantitative Simulation: Predictive Accuracy

Aim: Learn $\hat{w}$ using LS-SOWL, evaluate prediction error.
Generate samples:

- $x \sim \mathcal{N}(0, \Sigma)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$,
- $y = x^\top w^* + \epsilon$.

[4] The experiments followed the setup of Bondell and Reich [2008]

Quantitative Simulation: Predictive Accuracy

Aim: Learn $\hat{w}$ using LS-SOWL, evaluate prediction error.
Generate samples:

- $x \sim \mathcal{N}(0, \Sigma)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$,
- $y = x^\top w^* + \epsilon$.

Metric: $E[\|x^\top(w^* - \hat{w})\|_2] = (w^* - \hat{w})^\top \Sigma (w^* - \hat{w})$.

[4]The experiments followed the setup of Bondell and Reich [2008]

Quantitative Simulation: Predictive Accuracy

Aim: Learn $\hat{w}$ using LS-SOWL, evaluate prediction error.
Generate samples:

- $x \sim \mathcal{N}(0, \Sigma)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$,
- $y = x^\top w^* + \epsilon$.

Metric: $E[\|x^\top(w^* - \hat{w})\|_2] = (w^* - \hat{w})^\top \Sigma (w^* - \hat{w})$.

Data:[4] $w^* = [0_{10}^\top, 2_{10}^\top, 0_{10}^\top, 2_{10}^\top]^\top$.

- $n = 100$, $\sigma = 15$ and $\Sigma_{i,j} = 0.5$ if $i \neq j$ and 1 if $i = j$.

[4] The experiments followed the setup of Bondell and Reich [2008]

Quantitative Simulation: Predictive Accuracy

Aim: Learn $\hat{w}$ using LS-SOWL, evaluate prediction error.

Generate samples:

- $x \sim \mathcal{N}(0, \Sigma)$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$,
- $y = x^\top w^* + \epsilon$.

Metric: $E[\|x^\top(w^* - \hat{w})\|_2] = (w^* - \hat{w})^\top \Sigma (w^* - \hat{w})$.

Data:[4] $w^* = [0_{10}^\top, 2_{10}^\top, 0_{10}^\top, 2_{10}^\top]^\top$.

- $n = 100$, $\sigma = 15$ and $\Sigma_{i,j} = 0.5$ if $i \neq j$ and 1 if $i = j$.

Models with group variance:

- Measure $E[\|x^\top(\tilde{w}^* - \hat{w})\|_2]$.
  - $\tilde{w}^* = w^* + \tilde{\epsilon}$,
  - $\tilde{\epsilon} \sim \mathcal{U}[-\tau, \tau]$,
  - $\tau = 0, 0.2, 0.4$.

[4] The experiments followed the setup of Bondell and Reich [2008]

Predictive accuracy results

Algorithm	Med. MSE	MSE (10th Perc).	MSE (90th Perc)
LASSO	46.1 / 45.2 / 45.5	32.8 / 32.7 / 33.2	60.0 / 61.5 / 61.4
OWL	27.6 / 27.0 / 26.4	19.8 / 19.2 / 19.2	42.7 / 40.4 / 39.2
El. Net	30.8 / 30.7 / 30.6	21.9 / 22.6 / 23.0	42.4 / 43.0 / 41.4
$\Omega_{\mathcal{S}}$	**23.9** / **23.3** / **23.4**	**16.9** / **16.8** / **16.8**	**35.2** / **35.4** / **33.2**

Table: Each column has numbers for $\tau = 0, 0.2, 0.4$.

Summary

1. Proposed a new family of norms $\Omega_{\mathcal{S}}$.
2. Properties:
   - Equivalent to OWL in group identification.
   - Efficient computational tools
   - Equivalences to Group Lasso.
3. Illustrations on performance through simulations.

Questions ?

Thank you !!!

References I

F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

F. Bach. Structured sparsity-inducing norms through submodular functions. In *Neural Information Processing Systems*, 2010.

F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6-2-3:145–373, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, March 2009.

H. D. Bondell and B. J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, 64:115–123, 2008.

References II

P. Buhlmann, P. Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression : Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 43: 1835–1858, 2013.

A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40(1):120–145, May 2011.

M. A. T. Figueiredo and R. D. Nowak. Ordered weighted ℓ_1 regularized regression with strongly correlated covariates: Theoretical aspects. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

Mohamed Hebiri and Johannes Lederer. How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, 59 (3):1846–1854, 2013.

G. Obozinski and F. Bach. Convex relaxation for combinatorial penalties. Technical Report 00694765, HAL, 2012.

References III

R. Sankaran, F. Bach, and C. Bhattacharya. Identifying groups of strongly correlated variables through smoothed ordered weighted l_1-norms. In *Artificial Intelligence and Statistics*, pages 1123–1131, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions On Information Theory*, 55(5): 2183–2202, 2009.

X. Zeng and M. Figueiredo. The ordered weighted ℓ_1 norm: Atomic formulation, projections, and algorithms. *ArXiv preprint:1409.4271v5*, 2015.