RegML 2016
Class 6
Structured sparsity

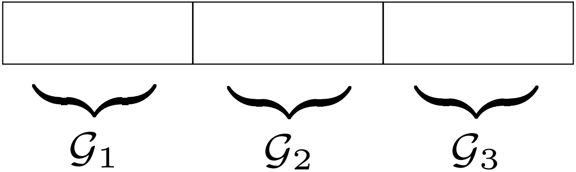
Lorenzo Rosasco
UNIGE-MIT-IIT

June 30, 2016

Exploiting structure

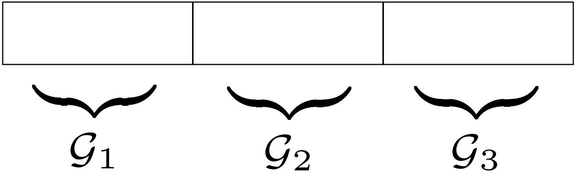
Building blocks of a function can be more structure than single variables

Sparsity



Variables divided in **non-overlapping** groups

Group sparsity



- $f(x) = \sum_{j=1}^{d} w_j x_j$
- $w = (\underbrace{w_1, \ldots,}_{w(1)} \ldots, \underbrace{\ldots, w_d}_{w(G)})$
- each group $\mathcal{G}_g$ has size $|\mathcal{G}_g|$, so $w(g) \in \mathbb{R}^{|\mathcal{G}_g|}$

Group sparsity regularization

Regularization exploiting structure

$$R_{\text{group}}(w) = \sum_{g=1}^{G} \|w(g)\| = \sum_{g=1}^{G} \sqrt{\sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^2}$$

Group sparsity regularization

Regularization exploiting structure

$$R_{\text{group}}(w) = \sum_{g=1}^{G} \|w(g)\| = \sum_{g=1}^{G} \sqrt{\sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^2}$$

Compare to

$$\sum_{g=1}^{G} \|w(g)\|^2 = \sum_{g=1}^{G} \sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^2$$

Group sparsity regularization

Regularization exploiting structure

$$R_{\text{group}}(w) = \sum_{g=1}^{G} \|w(g)\| = \sum_{g=1}^{G} \sqrt{\sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^2}$$

Compare to

$$\sum_{g=1}^{G} \|w(g)\|^2 = \sum_{g=1}^{G} \sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^2$$

or

$$\sum_{g=1}^{G} \|w(g)\|^2 = \sum_{g=1}^{G} \sum_{j=1}^{|\mathcal{G}_g|} |(w(g))_j|$$

$\ell_1 - \ell_2$ norm

We take the ℓ_2 norm of all the groups

$$(\|w(1)\|, \dots, \|w(G)\|)$$

and then the ℓ_1 norm of the above vector

$$\sum_{g=1}^{G} \|w(g)\|$$

Groups lasso

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \sum_{g=1}^{G} \|w(g)\|$$

- reduces to the Lasso if groups have cardinality one

Computations

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \underbrace{\sum_{g=1}^{G} \|w(g)\|}_{\text{non differentiable}}$$

Convex, non-smooth, but composite structure

$$w_{t+1} = \mathrm{Prox}_{\gamma\lambda R_{\text{group}}}\left(w_t - \gamma\frac{2}{n}\hat{X}^\top(\hat{X}w_t - \hat{y})\right)$$

Block thresholding

It can be shown that

$$\mathrm{Prox}_{\lambda R_{\mathrm{group}}}(w) = (\mathrm{Prox}_{\lambda\|\cdot\|}(w(1)), \dots, \mathrm{Prox}_{\lambda\|\cdot\|}(w(G))$$

$$(\mathrm{Prox}_{\lambda\|\cdot\|}(w(g)))^j = \begin{cases} w(g)^j - \lambda \frac{w(g)^j}{\|w(g)\|} & \|w(g)\| > \lambda \\ 0 & \|w(g)\| \leq \lambda \end{cases}$$

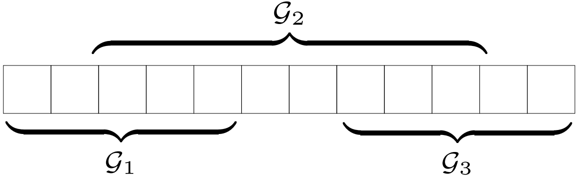
- Entire groups of coefficients set to zero!
- Reduces to softhresholding if groups have cardinality one

Other norms

$\ell_1 - \ell_p$ norms

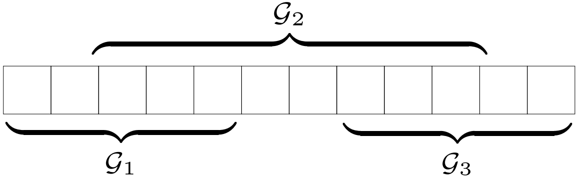
$$R(w) = \sum_{g=1}^{G} \|w(g)\|_p = \sum_{g=1}^{G} \left(\sum_{j=1}^{|\mathcal{G}_g|} (w(g))_j^p \right)^{\frac{1}{p}}$$

Overlapping groups



Variables divided in possibly **overlapping** groups

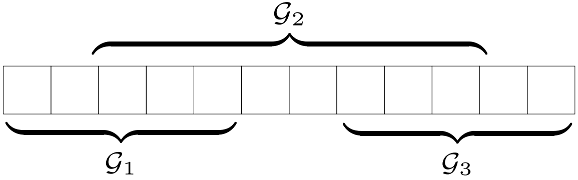
Regularization with overlapping groups



Group Lasso

$$R_{\mathrm{GL}}(w) = \sum_{g=1}^{G} \|w(g)\|$$

Regularization with overlapping groups

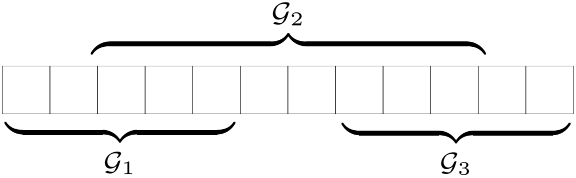


Group Lasso

$$R_{\mathrm{GL}}(w) = \sum_{g=1}^{G} \|w(g)\|$$

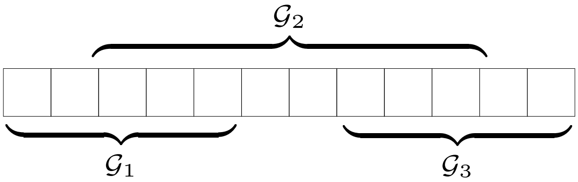
$\rightarrow$ The selected variables are **union** of group **complements**

Regularization with overlapping groups



Let $\bar{w}(g) \in \mathbb{R}^d$ be equal to $w(g)$ on group $\mathcal{G}_g$ and zero otherwise

Regularization with overlapping groups

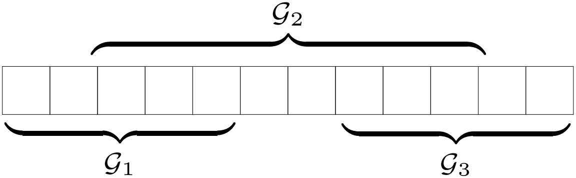


Let $\bar{w}(g) \in \mathbb{R}^d$ be equal to $w(g)$ on group $\mathcal{G}_g$ and zero otherwise

Group Lasso with overlap

$$R_{\text{GLO}}(w) = \inf \left\{ \sum_{g=1}^{G} \|w(g)\| \mid w(1), \ldots, w(g) \text{ s.t. } w = \sum_{g=1}^{G} \bar{w}(g) \right\}$$

Regularization with overlapping groups



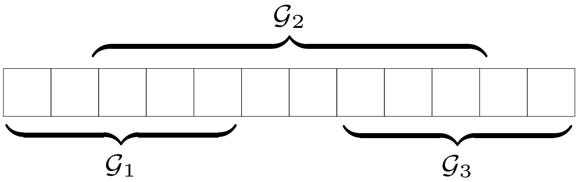
Let $\bar{w}(g) \in \mathbb{R}^d$ be equal to $w(g)$ on group $\mathcal{G}_g$ and zero otherwise

Group Lasso with overlap

$$R_{\text{GLO}}(w) = \inf \left\{ \sum_{g=1}^{G} \|w(g)\| \mid w(1), \ldots, w(g) \text{ s.t. } w = \sum_{g=1}^{G} \bar{w}(g) \right\}$$

- Multiple ways to write $w = \sum_{g=1}^{G} \bar{w}(g)$

Regularization with overlapping groups



Let $\bar{w}(g) \in \mathbb{R}^d$ be equal to $w(g)$ on group $\mathcal{G}_g$ and zero otherwise

Group Lasso with overlap

$$R_{\mathrm{GLO}}(w) = \inf \left\{ \sum_{g=1}^{G} \|w(g)\| \mid w(1), \dots, w(g) \text{ s.t. } w = \sum_{g=1}^{G} \bar{w}(g) \right\}$$

- Multiple ways to write $w = \sum_{g=1}^{G} \bar{w}(g)$
- Selected variables are groups!

An equivalence

It holds

$$\min_{w} \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda R_{\mathrm{GLO}}(w) \Leftrightarrow \min_{\tilde{w}} \frac{1}{n}\|\tilde{X}\tilde{w} - \hat{y}\|^2 + \lambda \sum_{g=1}^{G} \|w(g)\|$$

- $\tilde{X}$ is the matrix obtained by **replicating** columns/variables
- $\tilde{w} = (w(1), \dots, w(G))$, vector with (nonoverlapping!) groups

An equivalence (cont.)

Indeed

$$\min_w \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \inf_{\substack{w(1),\dots,w(g) \\ \text{s.t. } \sum_{g=1}^G \bar{w}(g)=w}} \sum_{g=1}^G \|w(g)\| \quad =$$

$$\inf_{\substack{w(1),\dots,w(g) \\ \text{s.t. } \sum_{g=1}^G \bar{w}(g)=w}} \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \sum_{g=1}^G \|w(g)\| =$$

$$\inf_{w(1),\dots,w(g)} \frac{1}{n}\|\hat{X}(\sum_{g=1}^G \bar{w}(g)) - \hat{y}\|^2 + \lambda \sum_{g=1}^G \|w(g)\| =$$

$$\inf_{w(1),\dots,w(g)} \frac{1}{n}\|\sum_{g=1}^G \hat{X}_{|\mathcal{G}_g} w(g) - \hat{y}\|^2 + \lambda \sum_{g=1}^G \|w(g)\| =$$

$$\min_{\tilde{w}} \frac{1}{n}\|\tilde{X}\tilde{w} - \hat{y}\|^2 + \lambda \sum_{g=1}^G \|w(g)\|$$

Computations

- Can use block thresholding with replicated variables $\implies$ potentially wasteful

- The proximal operator for R_{GLO} can be computed efficiently but **not** in closed form

More structure

Structured overlapping groups

- trees
- DAG
- ...

Structure can be exploited in computations...

Beyond linear models

Consider a dictionary made by union of distinct dictionaries

$$f(x) = \sum_{g=1}^{G} \underbrace{f_g(x)} = \sum_{g=1}^{G} \Phi_g(x)^\top w(g),$$

where each dictionary defines a feature map

$$\Phi_g(x) = (\phi_1^g(x), \dots, \phi_{p_g}^g(x))$$

Easy extension with usual change of variable...

Representer theorems

Let

$$f(x) = x^\top (\sum_{g=1}^{G} \bar{w}(g)) = \sum_{g=1}^{G} \bar{x}(g)^\top \bar{w}(g) = \sum_{g=1}^{G} f_g(x),$$

Representer theorems

Let

$$f(x) = x^\top (\sum_{g=1}^{G} \bar{w}(g)) = \sum_{g=1}^{G} \bar{x}(g)^\top \bar{w}(g) = \sum_{g=1}^{G} f_g(x),$$

Idea Show that

$$\bar{w}(g) = \sum_{i=1}^{n} \bar{x}(g)_i c(g)_i,$$

i.e.

$$f_g(x) = \sum_{i=1}^{n} \bar{x}(g)^\top \bar{x}(g)_i c(g)_i = \sum_{i=1}^{n} \underbrace{x(g)^\top x(g)_i}_{\Phi_g(x)^\top \Phi_g(x_i) = K_g(x, x_i)} c(g)_i$$

Representer theorems

Let

$$f(x) = x^\top (\sum_{g=1}^{G} \bar{w}(g)) = \sum_{g=1}^{G} \bar{x}(g)^\top \bar{w}(g) = \sum_{g=1}^{G} f_g(x),$$

Idea Show that

$$\bar{w}(g) = \sum_{i=1}^{n} \bar{x}(g)_i c(g)_i,$$

i.e.

$$f_g(x) = \sum_{i=1}^{n} \bar{x}(g)^\top \bar{x}(g)_i c(g)_i = \sum_{i=1}^{n} \underbrace{x(g)^\top x(g)_i}_{\Phi_g(x)^\top \Phi_g(x_i) = K_g(x, x_i)} c(g)_i$$

Note that in this case

$$\|f_g\|^2 = \|w(g)\|^2 = c(g)^\top \underbrace{\hat{X}(g)\hat{X}(g)^\top}_{\hat{K}(g)} c(g)$$

Coefficients update

$$c_{t+1} = \mathrm{Prox}_{\gamma\lambda R_{\mathrm{group}}}\left(c_t - \gamma(\hat{K}c_t - \hat{y})\right)$$

where $\hat{K} = (\hat{K}(1), \dots, \hat{K}(G))$, and $c_t = (c_t(1), \dots, c_t(G))$

Block Thresholding It can be shown that

$$(\mathrm{Prox}_{\lambda\|\cdot\|}(c(g)))^j = \begin{cases} c(g)^j - \lambda \dfrac{c(g)^j}{\underbrace{\sqrt{c(g)^\top \hat{K}(g) c(g)}}_{\|f_g\|}} & \|f_g\| > \lambda \\ 0 & \|f_g\| \leq \lambda \end{cases}$$

Non-parametric sparsity

$$f(x) = \sum_{g=1}^{G} f_g(x)$$

$$f_g(x) = \sum_{i=1}^{n} x(g)^\top x(g)_i (c(g))_i \quad \mapsto \quad f_g(x) = \sum_{i=1}^{n} K_g(x, x_i)(c(g))_i$$

$(K_1, \dots, K_G)$ family of kernels

$$\sum_{g=1}^{G} \|w(g)\| \implies \sum_{g=1}^{G} \|f_g\|_{K_g}$$

ℓ_1 MKL

$$\inf_{\substack{w(1),\dots,w(g) \\ \text{s.t. } \sum_{g=1}^{G} \bar{w}(g)=w}} \frac{1}{n}\|\hat{X}w - \hat{y}\|^2 + \lambda \sum_{g=1}^{G} \|w(g)\| =$$

$$\Downarrow$$

$$\min_{\substack{f_1,\dots,f_g \\ \text{s.t. } \sum_{g=1}^{G} f_g = f}} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \ + \ \lambda \sum_{g=1}^{G} \|f_g\|_{K_g}$$

ℓ_2 MKL

$$\sum_{g=1}^{G} \|w(g)\|^2 \implies \sum_{g=1}^{G} \|f_g\|_{K_g}^2$$

Corresponds to using the kernel

$$K(x, x') = \sum_{g=1}^{G} K_g(x, x')$$

ℓ_1 or ℓ_2 MKL

- ℓ_2 *much* faster
- ℓ_1 could be useful is only few kernels are relevant

Why MKL?

- Data fusion– different features
- Model selection, e.g. gaussian kernels with different widths
- Richer model– many kernels!

MKL & kernel learning

It can be shown that

$$\min_{\substack{f_1,\dots,f_g \\ \text{s.t.} \sum_{g=1}^G f_g = f}} \quad \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \; + \; \lambda \sum_{g=1}^G \|f_g\|_{K_g}$$

$$\Updownarrow$$

$$\min_{K\in\mathcal{K}} \min_{f\in\mathcal{H}_K} \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \; + \; \lambda \|f\|_K^2$$

where $\mathcal{K} = \{K \mid K = \sum_g K_g \alpha_g, \quad \alpha_g \geq 0\}$

Sparsity beyond vectors

Recall multi-variable regression

$$(x_i, y_i)_{i=1^n}, \quad x_i \in \mathbb{R}^d, \quad y_i \in \mathbb{R}^T$$

$$f(x) = x^\top \underbrace{W}_{d\times T}$$

$$\min_W \|\hat{X}W - \hat{Y}\|_F^2 + \lambda\, \mathbf{Tr}(WAW^\top)$$

Sparse regularization

- We have seen

$$\mathbf{Tr}(WW^\top) = \sum_{j=1}^{d} \sum_{t=1}^{T} (W_{t,j})^2$$

- We could consider now

$$\sum_{j=1}^{d} \sum_{t=1}^{T} |W_{t,j}|$$

- ...

Spectral Norms/p-Schatten norms

- We have seen

$$\mathbf{Tr}(WW^\top) = \sum_{t=1}^{\min\{d,T\}} \sigma_i^2$$

- We could consider now

$$R(W) = \|W\|_* = \sum_{t=1}^{\min\{d,T\}} \sigma_i, \qquad \text{nuclear norm}$$

- or

$$R(W) = (\sum_{t=1}^{\min\{d,T\}} (\sigma_i)^p)^{1/p}, \qquad \text{p-Schatten norm}$$

Nuclear norm regularization

$$\min_{W} \|\hat{X}W - \hat{Y}\|_F^2 + \lambda \|W\|_*$$

Computations

$$W_{t+1} = \mathrm{Prox}_{\gamma\lambda\|\cdot\|_*}\left(W_t - 2\gamma\hat{X}^\top(\widehat{X}W_t - \widehat{Y})\right)$$

Let $W = U\Sigma V^\top$, $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_p)$

$$\mathrm{Prox}_{\|\cdot\|_*}(W) = U\,\mathbf{diag}(\mathrm{Prox}_{\|\cdot\|_1}(\sigma_1, \ldots, \sigma_p))V^\top$$

This class

- Structured sparsity
- MKL
- Matrix sparsity

Next class

Data representation!