Learning the Structure for Structured Sparsity

Nino Shervashidze
joint work with Francis Bach

Journées MAS
31 August 2016

Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

Background

Context: We are interested in variable selection problems, where a small number of potentially overlapping groups of input variables explains the signal.

Examples:

- FMRI image classification (e.g., Jenatton *et al.*, 2011a).
- Multiple-loci genome-wide association studies (e.g., Azencott *et al.*, 2013).

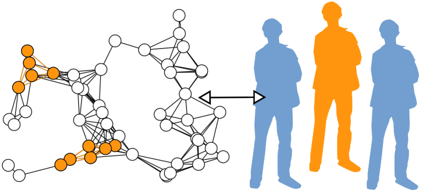


Figure by C.-A. Azencott

A standard approach: Regularization with sparsity-inducing norms.

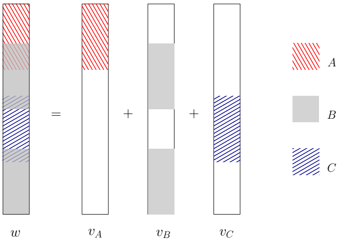
Latent group LASSO

In particular, Jacob *et al.* (2009) and Obozinski and Bach (2012) propose the norm

$$\Omega(w) = \min_{\substack{v_A \in \mathbb{R}^P, \\ \sum_{A \in \mathcal{G}} v_A = w}} \sum_{A \in \mathcal{G}} \|v_A\|_2 f(A)^{1/2},$$

where

- $\mathcal{G} \subseteq 2^{\{1,\dots,P\}}$ is the set of groups,
- P is the number of variables,
- $f(A)$ represents the prior belief in the subset A being relevant: If a group A is irrelevant, then $f(A) = +\infty$.



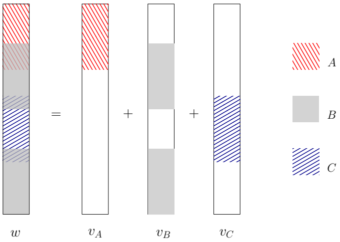
Latent group LASSO

In particular, Jacob *et al.* (2009) and Obozinski and Bach (2012) propose the norm

$$\Omega(w) = \min_{\substack{v_A \in \mathbb{R}^P, \\ \sum_{A \in \mathcal{G}} v_A = w}} \sum_{A \in \mathcal{G}} \|v_A\|_2 f(A)^{1/2},$$

where

- $\mathcal{G} \subseteq 2^{\{1,\dots,P\}}$ is the set of groups,
- P is the number of variables,
- $f(A)$ represents the prior belief in the subset A being relevant: If a group A is irrelevant, then $f(A) = +\infty$.



Goal: Learn the weights $f(A)$, unknown in practice.

Goal

Learn the set function $f : \mathcal{G} \mapsto \mathbb{R}_+ \cup \{+\infty\}$ from data
(in other words, learn the structure).

Goal

Learn the set function $f : \mathcal{G} \mapsto \mathbb{R}_+ \cup \{+\infty\}$ from data (in other words, learn the structure).

Remarks:

1. This requires a multi-task setting.
2. A relevant $\neq A$ "on" in every single task.

Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

What we build on: Sparse Bayesian learning

Our approach follows the pattern of sparse Bayesian models (Palmer *et al.*, 2006; Seeger and Nickisch, 2011, among others).

Main idea: Place a super-Gaussian sparsity prior on each component of the parameter vector and learn using variational inference.

We take these ideas two steps further:

- we propose a formulation suitable for structured sparsity with any family of groups, as opposed to classical sparsity,
- we learn the hyperparameters that are supposed to be fixed and common to all variables in existing work.

A probabilistic model for structured sparse linear regression

We consider K linear regression problems with

- design matrices $X^k \in \mathbb{R}^{N^k \times P}$,
- response vectors $y^k \in \mathbb{R}^{N^k}$,

$k \in \{1, \ldots, K\}$.

For each X^k and y^k, we assume

$$y^k \sim \mathcal{N}(X^k w^k, \sigma^2 I).$$

Example: X^k – genomes of individuals, y^k – phenotypes.

A probabilistic model for structured sparse linear regression

Let V be the set $\{1, \ldots, P\}$. For $\mathcal{G} \subseteq 2^V$, we assume

$$w^k = \sum_{A \in \mathcal{G}} v_A^k,$$

where, for each k,

- $\forall A \in \mathcal{G}, v_A^k$ is a vector in $\mathbb{R}^P$ supported on A,
- $\{v_A^k\}_{A \in \mathcal{G}}$ are jointly independent, and
- $\forall A \in \mathcal{G}, v_A^k$ has a density

$$p(v_A^k | f(A)) = q_A(\|v_A^k\|_2 f(A)^{1/2}) f(A)^{|A|/2},$$

where q_A is a zero-mean heavy-tailed distribution.

Group weights as hyperparameters of heavy-tailed priors

The inverse scale parameter of the distribution on v_A^k, $f(A)$, captures the relevance of the group A:

- The smaller $f(A)$, the more relevant the group, that is, the larger the values v_A^k is likely to take.
- Even if the group A is relevant, not all $v_A^k, k = 1, \ldots, K$ have to be large.

Link with sparsity-inducing norms

Note the resemblance between maximizing $\log p(w^k|f)$ for fixed f

$$\log p(w^k|f) = \sum_{A \in \mathcal{A}} \log q_A(\|v_A^k\|_2 f(A)^{1/2}) + \text{const}$$

and the latent group LASSO norm

$$\Omega(w^k) = \min_{\substack{v_A^k \in \mathbb{R}^P, \\ \sum_{A \in \mathcal{G}} v_A^k = w^k}} \sum_{A \in \mathcal{G}} \|v_A^k\|_2 f(A)^{1/2}.$$

When q_A is the *generalized Gaussian* density, the two expressions match exactly.

Goal: Maximize "type-II" likelihood

Find $f(A), A \in \mathcal{G}$, maximizing the likelihood

$$p(y^1, \ldots, y^K | f) = \prod_{k=1}^{K} \int p(y^k | X^k w^k, \sigma^2 I) \prod_{A \in \mathcal{G}} p(v_A^k | f(A)) dv_A^k,$$

where the v_A^k are marginalized over.

Super-Gaussian priors

We assume that q_A is a *scale mixture of Gaussians*:

$$q_A(u) = \int_0^\infty \mathcal{N}(u|0,s) r_A(s) ds.$$

Examples: Student's t, generalized Gaussian.

Why?

1. Heavy-tailed, hence suitable for modeling sparsity.
2. Amenable to variational optimization.

All Gaussian scale mixtures q_A are also super-Gaussian distributions (Palmer *et al.*, 2006):

- the logarithm of q_A is convex in u^2,
- the logarithm of q_A is non-increasing.

We can therefore write

$$\log q_A(u) = \sup_{s \geq 0} -\frac{u^2}{2s} - \phi_A(s),$$

where $\phi(s)$ is convex in $1/s$, by convex conjugacy.

Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

Inference

We use variational optimization to infer the set function f from data (building on work by Palmer *et al.* (2006) and Seeger and Nickisch (2011)).

The variational bound on the marginal likelihood $p(y|f)$ is amenable to optimization via alternating analytic updates, finding a local optimum.

The updates are equivalent to mean field updates, using the scale mixture representation of q_A (Palmer *et al.*, 2006).

Does it work?

- $K = 10,000$, $P = 1$, $X^k = 1$ for all $k \in \{1, \dots, K\}$, $\sigma^2 = 1$.
- $\mathcal{G} = \{\{1\}\}$. $y^k = w^k + \epsilon^k \in \mathbb{R}$.
- $f \in \{$14 equidistant values on the log. scale in $[0.02, 50]\}$.
- Goal: recover f.

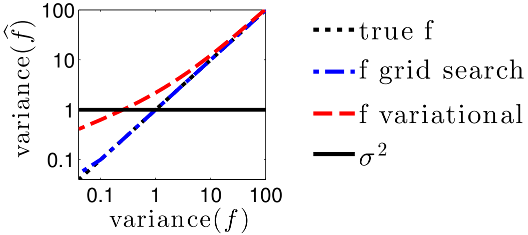


Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

Regularization

Use the improper hyperprior

$$p(f(A)) \propto f(A)^{\beta}$$

to encourage $f(A)$ to go to infinity when the variance of v_A^k is small.

The only update that changes is that for $f(A)$.

The effect of regularization

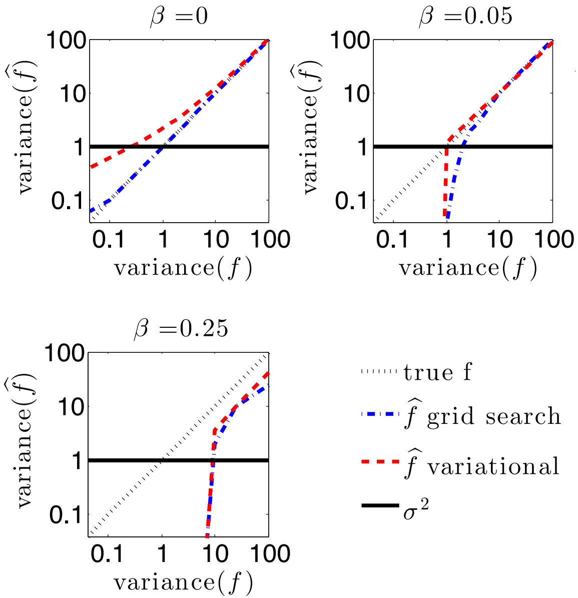


Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

Signal variance and noise variance

We measure the relevance of the group of variables A by the expectation of $\|v_A^k\|_2^2$,

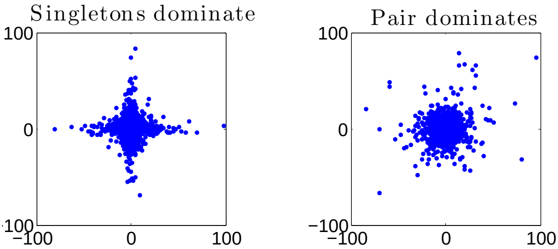
$$\mathbb{E}\left[\|v_A^k\|_2^2\right] = \frac{\mathbb{E}_{\|z\|_2 \sim q_A}\left[\|z\|_2^2\right]}{f(A)}.$$

As $\mathbb{E}\left[\|w^k\|_2^2\right] = \sum_{A \in \mathcal{A}} \mathbb{E}\left[\|v_A^k\|_2^2\right]$, $\mathbb{E}\left[\|v_A^k\|_2^2\right]$ allows us to measure the contribution of the group A w.r.t. $\mathbb{E}\left[\|w^k\|_2^2\right]$.

We call $\mathbb{E}\left[\|w^k\|_2^2\right]$ *total signal variance*, $\mathbb{E}\left[\|v_A^k\|_2^2\right]$ *signal variance coming from the group A*, and $P\sigma^2$ *total noise variance*.

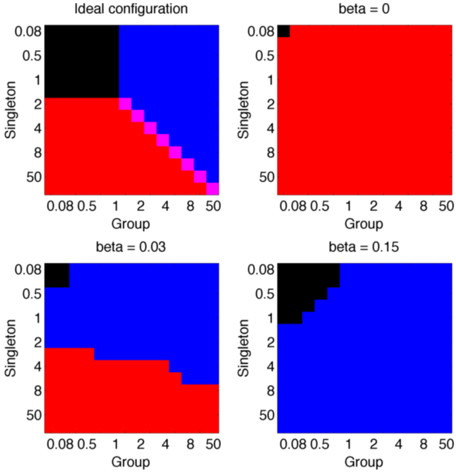
Structured sparsity with two variables

- $K = 5,000$, $P = 2$, $X^k = I$ for all $k \in \{1, \ldots, K\}$, $\mathcal{G} = \{\{1\}, \{2\}, \{1,2\}\}$, $\sigma^2 = 1$.
- Goal: recover $f(\{1\}), f(\{2\}), f(\{1,2\})$.



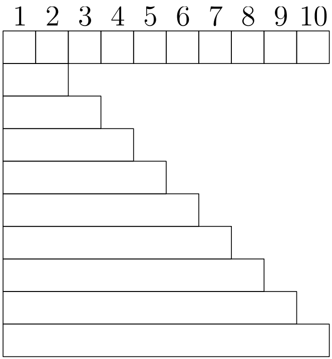
$w^k = v^k_{\{1\}} + v^k_{\{2\}}$ $\qquad$ $w^k = v^k_{\{1,2\}}$

Structured sparsity with two variables



Red – singletons dominate, blue – pair dominates.

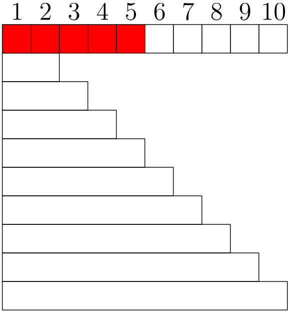
- $K = 10,000$, $P = 10$, $X^k = I$ for all $k \in \{1, \ldots, K\}$.
- $\mathcal{G} = \{\{Q\}_{Q=1,\ldots,P}, \{1, \ldots, Q\}_{Q=2,\ldots,P}\}$.
- Goal: Given $y^k, k \in 1, \ldots, K$, find the signals w^k.



Denoising with toy data: Setup

We consider three different ways of generating data:

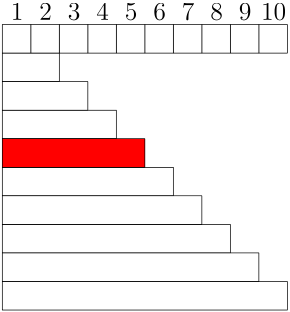
- **Singletons**: $\{1\}, \ldots, \{5\}$ relevant, all other groups irrelevant.



In all cases, σ^2 set so that the total signal variance equals the total noise variance.

We consider three different ways of generating data:

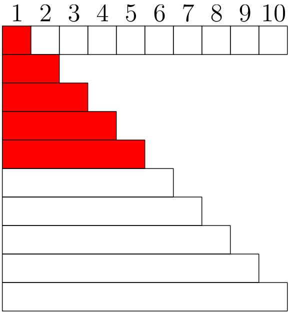
- **Singletons**: $\{1\}, \ldots, \{5\}$ relevant, all other groups irrelevant.
- **One group**: Only $\{1, 2, 3, 4, 5\}$ is relevant.



In all cases, σ^2 set so that the total signal variance equals the total noise variance.

We consider three different ways of generating data:

- **Singletons**: $\{1\}, \ldots, \{5\}$ relevant, all other groups irrelevant.
- **One group**: Only $\{1, 2, 3, 4, 5\}$ is relevant.
- **Overlapping groups**: The groups $\{1\}$, $\{1, 2\}$, ... , $\{1, 2, 3, 4, 5\}$ are relevant.

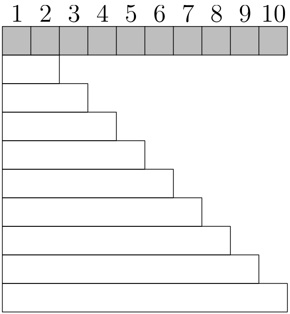


In all cases, σ^2 set so that the total signal variance equals the total noise variance.

Denoising with toy data: Setup

We consider four models for inference:

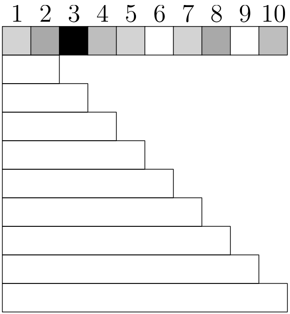
- **LASSO-like**:
  $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$, $f(A)$ constant across $\mathcal{G}$.



Goal: Given $y^k, k \in 1, \ldots, K$, find the clean signals w^k.

We consider four models for inference:

- **LASSO-like**:
  $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$, $f(A)$ constant across $\mathcal{G}$.
- **Weighted LASSO-like**:
  $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$.

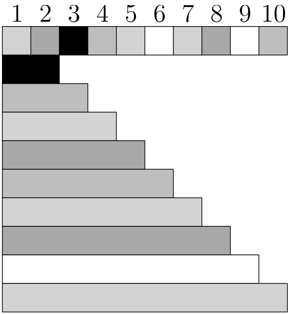


Goal: Given $y^k, k \in 1, \ldots, K$, find the clean signals w^k.

Denoising with toy data: Setup

We consider four models for inference:

- **LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$, $f(A)$ constant across $\mathcal{G}$.
- **Weighted LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$.
- **Structured**: $\mathcal{G} = \{\{Q\}_{Q=1,\ldots,P}, \{1, \ldots, Q\}_{Q=2,\ldots,P}\}$.



Goal: Given $y^k, k \in 1, \ldots, K$, find the clean signals w^k.

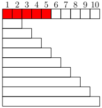
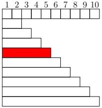
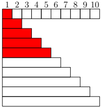
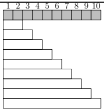
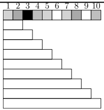
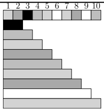
Denoising with toy data: Setup

We consider four models for inference:

- **LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$, $f(A)$ constant across $\mathcal{G}$.
- **Weighted LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$.
- **Structured**: $\mathcal{G} = \{\{Q\}_{Q=1,\ldots,P}, \{1, \ldots, Q\}_{Q=2,\ldots,P}\}$.
- **Structured(AS)**: $\mathcal{G}$ not specified in advance.

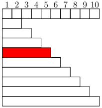
Goal: Given $y^k, k \in 1, \ldots, K$, find the clean signals w^k.

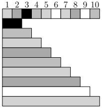
Denoising with toy data: Results

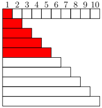
			
	18.5±0.3	18.6±0.4	58.4±1.1
	**14.5**±0.3	14.5±0.3	**42.8**±0.9
	14.8±0.3	**13.8**±0.3	43.0±0.9
Structured(AS)	14.6±0.3	14.0±0.3	**42.8**±0.9

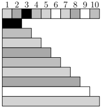
Mean squared error ± 95%-confidence error bars

Denoising with toy data: Results

Zoom into 

, 

and 

, 

:

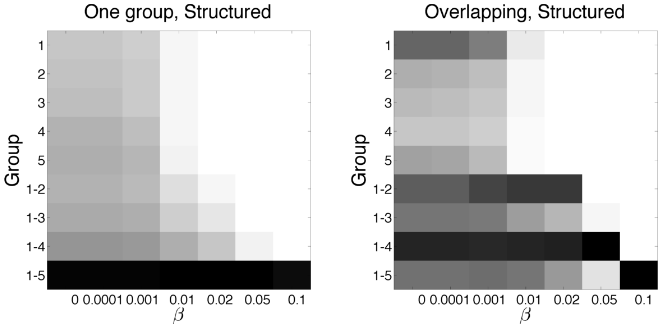


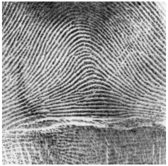
Image denoising with wavelets: Setup

- Each task is denoising a 32×32 image using wavelets ($P = 1024$).
- The Haar wavelet basis for 2-dimensional images (Mallat, 1998) can naturally be arranged in a rooted directed tree.
- We consider four models for inference:
  - **LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$, $f(A)$ constant across $\mathcal{G}$.
  - **Weighted LASSO-like**: $\mathcal{G} = \{\{1\}, \ldots, \{P\}\}$.
  - **Structured**: $\mathcal{G} = V \cup \{A | A \text{ is a path from the root in the wavelet tree.}\}$ (Jenatton *et al.* (2011b) have shown that structured sparsity-inducing norms with such groups improve over the ℓ_1 norm in this task.)
  - **Structured(AS)**: $\mathcal{G}$ not specified in advance.

Goal: Given $y^k, k \in 1, \ldots, K$, find the images w^k.

Image denoising with wavelets: Results






	Barbara	House	Fingerprint	Lena
LASSO-like	179.0±4.6 (0.001)	107.5±2.6 (0.001)	247.5±1.7 (0.005)	110.3±2.8 (0.001)
W.LASSO-like	163.3±5.1 (0)	93.7±2.6 (0)	195.0±1.8 (0.0001)	89.5±3.2 (0)
Structured	164.8±5.3 (0)	95.3±2.9 (0)	**193.6**±1.8 (0.0005)	90.3±3.5 (0)
Structured(AS)	163.1±5.0 (0.0001)	92.9±2.3 (0.0001)	194.9±1.8 (0.001)	89.5±2.8 (0.0001)
Tree-ℓ_2	**155.3**±6.4	93.3±3.8	214.9±2.4	88.7±3.7
LASSO	176.7±6.4	102.1±3.6	250.0±2.2	106.6±3.9

Table of contents

Introduction

Proposed model

Inference

Regularization

Experiments

Summary and outlook

Summary and outlook

- We propose a general model and an associated inference scheme to automatically learn group weights for structured sparse linear regression.
- We propose a regularization method that in practice circumvents the problems of the classical variational scheme for our model.
- We propose a heuristic allowing to explore a large set of groups.
- Experimental results in denoising show that learning group weights can make a difference.

Summary and outlook

- We propose a general model and an associated inference scheme to automatically learn group weights for structured sparse linear regression.
- We propose a regularization method that in practice circumvents the problems of the classical variational scheme for our model.
- We propose a heuristic allowing to explore a large set of groups.
- Experimental results in denoising show that learning group weights can make a difference.

- Other likelihood models (e.g., for y^k binary)?
- Avoid considering v_A^k explicitly (for computational efficiency)?
- GWAS application

N. Shervashidze and F. Bach. **Learning the structure for structured sparsity.** *IEEE Transactions on Signal Processing*, 63(18):4894-4902, 2015.

http://cbio.ensmp.fr/~nshervashidze/code/LLSS

Acknowledgements

Francis Bach

Guillaume Obozinski
Julien Mairal
Laurent Jacob
Sylvain Arlot

Thank you!

C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cut. *Bioinformatics*, 29(13):i171–i179, 2013.

L. Jacob, G. Obozinski, and J.-P. Vert. Group Lasso with overlap and graph Lasso. In *Proceedings of the International Conference on Machine Learning*, 2009.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824, 2011.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334, 2011.

S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

Bibliography II

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

J. A. Palmer, D. P. Wipf, K. Kreutz-Delgado, and B. D. Rao. Variational EM algorithms for non-gaussian latent variable models. In *Advances in Neural Information Processing Systems*, 2006.

M. Seeger and H. Nickisch. Large scale bayesian inference and experimental design for sparse linear models. *SIAM Journal on Imaging Sciences*, 4(1):166–199, 2011.