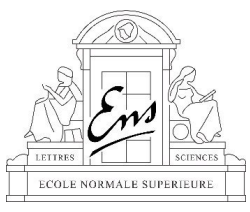
Structured Sparsity-Inducing Norms: Statistical and Algorithmic Properties with Applications to Neuroimaging

Rodolphe Jenatton
SIERRA project, INRIA Rocquencourt - Ecole Normale Supérieure

PhD defense at ENS Cachan, November 24, 2011

Advisors:
Jean-Yves Audibert
Francis Bach
Reviewers:
Laurent El Ghaoui
Massimiliano Pontil



Examinateurs:
Rémi Gribonval
Eric Moulines
Guillaume Obozinski
Bertrand Thirion

Acknowledgments


Francis Bach


Jean-Yves Audibert

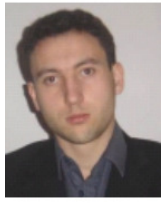

Guillaume Obozinski


Julien Mairal


Bertrand Thirion


Alexandre Gramfort


Gaël Varoqueaux


Vincent Michel


Rémi Gribonval

SIERRA/WILLOW

Sparsity

- Important concept in statistics, machine learning,...
  - Favour "simple models"
  - Easier interpretation, cheaper post-processing
  - Models with few parameters, feature selection

Sparsity

- Important concept in statistics, machine learning,...
  - Favour "simple models"
  - Easier interpretation, cheaper post-processing
  - Models with few parameters, feature selection

- Applications:
  - Compressed sensing [Candes and Tao, 2005]
  - Graphical models [Meinshausen and Bühlmann, 2006]
  - Signal/image processing tasks
    - E.g., denoising [Chen et al., 1998; Mairal, 2010]

Sparsity

- Important concept in statistics, machine learning,...
  - Favour "simple models"
  - Easier interpretation, cheaper post-processing
  - Models with few parameters, feature selection
- Applications:
  - Compressed sensing [Candes and Tao, 2005]
  - Graphical models [Meinshausen and Bühlmann, 2006]
  - Signal/image processing tasks
    - E.g., denoising [Chen et al., 1998; Mairal, 2010]
- Disregards structure:
  - Prior knowledge only about cardinality

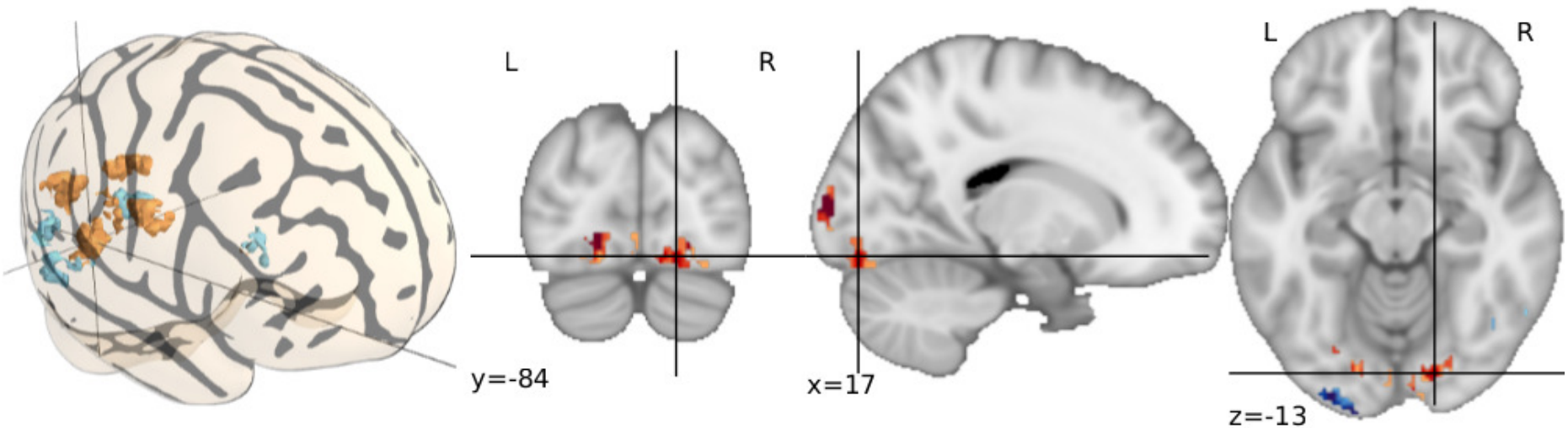
Structure? The example of neuroimaging

- In practice, know more than just cardinality!
- Sparsity comes with structural information

Structure? The example of neuroimaging

- In practice, know more than just cardinality!
- Sparsity comes with structural information

- Spatial structure in neuroimaging [Chklovskii and Koulakov, 2004; Gramfort et al., 2011]
  - Sparsity: few discriminative voxels
  - Spatiality: clusters according to the geometry of the brain



Structure? Other motivating examples

- **Spatial**:
  - Bioinformatics: contiguity due to the genome organization [Rapaport et al., 2008]
  - Image segmentation: neighboring pixels [Boykov et al., 2001]

Structure? Other motivating examples

- **Spatial**:
  - Bioinformatics: contiguity due to the genome organization [Rapaport et al., 2008]
  - Image segmentation: neighboring pixels [Boykov et al., 2001]
- **Temporal**:
  - Time series of gene expressions [Tibau Puig et al., 2011]

Structure? Other motivating examples

- **Spatial**:
  - Bioinformatics: contiguity due to the genome organization [Rapaport et al., 2008]
  - Image segmentation: neighboring pixels [Boykov et al., 2001]
- **Temporal**:
  - Time series of gene expressions [Tibau Puig et al., 2011]
- **Hierarchical**:
  - NLP: hierarchical selection of templates [Martins et al., 2011]
  - Graphical models: potential orders of interactions [Schmidt and Murphy, 2010]
  - Wavelet: zero-tree coding [Shapiro, 1993]

Structure? Other motivating examples

- **Spatial**:
  - Bioinformatics: contiguity due to the genome organization [Rapaport et al., 2008]
  - Image segmentation: neighboring pixels [Boykov et al., 2001]
- **Temporal**:
  - Time series of gene expressions [Tibau Puig et al., 2011]
- **Hierarchical**:
  - NLP: hierarchical selection of templates [Martins et al., 2011]
  - Graphical models: potential orders of interactions [Schmidt and Murphy, 2010]
  - Wavelet: zero-tree coding [Shapiro, 1993]
- **Physics of the sensors**:
  - Multipath radar signals [Rebafka et al., 2011]

Some questions addressed by the thesis

- Within a convex framework...
- How to take into account structure while inducing sparsity?
- Which structures can be considered?
- **Statistical** properties of the resulting estimators?
- Can we design efficient **algorithms** for practical problems?
- Do we observe improvements in **applications**?

Contribution 1

Topic of Part I

- New family of structured sparsity-inducing norms
- Analysis of
  - Type of structures which can be encoded
  - Automatic norm design
  - Statistical property: low- and high-dimensional consistency

Related publication:

R. Jenatton, J.-Y. Audibert, F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. In *Journal of Machine Learning Research*, 12, 2777-2824. 2011

Contribution 2

Topic of Part II

- Efficient optimization methods for structured sparsity:
  - Active-set algorithms
  - Reweighted-ℓ_2 schemes
  - Proximal-gradient techniques
- Efficient computations of proximal operator/dual norms
- Connections with operations research and network flows

Related publications:

R. Jenatton, J.-Y. Audibert, F. Bach. Structured Variable Selection with Sparsity-Inducing Norms. In *Journal of Machine Learning Research*, 12, 2777-2824. 2011

R. Jenatton, G. Obozinski, F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2010

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. 2010

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Hierarchical Sparse Coding. In *Journal of Machine Learning Research*, 12, 2297-2334. 2011

J. Mairal*, R. Jenatton*, G. Obozinski, F. Bach. Network Flow Algorithms for Structured Sparsity. Advances in Neural Information Processing Systems. 2010

J. Mairal*, R. Jenatton*, G. Obozinski, F. Bach. Convex and Network Flow Optimization for Structured Sparsity. In *Journal of Machine Learning Research*, 12, 2681-2720. 2011 (*equal contributions)

Contribution 3

Topic of Part III

- Sparse structured dictionary learning/matrix factorization:
  - Exploit previous norms for latent models
  - Learn representations adapted to structured signals
  - Flexible and rich framework
- Connections with probabilistic topic models [Blei et al., 2003]
- Applications in text/image processing and computer-vision

Related publications:

R. Jenatton, G. Obozinski, F. Bach. Structured sparse principal component analysis. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2010

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Sparse Hierarchical Dictionary Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. 2010

R. Jenatton*, J. Mairal*, G. Obozinski, F. Bach. Proximal Methods for Hierarchical Sparse Coding. In *Journal of Machine Learning Research*, 12, 2297-2334. 2011

Contribution 4

Topic of Part IV

- Applications of structured sparsity to neuroimaging
- Supervised: prediction of object sizes from fMRI signals
  - Large scale problem
  - Hierarchical norm to have multiscale representations of voxels
  - Gain in robustness for inter-subject validation
- Unsupervised: resting-state brain activity modeling
  - Sparse dictionary learning adapted to 3D structure
  - Generative model for model selection and evaluation

Related publications:

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion. Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*. 2011

R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, E. Eger, F. Bach, and B. Thirion. Multi-scale Mining of fMRI Data with Hierarchical Structured Sparsity. Preprint arXiv:1105.0363 *Submitted to SIAM Journal on Imaging Sciences*. 2011

G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, F. Bach, and B. Thirion. Sparse Structured Dictionary Learning for Brain Resting-State Activity Modeling. In *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*. 2010

Other contributions

- Characterization of the local minima of sparse coding
  - Work in progress with R. Gribonval and F. Bach
  - Non-asymptotic analysis of existence of local minima
  - Extend previous work limited to under-complete and/or noiseless signals [Gribonval and Schnass, 2010; Geng et al., 2011]

- Monographs about optimization for sparse models:

Related publications:

F. Bach, R. Jenatton, J. Mairal and G. Obozinski. Convex Optimization with Sparsity-Inducing Norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*, 2011.
F. Bach, R. Jenatton, J. Mairal and G. Obozinski. Optimization with Sparsity-Inducing Penalties. *To appear in Foundations and Trends in Machine Learning*, 2011.

- Efficient implementations for
  - Sparse structured PCA
  - Active-set algorithms
  - Available at `www.di.ens.fr/~jenatton/`

Part I
Structured sparsity-inducing norms: A guided tour

Traditional sparsity-inducing penalties

$$\min_{\mathbf{w} \in \mathbb{R}^p} \Big[f(\mathbf{w}) + \lambda \Omega(\mathbf{w}) \Big]$$

- f, usually a convex data-fitting term, e.g., square loss
- Ω, a regularization that encourages sparse solutions
- $\lambda \geq 0$, regularization parameter

Traditional sparsity-inducing penalties

$$\min_{\mathbf{w}\in\mathbb{R}^p} \Big[f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \Big]$$

- f, usually a convex data-fitting term, e.g., square loss
- Ω, a regularization that encourages sparse solutions
- $\lambda \geq 0$, regularization parameter
- Natural way to express sparsity

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_0 = \Big|\{j \in [\![1; p]\!];\ \mathbf{w}_j \neq 0\}\Big|$$

  - Leads to combinatorial problems [Natarajan, 1995]

Traditional sparsity-inducing penalties

$$\min_{\mathbf{w}\in\mathbb{R}^p} \Big[f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \Big]$$

- f, usually a convex data-fitting term, e.g., square loss
- Ω, a regularization that encourages sparse solutions
- $\lambda \geq 0$, regularization parameter
- Natural way to express sparsity

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_0 = \Big|\{j \in [\![1; p]\!];\ \mathbf{w}_j \neq 0\}\Big|$$

  - Leads to combinatorial problems [Natarajan, 1995]
- Convex relaxation via the ℓ_1-norm:

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{j=1}^{p} |\mathbf{w}_j|$$

  - For least-squares regression, Lasso [Tibshirani, 1996] and basis pursuit [Chen et al., 1998]

From plain sparsity to group sparsity

- ℓ_0 and ℓ_1: invariant w.r.t. permutations of $\mathbf{w}$
- Natural extension when variables can be grouped:

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \left[\sum_{j \in g} \mathbf{w}_j^2 \right]^{1/2} = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2, \text{ with } \mathcal{G} \text{ a partition of } [\![1; p]\!]$$

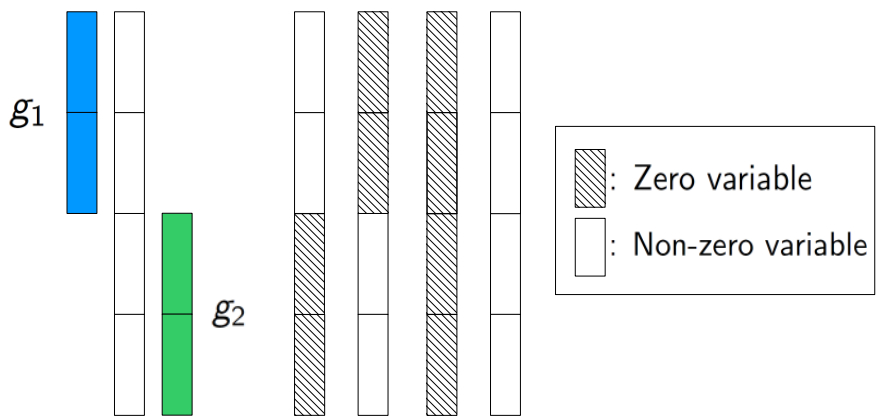
  - Mixed ℓ_1/ℓ_2-norm [Turlach et al., 2005; Yuan and Lin, 2006]
  - Can be extended to ℓ_1/ℓ_q-norms, for $q \in (1, \infty]$

From plain sparsity to group sparsity

- ℓ_0 and ℓ_1: invariant w.r.t. permutations of $\mathbf{w}$
- Natural extension when variables can be grouped:

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \Big[\sum_{j \in g} \mathbf{w}_j^2 \Big]^{1/2} = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2, \text{ with } \mathcal{G} \text{ a partition of } [\![1; p]\!]$$

  - Mixed ℓ_1/ℓ_2-norm [Turlach et al., 2005; Yuan and Lin, 2006]
  - Can be extended to ℓ_1/ℓ_q-norms, for $q \in (1, \infty]$
  - Put to zero entire groups of variables $g \in \mathcal{G}$



From plain sparsity to group sparsity

- ℓ_0 and ℓ_1: invariant w.r.t. permutations of $\mathbf{w}$
- Natural extension when variables can be grouped:

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \left[\sum_{j \in g} \mathbf{w}_j^2 \right]^{1/2} = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2, \text{ with } \mathcal{G} \text{ a partition of } [\![1; p]\!]$$

  - Mixed ℓ_1/ℓ_2-norm [Turlach et al., 2005; Yuan and Lin, 2006]
  - Can be extended to ℓ_1/ℓ_q-norms, for $q \in (1, \infty]$
  - Statistical gain when relevant prior knowledge
    - E.g., Stojnic et al. [2009]; Huang and Zhang [2010]

From plain sparsity to group sparsity

- ℓ_0 and ℓ_1: invariant w.r.t. permutations of $\mathbf{w}$
- Natural extension when variables can be grouped:

$$\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \Big[\sum_{j \in g} \mathbf{w}_j^2 \Big]^{1/2} = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2, \text{ with } \mathcal{G} \text{ a partition of } [\![1; p]\!]$$

  - Mixed ℓ_1/ℓ_2-norm [Turlach et al., 2005; Yuan and Lin, 2006]
  - Can be extended to ℓ_1/ℓ_q-norms, for $q \in (1, \infty]$
  - Statistical gain when relevant prior knowledge
    - E.g., Stojnic et al. [2009]; Huang and Zhang [2010]
  - Applications:
    - Encoding categorical variables [Roth and Fischer, 2008]
    - Multitask learning [Obozinski et al., 2009]

Overlapping groups of variables

- **Main idea:** $\mathcal{G}$ not a partition anymore
  - Ω is still a norm, $\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2$

- *Intuitively*, same behavior:
  - Some groups $g \in \mathcal{G}$ will be zeroed out
  - Sets of possible zero patterns and possible non-zero patterns?

Overlapping groups of variables

- **Main idea:** $\mathcal{G}$ not a partition anymore
  - Ω is still a norm, $\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2$

- *Intuitively*, same behavior:
  - Some groups $g \in \mathcal{G}$ will be zeroed out
  - Sets of possible zero patterns and possible non-zero patterns?

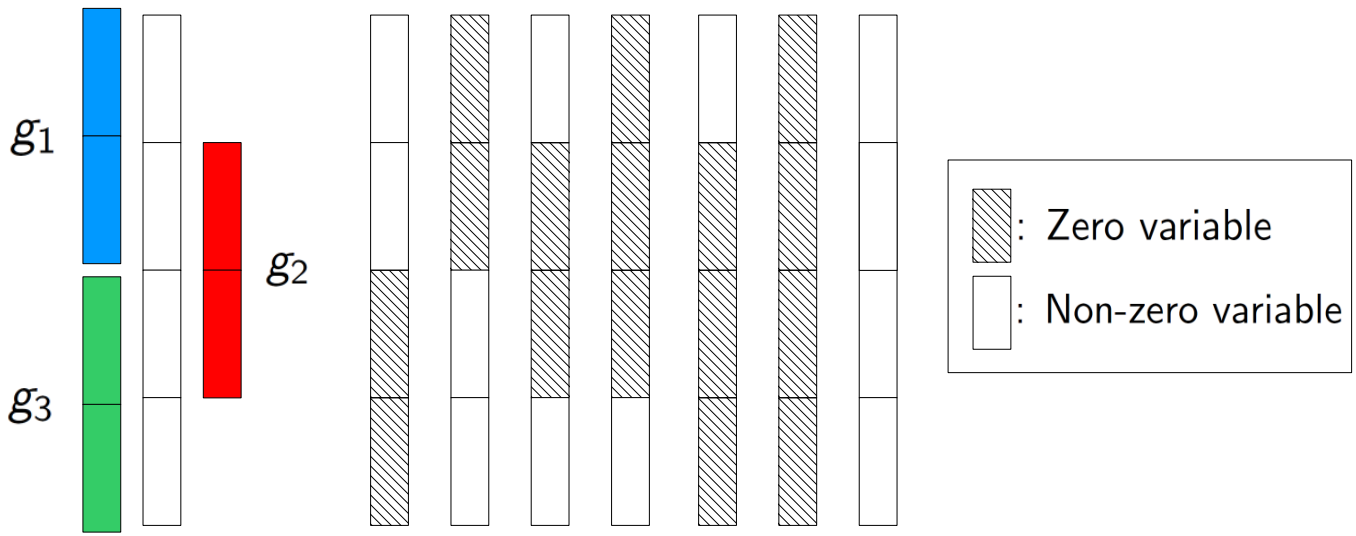
$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$

Overlapping groups of variables

- **Main idea:** $\mathcal{G}$ not a partition anymore
  - Ω is still a norm, $\Omega(\mathbf{w}) = \sum_{g \in \mathcal{G}} \|\mathbf{w}_g\|_2$

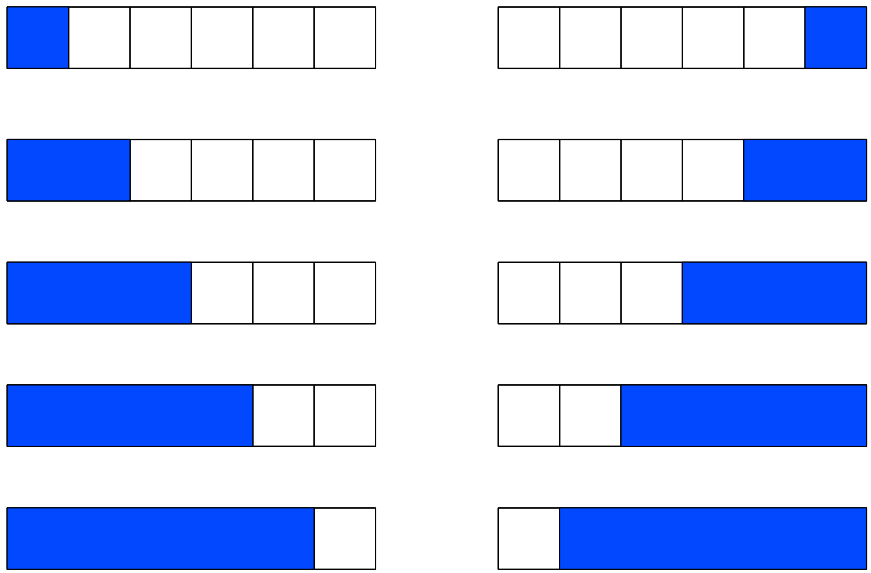
- *Intuitively*, same behavior:
  - Some groups $g \in \mathcal{G}$ will be zeroed out
  - Sets of possible zero patterns and possible non-zero patterns?

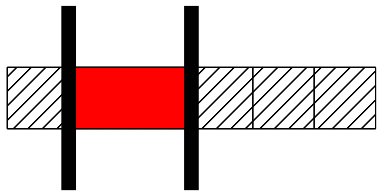
$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$



$$\mathcal{G} = \{g_1, g_2, g_3\}$$

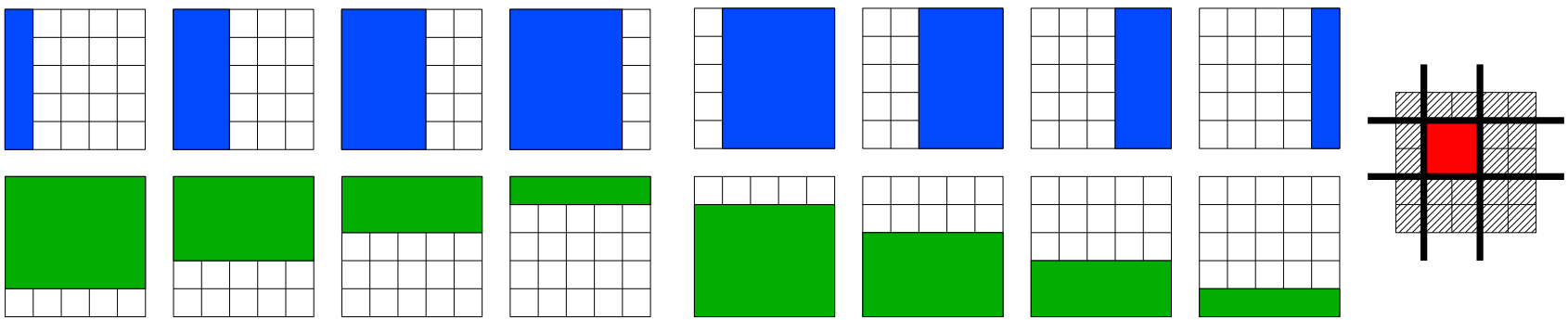
Example: Selection of contiguous patterns





The nonzero patterns are all the segments

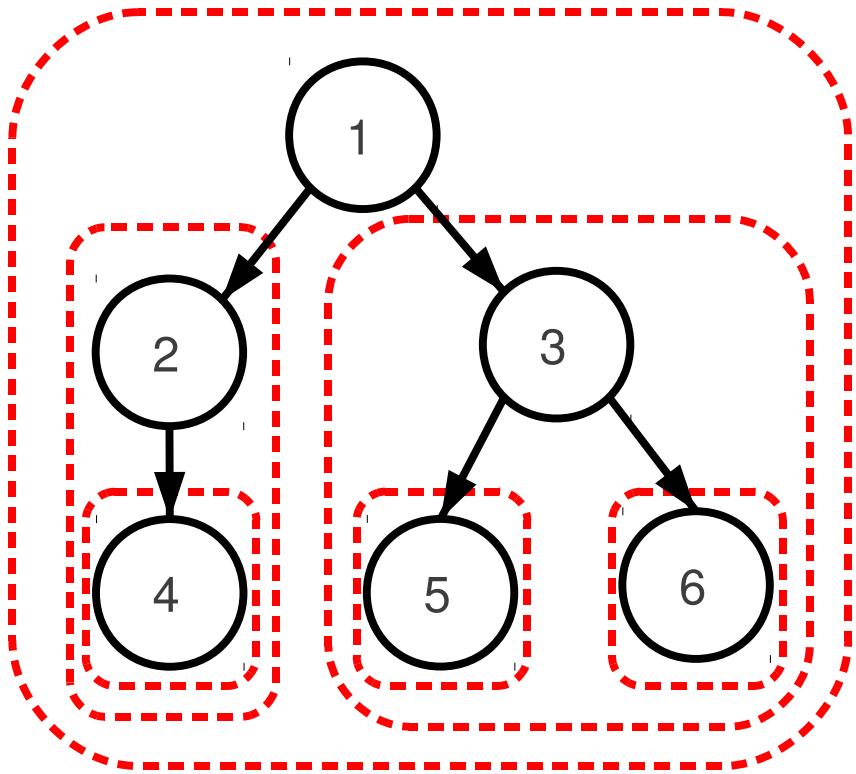
Example: Selection of rectangles



The nonzero patterns are all the rectangles

Example: Hierarchical structure

[Zhao et al., 2009]



- Groups in $\mathcal{G}$ are all the rooted subtrees
- **Selection rule**:
  - If a node is selected, the same goes for all its ancestors
  - If a node is not selected, then its descendants are not selected

Characterization of induced patterns

$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$

- Set of zero patterns $\mathcal{Z}$ closed under union
- Set of nonzero patterns $\mathcal{N}$ closed under intersection

Characterization of induced patterns

$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$

- Set of zero patterns $\mathcal{Z}$ closed under union
- Set of nonzero patterns $\mathcal{N}$ closed under intersection

- Relation between $\mathcal{G}$ and $\mathcal{Z}$, $\mathcal{N}$?

Characterization of induced patterns

$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$

- Set of zero patterns $\mathcal{Z}$ closed under union
- Set of nonzero patterns $\mathcal{N}$ closed under intersection

- Relation between $\mathcal{G}$ and $\mathcal{Z}$, $\mathcal{N}$?
- From $\mathcal{G}$ to $\mathcal{Z}$:
  - Union closure of $\mathcal{G}$

Characterization of induced patterns

$$\mathcal{Z} = \left\{ \bigcup_{g \in \mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\} \quad \text{and} \quad \mathcal{N} = \left\{ \bigcap_{g \in \mathcal{G}'} g^c;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$$

- Set of zero patterns $\mathcal{Z}$ closed under union
- Set of nonzero patterns $\mathcal{N}$ closed under intersection

- Relation between $\mathcal{G}$ and $\mathcal{Z}$, $\mathcal{N}$?
- From $\mathcal{G}$ to $\mathcal{Z}$:
  - Union closure of $\mathcal{G}$
- From (union-closed) $\mathcal{Z}$ to (minimal) $\mathcal{G}$:
  - Procedure from set theory [Doignon and Falmagne, 1998]
  - Automatic norm design

Some theoretical properties

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda \Omega(\mathbf{w}) \right] \quad (*)$$

- Supervised setting
  - $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - loss function l, convex, C^2

Some theoretical properties

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda \Omega(\mathbf{w}) \right] \quad (*)$$

- Supervised setting
  - $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - loss function l, convex, C^2
- Zero patterns:
  - Solutions of $(*)$ have zero patterns in $\mathcal{Z} = \left\{ \bigcup_{g\in\mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$

Some theoretical properties

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda \Omega(\mathbf{w}) \right] \quad (*)$$

- Supervised setting
  - $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - loss function l, convex, C^2
- Zero patterns:
  - Solutions of $(*)$ have zero patterns in $\mathcal{Z} = \left\{ \bigcup_{g\in\mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\}$
- Support recovery for least-squares regression:
  - For ℓ_1 [Zhao and Yu, 2006] and group Lasso [Bach, 2008]
  - Low- and high-dimensional settings

Related approaches to structured sparsity

- **Convex approaches**
  - Union-closed nonzero patterns [Jacob et al., 2009]
  - Via convex cones [Micchelli et al., 2010]
- **Submodular approaches**
  - Set-functions based on supports and level-sets [Bach, 2010a,b]
- **Non-convex approaches**
  - Information-theoretic criterion [Haupt and Nowak, 2006; Huang et al., 2009]
  - Union of supports [Baraniuk et al., 2010]
- **Bayesian approaches**
  - Hierarchical wavelet model [He and Carin, 2009]

Part II
Optimization for sparsity-inducing norms: The case of tree-structured groups

Which optimization tools?

$$\min_{\mathbf{w}\in\mathbb{R}^p} \Big[f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \Big]$$

- f convex, differentiable, Lipschitz-continuous gradient
- Ω convex, but nonsmooth

- Techniques blind to composite structure:
  - Subgradient descent
  - Interior-point for QP, SOCP

Which optimization tools?

$$\min_{\mathbf{w}\in\mathbb{R}^p} \Big[f(\mathbf{w}) + \lambda\Omega(\mathbf{w}) \Big]$$

- f convex, differentiable, Lipschitz-continuous gradient
- Ω convex, but nonsmooth

- Techniques blind to composite structure:
  - Subgradient descent
  - Interior-point for QP, SOCP

- Techniques developed in the thesis:
  - Active-set strategies
  - Reweighted-ℓ_2 schemes
  - Proximal methods

Proximal methods

- Forward-backward splitting [Combettes and Pesquet, 2010]
- At iteration k, update rule

$$\mathbf{w}^{(k+1)} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left[f(\mathbf{w}^{(k)}) + \nabla f(\mathbf{w}^{(k)})^\top (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w}^{(k)} - \mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

$$\leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{2}\left\| \mathbf{w} - \left(\mathbf{w}^{(k)} - \frac{1}{L}\nabla f(\mathbf{w}^{(k)})\right)\right\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

  - Linearization of f at $\mathbf{w}^{(k)}$
  - *Proximity* term, remain around $\mathbf{w}^{(k)}$

Proximal methods

- Forward-backward splitting [Combettes and Pesquet, 2010]
- At iteration k, update rule

$$\mathbf{w}^{(k+1)} \leftarrow \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[f(\mathbf{w}^{(k)}) + \nabla f(\mathbf{w}^{(k)})^\top (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w}^{(k)} - \mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

$$\leftarrow \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\left\| \mathbf{w} - (\mathbf{w}^{(k)} - \frac{1}{L}\nabla f(\mathbf{w}^{(k)})) \right\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

  - Linearization of f at $\mathbf{w}^{(k)}$
  - *Proximity* term, remain around $\mathbf{w}^{(k)}$
- If $\lambda = 0$, gradient update rule: $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \frac{1}{L}\nabla f(\mathbf{w}^{(k)})$

Proximal methods

- Forward-backward splitting [Combettes and Pesquet, 2010]
- At iteration k, update rule

$$\mathbf{w}^{(k+1)} \leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left[f(\mathbf{w}^{(k)}) + \nabla f(\mathbf{w}^{(k)})^\top (\mathbf{w} - \mathbf{w}^{(k)}) + \frac{L}{2}\|\mathbf{w}^{(k)} - \mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

$$\leftarrow \arg\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{2} \left\| \mathbf{w} - (\mathbf{w}^{(k)} - \frac{1}{L}\nabla f(\mathbf{w}^{(k)})) \right\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

  - Linearization of f at $\mathbf{w}^{(k)}$
  - *Proximity* term, remain around $\mathbf{w}^{(k)}$
- If $\lambda = 0$, gradient update rule: $\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \frac{1}{L}\nabla f(\mathbf{w}^{(k)})$
- There are also accelerated versions [Beck and Teboulle, 2009; Nesterov, 2007]
  - Linear combination of past iterates
  - Optimal convergence rates $O(\frac{1}{k^2})$ [Nesterov, 2004]

Key sub-problem: Proximal operator

- **Idea:** Ω only involved through its proximal operator:

$$\text{Prox}_{\lambda\Omega} : \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda\Omega(\mathbf{w})\right]$$

- Has to be solved efficiently and exactly [or with arbitrary precision, Schmidt et al., 2011]

Key sub-problem: Proximal operator

- **Idea:** Ω only involved through its proximal operator:

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda\Omega(\mathbf{w})\right]$$

- Has to be solved efficiently and exactly [or with arbitrary precision, Schmidt et al., 2011]
- For ℓ_1-norm, soft-thresholding [Donoho and Johnstone, 1995]:

$$\left[\text{Prox}_{\lambda\|.\|_1}(\mathbf{u})\right]_j = \text{sign}(\mathbf{u}_j)\max\{0, |\mathbf{u}_j| - \lambda\}$$

Key sub-problem: Proximal operator

- **Idea:** Ω only involved through its proximal operator:

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda\Omega(\mathbf{w})\right]$$

- Has to be solved efficiently and exactly [or with arbitrary precision, Schmidt et al., 2011]
- For ℓ_1-norm, soft-thresholding [Donoho and Johnstone, 1995]:

$$\left[\text{Prox}_{\lambda\|.\|_1}(\mathbf{u})\right]_j = \text{sign}(\mathbf{u}_j)\max\{0, |\mathbf{u}_j| - \lambda\}$$

- Also simple expressions for ℓ_2- and ℓ_∞-norms

Key sub-problem: Proximal operator

- **Idea:** Ω only involved through its proximal operator:

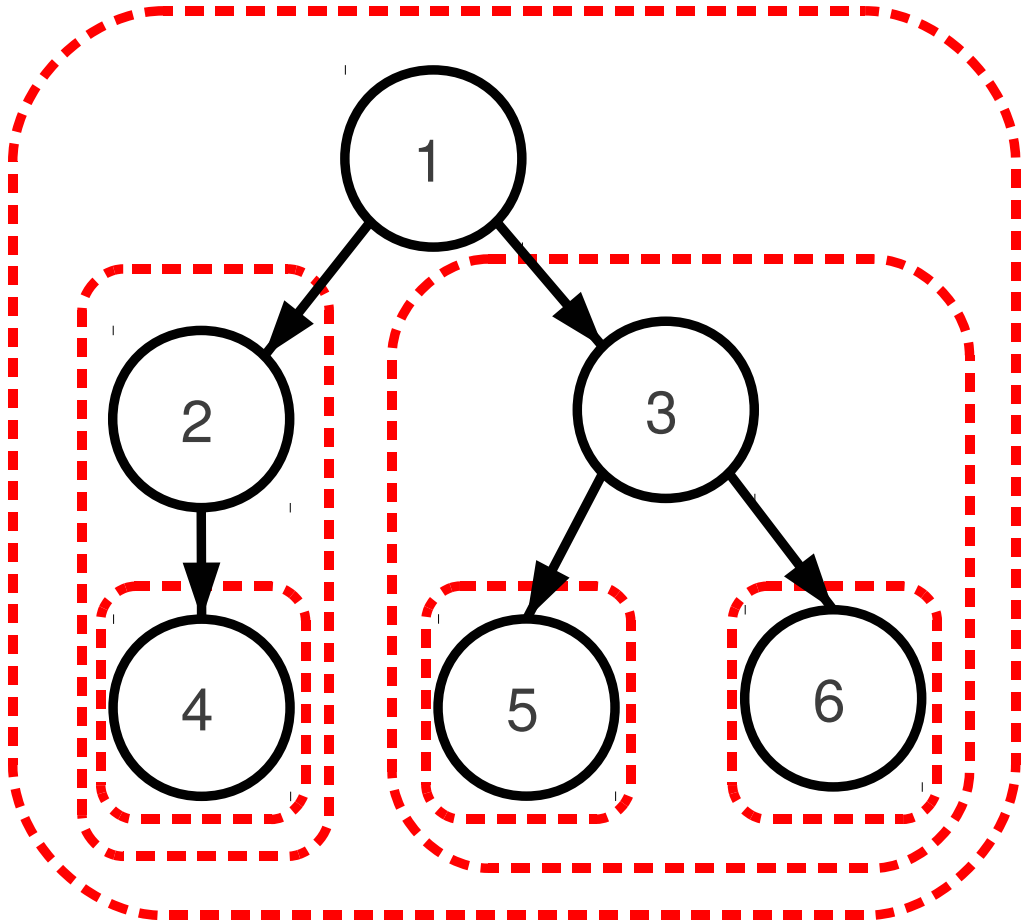
$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda\Omega(\mathbf{w}) \right]$$

- Has to be solved efficiently and exactly [or with arbitrary precision, Schmidt et al., 2011]
- For ℓ_1-norm, soft-thresholding [Donoho and Johnstone, 1995]:

$$\left[\text{Prox}_{\lambda\|.\|_1}(\mathbf{u})\right]_j = \text{sign}(\mathbf{u}_j)\max\{0, |\mathbf{u}_j| - \lambda\}$$

- Also simple expressions for ℓ_2- and ℓ_∞-norms
- When Ω has overlapping groups, no closed-form available...

Tree-structured set of groups



- Tree-structured set of groups $\mathcal{G}$:

$$\text{For any } g, h \in \mathcal{G}, \quad \Big[g \cap h \neq \varnothing \Big] \Rightarrow \Big[g \subseteq h \text{ or } h \subseteq g \Big]$$

Proximal operator for tree-structured $\mathcal{G}$

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q \right] \quad \text{for } q \in \{2,\infty\}$$

Proximal operator for tree-structured $\mathcal{G}$

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

- Consider

$$\text{Prox}_{g} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

Proximal operator for tree-structured $\mathcal{G}$

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

- Consider

$$\text{Prox}_{g} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

- Assume $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ ordered according to

$$g \preceq h \Rightarrow \Big[g \subseteq h \text{ or } g \cap h = \varnothing\Big] \quad (\text{“leaves up to the root”})$$

Proximal operator for tree-structured $\mathcal{G}$

$$\text{Prox}_{\lambda\Omega} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \sum_{g\in\mathcal{G}} \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

- Consider

$$\text{Prox}_{g} : \ \mathbf{u} \mapsto \underset{\mathbf{w}\in\mathbb{R}^p}{\arg\min} \left[\frac{1}{2}\|\mathbf{w}-\mathbf{u}\|_2^2 + \lambda \|\mathbf{w}_g\|_q\right] \quad \text{for } q \in \{2,\infty\}$$

- Assume $\mathcal{G} = \{g_1, \ldots, g_{|\mathcal{G}|}\}$ ordered according to

$$g \preceq h \Rightarrow \Big[g \subseteq h \text{ or } g \cap h = \varnothing\Big] \quad (\text{“leaves up to the root”})$$

Proposition (composition of prox)

$$Prox_{\lambda\Omega} = Prox_{g_{|\mathcal{G}|}} \circ \cdots \circ Prox_{g_1}$$

With careful implementation, complexity linear $O(p)$ when $q = 2$.

Extensions

- General overlapping groups for ℓ_1/ℓ_∞-norms
  - Mairal, Jenatton, Obozinski, and Bach [2011]
  - Connections with network flow problems
  - Conic duality + divide-and-conquer strategy

Part III
Some applications to dictionary learning

Dictionary learning in a nutshell

- **Data**: n signals, $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{m \times n}$
- **Dictionary**: p atoms, $\mathbf{D} = [\mathbf{d}^1, \ldots, \mathbf{d}^p] \in \mathbb{R}^{m \times p}$
- **Decomposition**: $\mathbf{A} = [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^n] \in \mathbb{R}^{p \times n}$
- **Goal:**

$$\mathbf{X} \approx \mathbf{DA}$$

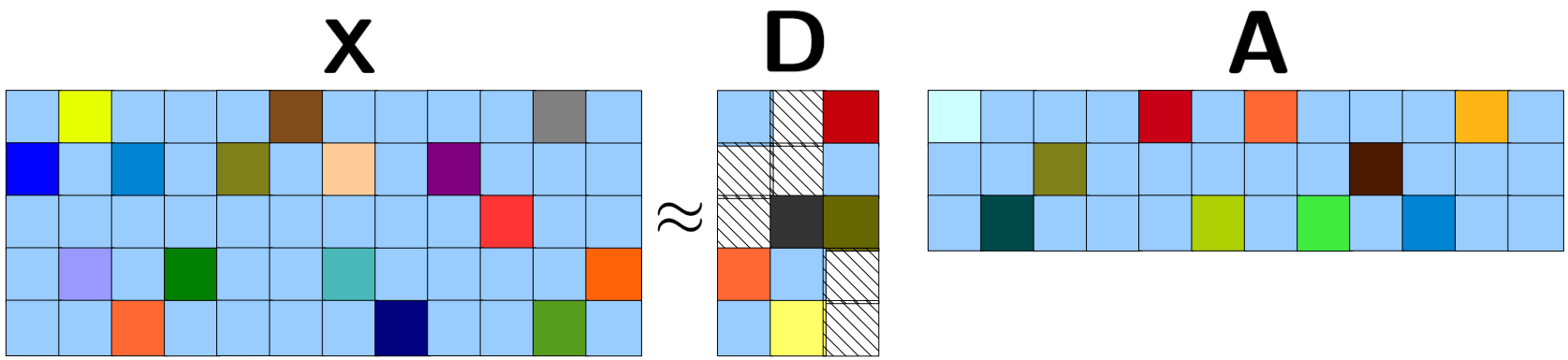
- Applied to various settings and classes of signals
  - Neuroscience [Olshausen and Field, 1997]
  - Image processing/Computer vision [Elad and Aharon, 2006; Peyré, 2009; Mairal, 2010]
  - Audio processing [Sprechmann et al., 2010; Lefèvre et al., 2011]
  - Topic modeling [Jenatton et al., 2011]
  - ...

Dictionary learning and structured sparsity

- Mainly, two settings

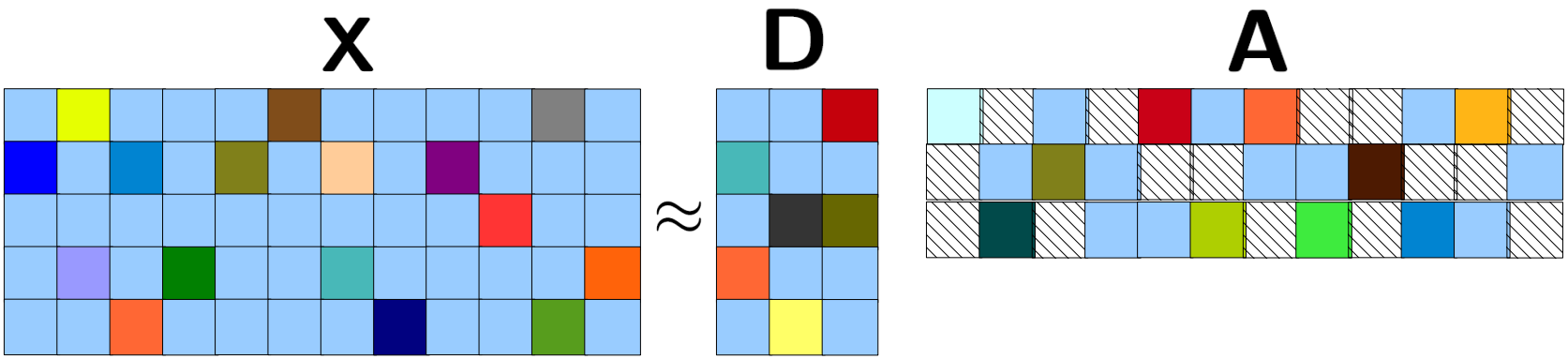
Dictionary learning and structured sparsity

- Mainly, two settings
- Regularization over $\mathbf{D}$: $\Omega\left(\mathbf{d}^j\right)$, for $j \in [\![1; p]\!]$
  - Sparse structured dictionary elements



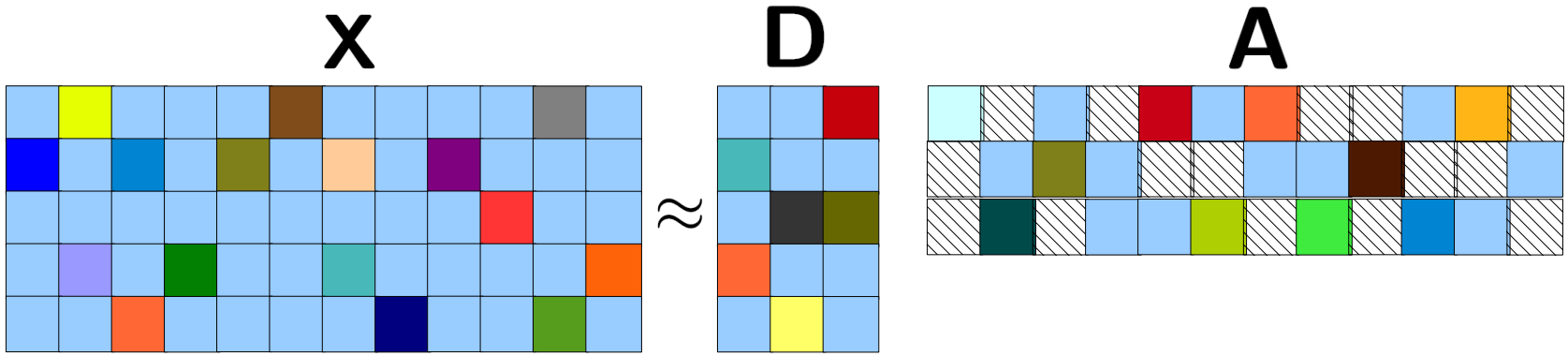
Dictionary learning and structured sparsity

- Regularization over **A**: $\Omega(\boldsymbol{\alpha}^i)$, for $i \in [\![1; n]\!]$
  - Each signal uses few atoms, sparse decomposition
  - Induced effect on **D**
  - Atoms organize and adapt to match the structure of Ω



Dictionary learning and structured sparsity

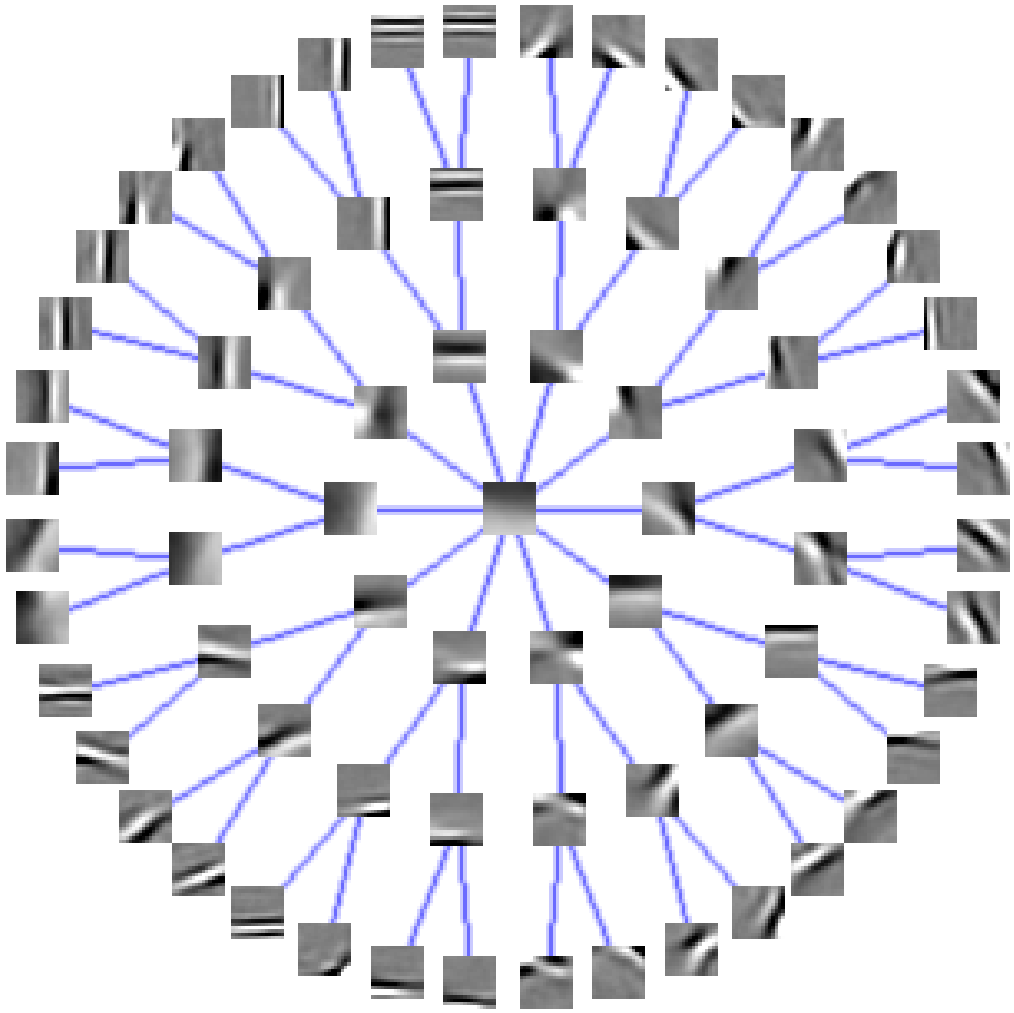
- Regularization over **A**: $\Omega(\boldsymbol{\alpha}^i)$, for $i \in [\![1; n]\!]$
  - Each signal uses few atoms, sparse decomposition
  - Induced effect on **D**
  - Atoms organize and adapt to match the structure of Ω



Example, the hierarchical case:

$$\min_{\mathbf{A}\in\mathbb{R}^{p\times n},\mathbf{D}\in\mathcal{D}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|\mathbf{x}^i - \mathbf{D}\boldsymbol{\alpha}^i\|_2^2 + \lambda\Omega_{\text{hierarchical}}(\boldsymbol{\alpha}^i)\right]$$

Learned dictionary of natural image patches



- Graph structure imposed by Ω
- Each patch at a node represents an atom

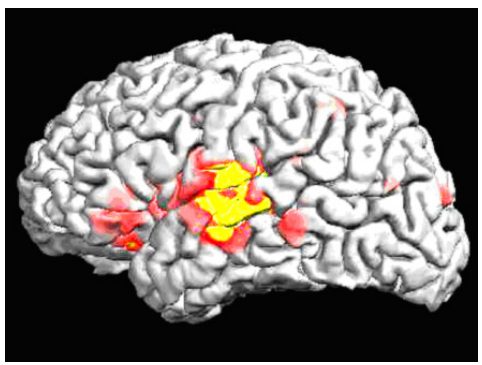
Part IV
Application to neuroimaging

Structured sparsity for neuroimaging

- Supervised problem
- $\mathbf{x} \in \mathbb{R}^p$ fMRI signal, with $p \approx 70\,000$ voxels
- y sizes of objects
- **Goal**: Predict y from brain activation-map $\mathbf{x}$, *across subjects*

Size y of the object?

fMRI signal $\mathbf{x}$



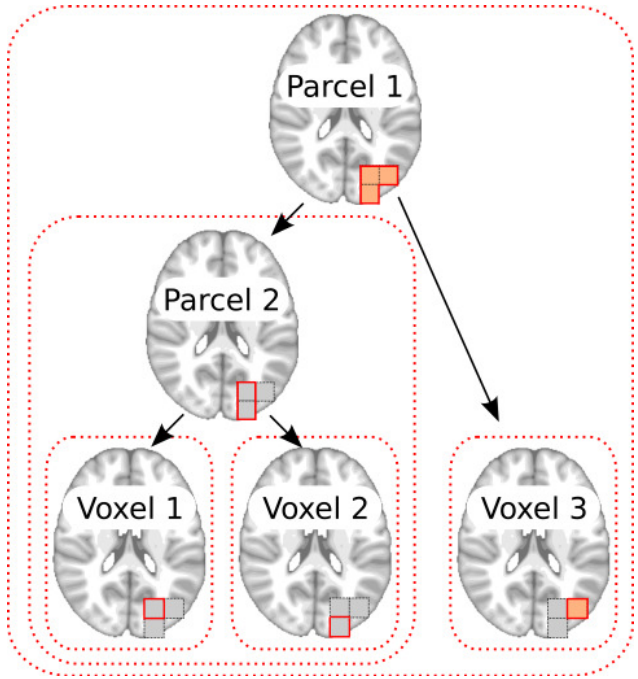
Predict $\longrightarrow$

Structured sparsity for neuroimaging

- Supervised problem
- $\mathbf{x} \in \mathbb{R}^p$ fMRI signal, with $p \approx 70\,000$ voxels
- y sizes of objects
- **Goal**: Predict y from brain activation-map $\mathbf{x}$, *across subjects*
- Some properties:
  - High-dimensional problem, only $n \approx 100$
  - A few voxels are useful to predict y (sparsity)
  - Relevant voxels appear as spatially-localized patterns
  - Across subjects, voxel-based predictions not reliable:
    - Misalignement issue

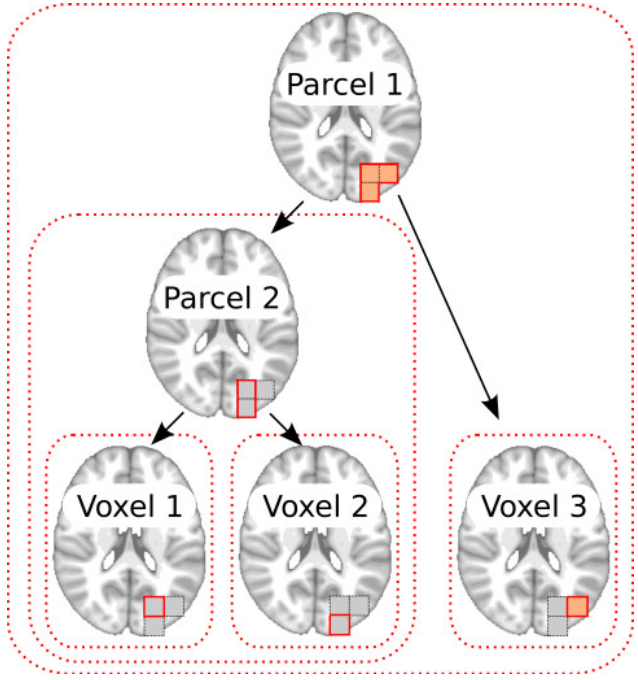
Proposed approach

- Instead of single voxels, consider local averages: parcels



Proposed approach

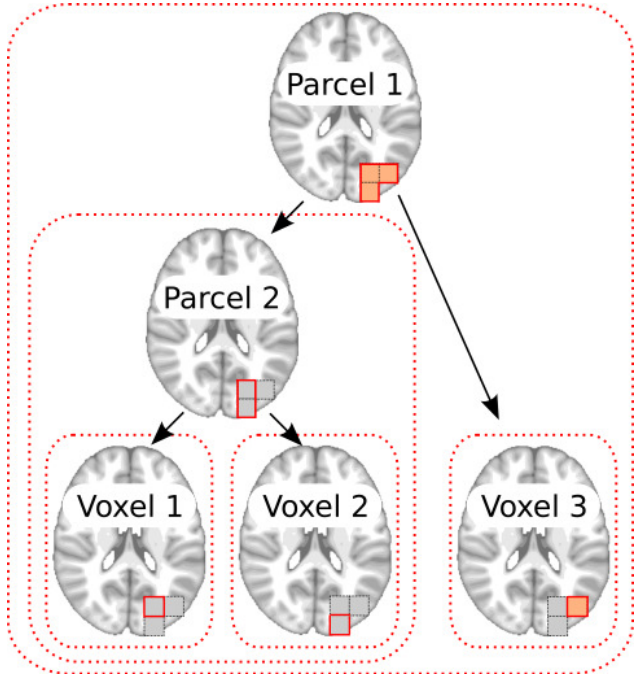
- Instead of single voxels, consider local averages: parcels



- Build a multi-scale representation of the data:

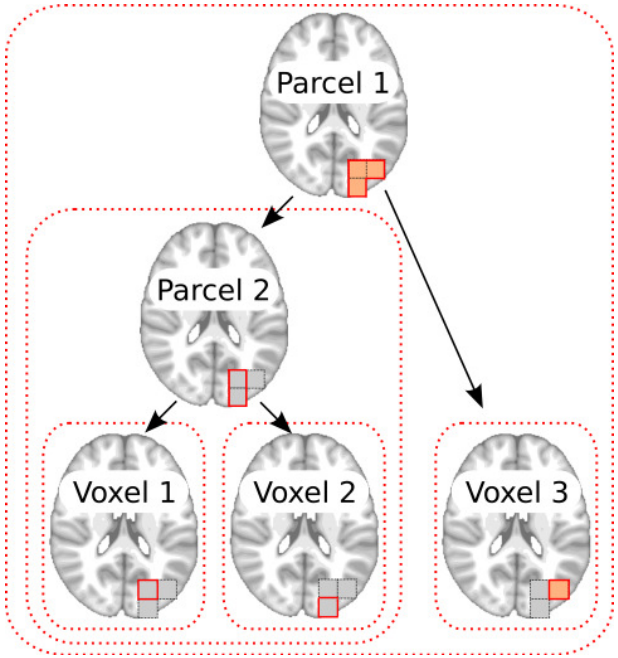
Proposed approach

- Instead of single voxels, consider local averages: parcels



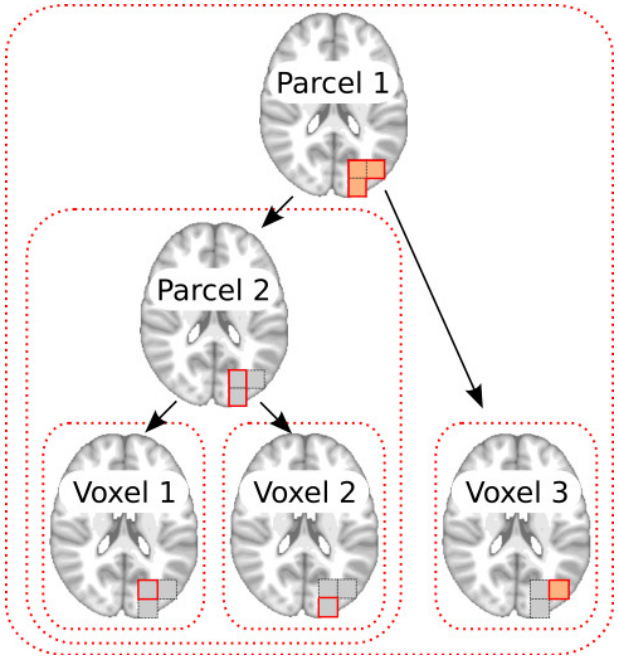
- Build a multi-scale representation of the data:
  - Ward's hierarchical clustering + spatial constraints
  - Leaves of the tree: single voxels
  - Root of the tree: all voxels

The new representation of the signals



- Original signal
  $\mathbf{x}_{\text{old}} = [\mathbf{x}_{\text{voxel1}}, \mathbf{x}_{\text{voxel2}}, \mathbf{x}_{\text{voxel3}}] \in \mathbb{R}^3$

The new representation of the signals



- Original signal
  $\mathbf{x}_{\text{old}} = [\mathbf{x}_{\text{voxel1}}, \mathbf{x}_{\text{voxel2}}, \mathbf{x}_{\text{voxel3}}] \in \mathbb{R}^3$
- New signal representation

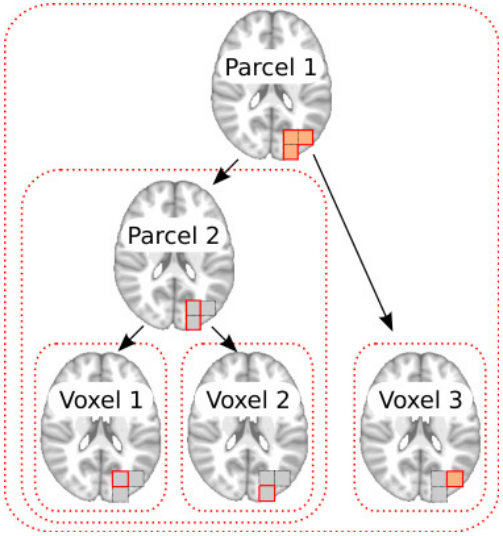
$$\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$$

with

$$\begin{cases} \mathbf{x}_{\text{parcel1}} = \text{mean}(\mathbf{x}_{\text{voxel1}}, \mathbf{x}_{\text{voxel2}}, \mathbf{x}_{\text{voxel3}}) \\ \mathbf{x}_{\text{parcel2}} = \text{mean}(\mathbf{x}_{\text{voxel1}}, \mathbf{x}_{\text{voxel2}}) \end{cases}$$

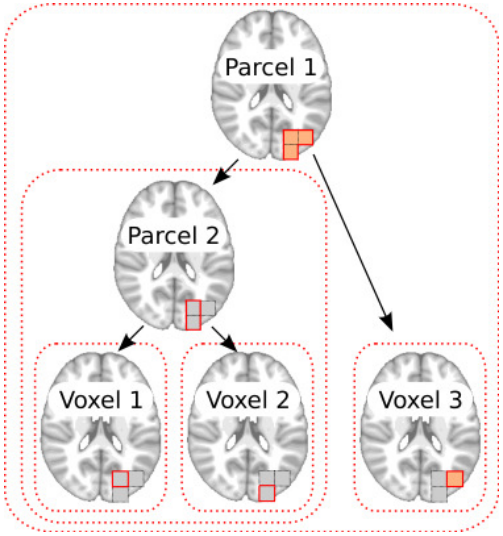
Reminiscent of spatial pyramid from Lazebnik et al. [2006]

How to exploit the new representation?



- $\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$
- Assume linear models: $\mathbf{w}^\top \mathbf{x}_{\text{new}}$
- **Idea**: The closer to the leaves, the less reliable across subjects
- First select $\mathbf{w}_j$ near the root

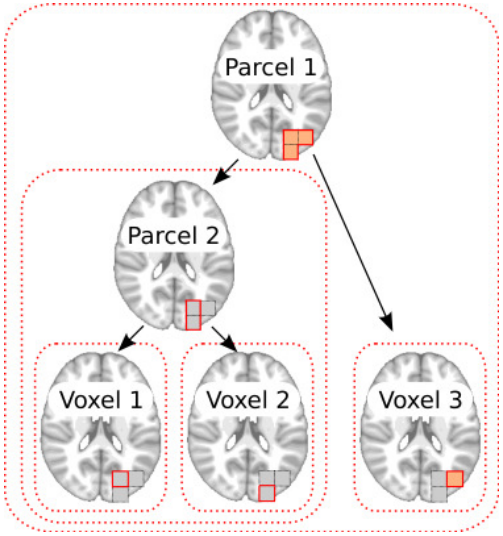
How to exploit the new representation?



- $\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$
- Assume linear models: $\mathbf{w}^\top \mathbf{x}_{\text{new}}$
- **Idea**: The closer to the leaves, the less reliable across subjects
- First select $\mathbf{w}_j$ near the root

- ℓ_1-norm would not distinguish between parcels/voxels

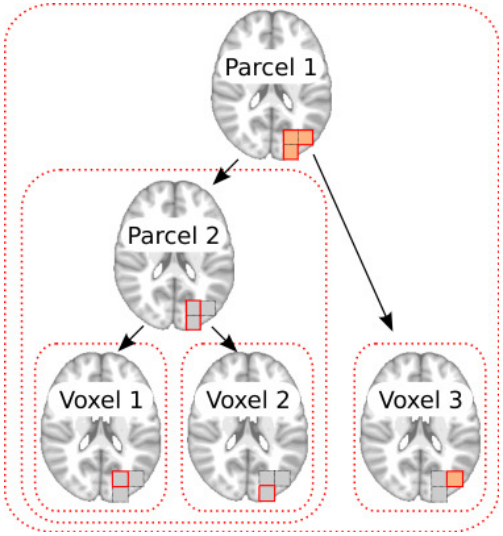
How to exploit the new representation?



- $\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$
- Assume linear models: $\mathbf{w}^\top \mathbf{x}_{\text{new}}$
- **Idea**: The closer to the leaves, the less reliable across subjects
- First select $\mathbf{w}_j$ near the root

- ℓ_1-norm would not distinguish between parcels/voxels
- We use a hierarchical sparsity-inducing norm

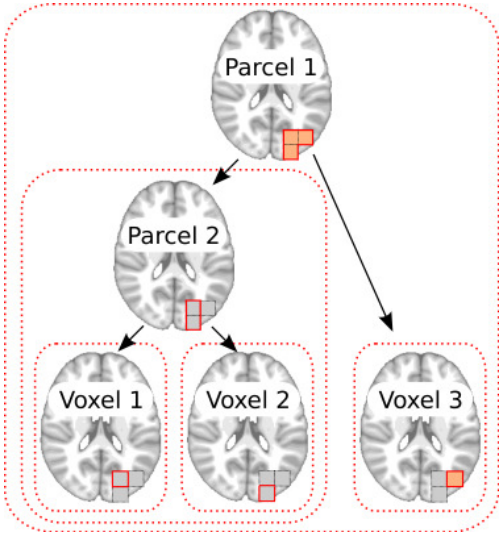
How to exploit the new representation?



- $\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$
- Assume linear models: $\mathbf{w}^\top \mathbf{x}_{\text{new}}$
- **Idea**: The closer to the leaves, the less reliable across subjects
- First select $\mathbf{w}_j$ near the root

- ℓ_1-norm would not distinguish between parcels/voxels
- We use a hierarchical sparsity-inducing norm
  - Cannot select a voxel without selecting ancestral parcels

How to exploit the new representation?



- $\mathbf{x}_{\text{new}} = [\mathbf{x}_{\text{old}}, \mathbf{x}_{\text{parcel1}}, \mathbf{x}_{\text{parcel2}}] \in \mathbb{R}^5$
- Assume linear models: $\mathbf{w}^\top \mathbf{x}_{\text{new}}$
- **Idea**: The closer to the leaves, the less reliable across subjects
- First select $\mathbf{w}_j$ near the root

- ℓ_1-norm would not distinguish between parcels/voxels
- We use a hierarchical sparsity-inducing norm
  - Cannot select a voxel without selecting ancestral parcels
  - If a parcel is zeroed out, the same goes for all its descendants

Regression results

Loss function:	Square	
	Error (mean,std)	P-value w.r.t. Tree ℓ_2
Regularization:		
ℓ_2 (Ridge)	(13.8, 7.6)	0.096
ℓ_1	(20.2, 10.9)	0.013*
$\ell_1 + \ell_2$ (Elastic net)	(14.4, 8.8)	0.065
Reweighted ℓ_1	(18.8, 14.6)	0.052
ℓ_1 (augmented space)	(14.2, 7.9)	0.096
ℓ_1 (tree weights)	(13.9, 7.9)	0.032*
Tree ℓ_2	(**11.8**, **6.7**)	-
Tree ℓ_∞	(12.8, 6.7)	0.137
Greedy [Michel et al., 2010]	(12.0, 5.5)	0.5

Classification results

Loss function:	Multinomial logistic	
	Error (mean,std)	P-value w.r.t. Tree ℓ_2
Regularization:		
ℓ_2 (Ridge)	(24.2, 9.2)	0.035*
ℓ_1	(25.8, 12.0)	0.004*
ℓ_1/ℓ_2 (Multi-task)	(26.7, 7.6)	0.007*
ℓ_1/ℓ_∞ (Multi-task)	(26.7, 11.6)	0.002*
Tree ℓ_2	(**16.7**, **10.4**)	-
Tree ℓ_∞	(22.5, 13.0)	0.156
Greedy [Michel et al., 2010]	(21.6, 14.5)	0.001*

Neuroimaging conclusions

- Best prediction accuracy
- Multi-scale representation
  - Robust to misalignment
  - Adapted to the spatial encoding of brain activations
- Explored in a convex way
  - No initialization issue
  - Minimum guarantee
- Improve supervised learning from brain imaging data:
  - fMRI decoding like shown here
  - Ability to classify patients vs. controls for diagnosis purposes

Conclusions and take-home messages

- Structured sparsity via convex tools
- Structure $\Leftrightarrow$ Constraints on possible sparsity patterns

- Plain and structured sparsity for the same computational cost!

- Different scenarios where structured sparsity applies:
  - Fixed dictionary with *preprocessing* to match Ω
    -e.g., neuroimaging with hierarchical clustering
  - Dictionary learning with *learned atoms* matching Ω
    -e.g., tree of image patches

- Connections between:
  - Sparse structured dictionary learning
  - Non-parametric Bayesian techniques

Perspectives and future work

- Deeper statistical analysis
  - Quantify statistical gain w.r.t. ℓ_1-norm
  - E.g., Negahban et al. [2009]; Maurer and Pontil [2011]
- Understand the role of regularizers in dictionary learning
  - Risk analysis [Maurer and Pontil, 2010]
  - Identifiability questions [Gribonval and Schnass, 2010]
- Monitor better the optimization in dictionary learning
- Further validate structured sparsity in NLP
  - Structures are plentiful [Martins et al., 2011]

Thank you all for your attention

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 *Advances in Neural Information Processing Systems*, 2010a.

F. Bach. Shaping level sets with submodular functions. Technical report, Preprint arXiv:1012.1501, 2010b.

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 56: 1982–2001, 2010.

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

D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

D. Blei, T. L. Griffiths, and M. I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2):1–30, 2010.

References II

Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, 2001.

E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

D. B. Chklovskii and A. A. Koulakov. Maps in the brain: What can we learn from them? *Annual Reviews in Neuroscience*, 27(1):369, 2004.

P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Springer, 2010.

J. P. Doignon and J. C. Falmagne. *Knowledge Spaces*. Springer-Verlag, 1998.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432), 1995.

References III

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 54(12):3736–3745, December 2006.

Q. Geng, H. Wang, and J. Wright. On the Local Correctness of L1 Minimization for Dictionary Learning. Technical report, Preprint arXiv:1101.5672, 2011.

A. Gramfort, D. Strohmeier, J. Haueisen, M. Hamalainen, and M. Kowalski. Functional brain imaging with M/EEG using structured sparsity in time-frequency dictionaries. In *Information Processing in Medical Imaging*, pages 600–611. Springer, 2011.

R. Gribonval and K. Schnass. Dictionary identification—sparse matrix-factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.

L. He and L. Carin. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 57:3488–3497, 2009.

J. Huang and T. Zhang. The benefit of group sparsity. *Annals of Statistics*, 38 (4):1978–2004, 2010.

References IV

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

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

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. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

A. Lefèvre, F. Bach, and C. Févotte. Itakura-saito nonnegative matrix factorization with group sparsity. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

J. Mairal. *Sparse coding for machine learning, image processing and computer vision*. PhD thesis, École normale supérieure de Cachan - ENS Cachan, 2010. Available at `http://tel.archives-ouvertes.fr/tel-00595312/fr/`.

References V

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Convex and network flow optimization for structured sparsity. *Journal of Machine Learning Research*, 12:2681–2720, 2011.

A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo. Structured sparsity in structured prediction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2011.

A. Maurer and M. Pontil. k-dimensional coding schemes in hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

A. Maurer and M. Pontil. Structured sparsity and generalization. Technical report, Preprint arXiv:1108.3476, 2011.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

C. A. Micchelli, J. M. Morales, and M. Pontil. A family of penalty functions for structured sparsity. In *Advances in Neural Information Processing Systems*, 2010.

V. Michel, E. Eger, C. Keribin, J.-B. Poline, and B. Thirion. A supervised clustering approach for extracting predictive information from brain activation images. *MMBIA'10*, 2010.

References VI

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24:227, 1995.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, 2009.

Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Kluwer Academic Publishers, 2004.

Y. Nesterov. Gradient methods for minimizing composite objective function. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, 2007.

G. Obozinski, B. Taskar, and M. I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, pages 1–22, 2009.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

G. Peyré. Sparse modeling of textures. *Journal of Mathematical Imaging and Vision*, 34(1):17–31, 2009.

References VII

F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.

T. Rebafka, C. Lévy-Leduc, and M. Charbit. Omp-type algorithm with structured sparsity patterns for multipath radar signals. Technical report, Preprint arXiv:1103.5158, 2011.

V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

M. Schmidt and K. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *Advances in Neural Information Processing Systems*, 2011.

J. M. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, 1993.

P. Sprechmann, I. Ramirez, P. Cancela, and G. Sapiro. Collaborative sources identification in mixed signals via hierarchical sparse modeling. Technical report, Preprint arXiv:1010.4893, 2010.

References VIII

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

A. Tibau Puig, A. Wiesel, A. Zaas, C. Woods, G. Ginsburg, G. Fleury, and A. Hero. Order-preserving factor analysis-application to longitudinal gene expression. *IEEE Transactions on Signal Processing*, 99(99):1–1, 2011.

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

B. A. Turlach, W. N. Venables, and S. J. Wright. Simultaneous variable selection. *Technometrics*, 47(3):349–363, 2005.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B*, 68(1):49–67, 2006.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A): 3468–3497, 2009.

Part V
Appendix

Consistency, "$p < n$" setting

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda\Omega(\mathbf{w})\right]$$

- Fixed-design model: $y = \mathbf{x}^\top \bar{\mathbf{w}} + \varepsilon$
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top = \bar{\mathbf{Q}} \succ 0$
- **Goal:** Recover $\bar{\mathrm{J}} = \{j \in [\![1; p]\!];\ \bar{\mathbf{w}}_j \neq 0\}$ by its estimate $\hat{\mathrm{J}}$:

$$\mathbb{P}(\hat{\mathrm{J}} = \bar{\mathrm{J}}) \underset{n\to+\infty}{\longrightarrow} 1 \quad (*)$$

Consistency, "$p < n$" setting

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \Omega(\mathbf{w}) \right]$$

- Fixed-design model: $y = \mathbf{x}^\top \bar{\mathbf{w}} + \varepsilon$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \bar{\mathbf{Q}} \succ 0$
- **Goal:** Recover $\bar{\mathrm{J}} = \{j \in [\![1; p]\!];\ \bar{\mathbf{w}}_j \neq 0\}$ by its estimate $\hat{\mathrm{J}}$:

$$\mathbb{P}(\hat{\mathrm{J}} = \bar{\mathrm{J}}) \underset{n\to+\infty}{\longrightarrow} 1 \quad (*)$$

- Key quantity:

Correlation bewteen "good" and "bad" variables

$$\theta_{\text{struct}} = \underbrace{\Omega^*}_{\text{dual norm of } \Omega} \Big(\overbrace{\bar{\mathbf{Q}}_{\bar{\mathrm{J}}^c \bar{\mathrm{J}}}}\ [\bar{\mathbf{Q}}_{\bar{\mathrm{J}}\bar{\mathrm{J}}}]^{-1} \mathbf{r}_{\bar{\mathrm{J}}} \Big), \text{ with } \mathbf{r}_j = \bar{\mathbf{w}}_j \sum_{\substack{g\in\mathcal{G}, g\ni j \\ g\cap\bar{\mathrm{J}}\neq\varnothing}} \|\bar{\mathbf{w}}_g\|_2^{-1}$$

Consistency, "$p < n$" setting

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \Omega(\mathbf{w}) \right]$$

- Fixed-design model: $y = \mathbf{x}^\top \bar{\mathbf{w}} + \varepsilon$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \bar{\mathbf{Q}} \succ 0$
- **Goal:** Recover $\bar{\mathrm{J}} = \{j \in [\![1; p]\!];\ \bar{\mathbf{w}}_j \neq 0\}$ by its estimate $\hat{\mathrm{J}}$:

$$\mathbb{P}(\hat{\mathrm{J}} = \bar{\mathrm{J}}) \underset{n\to+\infty}{\longrightarrow} 1 \quad (*)$$

- Key quantity:

Correlation bewteen "good" and "bad" variables

$$\theta_{\text{struct}} = \underbrace{\Omega^*}_{\text{dual norm of } \Omega} \Big(\overbrace{\bar{\mathbf{Q}}_{\bar{\mathrm{J}}^c \bar{\mathrm{J}}}}\ [\bar{\mathbf{Q}}_{\bar{\mathrm{J}}\bar{\mathrm{J}}}]^{-1} \mathbf{r}_{\bar{\mathrm{J}}} \Big), \text{ with } \mathbf{r}_j = \bar{\mathbf{w}}_j \sum_{\substack{g\in\mathcal{G}, g\ni j \\ g\cap\bar{\mathrm{J}}\neq\varnothing}} \|\bar{\mathbf{w}}_g\|_2^{-1}$$

For the ℓ_1-norm, $\theta_{\ell_1} = \left\| \bar{\mathbf{Q}}_{\bar{\mathrm{J}}^c \bar{\mathrm{J}}} [\bar{\mathbf{Q}}_{\bar{\mathrm{J}}\bar{\mathrm{J}}}]^{-1} \mathrm{sign}(\bar{\mathbf{w}})_{\bar{\mathrm{J}}} \right\|_\infty$ [Zhao and Yu, 2006]

Consistency, "$p < n$" setting

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda \Omega(\mathbf{w}) \right]$$

- Fixed-design model: $y = \mathbf{x}^\top \bar{\mathbf{w}} + \varepsilon$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \bar{\mathbf{Q}} \succ 0$
- **Goal:** Recover $\bar{\mathrm{J}} = \{j \in [\![1; p]\!];\ \bar{\mathbf{w}}_j \neq 0\}$ by its estimate $\hat{\mathrm{J}}$:

$$\mathbb{P}(\hat{\mathrm{J}} = \bar{\mathrm{J}}) \underset{n\to+\infty}{\longrightarrow} 1 \quad (*)$$

- Key quantity:

Correlation bewteen "good" and "bad" variables

$$\theta_{\text{struct}} = \underbrace{\Omega^*}_{\text{dual norm of } \Omega} \Big(\overbrace{\bar{\mathbf{Q}}_{\bar{\mathrm{J}}^c \bar{\mathrm{J}}}}\ [\bar{\mathbf{Q}}_{\bar{\mathrm{J}}\bar{\mathrm{J}}}]^{-1} \mathbf{r}_{\bar{\mathrm{J}}} \Big), \text{ with } \mathbf{r}_j = \bar{\mathbf{w}}_j \sum_{\substack{g\in\mathcal{G}, g\ni j \\ g\cap\bar{\mathrm{J}}\neq\varnothing}} \|\bar{\mathbf{w}}_g\|_2^{-1}$$

Proposition

If $\lambda \to 0, \lambda\sqrt{n} \to +\infty$, then if $\theta_{struct} < 1$, then $(*)$.
Conversely, if $(*)$, then $\theta_{struct} \leq 1$.

Formal characterization of zero patterns

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda\Omega(\mathbf{w})\right] \quad (*)$$

- $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ . \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p}$ fixed, $\{y_i\}_{i\in[\![1;n]\!]}$ random (with a density)

Formal characterization of zero patterns

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda\Omega(\mathbf{w})\right] \quad (*)$$

- $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ . \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p}$ fixed, $\{y_i\}_{i\in[\![1;n]\!]}$ random (with a density)
- Loss function, $l : (y, y') \mapsto l(y, y')$ convex, C^2, with
  - $\frac{\partial^2 l(y,y')}{\partial y \partial y'} \neq 0$ and $\frac{\partial^2 l(y,y')}{\partial^2 y'} > 0$

Formal characterization of zero patterns

$$\min_{\mathbf{w}\in\mathbb{R}^p} \left[\frac{1}{n}\sum_{i=1}^{n} l(y_i, \mathbf{w}^\top \mathbf{x}_i) + \lambda\Omega(\mathbf{w})\right] \quad (*)$$

- $\left\{(\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\right\}_{i\in[\![1;n]\!]}$, (input, output) data points
  - $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ . \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n\times p}$ fixed, $\{y_i\}_{i\in[\![1;n]\!]}$ random (with a density)
- Loss function, $l : (y, y') \mapsto l(y, y')$ convex, C^2, with
  - $\frac{\partial^2 l(y,y')}{\partial y \partial y'} \neq 0$ and $\frac{\partial^2 l(y,y')}{\partial^2 y'} > 0$

Proposition

Let k be the column rank of $\mathbf{X}$.
Any solution of $(*)$ with at most $k-1$ nonzero entries has almost surely its zero pattern in

$$\mathcal{Z} = \left\{ \bigcup_{g\in\mathcal{G}'} g;\ \mathcal{G}' \subseteq \mathcal{G} \right\}.$$

Benchmark: Denoising of image patches

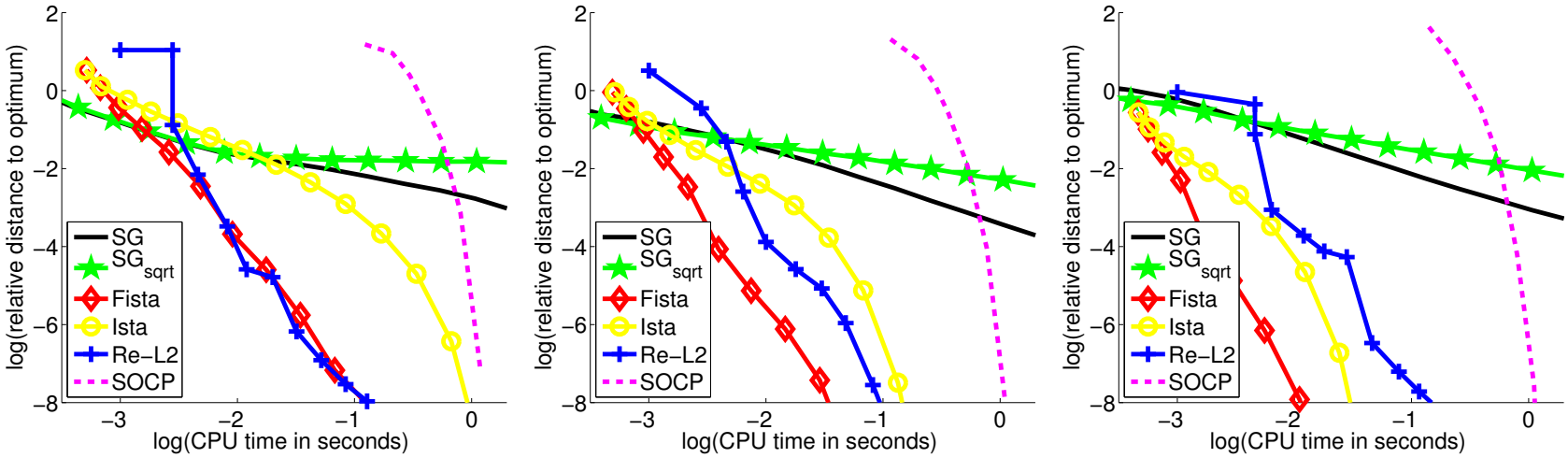


Figure: Sparsity level increasing from left to right.

- Small scale: Dictionary in $\mathbb{R}^{256\times 151}$
- Solved thousands of times in inner loop for dictionary learning

Hierarchical topic models

- NIPS proceedings (1714 documents, 8274 words).
- Each document is modeled through word frequencies.
- Alternative to probabilistic topic models [Blei et al., 2010].