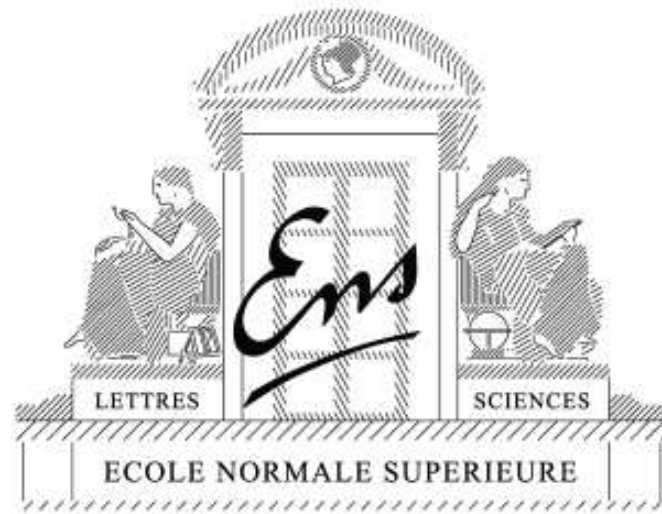


Structured sparsity through convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France



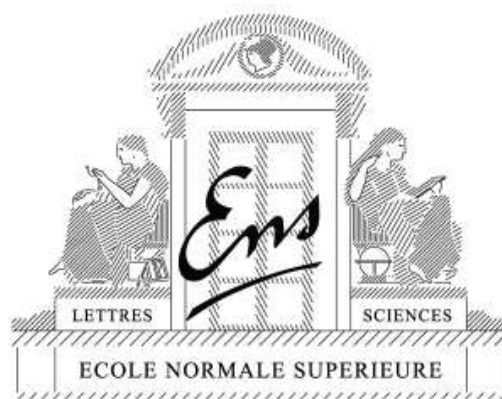
Joint work with R. Jenatton, J. Mairal, G. Obozinski
Journées INRIA - Apprentissage - December 2011

Outline

- **SIERRA project-team**
- **Introduction: Sparse methods for machine learning**
 - Need for structured sparsity: **Going beyond the ℓ_1 -norm**
- **Classical approaches to structured sparsity**
 - Linear combinations of ℓ_q -norms
- **Structured sparsity through submodular functions**
 - Relaxation of the penalization of supports
 - **Unified algorithms and analysis**

SIERRA - created January 1st, 2011

Composition of the INRIA/ENS/CNRS team



- **3 Researchers** (Sylvain Arlot, Francis Bach, Guillaume Obozinski)
- **4 Post-docs** (Simon Lacoste-Julien, Nicolas Le Roux, Ronny Luss, Mark Schmidt)
- **9 PhD students** (Louise Benoit, Florent Couzinie-Devy, Edouard Grave, Toby Hocking, Armand Joulin, Augustin Lefèvre, Anil Nelakanti, Fabian Pedregosa, Matthieu Solnon)

Machine learning

Computer science and applied mathematics

- **Modelisation, prediction and control from training examples**
- **Theory**
 - Analysis of statistical performance
- **Algorithms**
 - Numerical efficiency and stability
- **Applications**
 - Computer vision, bioinformatics, neuro-imaging, text, audio

Scientific objectives - SIERRA tenet

- Machine learning does not exist in the void**
- Specific domain knowledge must be exploited**

Scientific objectives - SIERRA tenet

- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

- Scientific challenges

- Fully automated data processing
- Incorporating structure
- Large-scale learning

Scientific objectives - SIERRA tenet

- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

- Scientific challenges

- Fully automated data processing
- Incorporating structure
- Large-scale learning

- Scientific objectives

- Supervised learning
- Parsimony
- Optimization
- Unsupervised learning

Scientific objectives - SIERRA tenet

- Machine learning does not exist in the void
- Specific domain knowledge must be exploited

- Scientific challenges

- Fully automated data processing
- Incorporating structure
- Large-scale learning

- Scientific objectives

- Supervised learning
- Parsimony
- Optimization
- Unsupervised learning

- Interdisciplinary collaborations

- Computer vision
- Bioinformatics
- Neuro-imaging
- Text, audio, natural language

Supervised learning

- Data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, $i = 1, \dots, n$
- **Goal:** predict $y \in \mathcal{Y}$ from $x \in \mathcal{X}$, i.e., find $f : \mathcal{X} \rightarrow \mathcal{Y}$
- Empirical risk minimization

$$\underbrace{\frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))}_{\text{Data-fitting}} + \underbrace{\frac{\lambda}{2} \|f\|^2}_{\text{Regularization}}$$

- **SIERRA Scientific objectives:**
 - Studying generalization error (S. Arlot, M. Solnon, F. Bach)
 - Improving calibration (S. Arlot, M. Solnon, F. Bach)
 - Two main types of norms: ℓ_2 vs. ℓ_1 (G. Obozinski, F. Bach)

Sparsity in supervised machine learning

- Observed data $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$
 - Response vector $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$
 - Design matrix $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$
- Regularized empirical risk minimization:

$$\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(y_i, w^\top x_i) + \lambda \Omega(w) = \boxed{\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \Omega(w)}$$

- Norm Ω to promote sparsity
 - square loss + ℓ_1 -norm \Rightarrow **basis pursuit** in signal processing (Chen et al., 2001), **Lasso** in statistics/machine learning (Tibshirani, 1996)
 - Proxy for **interpretability**
 - Allow **high-dimensional inference**: $\boxed{\log p = O(n)}$

Sparsity in **unsupervised** machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1, \dots, w^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(y^j, Xw^j) + \lambda \Omega(w^j) \right\}$$

Sparsity in **unsupervised** machine learning

- **Multiple** responses/signals $y = (y^1, \dots, y^k) \in \mathbb{R}^{n \times k}$

$$\min_{w^1, \dots, w^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(y^j, X w^j) + \lambda \Omega(w^j) \right\}$$

- **Only responses are observed** \Rightarrow **Dictionary learning**

– Learn $X = (x^1, \dots, x^p) \in \mathbb{R}^{n \times p}$ such that $\forall j, \|x^j\|_2 \leq 1$

$$\min_{X=(x^1, \dots, x^p)} \min_{w^1, \dots, w^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(y^j, X w^j) + \lambda \Omega(w^j) \right\}$$

– Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

- **sparse PCA**: replace $\|x^j\|_2 \leq 1$ by $\Theta(x^j) \leq 1$

Sparsity in signal processing

- **Multiple** responses/signals $x = (x^1, \dots, x^k) \in \mathbb{R}^{n \times k}$

$$\min_{\alpha^1, \dots, \alpha^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

- **Only responses are observed** \Rightarrow **Dictionary learning**

– Learn $D = (d^1, \dots, d^p) \in \mathbb{R}^{n \times p}$ such that $\forall j, \|d^j\|_2 \leq 1$

$$\min_{D=(d^1, \dots, d^p)} \min_{\alpha^1, \dots, \alpha^k \in \mathbb{R}^p} \sum_{j=1}^k \left\{ L(x^j, D\alpha^j) + \lambda \Omega(\alpha^j) \right\}$$

– Olshausen and Field (1997); Elad and Aharon (2006); Mairal et al. (2009a)

- **sparse PCA**: replace $\|d^j\|_2 \leq 1$ by $\Theta(d^j) \leq 1$

Why structured sparsity?

- **Interpretability**

- Structured dictionary elements (Jenatton et al., 2009b)
- Dictionary elements “organized” in a **tree** or a **grid** (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Structured sparse PCA (Jenatton et al., 2009b)



raw data



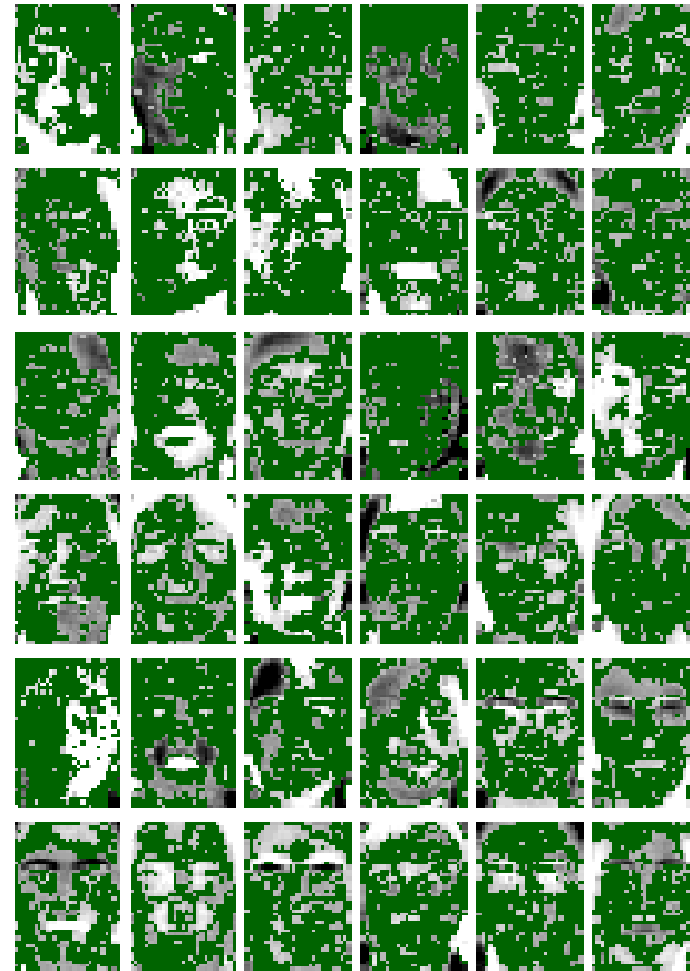
sparse PCA

- Unstructured sparse PCA \Rightarrow many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)



raw data



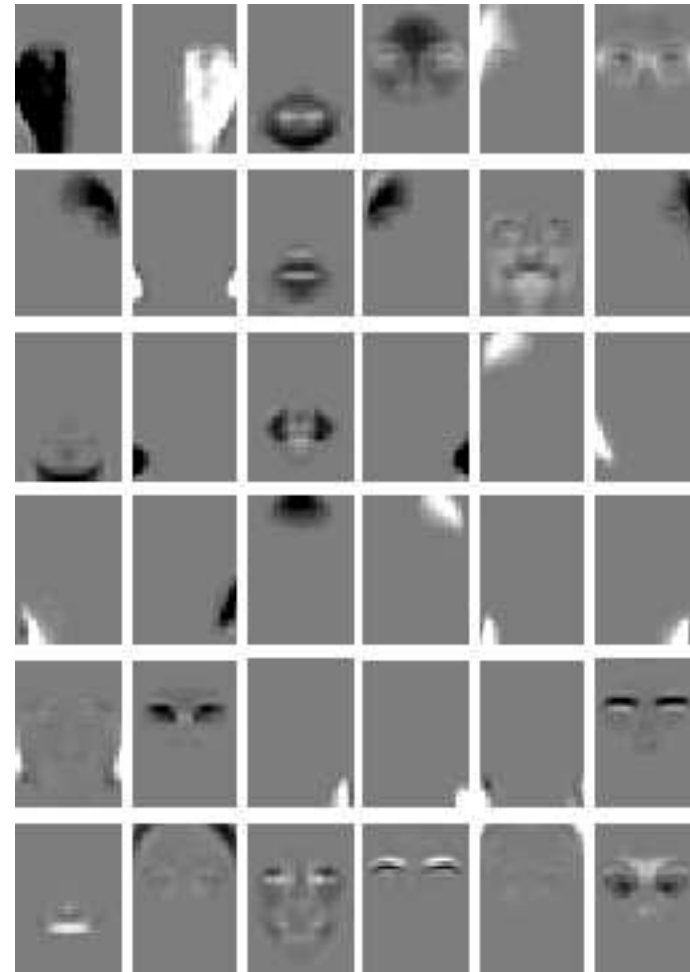
sparse PCA

- Unstructured sparse PCA \Rightarrow many zeros do not lead to better interpretability

Structured sparse PCA (Jenatton et al., 2009b)



raw data



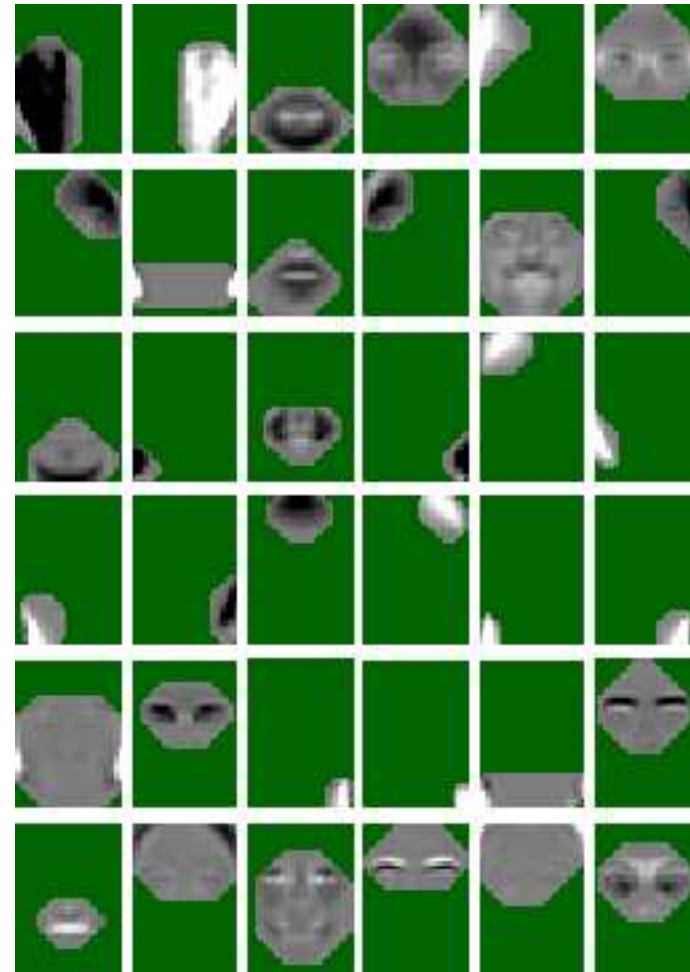
Structured sparse PCA

- Enforce selection of **convex** nonzero patterns \Rightarrow robustness to occlusion in face identification

Structured sparse PCA (Jenatton et al., 2009b)



raw data



Structured sparse PCA

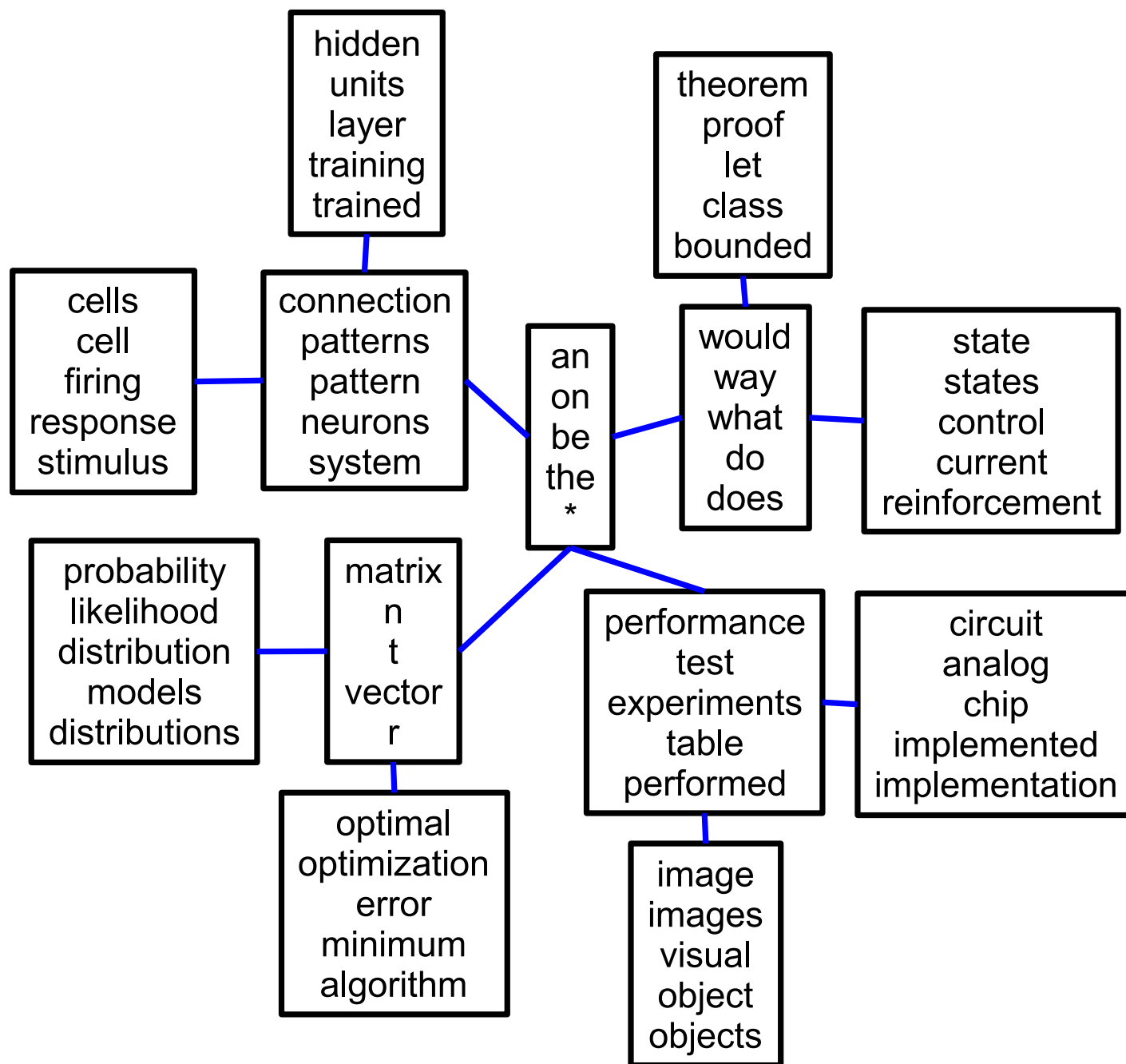
- Enforce selection of **convex** nonzero patterns \Rightarrow robustness to occlusion in face identification

Why structured sparsity?

- **Interpretability**

- Structured dictionary elements (Jenatton et al., 2009b)
- Dictionary elements “organized” in a **tree** or a **grid** (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Modelling of text corpora (Jenatton et al., 2010)



Why structured sparsity?

- **Interpretability**

- Structured dictionary elements (Jenatton et al., 2009b)
- Dictionary elements “organized” in a **tree** or a **grid** (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

Why structured sparsity?

- **Interpretability**

- Structured dictionary elements (Jenatton et al., 2009b)
- Dictionary elements “organized” in a **tree** or a **grid** (Kavukcuoglu et al., 2009; Jenatton et al., 2010; Mairal et al., 2010)

- **Stability and identifiability**

- Optimization problem $\min_{w \in \mathbb{R}^p} L(y, Xw) + \lambda \|w\|_1$ is unstable
- “Codes” w^j often used in later processing (Mairal et al., 2009c)

- **Prediction or estimation performance**

- When prior knowledge matches data (Haupt and Nowak, 2006; Baraniuk et al., 2008; Jenatton et al., 2009a; Huang et al., 2009)

- **Numerical efficiency**

- Non-linear variable selection with 2^p subsets (Bach, 2008)

Classical approaches to structured sparsity

- **Many application domains**

- Computer vision (Cevher et al., 2008; Mairal et al., 2009b)
- Neuro-imaging (Gramfort and Kowalski, 2009; Jenatton et al., 2011)
- Bio-informatics (Rapaport et al., 2008; Kim and Xing, 2010)

- **Non-convex approaches**

- Haupt and Nowak (2006); Baraniuk et al. (2008); Huang et al. (2009)

- **Convex approaches**

- Design of sparsity-inducing norms

Outline

- SIERRA project-team
- Introduction: Sparse methods for machine learning
 - Need for structured sparsity: Going beyond the ℓ_1 -norm
- Classical approaches to structured sparsity
 - Linear combinations of ℓ_q -norms
- Structured sparsity through submodular functions
 - Relaxation of the penalization of supports
 - Unified algorithms and analysis

Sparsity-inducing norms

- Popular choice for Ω

- The ℓ_1 - ℓ_2 norm,

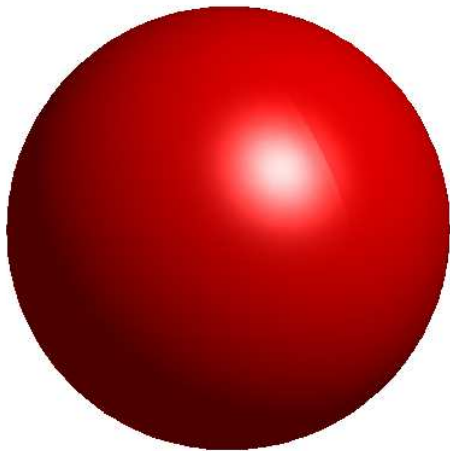
$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \left(\sum_{j \in G} w_j^2 \right)^{1/2}$$

- with \mathbf{H} a **partition** of $\{1, \dots, p\}$
- The ℓ_1 - ℓ_2 norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 -norm)
- For the square loss, group Lasso (Yuan and Lin, 2006)

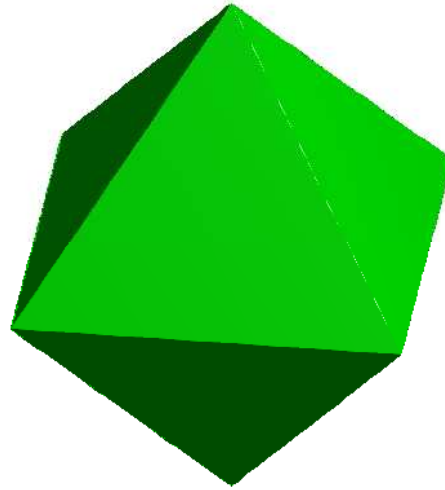


Unit norm balls

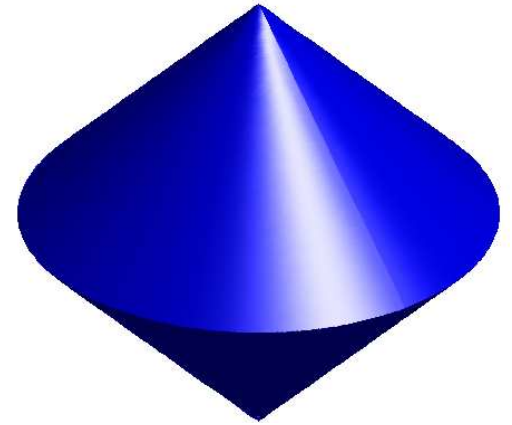
Geometric interpretation



$$\|w\|_2$$



$$\|w\|_1$$



$$\sqrt{w_1^2 + w_2^2} + |w_3|$$

Sparsity-inducing norms

- Popular choice for Ω

- The ℓ_1 - ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \left(\sum_{j \in G} w_j^2 \right)^{1/2}$$

- with \mathbf{H} a **partition** of $\{1, \dots, p\}$
- The ℓ_1 - ℓ_2 norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 -norm)
- For the square loss, group Lasso (Yuan and Lin, 2006)



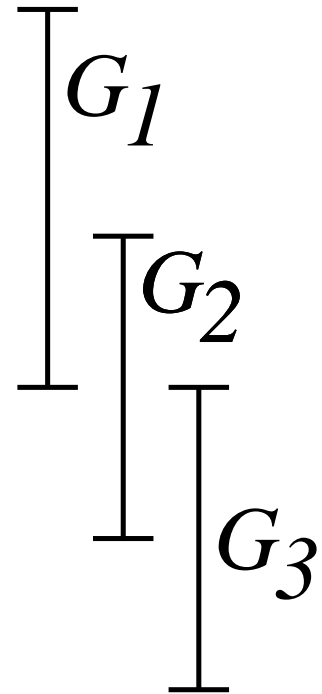
- However, the ℓ_1 - ℓ_2 norm encodes **fixed/static prior information**, requires to know in advance how to group the variables
- What happens if the set of groups \mathbf{H} is not a partition anymore?

Structured sparsity with **overlapping** groups (Jenatton, Audibert, and Bach, 2009a)

- When penalizing by the ℓ_1 - ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \left(\sum_{j \in G} w_j^2 \right)^{1/2}$$

- The ℓ_1 norm induces sparsity at the group level:
 - * Some w_G 's are set to zero
- Inside the groups, the ℓ_2 norm does not promote sparsity

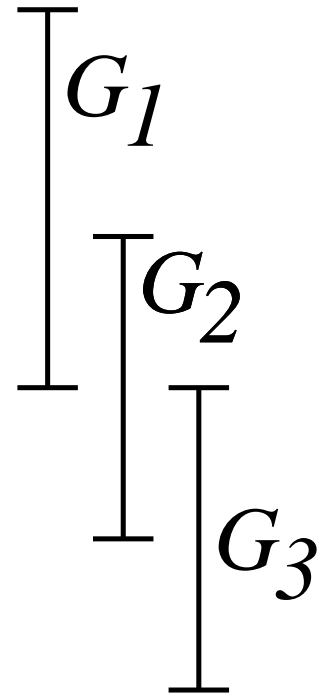


Structured sparsity with **overlapping** groups (Jenatton, Audibert, and Bach, 2009a)

- When penalizing by the ℓ_1 - ℓ_2 norm,

$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \left(\sum_{j \in G} w_j^2 \right)^{1/2}$$

- The ℓ_1 norm induces sparsity at the group level:
 - * Some w_G 's are set to zero
- Inside the groups, the ℓ_2 norm does not promote sparsity



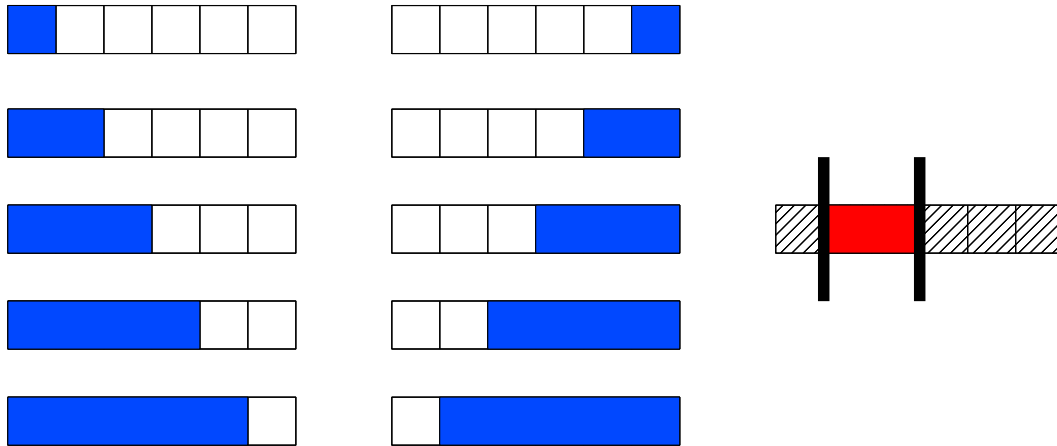
- The zero pattern of w is given by

$$\{j, w_j = 0\} = \bigcup_{G \in \mathbf{H}'} G \text{ for some } \mathbf{H}' \subseteq \mathbf{H}$$

- **Zero patterns are unions of groups**

Examples of set of groups \mathbf{H}

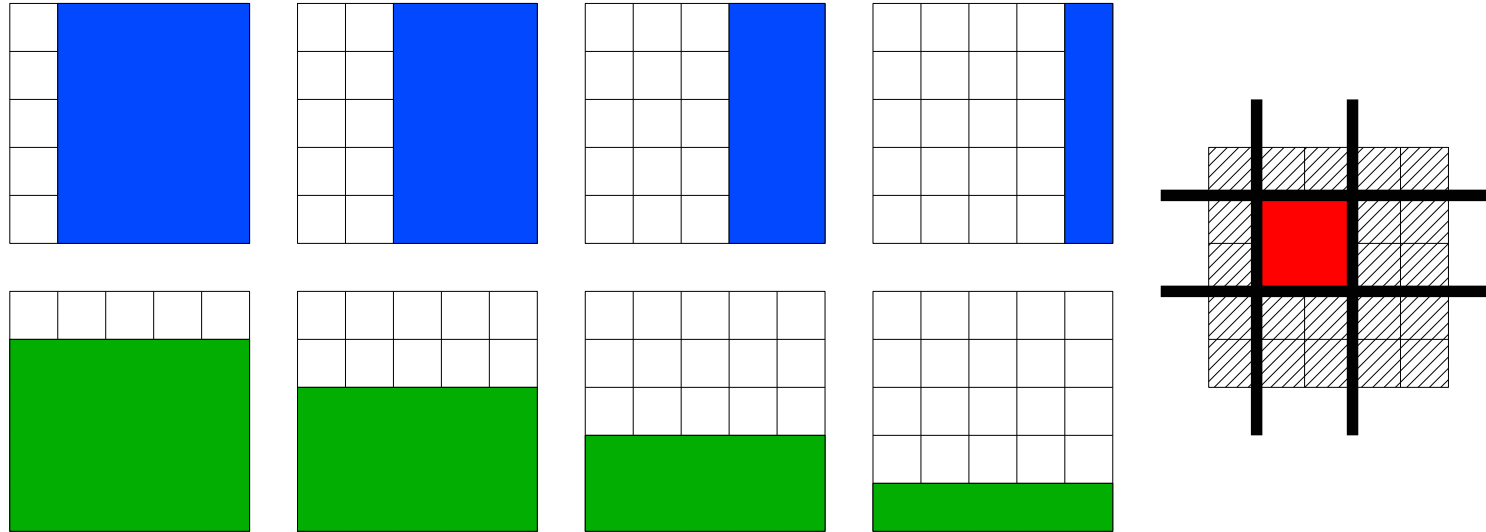
- Selection of contiguous patterns on a sequence, $p = 6$



- \mathbf{H} is the set of blue groups
- Any union of blue groups set to zero leads to the selection of a contiguous pattern

Examples of set of groups H

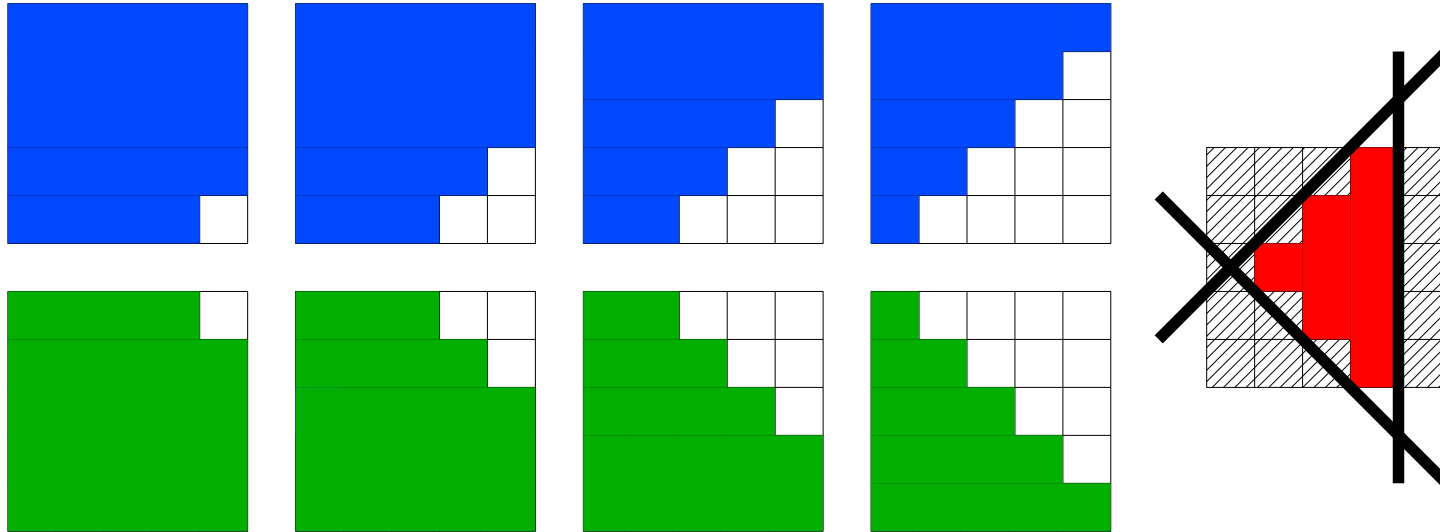
- Selection of rectangles on a 2-D grids, $p = 25$



- H is the set of blue/green groups (with their not displayed complements)
- Any union of blue/green groups set to zero leads to the selection of a rectangle

Examples of set of groups H

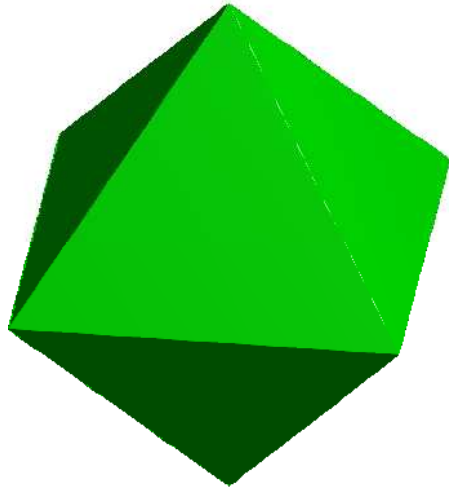
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



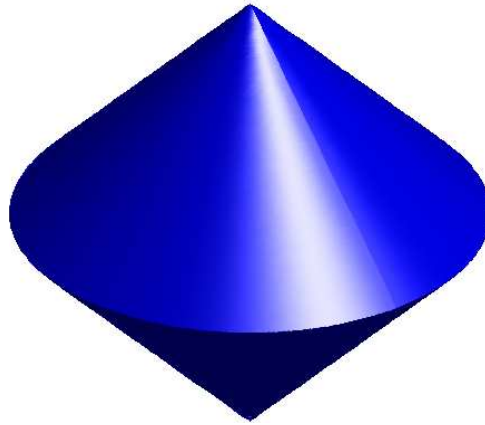
- It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

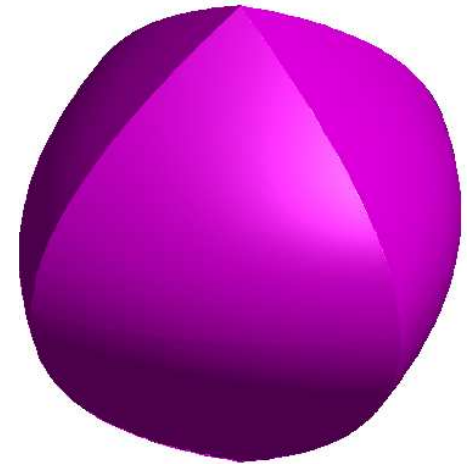
Geometric interpretation



$$\|w\|_1$$



$$\sqrt{w_1^2 + w_2^2} + |w_3|$$



$$\|w\|_2 + |w_1| + |w_2|$$

Optimization for sparsity-inducing norms

(see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)

- $w_{t+1} = \arg \min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2} \|w - w_t\|_2^2$
 - $w_{t+1} = w_t - \frac{1}{B} \nabla L(w_t)$

Optimization for sparsity-inducing norms

(see Bach, Jenatton, Mairal, and Obozinski, 2011)

- Gradient descent as a **proximal method** (differentiable functions)

- $w_{t+1} = \arg \min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{B}{2} \|w - w_t\|_2^2$
 - $w_{t+1} = w_t - \frac{1}{B} \nabla L(w_t)$

- Problems of the form:
$$\min_{w \in \mathbb{R}^p} L(w) + \lambda \Omega(w)$$

- $w_{t+1} = \arg \min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda \Omega(w) + \frac{B}{2} \|w - w_t\|_2^2$
 - $\Omega(w) = \|w\|_1 \Rightarrow$ **Thresholded gradient descent**

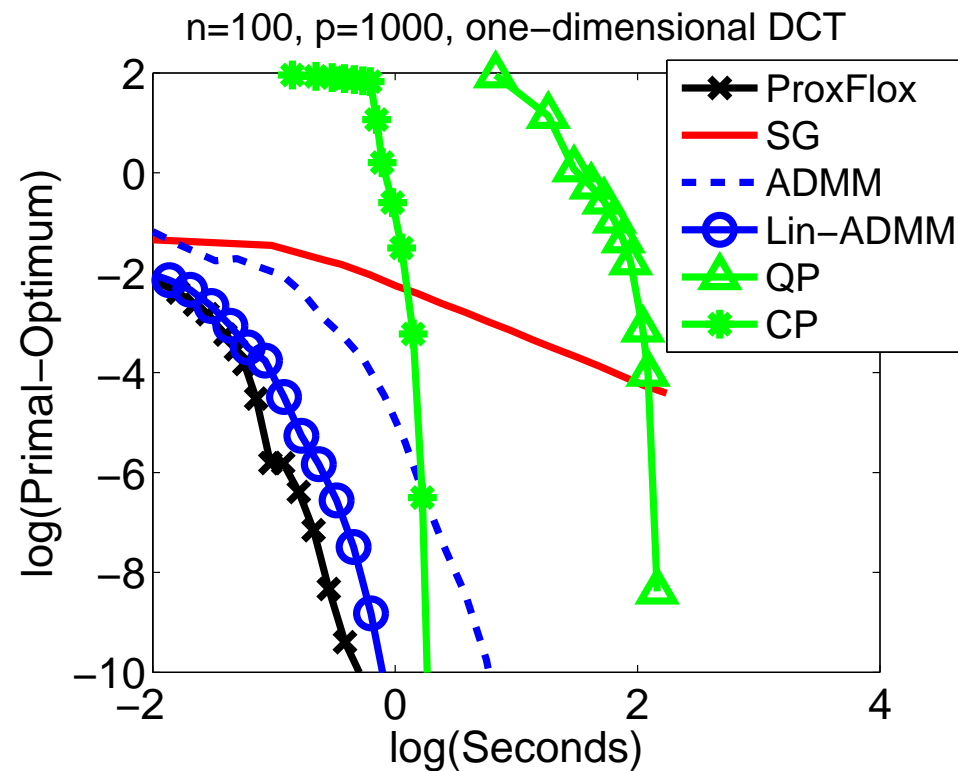
- Similar convergence rates than smooth optimization
 - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)

Comparison of optimization algorithms

(Mairal, Jenatton, Obozinski, and Bach, 2010)

Small scale

- Specific norms which can be implemented through network flows

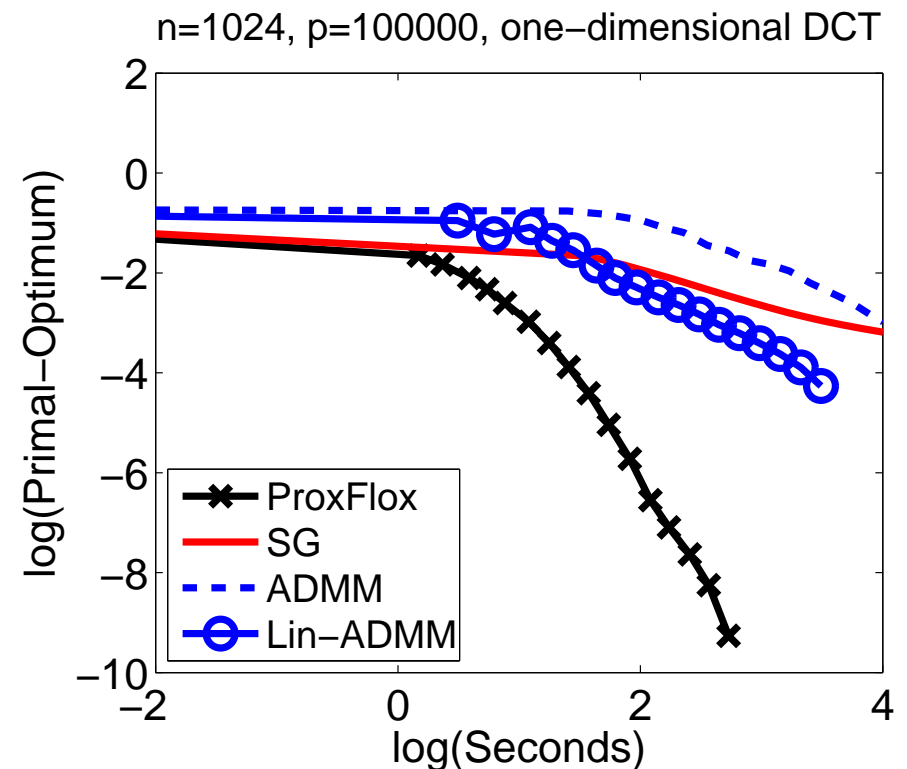
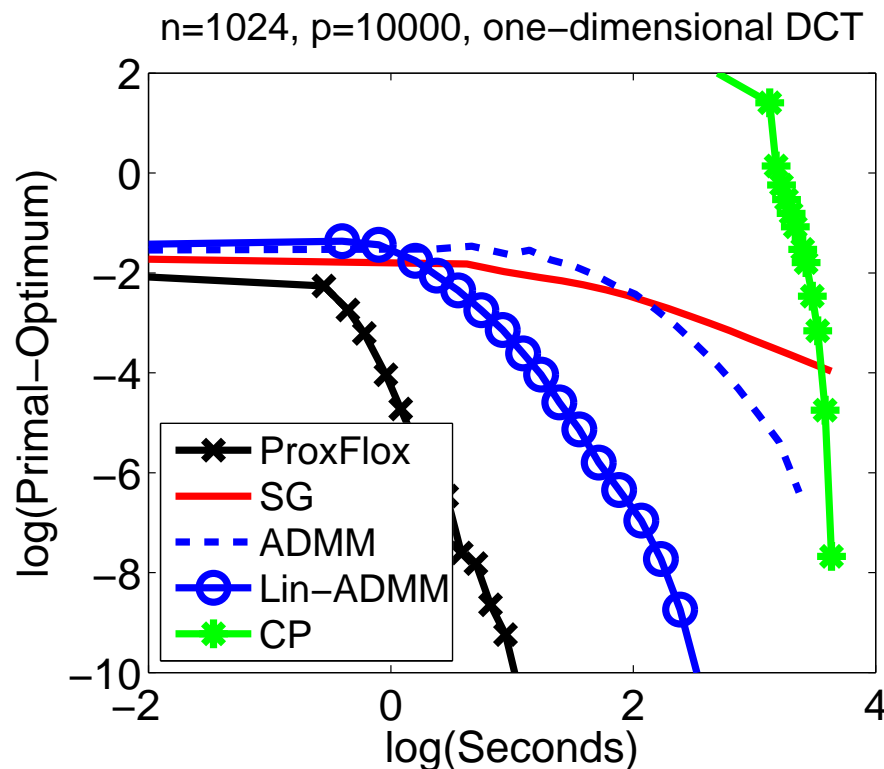


Comparison of optimization algorithms

(Mairal, Jenatton, Obozinski, and Bach, 2010)

Large scale

- Specific norms which can be implemented through network flows



Approximate proximal methods

(Schmidt, Le Roux, and Bach, 2011)

- Exact computation of proximal operator $\arg \min_{w \in \mathbb{R}^p} \frac{1}{2} \|w - z\|_2^2 + \lambda \Omega(w)$
 - Closed form for ℓ_1 -norm
 - Efficient for overlapping group norms (Jenatton et al., 2010; Mairal et al., 2010)
- Convergence rate: $O(1/t)$ and $O(1/t^2)$ (with acceleration)
- **Gradient or proximal operator may be only approximate**
 - Preserved convergence rate with appropriate control
 - Approximate gradient with non-random errors
 - Complex regularizers

Stochastic approximation

(Bach and Moulines, 2011)

- Loss = **generalization error** $L(w) = \mathbb{E}_{(x,y)} \ell(y, w^\top x)$
- **Stochastic approximation**: optimizing $L(w)$ given an sequence of samples (x_t, y_t)
- Context: large-scale learning
- Main algorithm: Stochastic gradient descent (a.k.a. Robbins-Monro)
 - Iteration: $w_t = w_{t-1} - \gamma_t \frac{\partial}{\partial w} \ell(y_t, w^\top x_t) \Big|_{w=w_{t-1}}$
 - Classical choice in machine learning: $\gamma_t = C/T \Rightarrow$ **Wrong choice**
- **Good choice**: **Use averaging of iterates with** $\gamma_t = C/t^{1/2}$
 - Robustness to difficulty of the problem and to the setting of C

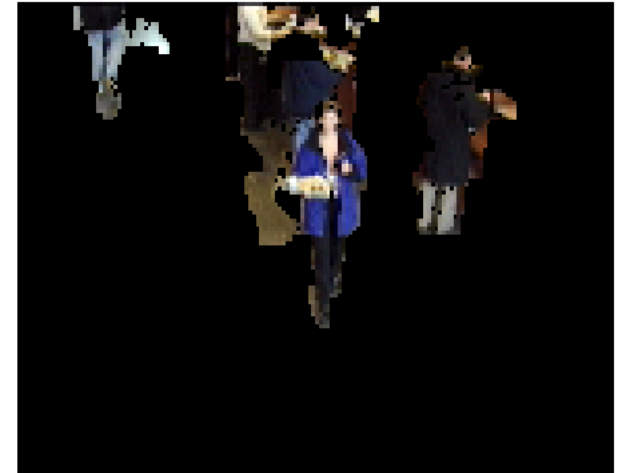
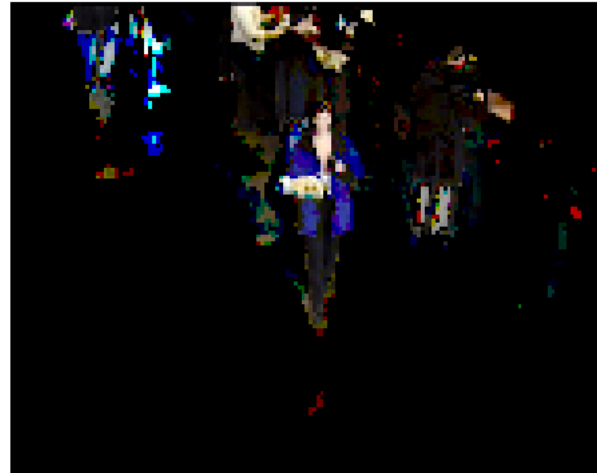
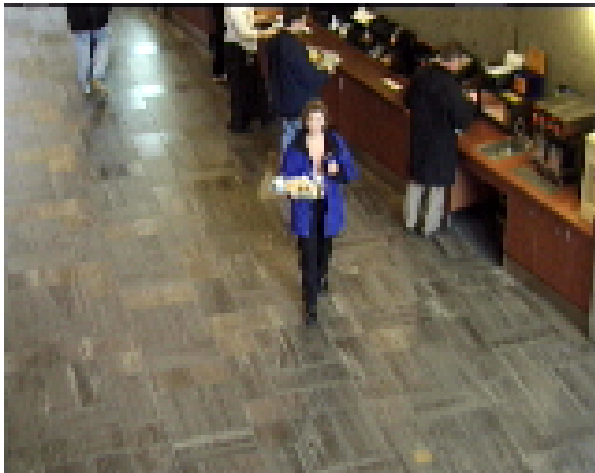
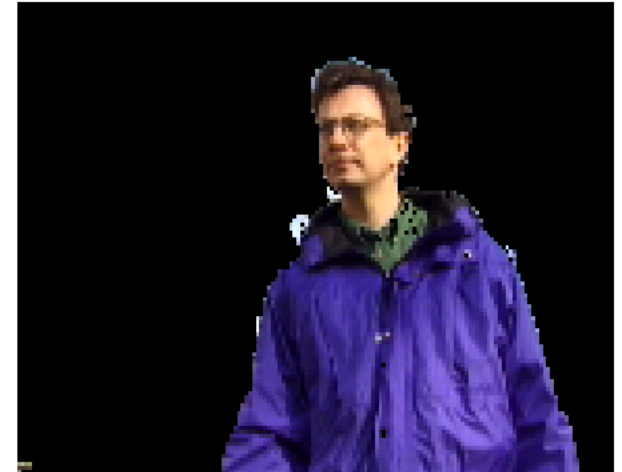
Application to background subtraction

(Mairal, Jenatton, Obozinski, and Bach, 2010)

Input

ℓ_1 -norm

Structured norm



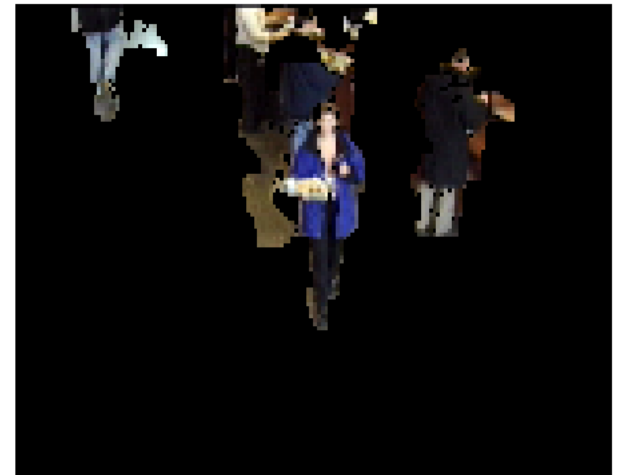
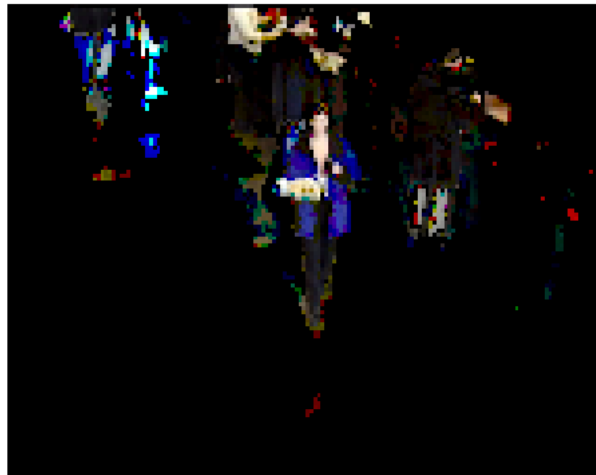
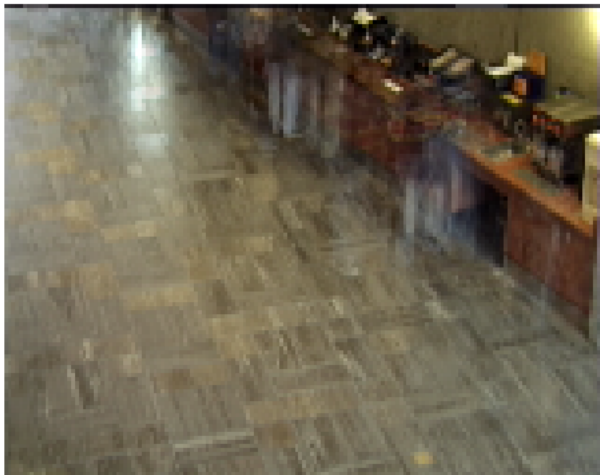
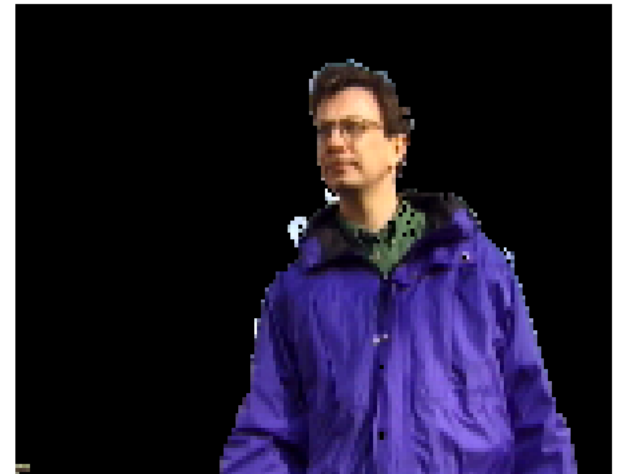
Application to background subtraction

(Mairal, Jenatton, Obozinski, and Bach, 2010)

Background

ℓ_1 -norm

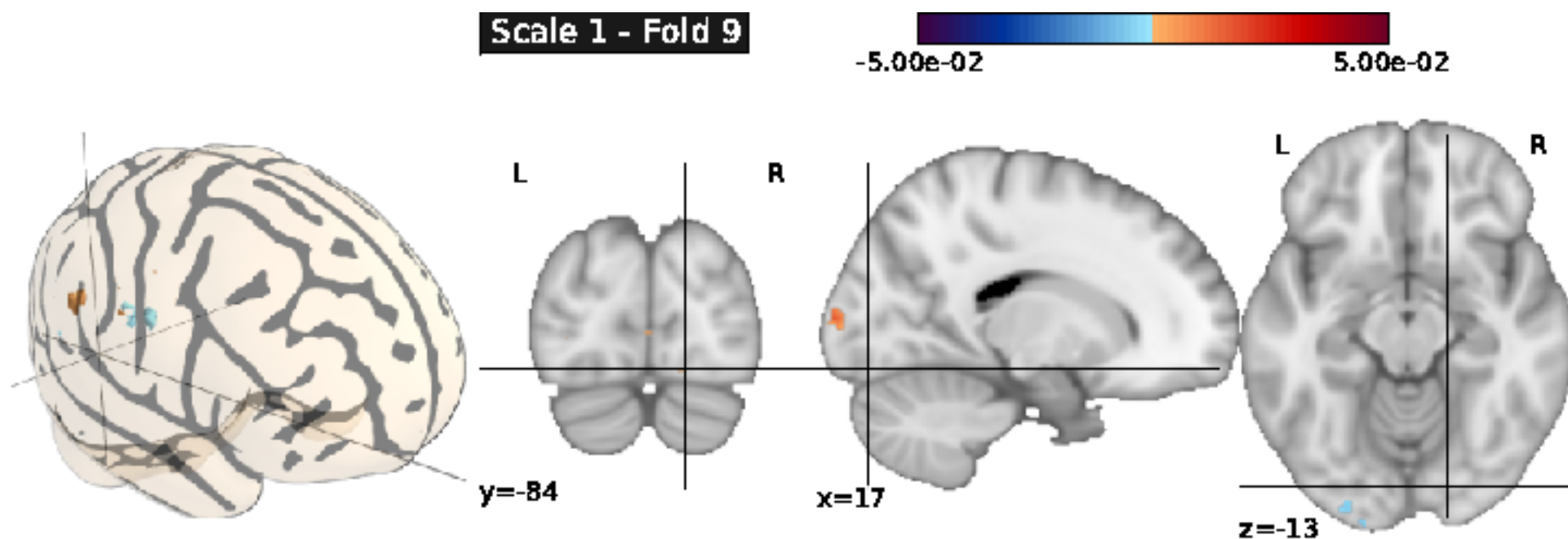
Structured norm



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

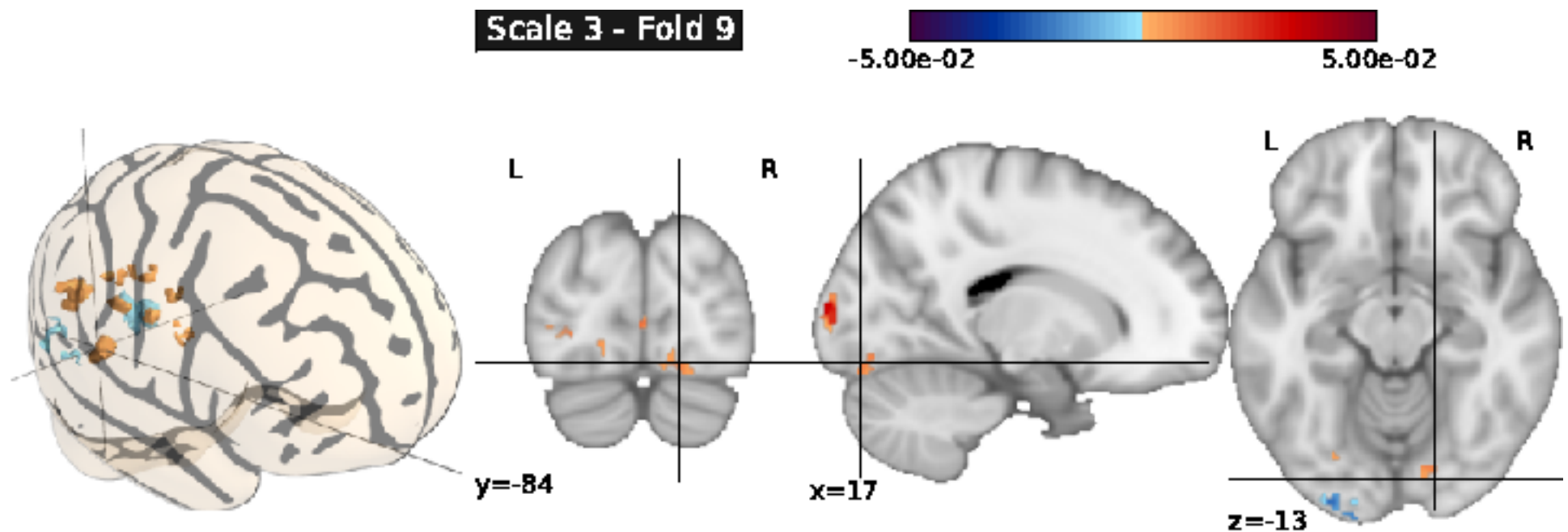
- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

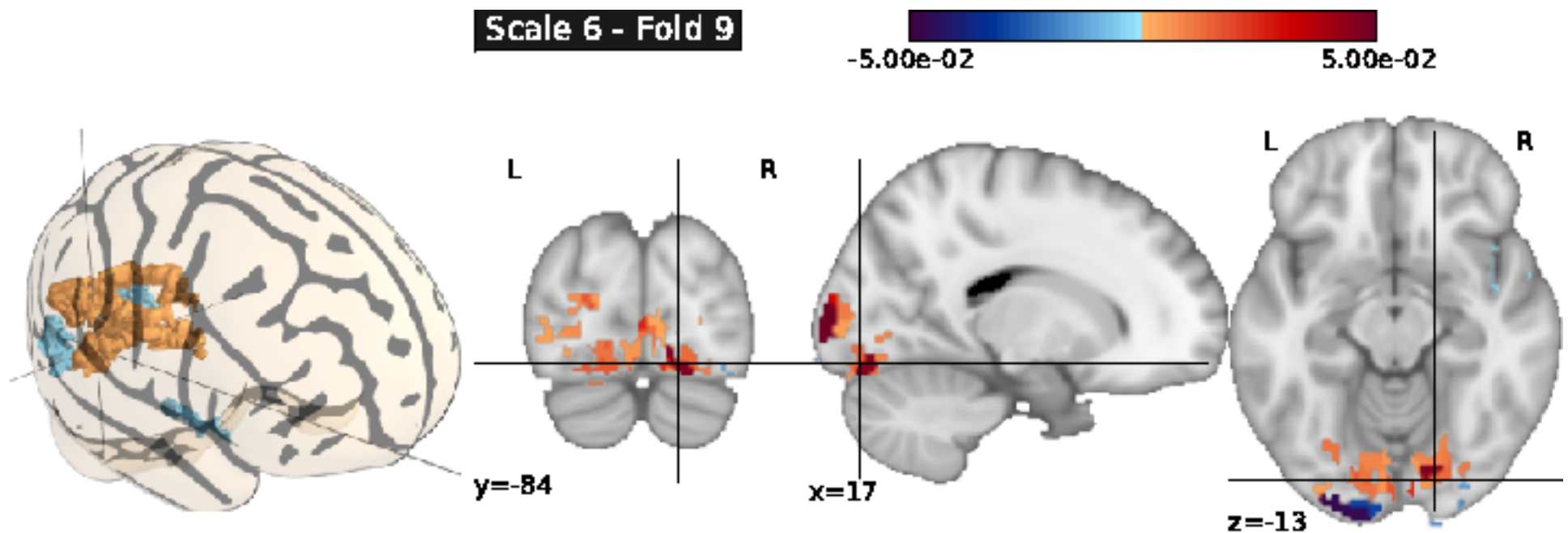
- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

Structured sparsity for fMRI (Jenatton et al., 2011)

- “Brain reading”: prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Sparse Structured PCA

(Jenatton, Obozinski, and Bach, 2009b)

- Learning **sparse and structured** dictionary elements:

$$\min_{W \in \mathbb{R}^{k \times n}, X \in \mathbb{R}^{p \times k}} \frac{1}{n} \sum_{i=1}^n \|y^i - Xw^i\|_2^2 + \lambda \sum_{j=1}^p \Omega(x^j) \text{ s.t. } \forall i, \|w^i\|_2 \leq 1$$

Application to face databases (1/3)



raw data



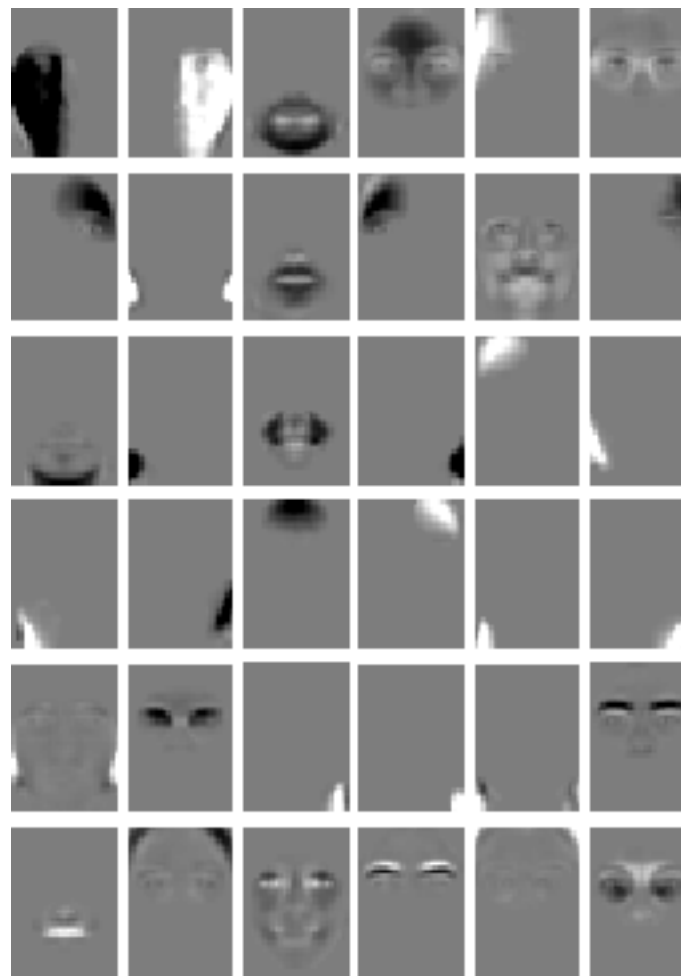
(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)



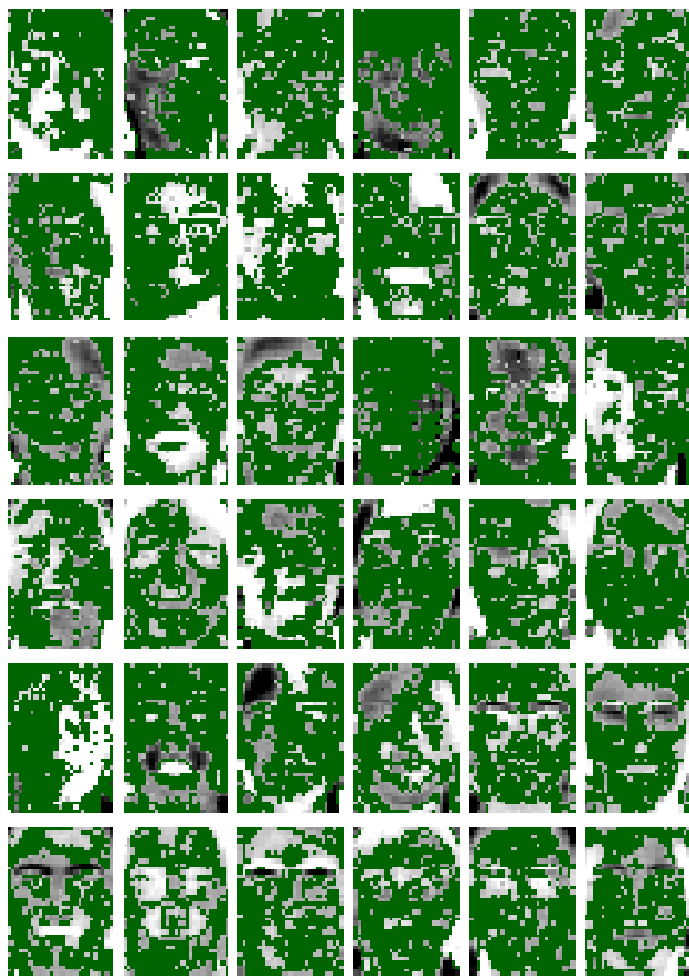
(unstructured) sparse PCA



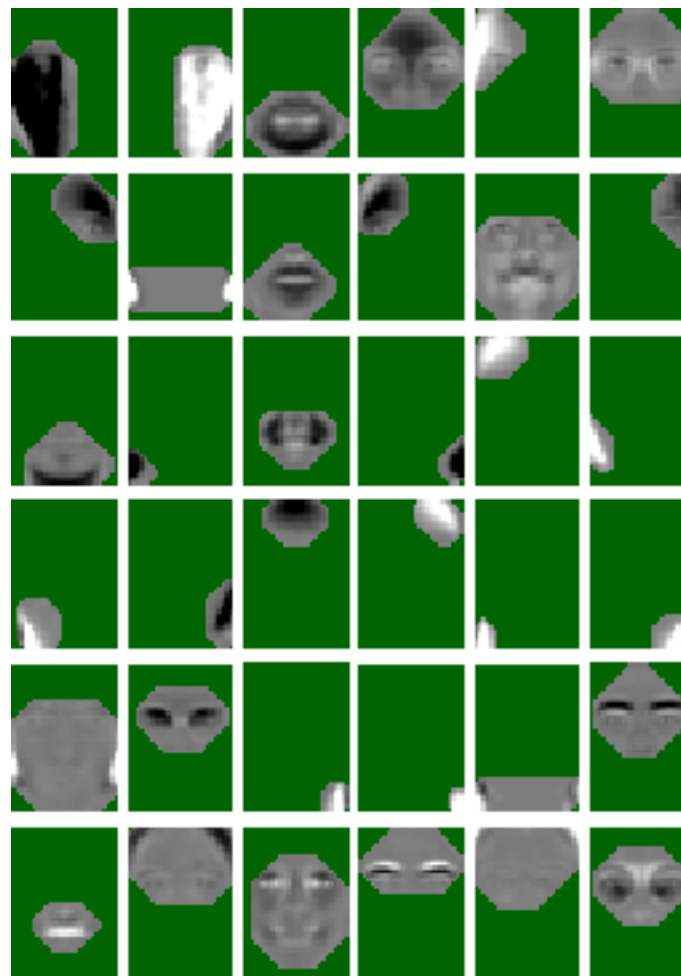
Structured sparse PCA

- Enforce selection of **convex** nonzero patterns \Rightarrow robustness to occlusion

Application to face databases (2/3)



(unstructured) sparse PCA

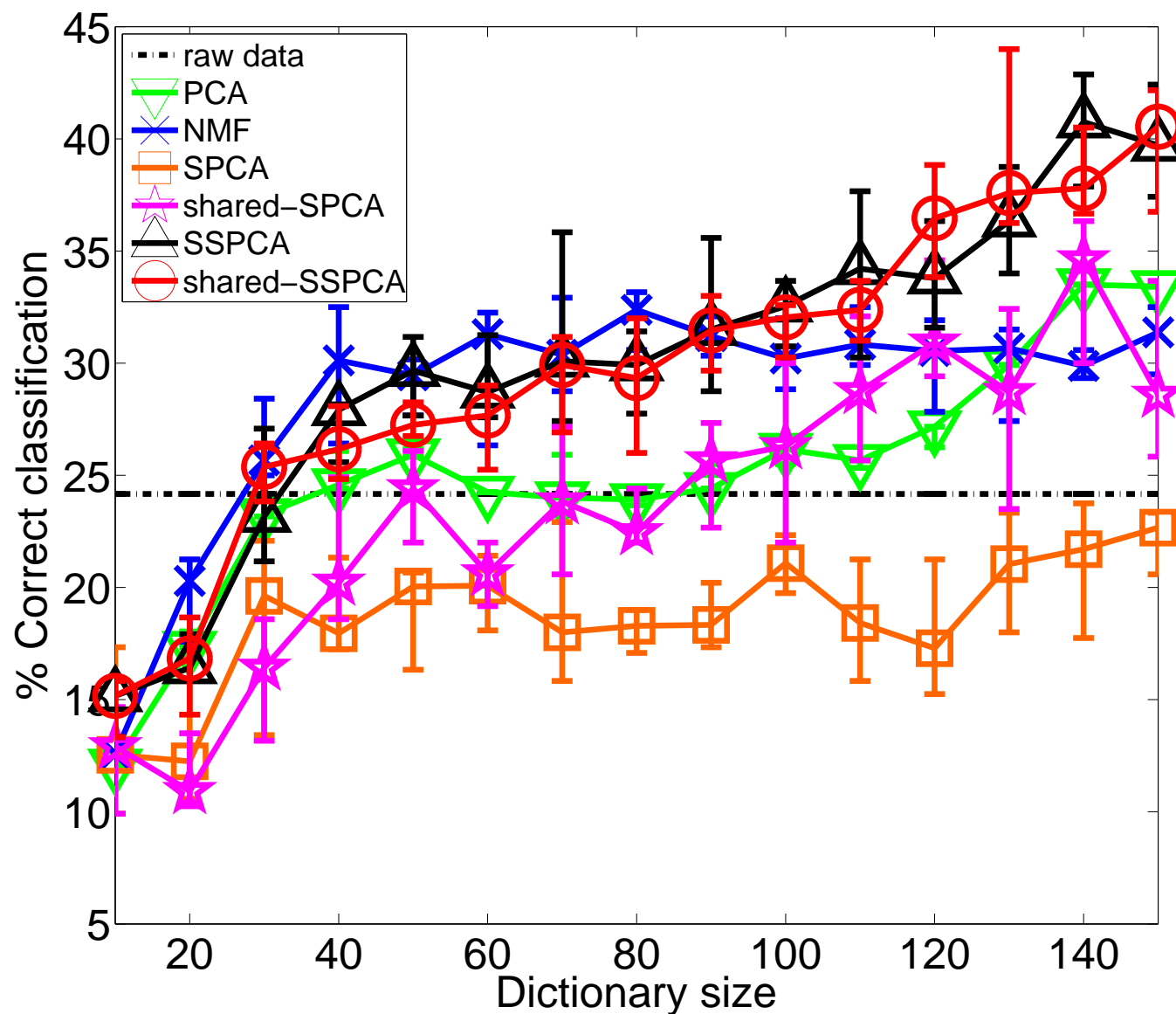


Structured sparse PCA

- Enforce selection of **convex** nonzero patterns \Rightarrow robustness to occlusion

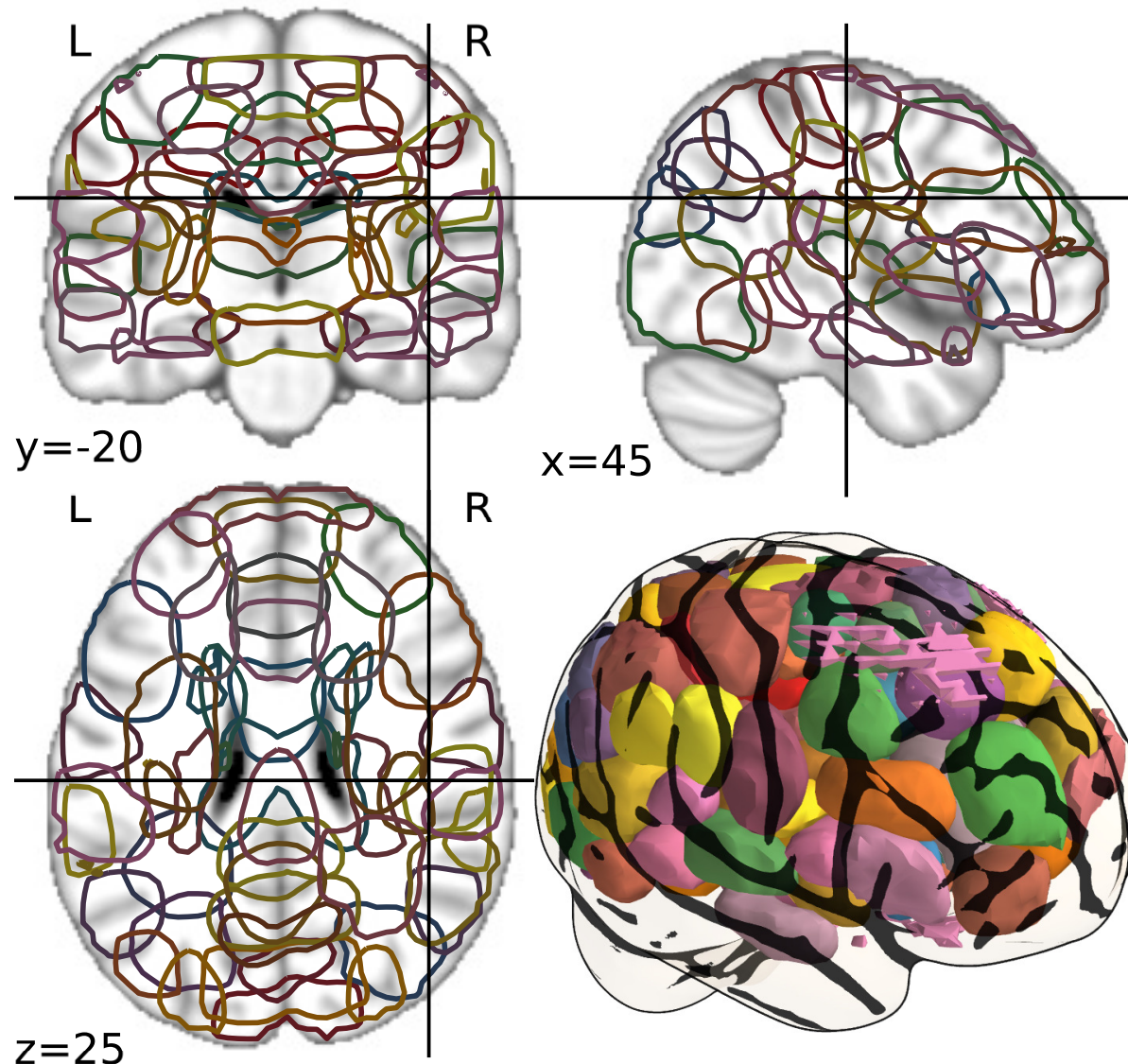
Application to face databases (3/3)

- Quantitative performance evaluation on classification task



Structured sparse PCA on resting state activity

(Varoquaux, Jenatton, Gramfort, Obozinski, Thirion, and Bach, 2010)



Dictionary learning vs. sparse structured PCA

Exchange roles of X and w

- Sparse structured PCA (**structured dictionary elements**):

$$\min_{W \in \mathbb{R}^{k \times n}, X \in \mathbb{R}^{p \times k}} \frac{1}{n} \sum_{i=1}^n \|y^i - X w^i\|_2^2 + \lambda \sum_{j=1}^k \Omega(x^j) \text{ s.t. } \forall i, \|w^i\|_2 \leq 1.$$

- Dictionary learning with **structured sparsity for codes** w :

$$\min_{W \in \mathbb{R}^{k \times n}, X \in \mathbb{R}^{p \times k}} \frac{1}{n} \sum_{i=1}^n \|y^i - X w^i\|_2^2 + \lambda \Omega(w^i) \text{ s.t. } \forall j, \|x^j\|_2 \leq 1.$$

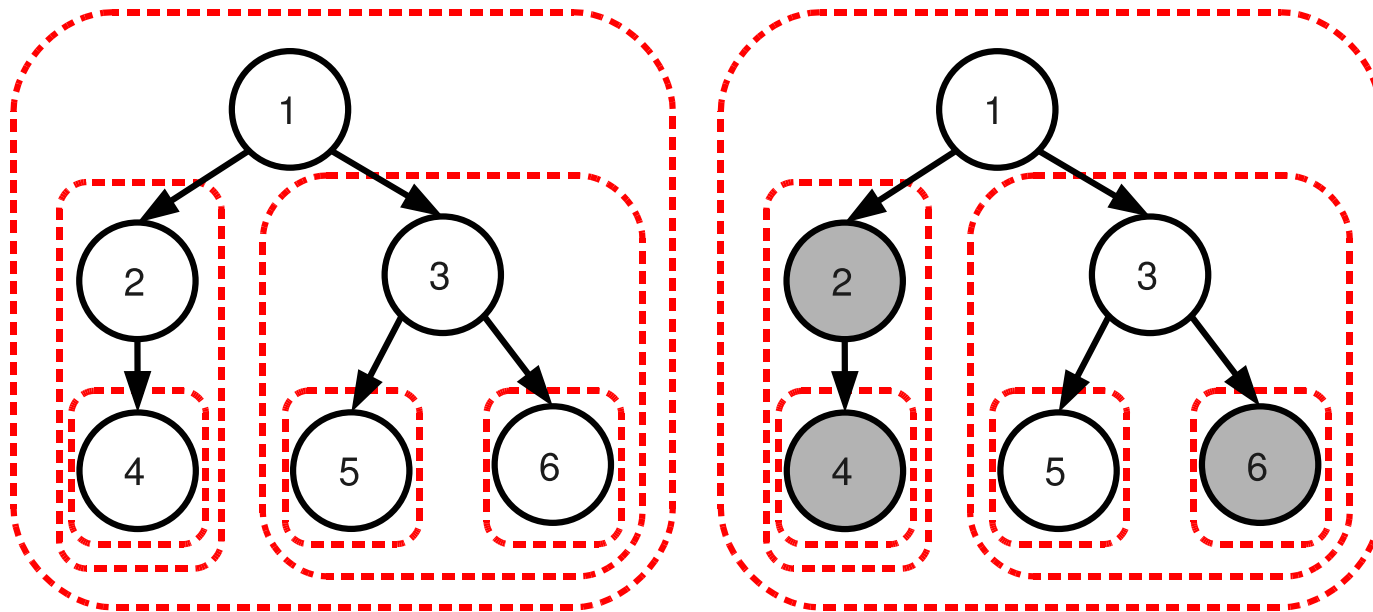
- **Optimization:**

- Alternating optimization
- **Modularity of implementation** if proximal step is efficient (Jenatton et al., 2010; Mairal et al., 2010)

Hierarchical dictionary learning

(Jenatton, Mairal, Obozinski, and Bach, 2010)

- Structure on codes w (not on dictionary X)
- Hierarchical penalization: $\Omega(w) = \sum_{G \in \mathbf{H}} \|w_G\|_2$ where groups G in \mathbf{H} are equal to **set of descendants** of some nodes in a tree



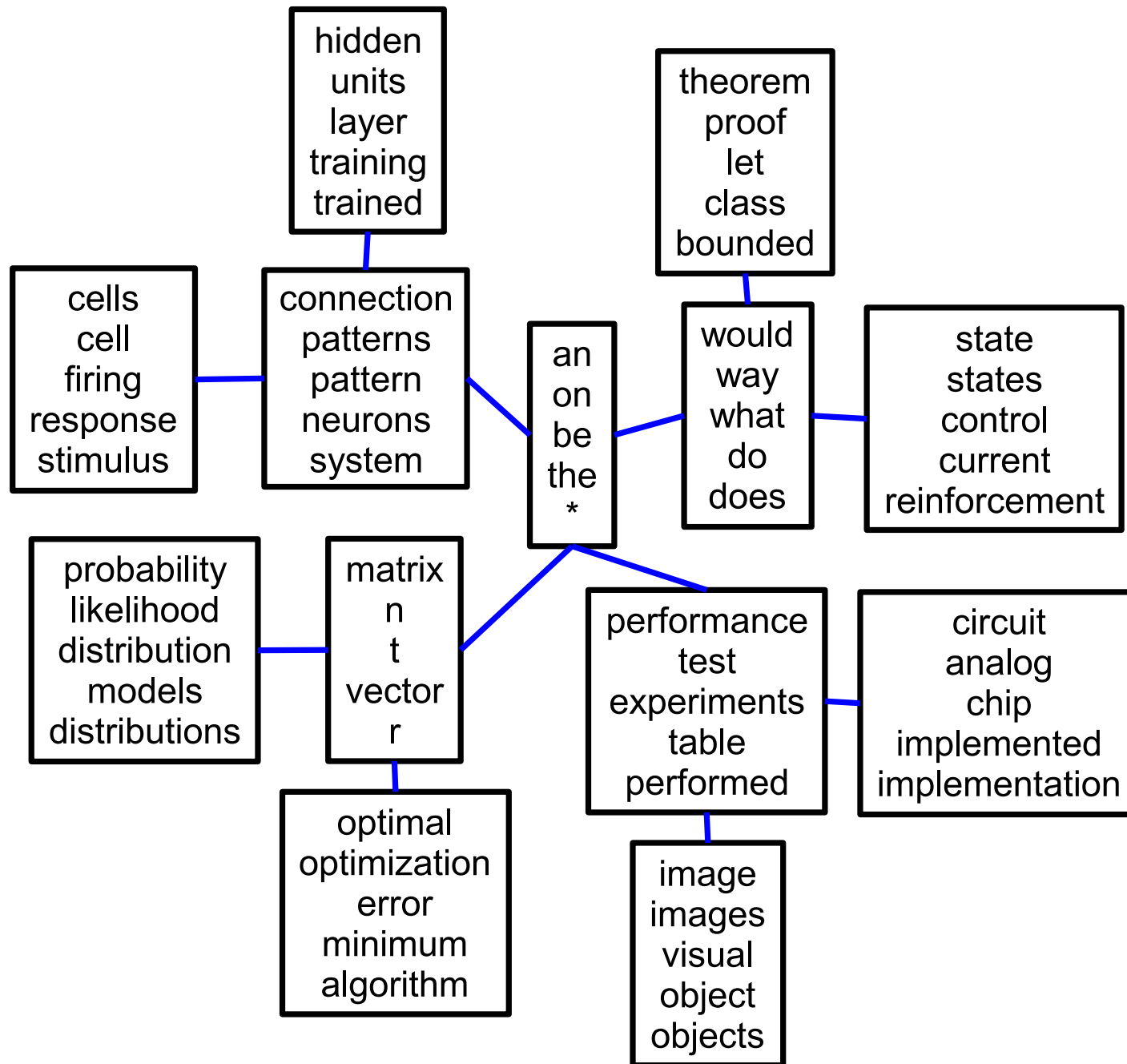
- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008)

Hierarchical dictionary learning

Modelling of text corpora

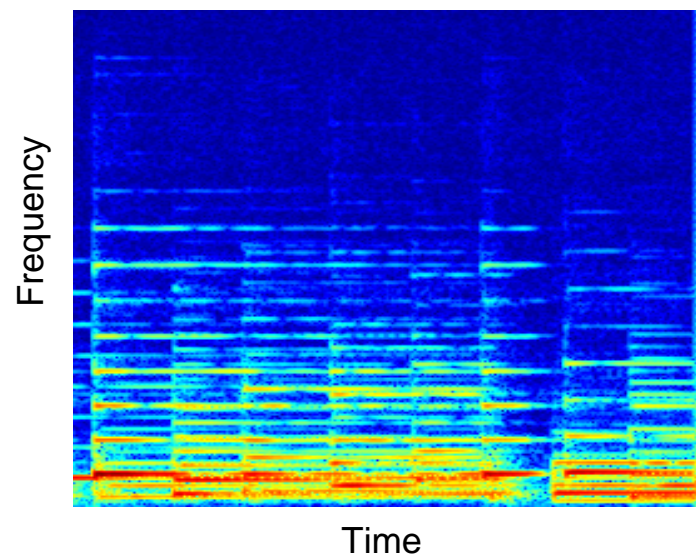
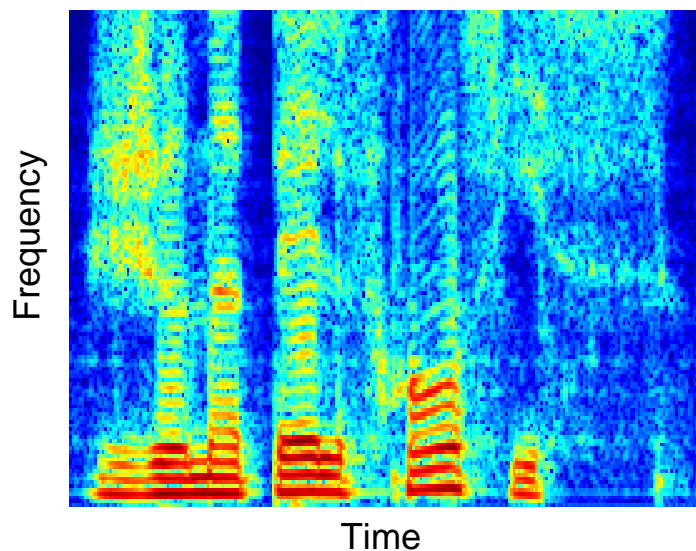
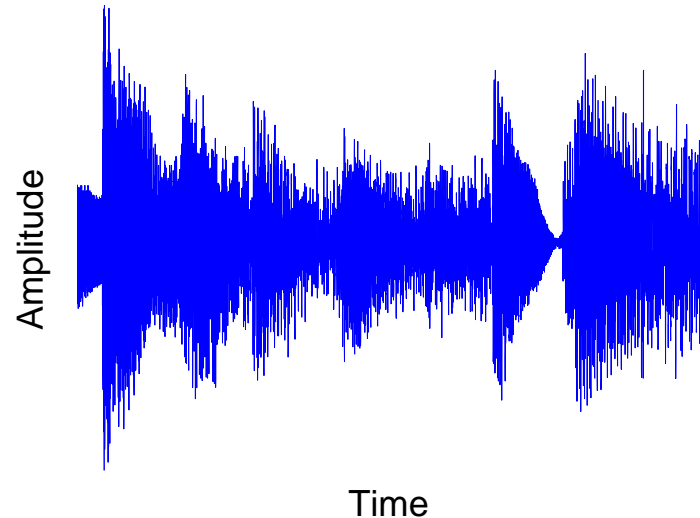
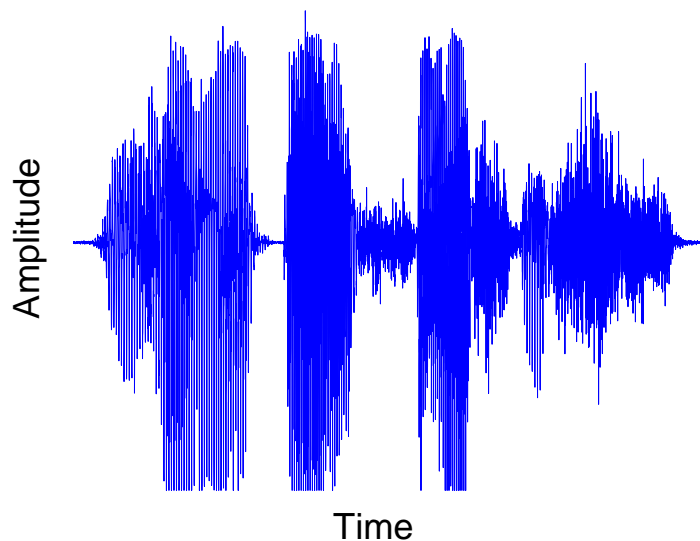
- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003)
 - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
 - **Can we achieve similar performance with simple matrix factorization formulation?**

Modelling of text corpora - Dictionary tree



Structured sparsity - **Audio processing**

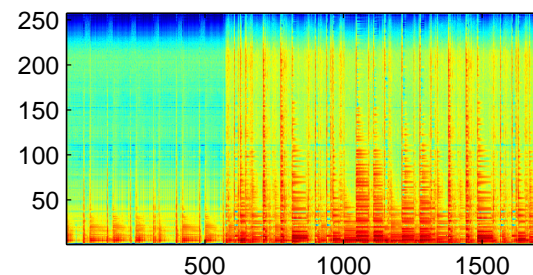
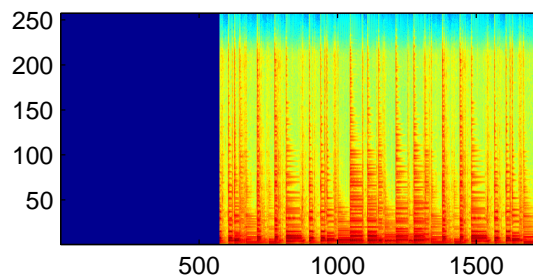
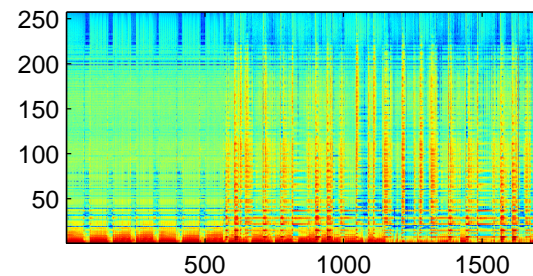
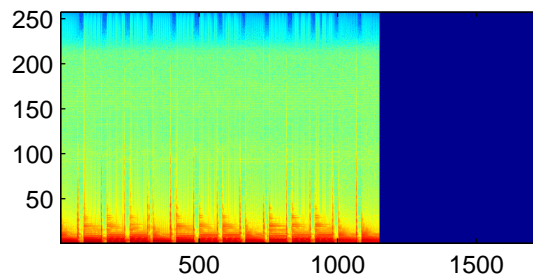
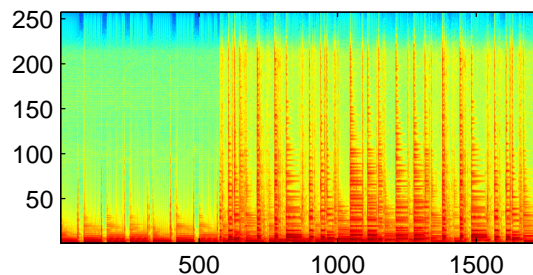
Source separation (Lefèvre et al., 2011)



Structured sparsity - Audio processing

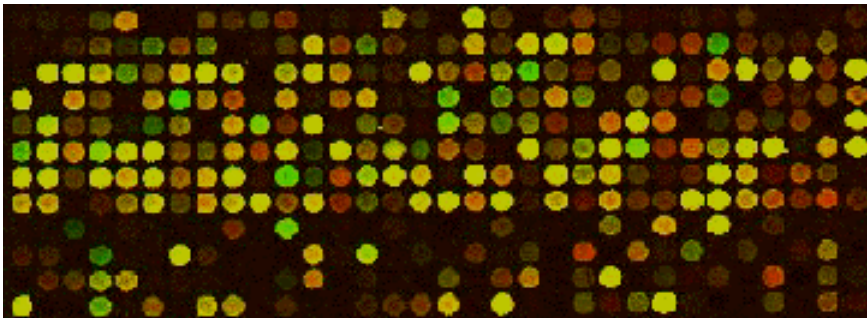
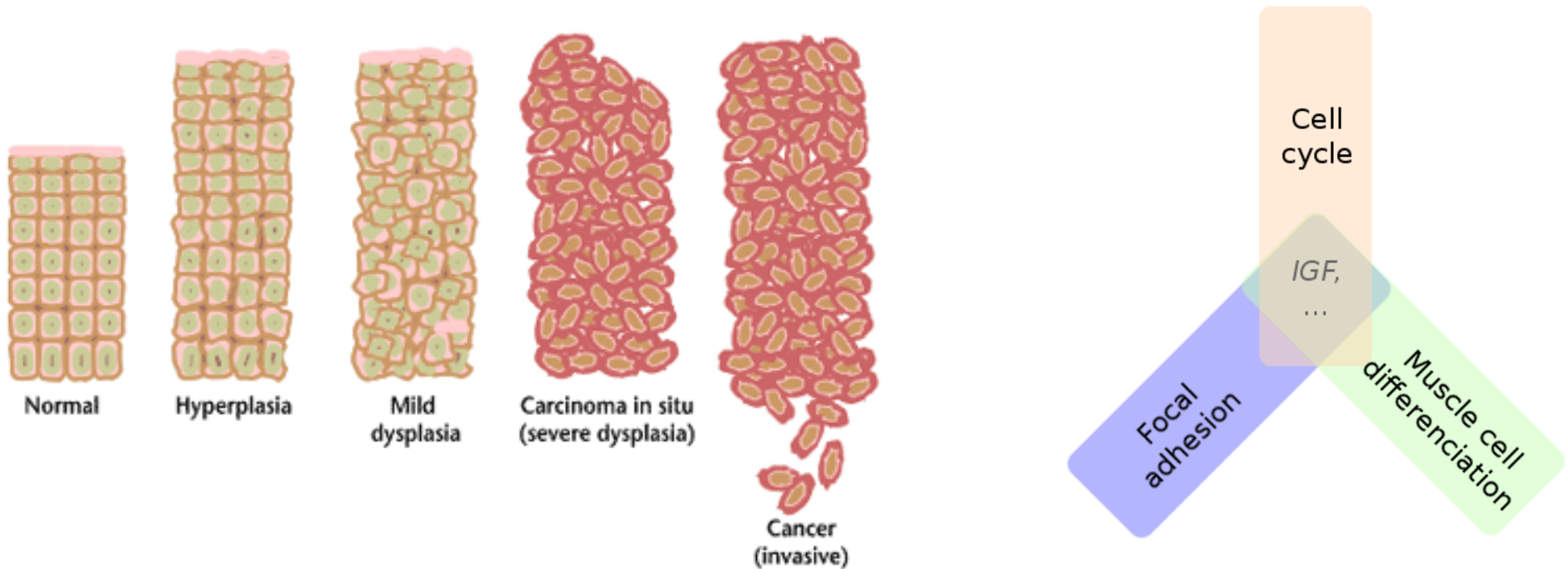
Musical instrument separation (Lefèvre et al., 2011)

- Unsupervised source separation with group-sparsity prior
 - Top: mixture
 - Left: source tracks (guitar, voice). Right: separated tracks.



Structured sparsity - Bioinformatics

- Collaboration with J.-P. Vert, Institut Curie (T. Hocking, G. Obozinski, F. Bach)
- Metastasis prediction from microarray data (G. Obozinski)



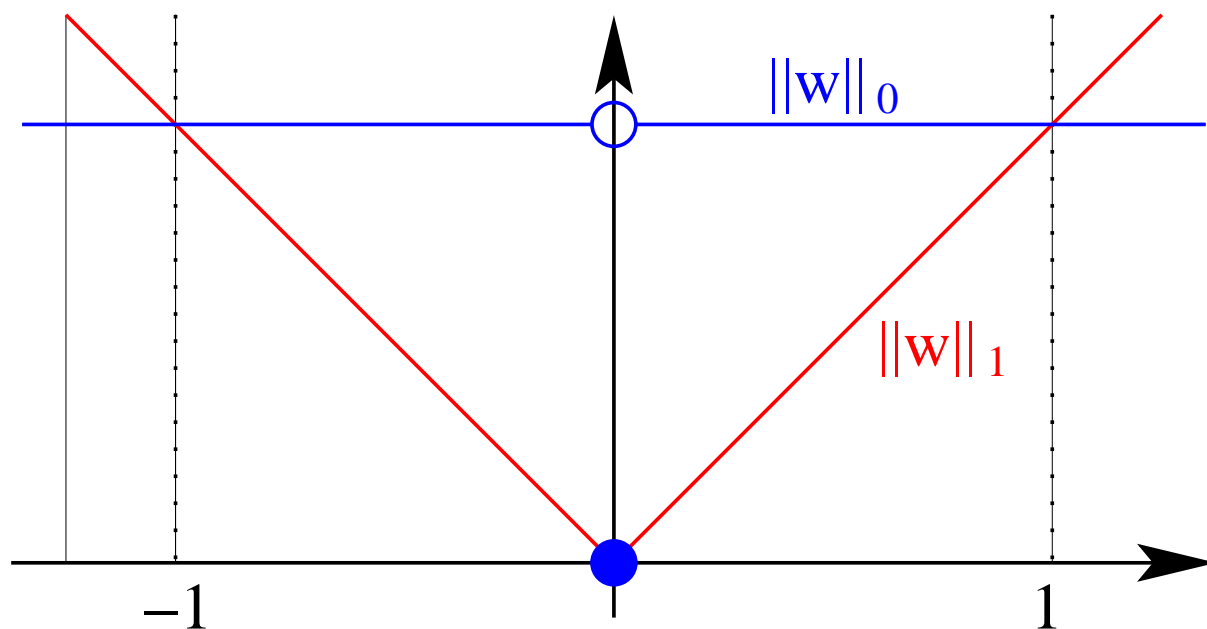
- Biological pathways
- Dedicated sparsity-inducing norm for better interpretability and prediction

Outline

- **SIERRA project-team**
- **Introduction: Sparse methods for machine learning**
 - Need for structured sparsity: **Going beyond the ℓ_1 -norm**
- **Classical approaches to structured sparsity**
 - Linear combinations of ℓ_q -norms
- **Structured sparsity through submodular functions**
 - Relaxation of the penalization of supports
 - **Unified algorithms and analysis**

ℓ_1 -norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\text{Supp}(w) = \{j \in V, w_j \neq 0\}$
- **Cardinality of support:** $\|w\|_0 = \text{Card}(\text{Supp}(w))$
- Convex envelope = largest convex lower bound (see, e.g., Boyd and Vandenberghe, 2004)



- ℓ_1 -norm = convex envelope of ℓ_0 -quasi-norm on the ℓ_∞ -ball $[-1, 1]^p$

Convex envelopes of general functions of the support (Bach, 2010)

- Let $F : 2^V \rightarrow \mathbb{R}$ be a **set-function**
 - Assume F is **non-decreasing** (i.e., $A \subset B \Rightarrow F(A) \leq F(B)$)
 - Explicit prior knowledge on supports (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)
- Define $\Theta(w) = F(\text{Supp}(w))$: **How to get its convex envelope?**
 1. Possible if F is also **submodular**
 2. Allows **unified** theory and algorithm
 3. Provides **new** regularizers

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$$

$$\Leftrightarrow \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$$

$$\Leftrightarrow \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1:** defined like concave functions (“diminishing returns”)
 - Example: $F : A \mapsto g(\text{Card}(A))$ is submodular if g is concave

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$$

$$\Leftrightarrow \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1:** defined like concave functions (“diminishing returns”)
 - Example: $F : A \mapsto g(\text{Card}(A))$ is submodular if g is concave
- **Intuition 2:** behave like convex functions
 - Polynomial-time minimization, conjugacy theory

Submodular functions (Fujishige, 2005; Bach, 2010b)

- $F : 2^V \rightarrow \mathbb{R}$ is **submodular** if and only if

$$\forall A, B \subset V, \quad F(A) + F(B) \geq F(A \cap B) + F(A \cup B)$$

$$\Leftrightarrow \forall k \in V, \quad A \mapsto F(A \cup \{k\}) - F(A) \text{ is non-increasing}$$

- **Intuition 1:** defined like concave functions (“diminishing returns”)
 - Example: $F : A \mapsto g(\text{Card}(A))$ is submodular if g is concave
- **Intuition 2:** behave like convex functions
 - Polynomial-time minimization, conjugacy theory
- Used in several areas of signal processing and machine learning
 - Total variation/graph cuts (Chambolle, 2005; Boykov et al., 2001)
 - Optimal design (Krause and Guestrin, 2005)

Submodular functions - Examples

- Concave functions of the cardinality: $g(|A|)$
- Cuts
- Entropies
 - $H((X_k)_{k \in A})$ from p random variables X_1, \dots, X_p
- Network flows
 - Efficient representation for set covers
- Rank functions of matroids

Submodular functions - Lovász extension

- Subsets may be identified with elements of $\{0, 1\}^p$
- Given **any** set-function F and w such that $w_{j_1} \geq \dots \geq w_{j_p}$, define:

$$f(w) = \sum_{k=1}^p w_{j_k} [F(\{j_1, \dots, j_k\}) - F(\{j_1, \dots, j_{k-1}\})]$$

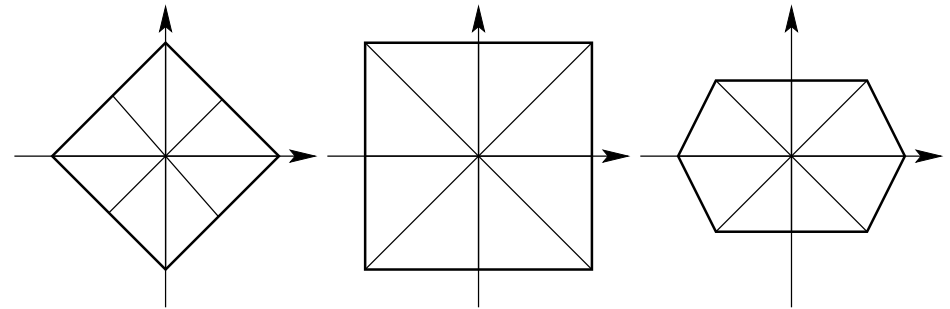
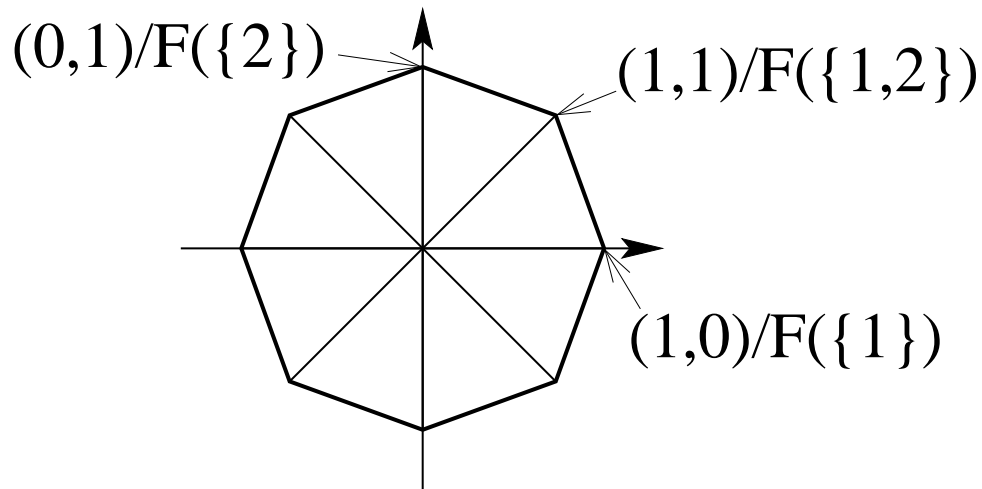
- If $w = 1_A$, $f(w) = F(A) \Rightarrow$ extension from $\{0, 1\}^p$ to \mathbb{R}^p
- f is piecewise affine and positively homogeneous
- **F is submodular if and only if f is convex** (Lovász, 1982)
 - Minimizing $f(w)$ on $w \in [0, 1]^p$ equivalent to minimizing F on 2^V

Submodular functions and structured sparsity

- Let $F : 2^V \rightarrow \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition:** the convex envelope of $\Theta : w \mapsto F(\text{Supp}(w))$ on the ℓ_∞ -ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F

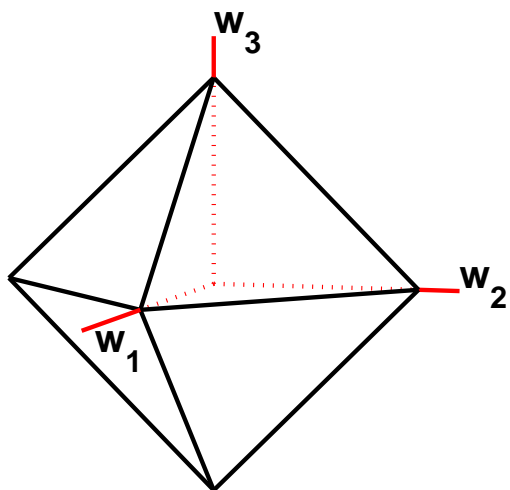
Submodular functions and structured sparsity

- Let $F : 2^V \rightarrow \mathbb{R}$ be a **non-decreasing submodular set-function**
- **Proposition:** the convex envelope of $\Theta : w \mapsto F(\text{Supp}(w))$ on the ℓ_∞ -ball is $\Omega : w \mapsto f(|w|)$ where f is the Lovász extension of F
- **Sparsity-inducing properties:** Ω is a **polyhedral** norm



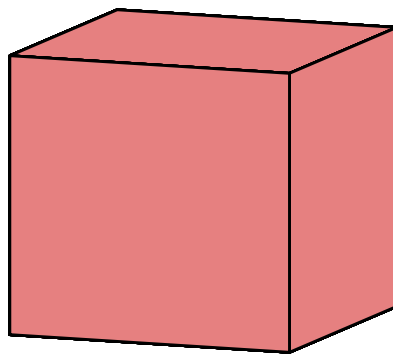
- A is stable if for all $B \supset A$, $B \neq A \Rightarrow F(B) > F(A)$
- With probability one, stable sets are the only allowed active sets

Polyhedral unit balls



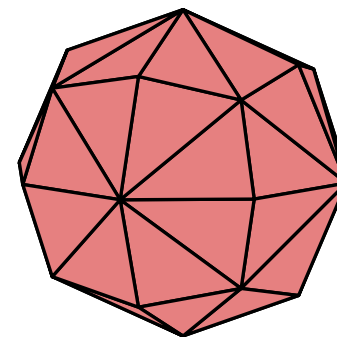
$$F(A) = |A|$$

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



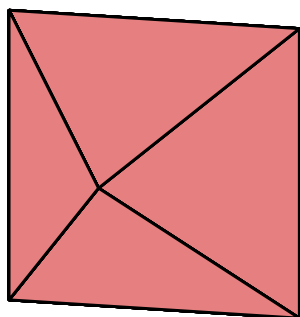
$$F(A) = \min\{|A|, 1\}$$

$$\Omega(w) = \|w\|_\infty$$



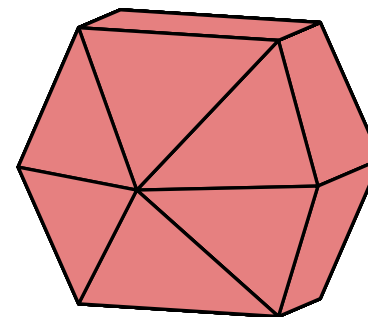
$$F(A) = |A|^{1/2}$$

all possible extreme points



$$F(A) = 1_{\{A \cap \{1\} \neq \emptyset\}} + 1_{\{A \cap \{2,3\} \neq \emptyset\}}$$

$$\Omega(w) = |w_1| + \|w_{\{2,3\}}\|_\infty$$



$$F(A) = 1_{\{A \cap \{1,2,3\} \neq \emptyset\}} + 1_{\{A \cap \{2,3\} \neq \emptyset\}} + 1_{\{A \cap \{3\} \neq \emptyset\}}$$

$$\Omega(w) = \|w\|_\infty + \|w_{\{2,3\}}\|_\infty + |w_3|$$

Submodular functions and structured sparsity

- **Unified theory and algorithms**

- Generic computation of proximal operator
- Unified oracle inequalities

- **Extensions**

- Shaping level sets through symmetric submodular function (Bach, 2010a)
- ℓ_q -relaxations of combinatorial penalties (Obozinski and Bach, 2011)

Conclusion

- **Structured sparsity for machine learning and statistics**
 - Many applications (image, audio, text, etc.)
 - May be achieved through structured sparsity-inducing norms
 - Link with submodular functions: unified analysis and algorithms

Conclusion

- **Structured sparsity for machine learning and statistics**
 - Many applications (image, audio, text, etc.)
 - May be achieved through structured sparsity-inducing norms
 - Link with submodular functions: unified analysis and algorithms
- **On-going/related work on structured sparsity**
 - **Norm design** beyond submodular functions
 - Complementary approach of Jacob, Obozinski, and Vert (2009)
 - Theoretical analysis of dictionary learning (Jenatton, Bach and Gribonval, 2011)
 - Achieving $\log p = O(n)$ algorithmically (Bach, 2008)

INRIA and machine learning

- **Machine learning is a relatively recent field**
 - Between applied mathematics and computer science
 - INRIA is a key actor (core ML + interactions)
- **What INRIA can do for machine learning**
 - **Junior researcher positions (CR)**
 - Invited professors

References

- F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008.
- F. Bach. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.
- F. Bach. Shaping level sets with submodular functions. Technical Report 00542949, HAL, 2010a.
- F. Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010b.
- F. Bach and E. Moulines. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. 2011.
- F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. Technical Report 00613125, HAL, 2011.
- R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, arXiv:0808.3572, 2008.
- 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. *The Journal of Machine Learning Research*, 3:993–1022, January 2003.
- D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*, 16:106, 2004.
- S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

- Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI*, 23(11):1222–1239, 2001.
- V. Cevher, M. F. Duarte, C. Hegde, and R. G. Baraniuk. Sparse signal recovery using markov random fields. In *Advances in Neural Information Processing Systems*, 2008.
- A. Chambolle. Total variation minimization and a class of binary MRF models. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 136–152. Springer, 2005.
- S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image Processing*, 15(12):3736–3745, 2006.
- S. Fujishige. *Submodular Functions and Optimization*. Elsevier, 2005.
- A. Gramfort and M. Kowalski. Improving M/EEG source localization with an inter-condition sparse prior. In *IEEE International Symposium on Biomedical Imaging*, 2009.
- J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Transactions on Information Theory*, 52(9):4036–4048, 2006.
- J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th 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 26th International Conference on Machine Learning (ICML)*, 2009.
- R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009a.

- R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009b.
- R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.
- 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. Technical report, Preprint arXiv:1105.0363, 2011. In submission to SIAM Journal on Imaging Sciences.
- K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of CVPR*, 2009.
- S. Kim and E. P. Xing. Tree-guided group Lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.
- A. Krause and C. Guestrin. Near-optimal nonmyopic value of information in graphical models. In *Proc. UAI*, 2005.
- L. Lovász. Submodular functions and convexity. *Mathematical programming: the state of the art, Bonn*, pages 235–257, 1982.
- J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. Technical report, arXiv:0908.0050, 2009a.
- J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009b.
- J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised dictionary learning. *Advances*

in *Neural Information Processing Systems (NIPS)*, 21, 2009c.

- J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *NIPS*, 2010.
- Y. Nesterov. Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Tech. Rep*, 76, 2007.
- G. Obozinski and F. Bach. Convex relaxation of combinatorial penalties. Technical report, HAL, 2011.
- 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.
- F. Rapaport, E. Barillot, and J.-P. Vert. Classification of arrayCGH data using fused SVM. *Bioinformatics*, 24(13):i375–i382, Jul 2008.
- M. Schmidt, N. Le Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. *Arxiv preprint arXiv:1109.2415*, 2011.
- R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.
- G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach. 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.
- 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, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.