Structured sparsity through convex optimization

Francis Bach

INRIA - Ecole Normale Supérieure, Paris, France

Joint work with R. Jenatton, J. Mairal, G. Obozinski, J. Ponce - ICVSS, July 2012

Outline

Sparse methods for machine learning and computer vision

- **Introduction**

- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis

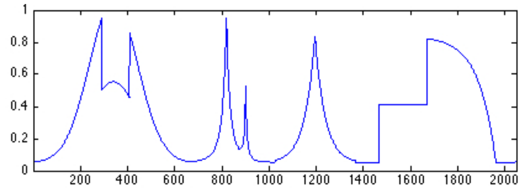
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering

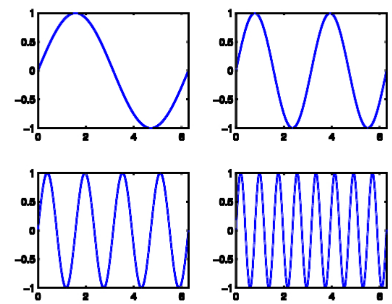
- **Sparsity for computer vision**
  - Task-driven dictionary learning

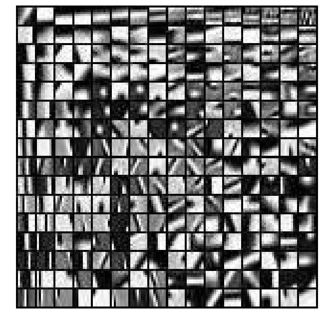
- **Structured sparsity**

Sparsity in signal processing

- Let $x \in \mathbb{R}^m$ be a signal







- Let $D = [d_1, \ldots, d_p] \in \mathbb{R}^{m \times p}$ be a set of normalized "basis vectors". We call it **dictionary**
- D is "adapted" to x if it can represent it with a few basis vectors:
  - there exists a sparse vector α in $\mathbb{R}^p$ such that $x \approx D\alpha$. We call α the **sparse code**.

$$\underbrace{\begin{pmatrix} x \end{pmatrix}}_{x \in \mathbb{R}^m} \approx \underbrace{\begin{pmatrix} d_1 & \big| & d_2 & \big| & \cdots & \big| & d_p \end{pmatrix}}_{D \in \mathbb{R}^{m \times p}} \underbrace{\begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix}}_{\alpha \in \mathbb{R}^p,\ \text{sparse}}$$

Sparsity in signal processing
Sparse decomposition problem

$$\min_{\alpha \in \mathbb{R}^p} \underbrace{\frac{1}{2}||x - D\alpha||_2^2}_{\text{data fitting term}} + \underbrace{\lambda\psi(\alpha)}_{\substack{\text{sparsity-inducing} \\ \text{regularization}}}$$

- The term ψ induces sparsity
  - the ℓ_0 "pseudo-norm": $||\alpha||_0 \triangleq \#\{i \ \text{ s.t. } \ \alpha_i \neq 0\}$ (NP-hard)
  - the ℓ_1 norm: $||\alpha||_1 \triangleq \sum_{i=1}^{p} |\alpha_i|$ (convex)
  - ...

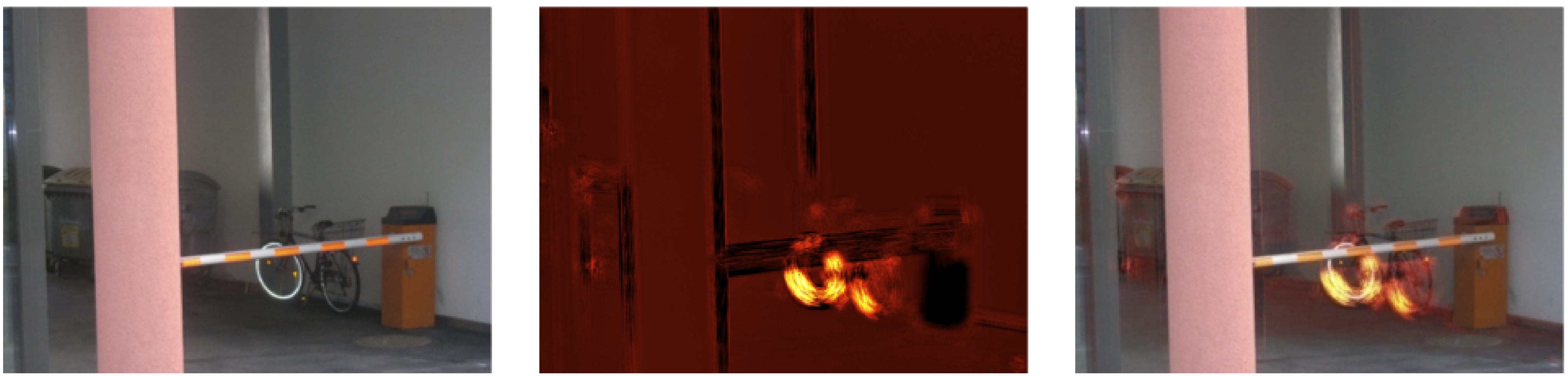
Sparsity in signal processing

- Simultaneously denoise all patches of a given image
- Example from Mairal, Bach, Ponce, Sapiro, and Zisserman (2009d)

Sparsity in signal processing
Applications to computer vision

- Uses the "code" α as representation of observations for subsequent processing (Raina et al., 2007; Yang et al., 2009b)

- Adapt dictionary elements to specific tasks (Mairal et al., 2009c)
  - Discriminative training for weakly supervised pixel classification (Mairal et al., 2008a)



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:

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

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: $\boxed{\log p = O(n)}$

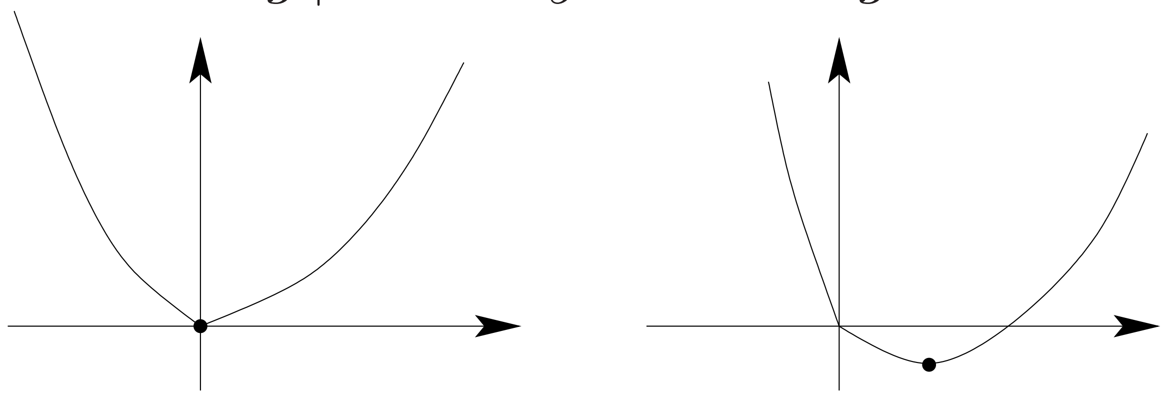
Outline

Sparse methods for machine learning and computer vision

- **Introduction**
- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering
- **Sparsity for computer vision**
  - Task-driven dictionary learning
- **Structured sparsity**

Why ℓ_1-norms lead to sparsity?

- **Example 1**: quadratic problem in 1D, i.e. $\boxed{\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|}$
- Piecewise quadratic function with a kink at zero
  - Derivative at $0+$: $g_+ = \lambda - y$ and $0-$: $g_- = -\lambda - y$

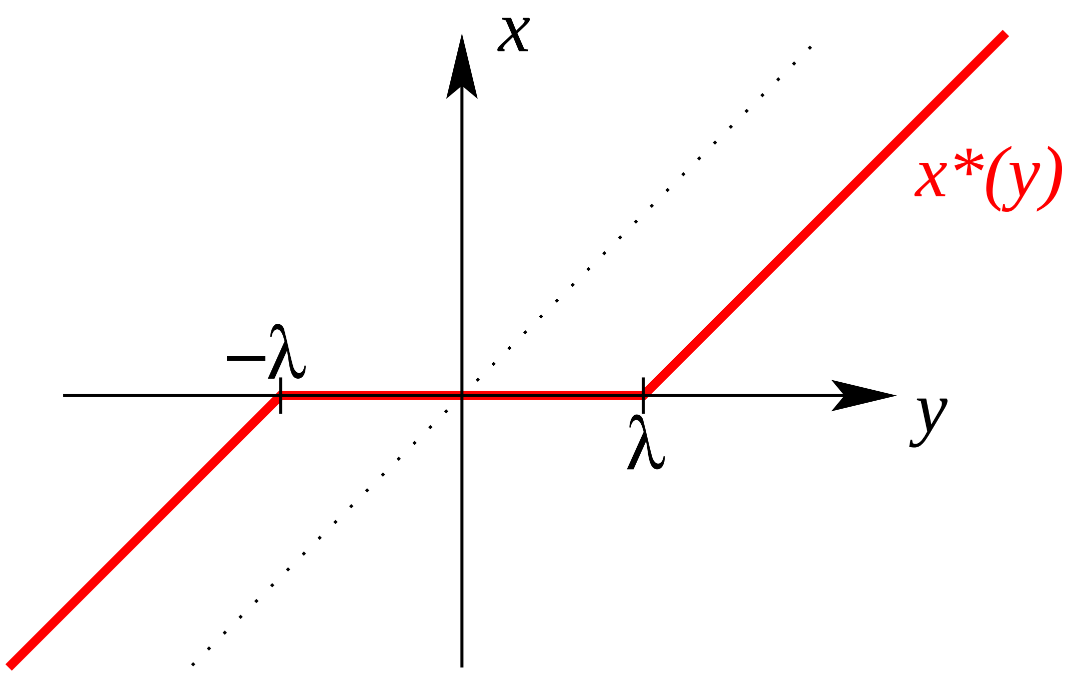


  - $x = 0$ is the solution iff $g_+ \geqslant 0$ and $g_- \leqslant 0$ (i.e., $|y| \leqslant \lambda$)
  - $x \geqslant 0$ is the solution iff $g_+ \leqslant 0$ (i.e., $y \geqslant \lambda$) $\Rightarrow x^* = y - \lambda$
  - $x \leqslant 0$ is the solution iff $g_- \leqslant 0$ (i.e., $y \leqslant -\lambda$) $\Rightarrow x^* = y + \lambda$

- Solution $\boxed{x^* = \mathrm{sign}(y)(|y| - \lambda)_+}$ = soft thresholding

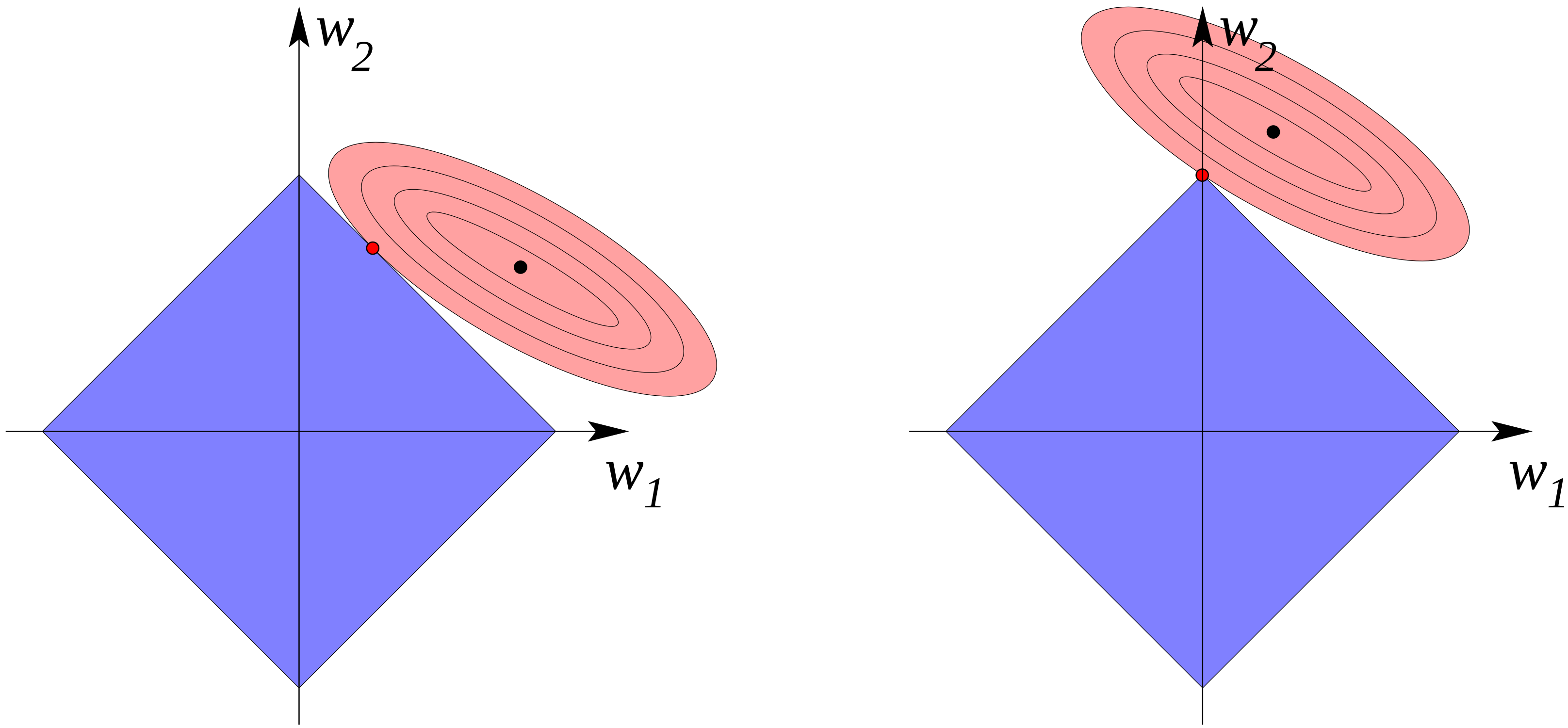
Why ℓ_1-norms lead to sparsity?

- **Example 1**: quadratic problem in 1D, i.e. $\min_{x \in \mathbb{R}} \frac{1}{2}x^2 - xy + \lambda|x|$
- Piecewise quadratic function with a kink at zero
- Solution $x^* = \text{sign}(y)(|y| - \lambda)_+$ = soft thresholding



Why ℓ_1-norms lead to sparsity?

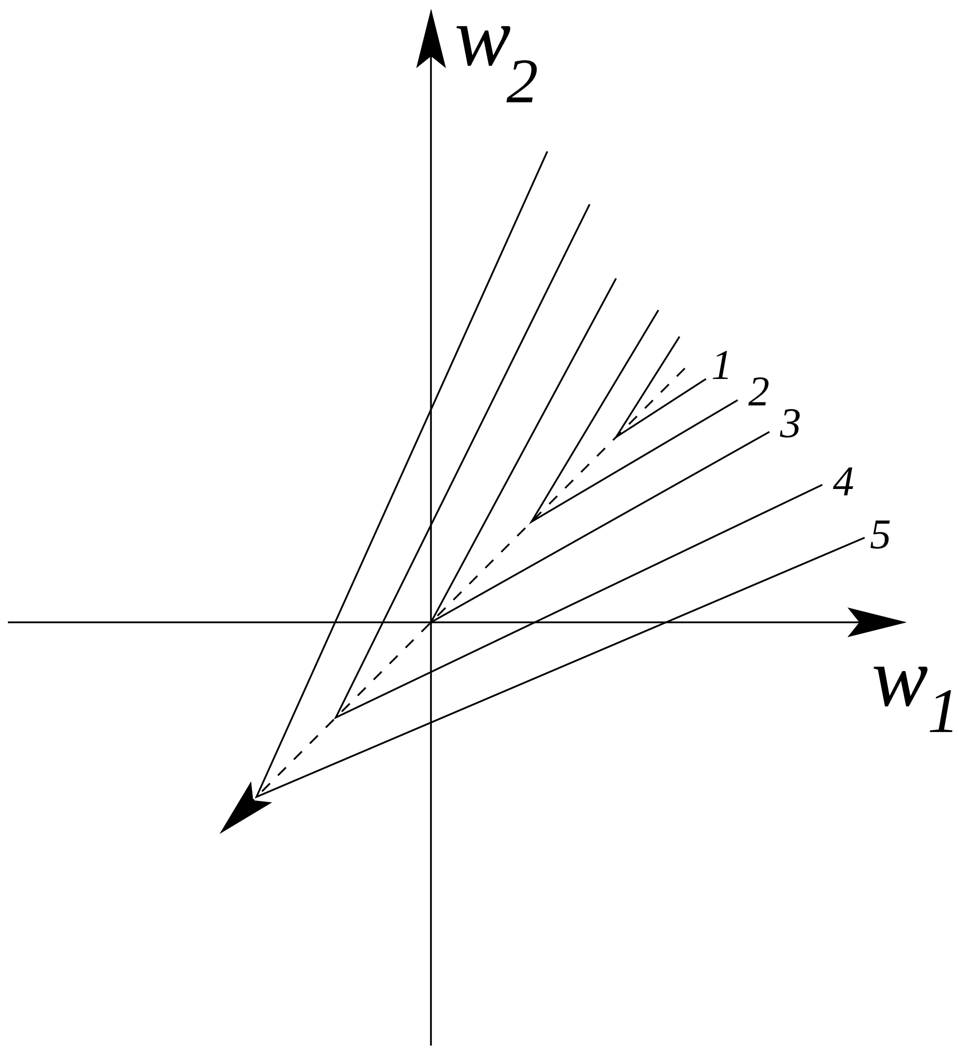
- **Example 2**: minimize quadratic function $Q(w)$ subject to $\|w\|_1 \leqslant T$.
  - coupled soft thresholding
- Geometric interpretation
  - NB : penalizing is "equivalent" to constraining



Non-smooth optimization

- **Simple techniques might not work!**
  - Gradient descent or coordinate descent
- **Special tools**
  - Subgradients or directional derivatives
- Typically slower than smooth optimization...
- ... except in some regularized problems

Counter-example
Coordinate descent for nonsmooth objectives



Regularized problems - Proximal methods

- Gradient descent as a proximal method (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{\mu}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{\mu} \nabla L(w_t)$

Regularized problems - Proximal methods

- Gradient descent as a proximal method (differentiable functions)
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \frac{\mu}{2}\|w - w_t\|_2^2$
  - $w_{t+1} = w_t - \frac{1}{\mu}\nabla L(w_t)$

- Problems of the form: $\boxed{\min_{w \in \mathbb{R}^p} L(w) + \lambda\Omega(w)}$
  - $w_{t+1} = \arg\min_{w \in \mathbb{R}^p} L(w_t) + (w - w_t)^\top \nabla L(w_t) + \lambda\Omega(w) + \frac{\mu}{2}\|w - w_t\|_2^2$
  - Thresholded gradient descent $w_{t+1} = \text{SoftThres}(w_t - \frac{1}{\mu}\nabla L(w_t))$

- Similar convergence rates than smooth optimization
  - Acceleration methods (Nesterov, 2007; Beck and Teboulle, 2009)
  - **depends on the condition number of the loss**

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**
- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding

Cheap (and not dirty) algorithms for all losses

- **Proximal methods**
- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding
- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 2009b)
  - Notice that $\sum_{j=1}^p |w_j| = \min_{\eta \geqslant 0} \frac{1}{2} \sum_{j=1}^p \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to η (closed-form $\eta_j = |w_j|$) and w (weighted squared ℓ_2-norm regularized problem)
  - Caveat: lack of continuity around $(w_i, \eta_i) = (0, 0)$: add ε / η_j

Cheap (and not dirty) algorithms for all losses

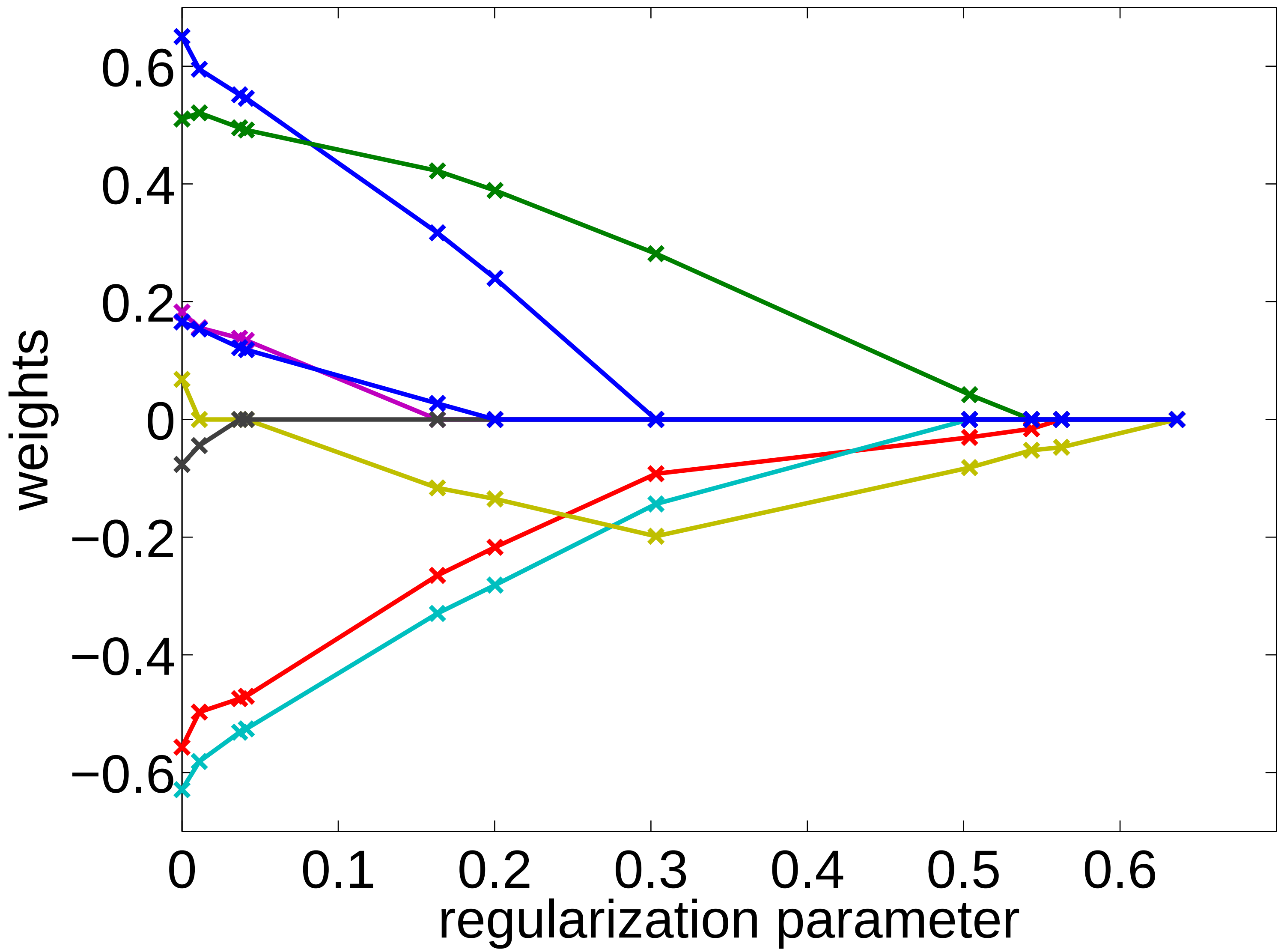
- **Proximal methods**

- **Coordinate descent** (Fu, 1998; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding

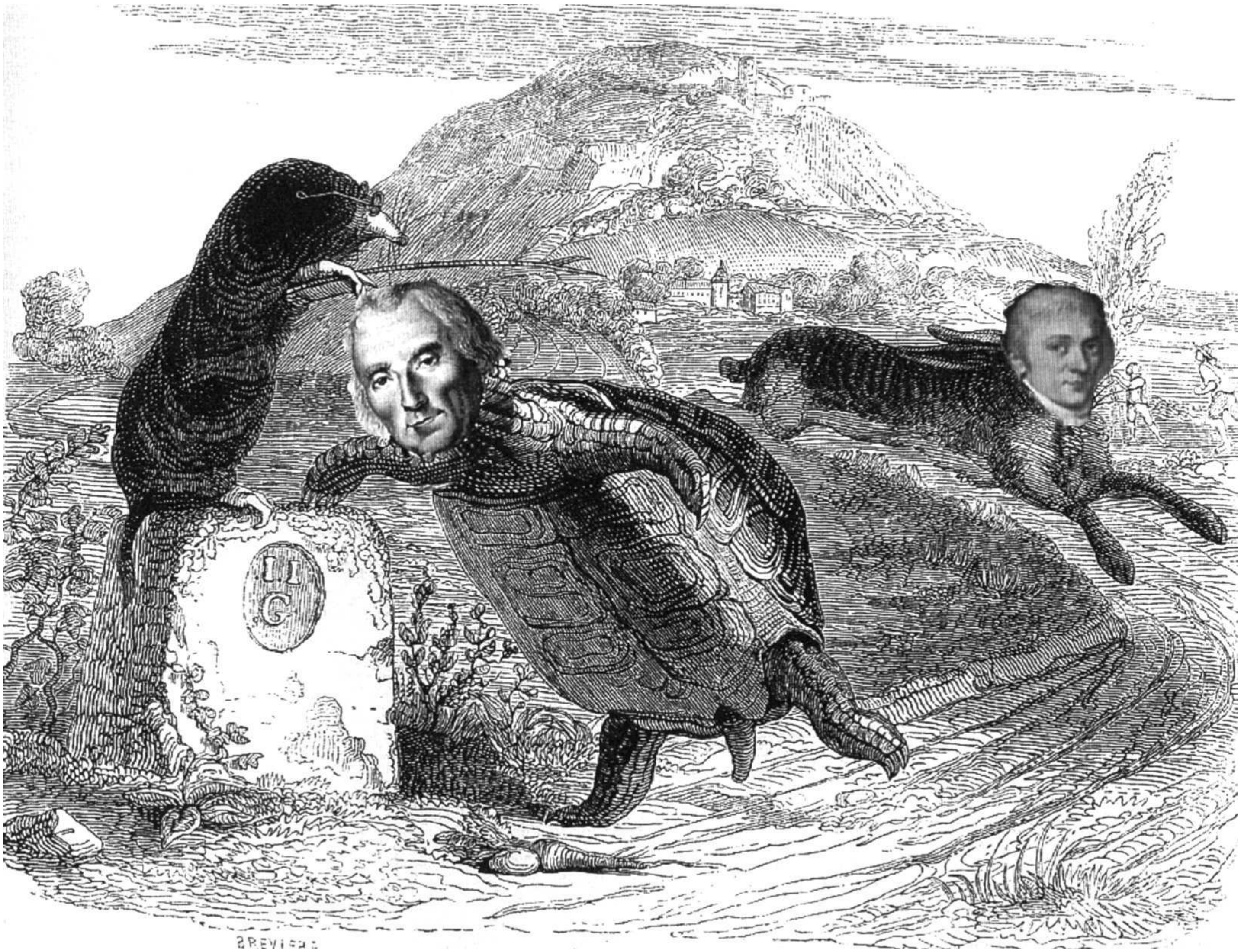
- **"η-trick"** (Rakotomamonjy et al., 2008; Jenatton et al., 2009b)
  - Notice that $\sum_{j=1}^p |w_j| = \min_{\eta \geqslant 0} \frac{1}{2} \sum_{j=1}^p \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to η (closed-form $\eta_j = |w_j|$) and w (weighted squared ℓ_2-norm regularized problem)
  - Caveat: lack of continuity around $(w_i, \eta_i) = (0, 0)$: add ε / η_i

- **Dedicated algorithms that use sparsity** (active sets/homotopy)

Piecewise linear paths



Gaussian hare vs. Laplacian tortoise



- Coord. descent and proximal: $O(pn)$ per iterations for ℓ_1 and ℓ_2
- "Exact" algorithms: $O(kpn)$ for ℓ_1 **vs.** $O(p^2n)$ for ℓ_2

Additional methods - Softwares

- Many contributions in signal processing, optimization, mach. learning
  - Extensions to stochastic setting (Bottou and Bousquet, 2008)

- **Extensions to other sparsity-inducing norms**
  - Computing proximal operator
  - F. Bach, R. Jenatton, J. Mairal, G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1-106, 2011.

- **Softwares**
  - Many available codes
  - SPAMS (SPArse Modeling Software)
    `http://www.di.ens.fr/willow/SPAMS/`

Outline

Sparse methods for machine learning and computer vision

- **Introduction**
- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering
- **Sparsity for computer vision**
  - Task-driven dictionary learning
- **Structured sparsity**

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if there are low correlations between relevant and irrelevant variables.

Model selection consistency (Lasso)

- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern

- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if
$$\boxed{\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\operatorname{sign}(\mathbf{w}_\mathbf{J})\|_\infty \leqslant 1}$$
where $\mathbf{Q} = \lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \operatorname{Supp}(\mathbf{w})$

Model selection consistency (Lasso)

- Assume $\mathbf{w}$ sparse and denote $\mathbf{J} = \{j, \mathbf{w}_j \neq 0\}$ the nonzero pattern

- **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if

$$\boxed{\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{JJ}}^{-1}\mathrm{sign}(\mathbf{w}_{\mathbf{J}})\|_\infty \leqslant 1}$$

  where $\mathbf{Q} = \lim_{n\to+\infty} \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$ and $\mathbf{J} = \mathrm{Supp}(\mathbf{w})$

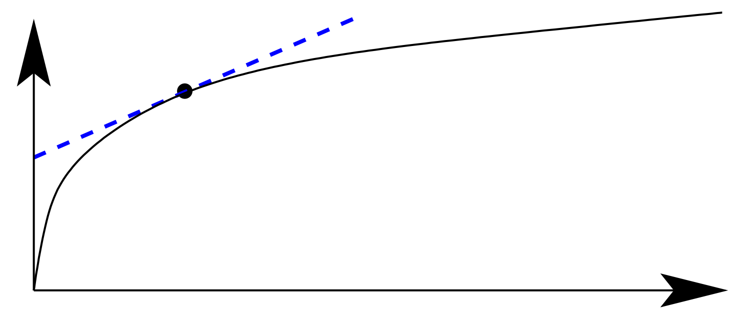
- **The Lasso is usually not model-consistent**
  - Selects more variables than necessary (see, e.g., Lv and Fan, 2009)
  - **Fixing the Lasso**: adaptive Lasso (Zou, 2006), relaxed Lasso (Meinshausen, 2008), thresholding (Lounici, 2008), Bolasso (Bach, 2008a), stability selection (Meinshausen and Bühlmann, 2008), Wasserman and Roeder (2009)

Adaptive Lasso and concave penalization

- **Adaptive Lasso** (Zou, 2006; Huang et al., 2008)
  - Weighted ℓ_1-norm: $\min_{w \in \mathbb{R}^p} L(w) + \lambda \sum_{j=1}^{p} \frac{|w_j|}{|\hat{w}_j|^{\alpha}}$
  - $\hat{w}$ estimator obtained from ℓ_2 or ℓ_1 regularization

- **Reformulation in terms of concave penalization**

$$\min_{w \in \mathbb{R}^p} L(w) + \sum_{j=1}^{p} g(|w_j|)$$



  - Example: $g(|w_j|) = |w_j|^{1/2}$ or $\log |w_j|$. Closer to the ℓ_0 penalty
  - Concave-convex procedure: replace $g(|w_j|)$ by affine upper bound
  - Better sparsity-inducing properties (Fan and Li, 2001; Zou and Li, 2008; Zhang, 2008b)

Lasso - Two main recent theoretical results

1. **Support recovery condition** (Zhao and Yu, 2006; Wainwright, 2009; Zou, 2006; Yuan and Lin, 2007): the Lasso is sign-consistent if and only if there are low correlations between relevant and irrelevant variables.

2. **Exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2009; Bickel et al., 2009; Lounici, 2008; Meinshausen and Yu, 2008): under appropriate assumptions, consistency is possible as long as

$$\log p = O(n)$$

High-dimensional inference
Variable selection without computational limits

- Approaches based on penalized criteria (close to BIC)

$$\min_{w \in \mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + C\sigma^2 \|w\|_0 \big(1 + \log \frac{p}{\|w\|_0}\big)$$

- **Oracle inequality** if data generated by $\mathbf{w}$ with k non-zeros (Massart, 2003; Bunea et al., 2007):

$$\frac{1}{n}\|X\hat{w} - X\mathbf{w}\|_2^2 \leqslant C\frac{k\sigma^2}{n}\big(1 + \log \frac{p}{k}\big)$$

- Gaussian noise - **No assumptions regarding correlations**
- **Scaling between dimensions**: $\dfrac{k \log p}{n}$ small

High-dimensional inference (Lasso)

- **Main result**: we only need $k \log p = O(n)$
  - if $\mathbf{w}$ is sufficiently sparse
  - **and** input variables are not too correlated

High-dimensional inference (Lasso)

- **Main result**: we only need $k \log p = O(n)$
  - if $\mathbf{w}$ is sufficiently sparse
  - **and** input variables are not too correlated

- Precise conditions on covariance matrix $\mathbf{Q} = \frac{1}{n} X^\top X$.
  - Mutual incoherence (Lounici, 2008)
  - Restricted eigenvalue conditions (Bickel et al., 2009)
  - Sparse eigenvalues (Meinshausen and Yu, 2008)
  - Null space property (Donoho and Tanner, 2005)

- Links with signal processing and compressed sensing (Candès and Wakin, 2008)

- **Slow rate if no assumptions**: $\sqrt{\frac{k \log p}{n}}$

Alternative sparse methods
Greedy methods

- Forward selection (a.k.a. orthogonal matching pursuit)
- Forward-backward selection
- Non-convex method
  - Harder to analyze
  - Simpler to implement
  - Problems of stability
- Positive theoretical results (Zhang, 2009, 2008a)
  - Similar sufficient conditions than for the Lasso

Comparing Lasso and other strategies for linear regression

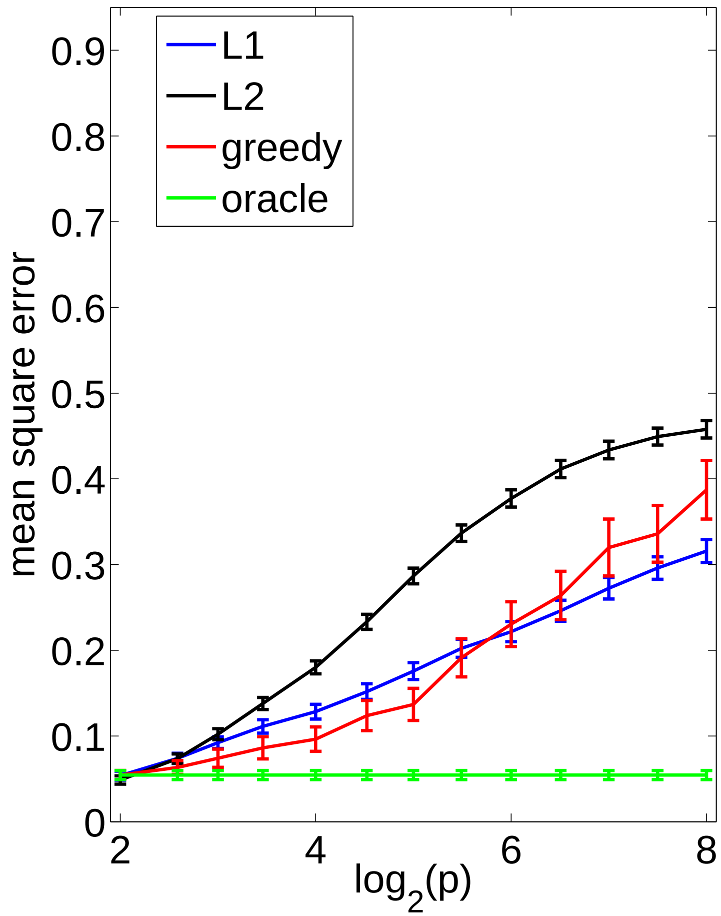
- Compared methods to reach the least-square solution
  - Ridge regression: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
  - Lasso: $\displaystyle \min_{w\in\mathbb{R}^p} \frac{1}{2}\|y - Xw\|_2^2 + \lambda\|w\|_1$
  - Forward greedy:
    * Initialization with empty set
    * Sequentially add the variable that best reduces the square loss

- Each method builds a path of solutions from 0 to ordinary least-squares solution

- Regularization parameters selected on the test set

Simulation results

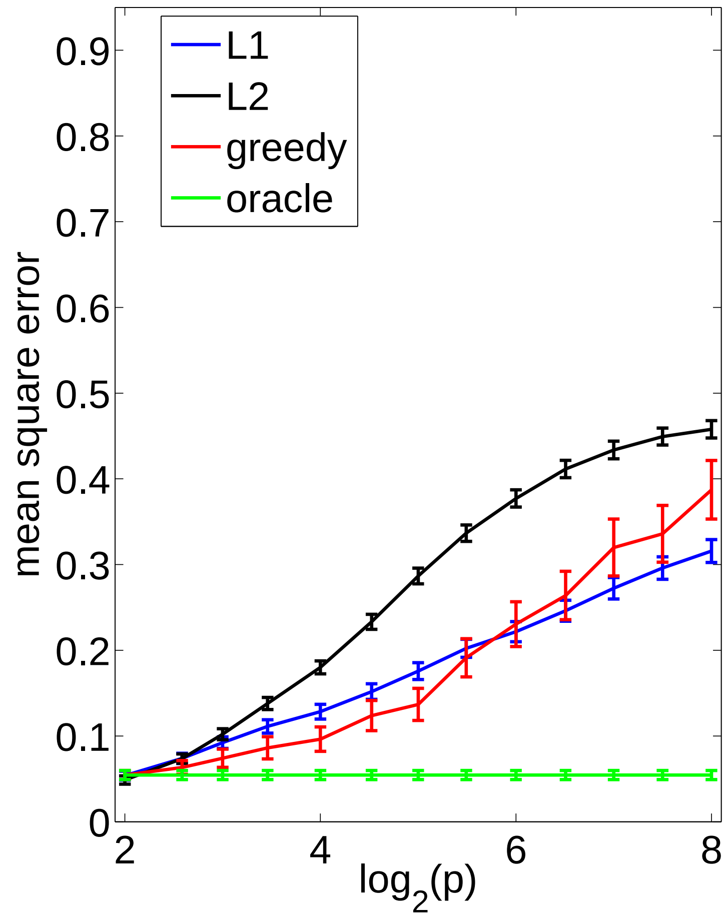
- i.i.d. Gaussian design matrix, $k = 4$, $n = 64$, $p \in [2, 256]$, SNR $= 1$
- Note stability to non-sparsity and variability



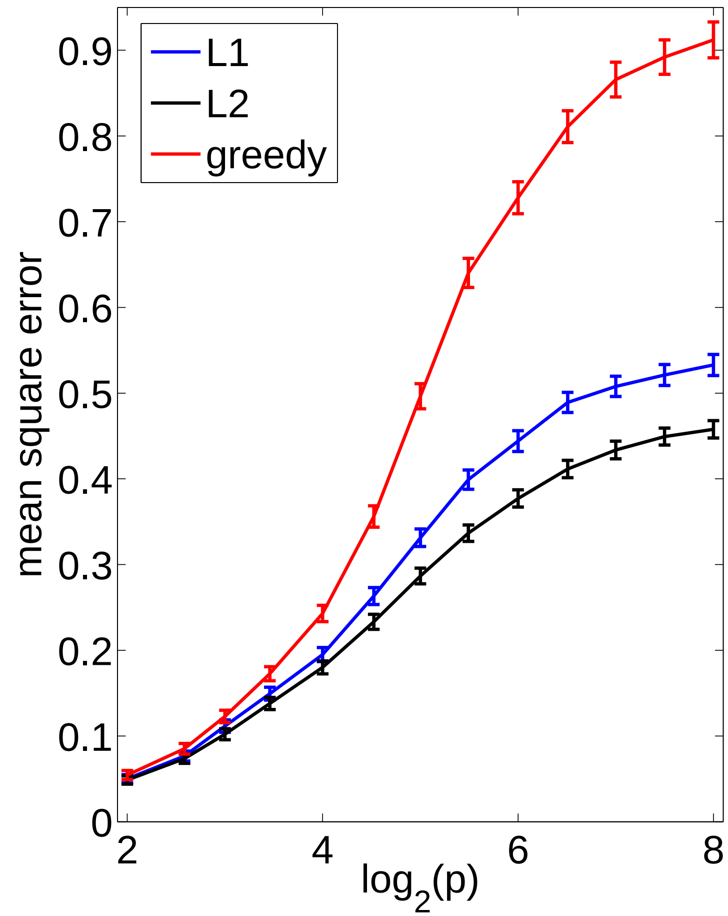
Sparse

Simulation results

- i.i.d. Gaussian design matrix, $k = 4$, $n = 64$, $p \in [2, 256]$, SNR $= 1$
- Note stability to non-sparsity and variability



Sparse



Rotated (non sparse)

Going beyond the Lasso

- ℓ_1-norm for **linear** feature selection in **high dimensions**
  - Lasso usually not applicable directly
- **Non-linearities**
  - Multiple kernel learning (Lanckriet et al., 2004; Bach et al., 2004)
- **Sparse learning on matrices**
  - Dictionary learning and matrix factorization
- **Dealing with structured set of features**
  - Specific sets of zeros

Outline
Sparse methods for machine learning and computer vision

- **Introduction**
- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering
- **Sparsity for computer vision**
  - Task-driven dictionary learning
- **Structured sparsity**

Learning on matrices - Image denoising

- Simultaneously denoise all patches of a given image
- Example from Mairal, Bach, Ponce, Sapiro, and Zisserman (2009d)

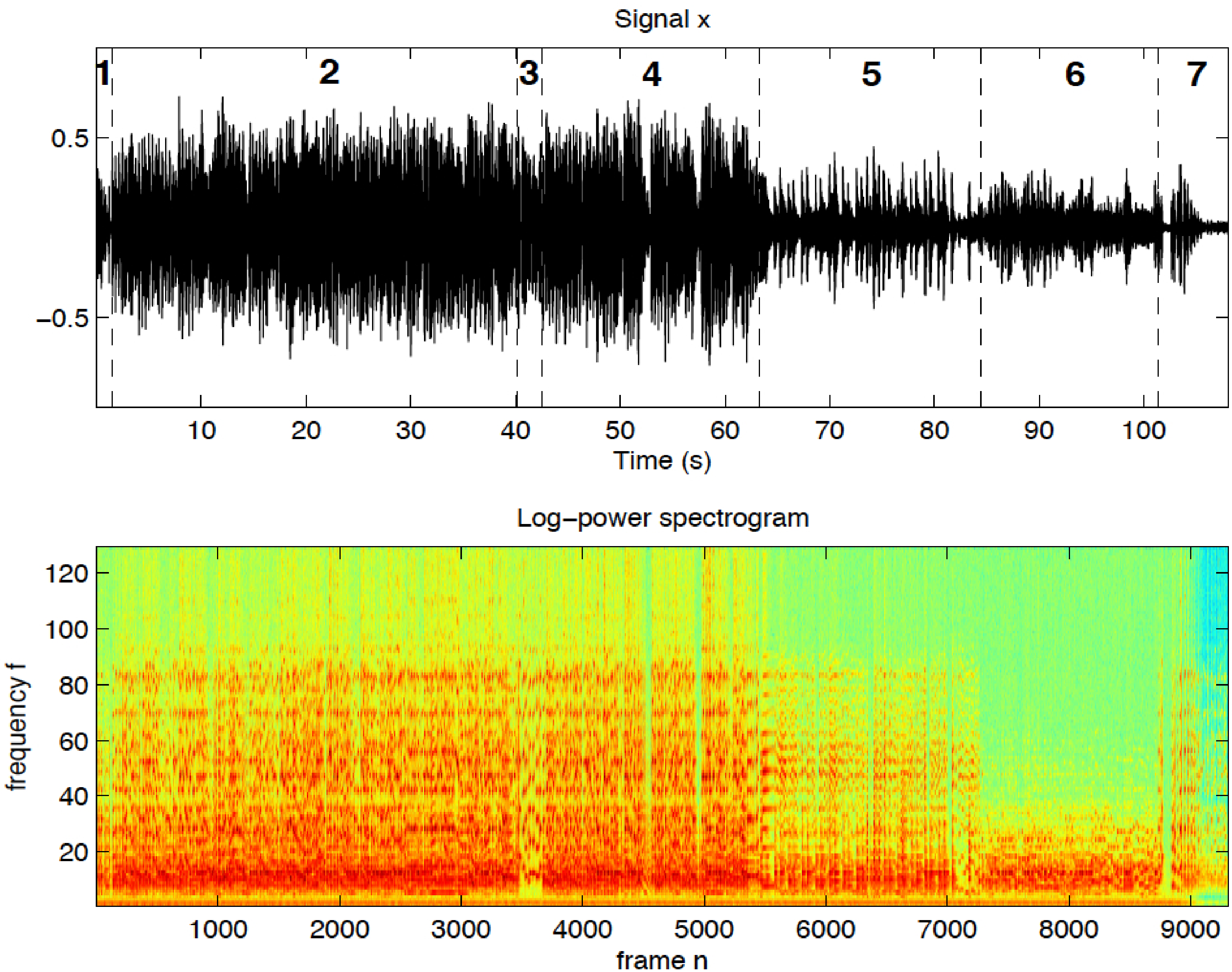
Learning on matrices - Collaborative filtering

- Given $n_{\mathcal{X}}$ "movies" $\mathbf{x} \in \mathcal{X}$ and $n_{\mathcal{Y}}$ "customers" $\mathbf{y} \in \mathcal{Y}$,
- predict the "rating" $z(\mathbf{x}, \mathbf{y}) \in \mathcal{Z}$ of customer $\mathbf{y}$ for movie $\mathbf{x}$
- Training data: large $n_{\mathcal{X}} \times n_{\mathcal{Y}}$ incomplete matrix $\mathbf{Z}$ that describes the known ratings of some customers for some movies
- **Goal**: complete the matrix.

	1				1								2							
											2				1				3	
	2									3						3				3
						1			2					1				1		
		1		3							1						1	3		
	1		2													2				
								3												

Learning on matrices - Source separation

- Single microphone (Benaroya et al., 2006; Févotte et al., 2009)



Learning on matrices - Multi-task learning

- k linear prediction tasks on same covariates $\mathbf{x} \in \mathbb{R}^p$
  - k weight vectors $\mathbf{w}_j \in \mathbb{R}^p$
  - Joint matrix of predictors $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$

- Classical application
  - Multi-category classification (one task per class) (Amit et al., 2007)

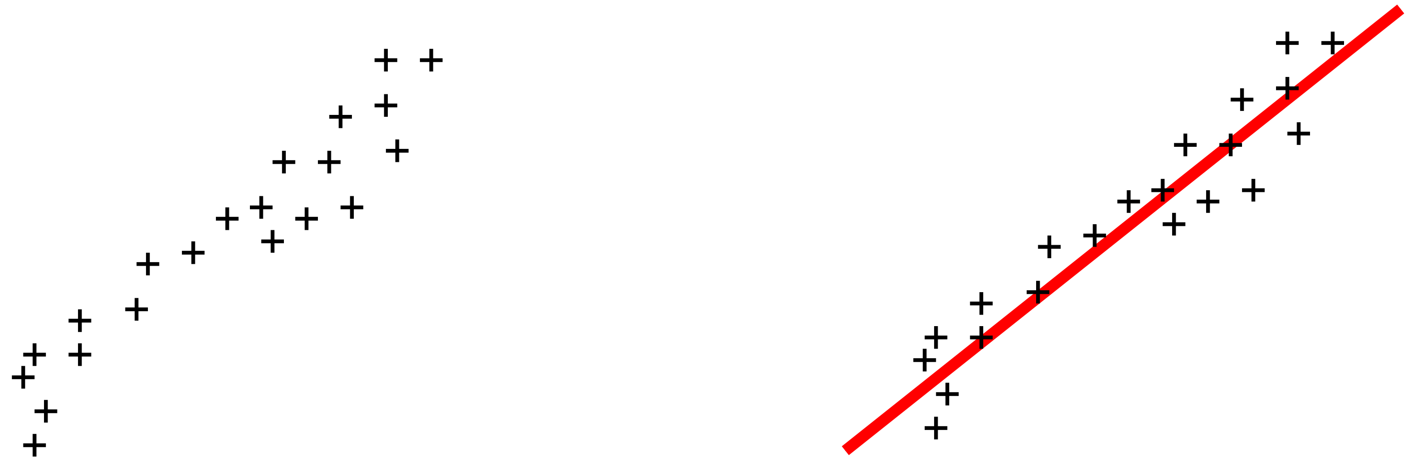
- **Share parameters between tasks**

- **Joint variable selection** (Obozinski et al., 2009)
  - Select variables which are predictive for all tasks

- **Joint feature selection** (Pontil et al., 2007)
  - Construct linear features common to all tasks

Matrix factorization - Dimension reduction

- Given data matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$
  - **Principal component analysis**: $\boxed{\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$

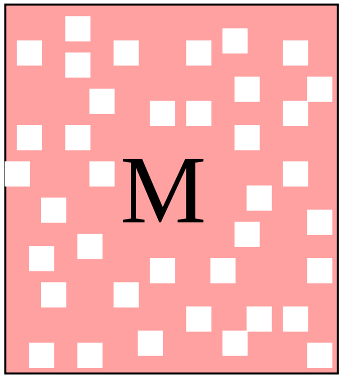


  - **K-means**: $\boxed{\mathbf{x}_i \approx \mathbf{d}_k \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}}$

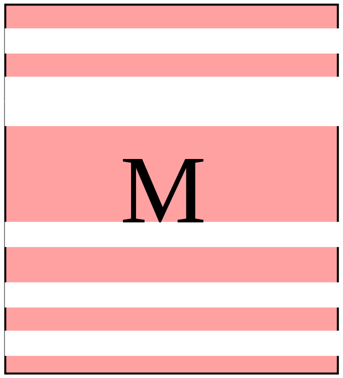
Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$

I - Directly on the elements of $\mathbf{M}$

- Many zero elements: $\mathbf{M}_{ij} = 0$

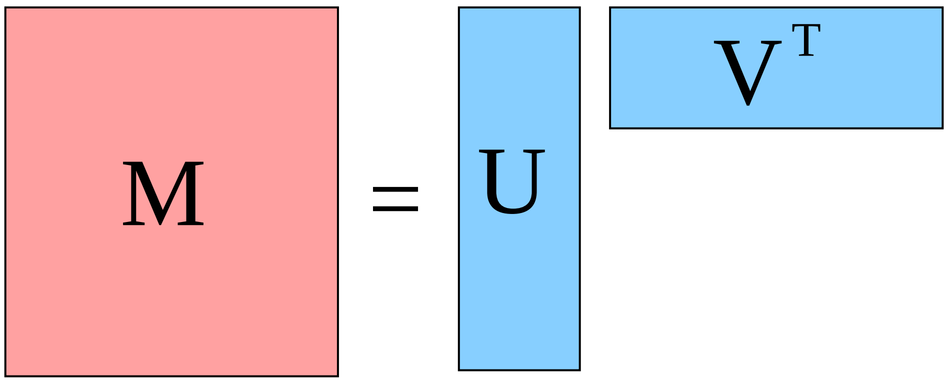


- Many zero rows (or columns): $(\mathbf{M}_{i1}, \ldots, \mathbf{M}_{ip}) = 0$

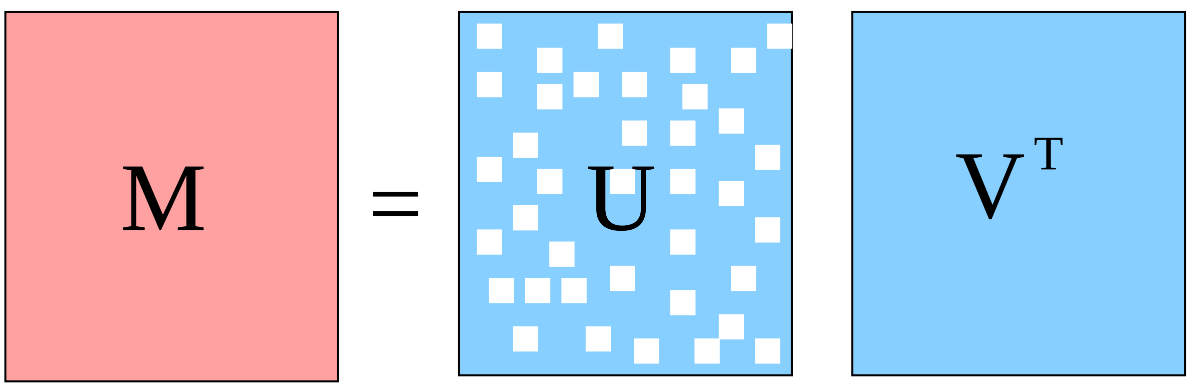


Two types of sparsity for matrices $\mathbf{M} \in \mathbb{R}^{n \times p}$
II - Through a factorization of $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$
- **Low rank**: m small



- **Sparse decomposition**: $\mathbf{U}$ sparse



Structured sparse matrix factorizations

- Matrix $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times k}$ and $\mathbf{V} \in \mathbb{R}^{p \times k}$

- **Structure on $\mathbf{U}$ and/or $\mathbf{V}$**
  - Low-rank: $\mathbf{U}$ and $\mathbf{V}$ have few columns
  - Dictionary learning / sparse PCA: $\mathbf{U}$ has many zeros
  - Clustering (k-means): $\mathbf{U} \in \{0,1\}^{n \times m}$, $\mathbf{U}\mathbf{1} = \mathbf{1}$
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros (Jenatton et al., 2010)
  - Low-rank + sparse (Candès et al., 2009)
  - etc.

- **Many applications**

- **Many open questions** (Algorithms, identifiability, etc.)

Multi-task learning

- Joint matrix of predictors $W = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k}$
- **Joint variable selection** (Obozinski et al., 2009)
  - Penalize by the sum of the norms of rows of W (group Lasso)
  - Select variables which are predictive for all tasks

Multi-task learning

- Joint matrix of predictors $W = (w_1, \dots, w_k) \in \mathbb{R}^{p \times k}$

- **Joint variable selection** (Obozinski et al., 2009)
  - Penalize by the sum of the norms of rows of W (group Lasso)
  - Select variables which are predictive for all tasks

- **Joint feature selection** (Pontil et al., 2007)
  - Penalize by the trace-norm (see later)
  - Construct linear features common to all tasks

- Theory: allows number of observations which is sublinear in the number of tasks (Obozinski et al., 2008; Lounici et al., 2009)

- Practice: more interpretable models, slightly improved performance

Low-rank matrix factorizations
Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values

- Rank of $\mathbf{M}$ equal to the number of non-zero singular values

Low-rank matrix factorizations
Trace norm

- Given a matrix $\mathbf{M} \in \mathbb{R}^{n \times p}$
  - Rank of $\mathbf{M}$ is the minimum size m of **all** factorizations of $\mathbf{M}$ into $\mathbf{M} = \mathbf{U}\mathbf{V}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times m}$ and $\mathbf{V} \in \mathbb{R}^{p \times m}$
  - Singular value decomposition: $\mathbf{M} = \mathbf{U}\,\mathrm{Diag}(\mathbf{s})\mathbf{V}^\top$ where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{s} \in \mathbb{R}_+^m$ are singular values

- Rank of $\mathbf{M}$ equal to the number of non-zero singular values

- **Trace-norm (a.k.a. nuclear norm)** $=$ sum of singular values

- Convex function, leads to a semi-definite program (Fazel et al., 2001)

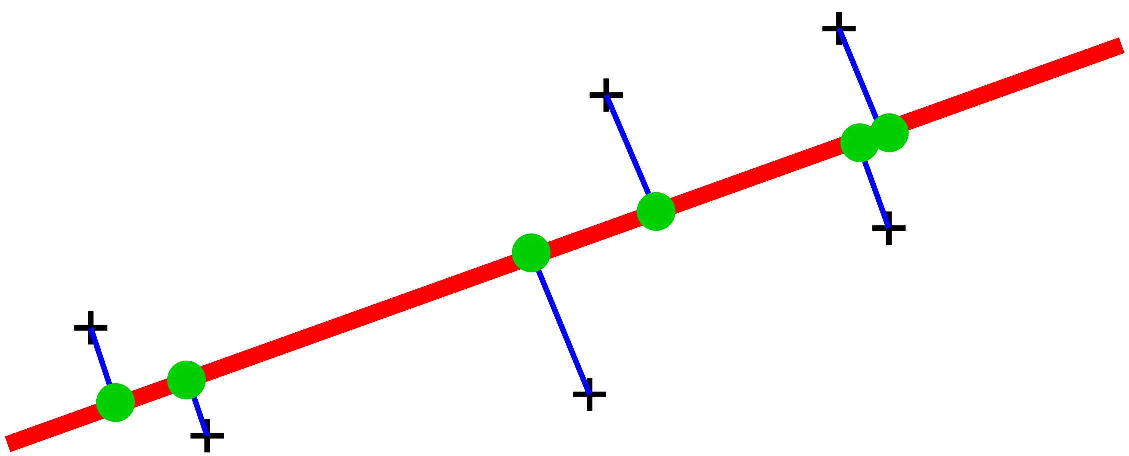
- First used for collaborative filtering (Srebro et al., 2005)

- Multi-category classif. (Amit et al., 2007; Harchaoui et al., 2012)

Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \ldots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent



Sparse principal component analysis

- Given data $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{p \times n}$, two views of PCA:
  - **Analysis view**: find the projection $\mathbf{d} \in \mathbb{R}^p$ of maximum variance (with deflation to obtain more components)
  - **Synthesis view**: find the basis $\mathbf{d}_1, \dots, \mathbf{d}_k$ such that all $\mathbf{x}_i$ have low reconstruction error when decomposed on this basis

- For regular PCA, the two views are equivalent

- **Sparse extensions**
  - Interpretability
  - High-dimensional inference
  - Two views are differents
    - For analysis view, see d'Aspremont, Bach, and El Ghaoui (2008)

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \ldots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

  – Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

Sparse principal component analysis
Synthesis view

- Find $\mathbf{d}_1, \dots, \mathbf{d}_k \in \mathbb{R}^p$ **sparse** so that

$$\sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^{k} (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^{n} \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

  – Look for $\mathbf{A} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n) \in \mathbb{R}^{k \times n}$ and $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k) \in \mathbb{R}^{p \times k}$ such that $\mathbf{D}$ is sparse and $\|\mathbf{X} - \mathbf{D}\mathbf{A}\|_F^2$ is small

- Sparse formulation (Witten et al., 2009; Bach et al., 2008)

  – Penalize/constrain $\mathbf{d}_j$ by the ℓ_1-norm for sparsity
  – Penalize/constrain $\boldsymbol{\alpha}_i$ by the ℓ_2-norm to avoid trivial solutions

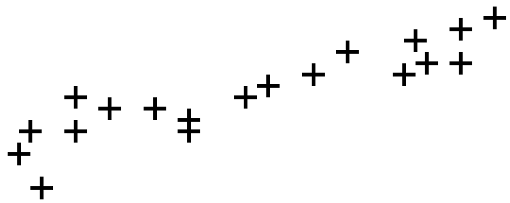
$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_1 \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_2 \leqslant 1$$

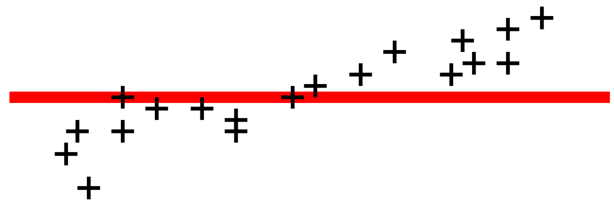
Sparse PCA vs. dictionary learning

- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse

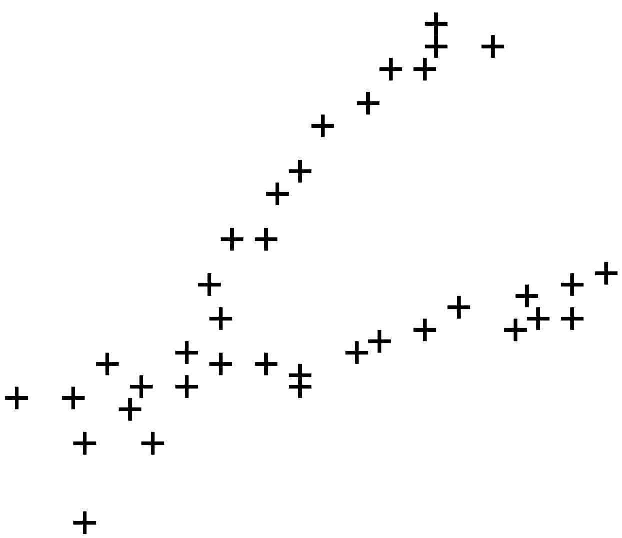
Sparse PCA vs. dictionary learning

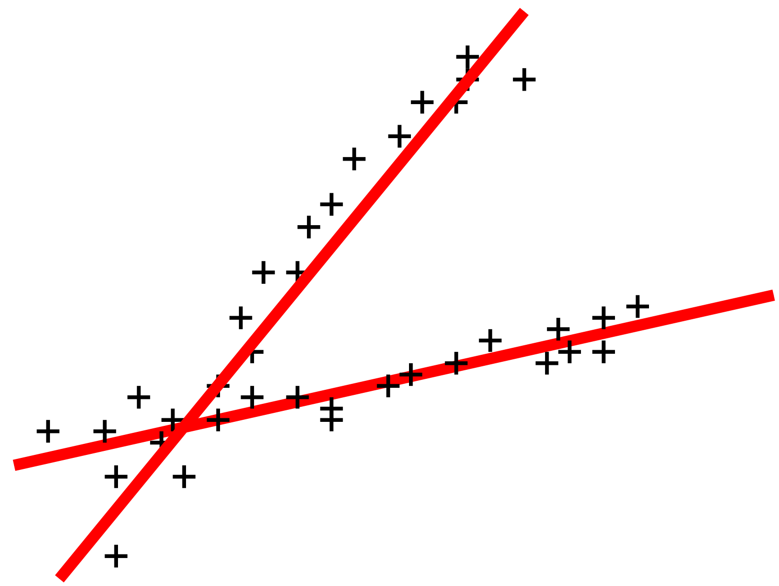
- **Sparse PCA**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\mathbf{D}$ sparse





- **Dictionary learning**: $\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i$, $\boldsymbol{\alpha}_i$ sparse





Structured matrix factorizations (Bach et al., 2008)

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^{k} \|\mathbf{d}_j\|_\star \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_\bullet \leqslant 1$$

$$\min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{i=1}^{n} \|\boldsymbol{\alpha}_i\|_\bullet \text{ s.t. } \forall j, \|\mathbf{d}_j\|_\star \leqslant 1$$

- Optimization by alternating minimization (non-convex)
- $\boldsymbol{\alpha}_i$ decomposition coefficients (or "code"), $\mathbf{d}_j$ dictionary elements
- Two related/equivalent problems:
  - **Sparse PCA = sparse dictionary** (ℓ_1-norm on $\mathbf{d}_j$)
  - **Dictionary learning = sparse decompositions** (ℓ_1-norm on $\boldsymbol{\alpha}_i$) (Olshausen and Field, 1997; Elad and Aharon, 2006; Lee et al., 2007)

Dictionary learning for image denoising

$$\underbrace{\mathbf{x}}_{\text{measurements}} = \underbrace{\mathbf{y}}_{\text{original image}} + \underbrace{\varepsilon}_{\text{noise}}$$

Dictionary learning for image denoising

- **Solving the denoising problem** (Elad and Aharon, 2006)
  - Extract all overlapping 8×8 patches $\mathbf{x}_i \in \mathbb{R}^{64}$
  - Form the matrix $\mathbf{X} = (\mathbf{x}_1^\top, \dots, \mathbf{x}_n^\top) \in \mathbb{R}^{n \times 64}$
  - Solve a matrix factorization problem:

$$\min_{\mathbf{D},\mathbf{A}} ||\mathbf{X} - \mathbf{D}\mathbf{A}||_F^2 = \min_{\mathbf{D},\mathbf{A}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2$$

    where $\mathbf{A}$ is **sparse**, and $\mathbf{D}$ is the **dictionary**
  - Each patch is decomposed into $\mathbf{x}_i = \mathbf{D}\boldsymbol{\alpha}_i$
  - Average the reconstruction $\mathbf{D}\boldsymbol{\alpha}_i$ of each patch $\mathbf{x}_i$ to reconstruct a full-sized image

- The number of patches n is large ($=$ number of pixels)

Online optimization for dictionary learning

$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \text{ s.t. } \forall j = 1,\ldots,k, \;\; ||\mathbf{d}_j||_2 \leqslant 1\}.$$

- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$
- Good results, but very slow !

Online optimization for dictionary learning

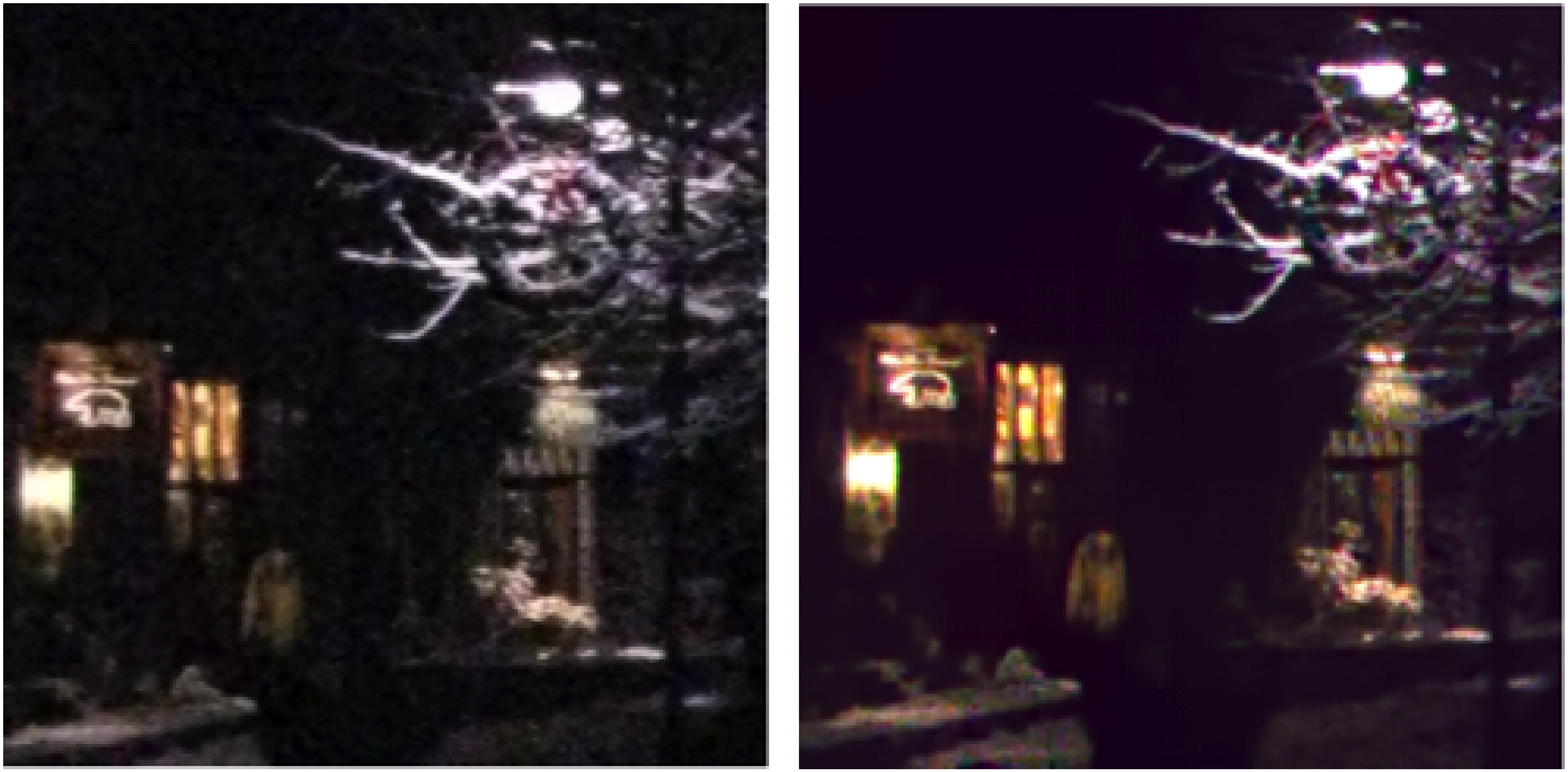
$$\min_{\mathbf{A}\in\mathbb{R}^{k\times n},\mathbf{D}\in\mathcal{D}} \sum_{i=1}^{n} ||\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i||_2^2 + \lambda||\boldsymbol{\alpha}_i||_1$$

$$\mathcal{D} \triangleq \{\mathbf{D} \in \mathbb{R}^{p\times k} \ \text{ s.t. } \ \forall j = 1,\ldots,k, \ \ ||\mathbf{d}_j||_2 \leqslant 1\}.$$

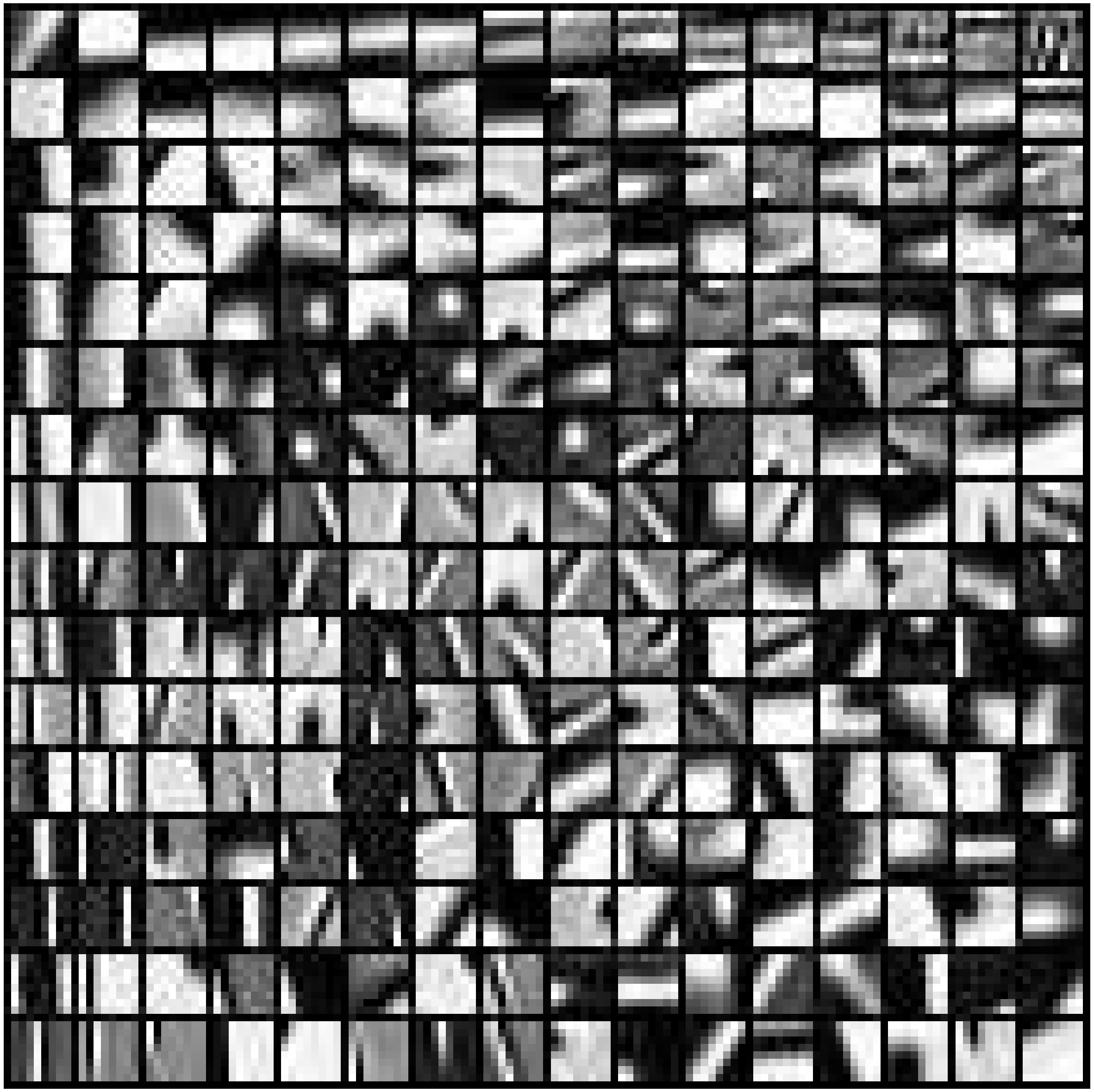
- Classical optimization alternates between $\mathbf{D}$ and $\mathbf{A}$.
- Good results, but very slow !
- **Online learning** (Mairal, Bach, Ponce, and Sapiro, 2009a) can
  - handle potentially infinite datasets
  - adapt to dynamic training sets
- **Simultaneous sparse coding** (Mairal et al., 2009d)
  - Links with NL-means (Buades et al., 2008)

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009d)

Denoising result
(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009d)



What does the dictionary $\mathbb{D}$ look like?



Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Inpainting a 12-Mpixel photograph

Additional methods - Softwares

- Many contributions in signal processing, optimization, mach. learning
  - Extensions to stochastic setting (Bottou and Bousquet, 2008)

- **Extensions to other sparsity-inducing norms**
  - Computing proximal operator
  - F. Bach, R. Jenatton, J. Mairal, G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1-106, 2011.

- **Softwares**
  - Many available codes
  - SPAMS (SPArse Modeling Software)
    `http://www.di.ens.fr/willow/SPAMS/`

Outline

Sparse methods for machine learning and computer vision

- **Introduction**
- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering
- **Sparsity for computer vision**
  - Task-driven dictionary learning
- **Structured sparsity**

Learning dictionaries with a discriminative cost function

- **Idea**: consider 2 sets S_-, S_+ of signals representing 2 different classes. Each set should admit a specific dictionary best adapted to its reconstruction.

- **Classification procedure** for a signal $x \in \mathbb{R}^n$:

$$\min(\mathbf{R}^\star(x, D_-), \mathbf{R}^\star(x, D_+))$$

$$\text{where } \mathbf{R}^\star(x, D) = \min_{\alpha \in \mathbb{R}^p} ||x - D\alpha||_2^2 \text{ s.t. } ||\alpha||_0 \leq L.$$

- **"Reconstructive" training**

$$\begin{cases} \min_{D_-} \sum_{i \in S_-} \mathbf{R}^\star(x_i, D_-) \\ \min_{D_+} \sum_{i \in S_+} \mathbf{R}^\star(x_i, D_+) \end{cases}$$

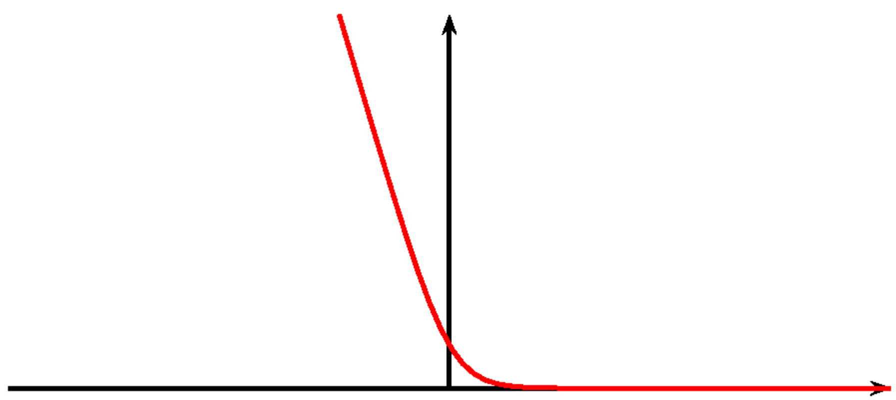
(Grosse et al., 2007; Huang and Aviyente, 2006; Sprechmann et al., 2010)

Learning dictionaries with a discriminative cost function

- **"Discriminative" training** (Mairal, Bach, Ponce, Sapiro, and Zisserman, 2008b)

$$\min_{D_-,D_+} \sum_i \mathcal{D}\Big(\lambda z_i \big(\mathbf{R}^\star(x_i, D_-) - \mathbf{R}^\star(x_i, D_+)\big)\Big),$$

where $z_i \in \{-1, +1\}$ is the label of $\mathbf{x}_i$.



Logistic regression function

Learning dictionaries with a discriminative cost function

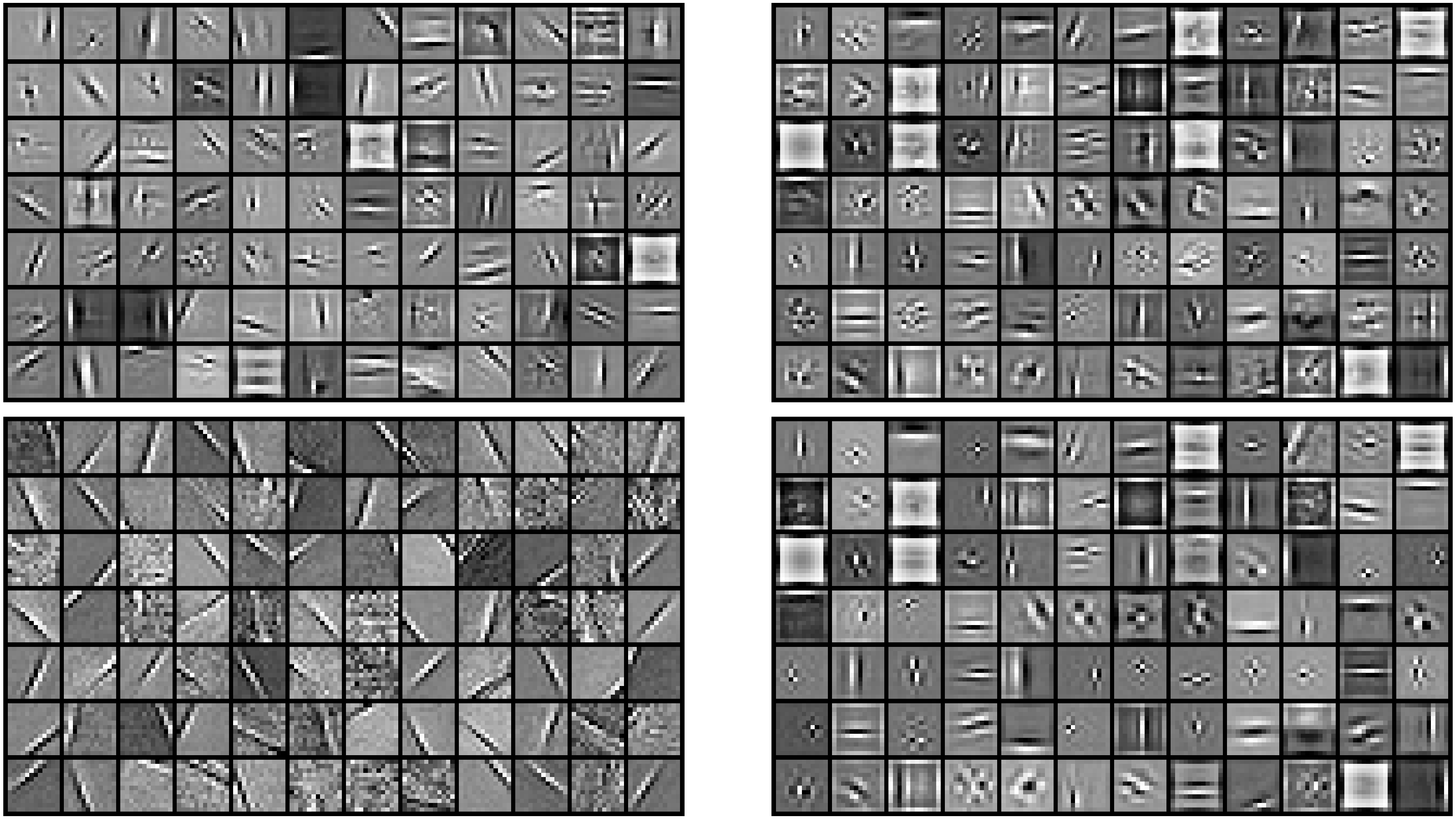
- **Mixed approach**

$$\min_{D_-,D_+} \sum_i \mathcal{D}\Big(\lambda z_i \big(\mathbf{R}^\star(x_i, D_-) - \mathbf{R}^\star(x_i, D_+)\big)\Big) + \mu \mathbf{R}^\star(x_i, D_{z_i}),$$

  where $z_i \in \{-1, +1\}$ is the label of $\mathbf{x}_i$.

- **Keys of the optimization framework**
  - Alternation of sparse coding and dictionary updates.
  - Continuation path with decreasing values of μ.
  - OMP to address the NP-hard sparse coding problem. . .
  - . . . or homotopy method when using ℓ_1.
  - Use softmax instead of logistic regression for $N > 2$ classes.

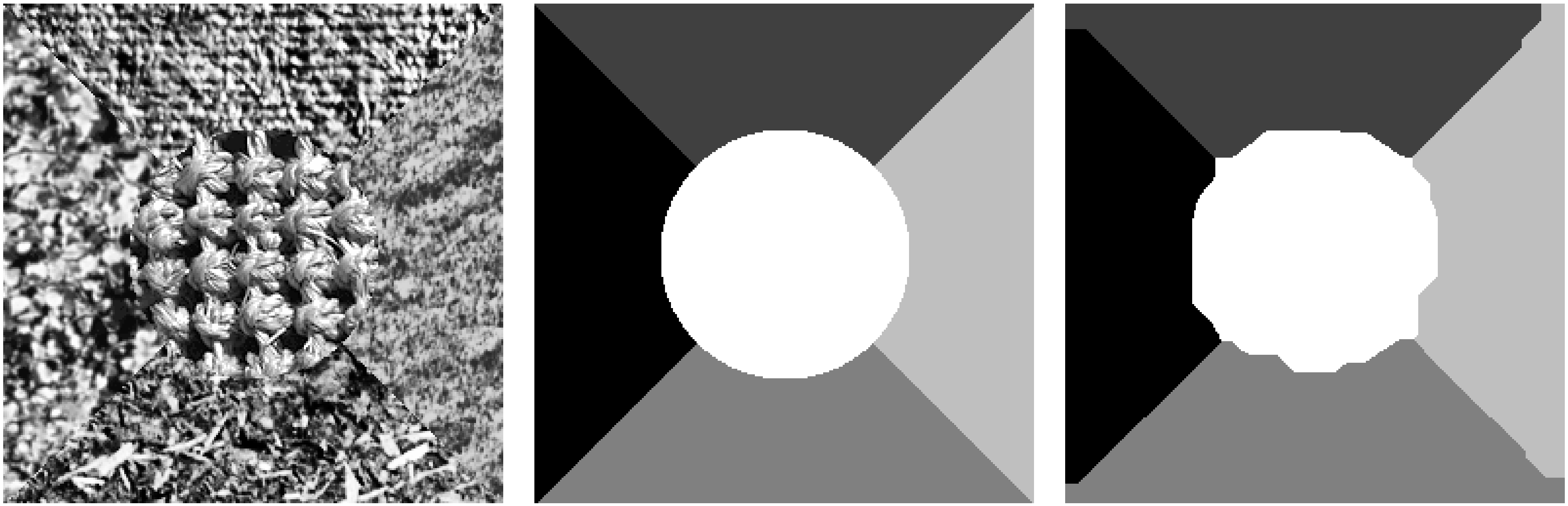
Learning dictionaries with a discriminative cost function - Examples of dictionaries



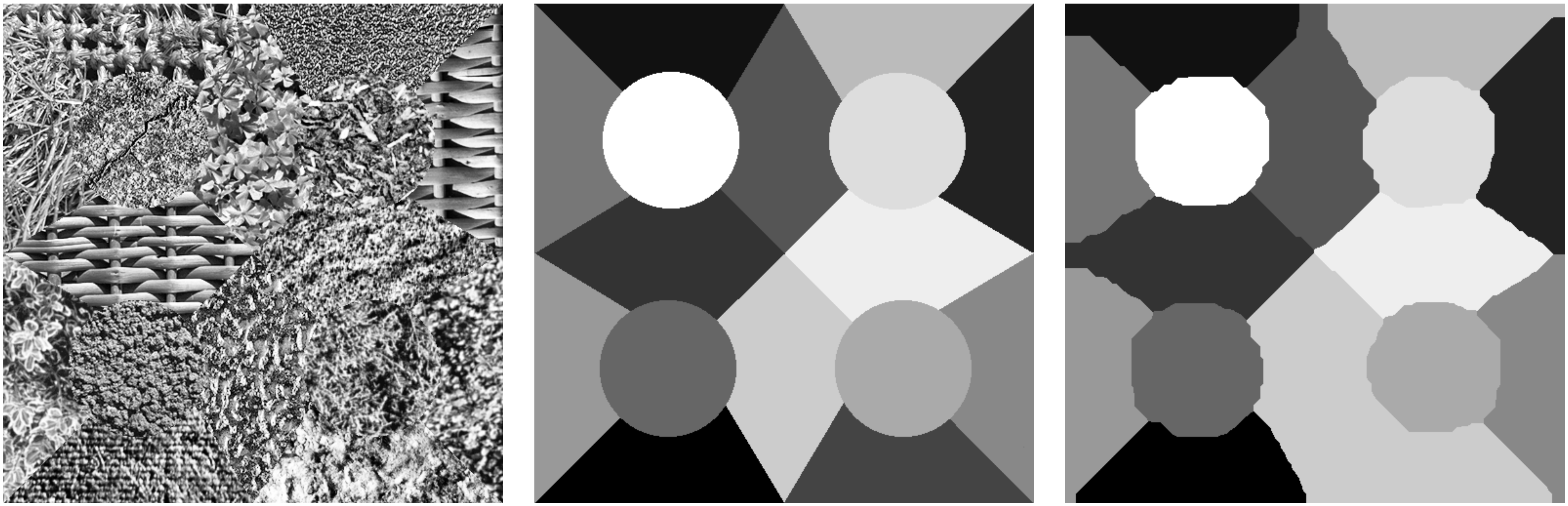
Top: reconstructive, Bottom: discriminative

Left: Bicycle, Right: Background

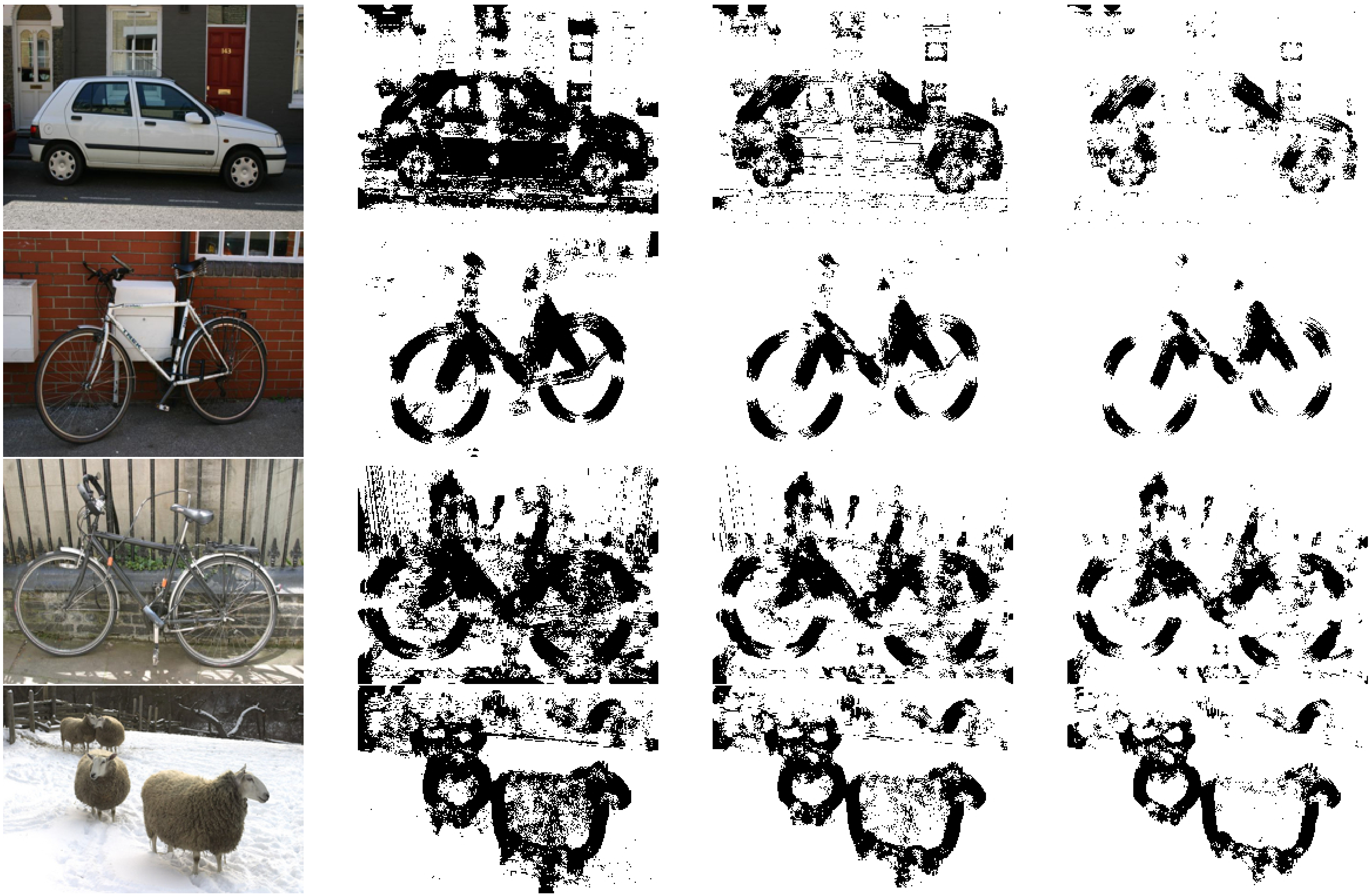
Learning dictionaries with a discriminative cost function - Texture segmentation



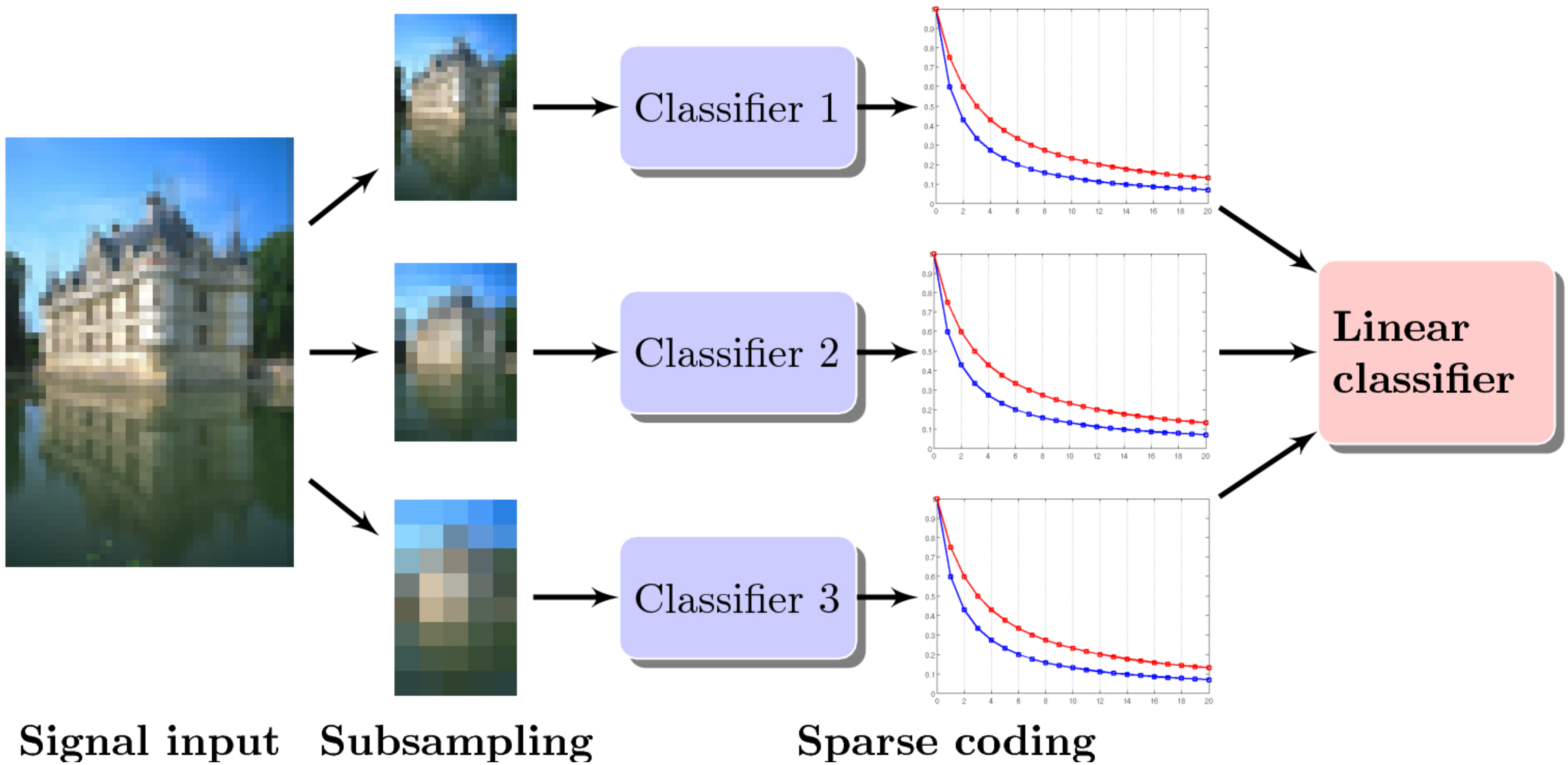
Learning dictionaries with a discriminative cost function - Texture segmentation



Learning dictionaries with a discriminative cost function - Pixelwise classification

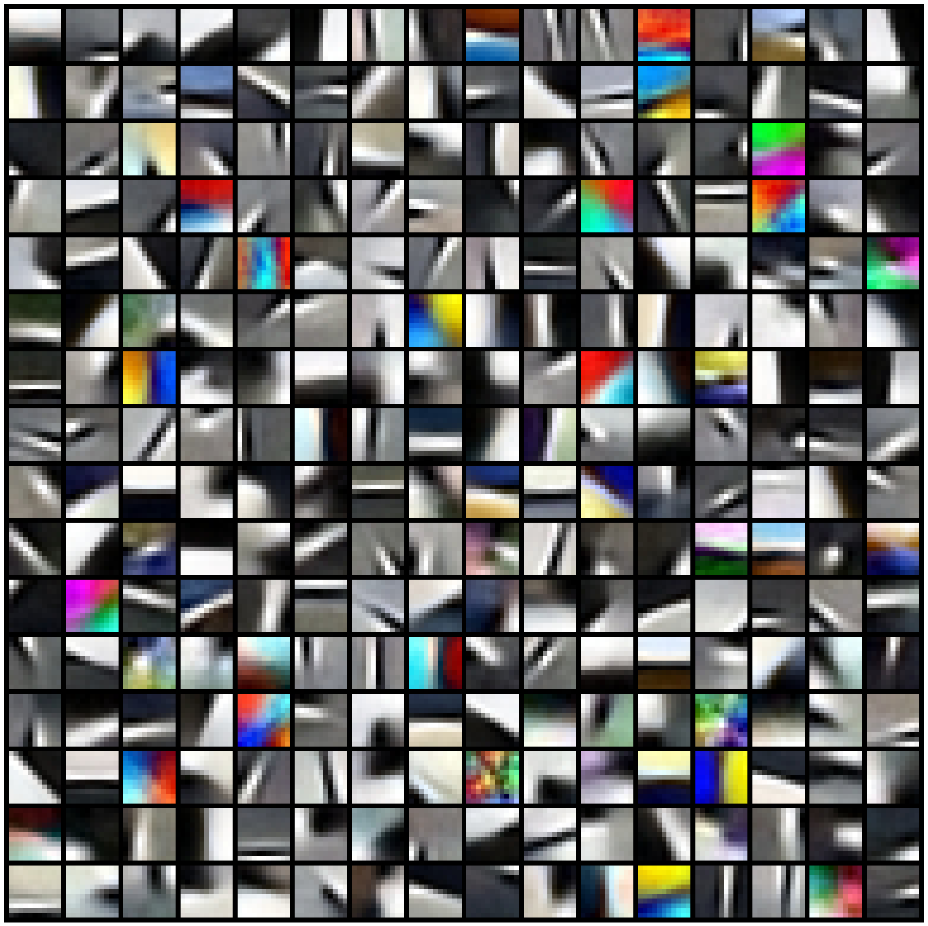


Learning dictionaries with a discriminative cost function - Multiscale scheme

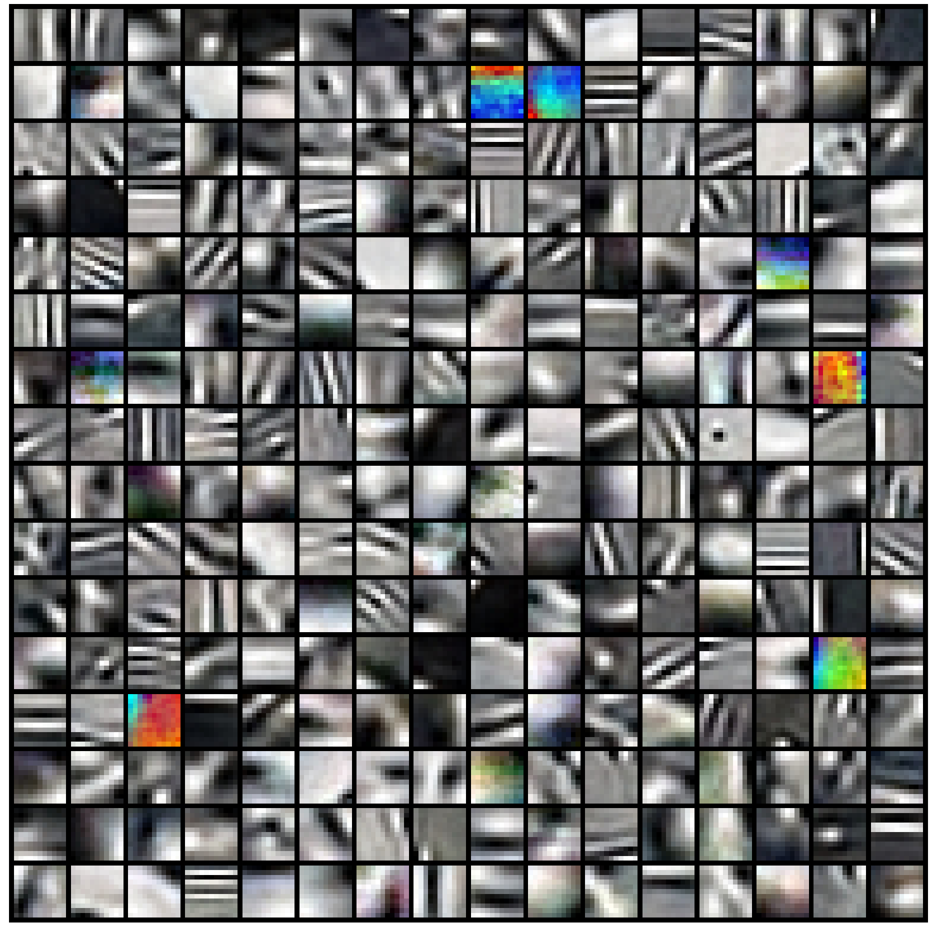


Learning dictionaries with a discriminative cost function - weakly-supervised pixel classification

Application to edge detection and classification (Mairal, Leordeanu, Bach, Hebert, and Ponce, 2008c)



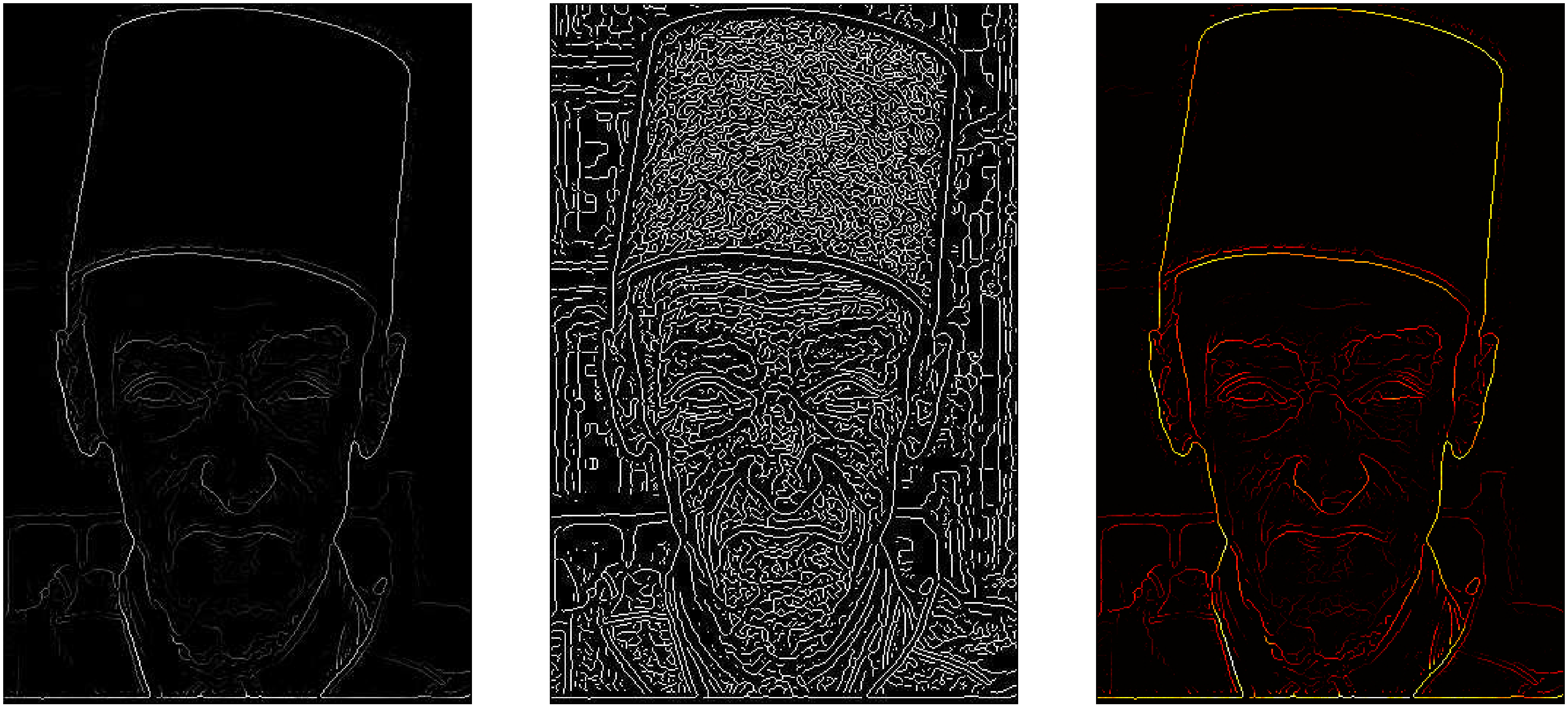
Good edges



Bad edges

Application to edge detection and classification

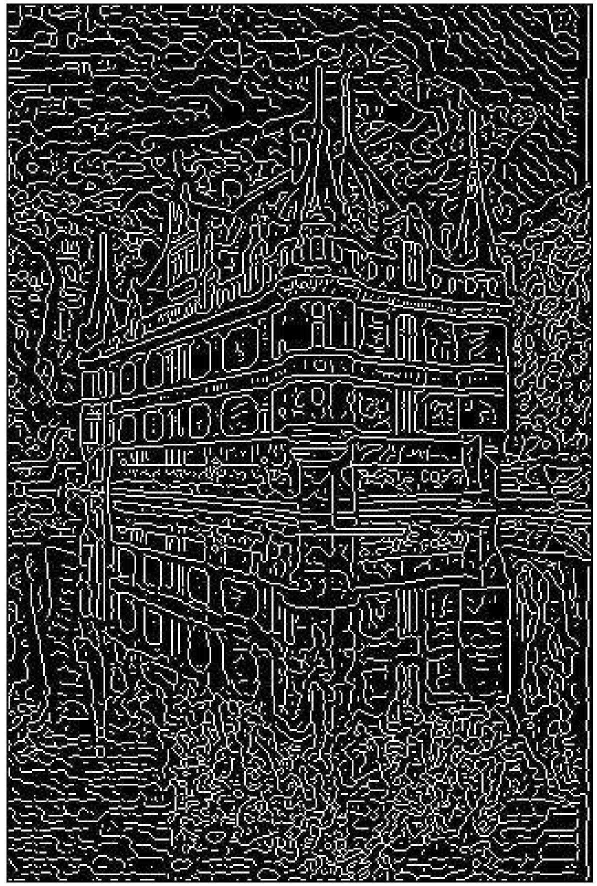
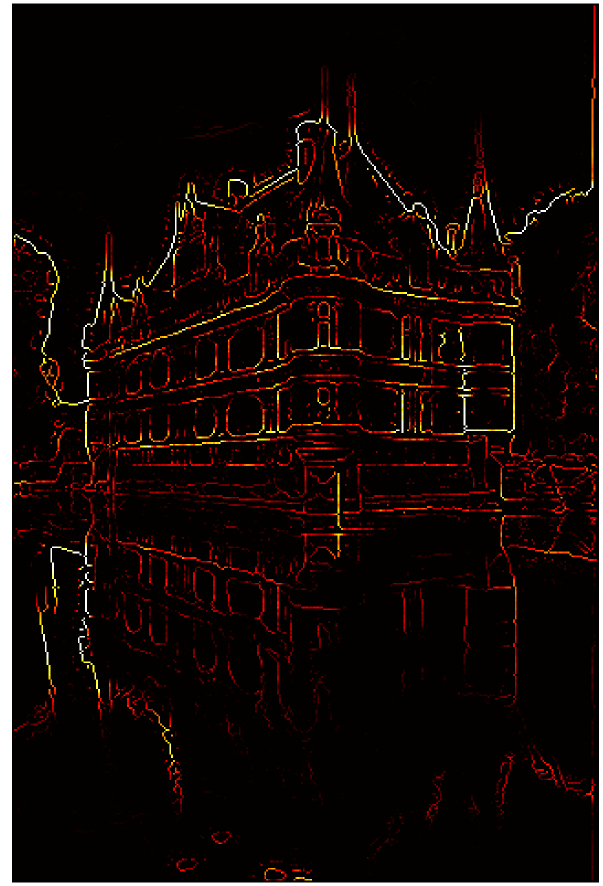
Berkeley segmentation benchmark



Raw edge detection on the right

Application to edge detection and classification

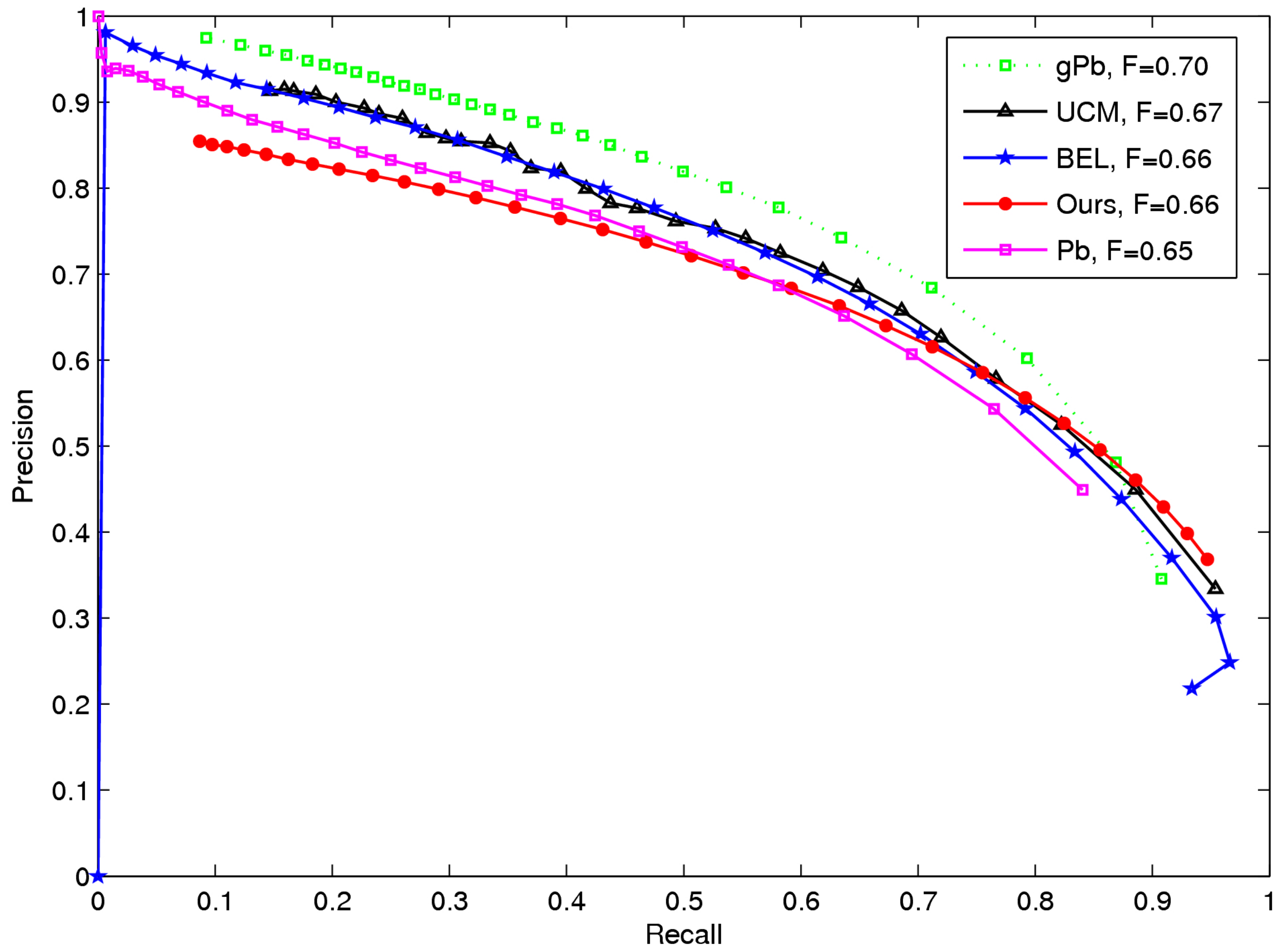
Berkeley segmentation benchmark

Raw edge detection on the right

Application to edge detection and classification

Berkeley segmentation benchmark

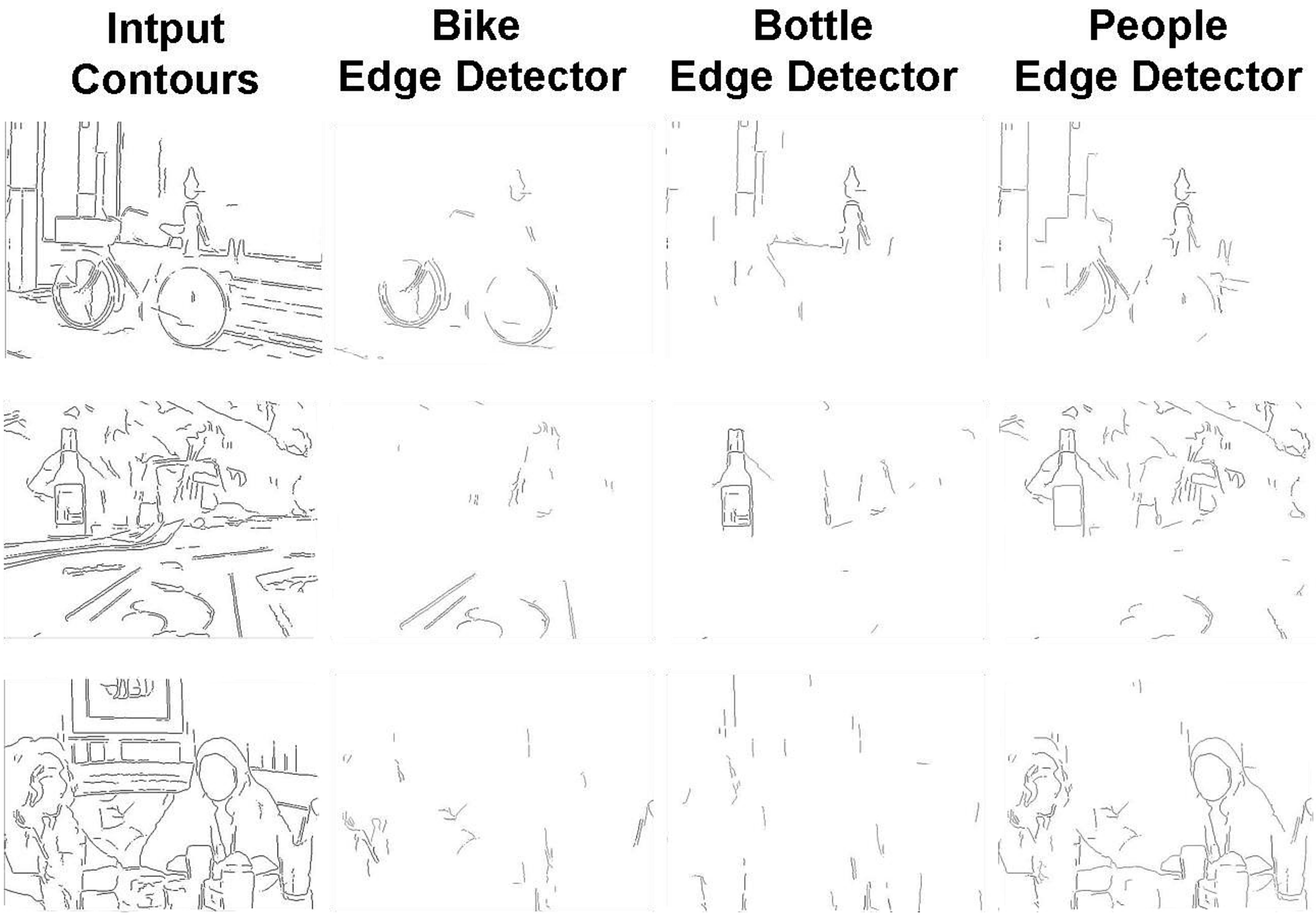


Application to edge detection and classification

Contour-based classifier (Leordeanu, Hebert, and Sukthankar, 2007)

Is there a bike, a motorbike, a car or a person on this image?

Application to edge detection and classification



Application to edge detection and classification
Performance gain due to the prefiltering

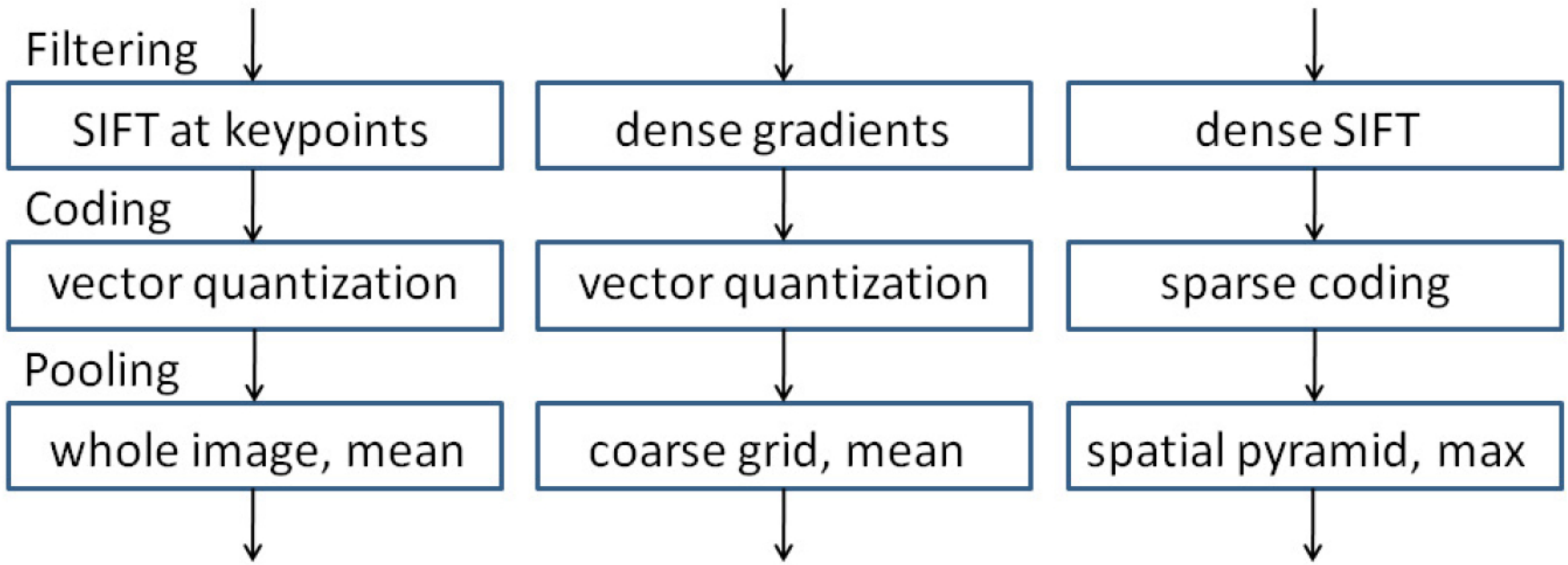
Ours + [Leordeanu '07]	[Leordeanu '07]	[Winn '05]
96.8%	89.4%	76.9%

Recognition rates for the same experiment as (Winn et al., 2005) on VOC 2005.

Category	Ours+[Leordeanu '07]	[Leordeanu '07]
Aeroplane	71.9%	61.9%
Boat	67.1%	56.4%
Cat	82.6%	53.4%
Cow	68.7%	59.2%
Horse	76.0%	67%
Motorbike	80.6%	73.6%
Sheep	72.9%	58.4%
Tvmonitor	87.7%	83.8%
Average	75.9%	64.2 %

Recognition performance at equal error rate for 8 classes on a subset of images from Pascal 07.

Learning Codebooks for Image Classification



- **Idea**: Replacing Vector Quantization by Learned Dictionaries!
  - unsupervised: (Yang et al., 2009a)
  - supervised: (Boureau et al., 2010; Yang et al., 2010) (CVPR '10)

Learning Codebooks for Image Classification

- Let an image be represented by a set of low-level descriptors $\mathbf{x}_i$ at N locations identified with their indices $i = 1, \ldots, N$
  - hard-quantization:

$$x_i \approx D\alpha_i, \quad \alpha_i \in \{0,1\}^p \text{ and } \sum_{j=1}^{p} (\alpha_i)_j = 1$$

  - soft-quantization:

$$(\alpha_i)_j = \frac{e^{-\beta \|x_i - d_j\|_2^2}}{\sum_{k=1}^{p} e^{-\beta \|x_i - d_k\|_2^2}}$$

  - sparse coding:

$$x_i \approx D\alpha_i, \quad \alpha_i = \operatorname*{argmin}_{\alpha} \frac{1}{2}\|x_i - D\alpha\|_2^2 + \lambda\|\alpha\|_1$$

Learning Codebooks for Image Classification
Table from Boureau, Bach, Lecun, and Ponce (2010)

Method	Caltech-101, 30 training examples Average Pool	Caltech-101, 30 training examples Max Pool	15 Scenes, 100 training examples Average Pool	15 Scenes, 100 training examples Max Pool
	Results with basic features, SIFT extracted each 8 pixels			
Hard quantization, linear kernel	51.4 ± 0.9 [256]	64.3 ± 0.9 [256]	73.9 ± 0.9 [1024]	80.1 ± 0.6 [1024]
Hard quantization, intersection kernel	64.2 ± 1.0 [256] **(1)**	64.3 ± 0.9 [256]	80.8 ± 0.4 [256] **(1)**	80.1 ± 0.6 [1024]
Soft quantization, linear kernel	57.9 ± 1.5 [1024]	69.0 ± 0.8 [256]	75.6 ± 0.5 [1024]	81.4 ± 0.6 [1024]
Soft quantization, intersection kernel	66.1 ± 1.2 [512] **(2)**	70.6 ± 1.0 [1024]	81.2 ± 0.4 [1024] **(2)**	83.0 ± 0.7 [1024]
Sparse codes, linear kernel	61.3 ± 1.3 [1024]	$\mathbf{71.5 \pm 1.1}$ [1024] **(3)**	76.9 ± 0.6 [1024]	83.1 ± 0.6 [1024] **(3)**
Sparse codes, intersection kernel	70.3 ± 1.3 [1024]	$\mathbf{71.8 \pm 1.0}$ [1024] **(4)**	83.2 ± 0.4 [1024]	$\mathbf{84.1 \pm 0.5}$ [1024] **(4)**
	Results with macrofeatures and denser SIFT sampling			
Hard quantization, linear kernel	55.6 ± 1.6 [256]	70.9 ± 1.0 [1024]	74.0 ± 0.5 [1024]	80.1 ± 0.5 [1024]
Hard quantization, intersection kernel	68.8 ± 1.4 [512]	70.9 ± 1.0 [1024]	81.0 ± 0.5 [1024]	80.1 ± 0.5 [1024]
Soft quantization, linear kernel	61.6 ± 1.6 [1024]	71.5 ± 1.0 [1024]	76.4 ± 0.7 [1024]	81.5 ± 0.4 [1024]
Soft quantization, intersection kernel	70.1 ± 1.3 [1024]	73.2 ± 1.0 [1024]	81.8 ± 0.4 [1024]	83.0 ± 0.4 [1024]
Sparse codes, linear kernel	65.7 ± 1.4 [1024]	$\mathbf{75.1 \pm 0.9}$ [1024]	78.2 ± 0.7 [1024]	83.6 ± 0.4 [1024]
Sparse codes, intersection kernel	73.7 ± 1.3 [1024]	$\mathbf{75.7 \pm 1.1}$ [1024]	83.5 ± 0.4 [1024]	$\mathbf{84.3 \pm 0.5}$ [1024]

	Unsup	Discr	Unsup	Discr
Linear	83.6 ± 0.4	$\mathbf{84.9 \pm 0.3}$	84.2 ± 0.3	$\mathbf{85.6 \pm 0.2}$
Intersect	84.3 ± 0.5	$\mathbf{84.7 \pm 0.4}$	84.6 ± 0.4	85.1 ± 0.5

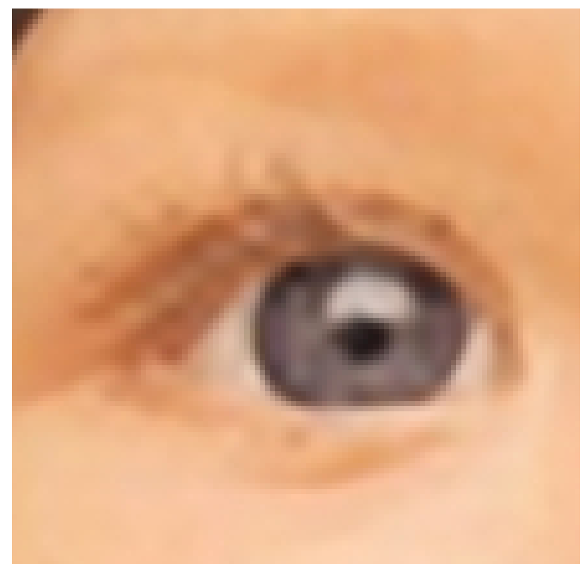
Yang et al. (2009a) have won the PASCAL VOC'09 challenge using this kind of techniques.

Task-driven dictionary learning (Mairal, Bach, and Ponce, 2010a)

- Define $\alpha^*(D, x) = \text{argmin}_\alpha \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1$
- α is used as a code for x
- **Direct optimization of $\alpha^*(D, x)$ with respect to D**
  - Application to image processing tasks such inverse half-toning (Mairal, Bach, and Ponce, 2010a)
  - Image super-resolution (Couzinie-Devy, Mairal, Bach, and Ponce, 2011)

Digital Zooming (Couzinie-Devy et al., 2011)

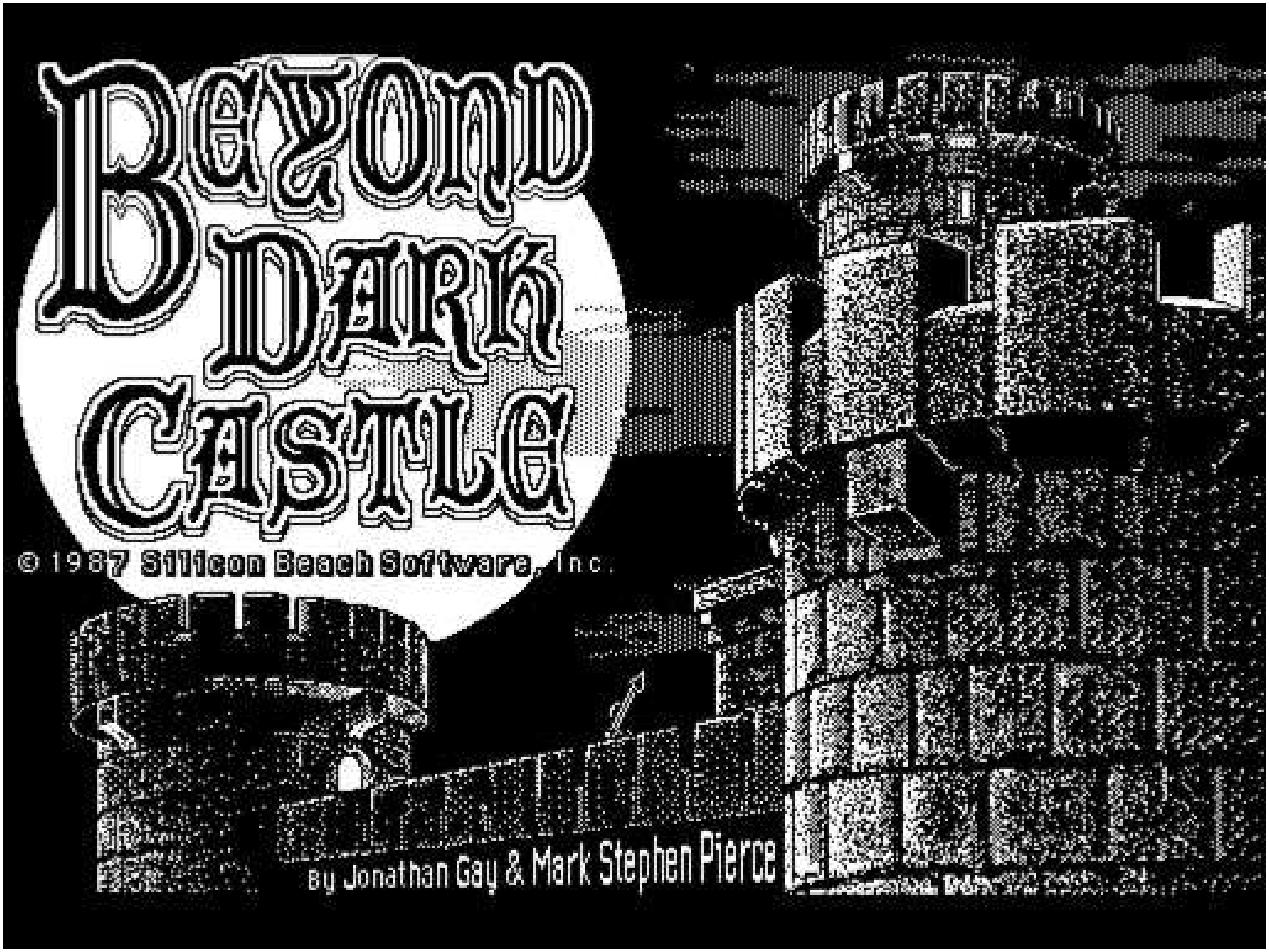
Digital Zooming (Couzinie-Devy et al., 2011)



Inverse half-toning (Mairal et al., 2010a)

Inverse half-toning (Mairal et al., 2010a)

Inverse half-toning (Mairal et al., 2010a)



Inverse half-toning (Mairal et al., 2010a)

Inverse half-toning (Mairal et al., 2010a)

Inverse half-toning (Mairal et al., 2010a)

Outline

Sparse methods for machine learning and computer vision

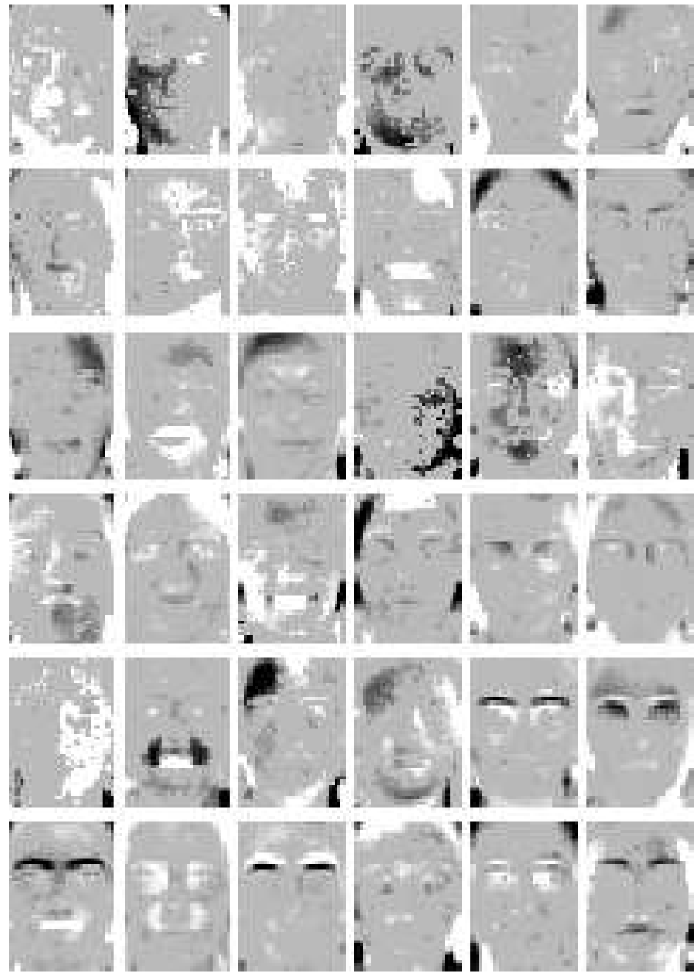
- **Introduction**
- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis
- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering
- **Sparsity for computer vision**
  - Task-driven dictionary learning
- **Structured sparsity**

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., 2010b)

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

raw data

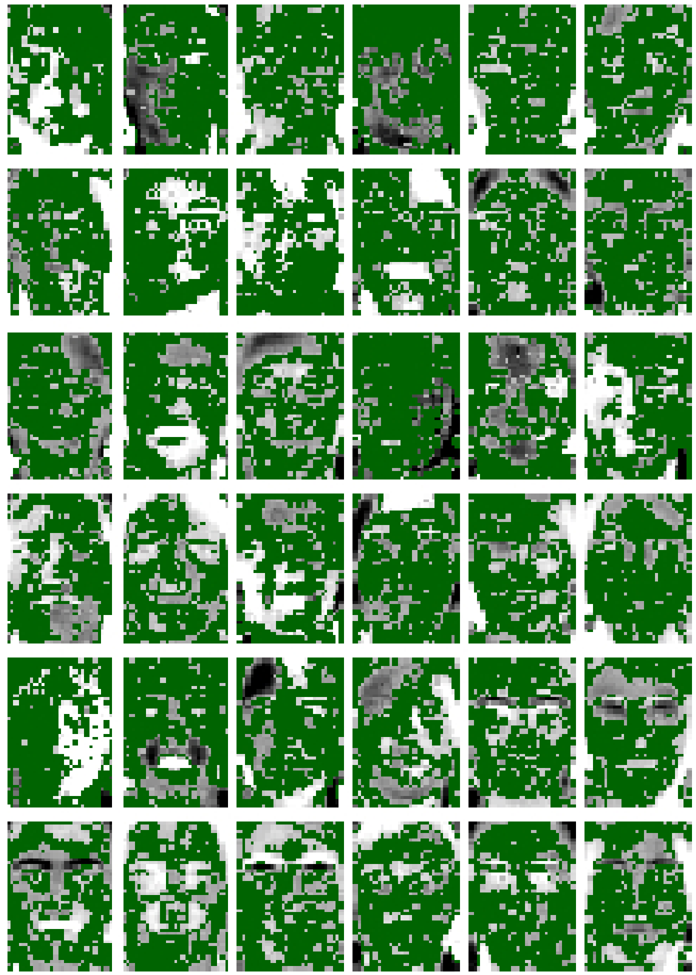


sparse PCA

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

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

raw data

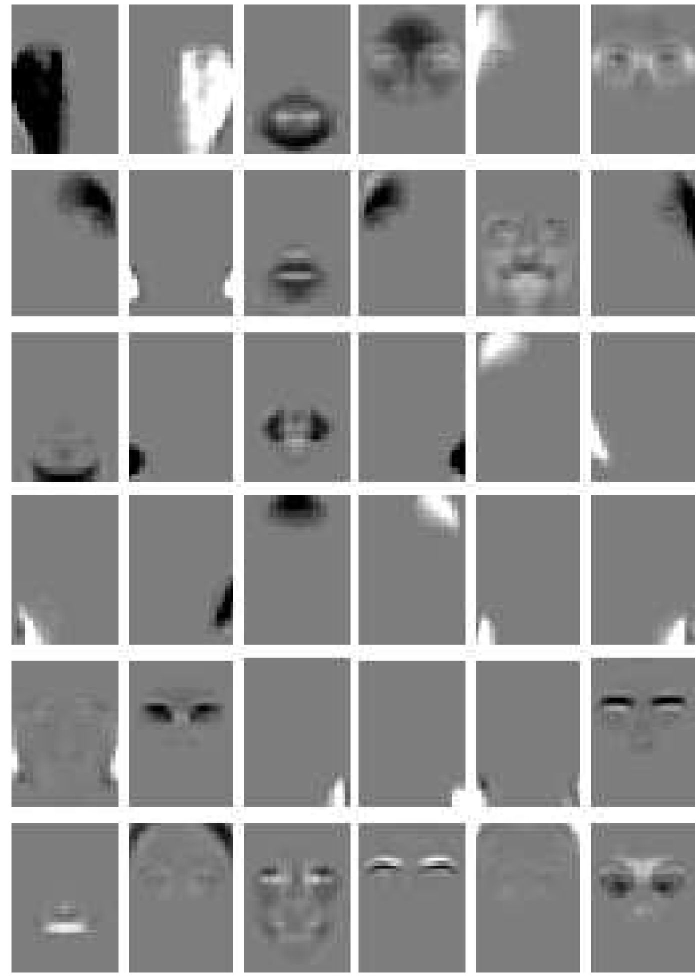


sparse PCA

- Unstructed sparse PCA ⇒ many zeros do not lead to better interpretability

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

raw data

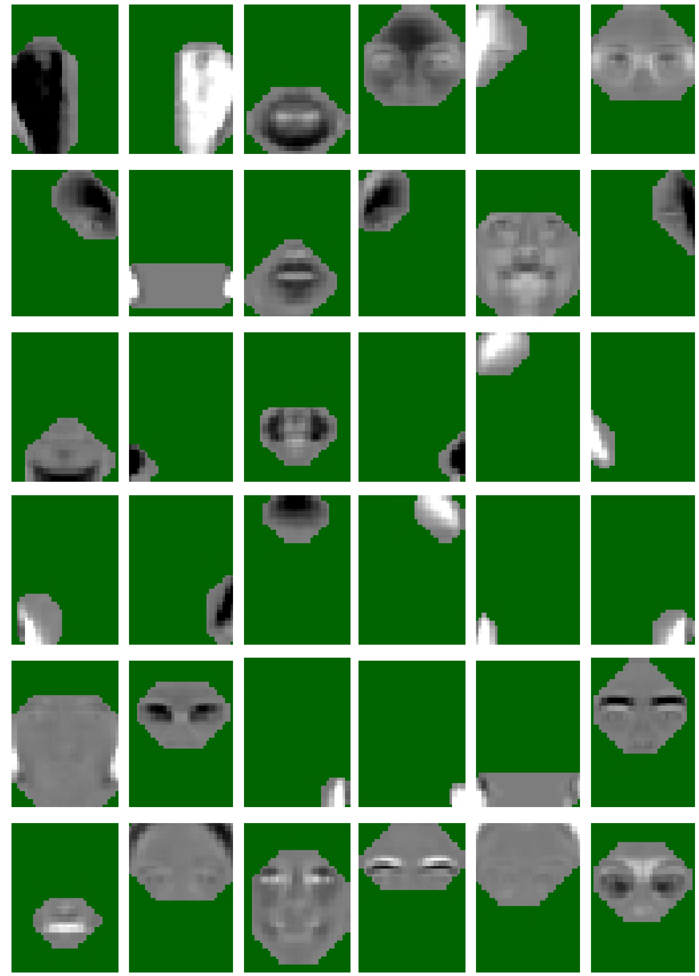


Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ 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., 2010b)

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., 2010b)
- **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, 2008b)

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)

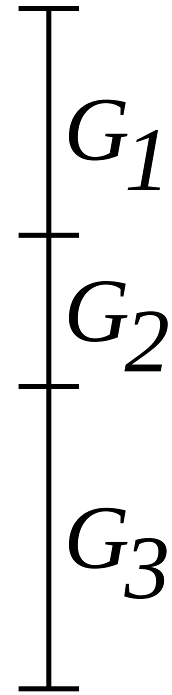
- **Non-convex approaches**
  - Haupt and Nowak (2006); Baraniuk et al. (2008); Huang et al. (2009)

- **Convex approaches**
  - Design of sparsity-inducing norms

Sparsity-inducing norms

- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

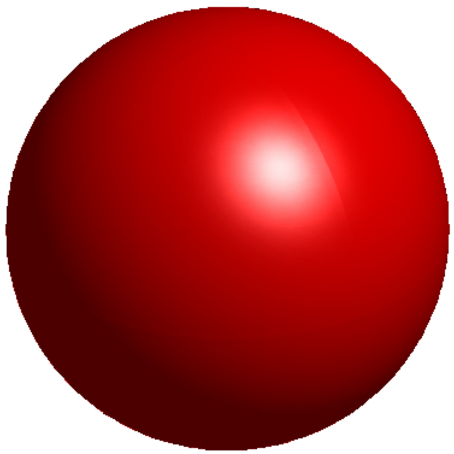
$$\sum_{G \in \mathbf{H}} \|w_G\|_2 = \sum_{G \in \mathbf{H}} \Big(\sum_{j \in G} w_j^2 \Big)^{1/2}$$



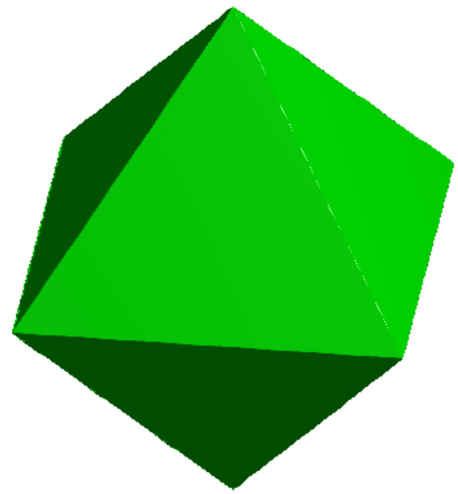
  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

Unit norm balls

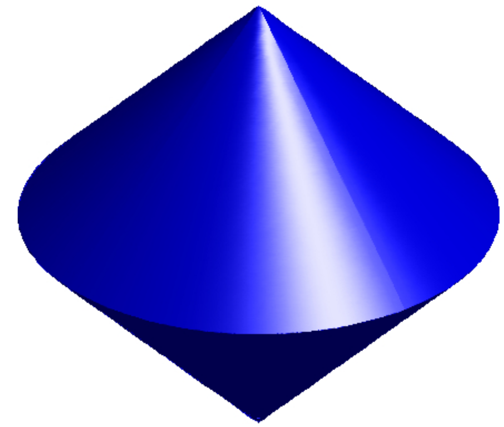
Geometric interpretation



$\|w\|_2$



$\|w\|_1$

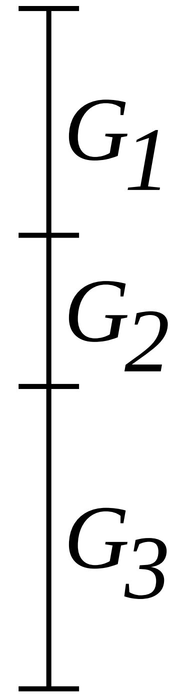


$\sqrt{w_1^2 + w_2^2} + |w_3|$

Sparsity-inducing norms

- **Popular choice for Ω**
  - The ℓ_1-ℓ_2 norm,

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

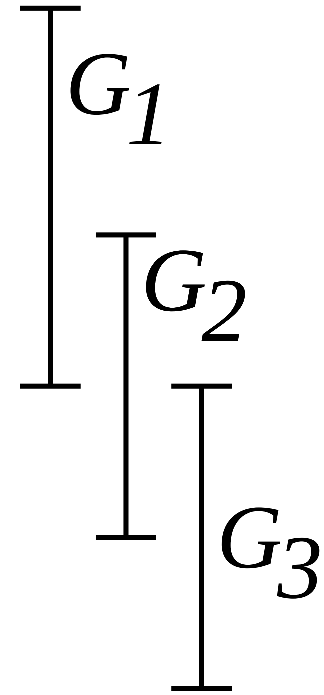


  - with $\mathbf{H}$ a partition of $\{1, \ldots, p\}$
  - The ℓ_1-ℓ_2 norm sets to zero groups of non-overlapping variables (as opposed to single variables for the ℓ_1-norm)
  - For the square loss, group Lasso (Yuan and Lin, 2006)

- 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)

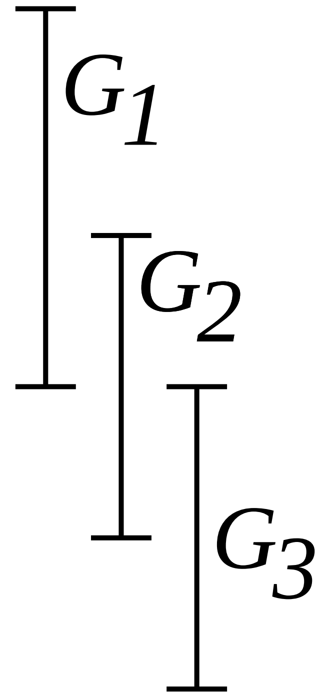


- When penalizing by the ℓ_1-ℓ_2 norm,

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

  - 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,

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

  – 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

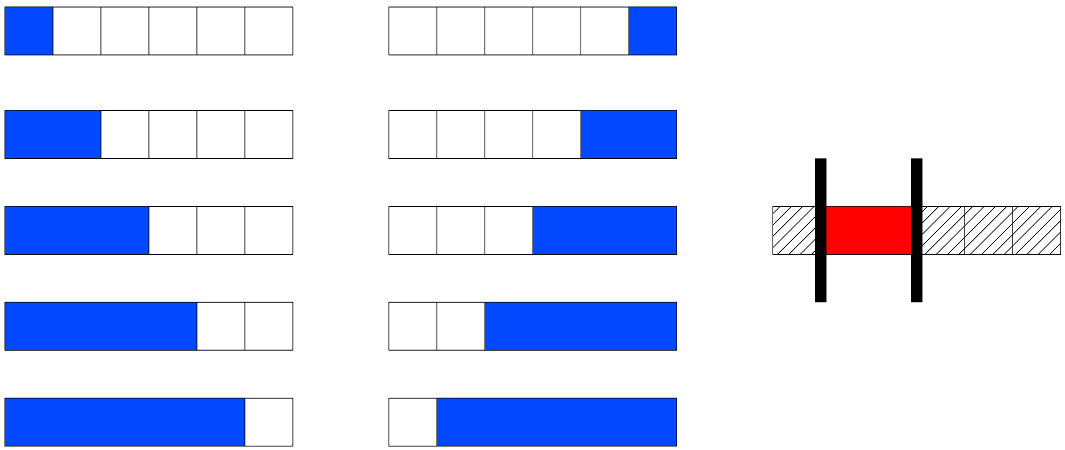
- The zero pattern of w is given by

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

- **Zero patterns are unions of groups**

Examples of set of groups $\mathbf{H}$

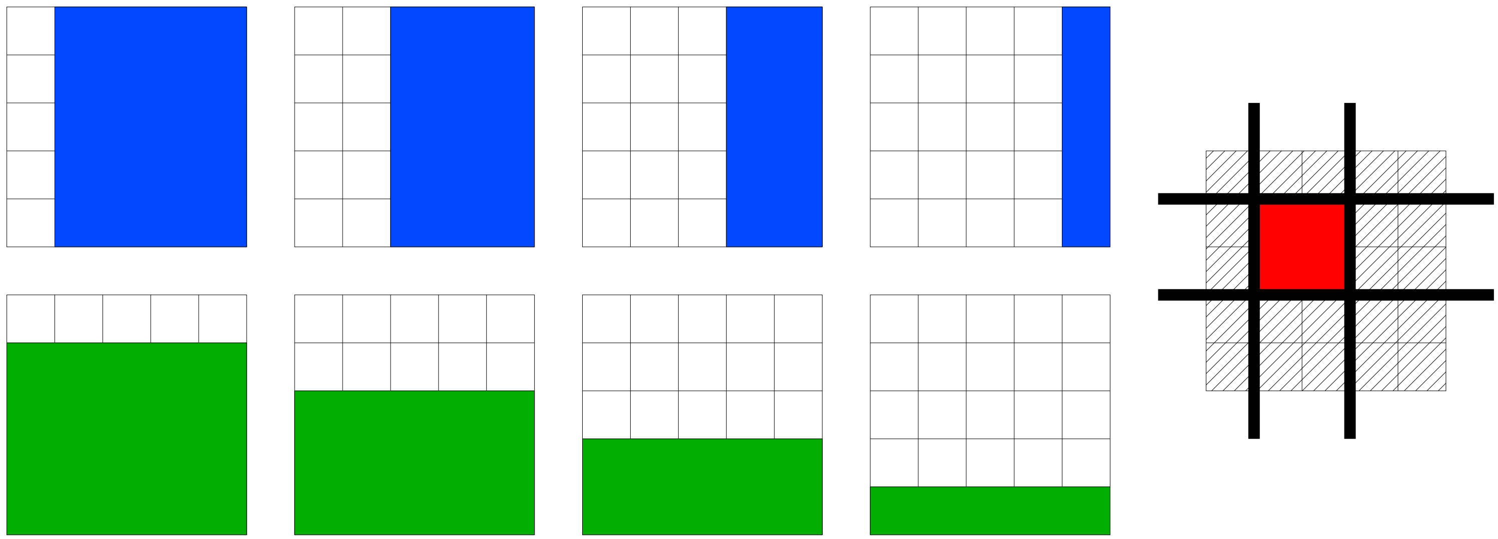
- Selection of contiguous patterns on a sequence, $p = 6$



  - $\mathbf{H}$ 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 $\mathbf{H}$

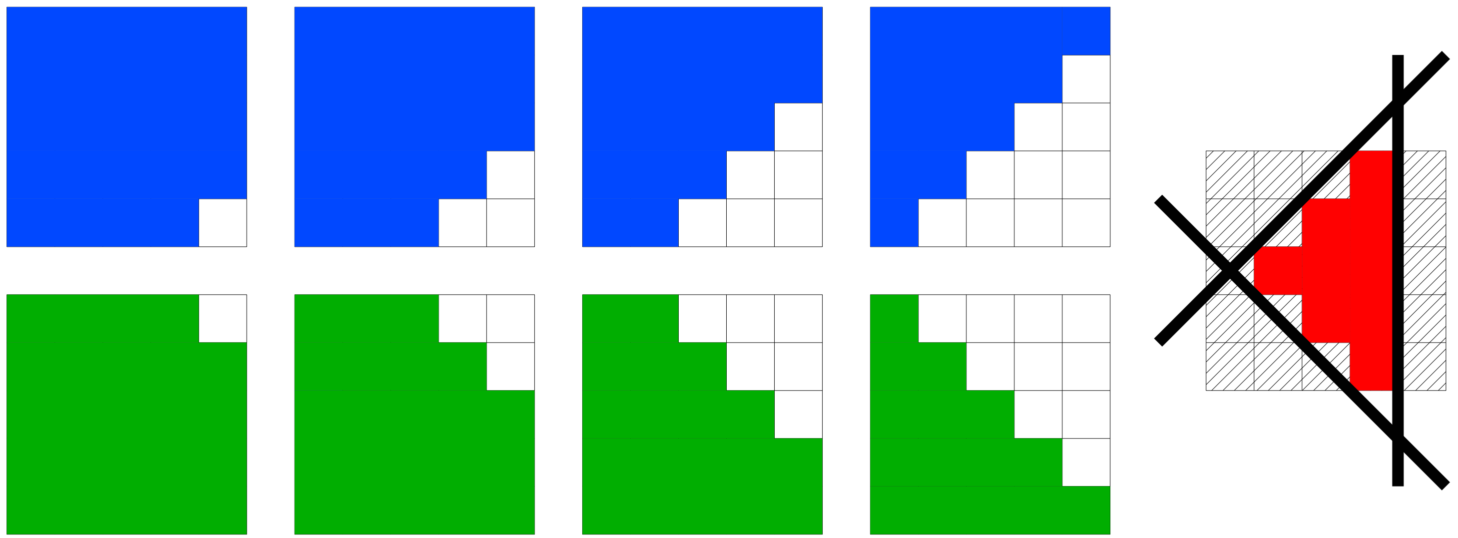
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{H}$ 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 $\mathbb{H}$

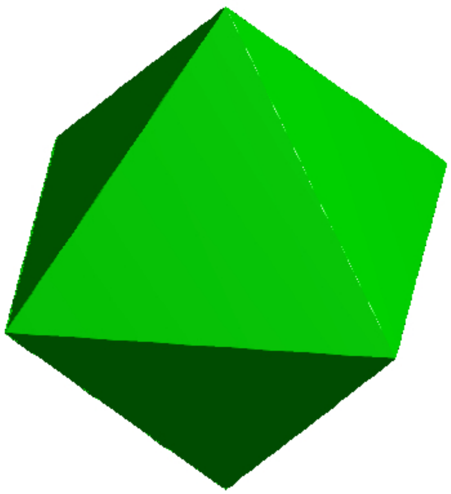
- Selection of diamond-shaped patterns on a 2-D grids, $p = 25$.



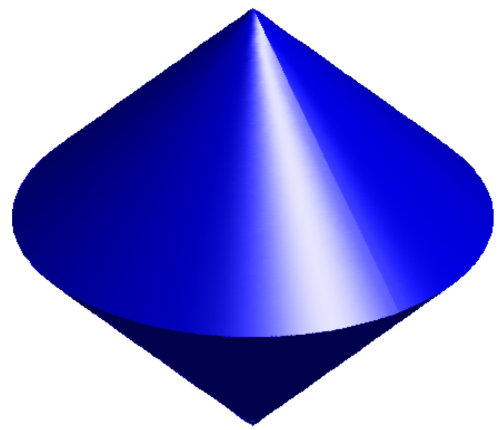
  - It is possible to extend such settings to 3-D space, or more complex topologies

Unit norm balls

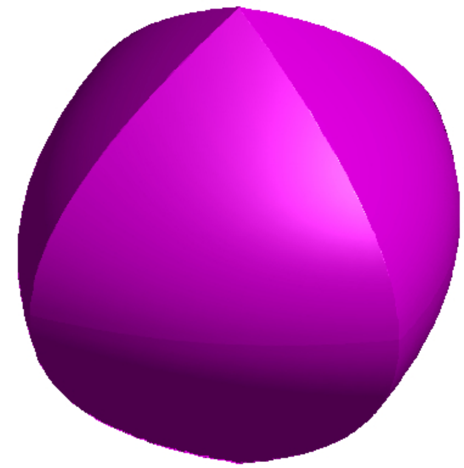
Geometric interpretation



$\|w\|_1$



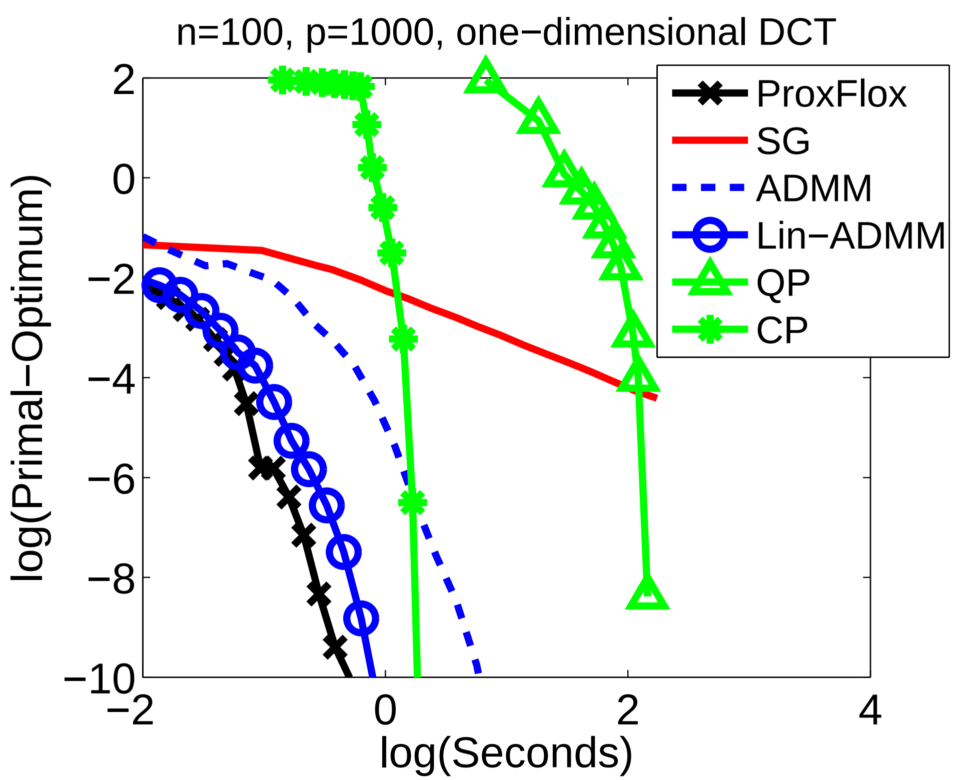
$\sqrt{w_1^2 + w_2^2} + |w_3|$



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

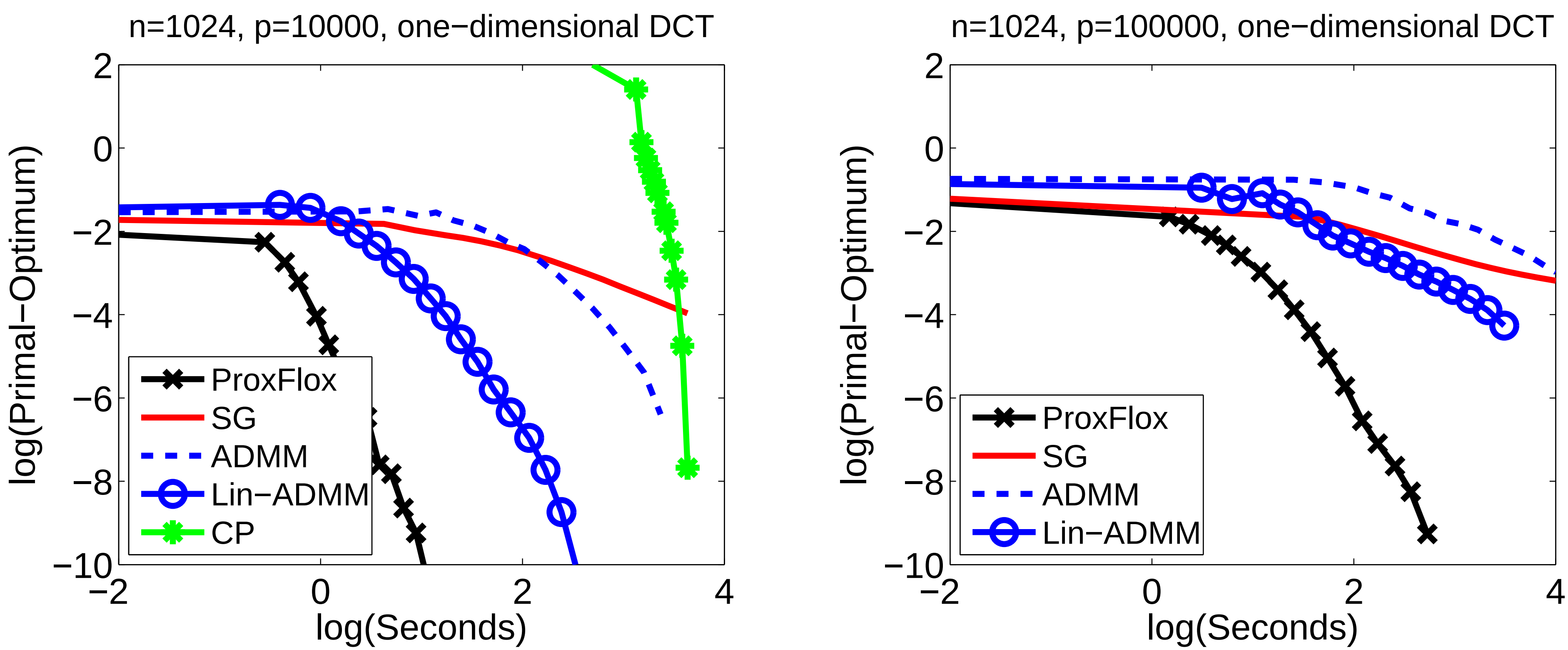
Comparison of optimization algorithms
(Mairal, Jenatton, Obozinski, and Bach, 2010b)
Small scale

- Specific norms which can be implemented through network flows



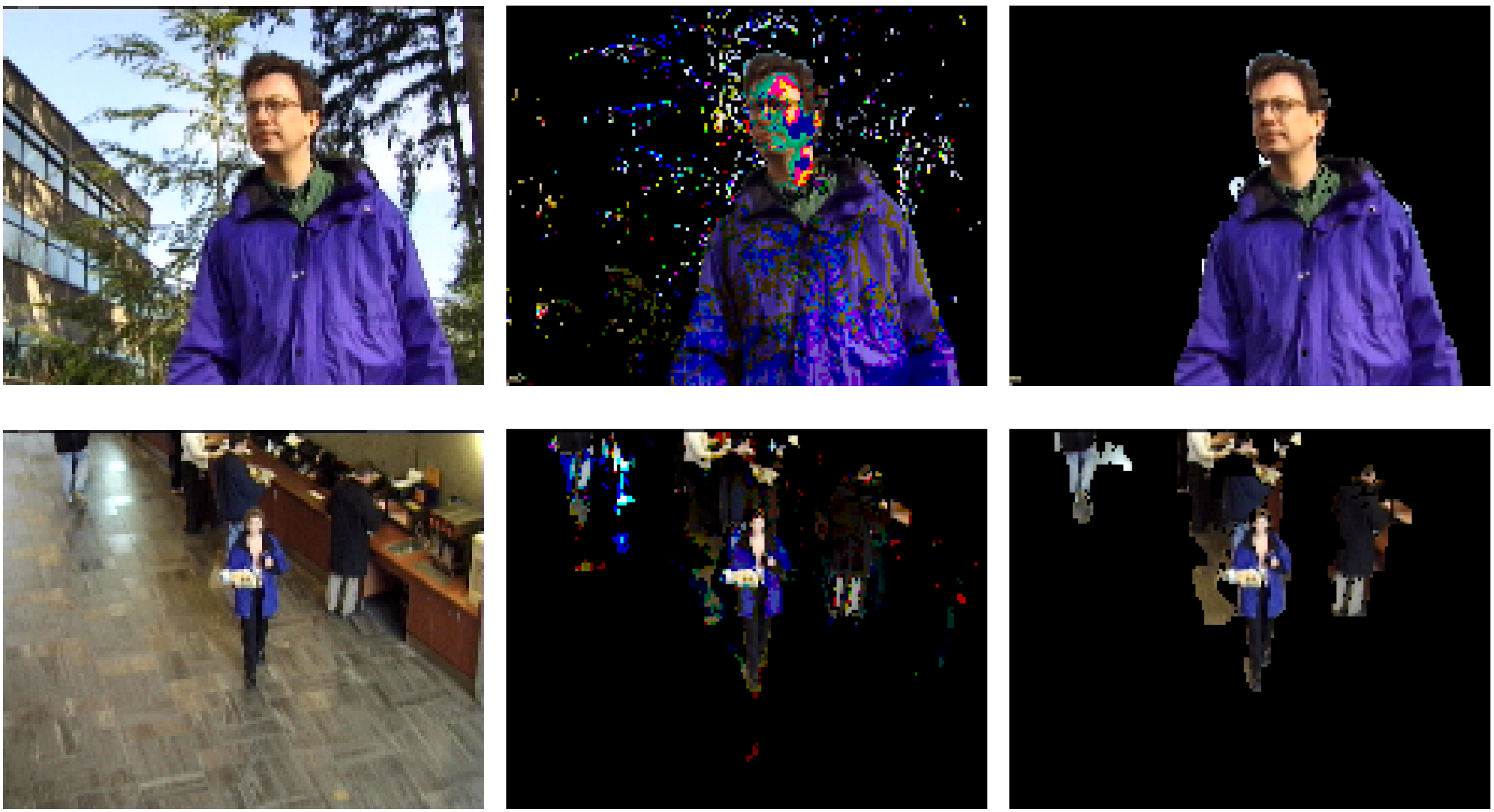
Comparison of optimization algorithms (Mairal, Jenatton, Obozinski, and Bach, 2010b) Large scale

- Specific norms which can be implemented through network flows



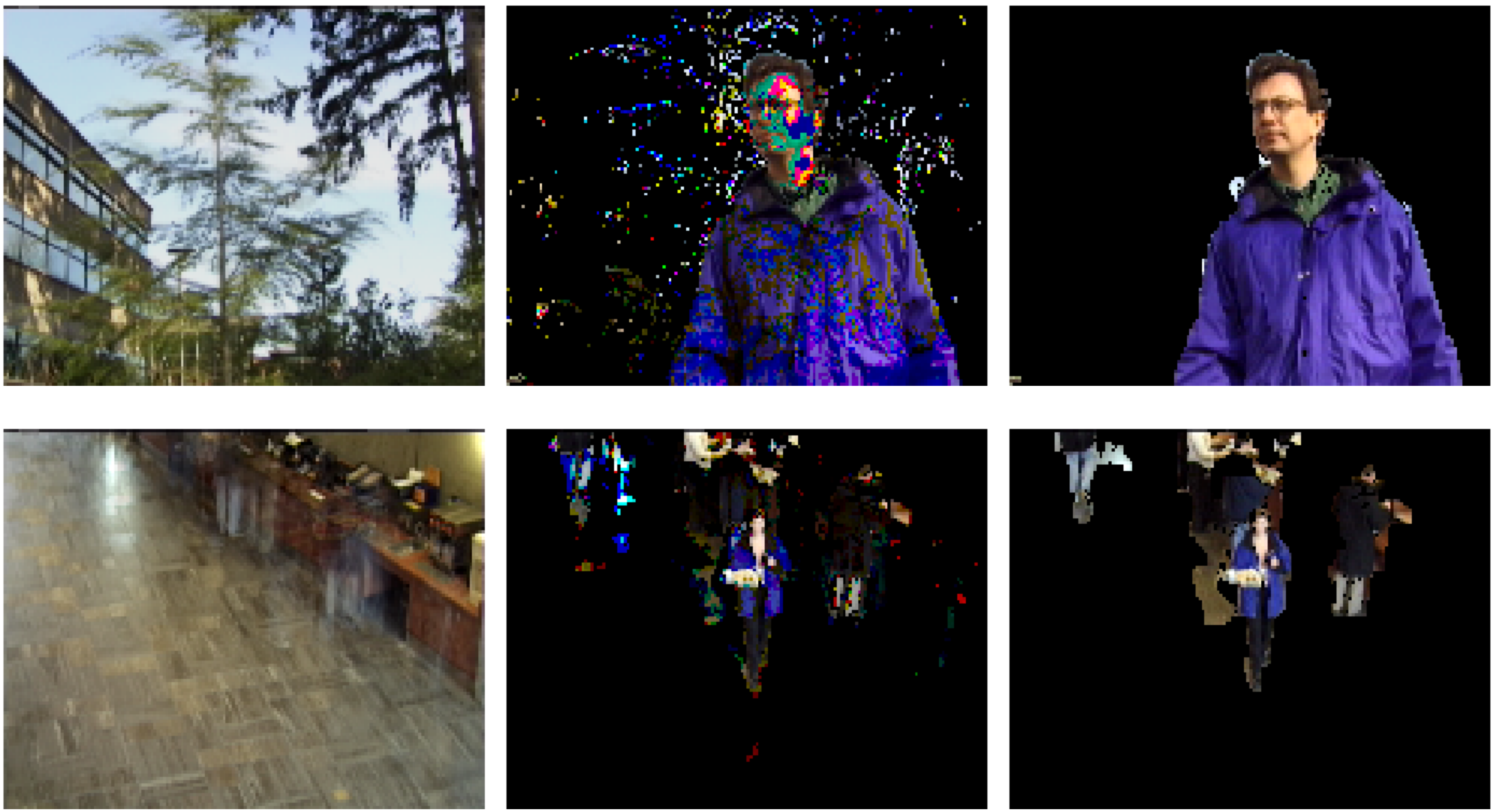
Application to background subtraction
(Mairal, Jenatton, Obozinski, and Bach, 2010b)

Input	ℓ_1-norm	Structured norm



Application to background subtraction
(Mairal, Jenatton, Obozinski, and Bach, 2010b)

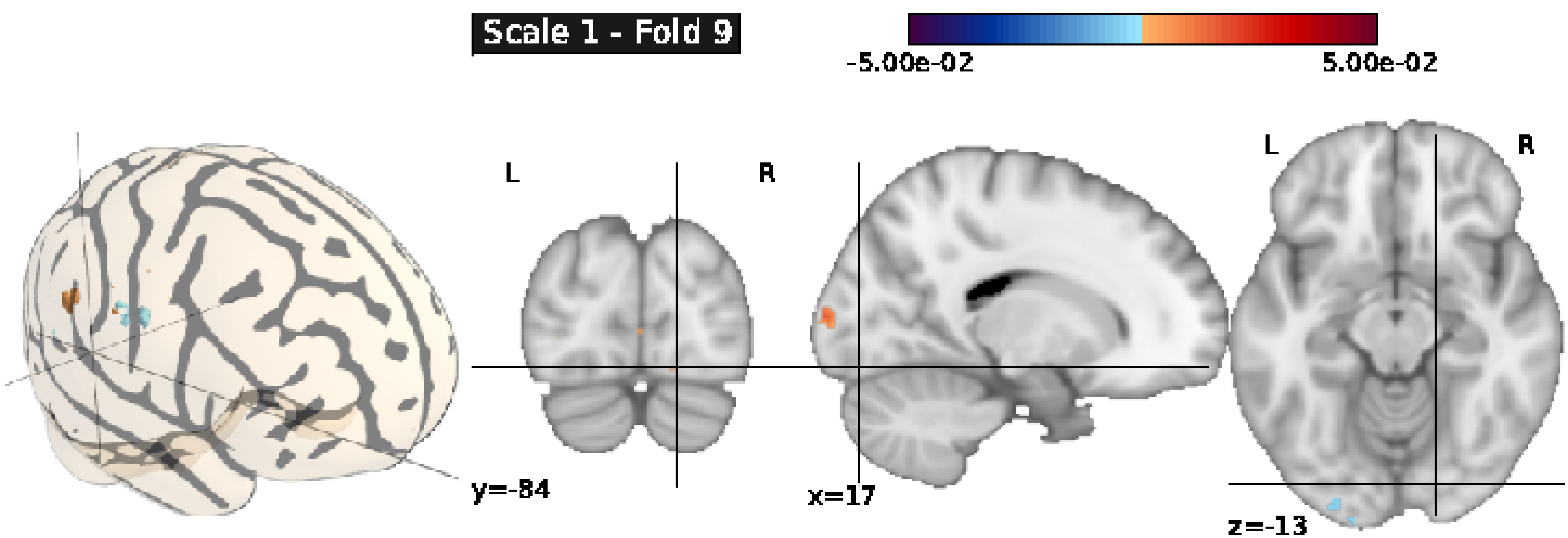
Background | ℓ_1-norm | Structured norm



Application to neuro-imaging

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

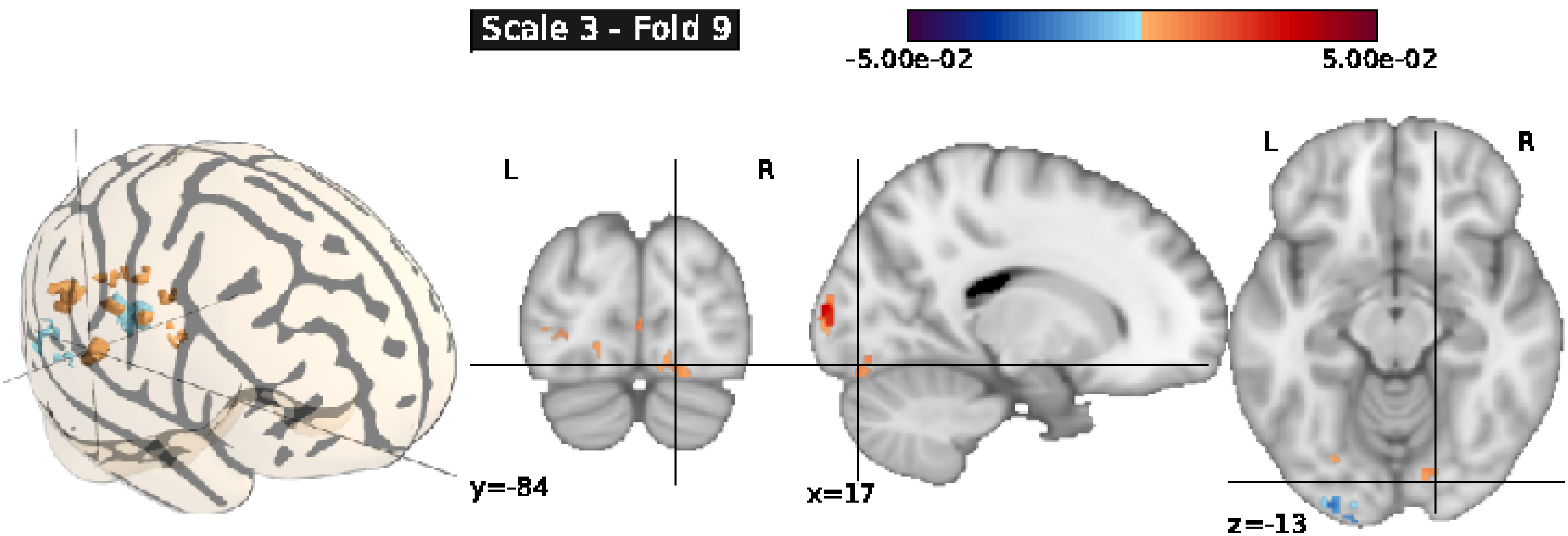
- "Brain reading": prediction of (seen) object size
- Multi-scale activity levels through hierarchical penalization



Application to neuro-imaging

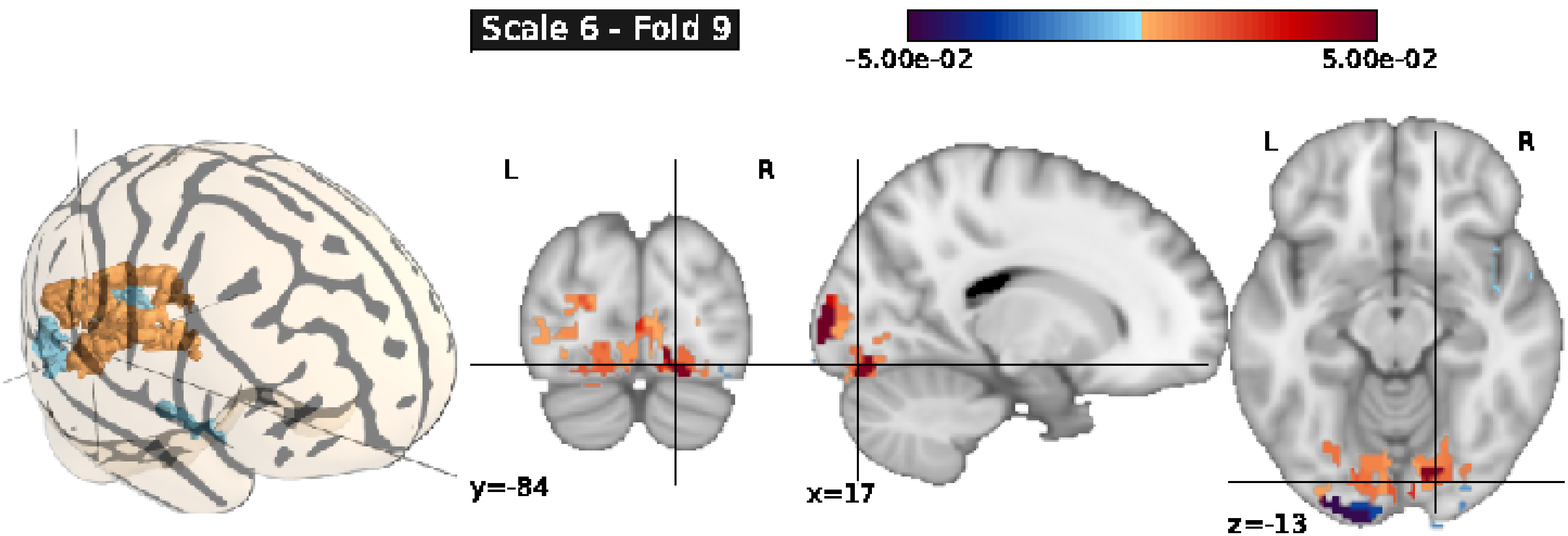
Structured sparsity for fMRI (Jenatton et al., 2011)

- "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

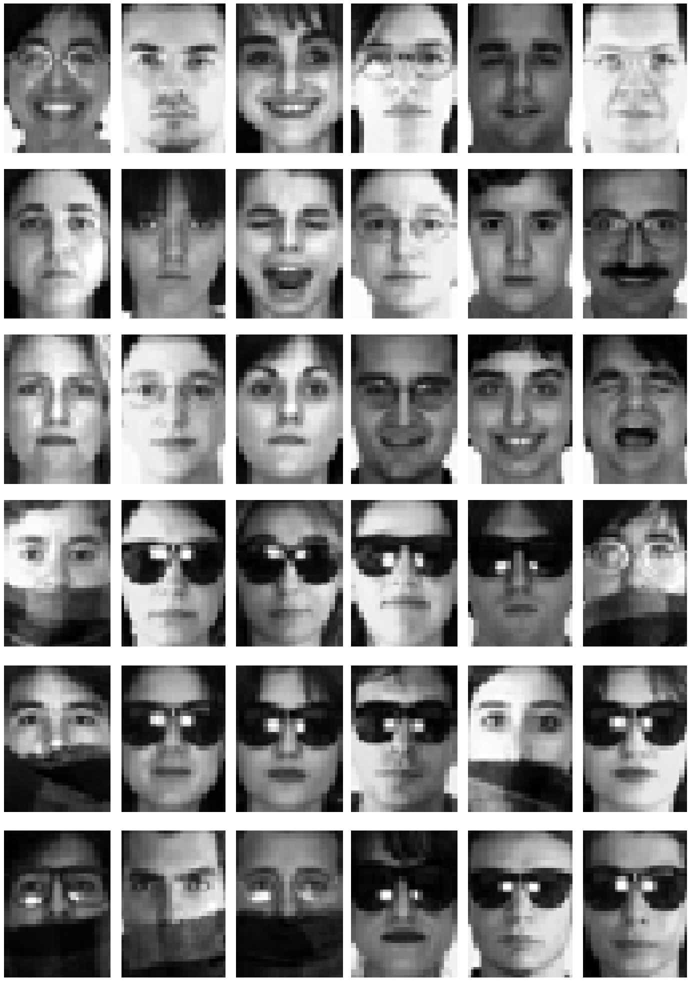


Sparse Structured PCA
(Jenatton, Obozinski, and Bach, 2009b)

- 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)

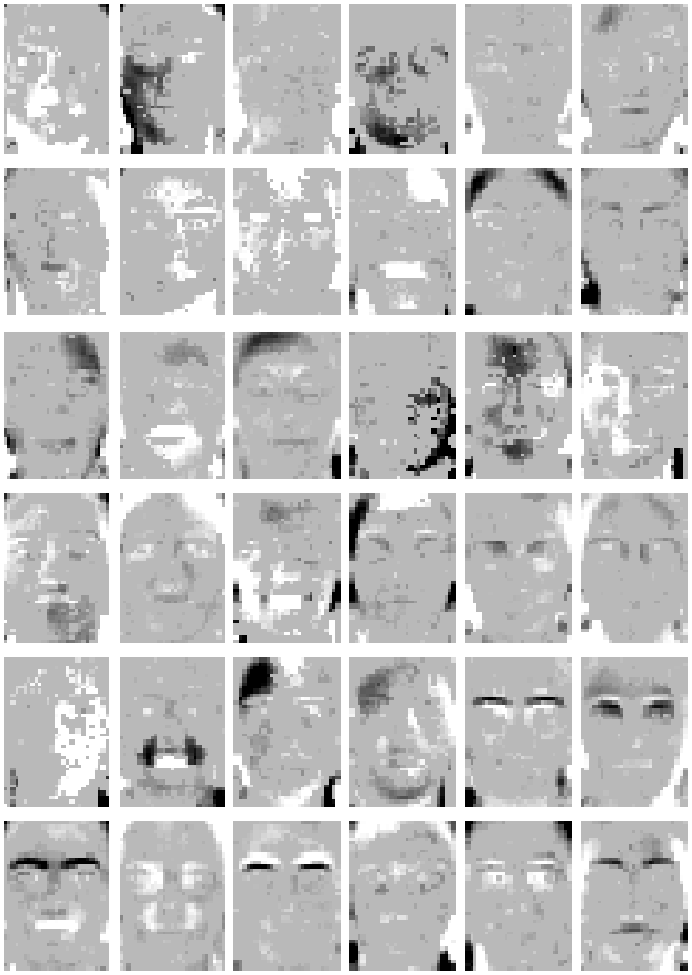


raw data

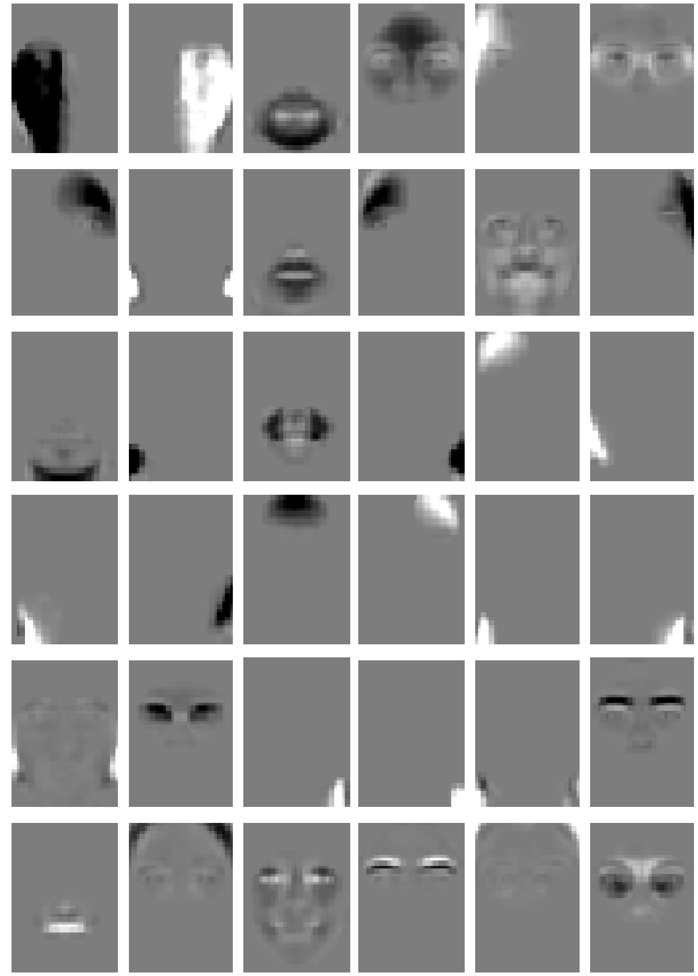
(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)



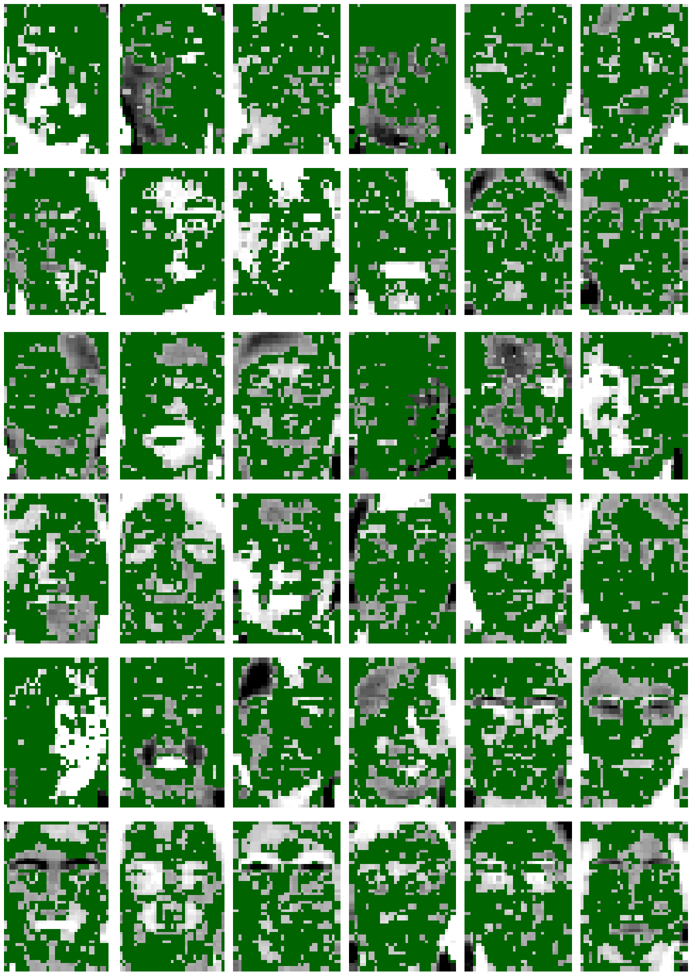
(unstructured) sparse PCA



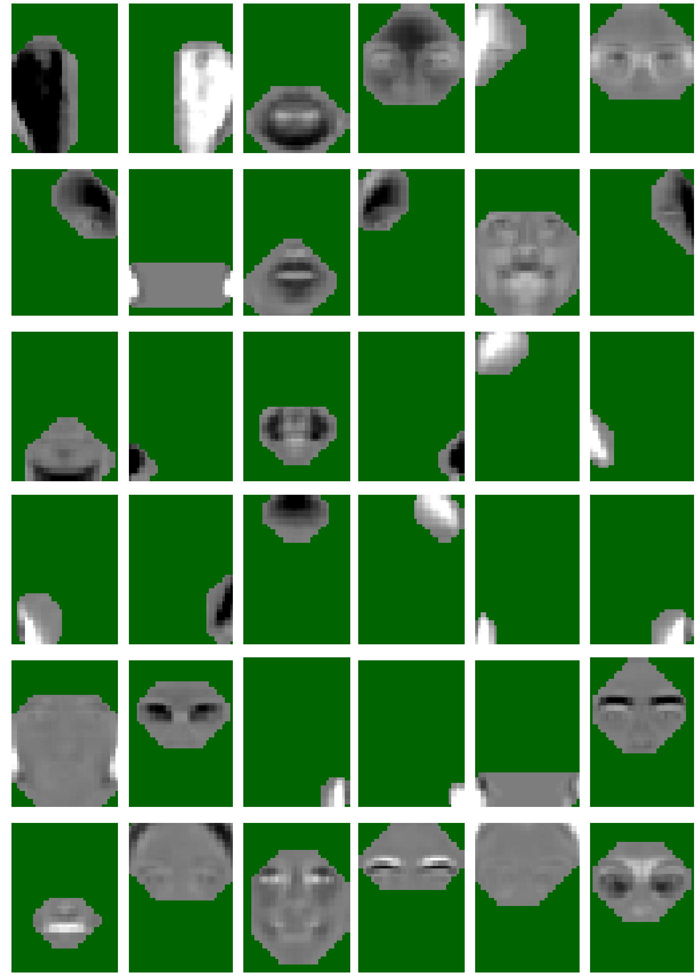
Structured sparse PCA

- Enforce selection of convex nonzero patterns ⇒ robustness to occlusion

Application to face databases (2/3)



(unstructured) sparse PCA

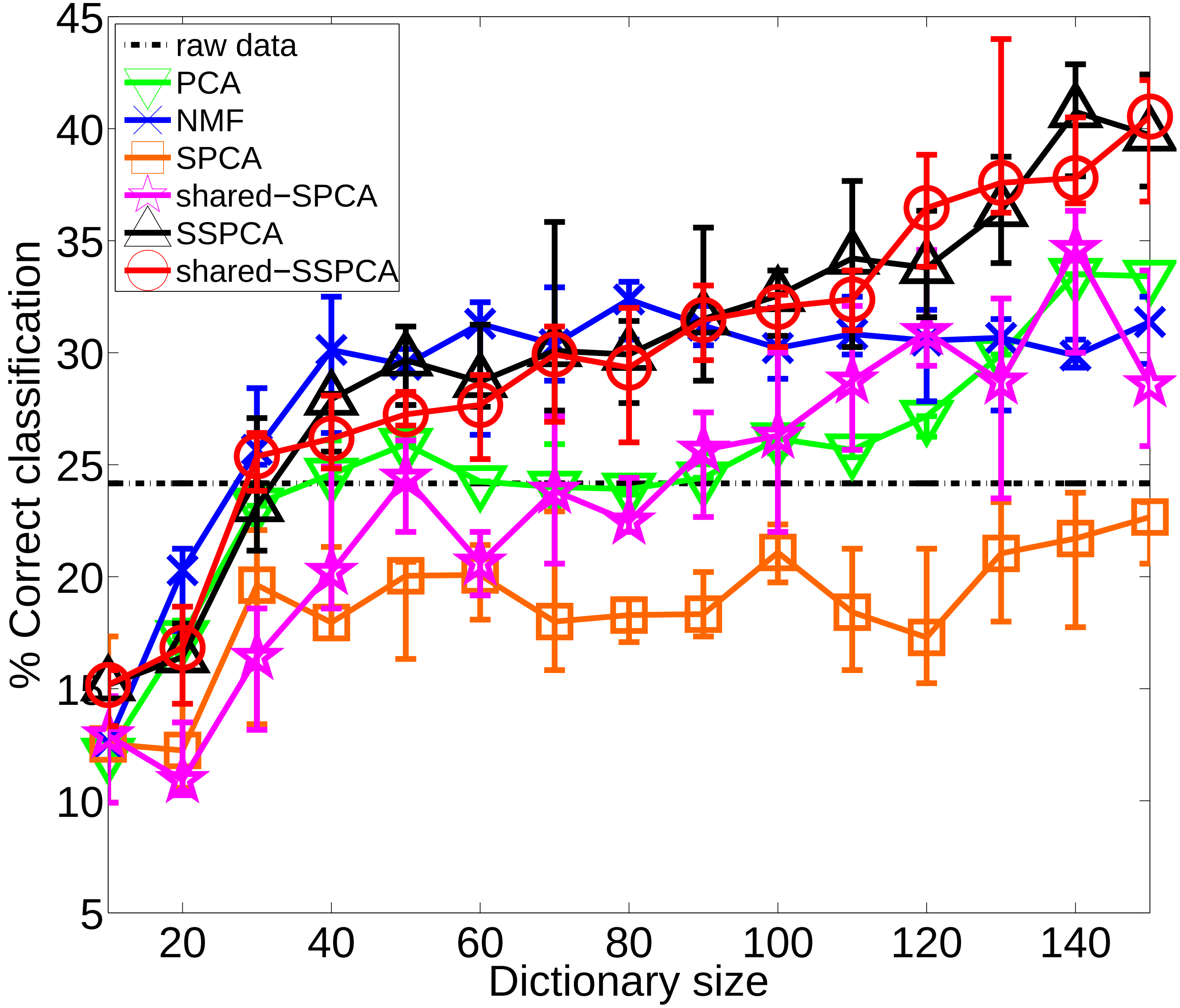


Structured sparse PCA

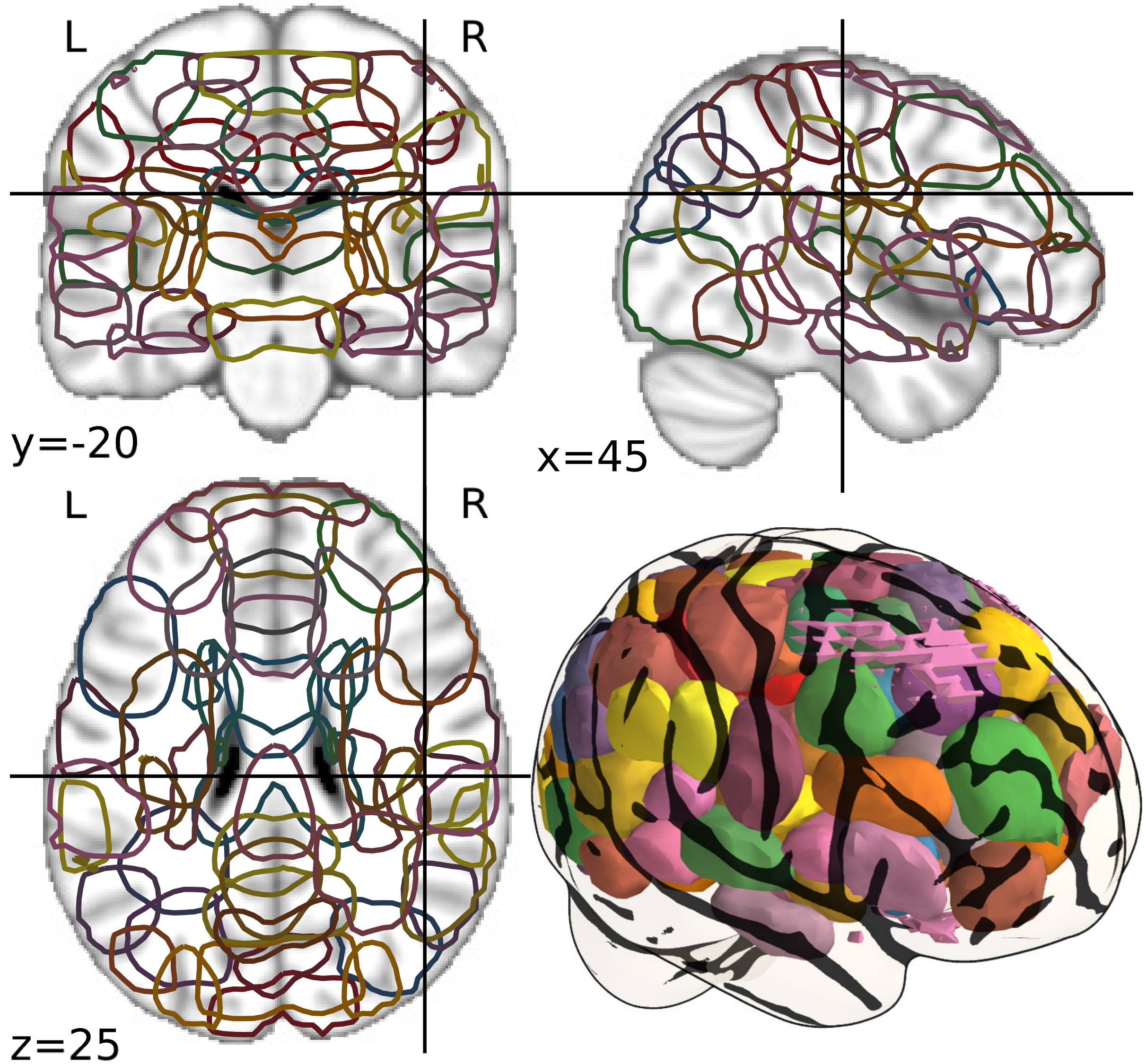
- Enforce selection of convex nonzero patterns $\Rightarrow$ robustness to occlusion

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**):

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

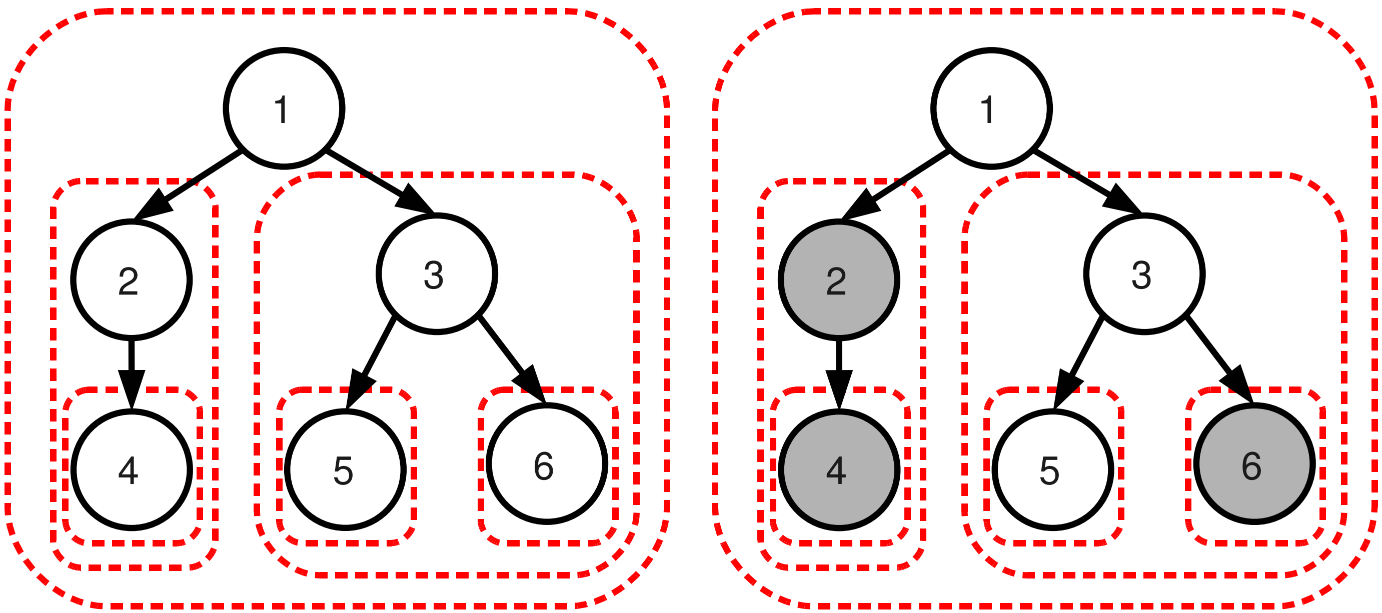
- Dictionary learning with **structured sparsity for codes** w:

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

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

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



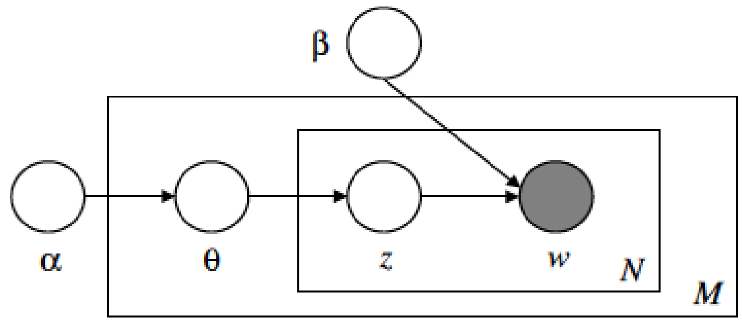
- Variable selected after its ancestors (Zhao et al., 2009; Bach, 2008b)

Hierarchical dictionary learning
Modelling of text corpora

- Each document is modelled through word counts
  - Low-rank matrix factorization of word-document matrix
  - Similar to NMF with multinomial loss

- Probabilistic topic models (Blei et al., 2003a)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**

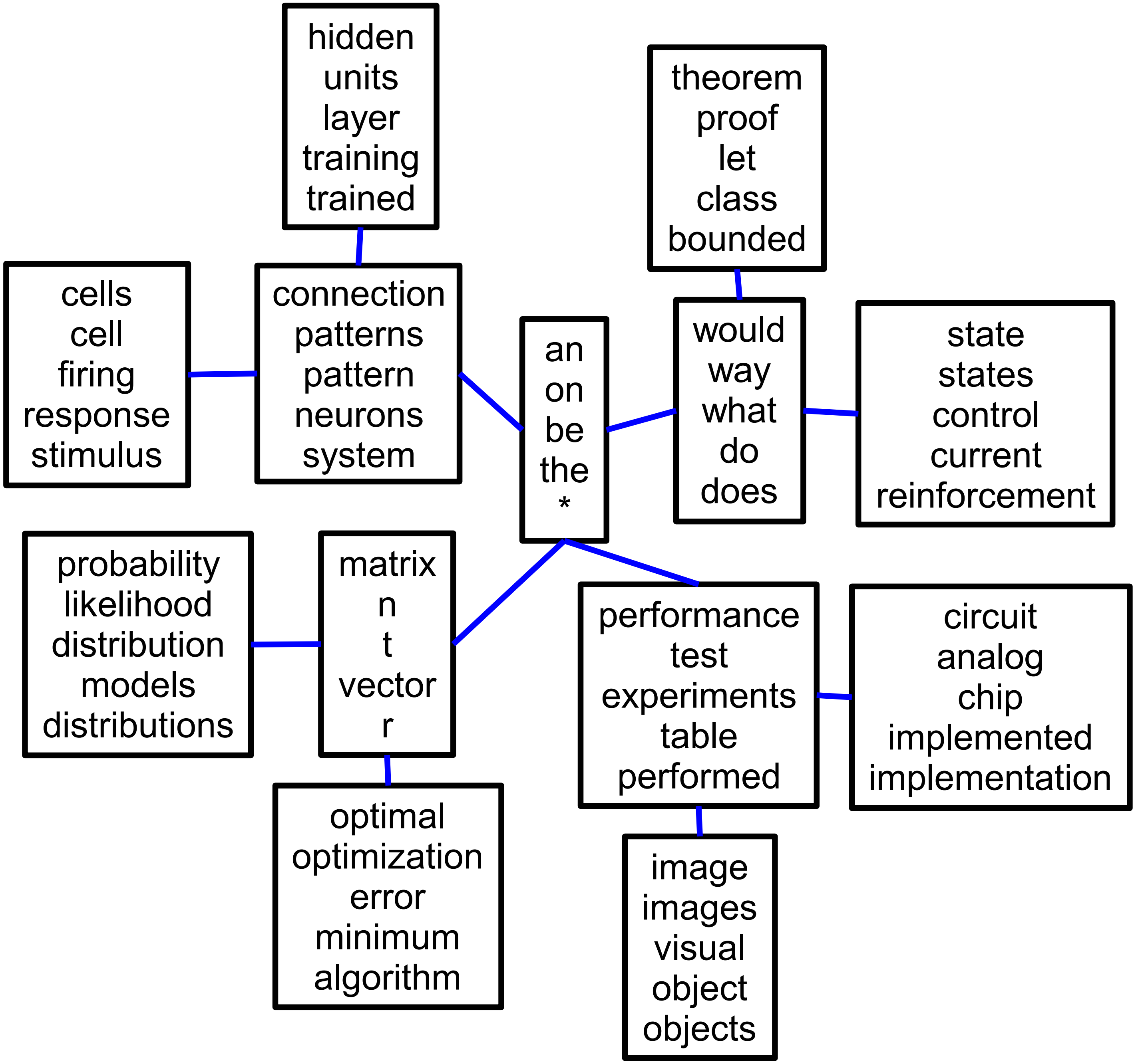
Topic models and matrix factorization



- **Latent Dirichlet allocation** (Blei et al., 2003b)
  - For a document, sample $\theta \in \mathbb{R}^k$ from a Dirichlet(α)
  - For the n-th word of the same document,
    - sample a topic z_n from a multinomial with parameter θ
    - sample a word w_n from a multinomial with parameter $\beta(z_n, :)$

- **Interpretation as multinomial PCA** (Buntine and Perttu, 2003)
  - Marginalizing over topic z_n, given θ, each word w_n is selected from a multinomial with parameter $\sum_{z=1}^k \theta_z \beta(z,:) = \beta^\top \theta$
  - Row of β = dictionary elements, θ code for a document

Modelling of text corpora - Dictionary tree



Topic models, NMF and matrix factorization

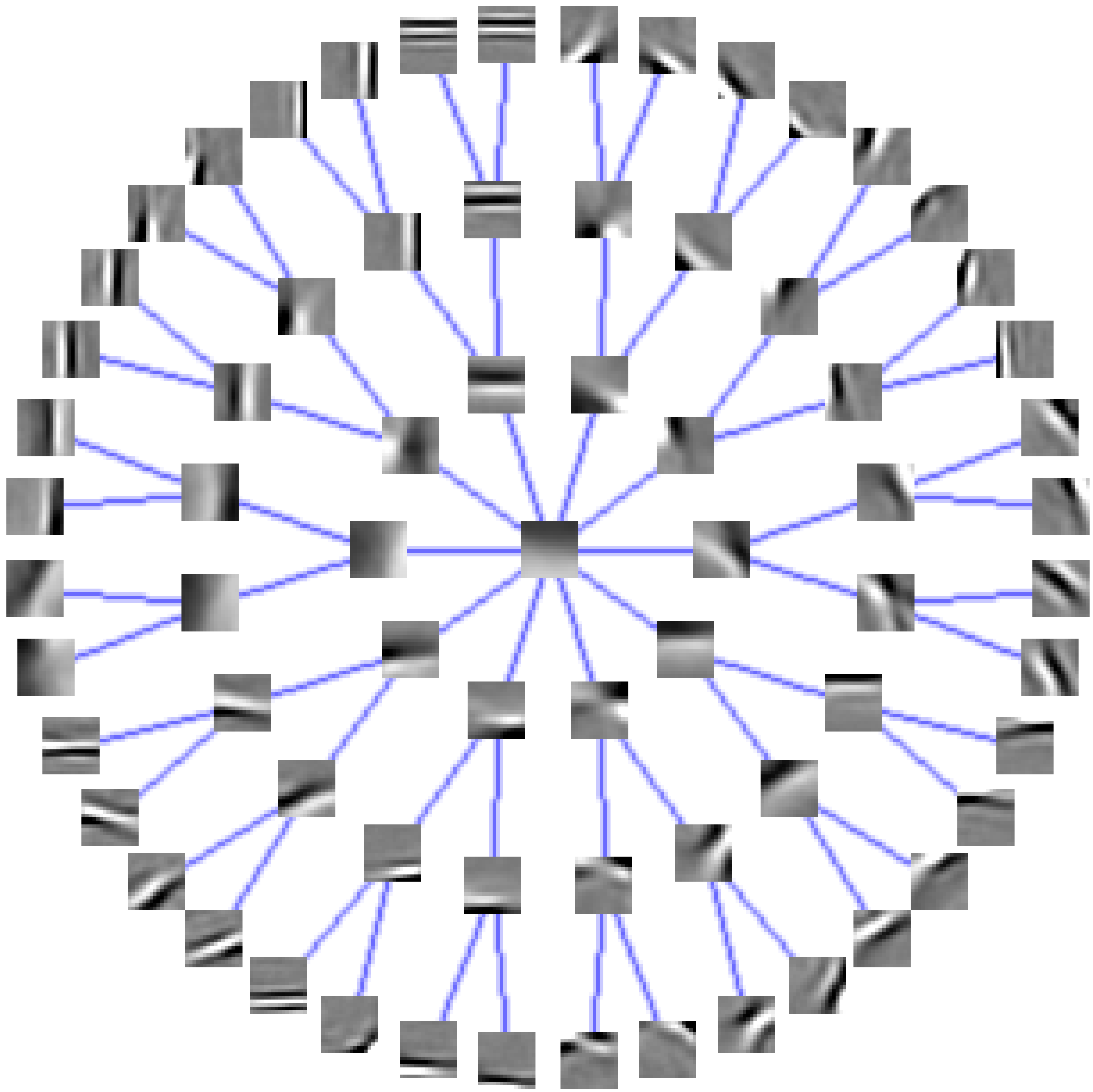
- **Three different views on the same problem**
  - Interesting parallels to be made
  - Common problems to be solved
- **Structure on dictionary/decomposition coefficients** with adapted priors, e.g., nested Chinese restaurant processes (Blei et al., 2004)
- **Learning hyperparameters from data**
- **Identifiability and interpretation/evaluation of results**
- **Discriminative tasks** (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008; Mairal et al., 2009c)
- **Optimization and local minima**

Structure on codes within dictionary learning

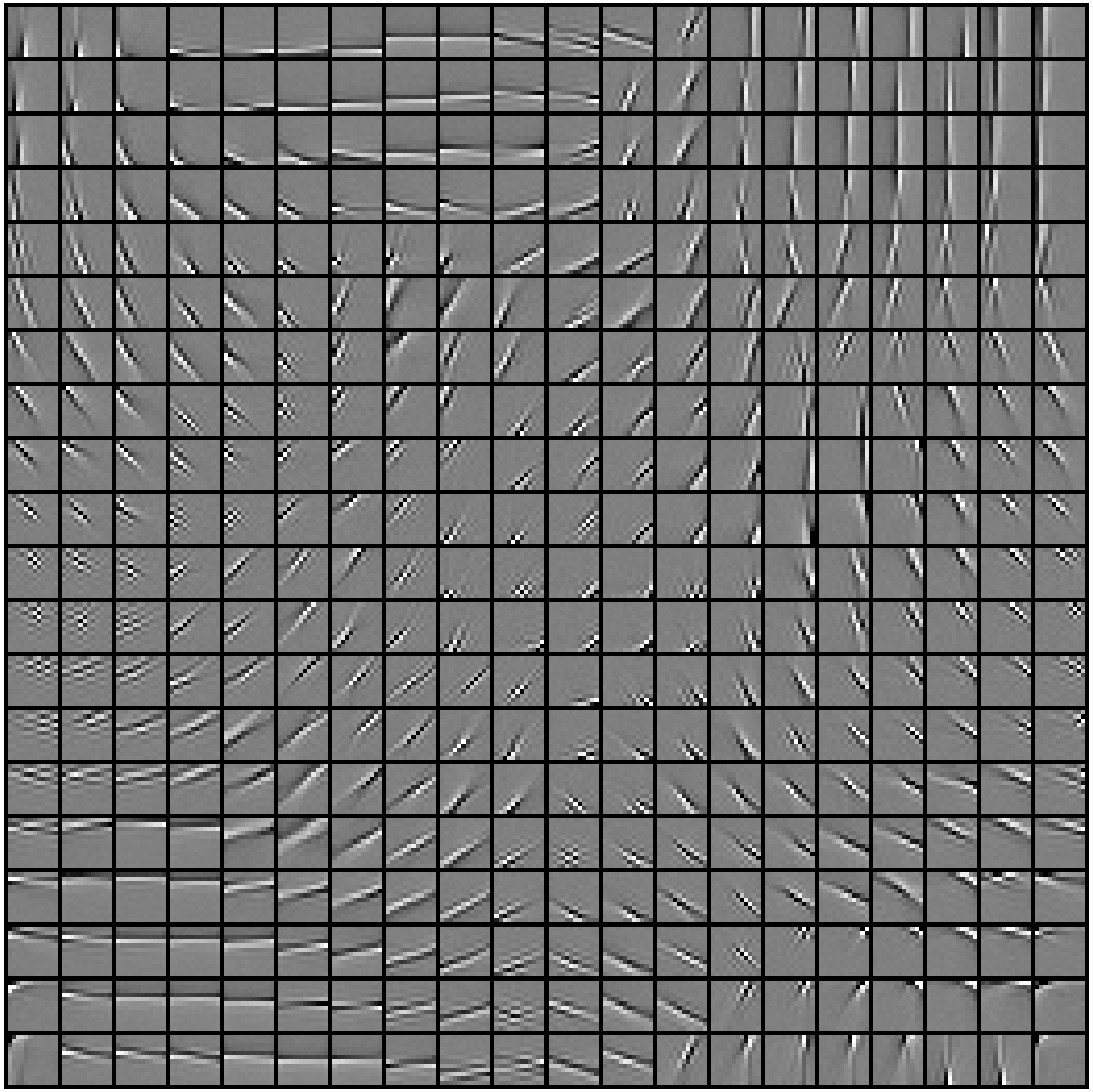
$$\min_{\substack{\mathbf{A}\in\mathbb{R}^{k\times n}\\ \mathbf{D}\in\mathbb{R}^{p\times k}}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda\psi(\boldsymbol{\alpha}_i) \text{ s.t. } \forall j,\ \|\mathbf{d}_j\|_2 \leq 1.$$

- **Impose topology between dictionary elements**
  - Hierarchical and topographic dictionaries for image patches
- **Grouping atoms**
  - Source separation

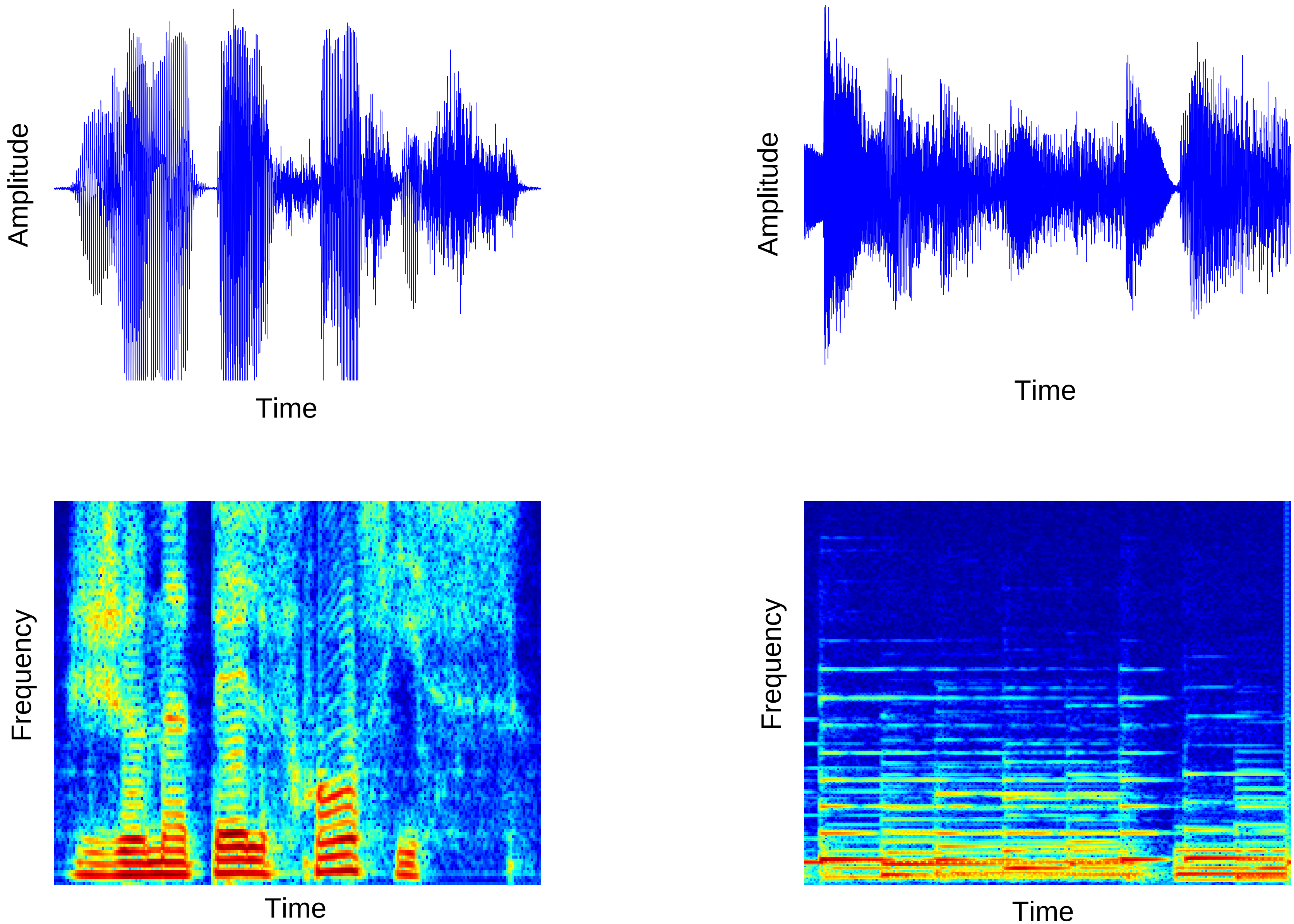
Hierarchical dictionaries (Jenatton et al., 2010)



Topographic dictionaries (Mairal et al., 2010b)

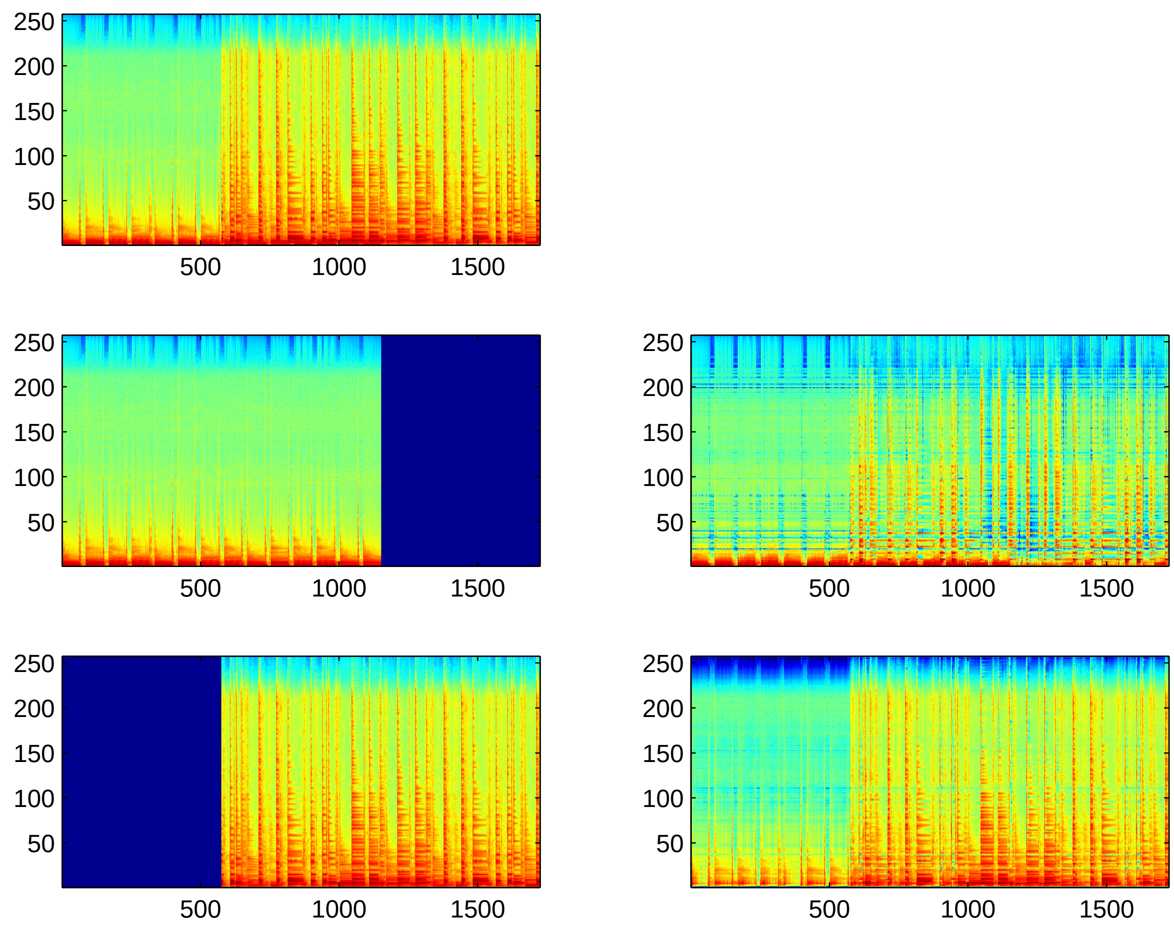


Structured sparsity - Audio processing
Source separation (Lefèvre et al., 2011)



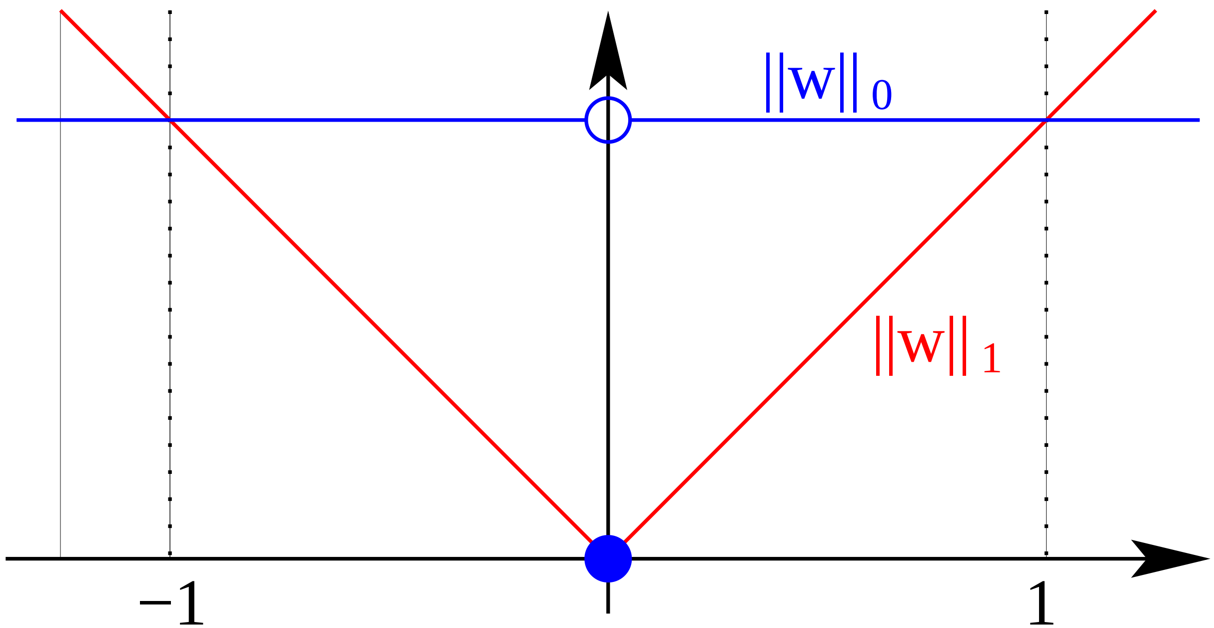
Structured sparsity - Audio processing
Musical instrument separation (Lefèvre et al., 2011)

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



ℓ_1-norm = convex envelope of cardinality of support

- Let $w \in \mathbb{R}^p$. Let $V = \{1, \dots, p\}$ and $\mathrm{Supp}(w) = \{j \in V,\ w_j \neq 0\}$
- **Cardinality of support**: $\|w\|_0 = \mathrm{Card}(\mathrm{Supp}(w))$
- 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 $F : 2^V \to \mathbb{R}$ be a **set-function**
  - Assume F is **non-decreasing** (i.e., $A \subset B \Rightarrow F(A) \leqslant F(B)$)
  - Explicit prior knowledge on supports (Haupt and Nowak, 2006; Baraniuk et al., 2008; Huang et al., 2009)

- Define $\Theta(w) = F(\mathrm{Supp}(w))$: How to get its convex envelope?

  1. Possible if F is also **submodular**
  2. Allows **unified** theory and algorithm
  3. Provides **new** regularizers

- References on submodular functions (Fujishige, 2005; Bach, 2010)

Outline

Sparse methods for machine learning and computer vision

- **Introduction**

- **Tutorial on sparse methods**
  - Non-smooth optimization
  - Theoretical analysis

- **Sparsity for matrices**
  - Dictionary learning and collaborative filtering

- **Sparsity for computer vision**
  - Task-driven dictionary learning

- **Structured sparsity**

Conclusion

- **Sparsity for machine learning and vision**
  - Many applications (image, audio, text, etc.)
  - May be achieved through **structured** sparsity-inducing norms
  - May be adapted to a **discriminative** task

Conclusion

- **Sparsity for machine learning and vision**
  - Many applications (image, audio, text, etc.)
  - May be achieved through **structured** sparsity-inducing norms
  - May be adapted to a **discriminative** task

- **On-going work on structured sparsity**
  - Norm design through submodular functions (Bach, 2010)
  - Large-scale learning (Le Roux et al., 2012)

References

Y. Amit, M. Fink, N. Srebro, and S. Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine Learning (ICML)*, 2007.

F. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*, 2008a.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008b.

F. Bach. Structured sparsity-inducing norms through submodular functions. In *NIPS*, 2010.

F. Bach. Convex analysis and optimization with submodular functions: a tutorial. Technical Report 00527714, HAL, 2010.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.

F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv, 2008.

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.

L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Speech and Audio Processing*, 14(1):191, 2006.

D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1995.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

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

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.

D.M. Blei and J. McAuliffe. Supervised topic models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

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

L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.

Y-L. Boureau, F. Bach, Y. Lecun, and J. Ponce. Learning mid-level features for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

S. P. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

A. Buades, B. Coll, and J.-M. Morel. Non-local image and movie denoising. *International Journal of Computer vision*, 76(2):123–139, 2008.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, 35(4):1674–1697, 2007.

W. Buntine and S. Perttu. Is multinomial PCA multi-faceted clustering or dimensionality reduction. In

International Workshop on Artificial Intelligence and Statistics (AISTATS), 2003.

E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.

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.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

Florent Couzinie-Devy, Julien Mairal, Francis Bach, and Jean Ponce. Dictionary Learning for Deblurring and Digital Zoom. Technical report, September 2011. URL `http://hal.inria.fr/inria-00627402`.

A. d'Aspremont, F. Bach, and L. El Ghaoui. Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, 9:1269–1294, 2008.

D.L. Donoho and J. Tanner. Neighborliness of randomly projected simplices in high dimensions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9452, 2005.

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.

J. Fan and R. Li. Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1361, 2001.

M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American Control Conference*, volume 6, pages 4734–4739, 2001.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the itakura-saito divergence. with application to music analysis. *Neural Computation*, 21(3), 2009.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332, 2007.

W. Fu. Penalized regressions: the bridge vs. the Lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998).

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.

R. Grosse, R. Raina, H. Kwong, and A. Y. Ng. Shift-invariant sparse coding for audio classification. In *Proceedings of the Twenty-third Conference on Uncertainty in Artificial Intelligence*, 2007.

Z. Harchaoui, M. Douze, M. Paulin, M. Dudik, and J. Malick. Large-scale classification with trace-norm regularization. In *Proc. CVPR*, 2012.

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

J. Huang, S. Ma, and C.H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18:1603–1618, 2008.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th*

International Conference on Machine Learning (ICML), 2009.

K. Huang and S. Aviyente. Sparse representation for signal classification. In *Advances in Neural Information Processing Systems*, Vancouver, Canada, December 2006.

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.

S. Lacoste-Julien, F. Sha, and M.I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems (NIPS) 21*, 2008.

G. R. G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

N. Le Roux, M. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for strongly-convex optimization with finite training sets. Technical Report -, HAL, 2012.

H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

M. Leordeanu, M. Hebert, and R. Sukthankar. Beyond local appearance: Category recognition from pairwise interactions of simple features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, 2:90–102, 2008.

K. Lounici, A.B. Tsybakov, M. Pontil, and S.A. van de Geer. Taking advantage of sparsity in multi-task learning. In *Conference on Computational Learning Theory (COLT)*, 2009.

J. Lv and Y. Fan. A unified approach to model selection and sparse recovery using regularized least squares. *Annals of Statistics*, 37(6A):3498–3528, 2009.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)*, 2008a.

J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008b.

J. Mairal, M. Leordeanu, F. Bach, M. Hebert, and J. Ponce. Discriminative sparse image models for class-specific edge detection and image interpretation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008c.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In

International Conference on Machine Learning (ICML), 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, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *International Conference on Computer Vision (ICCV)*, 2009d.

J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (99):1–1, 2010a.

J. Mairal, R. Jenatton, G. Obozinski, and F. Bach. Network flow algorithms for structured sparsity. In *NIPS*, 2010b.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. Technical report, arXiv: 0809.2932, 2008.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

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, M.J. Wainwright, and M.I. Jordan. High-dimensional union support recovery in multivariate regression. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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.

M. Pontil, A. Argyriou, and T. Evgeniou. Multi-task feature learning. In *Advances in Neural Information Processing Systems*, 2007.

R. Raina, A. Battle, H. Lee, B. Packer, and A.Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 2007.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

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

P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar. Collaborative hierarchical sparse modeling. Technical report, 2010. Preprint arXiv:1003.0400v1.

N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, 2005.

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. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE transactions on information theory*, 55(5):2183, 2009.

L. Wasserman and K. Roeder. High dimensional variable selection. *Annals of statistics*, 37(5A):2178, 2009.

J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.

D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009a.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009b.

J. Yang, K. Yu, , and T. Huang. Supervised translation-invariant sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 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.

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society*

Series B, 69(2):143–161, 2007.

T. Zhang. Adaptive forward-backward greedy algorithm for sparse learning with linear models. *Advances in Neural Information Processing Systems*, 22, 2008a.

T. Zhang. Multi-stage convex relaxation for learning with sparse regularization. *Advances in Neural Information Processing Systems*, 22, 2008b.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *The Journal of Machine Learning Research*, 10:555–568, 2009.

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. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.