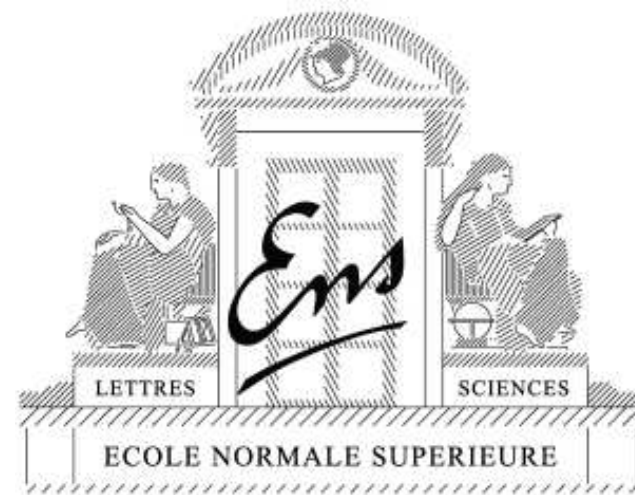


# Sparse methods for machine learning

## Theory and algorithms

**Francis Bach**

*Willow project, INRIA - Ecole Normale Supérieure*



Ecole d'été, Peyresq - June 2010

*Special thanks to R. Jenatton, J. Mairal, G. Obozinski*

# Supervised learning and regularization

- Data:  $x_i \in \mathcal{X}$ ,  $y_i \in \mathcal{Y}$ ,  $i = 1, \dots, n$
- Minimize with respect to function  $f : \mathcal{X} \rightarrow \mathcal{Y}$ :

$$\sum_{i=1}^n \ell(y_i, f(x_i)) \quad + \quad \frac{\lambda}{2} \|f\|^2$$

Error on data                      +                      Regularization

Loss & function space ?

Norm ?

- Two theoretical/algorithmic issues:
  1. Loss
  2. **Function space / norm**

# Regularizations

- **Main goal: avoid overfitting**
- **Two main lines of work:**
  1. **Euclidean** and **Hilbertian** norms (i.e.,  $\ell_2$ -norms)
    - Possibility of non linear predictors
    - Non parametric supervised learning and kernel methods
    - Well developed theory and algorithms (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)

# Regularizations

- **Main goal: avoid overfitting**

- **Two main lines of work:**

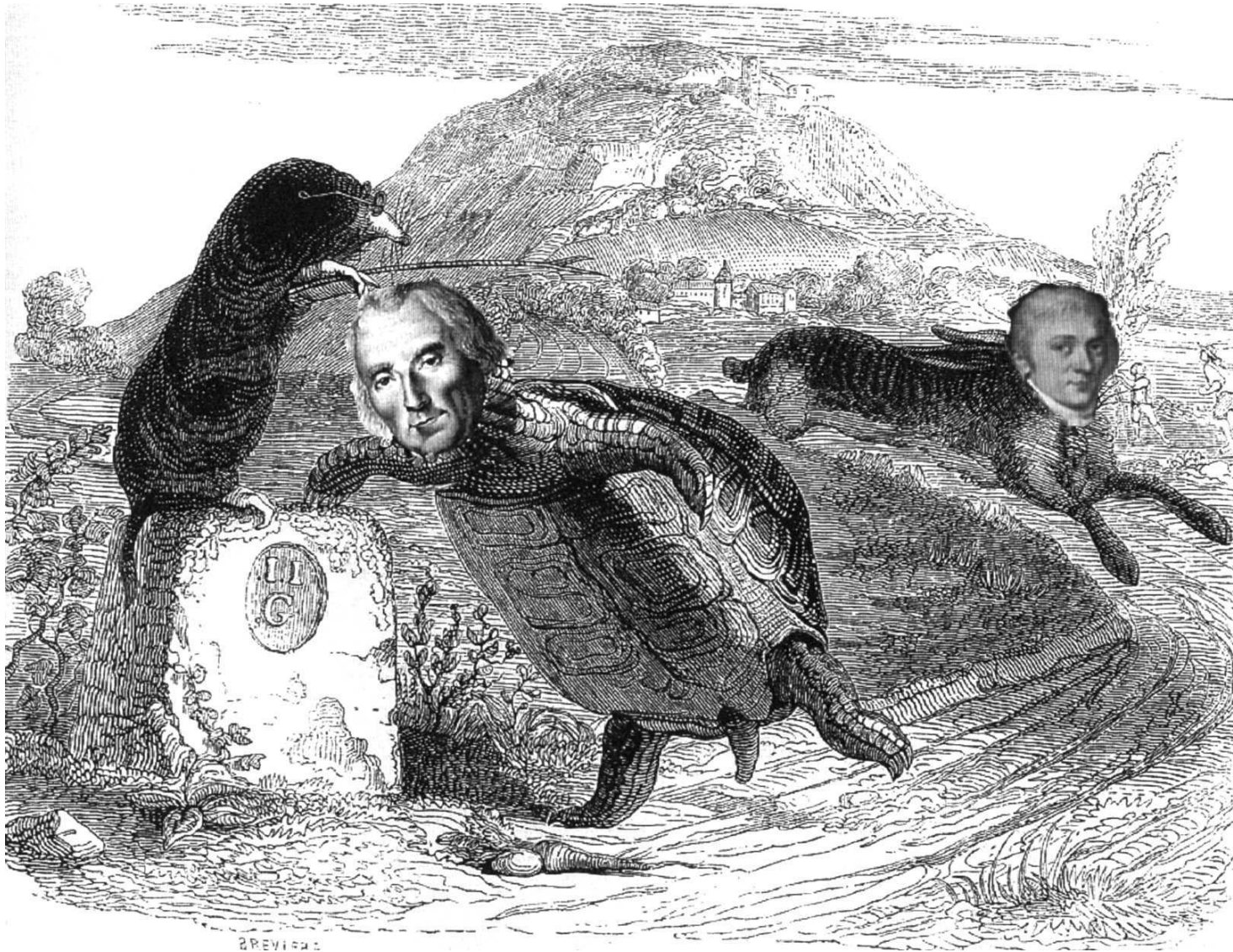
1. **Euclidean** and **Hilbertian** norms (i.e.,  $\ell_2$ -norms)

- Possibility of non linear predictors
- Non parametric supervised learning and kernel methods
- Well developed theory and algorithms (see, e.g., Wahba, 1990; Schölkopf and Smola, 2001; Shawe-Taylor and Cristianini, 2004)

2. **Sparsity-inducing** norms

- Usually restricted to linear predictors on vectors  $f(x) = w^\top x$
- Main example:  $\ell_1$ -norm  $\|w\|_1 = \sum_{i=1}^p |w_i|$
- Perform model selection as well as regularization
- **Theory and algorithms “in the making”**

# $l_2$ vs. $l_1$ - Gaussian hare vs. Laplacian tortoise



- First-order methods (Fu, 1998; Wu and Lange, 2008)
- Homotopy methods (Markowitz, 1956; Efron et al., 2004)

# 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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_{\infty} \leq 1,$$

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

# 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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_{\infty} \leq 1,$$

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

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

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

# Going beyond the Lasso

- $\ell_1$ -norm for **linear** feature selection in **high dimensions**
  - Lasso usually not applicable directly
- **Non-linearities**
- **Dealing with exponentially many features**
- **Sparse learning on matrices**



# Going beyond the Lasso

## Non-linearity - Multiple kernel learning

- **Multiple kernel learning**

- Learn sparse combination of matrices  $k(x, x') = \sum_{j=1}^p \eta_j k_j(x, x')$
- Mixing positive aspects of  $\ell_1$ -norms and  $\ell_2$ -norms

- **Equivalent to group Lasso**

- $p$  multi-dimensional features  $\Phi_j(x)$ , where

$$k_j(x, x') = \Phi_j(x)^\top \Phi_j(x')$$

- learn predictor  $\sum_{j=1}^p w_j^\top \Phi_j(x)$
- Penalization by  $\sum_{j=1}^p \|w_j\|_2$

# Going beyond the Lasso

## Structured set of features

- **Dealing with exponentially many features**
  - Can we design efficient algorithms for the case  $\log p \approx n$ ?
  - Use structure to reduce the number of allowed patterns of zeros
  - Recursivity, **hierarchies** and factorization
- **Prior information on sparsity patterns**
  - Grouped variables with overlapping groups

# Going beyond the Lasso

## Sparse methods on matrices

- **Learning problems on matrices**
  - Multi-task learning
  - Multi-category classification
  - Matrix completion
  - Image denoising
  - NMF, topic models, etc.
- **Matrix factorization**
  - Two types of sparsity (low-rank or dictionary learning)

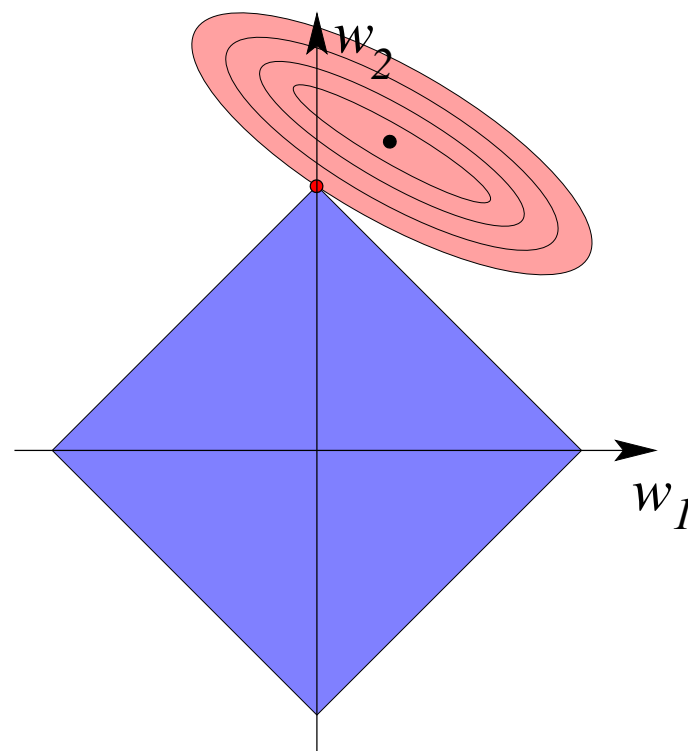
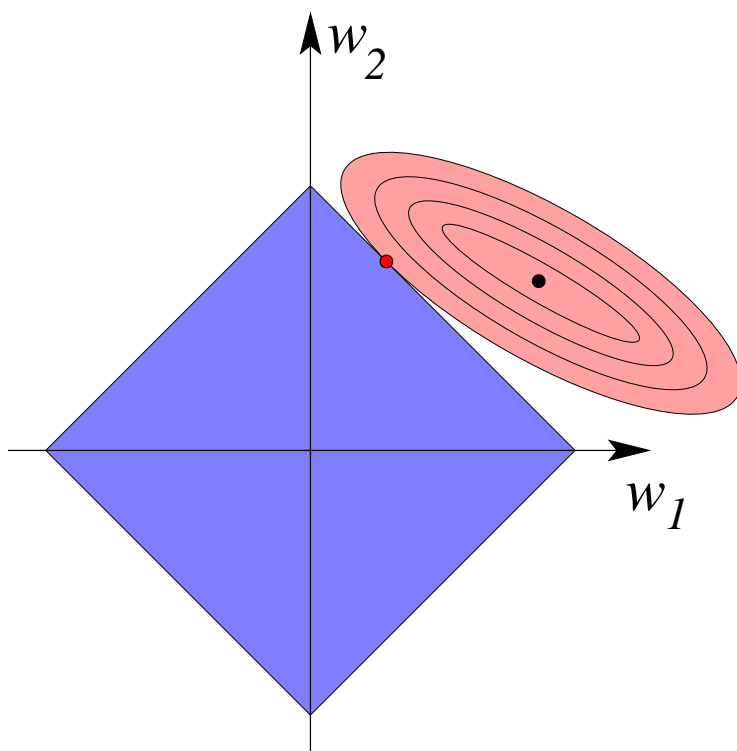
# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Why $\ell_1$ -norm constraints leads to sparsity?

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



# $\ell_1$ -norm regularization (linear setting)

- Data: covariates  $x_i \in \mathbb{R}^p$ , responses  $y_i \in \mathcal{Y}$ ,  $i = 1, \dots, n$
- Minimize with respect to loadings/weights  $w \in \mathbb{R}^p$ :

$$J(w) = \sum_{i=1}^n \ell(y_i, w^\top x_i) + \lambda \|w\|_1$$

Error on data + Regularization

- Including a constant term  $b$ ? Penalizing or constraining?
- square loss  $\Rightarrow$  basis pursuit in signal processing (Chen et al., 2001), Lasso in statistics/machine learning (Tibshirani, 1996)

# A review of nonsmooth convex analysis and optimization

- **Analysis:** optimality conditions
- **Optimization:** algorithms
  - First-order methods
- **Books:** Boyd and Vandenberghe (2004), Bonnans et al. (2003), Bertsekas (1995), Borwein and Lewis (2000)

# Optimality conditions for smooth optimization

## Zero gradient

- Example:  $\ell_2$ -regularization:  $\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top x_i) + \frac{\lambda}{2} \|w\|_2^2$ 
  - Gradient  $\nabla J(w) = \sum_{i=1}^n \ell'(y_i, w^\top x_i) x_i + \lambda w$  where  $\ell'(y_i, w^\top x_i)$  is the partial derivative of the loss w.r.t the second variable
  - If square loss,  $\sum_{i=1}^n \ell(y_i, w^\top x_i) = \frac{1}{2} \|y - Xw\|_2^2$ 
    - \* gradient =  $-X^\top (y - Xw) + \lambda w$
    - \* normal equations  $\Rightarrow w = (X^\top X + \lambda I)^{-1} X^\top y$



# Optimality conditions for smooth optimization

## Zero gradient

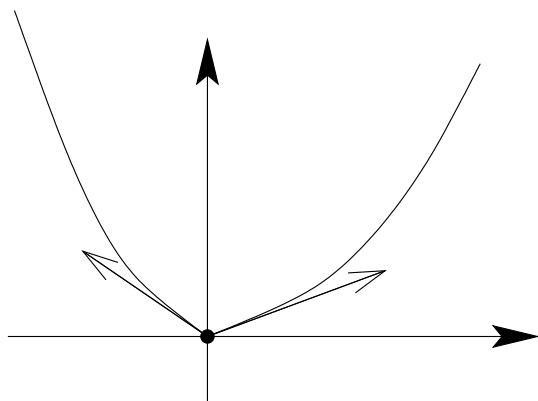
- Example:  $\ell_2$ -regularization: 
$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top x_i) + \frac{\lambda}{2} \|w\|_2^2$$
  - Gradient  $\nabla J(w) = \sum_{i=1}^n \ell'(y_i, w^\top x_i) x_i + \lambda w$  where  $\ell'(y_i, w^\top x_i)$  is the partial derivative of the loss w.r.t the second variable
  - If square loss,  $\sum_{i=1}^n \ell(y_i, w^\top x_i) = \frac{1}{2} \|y - Xw\|_2^2$ 
    - \* gradient =  $-X^\top (y - Xw) + \lambda w$
    - \* normal equations  $\Rightarrow w = (X^\top X + \lambda I)^{-1} X^\top y$
- $\ell_1$ -norm is non differentiable!
  - cannot compute the gradient of the absolute value
    - $\Rightarrow$  **Directional derivatives** (or subgradient)

# Directional derivatives - convex functions on $\mathbb{R}^p$

- **Directional derivative** in the direction  $\Delta$  at  $w$ :

$$\nabla J(w, \Delta) = \lim_{\varepsilon \rightarrow 0^+} \frac{J(w + \varepsilon \Delta) - J(w)}{\varepsilon}$$

- Always exist when  $J$  is convex and continuous
- Main idea: in non smooth situations, may need to look at all directions  $\Delta$  and not simply  $p$  independent ones



- **Proposition:**  $J$  is differentiable at  $w$ , if and only if  $\Delta \mapsto \nabla J(w, \Delta)$  is **linear**. Then,  $\nabla J(w, \Delta) = \nabla J(w)^\top \Delta$

# Optimality conditions for convex functions

- Unconstrained minimization (function defined on  $\mathbb{R}^p$ ):
  - **Proposition:**  $w$  is optimal **if and only if**  $\forall \Delta \in \mathbb{R}^p, \nabla J(w, \Delta) \geq 0$
  - Go up locally in all directions
- Reduces to zero-gradient for smooth problems
- Constrained minimization (function defined on a convex set  $K$ )
  - restrict  $\Delta$  to directions so that  $w + \varepsilon\Delta \in K$  for small  $\varepsilon$

# Directional derivatives for $\ell_1$ -norm regularization

- Function  $J(w) = \sum_{i=1}^n \ell(y_i, w^\top x_i) + \lambda \|w\|_1 = L(w) + \lambda \|w\|_1$

- $\ell_1$ -norm:  $\|w + \varepsilon \Delta\|_1 - \|w\|_1 = \sum_{j, w_j \neq 0} \{|w_j + \varepsilon \Delta_j| - |w_j|\} + \sum_{j, w_j = 0} |\varepsilon \Delta_j|$

- Thus,

$$\begin{aligned} \nabla J(w, \Delta) &= \nabla L(w)^\top \Delta + \lambda \sum_{j, w_j \neq 0} \text{sign}(w_j) \Delta_j + \lambda \sum_{j, w_j = 0} |\Delta_j| \\ &= \sum_{j, w_j \neq 0} [\nabla L(w)_j + \lambda \text{sign}(w_j)] \Delta_j + \sum_{j, w_j = 0} [\nabla L(w)_j \Delta_j + \lambda |\Delta_j|] \end{aligned}$$

- Separability of optimality conditions

# Optimality conditions for $\ell_1$ -norm regularization

- **General loss:**  $w$  optimal if and only if for all  $j \in \{1, \dots, p\}$ ,

$$\text{sign}(w_j) \neq 0 \Rightarrow \nabla L(w)_j + \lambda \text{sign}(w_j) = 0$$

$$\text{sign}(w_j) = 0 \Rightarrow |\nabla L(w)_j| \leq \lambda$$

- **Square loss:**  $w$  optimal if and only if for all  $j \in \{1, \dots, p\}$ ,

$$\text{sign}(w_j) \neq 0 \Rightarrow -X_j^\top (y - Xw) + \lambda \text{sign}(w_j) = 0$$

$$\text{sign}(w_j) = 0 \Rightarrow |X_j^\top (y - Xw)| \leq \lambda$$

- For  $J \subset \{1, \dots, p\}$ ,  $X_J \in \mathbb{R}^{n \times |J|} = X(:, J)$  denotes the columns of  $X$  indexed by  $J$ , i.e., variables indexed by  $J$

# First order methods for convex optimization on $\mathbb{R}^p$

## Smooth optimization

- **Gradient descent:**  $w_{t+1} = w_t - \alpha_t \nabla J(w_t)$ 
  - with line search: search for a decent (not necessarily best)  $\alpha_t$
  - fixed diminishing step size, e.g.,  $\alpha_t = a(t + b)^{-1}$
- Convergence of  $f(w_t)$  to  $f^* = \min_{w \in \mathbb{R}^p} f(w)$  (Nesterov, 2003)
  - $f$  convex and  $M$ -Lipschitz:  $f(w_t) - f^* = O(M/\sqrt{t})$
  - and, differentiable with  $L$ -Lipschitz gradient:  $f(w_t) - f^* = O(L/t)$
  - and,  $f$   $\mu$ -strongly convex:  $f(w_t) - f^* = O(L \exp(-4t\frac{\mu}{L}))$
- $\frac{\mu}{L} =$  condition number of the optimization problem
- Coordinate descent: similar properties
- NB: “optimal scheme”  $f(w_t) - f^* = O(L \min\{\exp(-4t\sqrt{\mu/L}), t^{-2}\})$

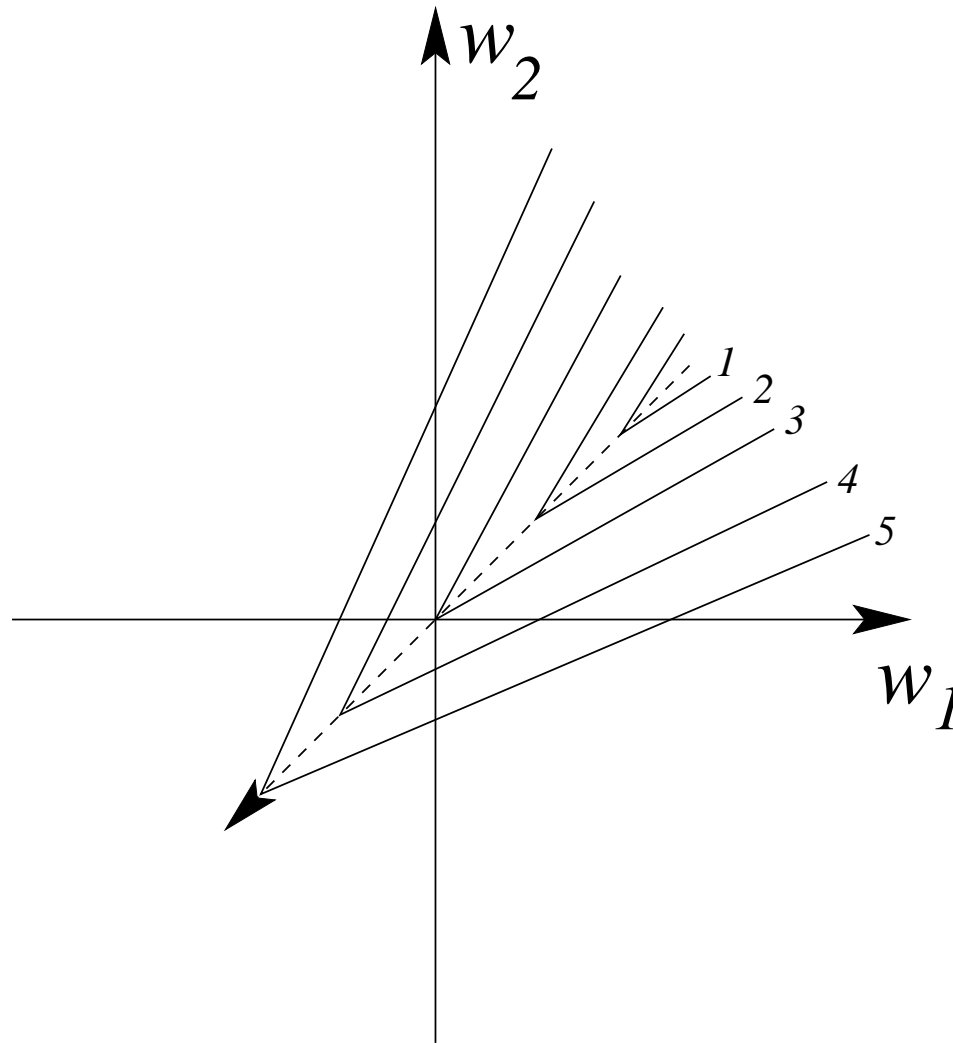
# First-order methods for convex optimization on $\mathbb{R}^p$

## Non smooth optimization

- First-order methods for **non differentiable objective**
  - Subgradient descent:  $w_{t+1} = w_t - \alpha_t g_t$ , with  $g_t \in \partial J(w_t)$ , i.e., such that  $\forall \Delta, g_t^\top \Delta \leq \nabla J(w_t, \Delta)$ 
    - \* with exact line search: not always convergent (see counter-example)
    - \* diminishing step size, e.g.,  $\alpha_t = a(t + b)^{-1}$ : convergent
  - Coordinate descent: not always convergent (show counter-example)
- Convergence rates ( $f$  convex and  $M$ -Lipschitz):  $f(w_t) - f^* = O\left(\frac{M}{\sqrt{t}}\right)$

# Counter-example

## Coordinate descent for nonsmooth objectives

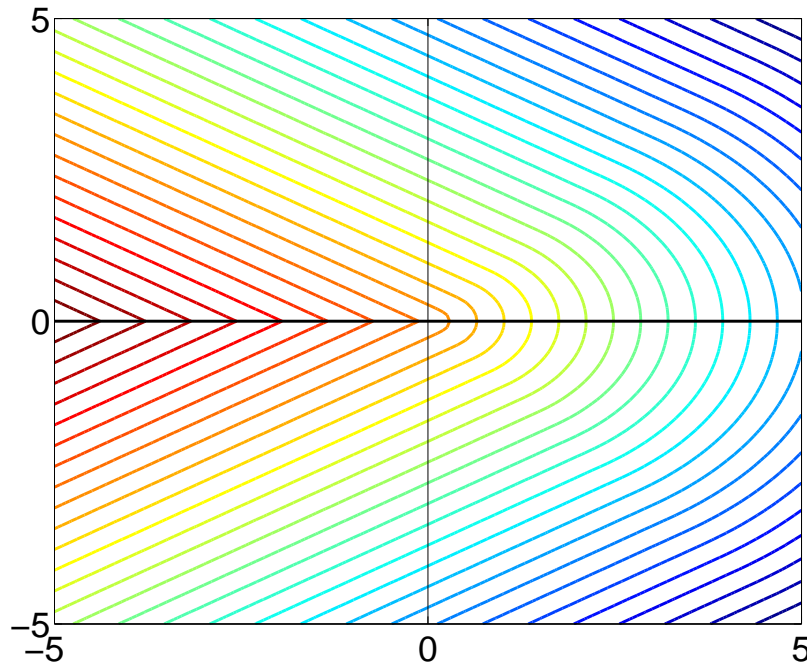




# Counter-example (Bertsekas, 1995)

## Steepest descent for nonsmooth objectives

- $q(x_1, x_2) = \begin{cases} -5(9x_1^2 + 16x_2^2)^{1/2} & \text{if } x_1 > |x_2| \\ -(9x_1 + 16|x_2|)^{1/2} & \text{if } x_1 \leq |x_2| \end{cases}$
- Steepest descent starting from any  $x$  such that  $x_1 > |x_2| > (9/16)^2|x_1|$



# Regularized problems - Proximal methods

- Gradient descent as a proximal method (differentiable functions)

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

- $w_{t+1} = w_t - \frac{1}{L} \nabla J(w_t)$

- Problems of the form:

$$\min_{w \in \mathbb{R}^p} L(w) + \lambda \Omega(w)$$

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

- Thresholded gradient descent

- 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; Wu and Lange, 2008; 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; Wu and Lange, 2008; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding
- “ $\eta$ -trick” (Rakotomamonjy et al., 2008; Jenatton et al., 2009b)
  - Notice that  $\sum_{j=1}^p |w_j| = \min_{\eta \geq 0} \frac{1}{2} \sum_{j=1}^p \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to  $\eta$  (closed-form) and  $w$  (weighted squared  $\ell_2$ -norm regularized problem)

# Cheap (and not dirty) algorithms for all losses

- **Proximal methods**
- **Coordinate descent** (Fu, 1998; Wu and Lange, 2008; Friedman et al., 2007)
  - convergent **here** under reasonable assumptions! (Bertsekas, 1995)
  - separability of optimality conditions
  - equivalent to iterative thresholding
- **“ $\eta$ -trick”** (Rakotomamonjy et al., 2008; Jenatton et al., 2009b)
  - Notice that  $\sum_{j=1}^p |w_j| = \min_{\eta \geq 0} \frac{1}{2} \sum_{j=1}^p \left\{ \frac{w_j^2}{\eta_j} + \eta_j \right\}$
  - Alternating minimization with respect to  $\eta$  (closed-form) and  $w$  (weighted squared  $\ell_2$ -norm regularized problem)
- **Dedicated algorithms that use sparsity** (active sets/homotopy)

## Special case of square loss

- **Quadratic programming formulation:** minimize

$$\frac{1}{2} \|y - Xw\|^2 + \lambda \sum_{j=1}^p (w_j^+ + w_j^-) \text{ such that } w = w^+ - w^-, w^+ \geq 0, w^- \geq 0$$

## Special case of square loss

- **Quadratic programming formulation:** minimize

$$\frac{1}{2} \|y - Xw\|^2 + \lambda \sum_{j=1}^p (w_j^+ + w_j^-) \text{ such that } w = w^+ - w^-, w^+ \geq 0, w^- \geq 0$$

- **generic toolboxes  $\Rightarrow$  very slow**
- **Main property:** if the sign pattern  $s \in \{-1, 0, 1\}^p$  of the solution is known, the solution can be obtained in closed form
  - Lasso equivalent to minimizing  $\frac{1}{2} \|y - X_J w_J\|^2 + \lambda s_J^\top w_J$  w.r.t.  $w_J$  where  $J = \{j, s_j \neq 0\}$ .
  - Closed form solution  $w_J = (X_J^\top X_J)^{-1} (X_J^\top y - \lambda s_J)$
- **Algorithm: “Guess”  $s$  and check optimality conditions**



# Optimality conditions for the sign vector $s$ (Lasso)

- For  $s \in \{-1, 0, 1\}^p$  sign vector,  $J = \{j, s_j \neq 0\}$  the nonzero pattern
- potential closed form solution:  $w_J = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$  and  $w_{J^c} = 0$
- $s$  is optimal if and only if
  - active variables:  $\text{sign}(w_J) = s_J$
  - inactive variables:  $\|X_{J^c}^\top (y - X_J w_J)\|_\infty \leq \lambda$
- **Active set algorithms** (Lee et al., 2007; Roth and Fischer, 2008)
  - Construct  $J$  iteratively by adding variables to the active set
  - Only requires to invert small linear systems

# Homotopy methods for the square loss (Markowitz, 1956; Osborne et al., 2000; Efron et al., 2004)

- **Goal:** Get **all** solutions for **all** possible values of the regularization parameter  $\lambda$
- Same idea as before: if the sign vector is known,

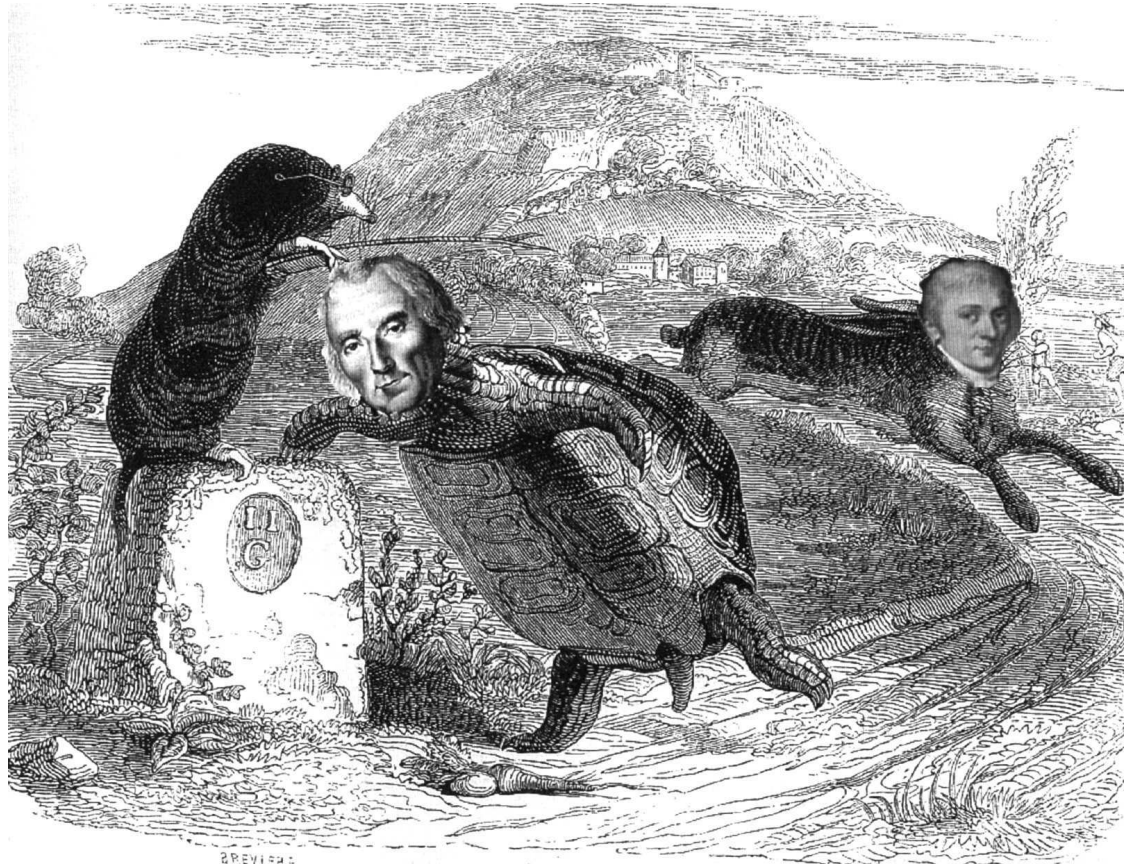
$$w_J^*(\lambda) = (X_J^\top X_J)^{-1}(X_J^\top y - \lambda s_J)$$

valid, as long as,

- sign condition:  $\text{sign}(w_J^*(\lambda)) = s_J$
  - subgradient condition:  $\|X_{J^c}^\top (X_J w_J^*(\lambda) - y)\|_\infty \leq \lambda$
  - this defines an interval on  $\lambda$ : the path is thus **piecewise affine**
- Simply need to find break points and directions



# Algorithms for $\ell_1$ -norms (square loss): Gaussian hare vs. Laplacian tortoise



- Coordinate descent:  $O(pn)$  per iterations for  $\ell_1$  and  $\ell_2$
- “Exact” algorithms:  $O(kpn)$  for  $\ell_1$  **vs.**  $O(p^2n)$  for  $\ell_2$

# Additional methods - Softwares

- Many contributions in signal processing, optimization, machine learning
  - Extensions to stochastic setting (Bottou and Bousquet, 2008)
- Extensions to other sparsity-inducing norms
  - Computing proximal operator
- **Softwares**
  - Many available codes
  - **SPAMS (SPArse Modeling Software)** - note difference with SpAM (Ravikumar et al., 2008)  
<http://www.di.ens.fr/willow/SPAMS/>

# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Theoretical results - Square loss

- Main assumption: data generated from a certain sparse  $\mathbf{w}$
- Three main problems:
  1. **Regular consistency**: convergence of **estimator**  $\hat{\mathbf{w}}$  to  $\mathbf{w}$ , i.e.,  $\|\hat{\mathbf{w}} - \mathbf{w}\|$  tends to zero when  $n$  tends to  $\infty$
  2. **Model selection consistency**: convergence of the **sparsity pattern** of  $\hat{\mathbf{w}}$  to the pattern  $\mathbf{w}$
  3. **Efficiency**: convergence of **predictions** with  $\hat{\mathbf{w}}$  to the predictions with  $\mathbf{w}$ , i.e.,  $\frac{1}{n}\|X\hat{\mathbf{w}} - X\mathbf{w}\|_2^2$  tends to zero
- Main results:
  - **Condition for model consistency (support recovery)**
  - **High-dimensional inference**

# 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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_{\infty} \leq 1$$

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



# 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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_{\infty} \leq 1$$

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

- Condition depends on  $\mathbf{w}$  and  $\mathbf{J}$  (may be relaxed)
  - may be relaxed by maximizing out  $\text{sign}(\mathbf{w})$  or  $\mathbf{J}$
- Valid in low and high-dimensional settings
- Requires lower-bound on magnitude of nonzero  $\mathbf{w}_j$

# 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

$$\|\mathbf{Q}_{\mathbf{J}^c\mathbf{J}}\mathbf{Q}_{\mathbf{J}\mathbf{J}}^{-1}\text{sign}(\mathbf{w}_{\mathbf{J}})\|_{\infty} \leq 1$$

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

- **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)

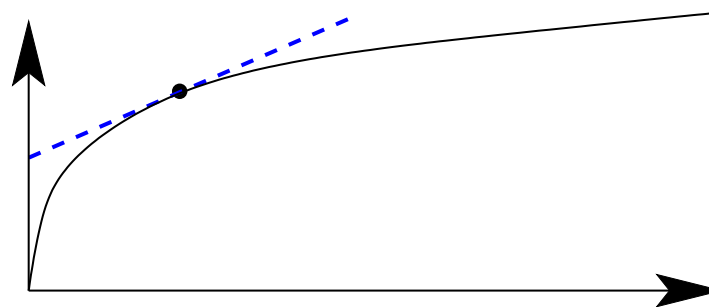
# Adaptive Lasso and concave penalization

- **Adaptive Lasso** (Zou, 2006; Huang et al., 2008)

- Weighted  $\ell_1$ -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  $\ell_2$  or  $\ell_1$  regularization

- **Reformulation in terms of concave penalization**

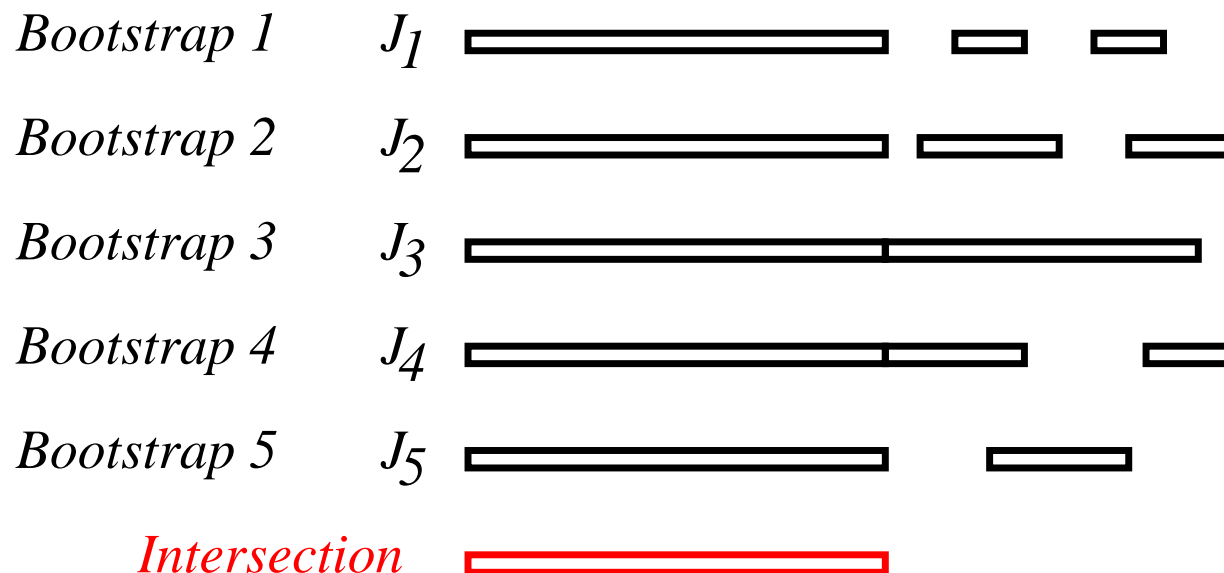
$$\min_{w \in \mathbb{R}^p} L(w) + \sum_{j=1}^p g(|w_j|)$$



- Example:  $g(|w_j|) = |w_j|^{1/2}$  or  $\log |w_j|$ . Closer to the  $\ell_0$  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)

# Bolasso (Bach, 2008a)

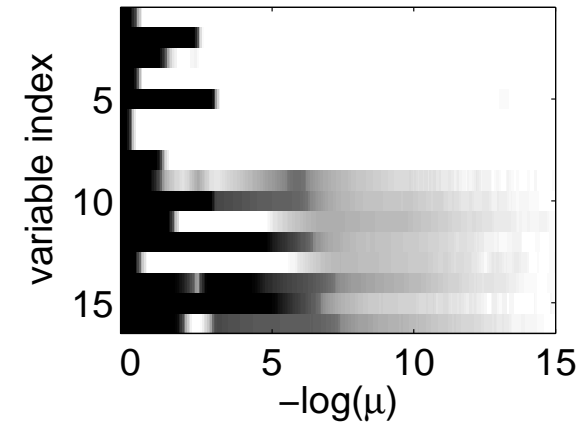
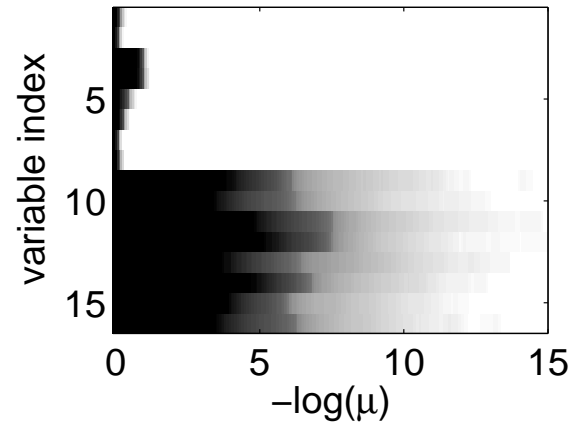
- **Property:** for a specific choice of regularization parameter  $\lambda \approx \sqrt{n}$ :
  - all variables in  $\mathbf{J}$  are always selected with high probability
  - all other ones selected with probability in  $(0, 1)$
- Use the bootstrap to simulate several replications
  - Intersecting supports of variables
  - Final estimation of  $w$  on the entire dataset



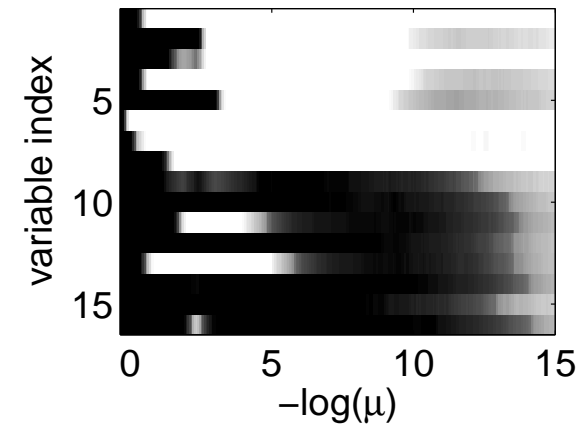
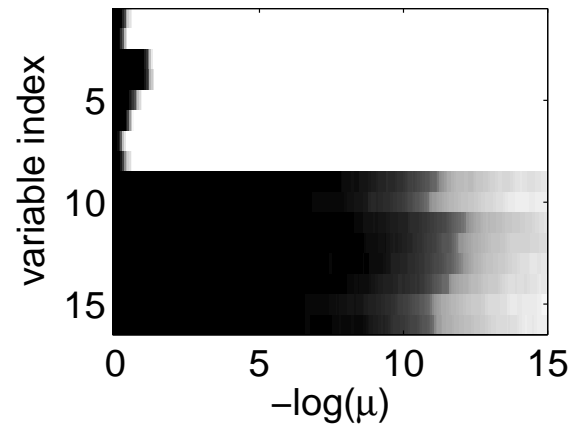
# Model selection consistency of the Lasso/Bolasso

- probabilities of selection of each variable vs. regularization param.  $\mu$

LASSO



BOLASSO



Support recovery condition **satisfied**

**not satisfied**

# High-dimensional inference

## Going beyond exact support recovery

- Theoretical results usually assume that non-zero  $w_j$  are large enough, i.e.,  $|w_j| \geq \sigma \sqrt{\frac{\log p}{n}}$
- **May include too many variables but still predict well**
- Oracle inequalities
  - Predict as well as the estimator obtained with the knowledge of  $\mathbf{J}$
  - Assume i.i.d. Gaussian noise with variance  $\sigma^2$
  - We have:

$$\frac{1}{n} \mathbb{E} \|X \hat{w}_{\text{oracle}} - X \mathbf{w}\|_2^2 = \frac{\sigma^2 |J|}{n}$$

# High-dimensional inference

## Variable selection without computational limits

- Approaches based on penalized criteria (close to BIC)

$$\min_{J \subset \{1, \dots, p\}} \left\{ \min_{w_J \in \mathbb{R}^{|J|}} \|y - X_J w_J\|_2^2 \right\} + C \sigma^2 |J| \left(1 + \log \frac{p}{|J|}\right)$$

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

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

- Gaussian noise - **No assumptions regarding correlations**
- **Scaling between dimensions:**  $\frac{k \log p}{n}$  small
- Optimal in the minimax sense

## Where does the $\log p$ come from?

- Expectation of the maximum of  $p$  Gaussian variables  $\approx \sqrt{\log p}$
- Union-bound:

$$\begin{aligned}\mathbb{P}(\exists j \in \mathbf{J}^c, |X_j^\top \varepsilon| \geq \lambda) &\leq \sum_{j \in \mathbf{J}^c} \mathbb{P}(|X_j^\top \varepsilon| \geq \lambda) \\ &\leq |\mathbf{J}^c| e^{-\frac{\lambda^2}{2n\sigma^2}} \leq p e^{-\frac{A^2}{2} \log p} = p^{1-\frac{A^2}{2}}\end{aligned}$$



# 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  $\mathbf{Q} = \frac{1}{n} \mathbf{X}^\top \mathbf{X}$ .
  - **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)
- Assume that  $\mathbf{Q}$  has unit diagonal

# Mutual incoherence (uniform low correlations)

- **Theorem** (Lounici, 2008):

- $y_i = \mathbf{w}^\top x_i + \varepsilon_i$ ,  $\varepsilon$  i.i.d. normal with mean zero and variance  $\sigma^2$
- $\mathbf{Q} = X^\top X/n$  with unit diagonal and **cross-terms less than  $\frac{1}{14k}$**
- if  $\|\mathbf{w}\|_0 \leq k$ , and  $A^2 > 8$ , then, with  $\lambda = A\sigma\sqrt{n \log p}$

$$\mathbb{P}\left(\|\hat{\mathbf{w}} - \mathbf{w}\|_\infty \leq 5A\sigma \left(\frac{\log p}{n}\right)^{1/2}\right) \geq 1 - p^{1-A^2/8}$$

- Model consistency by thresholding if  $\min_{j, \mathbf{w}_j \neq 0} |\mathbf{w}_j| > C\sigma\sqrt{\frac{\log p}{n}}$
- Mutual incoherence condition depends *strongly* on  $k$
- Improved result by averaging over sparsity patterns (Candès and Plan, 2009b)

## Where does the $\log p$ come from?

- Expectation of the maximum of  $p$  Gaussian variables  $\approx \sqrt{\log p}$
- Union-bound:

$$\begin{aligned}\mathbb{P}(\exists j \in \mathbf{J}^c, |\mathbf{X}_j^\top \varepsilon| \geq \lambda) &\leq \sum_{j \in \mathbf{J}^c} \mathbb{P}(|\mathbf{X}_j^\top \varepsilon| \geq \lambda) \\ &\leq |\mathbf{J}^c| e^{-\frac{\lambda^2}{2n\sigma^2}} \leq p e^{-\frac{A^2}{2} \log p} = p^{1-\frac{A^2}{2}}\end{aligned}$$

# Restricted eigenvalue conditions

- **Theorem** (Bickel et al., 2009):

– assume  $\kappa(k)^2 = \min_{|J| \leq k} \min_{\Delta, \|\Delta_{J^c}\|_1 \leq \|\Delta_J\|_1} \frac{\Delta^\top \mathbf{Q} \Delta}{\|\Delta_J\|_2^2} > 0$

– assume  $\lambda = A\sigma\sqrt{n \log p}$  and  $A^2 > 8$

– then, with probability  $1 - p^{1-A^2/8}$ , we have

estimation error  $\|\hat{\mathbf{w}} - \mathbf{w}\|_1 \leq \frac{16A}{\kappa^2(k)} \sigma k \sqrt{\frac{\log p}{n}}$

prediction error  $\frac{1}{n} \|X\hat{\mathbf{w}} - X\mathbf{w}\|_2^2 \leq \frac{16A^2}{\kappa^2(k)} \frac{\sigma^2 k}{n} \log p$

- Condition imposes a potentially hidden scaling between  $(n, p, k)$
- Condition always satisfied for  $\mathbf{Q} = I$

# Checking sufficient conditions

- **Most of the conditions are not computable in polynomial time**
- **Random matrices**
  - Sample  $X \in \mathbb{R}^{n \times p}$  from the Gaussian ensemble
  - Conditions satisfied with high probability for certain  $(n, p, k)$
  - Example from Wainwright (2009):  $n \geq Ck \log p$
- **Checking with convex optimization**
  - Relax conditions to convex optimization problems (d'Aspremont et al., 2008; Juditsky and Nemirovski, 2008; d'Aspremont and El Ghaoui, 2008)
  - Example: sparse eigenvalues  $\min_{|J| \leq k} \lambda_{\min}(\mathbf{Q}_{JJ})$
  - **Open problem: verifiable assumptions still lead to weaker results**

# Sparse methods

## Common extensions

- **Removing bias of the estimator**
  - Keep the active set, and perform **unregularized** restricted estimation (Candès and Tao, 2007)
  - Better theoretical bounds
  - Potential problems of robustness
- **Elastic net** (Zou and Hastie, 2005)
  - Replace  $\lambda\|w\|_1$  by  $\lambda\|w\|_1 + \varepsilon\|w\|_2^2$
  - Make the optimization strongly convex with unique solution
  - Better behavior with heavily correlated variables

# Relevance of theoretical results

- **Most results only for the square loss**
  - Extend to other losses (Van De Geer, 2008; Bach, 2009b)
- **Most results only for  $\ell_1$ -regularization**
  - May be extended to other norms (see, e.g., Huang and Zhang, 2009; Bach, 2008b)
- **Condition on correlations**
  - very restrictive, far from results for BIC penalty
- **Non sparse generating vector**
  - little work on robustness to lack of sparsity
- **Estimation of regularization parameter**
  - No satisfactory solution  $\Rightarrow$  open problem

# Alternative sparse methods

## Greedy methods

- Forward selection
- 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



# Alternative sparse methods

## Bayesian methods

- Lasso: minimize  $\sum_{i=1}^n (y_i - w^\top x_i)^2 + \lambda \|w\|_1$ 
  - Equivalent to MAP estimation with Gaussian likelihood and factorized **Laplace** prior  $p(w) \propto \prod_{j=1}^p e^{-\lambda |w_j|}$  (Seeger, 2008)
  - **However, posterior puts zero weight on exact zeros**
- Heavy-tailed distributions as a proxy to sparsity
  - Student distributions (Caron and Doucet, 2008)
  - Generalized hyperbolic priors (Archambeau and Bach, 2008)
  - Instance of automatic relevance determination (Neal, 1996)
- Mixtures of “Diracs” and another absolutely continuous distributions, e.g., “spike and slab” (Ishwaran and Rao, 2005)
- Less theory than frequentist methods

# Comparing Lasso and other strategies for linear regression

- Compared methods to reach the least-square solution

- Ridge regression:  $\min_{w \in \mathbb{R}^p} \frac{1}{2} \|y - Xw\|_2^2 + \frac{\lambda}{2} \|w\|_2^2$

- Lasso:  $\min_{w \in \mathbb{R}^p} \frac{1}{2} \|y - Xw\|_2^2 + \lambda \|w\|_1$

- Forward greedy:

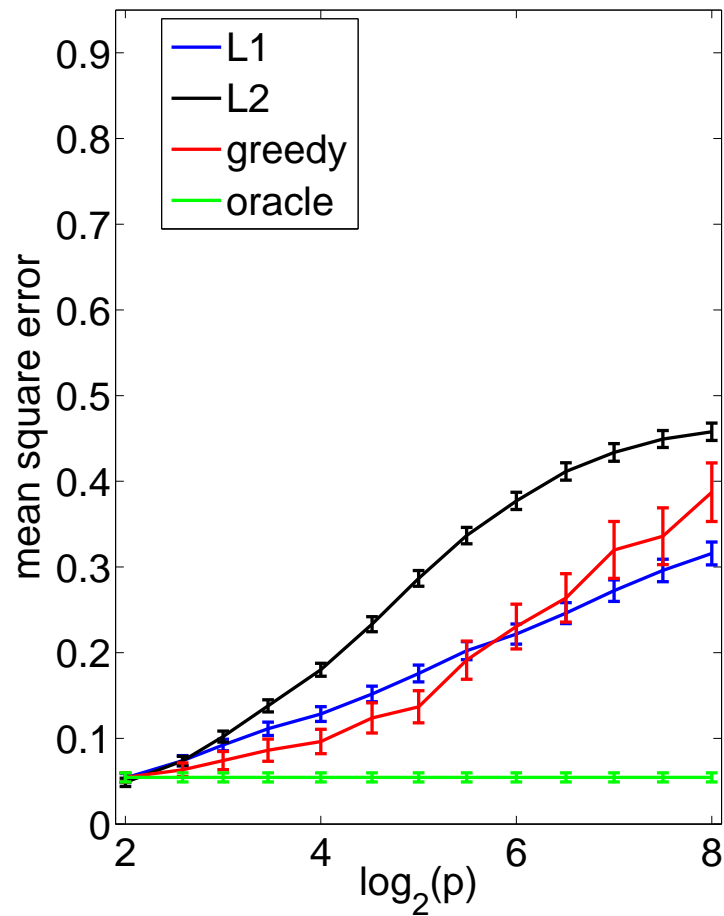
- \* Initialization with empty set

- \* Sequentially add the variable that best reduces the square loss

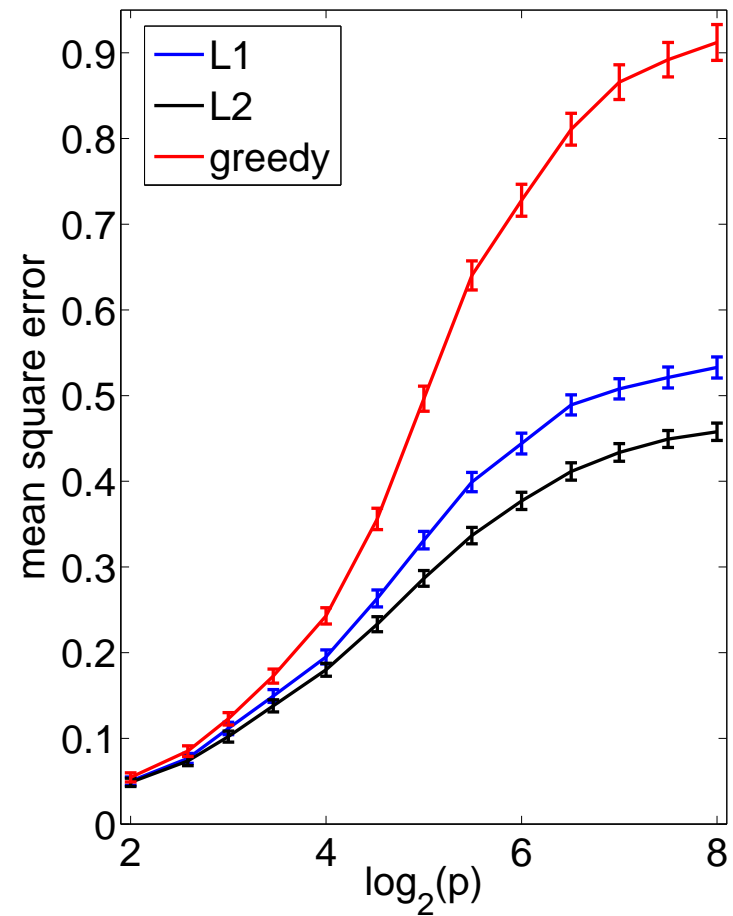
- Each method builds a path of solutions from 0 to ordinary least-squares solution
- Regularization parameters selected on the test set

# 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



Rotated (non sparse)

# Summary

## $\ell_1$ -norm regularization

- $\ell_1$ -norm regularization leads to **nonsmooth optimization problems**
  - analysis through directional derivatives or subgradients
  - optimization may or may not take advantage of sparsity
- $\ell_1$ -norm regularization allows **high-dimensional inference**
- Interesting problems for  $\ell_1$ -regularization
  - Stable variable selection
  - Weaker sufficient conditions (for weaker results)
  - Estimation of regularization parameter (all bounds depend on the unknown noise variance  $\sigma^2$ )

# Extensions

- **Sparse methods are not limited to the square loss**
  - logistic loss: algorithms (Beck and Teboulle, 2009) and theory (Van De Geer, 2008; Bach, 2009b)
- **Sparse methods are not limited to supervised learning**
  - Learning the structure of Gaussian graphical models (Meinshausen and Bühlmann, 2006; Banerjee et al., 2008)
  - Sparsity on matrices (last part of the tutorial)
- **Sparse methods are not limited to variable selection in a linear model**
  - **See next part of the tutorial**

**Questions?**

# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Penalization with grouped variables (Yuan and Lin, 2006)

- Assume that  $\{1, \dots, p\}$  is **partitioned** into  $m$  groups  $G_1, \dots, G_m$
- Penalization by  $\sum_{i=1}^m \|w_{G_i}\|_2$ , often called  $\ell_1$ - $\ell_2$  norm
- Induces group sparsity
  - Some groups entirely set to zero
  - no zeros within groups
- In this tutorial:
  - Groups may have infinite size  $\Rightarrow$  **MKL**
  - Groups may overlap  $\Rightarrow$  **structured sparsity**



# Linear vs. non-linear methods

- All methods in this tutorial are **linear in the parameters**
- By replacing  $x$  by features  $\Phi(x)$ , they can be made **non linear in the data**
- **Implicit vs. explicit features**
  - $\ell_1$ -norm: explicit features
  - $\ell_2$ -norm: representer theorem allows to consider implicit features if their dot products can be computed easily (kernel methods)

# Kernel methods: regularization by $\ell_2$ -norm

- Data:  $x_i \in \mathcal{X}$ ,  $y_i \in \mathcal{Y}$ ,  $i = 1, \dots, n$ , with **features**  $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$ 
  - Predictor  $f(x) = w^\top \Phi(x)$  linear in the features

- Optimization problem:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\lambda}{2} \|w\|_2^2$$

# Kernel methods: regularization by $\ell_2$ -norm

- Data:  $x_i \in \mathcal{X}$ ,  $y_i \in \mathcal{Y}$ ,  $i = 1, \dots, n$ , with **features**  $\Phi(x) \in \mathcal{F} = \mathbb{R}^p$ 
  - Predictor  $f(x) = w^\top \Phi(x)$  linear in the features

- Optimization problem:

$$\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\lambda}{2} \|w\|_2^2$$

- **Representer theorem** (Kimeldorf and Wahba, 1971): solution must be of the form  $w = \sum_{i=1}^n \alpha_i \Phi(x_i)$

- Equivalent to solving:

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\lambda}{2} \alpha^\top K \alpha$$

- Kernel matrix  $K_{ij} = k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

# Multiple kernel learning (MKL)

(Lanckriet et al., 2004b; Bach et al., 2004a)

- Sparse methods are linear!
- Sparsity with non-linearities
  - replace  $f(x) = \sum_{j=1}^p w_j^\top x_j$  with  $x \in \mathbb{R}^p$  and  $w_j \in \mathbb{R}$
  - by  $f(x) = \sum_{j=1}^p w_j^\top \Phi_j(x)$  with  $x \in \mathcal{X}$ ,  $\Phi_j(x) \in \mathcal{F}_j$  and  $w_j \in \mathcal{F}_j$
- Replace the  $\ell_1$ -norm  $\sum_{j=1}^p |w_j|$  by “block”  $\ell_1$ -norm  $\sum_{j=1}^p \|w_j\|_2$
- Remarks
  - Hilbert space extension of the group Lasso (Yuan and Lin, 2006)
  - Alternative sparsity-inducing norms (Ravikumar et al., 2008)

# Multiple kernel learning (MKL)

(Lanckriet et al., 2004b; Bach et al., 2004a)

- Multiple feature maps / kernels on  $x \in \mathcal{X}$ :
  - $p$  “feature maps”  $\Phi_j : \mathcal{X} \mapsto \mathcal{F}_j, j = 1, \dots, p$ .
  - Minimization with respect to  $w_1 \in \mathcal{F}_1, \dots, w_p \in \mathcal{F}_p$
  - Predictor:  $f(x) = w_1^\top \Phi_1(x) + \dots + w_p^\top \Phi_p(x)$

$$\begin{array}{ccccc}
 & & \Phi_1(x)^\top & w_1 & \\
 & \nearrow & \vdots & \vdots & \searrow \\
 x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \dots + w_p^\top \Phi_p(x) \\
 & \searrow & \vdots & \vdots & \nearrow \\
 & & \Phi_p(x)^\top & w_p & 
 \end{array}$$

- Generalized additive models (Hastie and Tibshirani, 1990)

## Regularization for multiple features

$$\begin{array}{ccc} & \Phi_1(x)^\top & w_1 \\ & \vdots & \vdots \\ x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \dots + w_p^\top \Phi_p(x) \\ & \searrow & \vdots & \vdots & \nearrow & \\ & & \Phi_p(x)^\top & w_p & & \end{array}$$

- Regularization by  $\sum_{j=1}^p \|w_j\|_2^2$  is equivalent to using  $K = \sum_{j=1}^p K_j$ 
  - Summing kernels is equivalent to concatenating feature spaces

# Regularization for multiple features

$$\begin{array}{ccc} & \Phi_1(x)^\top & w_1 \\ & \vdots & \vdots \\ x & \longrightarrow & \Phi_j(x)^\top & w_j & \longrightarrow & w_1^\top \Phi_1(x) + \dots + w_p^\top \Phi_p(x) \\ & \searrow & \vdots & \vdots & \nearrow & \\ & \Phi_p(x)^\top & w_p & & & \end{array}$$

- Regularization by  $\sum_{j=1}^p \|w_j\|_2^2$  is equivalent to using  $K = \sum_{j=1}^p K_j$
- Regularization by  $\sum_{j=1}^p \|w_j\|_2$  imposes sparsity at the group level
- **Main questions when regularizing by block  $\ell_1$ -norm:**
  1. Algorithms
  2. Analysis of sparsity inducing properties (Ravikumar et al., 2008; Bach, 2008b)
  3. Does it correspond to a specific combination of kernels?

# General kernel learning

- **Proposition** (Lanckriet et al, 2004, Bach et al., 2005, Micchelli and Pontil, 2005):

$$\begin{aligned} G(K) &= \min_{w \in \mathcal{F}} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\lambda}{2} \|w\|_2^2 \\ &= \max_{\alpha \in \mathbb{R}^n} - \sum_{i=1}^n \ell_i^*(\lambda \alpha_i) - \frac{\lambda}{2} \alpha^\top K \alpha \end{aligned}$$

is a **convex** function of the **kernel matrix**  $K$

- Theoretical learning bounds (Lanckriet et al., 2004, Srebro and Ben-David, 2006)
  - Less assumptions than sparsity-based bounds, but slower rates



# Equivalence with kernel learning (Bach et al., 2004a)

- Block  $\ell_1$ -norm problem:

$$\sum_{i=1}^n \ell(y_i, w_1^\top \Phi_1(x_i) + \cdots + w_p^\top \Phi_p(x_i)) + \frac{\lambda}{2} (\|w_1\|_2 + \cdots + \|w_p\|_2)^2$$

- **Proposition:** Block  $\ell_1$ -norm regularization is equivalent to minimizing with respect to  $\eta$  the optimal value  $G(\sum_{j=1}^p \eta_j K_j)$
- (sparse) weights  $\eta$  obtained from optimality conditions
- dual parameters  $\alpha$  optimal for  $K = \sum_{j=1}^p \eta_j K_j$ ,
- **Single optimization problem for learning both  $\eta$  and  $\alpha$**

# Proof of equivalence

$$\begin{aligned}
& \min_{w_1, \dots, w_p} \sum_{i=1}^n \ell\left(y_i, \sum_{j=1}^p w_j^\top \Phi_j(x_i)\right) + \lambda \left(\sum_{j=1}^p \|w_j\|_2\right)^2 \\
= & \min_{w_1, \dots, w_p} \min_{\sum_j \eta_j = 1} \sum_{i=1}^n \ell\left(y_i, \sum_{j=1}^p w_j^\top \Phi_j(x_i)\right) + \lambda \sum_{j=1}^p \|w_j\|_2^2 / \eta_j \\
= & \min_{\sum_j \eta_j = 1} \min_{\tilde{w}_1, \dots, \tilde{w}_p} \sum_{i=1}^n \ell\left(y_i, \sum_{j=1}^p \eta_j^{1/2} \tilde{w}_j^\top \Phi_j(x_i)\right) + \lambda \sum_{j=1}^p \|\tilde{w}_j\|_2^2 \text{ with } \tilde{w}_j = w_j \eta_j^{-1/2} \\
= & \min_{\sum_j \eta_j = 1} \min_{\tilde{w}} \sum_{i=1}^n \ell\left(y_i, \tilde{w}^\top \Psi_\eta(x_i)\right) + \lambda \|\tilde{w}\|_2^2 \text{ with } \Psi_\eta(x) = (\eta_1^{1/2} \Phi_1(x), \dots, \eta_p^{1/2} \Phi_p(x))
\end{aligned}$$

- We have:  $\Psi_\eta(x)^\top \Psi_\eta(x') = \sum_{j=1}^p \eta_j k_j(x, x')$  with  $\sum_{j=1}^p \eta_j = 1$  (and  $\eta \geq 0$ )

# Algorithms for the group Lasso / MKL

- Group Lasso
  - Block coordinate descent (Yuan and Lin, 2006)
  - Active set method (Roth and Fischer, 2008; Obozinski et al., 2009)
  - Nesterov's accelerated method (Liu et al., 2009)
- MKL
  - Dual ascent, e.g., sequential minimal optimization (Bach et al., 2004a)
  - $\eta$ -trick + cutting-planes (Sonnenburg et al., 2006)
  - $\eta$ -trick + projected gradient descent (Rakotomamonjy et al., 2008)
  - Active set (Bach, 2008c)

# Applications of multiple kernel learning

- Selection of hyperparameters for kernel methods
- Fusion from heterogeneous data sources (Lanckriet et al., 2004a)
- Two strategies for kernel combinations:
  - Uniform combination  $\Leftrightarrow \ell_2$ -norm
  - Sparse combination  $\Leftrightarrow \ell_1$ -norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - \* In particular, with few well-designed kernels
    - \* Be careful with normalization of kernels (Bach et al., 2004b)

# Applications of multiple kernel learning

- **Selection of hyperparameters for kernel methods**
- **Fusion from heterogeneous data sources** (Lanckriet et al., 2004a)
- Two strategies for kernel combinations:
  - Uniform combination  $\Leftrightarrow \ell_2$ -norm
  - Sparse combination  $\Leftrightarrow \ell_1$ -norm
  - MKL always leads to more interpretable models
  - MKL does not always lead to better predictive performance
    - \* In particular, with few well-designed kernels
    - \* Be careful with normalization of kernels (Bach et al., 2004b)
- **Sparse methods: new possibilities and new features**

# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Lasso - Two main recent theoretical results

1. **Support recovery condition**
2. **Exponentially many irrelevant variables:** under appropriate assumptions, consistency is possible as long as

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

# Lasso - Two main recent theoretical results

1. **Support recovery condition**
2. **Exponentially many irrelevant variables:** under appropriate assumptions, consistency is possible as long as

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

- Question: is it possible to build a sparse algorithm that can learn from more than  $10^{80}$  features?



# Lasso - Two main recent theoretical results

1. **Support recovery condition**
2. **Exponentially many irrelevant variables:** under appropriate assumptions, consistency is possible as long as

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

- Question: is it possible to build a sparse algorithm that can learn from more than  $10^{80}$  features?
  - **Some type of recursivity/factorization is needed!**

# Non-linear variable selection

- Given  $x = (x_1, \dots, x_q) \in \mathbb{R}^q$ , find function  $f(x_1, \dots, x_q)$  which depends only on a few variables
- Sparse generalized additive models (e.g., MKL):
  - restricted to  $f(x_1, \dots, x_q) = f_1(x_1) + \dots + f_q(x_q)$
- Cosso (Lin and Zhang, 2006):
  - restricted to  $f(x_1, \dots, x_q) = \sum_{J \subset \{1, \dots, q\}, |J| \leq 2} f_J(x_J)$

# Non-linear variable selection

- Given  $x = (x_1, \dots, x_q) \in \mathbb{R}^q$ , find function  $f(x_1, \dots, x_q)$  which depends only on a few variables
- Sparse generalized additive models (e.g., MKL):
  - restricted to  $f(x_1, \dots, x_q) = f_1(x_1) + \dots + f_q(x_q)$
- Cosso (Lin and Zhang, 2006):
  - restricted to  $f(x_1, \dots, x_q) = \sum_{J \subset \{1, \dots, q\}, |J| \leq 2} f_J(x_J)$
- Universally consistent non-linear selection requires all  $2^q$  subsets

$$f(x_1, \dots, x_q) = \sum_{J \subset \{1, \dots, q\}} f_J(x_J)$$

# Hierarchical kernel learning (Bach, 2008c)

- Many kernels can be decomposed as a sum of many “small” kernels

indexed by a certain set  $V$ :

$$k(x, x') = \sum_{v \in V} k_v(x, x')$$

- Example with  $x = (x_1, \dots, x_q) \in \mathbb{R}^q$  ( $\Rightarrow$  **non linear variable selection**)

– Gaussian/ANOVA kernels:  $p = \#(V) = 2^q$

$$\prod_{j=1}^q \left(1 + e^{-\alpha(x_j - x'_j)^2}\right) = \sum_{J \subset \{1, \dots, q\}} \prod_{j \in J} e^{-\alpha(x_j - x'_j)^2} = \sum_{J \subset \{1, \dots, q\}} e^{-\alpha \|x_J - x'_J\|_2^2}$$

– NB: decomposition is related to Cosso (Lin and Zhang, 2006)

- **Goal:** learning sparse combination  $\sum_{v \in V} \eta_v k_v(x, x')$

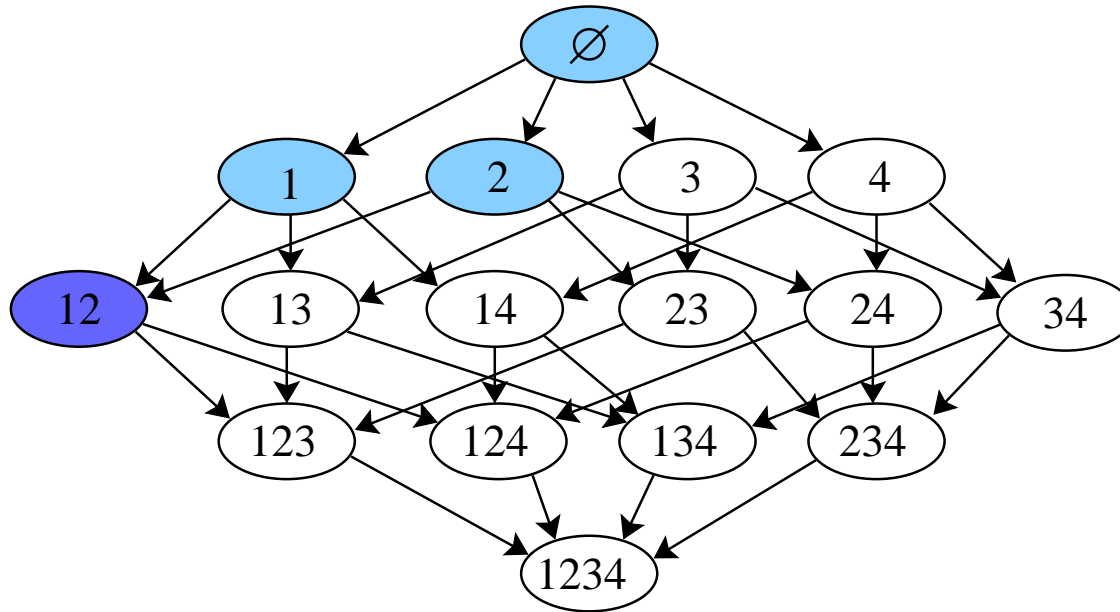
- **Universally consistent non-linear variable selection requires all subsets**

# Restricting the set of active kernels

- Assume one separate predictor  $w_v$  for each kernel  $k_v$ 
  - Final prediction:  $f(x) = \sum_{v \in V} w_v^\top \Phi_v(x)$
- With flat structure
  - Consider block  $\ell_1$ -norm:  $\sum_{v \in V} d_v \|w_v\|_2$
  - cannot avoid being linear in  $p = \#(V) = 2^q$
- Using the structure of the small kernels
  1. for computational reasons
  2. to allow more irrelevant variables

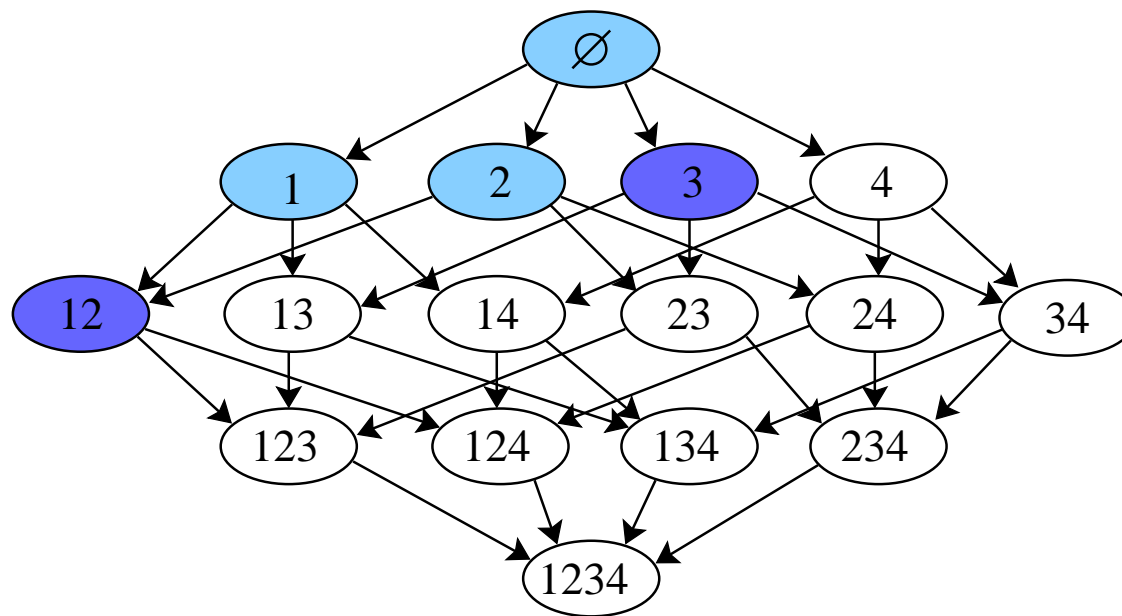
# Restricting the set of active kernels

- $V$  is endowed with a directed acyclic graph (DAG) structure:  
**select a kernel only after all of its ancestors have been selected**
- Gaussian kernels:  $V =$  power set of  $\{1, \dots, q\}$  with **inclusion** DAG
  - Select a subset only after all its subsets have been selected



# Restricting the set of active kernels

- $V$  is endowed with a directed acyclic graph (DAG) structure:  
**select a kernel only after all of its ancestors have been selected**
- Gaussian kernels:  $V =$  power set of  $\{1, \dots, q\}$  with **inclusion** DAG
  - Select a subset only after all its subsets have been selected



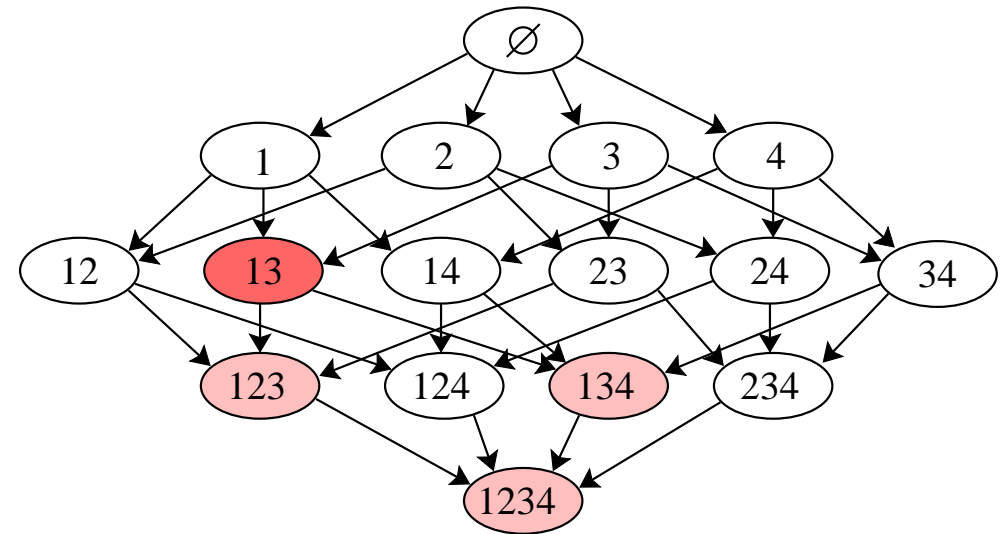
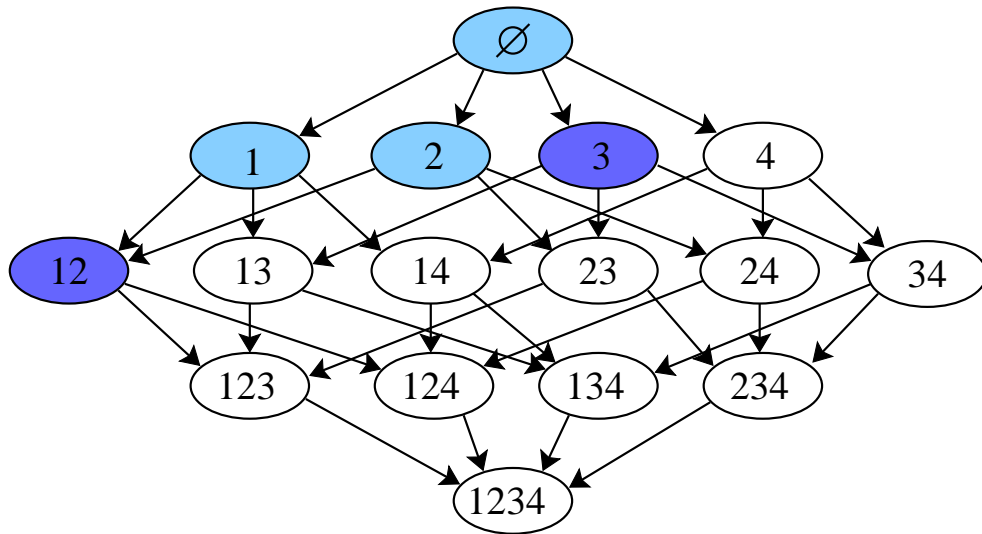
# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

–  $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} \|w_{D(v)}\|_2 = \sum_{v \in V} \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors





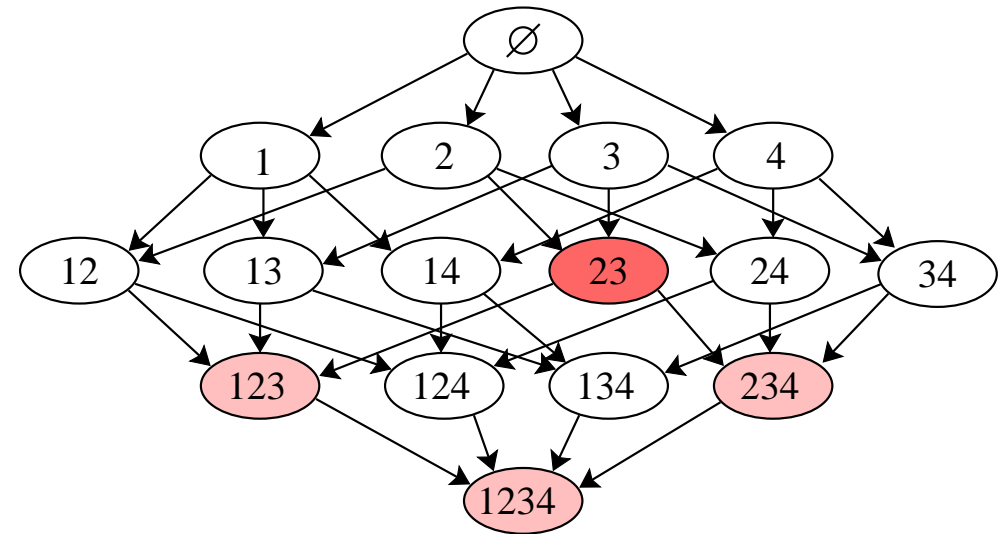
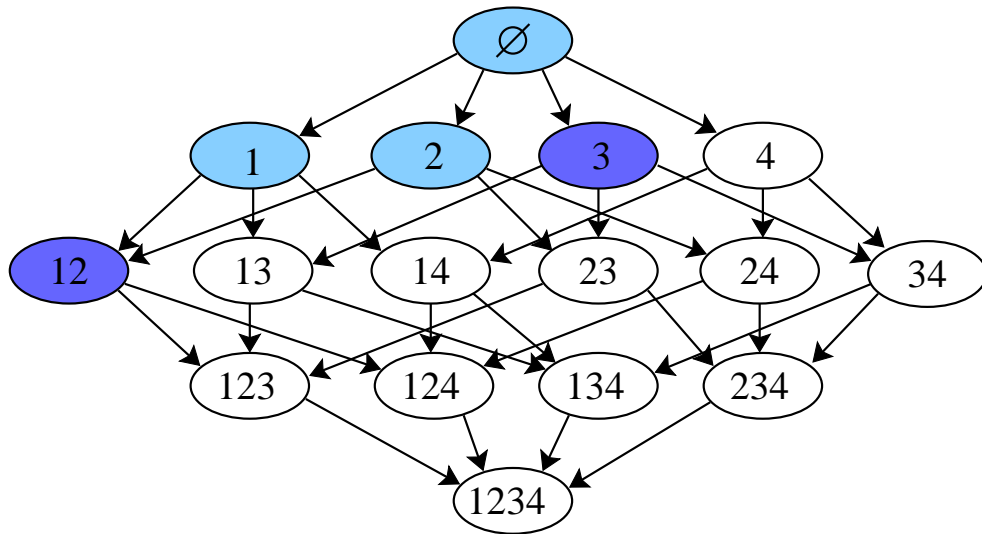
# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

–  $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} \|w_{D(v)}\|_2 = \sum_{v \in V} \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors



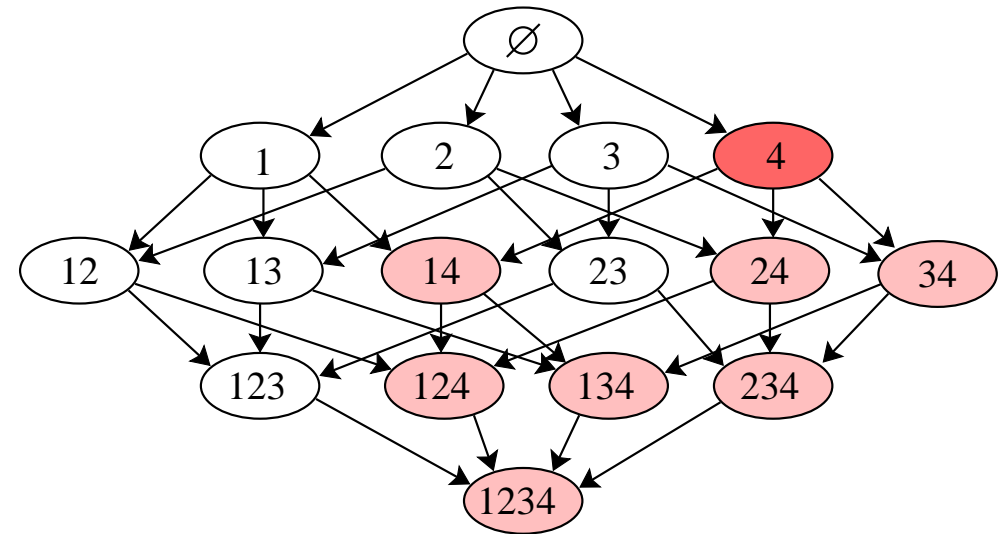
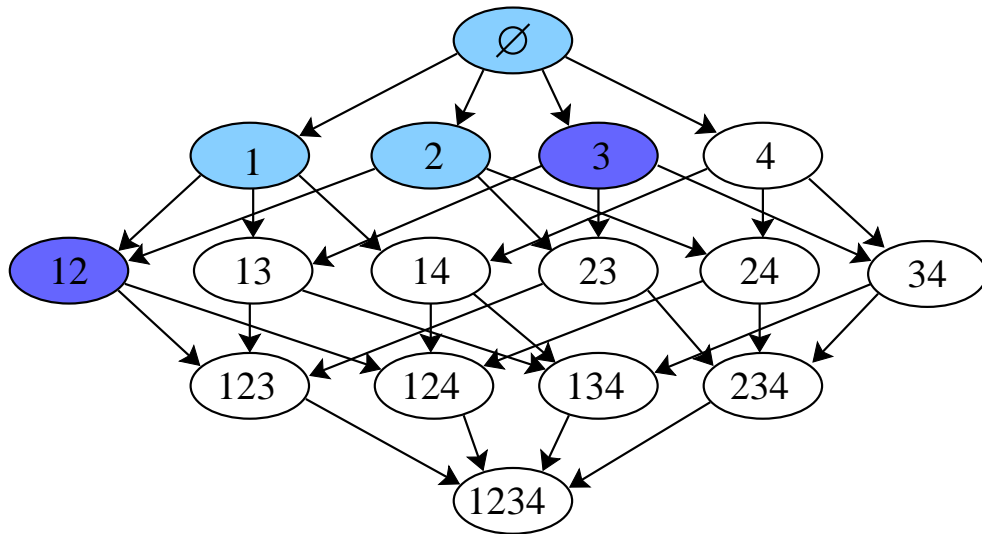
# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

–  $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} \|w_{D(v)}\|_2 = \sum_{v \in V} \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors



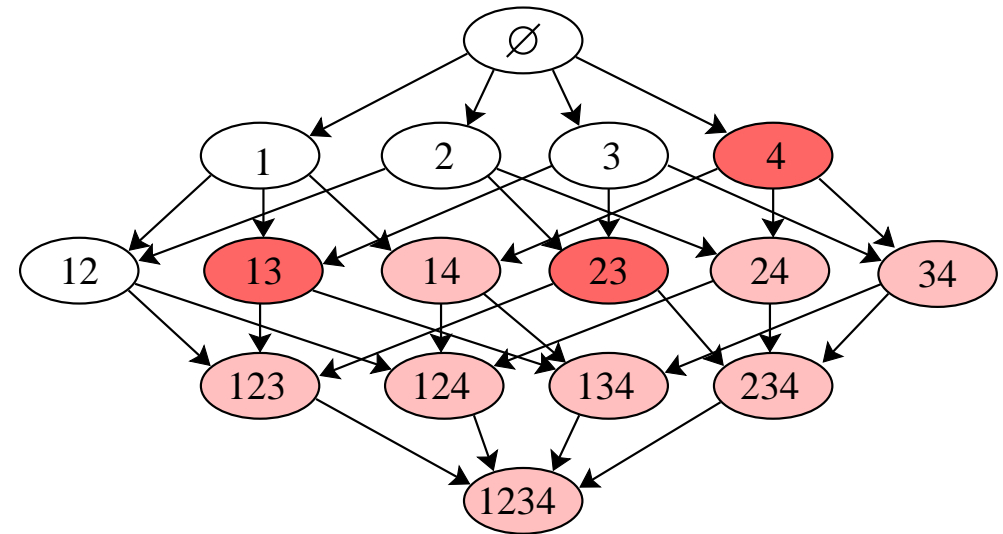
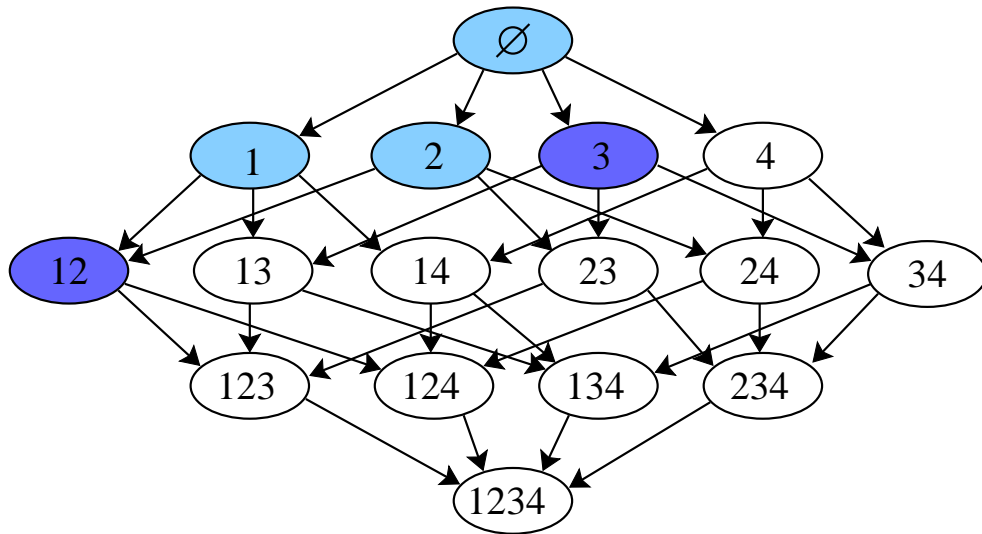
# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

–  $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} \|w_{D(v)}\|_2 = \sum_{v \in V} \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors



# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

- $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} \|w_{D(v)}\|_2 = \sum_{v \in V} \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors
- **Hierarchical kernel learning** (Bach, 2008c) :
  - **polynomial-time** algorithm for this norm
  - **necessary/sufficient conditions** for consistent kernel selection
  - **Scaling between  $p, q, n$**  for consistency
  - **Applications** to variable selection or other kernels

# DAG-adapted norm (Zhao et al., 2009; Bach, 2008c)

- Graph-based structured regularization

- $D(v)$  is the set of descendants of  $v \in V$ :

$$\sum_{v \in V} d_v \|w_{D(v)}\|_2 = \sum_{v \in V} d_v \left( \sum_{t \in D(v)} \|w_t\|_2^2 \right)^{1/2}$$

- Main property: If  $v$  is selected, so are all its ancestors

- **Hierarchical kernel learning** (Bach, 2008c) :

- **polynomial-time** algorithm for this norm
- **necessary/sufficient conditions** for consistent kernel selection
- **Scaling between p, q, n** for consistency
- **Applications** to variable selection or other kernels

## Scaling between $p$ , $n$ and other graph-related quantities

$n$  = number of observations

$p$  = number of vertices in the DAG

$\deg(V)$  = maximum out degree in the DAG

$\text{num}(V)$  = number of connected components in the DAG

- **Proposition** (Bach, 2009a): Assume consistency condition satisfied, Gaussian noise and data generated from a sparse function, then the support is recovered with high-probability as soon as:

$$\log \deg(V) + \log \text{num}(V) = O(n)$$

# Scaling between $p$ , $n$ and other graph-related quantities

$n$  = number of observations

$p$  = number of vertices in the DAG

$\deg(V)$  = maximum out degree in the DAG

$\text{num}(V)$  = number of connected components in the DAG

- **Proposition** (Bach, 2009a): Assume consistency condition satisfied, Gaussian noise and data generated from a sparse function, then the support is recovered with high-probability as soon as:

$$\log \deg(V) + \log \text{num}(V) = O(n)$$

- **Unstructured case:**  $\text{num}(V) = p \Rightarrow \boxed{\log p = O(n)}$

- **Power set of  $q$  elements:**  $\deg(V) = q \Rightarrow \boxed{\log q = \log \log p = O(n)}$

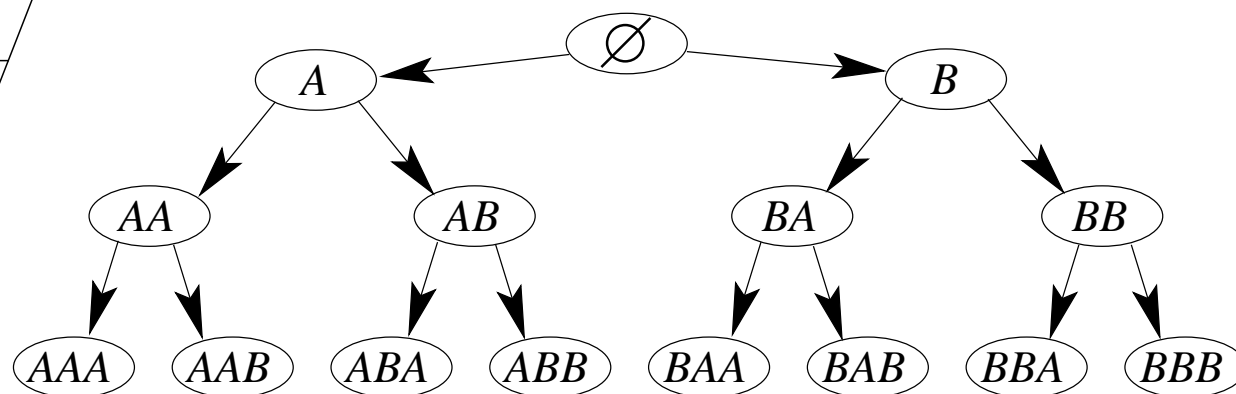
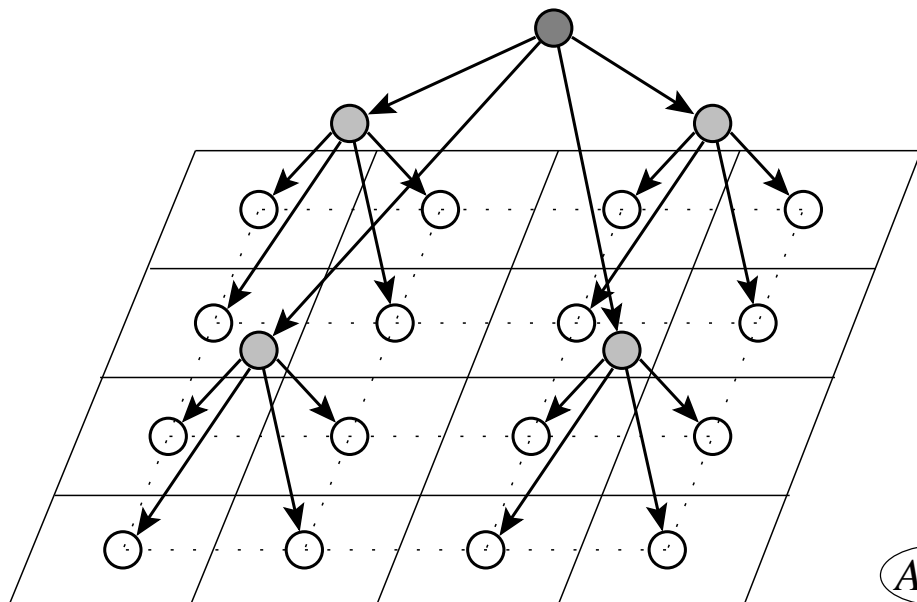
## Mean-square errors (regression)

dataset	$n$	$p$	$k$	$\#(V)$	L2	greedy	MKL	HKL
abalone	4177	10	pol4	$\approx 10^7$	44.2±1.3	43.9±1.4	44.5±1.1	<b>43.3±1.0</b>
abalone	4177	10	rbf	$\approx 10^{10}$	<b>43.0±0.9</b>	45.0±1.7	43.7±1.0	43.0±1.1
boston	506	13	pol4	$\approx 10^9$	<b>17.1±3.6</b>	24.7±10.8	22.2±2.2	18.1±3.8
boston	506	13	rbf	$\approx 10^{12}$	<b>16.4±4.0</b>	32.4±8.2	20.7±2.1	17.1±4.7
pumadyn-32fh	8192	32	pol4	$\approx 10^{22}$	57.3±0.7	56.4±0.8	<b>56.4±0.7</b>	56.4±0.8
pumadyn-32fh	8192	32	rbf	$\approx 10^{31}$	57.7±0.6	72.2±22.5	56.5±0.8	<b>55.7±0.7</b>
pumadyn-32fm	8192	32	pol4	$\approx 10^{22}$	6.9±0.1	6.4±1.6	7.0±0.1	<b>3.1±0.0</b>
pumadyn-32fm	8192	32	rbf	$\approx 10^{31}$	5.0±0.1	46.2±51.6	7.1±0.1	<b>3.4±0.0</b>
pumadyn-32nh	8192	32	pol4	$\approx 10^{22}$	84.2±1.3	73.3±25.4	83.6±1.3	<b>36.7±0.4</b>
pumadyn-32nh	8192	32	rbf	$\approx 10^{31}$	56.5±1.1	81.3±25.0	83.7±1.3	<b>35.5±0.5</b>
pumadyn-32nm	8192	32	pol4	$\approx 10^{22}$	60.1±1.9	69.9±32.8	77.5±0.9	<b>5.5±0.1</b>
pumadyn-32nm	8192	32	rbf	$\approx 10^{31}$	15.7±0.4	67.3±42.4	77.6±0.9	<b>7.2±0.1</b>



# Extensions to other kernels

- Extension to graph kernels, string kernels, pyramid match kernels



- Exploring large feature spaces with structured sparsity-inducing norms
  - Opposite view than traditional kernel methods
  - Interpretable models
- Other structures than hierarchies or DAGs

# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Grouped variables

- Supervised learning with known groups:

- The  $\ell_1$ - $\ell_2$  norm

$$\sum_{G \in \mathcal{G}} \|w_G\|_2 = \sum_{G \in \mathcal{G}} \left( \sum_{j \in G} w_j^2 \right)^{1/2}, \text{ with } \mathbf{G} \text{ a partition of } \{1, \dots, p\}$$

- The  $\ell_1$ - $\ell_2$  norm sets to zero **non-overlapping groups of variables** (as opposed to single variables for the  $\ell_1$  norm)

# Grouped variables

- Supervised learning with known groups:

- The  $\ell_1$ - $\ell_2$  norm

$$\sum_{G \in \mathbf{G}} \|w_G\|_2 = \sum_{G \in \mathbf{G}} \left( \sum_{j \in G} w_j^2 \right)^{1/2}, \text{ with } \mathbf{G} \text{ a partition of } \{1, \dots, p\}$$

- The  $\ell_1$ - $\ell_2$  norm sets to zero **non-overlapping groups of variables** (as opposed to single variables for the  $\ell_1$  norm).
- However, the  $\ell_1$ - $\ell_2$  norm encodes **fixed/static prior information**, requires to know in advance how to group the variables
- What happens if the set of groups  $\mathbf{G}$  is not a partition anymore?

# Structured Sparsity (Jenatton et al., 2009a)

- When penalizing by the  $\ell_1$ - $\ell_2$  norm

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

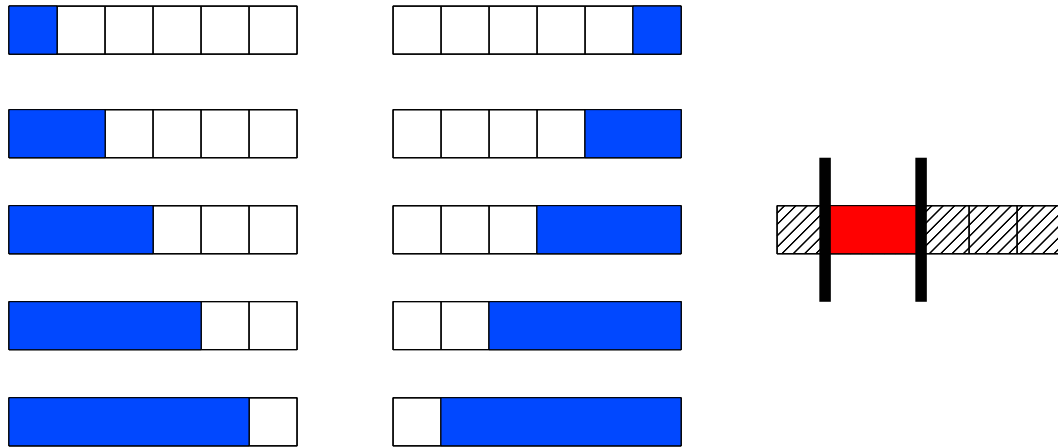
- The  $\ell_1$  norm induces sparsity at the group level:
    - \* Some  $w_G$ 's are set to zero
  - Inside the groups, the  $\ell_2$  norm does not promote sparsity
- Intuitively, the zero pattern of  $w$  is given by

$$\{j \in \{1, \dots, p\}; w_j = 0\} = \bigcup_{G \in \mathbf{G}'} G \text{ for some } \mathbf{G}' \subseteq \mathbf{G}.$$

- This intuition is actually true and can be formalized

# Examples of set of groups $G$ (1/3)

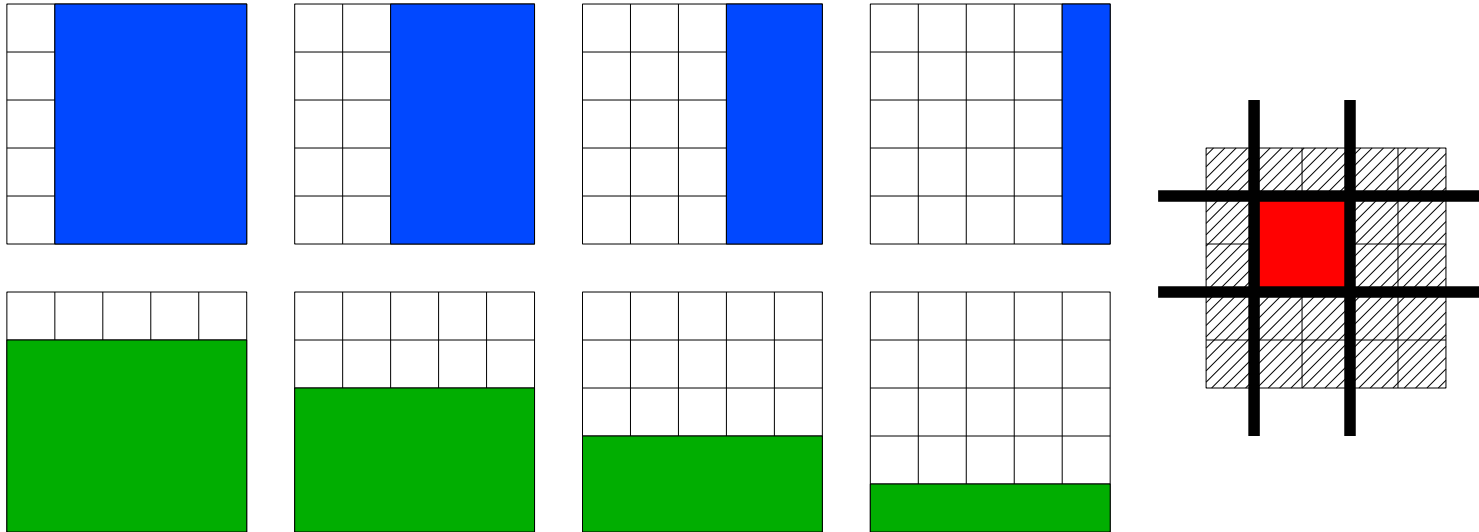
- Selection of contiguous patterns on a sequence,  $p = 6$



- $G$  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 $G$ (2/3)

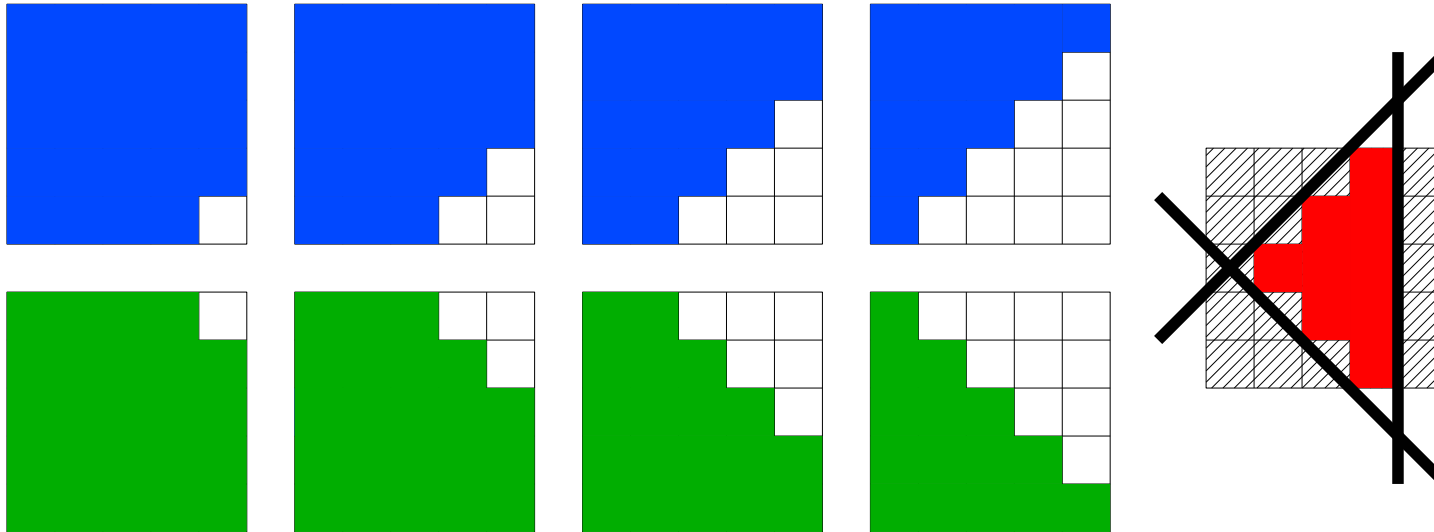
- Selection of rectangles on a 2-D grids,  $p = 25$



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

## Examples of set of groups $G$ (3/3)

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



- It is possible to extent such settings to 3-D space, or more complex topologies
- **See applications later (sparse PCA)**



# Relationship between $\mathbb{G}$ and Zero Patterns (Jenatton, Audibert, and Bach, 2009a)

- $\mathbb{G} \rightarrow$  Zero patterns:
  - by generating the **union-closure** of  $\mathbb{G}$
- Zero patterns  $\rightarrow \mathbb{G}$ :
  - Design groups  $\mathbb{G}$  from any **union-closed set** of zero patterns
  - Design groups  $\mathbb{G}$  from any **intersection-closed set** of **non-zero** patterns

# Overview of other work on structured sparsity

- Specific hierarchical structure (Zhao et al., 2009; Bach, 2008c)
- **Union-closed** (as opposed to intersection-closed) family of nonzero patterns (Jacob et al., 2009; Baraniuk et al., 2008)
- Nonconvex penalties based on information-theoretic criteria with greedy optimization (Huang et al., 2009)

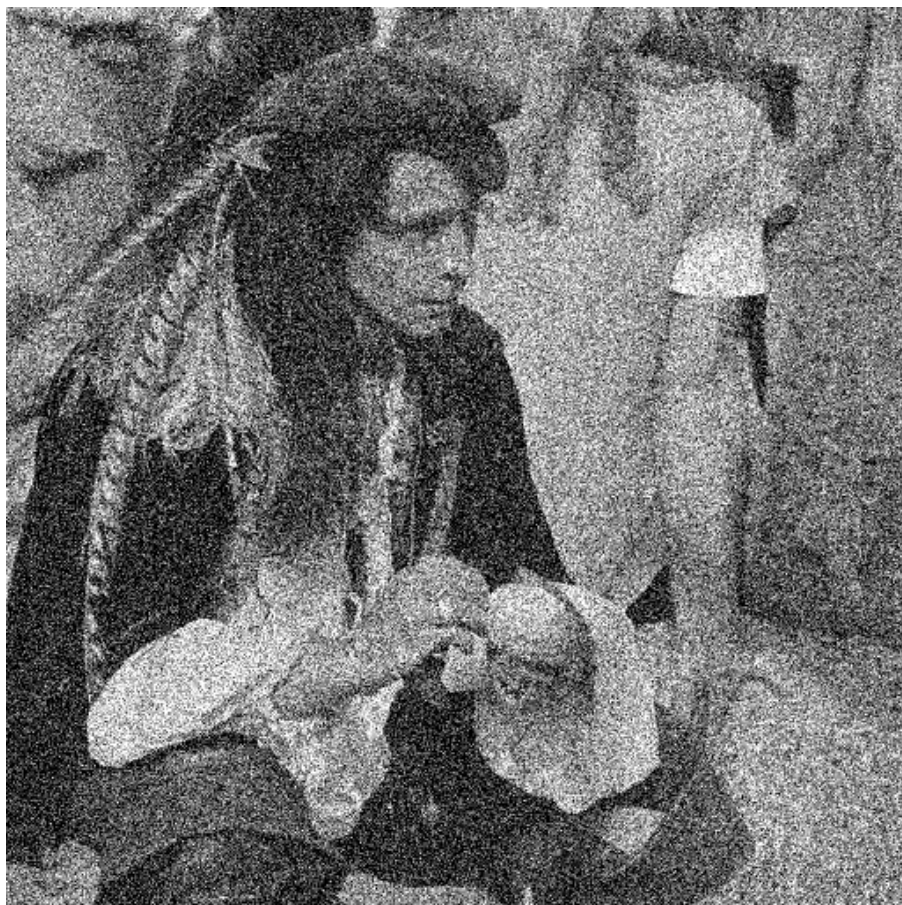
# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

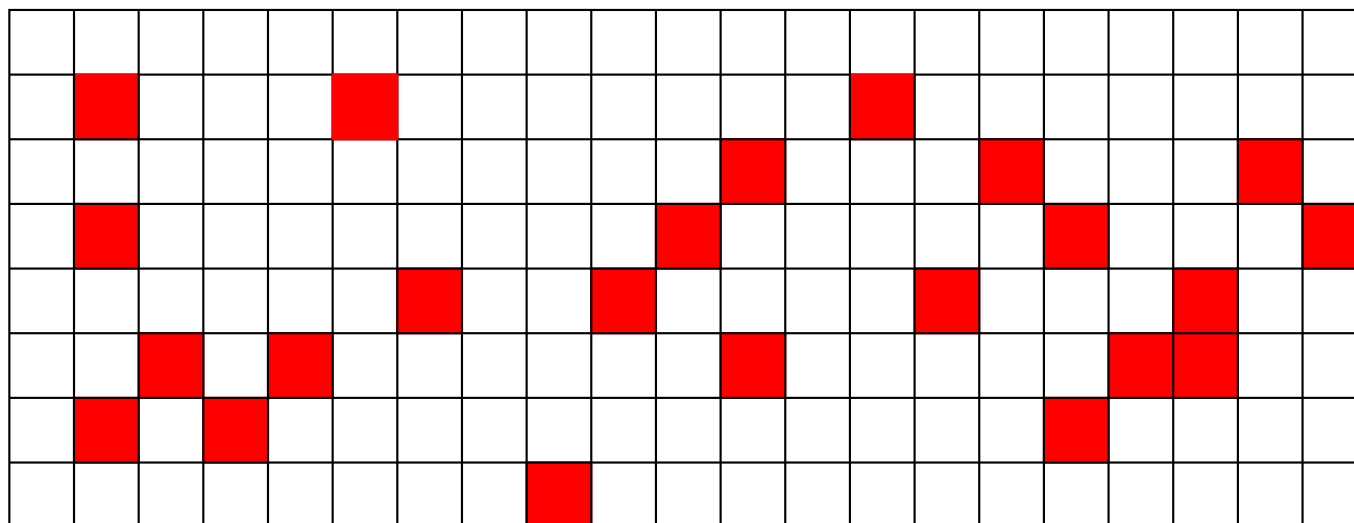
## Learning on matrices - Image denoising

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



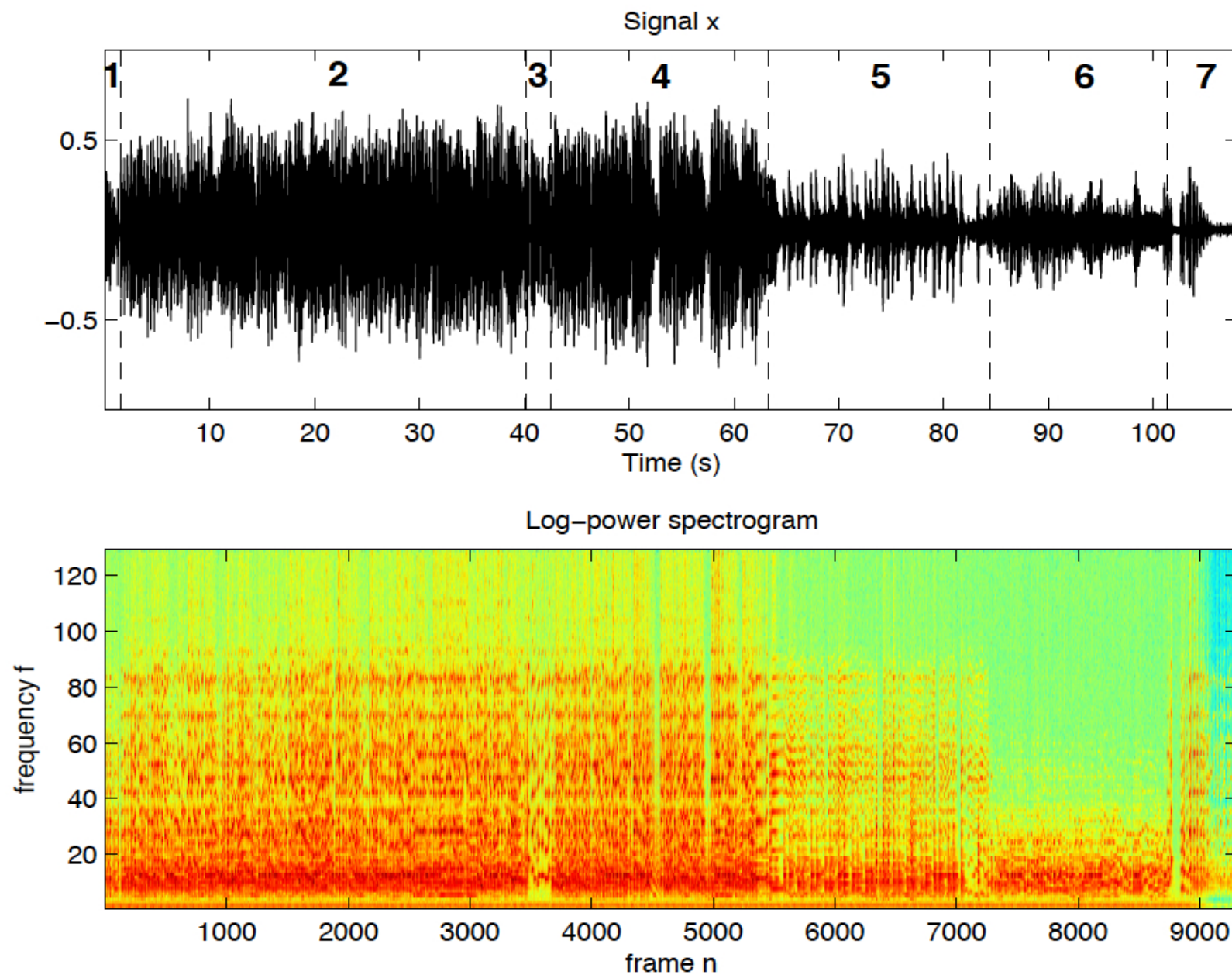
# Learning on matrices - Collaborative filtering

- Given  $n_x$  “movies”  $\mathbf{x} \in \mathcal{X}$  and  $n_y$  “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  $n_x \times n_y$  incomplete matrix  $\mathbf{Z}$  that describes the known ratings of some customers for some movies
- **Goal:** complete the matrix.



# Learning on matrices - Source separation

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



# Learning on matrices - Multi-task learning

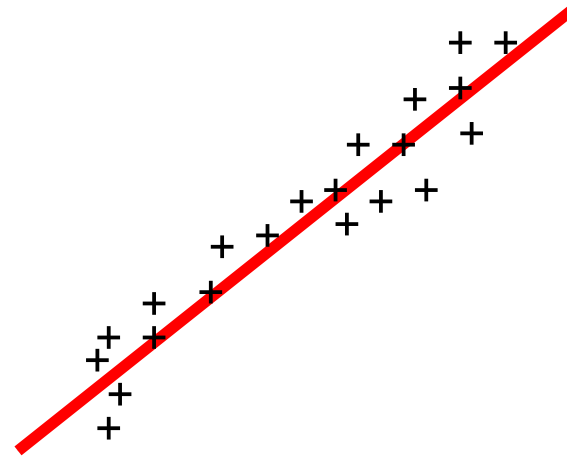
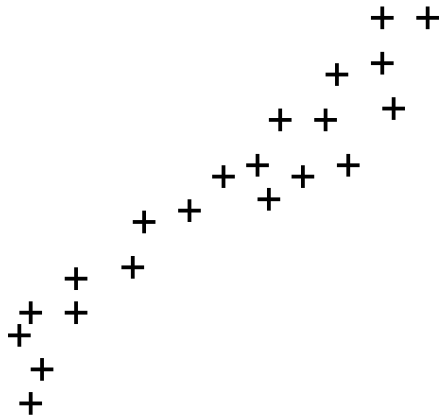
- $k$  linear prediction tasks on same covariates  $\mathbf{x} \in \mathbb{R}^p$ 
  - $k$  weight vectors  $\mathbf{w}_j \in \mathbb{R}^p$
  - Joint matrix of predictors  $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k) \in \mathbb{R}^{p \times k}$
- Classical application
  - Multi-category classification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

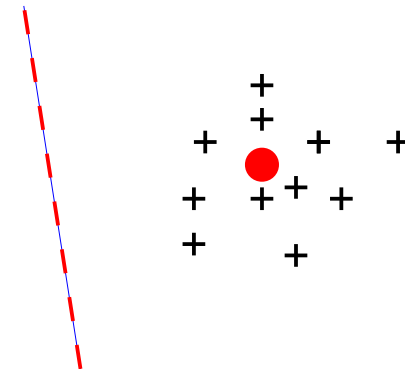
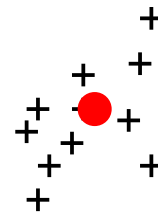
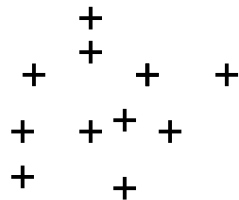
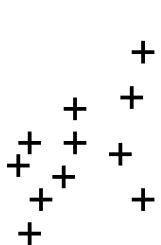
– Principal component analysis:

$$\mathbf{x}_i \approx \mathbf{D}\alpha_i \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}$$



– K-means:

$$\mathbf{x}_i \approx \mathbf{d}_k \Rightarrow \mathbf{X} = \mathbf{D}\mathbf{A}$$

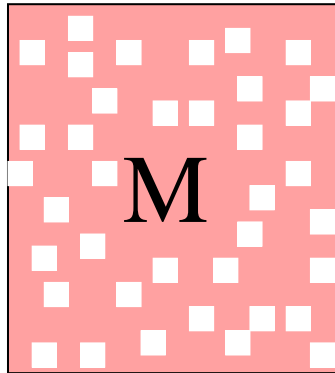




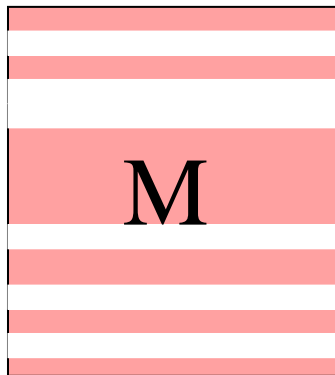
# Two types of sparsity for matrices $M \in \mathbb{R}^{n \times p}$

## I - Directly on the elements of $M$

- Many zero elements:  $M_{ij} = 0$



- Many zero rows (or columns):  $(M_{i1}, \dots, M_{ip}) = 0$

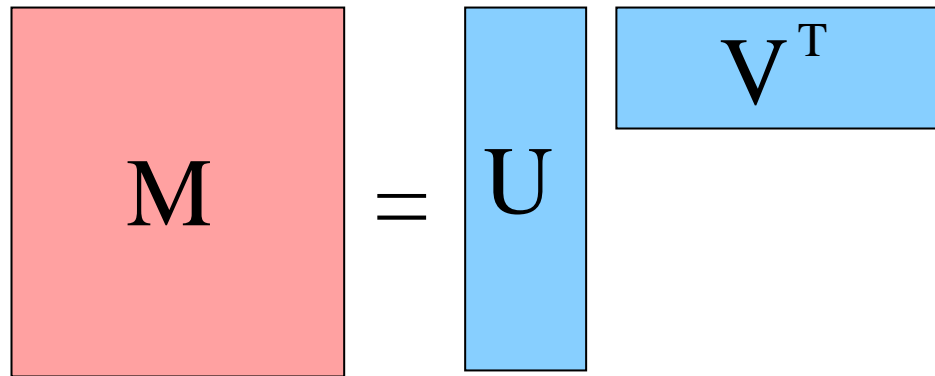


# Two types of sparsity for matrices $M \in \mathbb{R}^{n \times p}$

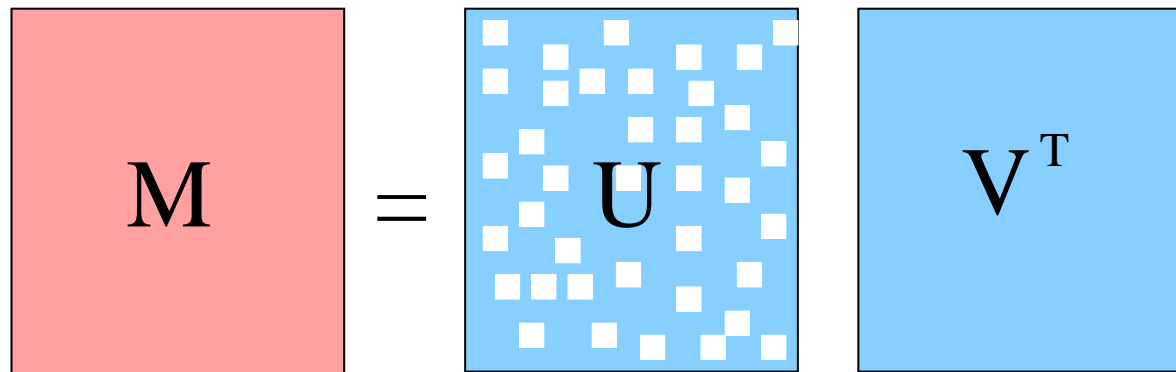
## II - Through a factorization of $M = UV^T$

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

- Low rank:  $m$  small



- Sparse decomposition:  $U$  sparse



# Structured sparse matrix factorizations

- Matrix  $\mathbf{M} = \mathbf{UV}^\top$ ,  $\mathbf{U} \in \mathbb{R}^{n \times k}$  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  $W = (w_1, \dots, w_k) \in \mathbb{R}^{p \times k}$
- **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

# 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:  $\mathbf{M} = \mathbf{U} \text{Diag}(\mathbf{s}) \mathbf{V}^\top$  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:  $\mathbf{M} = \mathbf{U} \text{Diag}(\mathbf{s}) \mathbf{V}^\top$  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)

## Results for the trace norm

- Rank recovery condition (Bach, 2008d)
  - The Hessian of the loss around the asymptotic solution should be close to diagonal
- Sufficient condition for exact rank minimization (Recht et al., 2009)
- High-dimensional inference for noisy matrix completion (Srebro et al., 2005; Candès and Plan, 2009a)
  - May recover entire matrix from slightly more entries than the minimum of the two dimensions
- **Efficient algorithms:**
  - First-order methods based on the singular value decomposition (see, e.g., Mazumder et al., 2009)
  - Low-rank formulations (Rennie and Srebro, 2005; Abernethy et al., 2009)



# Spectral regularizations

- Extensions to any functions of singular values
- Extensions to **bilinear forms** (Abernethy et al., 2009)

$$(\mathbf{x}, \mathbf{y}) \mapsto \Phi(\mathbf{x})^\top B \Psi(\mathbf{y})$$

on features  $\Phi(\mathbf{x}) \in \mathbb{R}^{f_x}$  and  $\Psi(\mathbf{y}) \in \mathbb{R}^{f_y}$ , and  $B \in \mathbb{R}^{f_x \times f_y}$

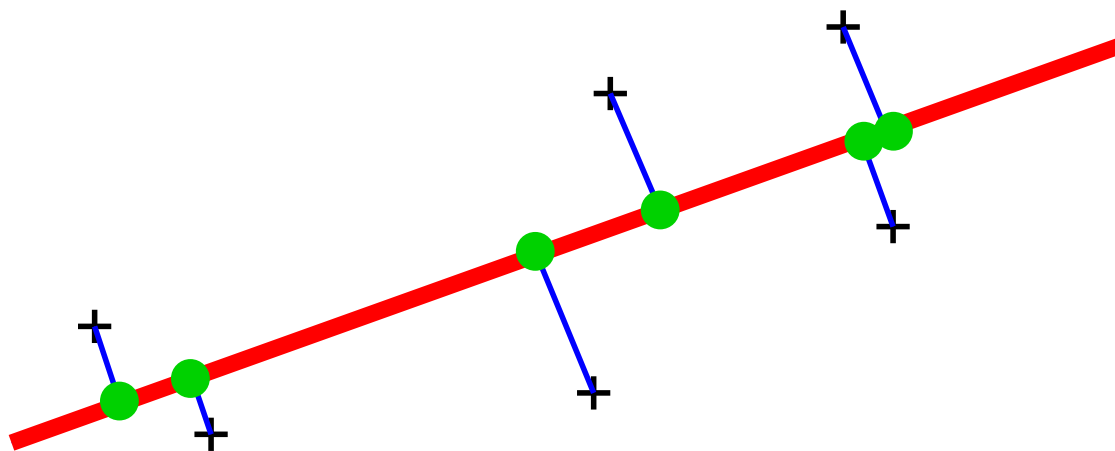
- Collaborative filtering with attributes
- **Representer theorem**: the solution must be of the form

$$B = \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} \alpha_{ij} \Psi(\mathbf{x}_i) \Phi(\mathbf{y}_j)^\top$$

- Only norms invariant by orthogonal transforms (Argyriou et al., 2009)

# 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 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 different
    - \* For analysis view, see d'Aspremont, Bach, and El Ghaoui (2008)

# 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 principal component analysis

## Synthesis view

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

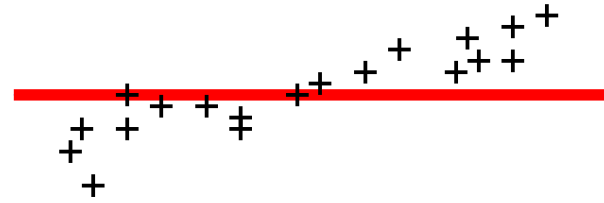
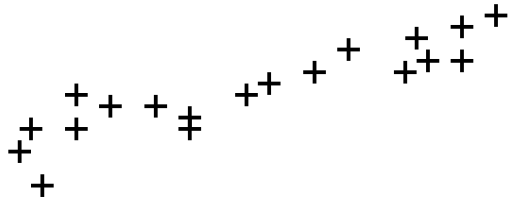
$$\sum_{i=1}^n \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \sum_{j=1}^k (\boldsymbol{\alpha}_i)_j \mathbf{d}_j \right\|_2^2 = \sum_{i=1}^n \min_{\boldsymbol{\alpha}_i \in \mathbb{R}^m} \left\| \mathbf{x}_i - \mathbf{D} \boldsymbol{\alpha}_i \right\|_2^2 \text{ is small}$$

- 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  $\ell_1$ -norm for sparsity
  - Penalize/constrain  $\boldsymbol{\alpha}_i$  by the  $\ell_2$ -norm to avoid trivial solutions

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

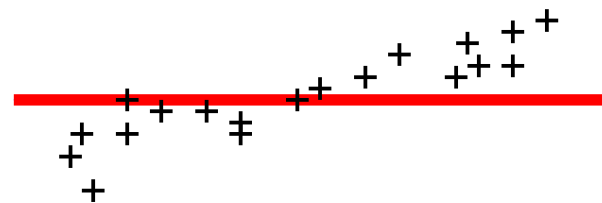
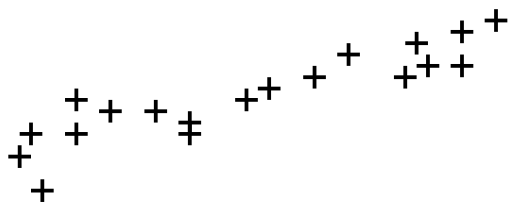
# Sparse PCA vs. dictionary learning

- Sparse PCA:  $\mathbf{x}_i \approx \mathbf{D}\alpha_i$ ,  $\mathbf{D}$  sparse

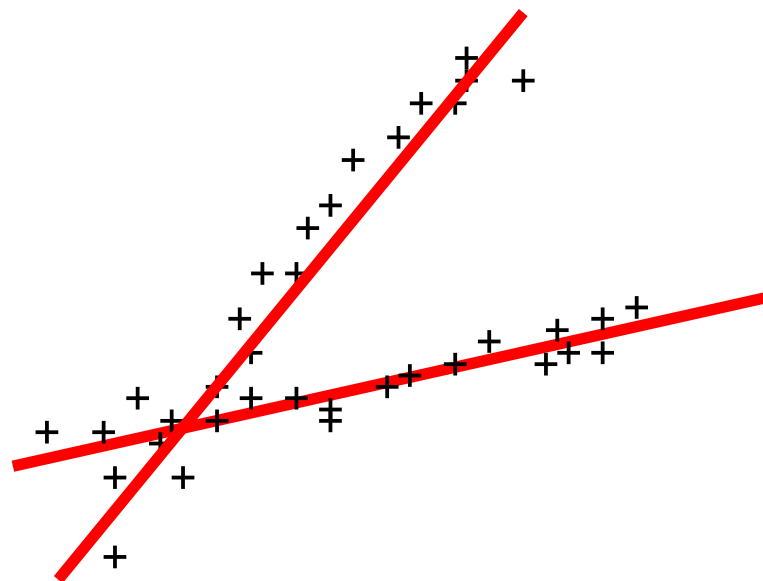
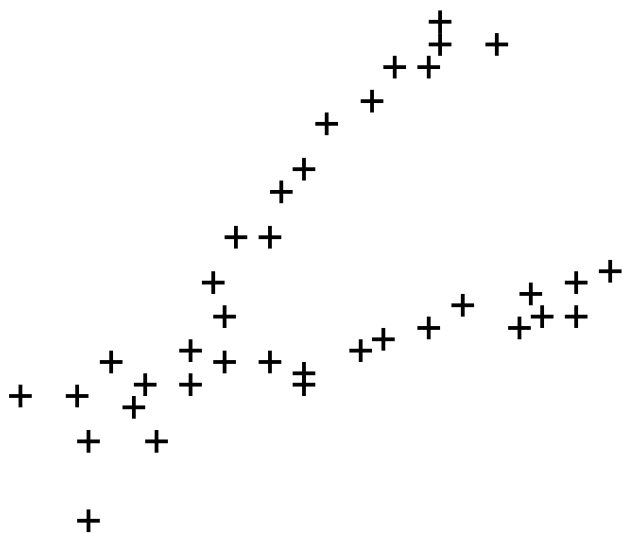


# Sparse PCA vs. dictionary learning

- Sparse PCA:  $\mathbf{x}_i \approx \mathbf{D}\alpha_i$ ,  $\mathbf{D}$  sparse



- Dictionary learning:  $\mathbf{x}_i \approx \mathbf{D}\alpha_i$ ,  $\alpha_i$  sparse



# Structured matrix factorizations (Bach et al., 2008)

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

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

- 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 ( $\ell_1$ -norm on  $\mathbf{d}_j$ )
  - **Dictionary learning** = sparse decompositions ( $\ell_1$ -norm on  $\boldsymbol{\alpha}_i$ )  
(Olshausen and Field, 1997; Elad and Aharon, 2006; Lee et al., 2007)



# Dictionary learning for image denoising



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

# Dictionary learning for image denoising

- **Solving the denoising problem** (Elad and Aharon, 2006)

- Extract all overlapping  $8 \times 8$  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

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

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

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

# Online optimization for dictionary learning

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

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

- 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., 2009c)
  - Links with NL-means (Buades et al., 2008)

## Denoising result

(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009c)

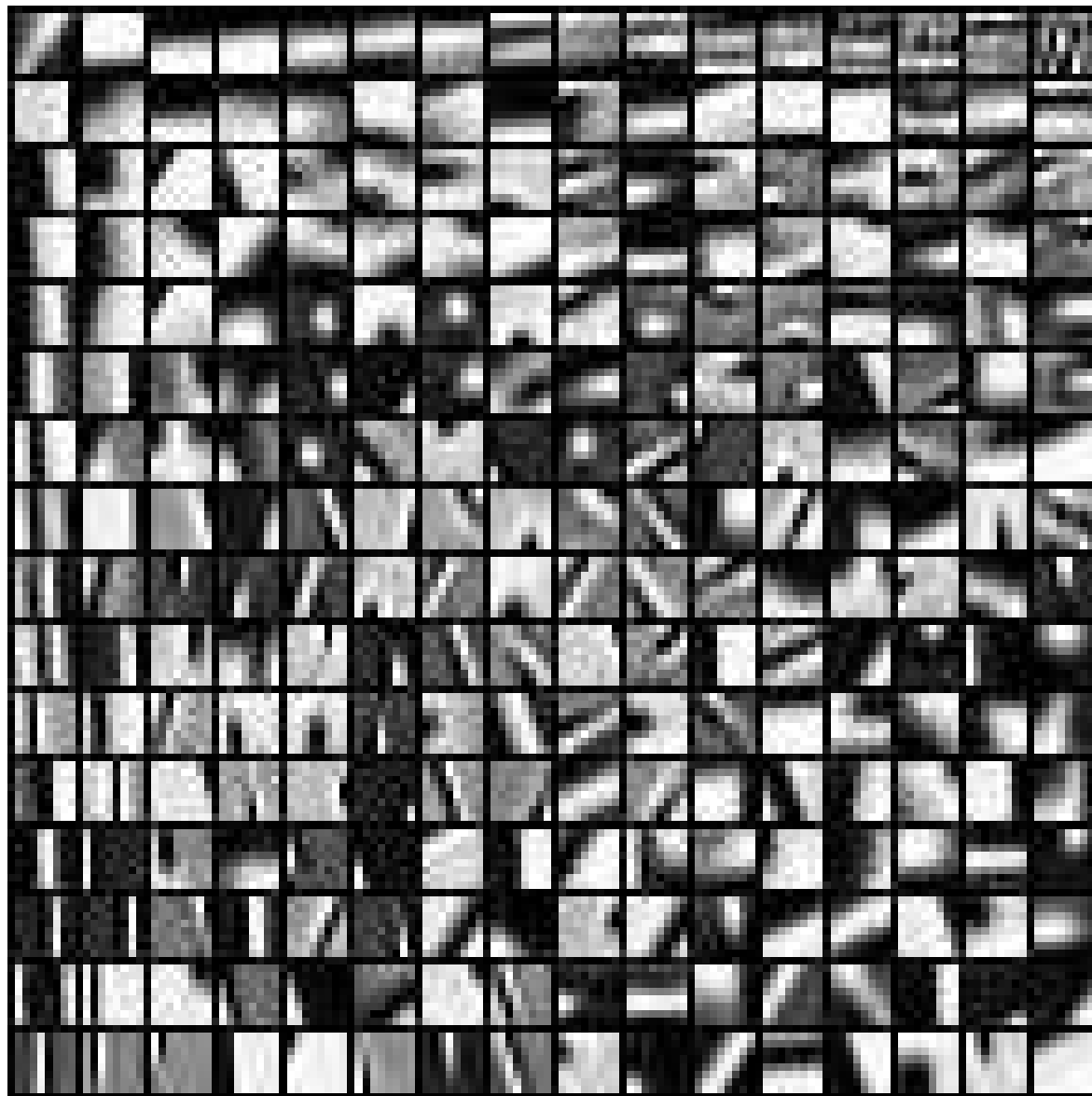


## Denoising result

(Mairal, Bach, Ponce, Sapiro, and Zisserman, 2009c)



What does the dictionary  $D$  look like?



# Inpainting a 12-Mpixel photograph

THE SALINAS VALLEY is in Northern California. It is a long narrow swale between two ranges of mountains, and the Salinas River winds and twists up the center until it falls at last into Monterey Bay.

I remember my childhood names for grasses and secret flowers. I remember where a toad may live and what time the birds awaken in the summer-and what trees and seasons smelled like-how people looked and walked and smelled even. The memory of odors is very rich.

I remember that the Gabilan Mountains to the east of the valley were light gay mountains full of sun and loveliness and a kind of invitation, so that you wanted to climb into their warm foothills almost as you want to climb into the lap of a beloved mother. They were beckoning mountains with a brown grass love. The Santa Lucias stood up against the sky to the west and kept the valley from the open sea, and they were dark and brooding-unfriendly and dangerous. I always found in myself a dread of west and a love of east. Where I ever got such an idea I cannot say, unless it could be that the morning came over the peaks of the Gabilans and the night drifted back from the ridges of the Santa Lucias. It may be that the birth and death of the day had some part in my feeling about the two ranges of mountains.

From both sides of the valley little streams slipped out of the hill canyons and fell into the bed of the Salinas River. In the winter of wet years the streams ran full-freshet, and they swelled the river until sometimes it raged and boiled, bank full, and then it was a destroyer. The river tore the edges of the farm lands and washed whole acres down; it toppled barns and houses into itself, to go floating and bobbing away. It trapped cows and pigs and sheep and drowned them in its muddy brown water and carried them to the sea. Then when the late spring came, the river drew in from its edges and the sand banks appeared. And in the summer the river didn't run at all above ground. Some pools would be left in the deep swirl places under a high bank. The tules and grasses grew back, and willows straightened up with the flood debris in their upper branches. The Salinas was only a part-time river. The summer sun drove it underground. It was not a fine river at all, but it was the only one we had and so we boasted about it-how dangerous it was in a wet winter and how dry it was in a dry summer. You can boast about anything if it's all you have. Maybe the less you have, the more you are required to boast.

The floor of the Salinas Valley, between the ranges and below the foothills, is level because this valley used to be the bottom of a hundred-mile inlet from the sea. The river mouth at Moss Landing was centuries ago the entrance to this long inland water. Once, fifty miles down the valley, my father bored a well. The drill came up first with topsoil and then with gravel and then with white sea sand full of shells and even pi...





# Inpainting a 12-Mpixel photograph



## Inpainting a 12-Mpixel photograph



## Inpainting a 12-Mpixel photograph



# Sparse structured PCA (Jenatton, Obozinski, and Bach, 2009b)

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

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

- Structure of the dictionary elements determined by the choice of overlapping groups  $\mathbf{G}$  (and thus  $\psi$ )
- Efficient learning procedures through “ $\eta$ -tricks”

– Reweighted  $\ell_2$ : 
$$\sum_{G \in \mathbf{G}} \|\mathbf{y}_G\|_2 = \min_{\eta_G \geq 0, G \in \mathbf{G}} \frac{1}{2} \sum_{G \in \mathbf{G}} \left\{ \frac{\|\mathbf{y}_G\|_2^2}{\eta_G} + \eta_G \right\}$$

# Application to face databases (1/3)



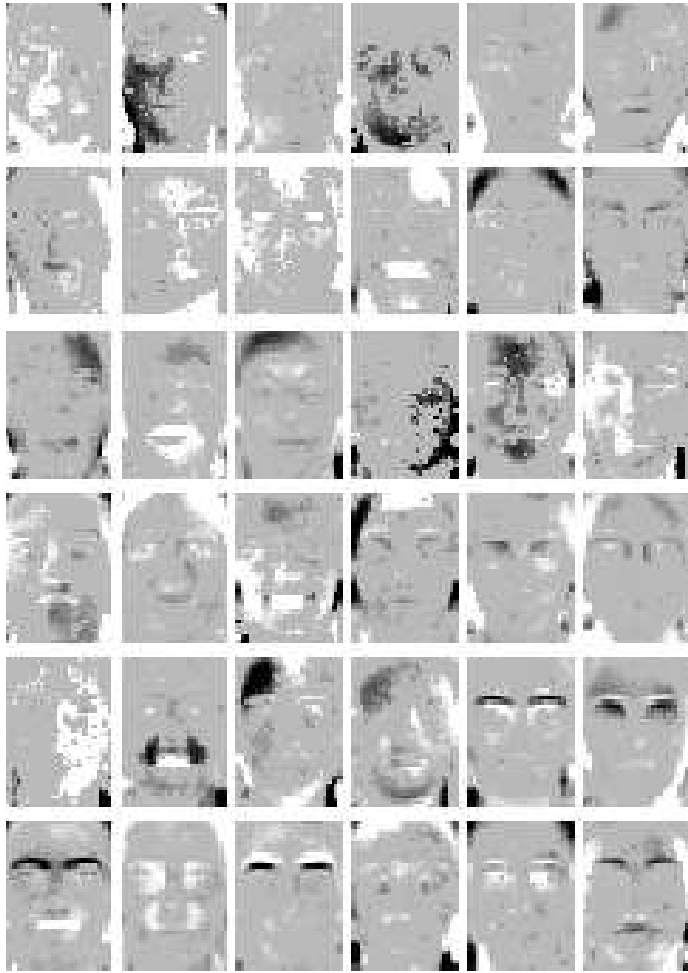
raw data



(unstructured) NMF

- NMF obtains partially local features

## Application to face databases (2/3)



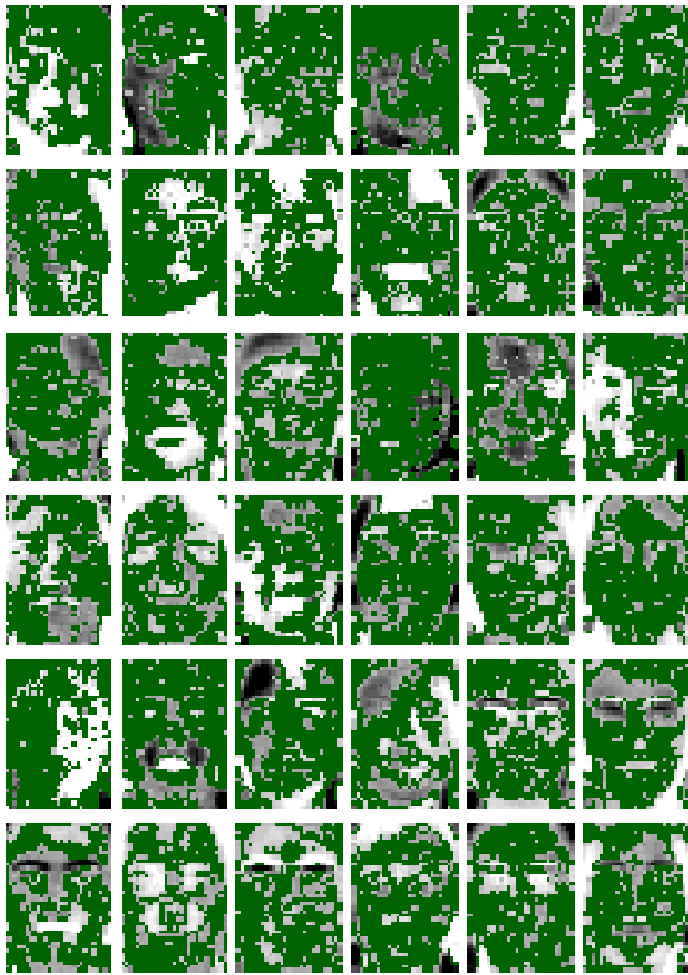
(unstructured) sparse PCA



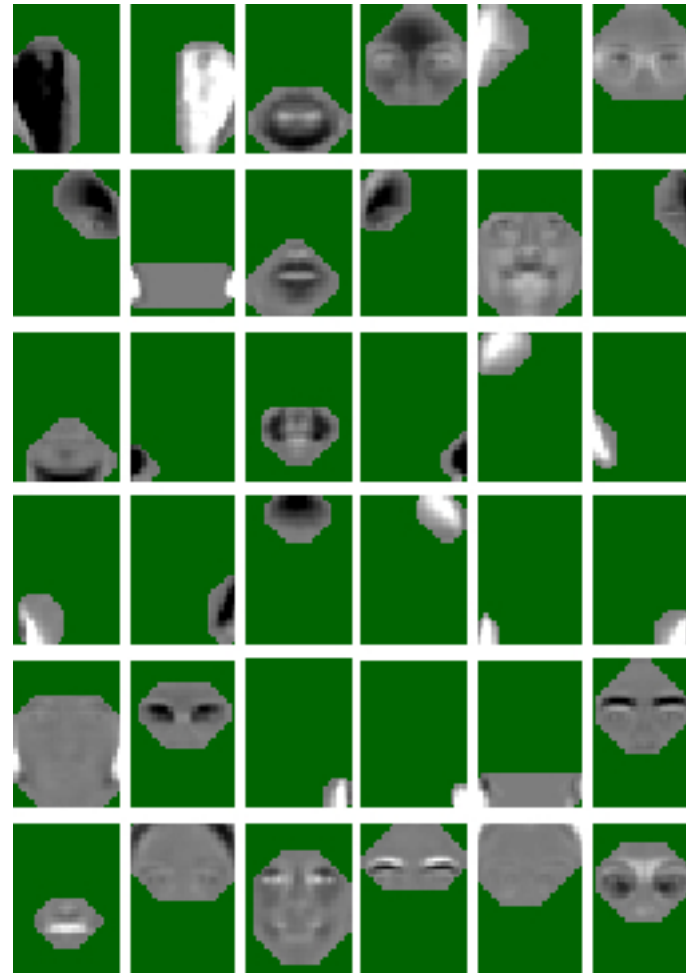
Structured sparse PCA

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

## Application to face databases (2/3)



(unstructured) sparse PCA

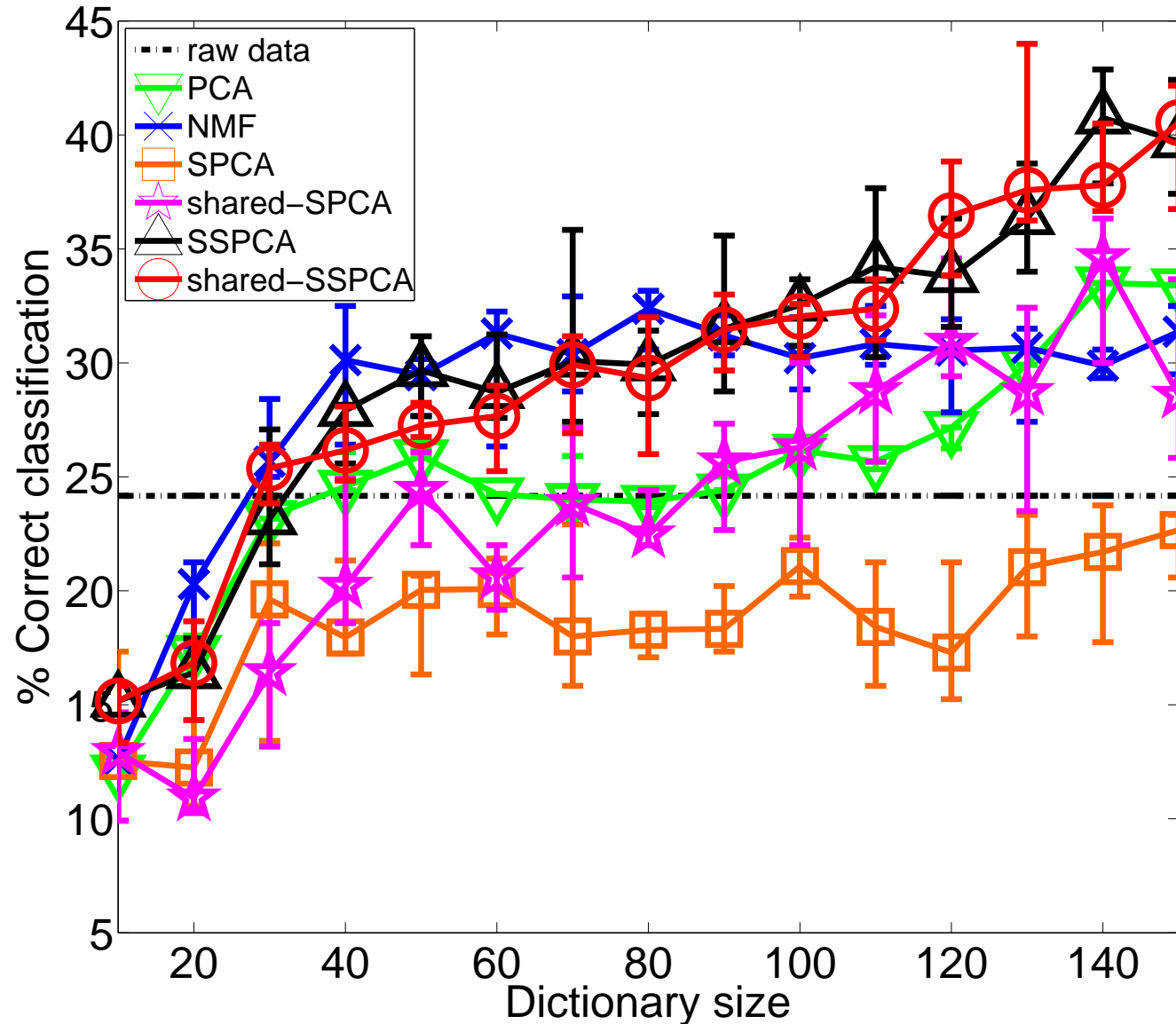


Structured sparse PCA

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

# Application to face databases (3/3)

- Quantitative performance evaluation on classification task





# Dictionary learning vs. sparse structured PCA

## Exchange roles of $\mathbf{D}$ and $\mathbf{A}$

- Sparse structured PCA (sparse and structured dictionary elements):

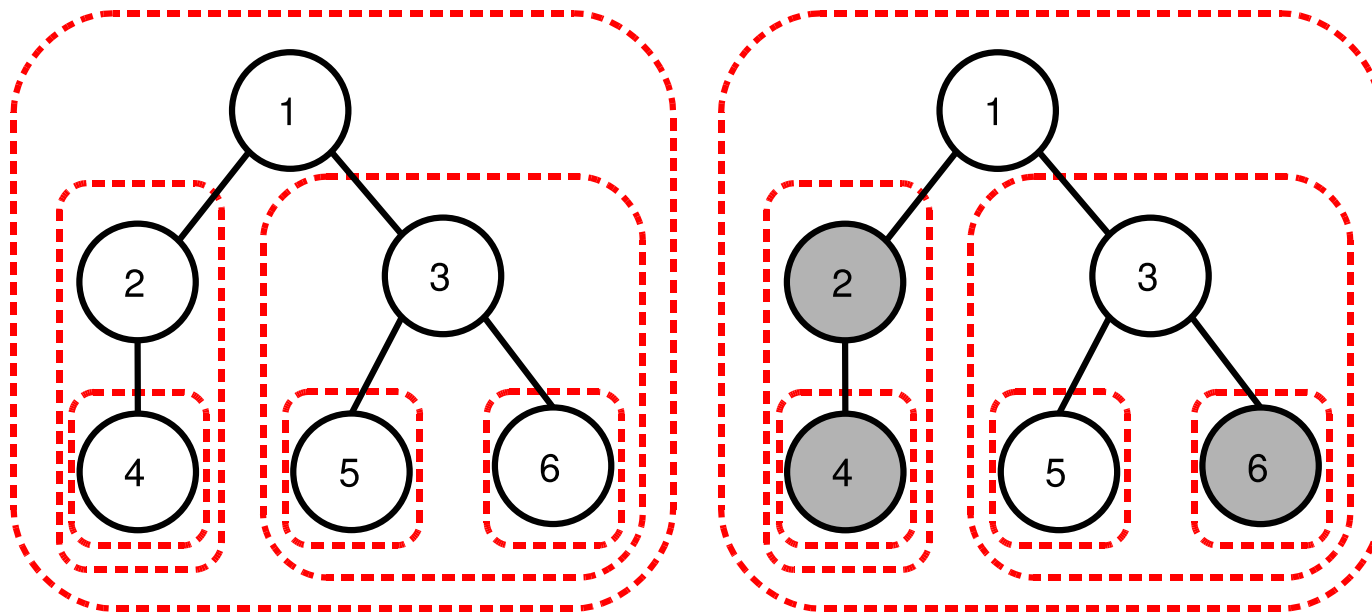
$$\min_{\substack{\mathbf{A} \in \mathbb{R}^{k \times n} \\ \mathbf{D} \in \mathbb{R}^{p \times k}}} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \lambda \sum_{j=1}^k \psi(\mathbf{d}_j) \text{ s.t. } \forall i, \|\boldsymbol{\alpha}_i\|_2 \leq 1.$$

- Dictionary learning with structured sparsity for  $\boldsymbol{\alpha}$ :

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

# Hierarchical dictionary learning (Jenatton, Mairal, Obozinski, and Bach, 2010)

- Structure on codes  $\alpha$  (not on dictionary  $\mathbf{D}$ )
- Hierarchical penalization:  $\psi(\alpha) = \sum_{G \in \mathcal{G}} \|\alpha_G\|_2$  where groups  $G$  in  $\mathcal{G}$  are equal to **set of descendants** of some nodes in a tree



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

# Hierarchical dictionary learning

## Efficient optimization

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

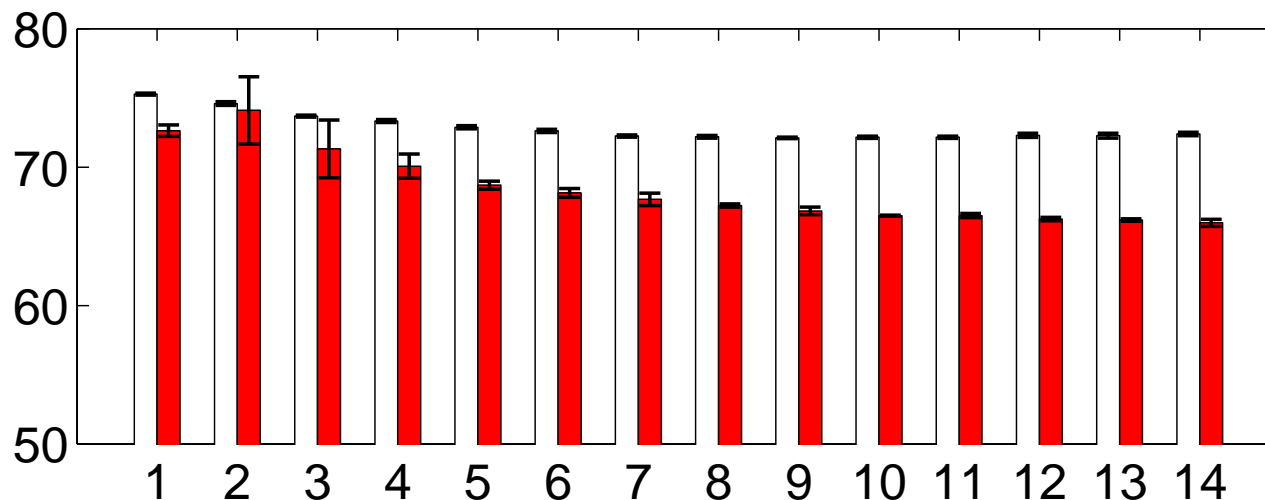
- Minimization with respect to  $\boldsymbol{\alpha}_i$  : regularized least-squares
  - Many algorithms dedicated to the  $\ell_1$ -norm  $\psi(\boldsymbol{\alpha}) = \|\boldsymbol{\alpha}\|_1$
- **Proximal methods** : first-order methods with optimal convergence rate (Nesterov, 2007; Beck and Teboulle, 2009)
  - Requires solving many times  $\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{1}{2}\|\mathbf{y} - \boldsymbol{\alpha}\|_2^2 + \lambda\psi(\boldsymbol{\alpha})$
- **Tree-structured regularization** : **Efficient linear time algorithm based on primal-dual decomposition** (Jenatton et al., 2010)

# Hierarchical dictionary learning

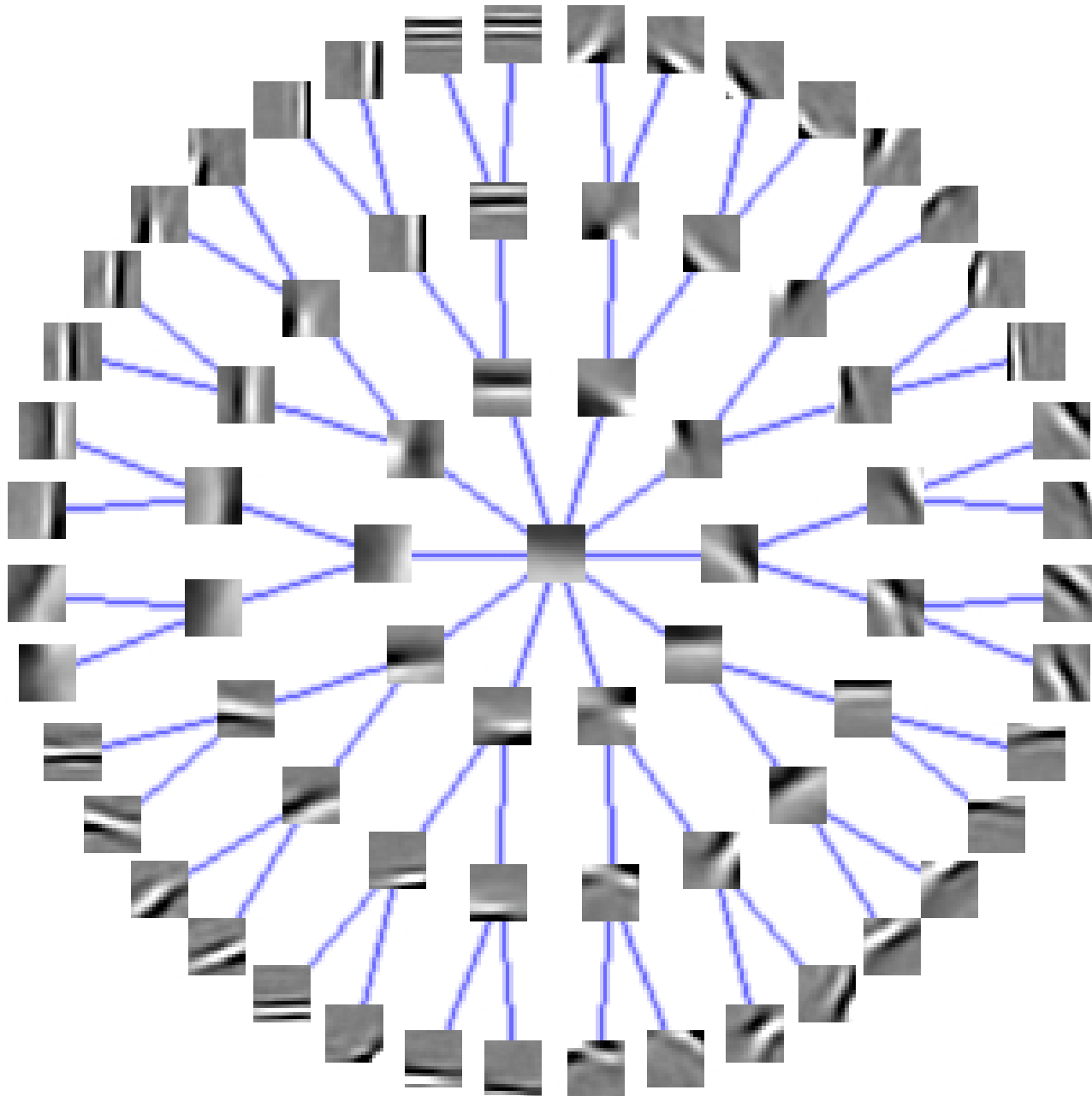
## Application to image denoising

- Reconstruction of 100,000  $8 \times 8$  natural images patches
  - Remove randomly subsampled pixels
  - Reconstruct with matrix factorization and structured sparsity

noise	50 %	60 %	70 %	80 %	90 %
flat	$19.3 \pm 0.1$	$26.8 \pm 0.1$	$36.7 \pm 0.1$	$50.6 \pm 0.0$	$72.1 \pm 0.0$
tree	$18.6 \pm 0.1$	$25.7 \pm 0.1$	$35.0 \pm 0.1$	$48.0 \pm 0.0$	$65.9 \pm 0.3$



# Application to image denoising - Dictionary tree



# Hierarchical dictionary learning

## Modelling of text corpora

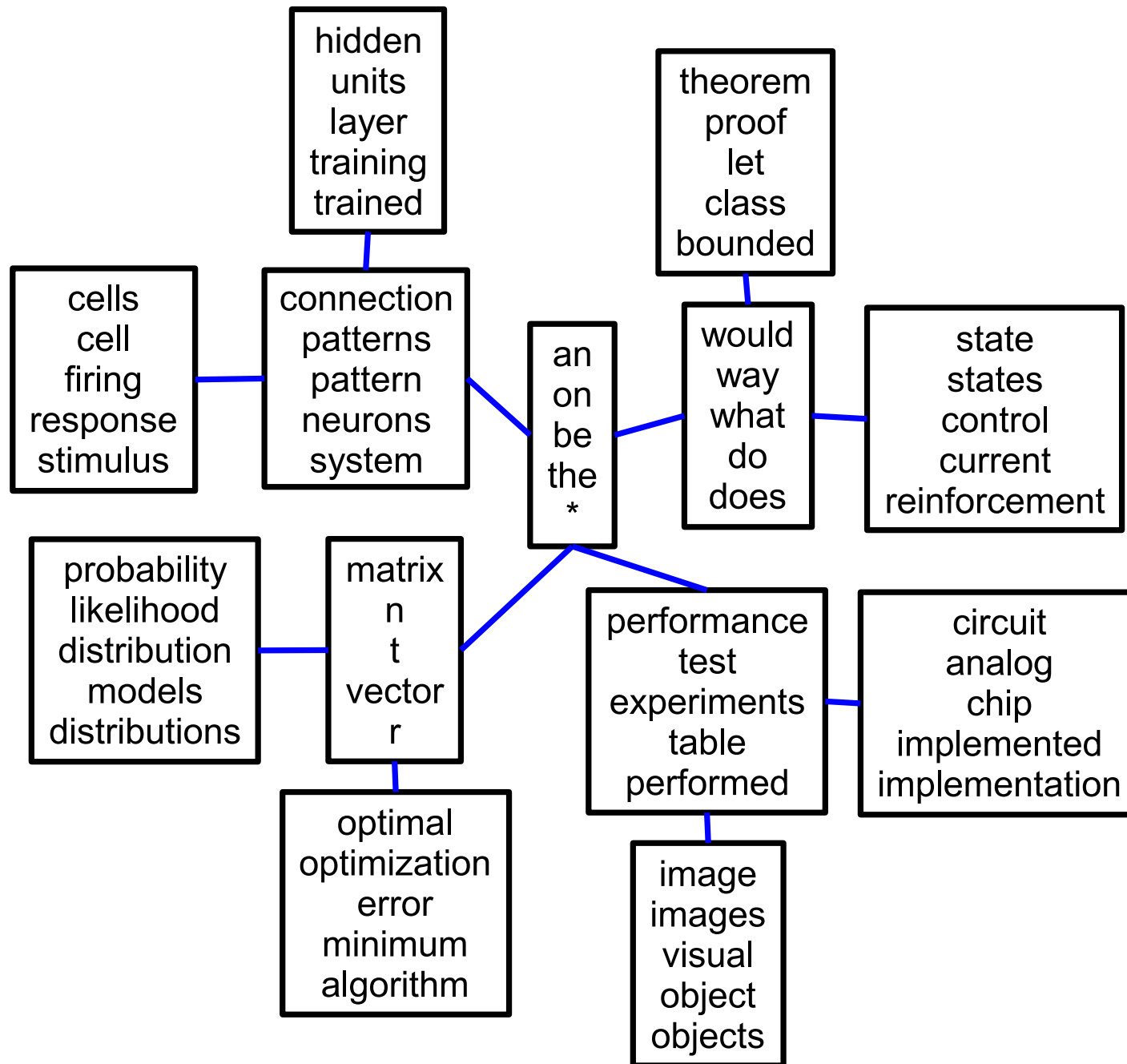
- Each document is modelled through word counts
- Low-rank matrix factorization of word-document matrix
- Probabilistic topic models (Blei et al., 2003a)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization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

# 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., 2003a)
  - Similar structures based on non parametric Bayesian methods (Blei et al., 2004)
  - **Can we achieve similar performance with simple matrix factorization formulation?**
- Experiments:
  - Qualitative: NIPS abstracts (1714 documents, 8274 words)
  - Quantitative: newsgroup articles (1425 documents, 13312 words)

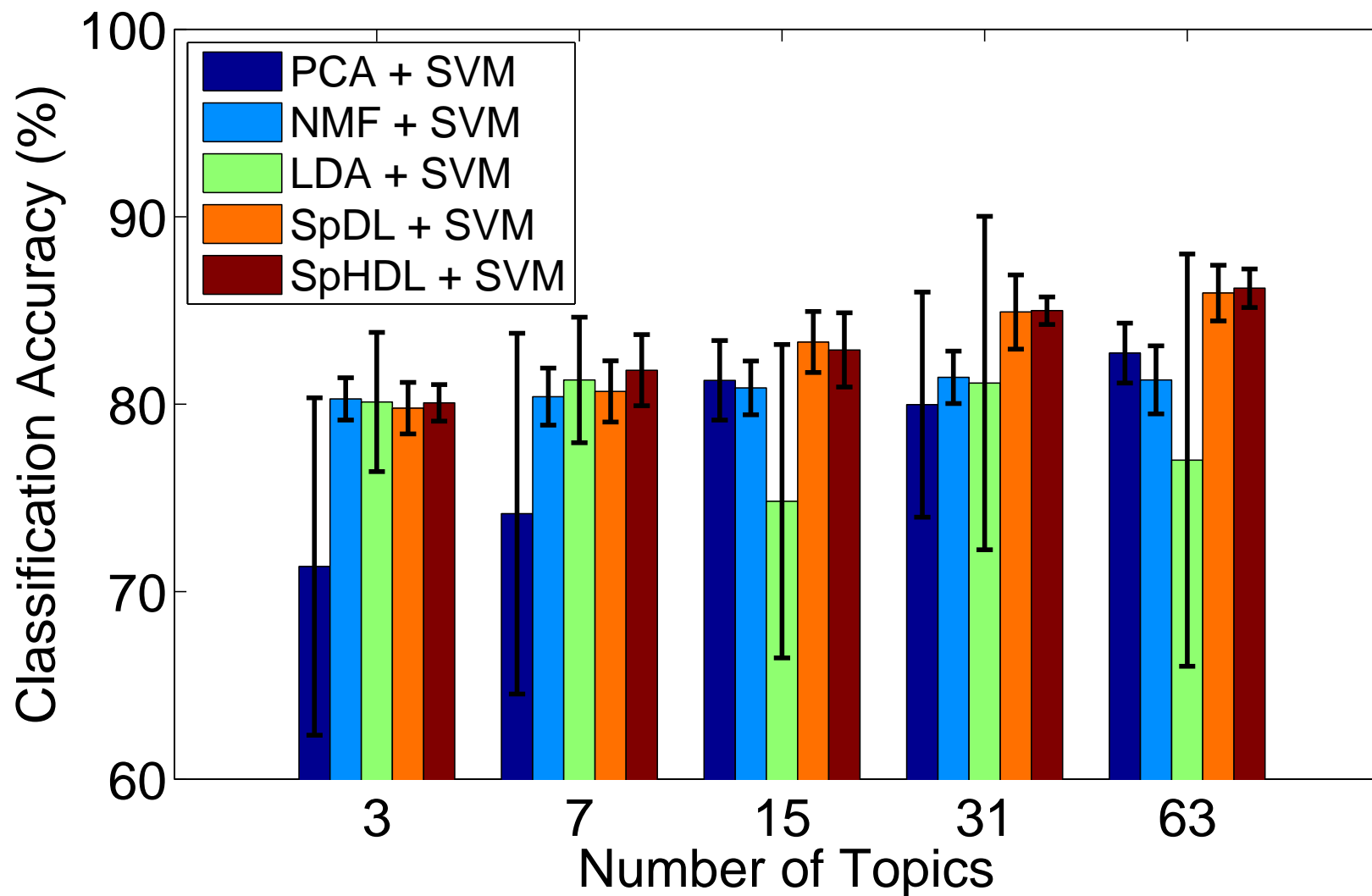
# Modelling of text corpora - Dictionary tree





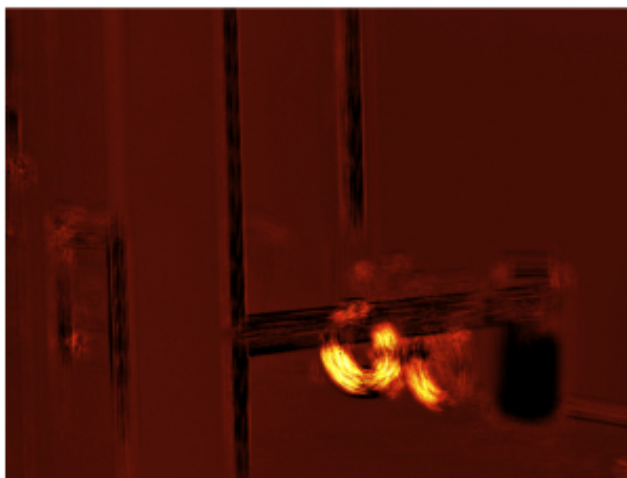
# Modelling of text corpora

- Comparison on predicting newsgroup article subjects:



# Alternative usages of dictionary learning

- Uses the “code”  $U$  as representation of observations for subsequent processing (Raina et al., 2007; Yang et al., 2009)
- Adapt dictionary elements to specific tasks (Mairal et al., 2009b)
  - Discriminative training for weakly supervised pixel classification (Mairal et al., 2008)



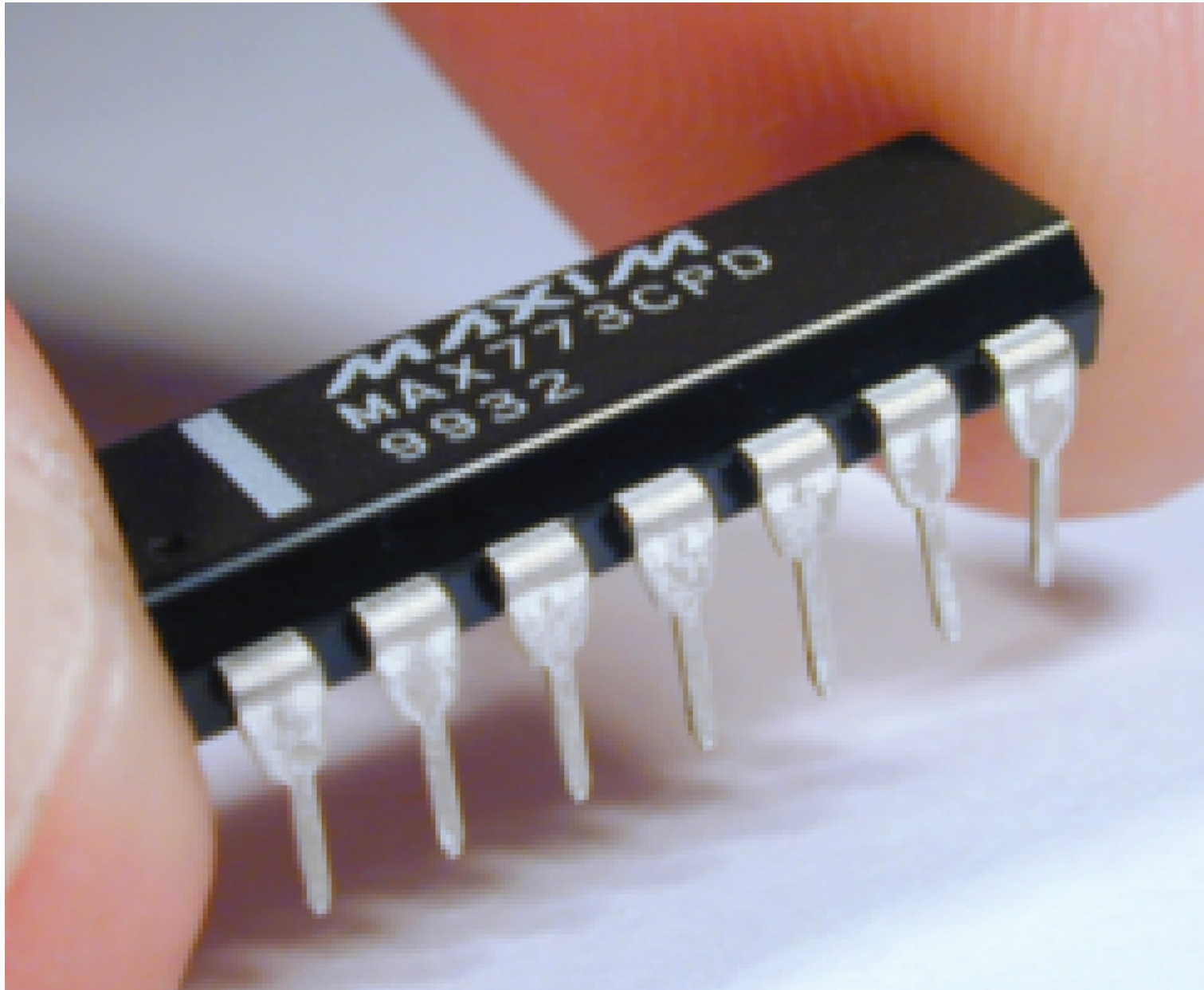
## Ongoing Work - Digital Zooming



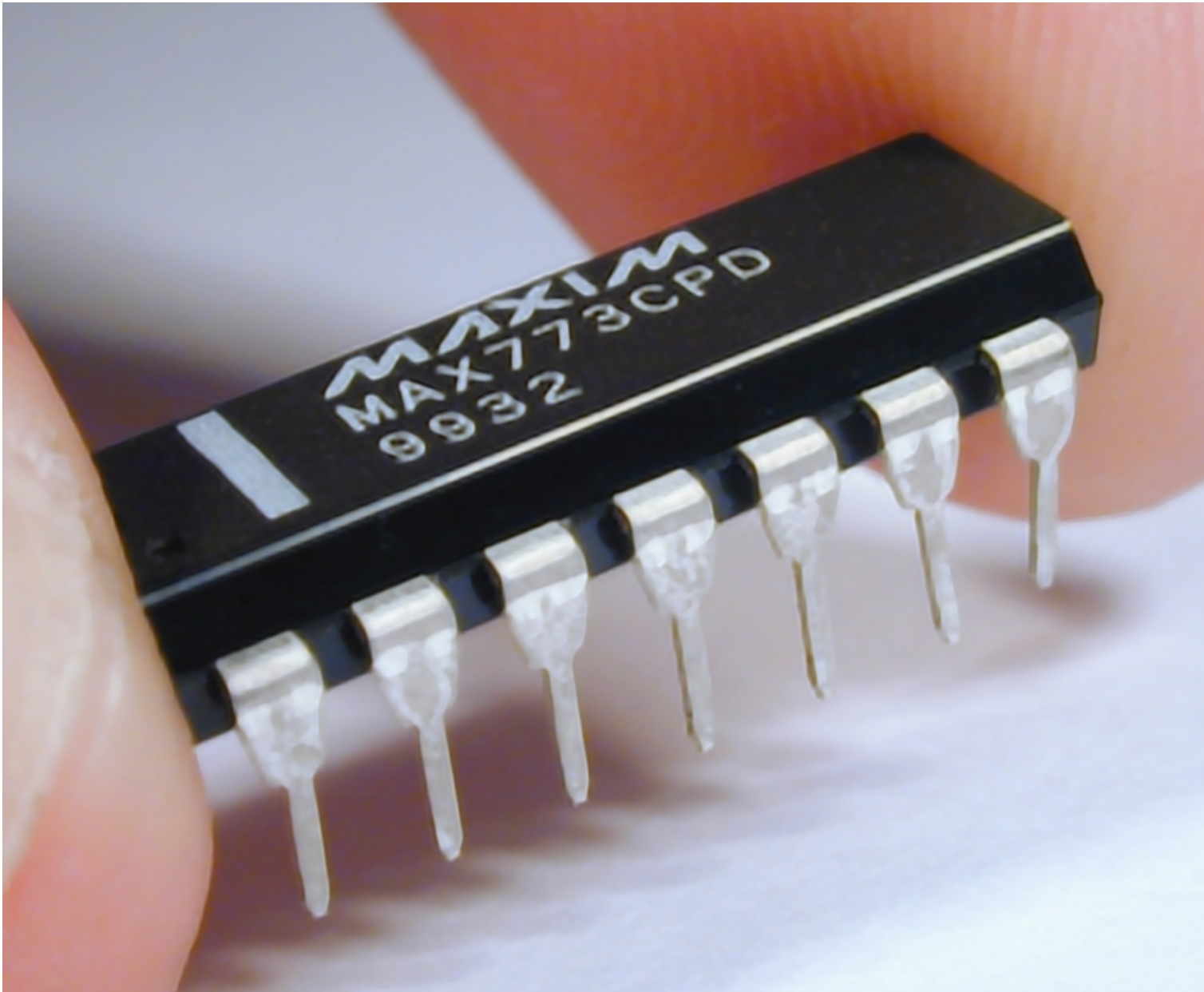
# Digital Zooming (Couzinie-Devy et al., 2010)



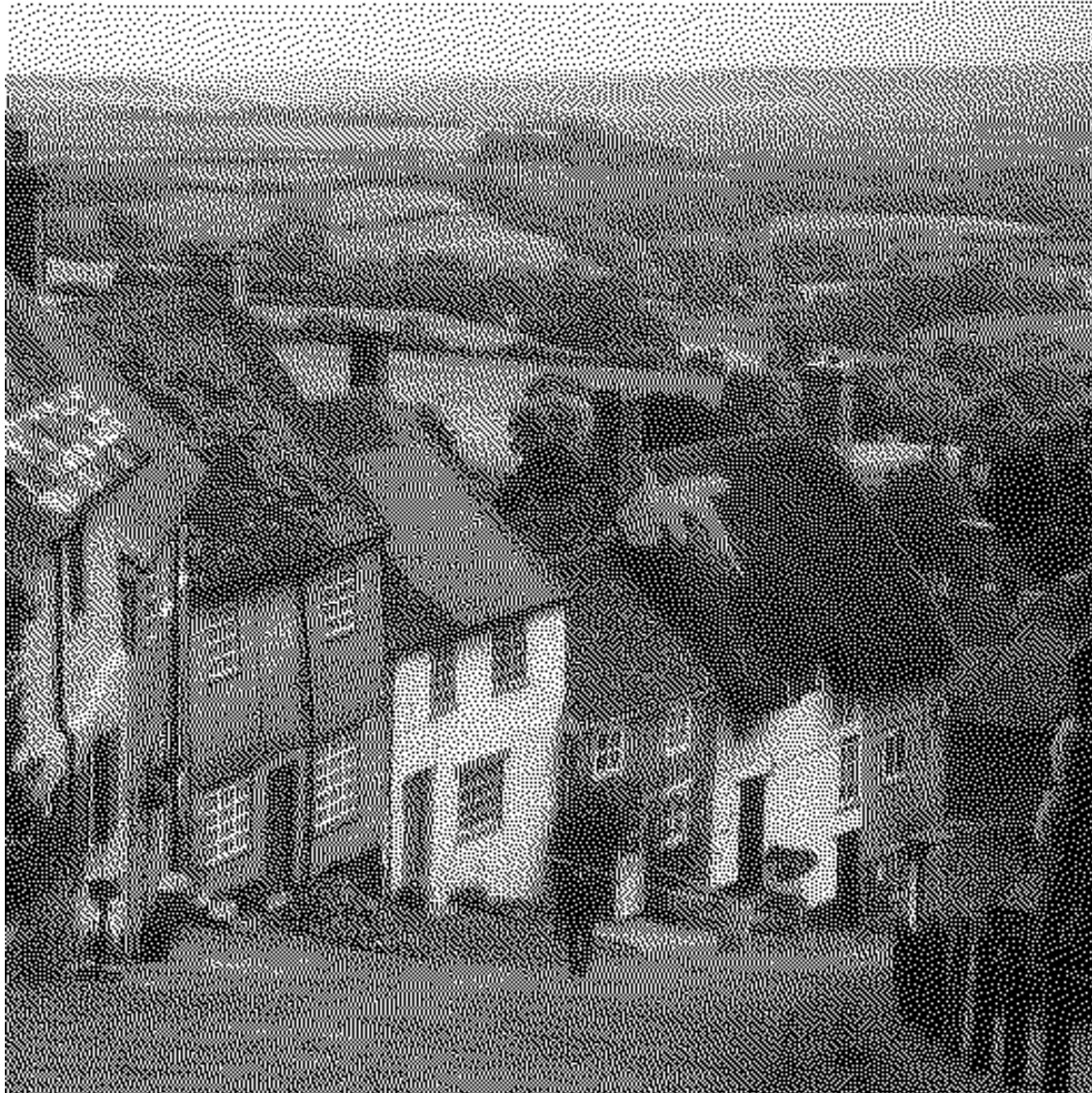
# Digital Zooming (Couzinie-Devy et al., 2010)



# Digital Zooming (Couzinie-Devy et al., 2010)



**Ongoing Work - Task-driven dictionaries**  
**inverse half-toning** (Mairal et al., 2010)



**Ongoing Work - Task-driven dictionaries**  
**inverse half-toning** (Mairal et al., 2010)





# Ongoing Work - Inverse half-toning



# Ongoing Work - Inverse half-toning



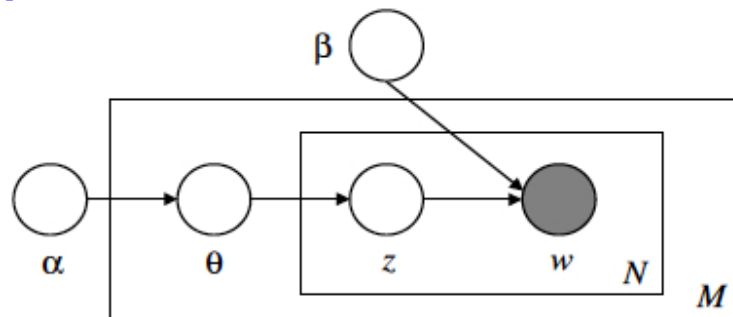
# Ongoing Work - Inverse half-toning



# Ongoing Work - Inverse half-toning

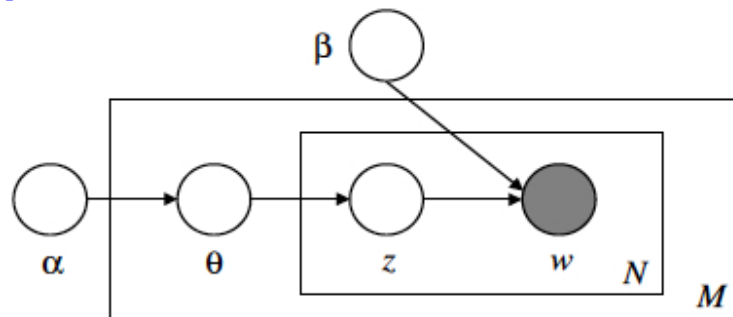


# Probabilistic topic models and matrix factorization



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

# Probabilistic topic models and matrix factorization



- **Latent Dirichlet allocation** (Blei et al., 2003b)

- For a document, sample  $\theta \in \mathbb{R}^k$  from a Dirichlet( $\alpha$ )
- For the  $n$ -th word of the same document,
  - \* sample a topic  $z_n$  from a multinomial with parameter  $\theta$
  - \* 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  $\theta$ , each word  $w_n$  is selected from a multinomial with parameter  $\sum_{z=1}^k \theta_k \beta(z, :) = \beta^\top \theta$
- Row of  $\beta$  = dictionary elements,  $\theta$  code for a document

# Probabilistic topic models and matrix factorization

- **Two different views on the same problem**
  - Interesting parallels to be made
  - Common problems to be solved
- **Structure on dictionary/decomposition coefficients** with adapted priors (Blei et al., 2004; Jenatton et al., 2010)
- **Identifiability and interpretation/evaluation of results**
- **Discriminative tasks** (Blei and McAuliffe, 2008; Lacoste-Julien et al., 2008; Mairal et al., 2009b)
- **Optimization and local minima**
  - Online learning (Mairal et al., 2009d)

# Sparsifying linear methods

- **Same pattern than with kernel methods**
  - High-dimensional inference rather than non-linearities
- Main difference: in general no unique way
- Sparse CCA (Sriperumbudur et al., 2009; Haroon and Shawe-Taylor, 2008; Archambeau and Bach, 2008)
- Sparse LDA (Moghaddam et al., 2006)
- Sparse ...



# Sparse methods for matrices

## Summary

- Structured matrix factorization has many applications
- Algorithmic issues
  - Dealing with large datasets
  - Dealing with structured sparsity
- Theoretical issues
  - Identifiability of structures, dictionaries or codes
  - Other approaches to sparsity and structure
- Non-convex optimization versus convex optimization
  - Convexification through unbounded dictionary size (Bach et al., 2008) - few performance improvements

# Sparse methods for machine learning

## Outline

- **Introduction - Overview**
- **Sparse linear estimation with the  $\ell_1$ -norm**
  - Convex optimization and algorithms
  - Theoretical results
- **Structured sparse methods on vectors**
  - Groups of features / Multiple kernel learning
  - Extensions (hierarchical or overlapping groups)
- **Sparse methods on matrices**
  - Multi-task learning
  - Matrix factorization (low-rank, sparse PCA, dictionary learning)

# Links with compressed sensing

## (Baraniuk, 2007; Candès and Wakin, 2008)

- Goal of compressed sensing: recover a signal  $w \in \mathbb{R}^p$  from only  $n$  measurements  $y = Xw \in \mathbb{R}^n$
- Assumptions: the signal is  $k$ -sparse,  $n$  much smaller than  $p$
- Algorithm:  $\min_{w \in \mathbb{R}^p} \|w\|_1$  such that  $y = Xw$
- Sufficient condition on  $X$  and  $(k, n, p)$  for perfect recovery:
  - Restricted isometry property (all small submatrices of  $X^\top X$  must be well-conditioned)
  - Such matrices are hard to come up with deterministically, but random ones are OK with  $k \log p = O(n)$
- Random  $X$  for machine learning?

# Sparse methods for machine learning

## Why use sparse methods?

- **Sparsity as a proxy to interpretability**
  - Structured sparsity
- **Sparsity for high-dimensional inference**
  - Influence on feature design
- **Sparse methods are not limited to least-squares regression**
- **Faster training/testing**
- **Better predictive performance?**
  - Problems are sparse if you look at them the right way

# Conclusion - Interesting questions/issues

- Implicit vs. explicit features
  - Can we algorithmically achieve  $\log p = O(n)$  with explicit unstructured features?
- Norm design
  - What type of behavior may be obtained with sparsity-inducing norms?
- Overfitting convexity
  - Do we actually need convexity for matrix factorization problems?

# References

- J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10:803–826, 2009.
- 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.
- C. Archambeau and F. Bach. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21 (NIPS)*, 2008.
- A. Argyriou, C.A. Micchelli, and M. Pontil. On spectral learning. *Journal of Machine Learning Research*, 2009. To appear.
- F. Bach. High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning. Technical Report 0909.0844, arXiv, 2009a.
- 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. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008b.
- F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in Neural Information Processing Systems*, 2008c.
- F. Bach. Self-concordant analysis for logistic regression. Technical Report 0910.4627, ArXiv, 2009b.
- F. Bach. Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048, 2008d.

- 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)*, 2004a.
- F. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems 17*, 2004b.
- F. Bach, J. Mairal, and J. Ponce. Convex sparse matrix factorizations. Technical Report 0812.1869, ArXiv, 2008.
- O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9: 485–516, 2008.
- R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.
- 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.
- J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizbal. *Numerical Optimization Theoretical and Practical Aspects*. Springer, 2003.
- J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer-Verlag, 2000.
- L. Bottou and O. Bousquet. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems (NIPS)*, volume 20, 2008.
- 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 T. Tao. The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ . *Annals of Statistics*, 35(6):2313–2351, 2007.



- E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.
- E.J. Candès and Y. Plan. Matrix completion with noise. 2009a. Submitted.
- E.J. Candès and Y. Plan. Near-ideal model selection by  $l_1$  minimization. *The Annals of Statistics*, 37(5A):2145–2177, 2009b.
- E.J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Arxiv preprint arXiv:0912.3599*, 2009.
- F. Caron and A. Doucet. Sparse Bayesian nonparametric regression. In *25th International Conference on Machine Learning (ICML)*, 2008.
- S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.
- A. d’Aspremont and L. El Ghaoui. Testing the nullspace property using semidefinite programming. Technical Report 0807.3520v5, arXiv, 2008.
- 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.
- B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2):407–451, 2004.
- 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. Hofling, 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).
- D. R. Hardoon and J. Shawe-Taylor. Sparse canonical correlation analysis. In *Sparsity and Inverse Problems in Statistical Theory and Econometrics*, 2008.
- T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.
- J. Huang and T. Zhang. The benefit of group sparsity. Technical Report 0901.2962v2, ArXiv, 2009.
- 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.
- H. Ishwaran and J.S. Rao. Spike and slab variable selection: frequentist and Bayesian strategies. *The*

*Annals of Statistics*, 33(2):730–773, 2005.

- 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 *Proc. ICML*, 2010.
- A. Juditsky and A. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via  $\ell_1$ -minimization. Technical Report 0809.2650v1, ArXiv, 2008.
- G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.
- 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, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004a.
- 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, 2004b.
- H. Lee, A. Battle, R. Raina, and A. Ng. Efficient sparse coding algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, 2007.

- Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.
- J. Liu, S. Ji, and J. Ye. Multi-Task Feature Learning Via Efficient  $l_{2,1}$ -Norm Minimization. *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.
- 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)*, 2008.
- 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. Supervised dictionary learning. *Advances in Neural Information Processing Systems (NIPS)*, 21, 2009b.
- 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)*, 2009c.
- Julien Mairal, F. Bach, J. Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 2009d.
- H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval*

*Research Logistics Quarterly*, 3:111–133, 1956.

- P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.
- R. Mazumder, T. Hastie, and R. Tibshirani. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. 2009. Submitted.
- N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, 52(1):374–393, 2008.
- N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of statistics*, 34(3):1436, 2006.
- 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.
- B. Moghaddam, Y. Weiss, and S. Avidan. Generalized spectral bounds for sparse LDA. In *Proceedings of the 23rd international conference on Machine Learning (ICML)*, 2006.
- R.M. Neal. *Bayesian learning for neural networks*. Springer Verlag, 1996.
- Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Kluwer Academic Pub, 2003.
- 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. R. Osborne, B. Presnell, and B. A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, 9(2):319–337, 2000.
- 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.
- P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.
- B. Recht, W. Xu, and B. Hassibi. Null Space Conditions and Thresholds for Rank Minimization. 2009. Submitted.
- J. D. M. Rennie and N. Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine Learning (ICML)*, 2005.
- V. Roth and B. Fischer. The group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning*

(ICML), 2008.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2001.

M.W. Seeger. Bayesian inference and optimal design for the sparse linear model. *The Journal of Machine Learning Research*, 9:759–813, 2008.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

S. Sonnenburg, G. Raetsch, C. Schaefer, and B. Schoelkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

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

B. K. Sriperumbudur, D. A. Torres, and G. R. G. Lanckriet. A d.c. programming approach to the sparse generalized eigenvalue problem. Technical Report 0901.1504v2, ArXiv, 2009.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

S. A. Van De Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, 36(2):614, 2008.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -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.

- 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.
- T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.
- 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)*, 2009.
- 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 T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, 67(2):301–320, 2005.
- H. Zou and R. Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509–1533, 2008.