Hierarchical kernel learning

Francis Bach

Willow project, INRIA - Ecole Normale Supérieure

March 2010

Outline

- **Supervised learning and regularization**
  - Kernel methods vs. sparse methods
- **MKL: Multiple kernel learning**
  - Non linear sparse methods
- **HKL: Hierarchical kernel learning**
  - Non linear variable selection
- **Extensions**
  - Structured sparsity, sparse PCA (dictionary learning)

Supervised learning and regularization

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

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

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., ℓ^2-norms)
     - Non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped theory (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., ℓ^2-norms)
     - Non linear predictors
     - Non parametric supervised learning and kernel methods
     - Well developped theory (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: ℓ_1-norm $\|w\|_1 = \sum_{i=1}^{p} |w_i|$
     - Perform model selection as well as regularization
     - Theory "in the making"

Kernel methods: regularization by ℓ^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: $\boxed{\min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|_{\mathbf{2}}^2}$

Kernel methods: regularization by ℓ^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: $\boxed{\displaystyle \min_{w \in \mathbb{R}^p} \sum_{i=1}^n \ell(y_i, w^\top \Phi(x_i)) + \frac{\mu}{2} \|w\|_{\mathbf{2}}^2}$
- **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: $\boxed{\displaystyle \min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n \ell(y_i, (K\alpha)_i) + \frac{\mu}{2} \alpha^\top K \alpha}$
  - Kernel matrix $K_{ij} = k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$

Kernel methods: regularization by ℓ^2-norm

- Running time $O(n^2\kappa + n^3)$ where κ complexity of one kernel evaluation (often much less) - **independent from p**

- **Kernel trick**: implicit mapping if $\kappa = o(p)$ by using only $k(x_i, x_j)$ instead of $\Phi(x_i)$

- Examples:
  - Polynomial kernel: $k(x,y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} =$ polynomials
  - Gaussian kernel: $k(x,y) = e^{-\alpha\|x-y\|_2^2} \quad \Rightarrow \mathcal{F} =$ smooth functions
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)

Kernel methods: regularization by ℓ^2-norm

- Running time $O(n^2\kappa + n^3)$ where κ complexity of one kernel evaluation (often much less) - **independent from p**

- **Kernel trick**: implicit mapping if $\kappa = o(p)$ by using only $k(x_i, x_j)$ instead of $\Phi(x_i)$

- Examples:
  - Polynomial kernel: $k(x,y) = (1 + x^\top y)^d \Rightarrow \mathcal{F} =$ polynomials
  - Gaussian kernel: $k(x,y) = e^{-\alpha\|x-y\|_2^2} \Rightarrow \mathcal{F} =$ smooth functions
  - Kernels on structured data (see Shawe-Taylor and Cristianini, 2004)

- $+$: Implicit non linearities and high-dimensionality

- $-$: Problems of interpretability, dimension too high?

ℓ_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 $w \in \mathbb{R}^p$:

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

$$\text{Error on data} \quad + \quad \text{Regularization}$$

- square loss $\Rightarrow$ basis pursuit (signal processing) (Chen et al., 2001), Lasso (statistics/machine learning) (Tibshirani, 1996)

ℓ^2-norm vs. ℓ^1-norm

- ℓ^1-norms lead to **sparse**/interpretable models
- ℓ^2-norms can be run implicitly with very large feature spaces

ℓ^2-norm vs. ℓ^1-norm

- ℓ^1-norms lead to **sparse**/interpretable models
- ℓ^2-norms can be run implicitly with very large feature spaces
- **Algorithms**:
  - Smooth convex optimization vs. nonsmooth convex optimization
  - First-order methods (Fu, 1998; Wu and Lange, 2008)
  - Homotopy methods (Markowitz, 1956; Efron et al., 2004)
- **Theory**:
  - Advantages of parsimony?
  - Consistent estimation of the support?

Lasso - Two main recent theoretical results

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

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

where $\mathbf{Q} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$.

Lasso - Two main recent theoretical results

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

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

where $\mathbf{Q} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$.

- **The Lasso alone cannot find in general the good model**
- Two step-procedures
  - Adaptive Lasso (Zou, 2006; van de Geer et al., 2010)
    $\Rightarrow$ penalize by $\sum_{j=1}^p \frac{|w_j|}{|\hat{w}_j|}$
  - Resampling (Bach, 2008a; Meinshausen and Bühlmann, 2008)
    $\Rightarrow$ use the bootstrap to select the model

Lasso - Two main recent theoretical results

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

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

   where $\mathbf{Q} = \frac{1}{n}\sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{p\times p}$.

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

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

Outline

- **Supervised learning and regularization**
  - Kernel methods vs. sparse methods
- **MKL: Multiple kernel learning**
  - Non linear sparse methods
- **HKL: Hierarchical kernel learning**
  - Non linear variable selection
- **Extensions**
  - Structured sparsity, sparse PCA (dictionary learning)

Multiple kernel learning (MKL) (Lanckriet et al., 2004; Bach et al., 2004a)

- Sparse methods are most often linear

- Sparsity with non-linearities
  - replace $f(x) = \sum_{j=1}^{p} w_j^\top x_j$ with $x_j \in \mathbb{R}$ and $w_j \in \mathbb{R}$
  - by $f(x) = \sum_{j=1}^{p} w_j^\top \Phi_j(x)$ with $\Phi_j(x) \in \mathcal{F}_j$ an $w_j \in \mathcal{F}_j$

- Replace the ℓ^1-norm $\sum_{j=1}^{p} |w_j|$ by "block" ℓ^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., 2004; 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: $f(x) = w_1{}^\top \Phi_1(x) + \cdots + w_p{}^\top \Phi_p(x)$

$$
\begin{tikzcd}
 & \Phi_1(x)^\top \quad w_1 \arrow[dr] & \\
 & \vdots \qquad\quad \vdots & \\
x \arrow[uur] \arrow[r] \arrow[ddr] & \Phi_j(x)^\top \quad w_j \arrow[r] & w_1^\top \Phi_1(x) + \cdots + w_p^\top \Phi_p(x) \\
 & \vdots \qquad\quad \vdots & \\
 & \Phi_p(x)^\top \quad w_p \arrow[ur] &
\end{tikzcd}
$$

  - Generalized additive models (Hastie and Tibshirani, 1990)
  - **Link between regularization and kernel matrices**

Regularization for multiple features

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

- 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

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

- 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 ℓ^1-norm**:

1. Algorithms (Bach et al., 2004a,b; Rakotomamonjy et al., 2008)
2. Analysis of sparsity inducing properties (Bach, 2008b)
3. Sparse kernel combinations $\sum_{j=1}^p \eta_j K_j$ (Bach et al., 2004a)
4. Application to data fusion and hyperparameter learning

Outline

- **Supervised learning and regularization**
  - Kernel methods vs. sparse methods
- **MKL: Multiple kernel learning**
  - Non linear sparse methods
- **HKL: Hierarchical kernel learning**
  - Non linear variable selection
- **Extensions**
  - Structured sparsity, sparse PCA (dictionary learning)

Lasso - Two main recent theoretical results

1. **Support recovery condition**

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

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

Lasso - Two main recent theoretical results

1. **Support recovery condition**

2. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): 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. **(sub-)exponentially many irrelevant variables** (Zhao and Yu, 2006; Wainwright, 2006; Bickel et al., 2008; Lounici, 2008; Meinshausen and Yu, 2009): 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!**

Hierarchical kernel learning (Bach, 2008c)

- Many kernels can be decomposed as a sum of many "small" kernels indexed by a certain set V: $\boxed{k(x,x') = \sum_{v \in V} k_v(x,x')}$
- 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

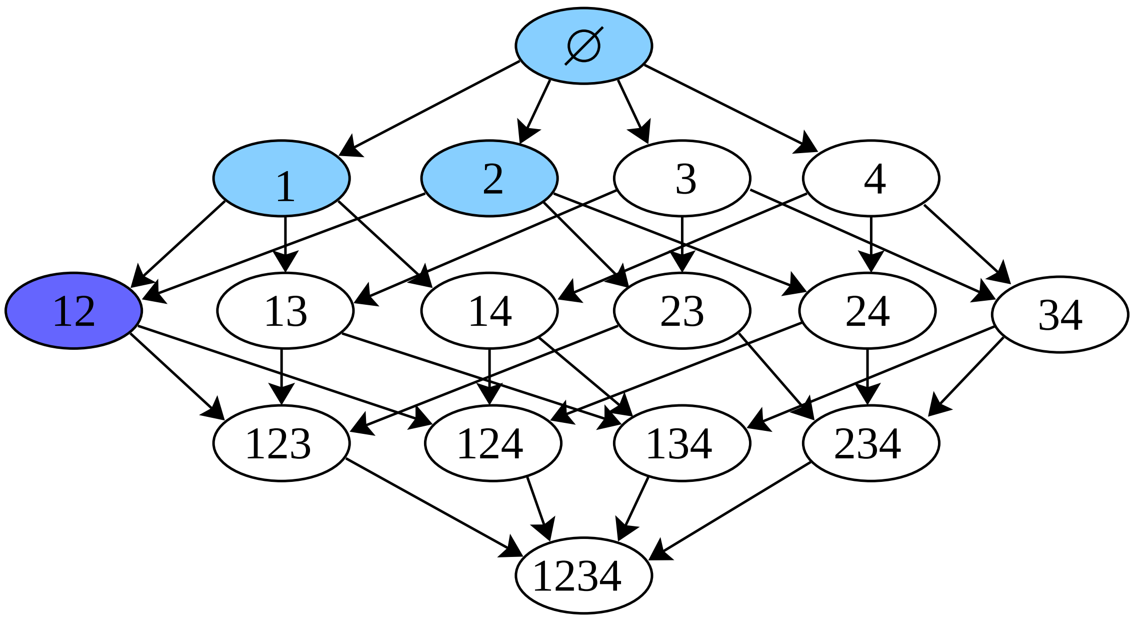
Restricting the set of active kernels

- 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 ℓ_1-norm: $\sum_{v \in V} \|w_v\|_2$
  - 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 $\{1, \ldots, q\}$ with inclusion DAG
  – Select a subset only after all its subsets have been selected

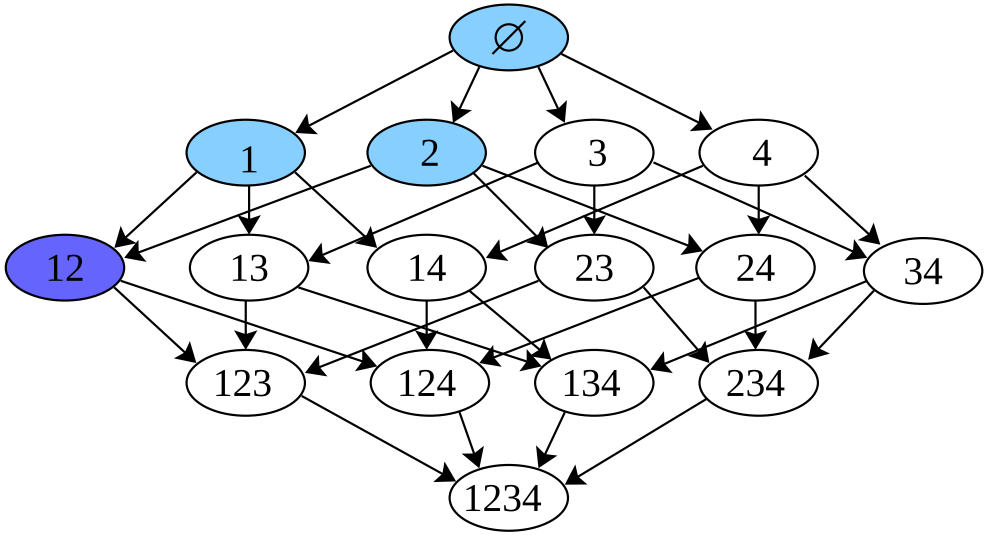


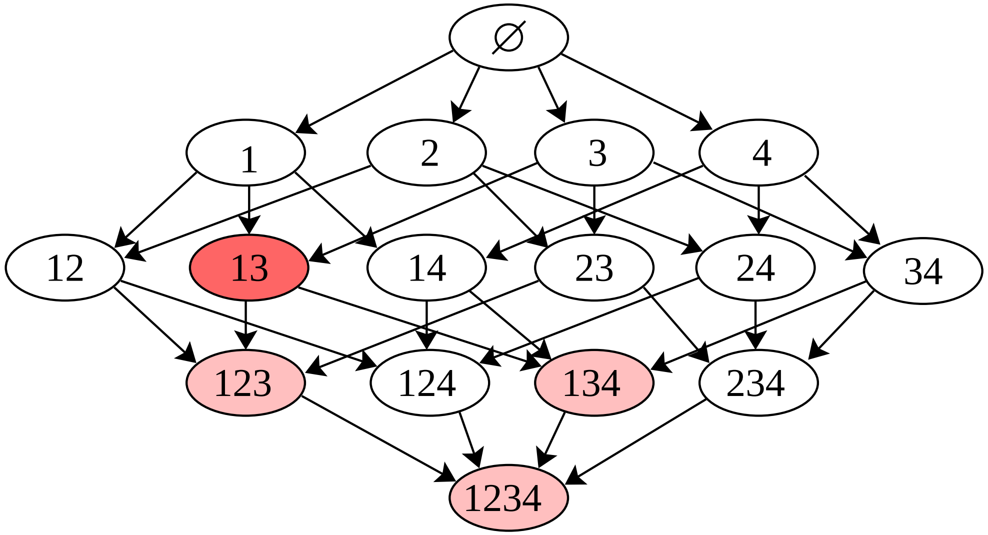
DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  - $\mathrm{D}(v)$ is the set of descendants of $v \in V$:

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

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





DAG-adapted norm (Zhao & Yu, 2008)

- Graph-based structured regularization
  - $\mathrm{D}(v)$ is the set of descendants of $v \in V$:

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

- 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

$$\begin{array}{lcl} n & = & \text{number of observations} \\ p & = & \text{number of vertices in the DAG} \\ \deg(V) & = & \text{maximum out degree in the DAG} \\ \mathrm{num}(V) & = & \text{number of connected components in the DAG} \end{array}$$

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

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

Scaling between p, n and other graph-related quantities

$$
\begin{array}{lcl}
n & = & \text{number of observations} \\
p & = & \text{number of vertices in the DAG} \\
\deg(V) & = & \text{maximum out degree in the DAG} \\
\text{num}(V) & = & \text{number of connected components in the DAG}
\end{array}
$$

- **Proposition** (Bach, 2009): 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: $\text{num}(V) = p \Rightarrow \boxed{\log p = O(n)}$

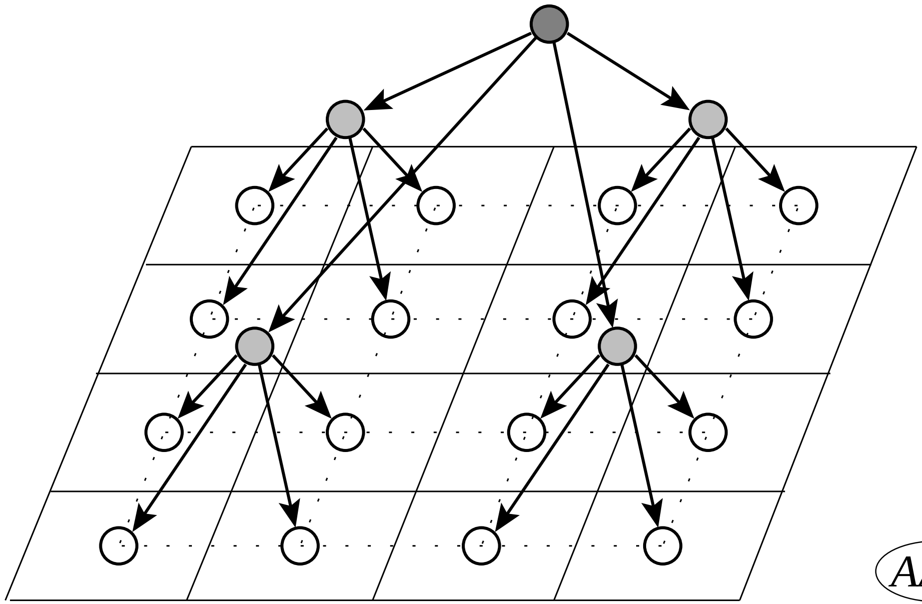
- Power set of q elements: $\deg(V) = q \Rightarrow \boxed{\log q = \log \log p = O(n)}$

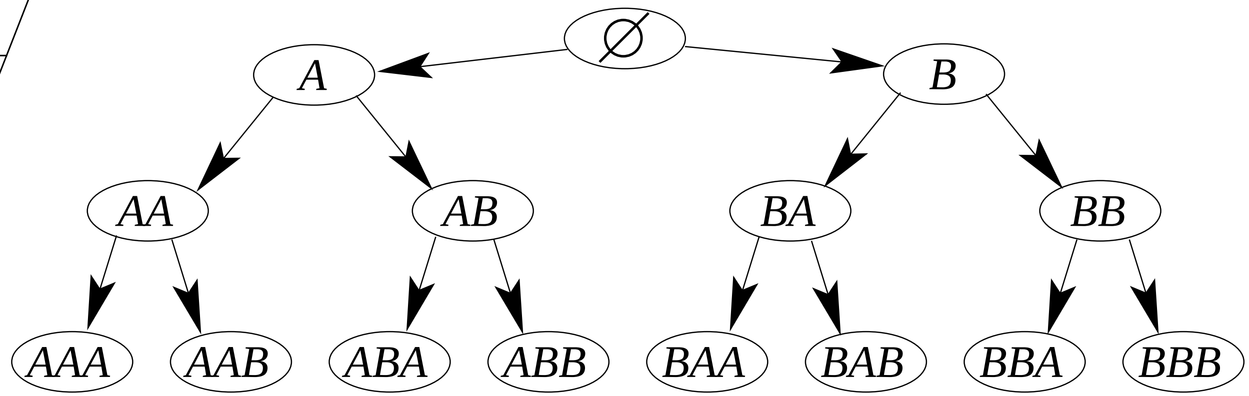
Mean-square errors (regression)

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	**43.3±1.0**
abalone	4177	10	rbf	$\approx 10^{10}$	**43.0±0.9**	45.0±1.7	43.7±1.0	43.0±1.1
boston	506	13	pol4	$\approx 10^{9}$	**17.1±3.6**	24.7±10.8	22.2±2.2	18.1±3.8
boston	506	13	rbf	$\approx 10^{12}$	**16.4±4.0**	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	**56.4±0.7**	56.4±0.8
pumadyn-32fh	8192	32	rbf	$\approx 10^{31}$	57.7±0.6	72.2±22.5	56.5±0.8	**55.7±0.7**
pumadyn-32fm	8192	32	pol4	$\approx 10^{22}$	6.9±0.1	6.4±1.6	7.0±0.1	**3.1±0.0**
pumadyn-32fm	8192	32	rbf	$\approx 10^{31}$	5.0±0.1	46.2±51.6	7.1±0.1	**3.4±0.0**
pumadyn-32nh	8192	32	pol4	$\approx 10^{22}$	84.2±1.3	73.3±25.4	83.6±1.3	**36.7±0.4**
pumadyn-32nh	8192	32	rbf	$\approx 10^{31}$	56.5±1.1	81.3±25.0	83.7±1.3	**35.5±0.5**
pumadyn-32nm	8192	32	pol4	$\approx 10^{22}$	60.1±1.9	69.9±32.8	77.5±0.9	**5.5±0.1**
pumadyn-32nm	8192	32	rbf	$\approx 10^{31}$	15.7±0.4	67.3±42.4	77.6±0.9	**7.2±0.1**

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**

Grouped variables

- Supervised learning with known groups:
  - The ℓ_1-ℓ_2 norm

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

  - The ℓ_1-ℓ_2 norm sets to zero non-overlapping groups of variables (as opposed to single variables for the ℓ_1 norm)

Grouped variables

- Supervised learning with known groups:
  - The ℓ_1-ℓ_2 norm

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

  - The ℓ_1-ℓ_2 norm sets to zero non-overlapping groups of variables (as opposed to single variables for the ℓ_1 norm)
- However, the ℓ_1-ℓ_2 norm encodes **fixed/static prior information**, requires to know in advance how to group the variables
- What happens if the set of groups $\mathbf{G}$ is not a partition anymore?

Structured Sparsity (Jenatton et al., 2009a)

- When penalizing by the ℓ_1-ℓ_2 norm

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

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

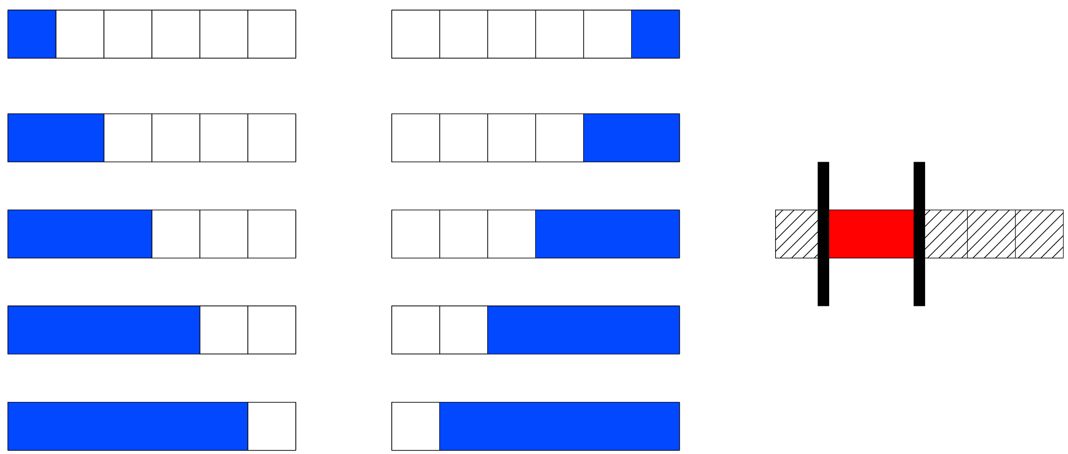
- Intuitively, the zero pattern of w is given by

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

- This intuition is actually true and can be formalized

Examples of set of groups $\mathbf{G}$ (1/3)

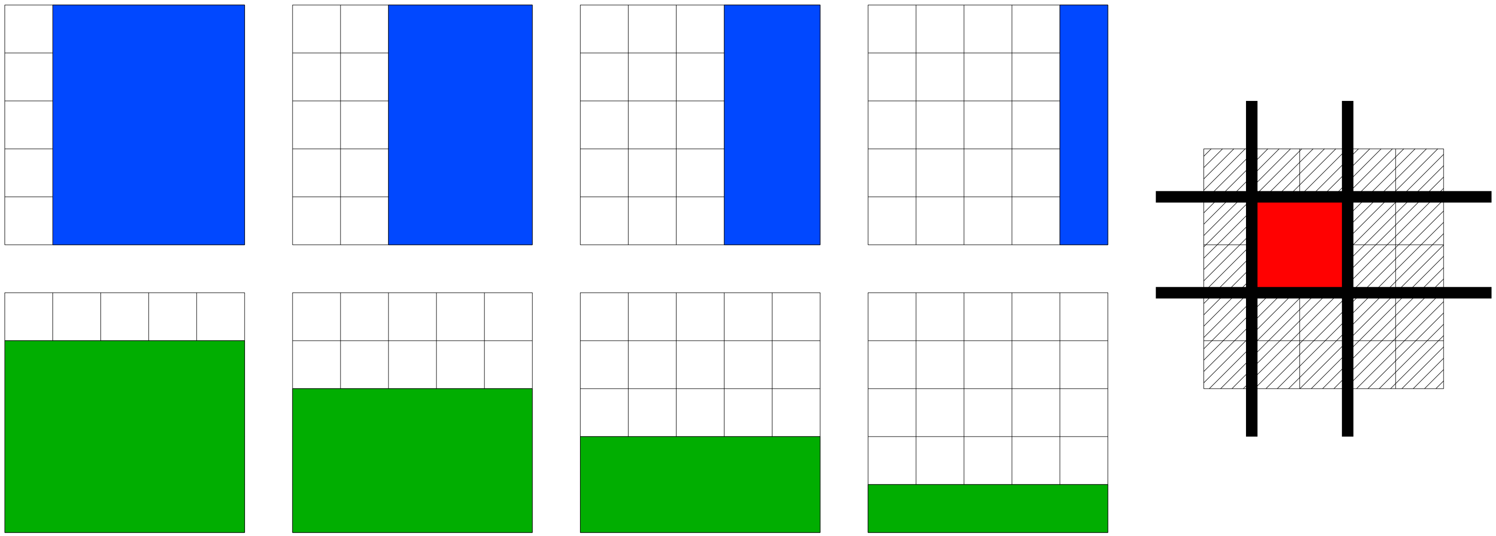
- Selection of contiguous patterns on a sequence, $p = 6$



  - $\mathbf{G}$ 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 $\mathbb{G}$ (2/3)

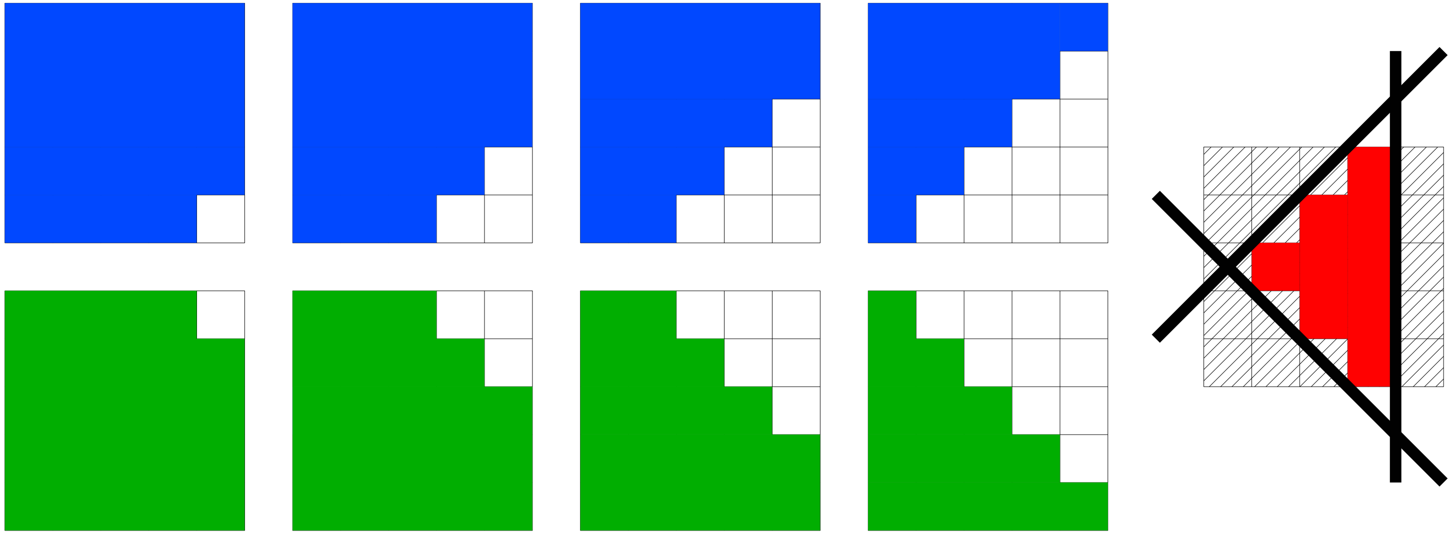
- Selection of rectangles on a 2-D grids, $p = 25$



  - $\mathbf{G}$ 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 $\mathbb{G}$ (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
  - **Applications to sparse PCA / dictionary learning**

Structured matrix factorizations (Bach et al., 2008)

- Data $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \mathbb{R}^{p \times n}$ to decompose in $\mathbf{D} = (\mathbf{d}_1, \dots, \mathbf{d}_k)$

$$\min_{\mathbf{D}, \boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}_i\|_2^2 + \mu \sum_{i=1}^n \|\boldsymbol{\alpha}_i\|_{\bullet} \text{ s.t. } \forall j, \|\mathbf{d}_j\|_{\star} \leqslant 1$$

- $\boldsymbol{\alpha}_i$ decomposition coefficients (or "code"), $\mathbf{d}_j$ dictionary elements

- Two related/equivalent problems:
  - **Sparse PCA = sparse dictionary** (ℓ_1-norm on $\mathbf{d}_j$)
  - **Dictionary learning = sparse decompositions** (ℓ_1-norm on $\boldsymbol{\alpha}_i$) (Olshausen and Field, 1997; Elad and Aharon, 2006)

- **Structured regularization** on $\mathbf{d}_j$ or $\boldsymbol{\alpha}_i$ (Jenatton, Obozinski, and Bach, 2009b; Jenatton, Mairal, Obozinski, and Bach, 2010)

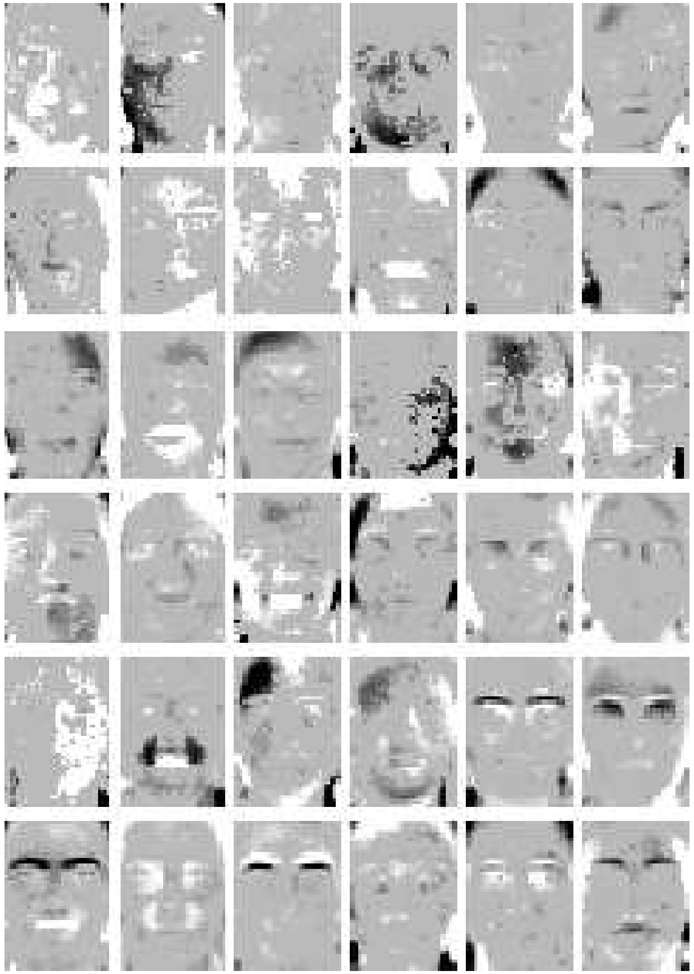
Application to face databases (1/3)

raw data

(unstructured) NMF

- NMF obtains partially local features

Application to face databases (2/3)

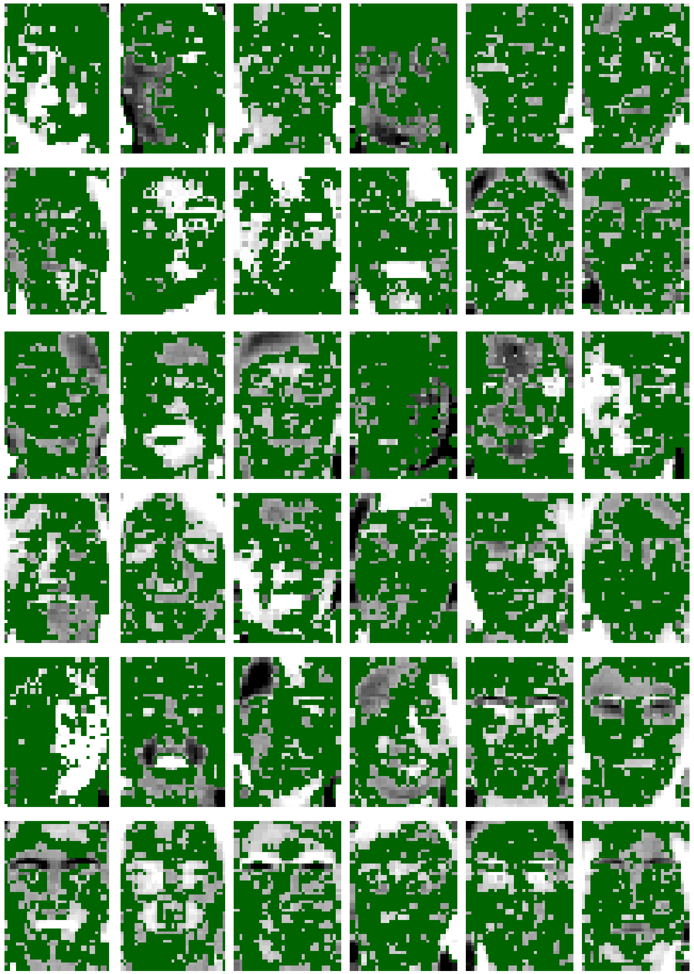


(unstructured) sparse PCA

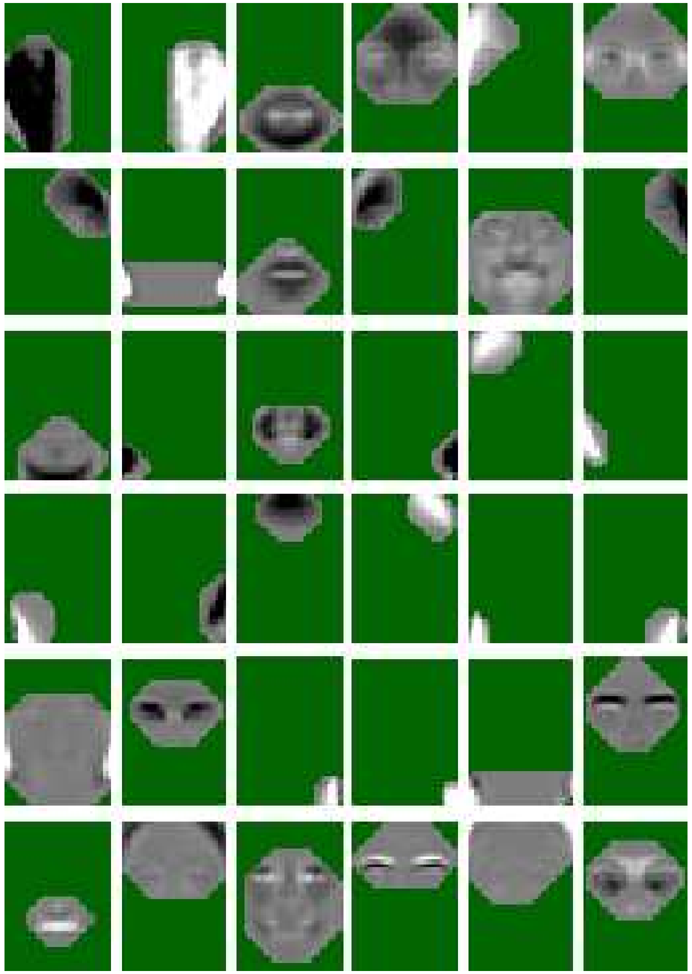
Structured sparse PCA

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

Application to face databases (2/3)



(unstructured) sparse PCA

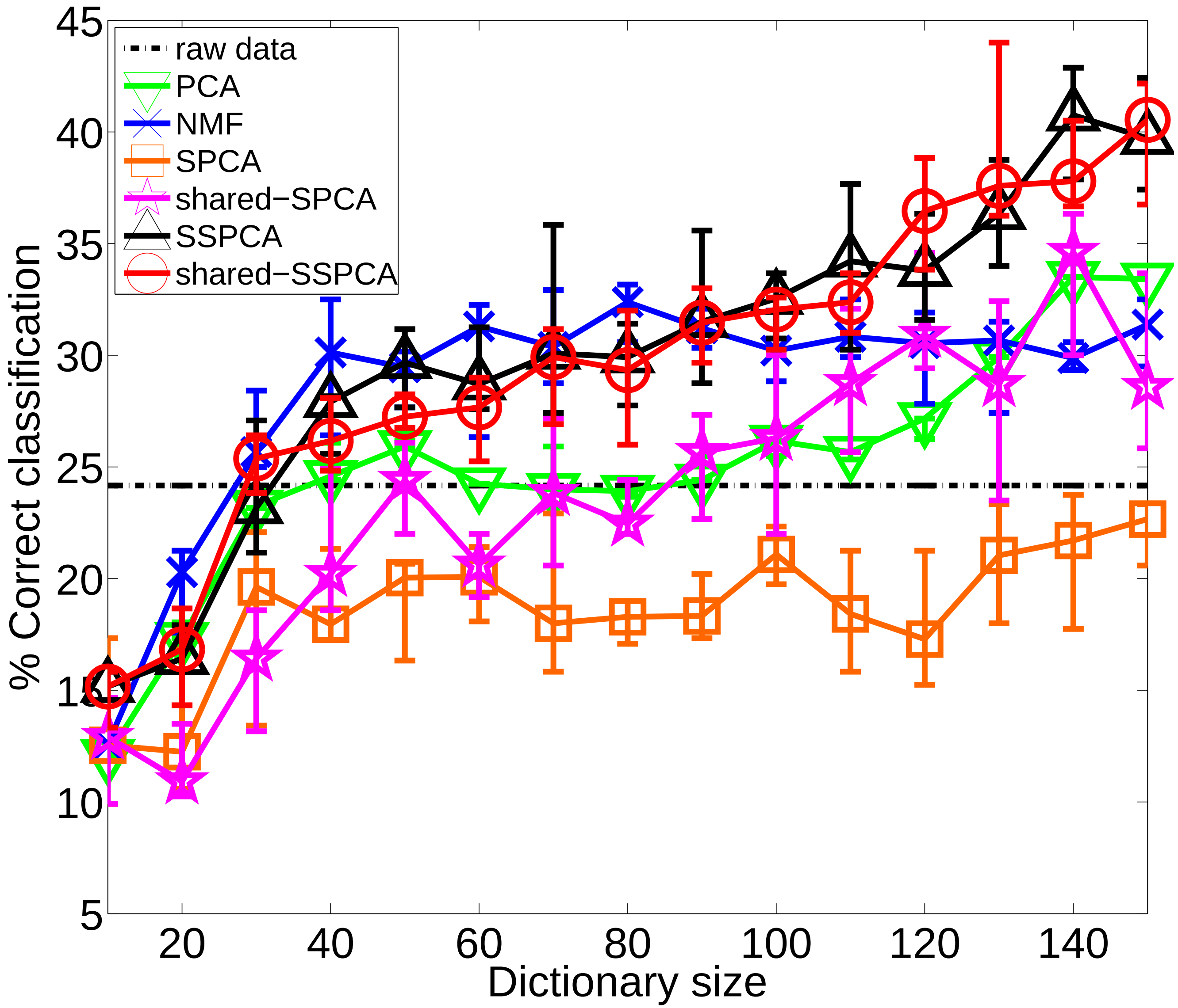


Structured sparse PCA

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

Application to face databases (3/3)

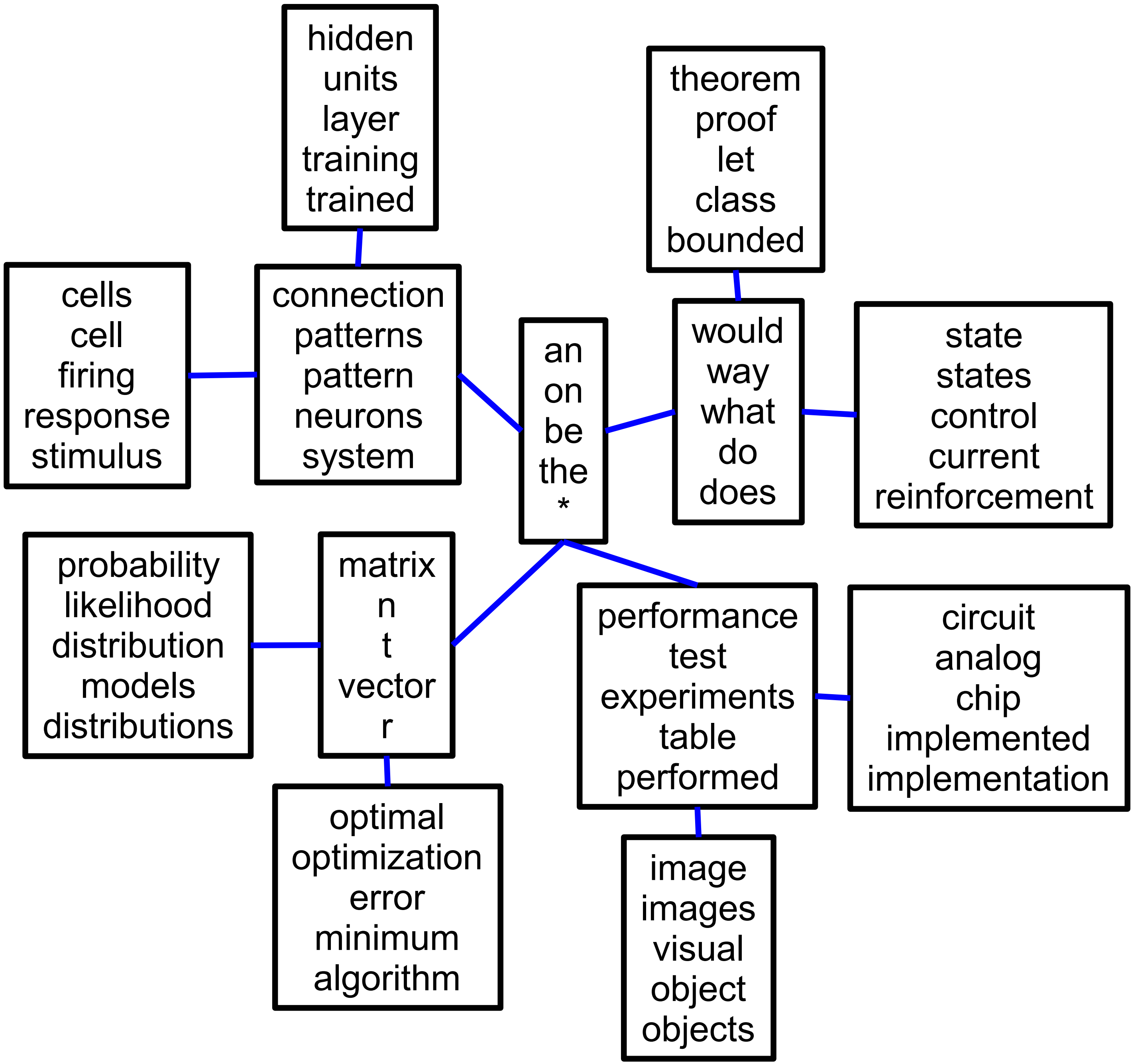
- Quantitative performance evaluation on classification task



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

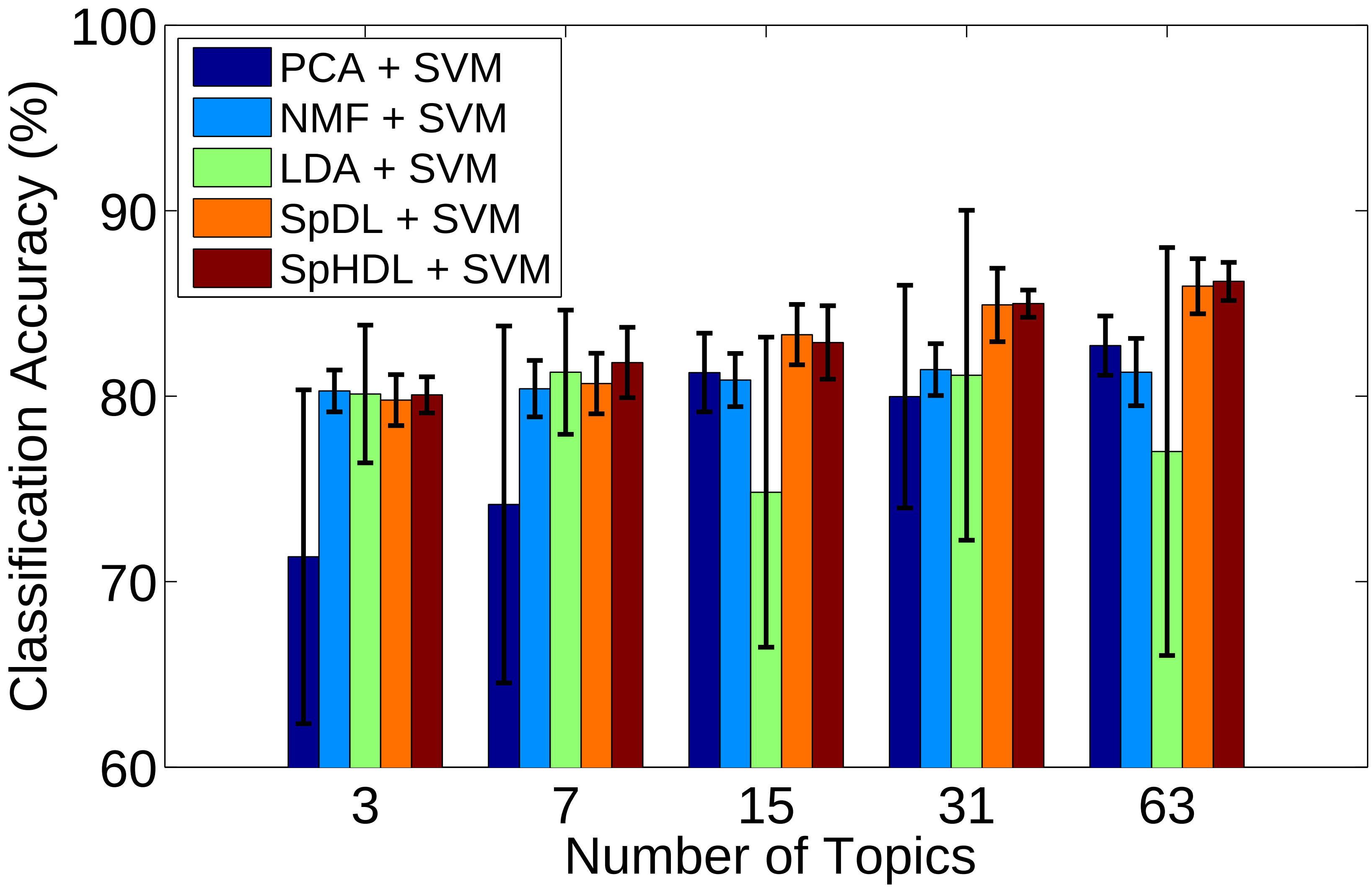
- **Hierarchical norms** on decomposition coefficients $\boldsymbol{\alpha}_i$
  - Equivalent to assume tree-structure among dictionary elements
  - Efficient optimization through proximal methods
- **Modelling of text corpora**
  - Each document is modelled through word counts
  - Low-rank matrix factorization of word-document matrix
- **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



Conclusion

- **Structured sparsity**
  - Sparsity-inducing norms
  - Supervised learning: non-linear variable selection
  - Unsupervised learning: dictionary learning

- **Further/current work**
  - Universal consistency of non-linear variable selection
  - Algorithms
  - Norm design, norms on matrices
  - Applications to computer vision, audio, neuroscience

References

F. Bach. High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning. Technical Report 0909.0844, arXiv, 2009.

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*, pages 1179–1225, 2008b.

F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Adv. NIPS*, 2008c.

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.

P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2008. To appear.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43(1):129–159, 2001. ISSN 0036-1445.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Stat.*, 32:407, 2004.

M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Proc.*, 15(12):3736–3745, 2006.

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

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

R. Jenatton, J.Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, arXiv:0904.3523, 2009a.

R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. Technical report, arXiv:0909.1440, 2009b.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Submitted to ICML*, 2010.

G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applicat.*, 33:82–95, 1971.

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

Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.

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

H. M. Markowitz. The optimization of a quadratic function subject to linear constraints. *Naval Research Logistics Quarterly*, 3:111–133, 1956.

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. *Ann. Stat.*, 2009. to appear.

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.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *Journal of Machine Learning Research*, to appear, 2008.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

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

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Camb. U. P., 2004.

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

S. van de Geer, P. Buhlmann, and S. Zhou. Prediction and variable selection with the adaptive lasso. Technical Report 1001.5176, ArXiv, 2010.

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

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. Technical Report 709, Dpt. of Statistics, UC Berkeley, 2006.

T. T. Wu and K. Lange. Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, 2(1):224–244, 2008.

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.

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

H. Zou. The adaptive Lasso and its oracle properties. *JASA*, 101:1418–1429, 2006.