Sparse Coding and Dictionary Learning for Image Analysis

Part IV: Recent Advances in Computer Vision and New Models

Francis Bach, Julien Mairal, Jean Ponce and Guillermo Sapiro

CVPR'10 tutorial, San Francisco, 14th June 2010

What this part is about

- Learning dictionaries for discriminative tasks...
- ...and adapted to image classification tasks.
- Structured Sparse Models.

Learning dictionaries with a discriminative cost function

Idea:

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

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

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

where

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

"Reconstructive" training

$$\begin{cases} \min_{\mathbf{D}_-} \sum_{i \in S_-} \mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_-) \\ \min_{\mathbf{D}_+} \sum_{i \in S_+} \mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_+) \end{cases}$$

[Grosse et al., 2007], [Huang and Aviyente, 2006], [Sprechmann et al., 2010b] for unsupervised clustering (CVPR '10)

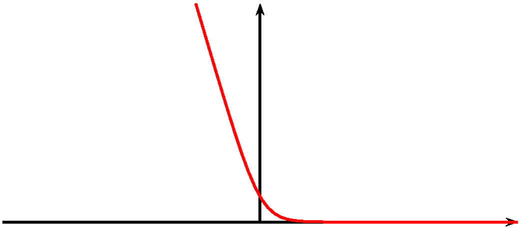
Learning dictionaries with a discriminative cost function

"Discriminative" training

[Mairal, Bach, Ponce, Sapiro, and Zisserman, 2008a]

$$\min_{\mathbf{D}_-,\mathbf{D}_+} \sum_i \mathcal{C}\Big(\lambda z_i \big(\mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_-) - \mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_+)\big)\Big),$$

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



Logistic regression function

Learning dictionaries with a discriminative cost function

Mixed approach

$$\min_{\mathbf{D}_-,\mathbf{D}_+} \sum_i \mathcal{C}\Big(\lambda z_i \big(\mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_-) - \mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_+)\big)\Big) + \mu \mathbf{R}^\star(\mathbf{x}_i, \mathbf{D}_{z_i}),$$

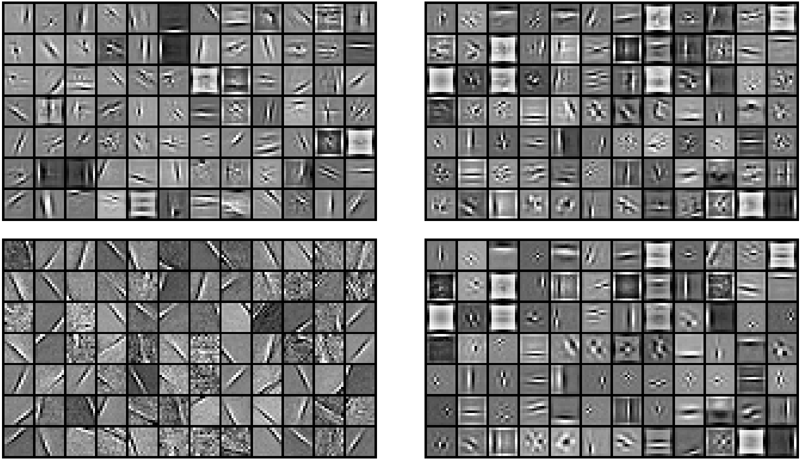
where $z_i \in \{-1, +1\}$ is the label of $\mathbf{x}_i$.

Keys of the optimization framework

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

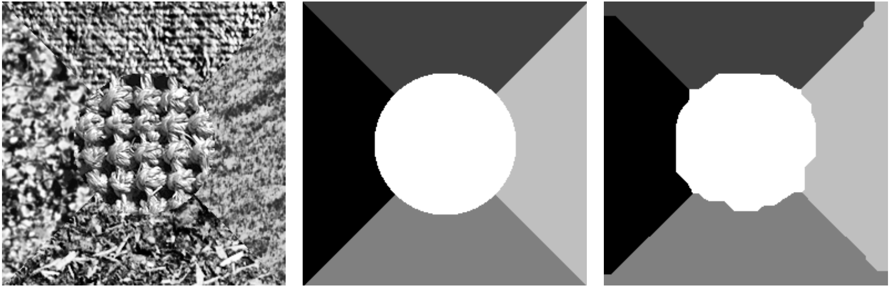
Learning dictionaries with a discriminative cost function

Examples of dictionaries



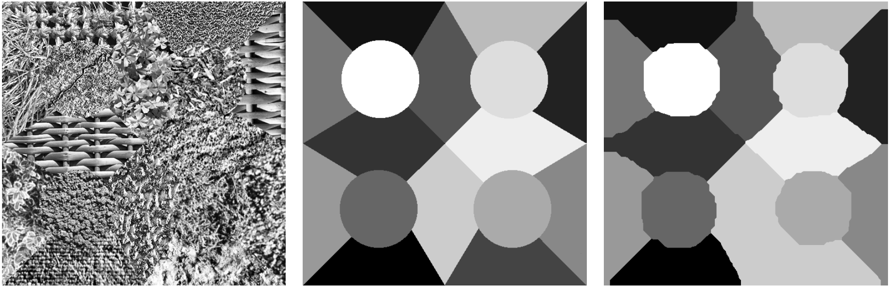
Top: reconstructive, Bottom: discriminative, Left: Bicycle, Right: Background.

Learning dictionaries with a discriminative cost function
Texture segmentation

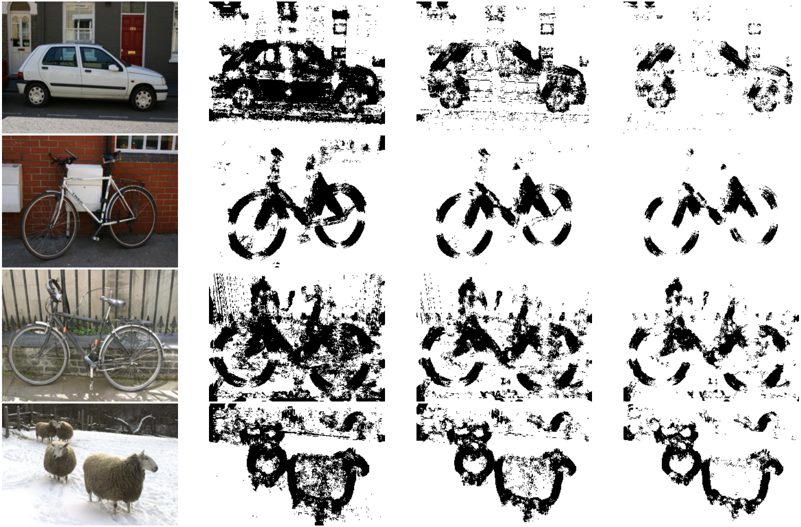


Learning dictionaries with a discriminative cost function

Texture segmentation

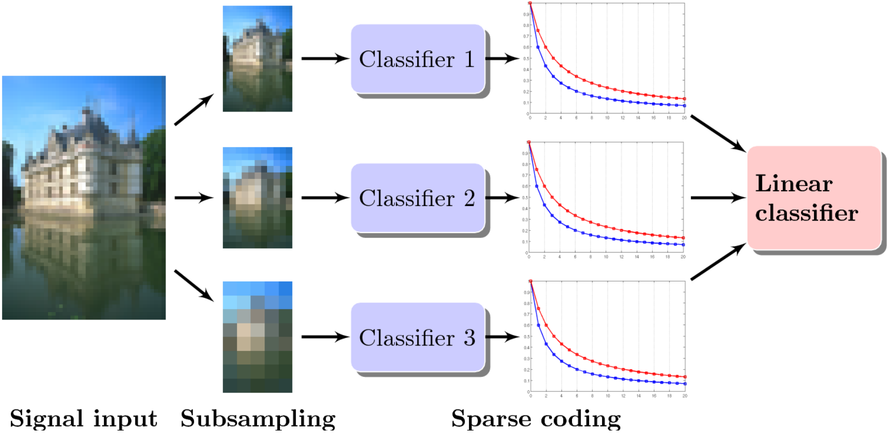


Learning dictionaries with a discriminative cost function
Pixelwise classification



Learning dictionaries with a discriminative cost function

Multiscale scheme

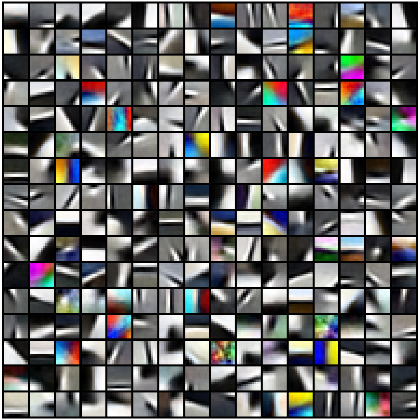


Learning dictionaries with a discriminative cost function

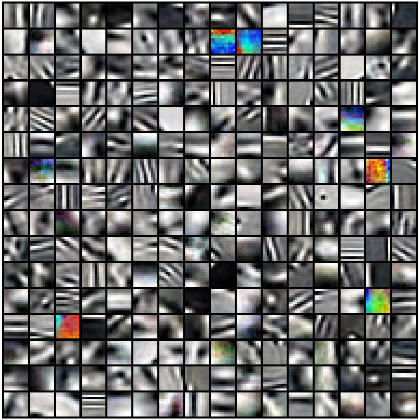
weakly-supervised pixel classification

Application to edge detection and classification

[Mairal, Leordeanu, Bach, Hebert, and Ponce, 2008b]

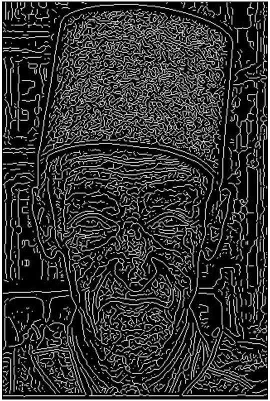


Good edges



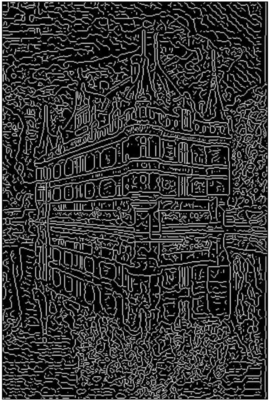
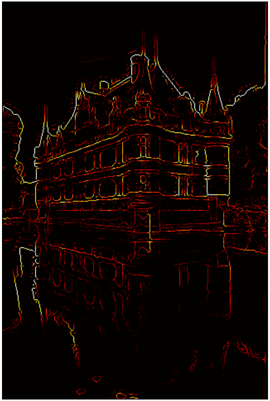
Bad edges

Application to edge detection and classification
Berkeley segmentation benchmark



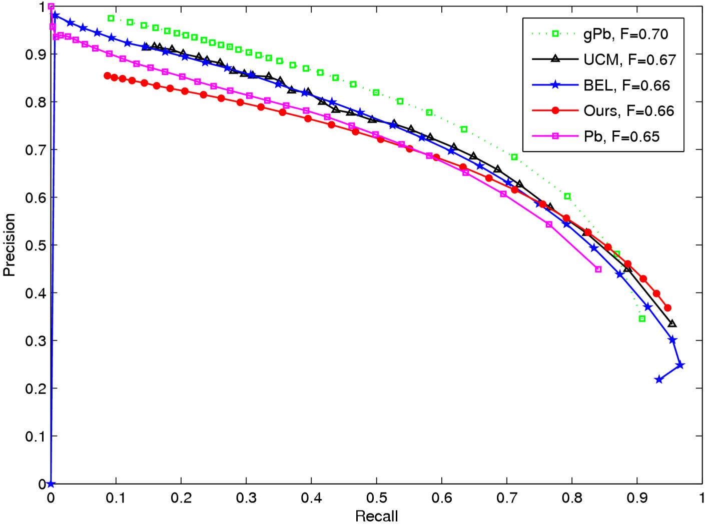
Raw edge detection on the right

Application to edge detection and classification
Berkeley segmentation benchmark

Raw edge detection on the right

Application to edge detection and classification
Berkeley segmentation benchmark

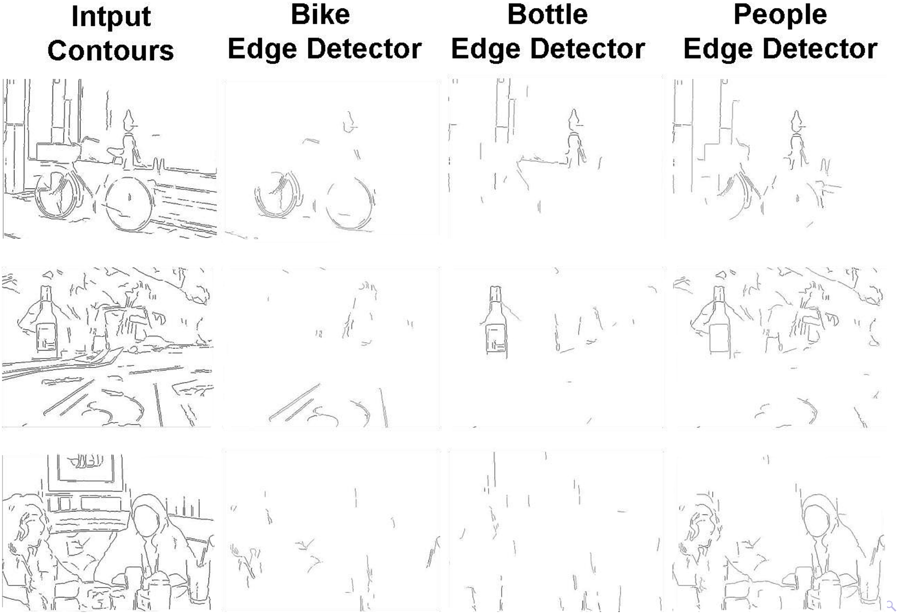


Application to edge detection and classification

Contour-based classifier: [Leordeanu, Hebert, and Sukthankar, 2007]

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

Application to edge detection and classification



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

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

Recognition rates for the same experiment as [Winn et al., 2005] on VOC 2005.

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

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

Digital Art Authentification

Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake

Digital Art Authentification

Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake

Fake

Digital Art Authentification

Data Courtesy of Hugues, Graham, and Rockmore [2009]

Authentic

Fake

Authentic

Image Half-Toning

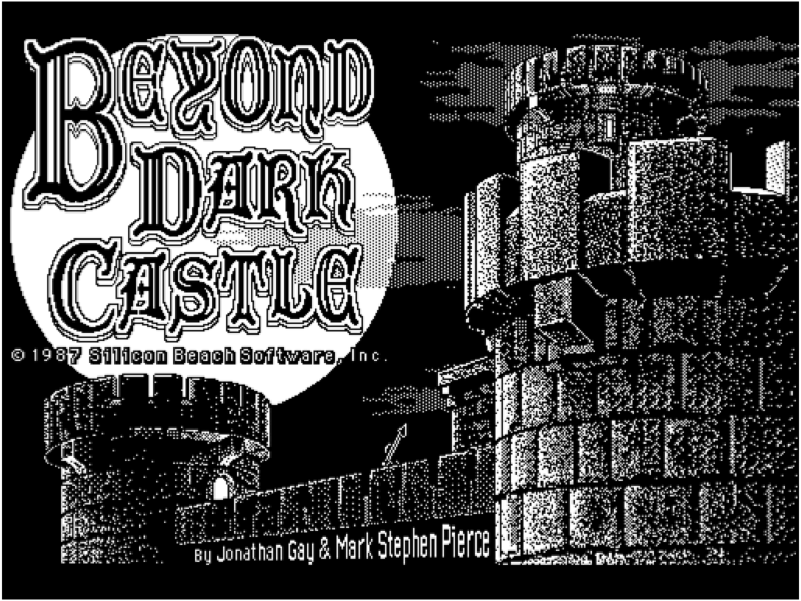
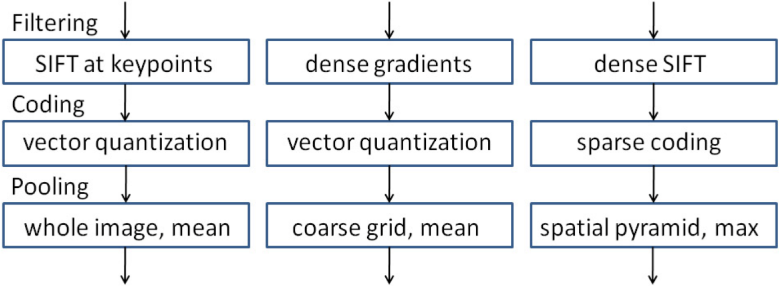


Image Half-Toning

Learning Codebooks for Image Classification



Idea

Replacing Vector Quantization by Learned Dictionaries!

- unsupervised: [Yang et al., 2009]
- supervised: [Boureau et al., 2010, Yang et al., 2010] (CVPR '10)

Learning Codebooks for Image Classification

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

- hard-quantization:

$$\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i, \quad \boldsymbol{\alpha}_i \in \{0,1\}^p \text{ and } \sum_{j=1}^{p} \boldsymbol{\alpha}_i[j] = 1$$

- soft-quantization:

$$\boldsymbol{\alpha}_i[j] = \frac{e^{-\beta\|\mathbf{x}_i - \mathbf{d}_j\|_2^2}}{\sum_{k=1}^{p} e^{-\beta\|\mathbf{x}_i - \mathbf{d}_k\|_2^2}}$$

- sparse coding:

$$\mathbf{x}_i \approx \mathbf{D}\boldsymbol{\alpha}_i, \quad \boldsymbol{\alpha}_i = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

Learning Codebooks for Image Classification

Table from Boureau et al. [2010]

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

	Unsup	Discr
Linear	83.6 ± 0.4	$\mathbf{84.9 \pm 0.3}$
Intersect	84.3 ± 0.5	$\mathbf{84.7 \pm 0.4}$

Yang et al. [2009] have won the PASCAL VOC'09 challenge using this kind of techniques.

Summary so far

- Learned dictionaries are well adapted to model images.
- They can be used to learn dictionaries of SIFT features.
- They are also adapted to discriminative tasks.

Sparse Structured Linear Model

- We focus again on linear models

$$\mathbf{x} \approx \mathbf{D}\boldsymbol{\alpha}.$$

- $\mathbf{x} \in \mathbb{R}^m$, vector of m observations.
- $\mathbf{D} \in \mathbb{R}^{m \times p}$, dictionary or data matrix.
- $\boldsymbol{\alpha} \in \mathbb{R}^p$, loading vector.

Assumptions:

- $\boldsymbol{\alpha}$ is **sparse**, i.e., it has a small support

$$|\Gamma| \ll p, \quad \Gamma = \{j \in \{1, \dots, p\};\ \boldsymbol{\alpha}_j \neq 0\}.$$

- The support, or nonzero pattern, Γ is **structured**:
    - Γ reflects spatial/geometrical/temporal...information.
    - e.g., 2-D grid for features associated to the pixels of an image.

Sparsity-Inducing Norms (1/2)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^p} \overbrace{f(\boldsymbol{\alpha})}^{\text{data fitting term}} + \lambda \underbrace{\psi(\boldsymbol{\alpha})}_{\text{sparsity-inducing norm}}$$

Standard approach to enforce sparsity in learning procedures:

- Regularizing by a **sparsity-inducing norm** ψ.
- The effect of ψ is to set some $\boldsymbol{\alpha}_j$'s to zero, depending on the regularization parameter $\lambda \geq 0$.

The most popular choice for ψ:

- The ℓ_1 norm, $\|\boldsymbol{\alpha}\|_1 = \sum_{j=1}^{p} |\boldsymbol{\alpha}_j|$.
- For the square loss, Lasso [Tibshirani, 1996].
- However, the ℓ_1 norm encodes poor information, just **cardinality**!

Sparsity-Inducing Norms (2/2)

Another popular choice for ψ:

- The ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathcal{G}} \|\boldsymbol{\alpha}_G\|_2 = \sum_{G \in \mathcal{G}} \Big(\sum_{j \in G} \boldsymbol{\alpha}_j^2 \Big)^{1/2}, \text{ with } \mathcal{G} \text{ a } \textbf{partition} \text{ of } \{1, \ldots, p\}.$$

- The ℓ_1-ℓ_2 norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 norm).
- For the square loss, group Lasso [Yuan and Lin, 2006].
- However, the ℓ_1-ℓ_2 norm encodes fixed/static prior information, requires to know in advance how to group the variables !

Questions:

- What happen if the set of groups $\mathcal{G}$ is not a partition anymore?
- What is the relationship between $\mathcal{G}$ and the sparsifying effect of ψ?

Structured Sparsity

[Jenatton et al., 2009]

Case of general overlapping groups.

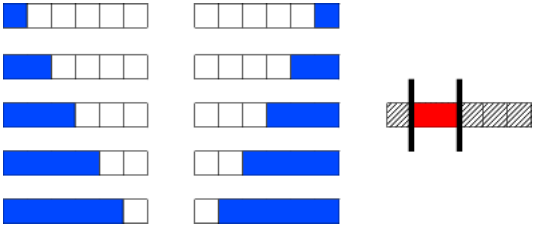
When penalizing by the ℓ_1-ℓ_2 norm,

$$\sum_{G \in \mathcal{G}} \|\boldsymbol{\alpha}_G\|_2 = \sum_{G \in \mathcal{G}} \Big(\sum_{j \in G} \boldsymbol{\alpha}_j^2 \Big)^{1/2}$$

- The ℓ_1 norm induces sparsity at the group level:
  - Some $\boldsymbol{\alpha}_G$'s are set to zero.
- Inside the groups, the ℓ_2 norm does not promote sparsity.
- Intuitively, variables belonging to the same groups are encouraged to be set to zero together.
- Optimization via reweighted least-squares, proximal methods, etc...

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

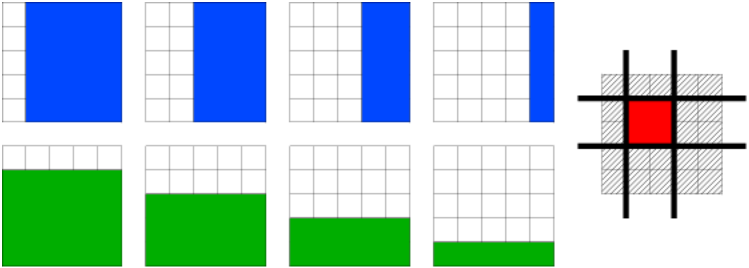
Selection of contiguous patterns on a sequence, $p = 6$.



- $\mathcal{G}$ 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 $\mathcal{G}$ (2/3)

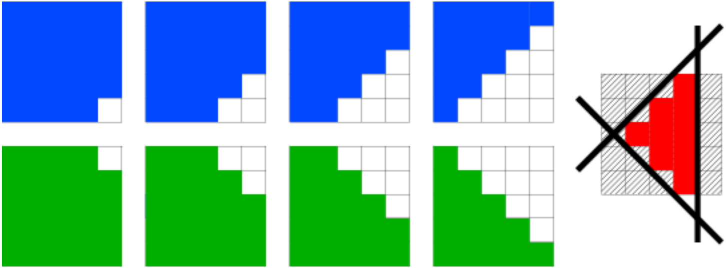
Selection of rectangles on a 2-D grids, $p = 25$.



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

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

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



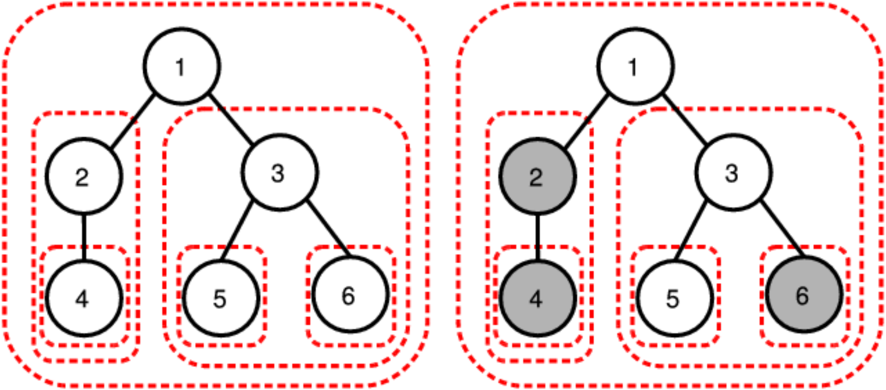
- It is possible to extent such settings to 3-D space, or more complex topologies.

Overview of other work on structured sparsity

- Specific hierarchical structure [Zhao et al., 2009, Bach, 2008].
- **Union-closed** (as opposed to intersection-closed) family of nonzero patterns [Baraniuk et al., 2010, Jacob et al., 2009].
- Nonconvex penalties based on information-theoretic criteria with greedy optimization [Huang et al., 2009].
- Structure expressed through a Bayesian prior, e.g., [He and Carin, 2009].

Hierarchical Dictionaries

[Jenatton, Mairal, Obozinski, and Bach, 2010]

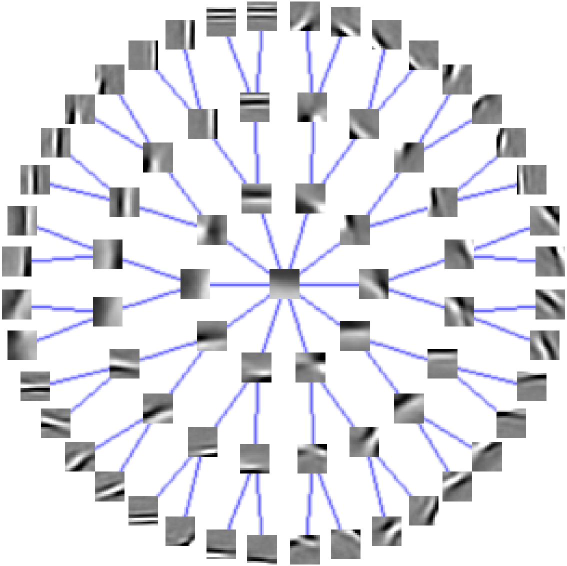


A node can be active only if its **ancestors are active**.
The selected patterns are **rooted subtrees.**

Optimization via efficient proximal methods (same cost as ℓ_1)

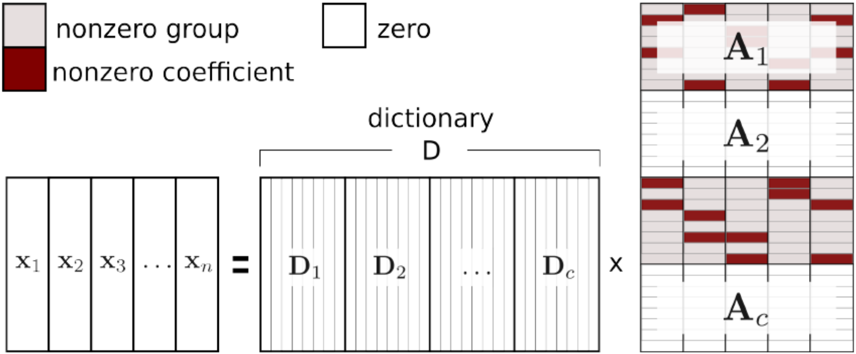
Hierarchical Dictionaries

[Jenatton, Mairal, Obozinski, and Bach, 2010]



Group Lasso + ℓ_1 = Collaborative Hierarchical Lasso

[Sprechmann, Ramirez, Sapiro, and Eldar, 2010a]



Optimization also via proximal methods

Topographic Dictionaries

"Topographic" dictionaries [Hyvarinen and Hoyer, 2001, Kavukcuoglu et al., 2009] are a specific case of dictionaries learned with a structured sparsity regularization for α.

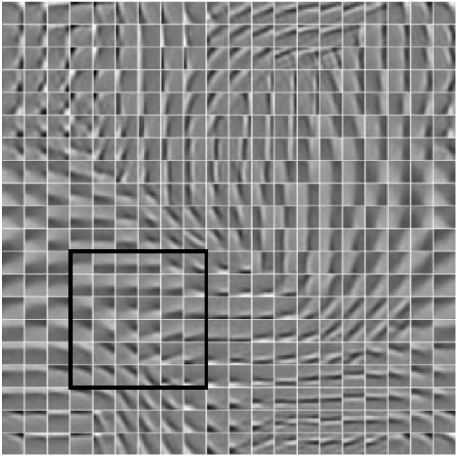


Figure: Image obtained from [Kavukcuoglu et al., 2009]

References I

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

R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. *IEEE Transactions on Information Theory*, 2010. to appear.

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

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

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

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

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

References II

J. M. Hugues, D. J. Graham, and D. N. Rockmore. Quantification of artistic style through sparse coding analysis in the drawings of Pieter Bruegel the Elder. *Proceedings of the National Academy of Science, TODO USA*, 107(4):1279–1283, 2009.

A. Hyvarinen and P. Hoyer. A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images. *Vision Research*, 41(18):2413–2423, 2001.

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

R. Jenatton, J-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical report, 2009. preprint arXiv:0904.3523v1.

R. Jenatton, J. Mairal, G. Obozinski, and F. Bach. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun. Learning invariant features through topographic filter maps. In *Proceedings of CVPR*, 2009.

References III

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

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

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

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

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

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

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

References IV

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

J. Yang, K. Yu, , and T. Huang. Supervised translation-invariant sparse coding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

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

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