Sparse Coding and Dictionary Learning for Image Analysis

Part IV: New sparse models

Francis Bach, Julien Mairal, Jean Ponce and Guillermo Sapiro

ICCV'09 tutorial, Kyoto, 28th September 2009

Sparse Structured Linear Model

- We focus 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

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

- The support, or nonzero pattern, Γ is **structured**:
  - Γ reflects spatial/geometrical/temporal... information about the data.
  - e.g., 2-D grid structure 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,

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

- The ℓ_1-ℓ_2 norm sets to zero **groups of non-overlapping variables** (as opposed to single variables for the ℓ_1 norm).
- For the square loss, group Lasso [Yuan and Lin, 2006, Bach, 2008a].
- 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]

Assumption: $\bigcup_{G \in \mathcal{G}} G = \{1, \ldots, p\}$.

When penalizing by the ℓ_1-ℓ_2 norm,

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

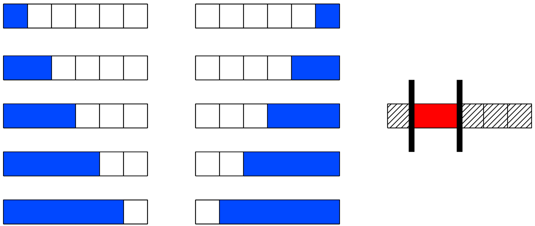
- The ℓ_1 norm induces sparsity at the group level:
  - Some $\boldsymbol{\alpha}_G$'s are set to zero.
- Inside the groups, the ℓ_2 norm does not promote sparsity.
- Intuitively, the zero pattern of w is given by

$$\{j \in \{1, \ldots, p\};\ \boldsymbol{\alpha}_j = 0\} = \bigcup_{G \in \mathcal{G}'} G \quad \text{for some } \mathcal{G}' \subseteq \mathcal{G}.$$

This intuition is actually true and can be formalized (see [Jenatton et al., 2009]).

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

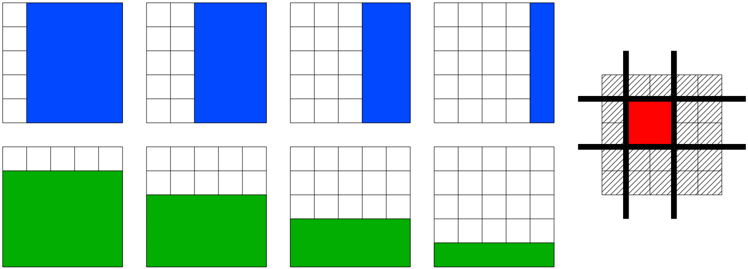
Selection of contiguous patterns on a sequence, $p = 6$.



- $\mathcal{G}$ 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)

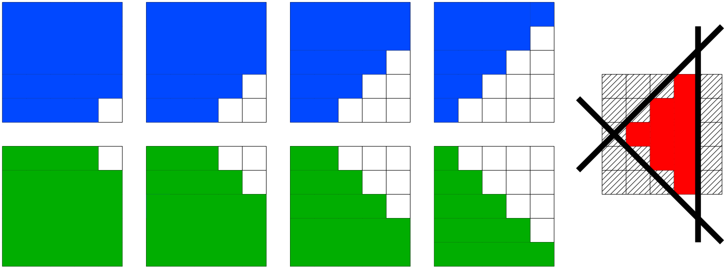
Selection of rectangles on a 2-D grids, $p = 25$.



- $\mathcal{G}$ 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.

Relationship bewteen $\mathcal{G}$ and Zero Patterns (1/2)

[Jenatton et al., 2009]

To sum up, given $\mathcal{G}$, the variables set to zero by ψ belong to

$$\Big\{ \bigcup_{G \in \mathcal{G}'} G;\ \mathcal{G}' \subseteq \mathcal{G} \Big\},$$ i.e., are **a union of elements of $\mathcal{G}$**.

In particular, the set of nonzero patterns allowed by ψ is **closed under intersection**.

Relationship bewteen $\mathcal{G}$ and Zero Patterns (2/2)

[Jenatton et al., 2009]

$\mathcal{G} \rightarrow$ **Zero patterns**:

- We have seen how we can go from $\mathcal{G}$ to the zero patterns induced by ψ (i.e., by generating the **union-closure** of $\mathcal{G}$).

Zero patterns $\rightarrow$ $\mathcal{G}$:

- Conversely, it is possible to go from a desired set of zero patterns to the **minimal** set of groups $\mathcal{G}$ generating these zero patterns.

The latter property is central to our structured sparsity: we can design norms, in form of allowed zero patterns.

Overview of other work on structured sparsity

- Specific hierarchical structure [Zhao et al., 2008, Bach, 2008b].
- **Union-closed** (as opposed to intersection-closed) family of nonzero patterns [Baraniuk et al., 2008, 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].

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 $\boldsymbol{\alpha}$.

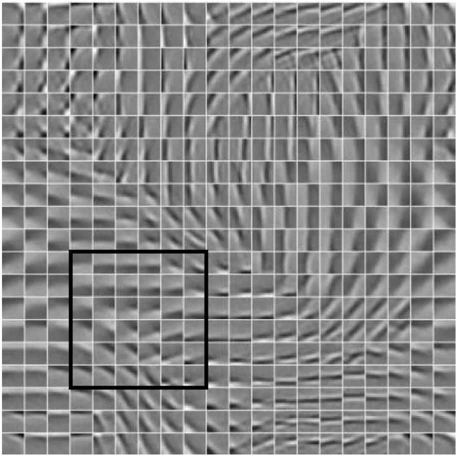


Figure: Image obtained from [Kavukcuoglu et al., 2009]

Dictionary Learning vs Sparse Structured PCA

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

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

- Let us transpose: Sparse Structured PCA (sparse and structured dictionary elements):

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

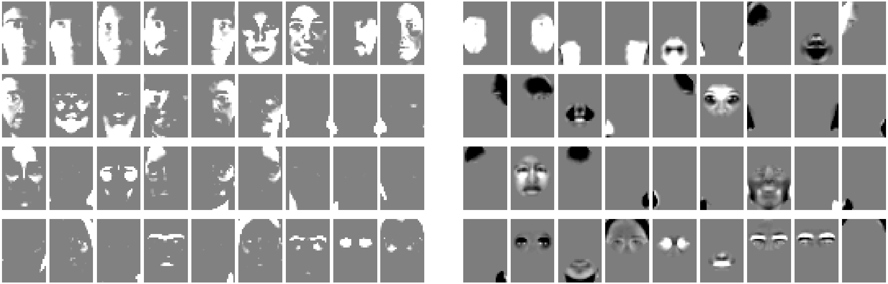
Sparse Structured PCA

We are interested in learning **sparse and structured** dictionary elements:

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

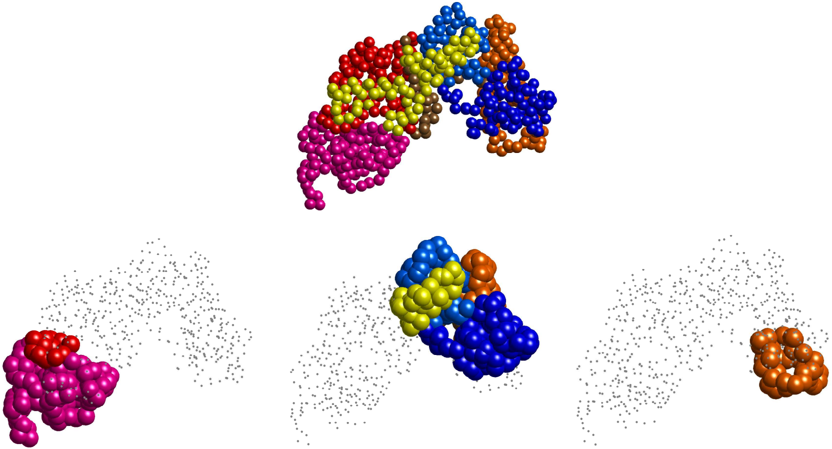
- The columns of $\boldsymbol{\alpha}$ are kept bounded to avoid degenerated solutions.
- The structure of the dictionary elements is determined by the choice of $\mathcal{G}$ (and ψ).

Some results (1/2)



- Application on the AR Face Database [Martinez and Kak, 2001].
- $r = 36$ dictionary elements.
- Left, NMF - Right, our approach.
- We enforce the selection of **convex** nonzero patterns.

Some results (2/2)



- Study the dynamics of protein complexes [Laine et al., 2009].
- Find small **convex** regions in the complex that summerize the dynamics of the whole complex.
- $\mathcal{G}$ represents the 3-D structure of the problem.

Conclusion

- We have shown how sparsity-inducing norms can encode structure.
- The structure prior is expressed in terms of **allowed patterns** by the regularization norm ψ.

Future directions:

- Can be used in many learning tasks, as soon as structure information about the sparse decomposition is known.
- e.g., multi-taks learning or multiple-kernel learning.

References I

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008a.

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

R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde. Model-based compressive sensing. Technical report, 2008. Submitted to IEEE Transactions on Information Theory.

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

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the 26th International Conference on Machine Learning*, 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 overlaps and graph Lasso. In *Proceedings of the 26th International Conference on Machine learning*, 2009.

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

References II

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

E. Laine, A. Blondel, and T. E. Malliavin. Dynamics and energetics: A consensus analysis of the impact of calcium on ef-cam protein complex. *Biophysical Journal*, 96(4):1249–1263, 2009.

A. M. Martinez and A. C. Kak. PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):228–233, 2001.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B*, pages 267–288, 1996.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67, 2006.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 2008. To appear.