Matrix Sparsity - Structured Sparsity

Francis Bach - Guillaume Obozinski
Willow group - INRIA - ENS - Paris

ECML 2010, Barcelona, September 20th
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3 Conclusion
Learning on matrices - Collaborative Filtering (CF)

- Given $n_X$ “movies” $x \in X$ and $n_Y$ “customers” $y \in Y$,
- predict the “rating” $z(x, y) \in Z$ of customer $y$ for movie $x$
- Training data: large $n_X \times n_Y$ incomplete matrix $Z$ that describes the known ratings of some customers for some movies
- Goal: complete the matrix.
Learning on matrices - Multivariate problems

- Multivariate linear regression

\[ Y_{n \times K} = X_{n \times p} \cdot W^*_{p \times K} + \varepsilon_{n \times K} \]

- Multivariate output
- Design matrix
- Coefficient matrix
- Noise

Multiclass classification

\[
\min_{W} \sum_{i=1}^{n} \ell(w^\top \mathbf{x}(i), \ldots, w^\top K \mathbf{x}(i), y(i))
\]

with $y(i) \in \{0, 1\}^K$

One parameter vector $w_k \in \mathbb{R}^p$ per class

$\ell$ is e.g. the multiclass logistic loss
Learning on matrices - Multivariate problems

- Multivariate linear regression

\[ Y_{n \times K} = X_{n \times p} \cdot W^*_{p \times K} + \varepsilon_{n \times K} \]

- Multiclass classification

\[
\min_W \sum_{i=1}^{n} \frac{1}{n} \ell(w_1^T x^{(i)}, \ldots, w_K^T x^{(i)}, y^{(i)})
\]

with

- \( y^{(i)} \in \{0, 1\}^K \)
- One parameter vector \( w_k \in \mathbb{R}^p \) per class
- \( \ell \) is e.g. the multiclass logistic loss
Learning on matrices - Multi-task learning

- $k$ prediction tasks on same covariates $x \in \mathbb{R}^p$
- Each model parameterized by: $w^k \in \mathbb{R}^p$, $1 \leq k \leq K$
Learning on matrices - Multi-task learning

- $k$ prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p$, $1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k(w^k \top x^k_i, y^k_i)$

- Many applications
  - Multi-category classification (one task per class) (Amit et al., 2007)
  - Share parameters between various tasks similar to fixed effect/random effect models (Raudenbush and Bryk, 2002)
Learning on matrices - Multi-task learning

- $k$ prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p, \quad 1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k(w^k \top x^k_i, y^k_i)$
  - All parameters form a matrix:

$$W = [w^1, \ldots, w^K] = \begin{bmatrix} w_1^1 & \cdots & w_1^K \\ \vdots & w_j^k & \vdots \\ w_p^1 & \cdots & w^K_p \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p \times K}$$
Learning on matrices - Multi-task learning

- $k$ prediction tasks on same covariates $x \in \mathbb{R}^p$
  - Each model parameterized by: $w^k \in \mathbb{R}^p$, $1 \leq k \leq K$
  - Empirical risks: $L_k(w^k) = \frac{1}{n} \sum_{i=1}^{n} \ell_k(w^k \top x^k_i, y^k_i)$

- All parameters form a matrix:
  $$W = [w^1, \ldots, w^K] = \begin{bmatrix} w^1_1 & \cdots & w^K_1 \\ \vdots & \ddots & \vdots \\ w^1_p & \cdots & w^K_p \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \end{bmatrix} \in \mathbb{R}^{p \times K}$$

- Many applications
  - Multi-category classification (one task per class) (Amit et al., 2007)
  - Share parameters between various tasks
    - similar to fixed effect/random effect models (Raudenbush and Bryk, 2002)
Learning on matrices - Image denoising

- Simultaneously denoise all patches of a given image
- Example from Mairal et al. (2009b)
Outline

1 Matrix Sparsity
   - Learning on matrices
   - **Forms of sparsity for matrices**
     - Multivariate learning and row sparsity
     - Sparse spectrum
     - Sparse Principal Component Analysis
     - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3 Conclusion
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): $(M_{i1}, \ldots, M_{ip}) = 0$
Two types of sparsity for matrices $M \in \mathbb{R}^{n \times p}$

II - Through a factorization of $M = UV^\top$

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

- **Sparse decomposition**: $U$ sparse

- Same as dictionary learning with notations $M = X$, $V = D$ and $A = U^\top$. 

Sparsity tutorial II, ECML 2010, Barcelona
Outline

1. Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2. Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3. Conclusion
Joint variable selection (Obozinski et al., 2009)

\[ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k(w^k x_i^k, y_i^k) + \lambda \Omega(W) \]

- Joint matrix of predictors \( W = (w_1, \ldots, w_k) \in \mathbb{R}^{p \times k} \):
  \[
  W = [w^1, \ldots, w^K] = \\
  \begin{bmatrix} 
  w^1_1 & \ldots & w^K_1 \\
  \vdots & \ddots & \vdots \\
  w^1_p & \ldots & w^K_p 
  \end{bmatrix} \\
  \begin{bmatrix} 
  w_1 \\
  \vdots \\
  w_p 
  \end{bmatrix} \in \mathbb{R}^{p \times K} \rightarrow W^* 
  
- Select all variables that are relevant to at least one task

\[ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \ell_k(w^k x_i^k, y_i^k) + \lambda \sum_{j=1}^{p} \|w_j\|_2 \]

- Can improve performance over \( \ell_1 \)-regularization (Obozinski et al., 2008; Lounici et al., 2009)
Applications for simultaneous selection

Multi-class image classification (Quattoni et al., 2008)
→ algorithms for the regularization by a sum of $\ell_\infty$-norm ($\ell_1/\ell_\infty$).
→ increase in performance

Multi-class tumor classification based on gene expression data (Obozinski et al., 2009)
→ smaller gene signatures

Source localization in M/EEG inverse problems from several experiments (Gramfort, 2010)
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
     - Sparse Principal Component Analysis
     - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3 Conclusion
Rank constraints and sparsity of the spectrum

Rank

Given a matrix $M \in \mathbb{R}^{n \times p}$

- Singular value decomposition (SVD): $M = U \text{Diag}(s) V^\top$
  where $U, V$ orthogonal, $s \in \mathbb{R}^m_+$ are singular values
- $\text{Rank}(M) = \|s\|_0$
- Rank of $M$ is the minimum size $m$ of all factorizations of $M$ into $M = UV^\top$, $U \in \mathbb{R}^{n \times m}$ and $V \in \mathbb{R}^{p \times m}$
Rank constraints and sparsity of the spectrum

Rank

Given a matrix \( M \in \mathbb{R}^{n \times p} \)

- Singular value decomposition (SVD): \( M = U \text{Diag}(s) V^\top \)
  where \( U, V \) orthogonal, \( s \in \mathbb{R}^m_+ \) are singular values
- \( \text{Rank}(M) = \|s\|_0 \)
- Rank of \( M \) is the minimum size \( m \) of all factorizations of \( M \) into \( M = UV^\top \), \( U \in \mathbb{R}^{n \times m} \) and \( V \in \mathbb{R}^{p \times m} \)

Rank constrained Learning

\[
\min_{W \in \mathbb{R}^{p \times k}} L(W) \quad \text{s.t.} \quad \text{rank}(W) \leq m
\]

Examples:

- Collaborative filtering
- Multi-task learning with task parameters assumed in a low dimensional subspace (Argyriou et al., 2009)
Low-rank via factorization

Reduced-rank multivariate regression

\[
\min_{W} \| Y - XW \|_F^2 \quad \text{s.t. rank}(W) \leq k
\]

- Well studied (Anderson, 1951; Izenman, 1975; Reinsel and Velu, 1998)
- Is solved directly using the SVD (by OLS + SVD + projection)

General factorization

\[
\min_{U \in \mathbb{R}^{p \times m}, V \in \mathbb{R}^{k \times m}} L(UV^T)
\]

- Still non-convex but convex w.r.t. U and V separately
- Optimization by alternating procedures
Trace norm relaxation

With SVD $W = U \text{Diag}(s)V^\top$, $\operatorname{rank}(W) = \|s\|_0 \xrightarrow{\text{Relax}} \|s\|_1$.

- $M \mapsto \|s\|_1$ is actually a \textit{unitary invariant} norm: the trace norm, nuclear norm or unitary norm
- Write it $M \mapsto \|M\|_{\text{tr}}$
- Dual norm to the spectral norm $\|M\|_2 = \|s\|_\infty$

Trace norm regularization

$$\min_{W \in \mathbb{R}^{p \times k}} L(W) + \lambda \|W\|_{\text{tr}}$$

- Convex problem
- Algorithms:
  - Proximal methods
  - Iterated Reweighted Least-Square (Argyriou et al., 2009)
  - Common bottleneck: require iterative SVD
Trace norm and collaborative filtering

$$\min_{M \in \mathbb{R}^{p \times n}} \sum_{(i,j) \in S} \| M_{i,j} - M_{i,j}^0 \|^2_2 + \lambda \| M \|_{tr}$$

- semi-definite program (Fazel et al., 2001)
- see also max-margin approaches to CF (Srebro et al., 2005)
- Statistical results:
  - High-dimensional inference for noisy matrix completion (Srebro et al., 2005; Candès and Plan, 2009)
  - May recover entire matrix from slightly more entries than the minimum of the two dimensions
Outline

1 Matrix Sparsity
   * Learning on matrices
   * Forms of sparsity for matrices
   * Multivariate learning and row sparsity
   * Sparse spectrum
   * **Sparse Principal Component Analysis**
   * Dictionary learning, image denoising and inpainting

2 Structured sparsity
   * Overview
   * Sparsity patterns stable by union
   * Sparse Structured PCA
   * Hierarchical Dictionary Learning

3 Conclusion
Two different views of PCA

Given data matrix $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p}$,
Two different views of PCA

Given data matrix $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p}$,

**Analysis view**

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$
\max_{v \in \mathbb{R}^p} \quad v^T X^T X v \\
\text{s.t.} \quad \|v\|_2 \leq 1
$$

$\rightarrow$ deflate and iterate to obtain more components.

For regular PCA, the two views are equivalent!

Not true if constraints on $u, v$ change
Two different views of PCA

Given data matrix $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p}$,

**Analysis view**

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\max_{v \in \mathbb{R}^p} \quad v^T X^T X \ v$$

s.t. $\|v\|_2 \leq 1$

$\rightarrow$ deflate and iterate to obtain more components.

**Synthesis view**

Find $V = [v_1, \ldots, v_k]$ s.t. $x_i$ have low reconstruction error on $\text{span}(V)$:

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^k u_i v_i^T\|_F^2$$
Two different views of PCA

Given data matrix $X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p}$,

**Analysis view**

Find projection $v \in \mathbb{R}^p$ maximizing variance:

$$\max_{v \in \mathbb{R}^p} v^T X^T X v \quad \text{s.t.} \quad \|v\|_2 \leq 1$$

→ deflate and iterate to obtain more components.

**Synthesis view**

Find $V = [v_1, \ldots, v_k]$ s.t. $x_i$ have low reconstruction error on span($V$):

$$\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^k u_i v_i^T\|_F^2$$
Two different views of PCA

Given data matrix \( X = (x_1^T, \ldots, x_n^T)^T \in \mathbb{R}^{n \times p} \),

**Analysis view**

Find projection \( v \in \mathbb{R}^p \) maximizing variance:

\[
\max_{v \in \mathbb{R}^p} v^T X^T X v \\
\text{s.t.} \quad \|v\|_2 \leq 1
\]

\( \rightarrow \) deflate and iterate to obtain more components.

**Synthesis view**

Find \( V = [v_1, \ldots, v_k] \) s.t. \( x_i \) have low reconstruction error on \( \text{span}(V) \):

\[
\min_{u_i, v_i \in \mathbb{R}^p} \|X - \sum_{i=1}^k u_i v_i^T\|_F^2
\]

For regular PCA, the two views are equivalent!

Not true if constraints on \( u, v \) change.
Sparse PCA - Analysis view

Add sparsity constraint:

\[
\max \quad v^\top X^\top X v \\
\text{subject to } \|v\|_2 = 1, \|v\|_0 \leq k
\]
Sparse PCA - Analysis view

Add sparsity constraint:

$$\max_{\|v\|_2=1, \|v\|_0 \leq k} v^T X^T X v$$

Convex relaxation **DSPCA** (d’Aspremont et al., 2007)

relaxed into $$\max_{\|v\|_2=1, \|v\|_1 \leq k^{1/2}} v^T X^T X v$$

then relaxed into $$\max_{M \succeq 0, \text{tr}(M)=1, 1^T |M|_1 \leq k} \text{tr}(X^T X M)$$, using $$M = vv^T$$.
Sparse PCA - Analysis view

Add sparsity constraint:

$$\max \quad v^T X^T X v$$
$$\|v\|_2 = 1, \|v\|_0 \leq k$$

Convex relaxation **DSPCA** (d’Aspremont et al., 2007)

relaxed into

$$\max \quad v^T X^T X v$$
$$\|v\|_2 = 1, \|v\|_1 \leq k^{1/2}$$

then relaxed into

$$\max \quad \text{tr}(X^T XM), \quad \text{using } M = vv^T.$$ 

- Requires deflation for multiple components (Mackey, 2009)
- More refined convex relaxation (d’Aspremont et al., 2008)
- Analysis of non-convex formulation (Moghaddam et al., 2006)
Sparse PCA - Synthesis view

Find $V = [v_1, \ldots, v_m] \in \mathbb{R}^{p \times n}$ sparse and $U = [u_1, \ldots, u_m] \in \mathbb{R}^{n \times n}$ s.t.

$$
\sum_{i=1}^{n} \left\| x_i - \sum_{j=1}^{m} u_{ij} v_j \right\|_2^2 \text{ is small } \iff \|X - UV^\top\|_F^2, \text{ is small}
$$

Sparse matrix factorization (Witten et al., 2009; Bach et al., 2008)

- Penalize columns $v_i$ of $V$ by the $\ell_1$-norm for sparsity
- Penalize columns $u_i$ of $U$ by the $\ell_2$-norm to avoid trivial solutions

$$
\begin{align*}
\min_{U,V} & \|X - UV^\top\|_F^2 + \frac{\lambda}{2} \sum_{i=1}^{m} \left\{ \|u_i\|_2^2 + \|v_i\|_1^2 \right\} \\
\min_{U,V} & \|X - UV^\top\|_F^2 + \lambda \sum_{i} \|u_i\|_2 \|v_i\|_1 \\
\min_{U,V} & \|X - UV^\top\|_F^2 + \lambda \sum_{i} \|v_i\|_1 \text{ s.t. } \|u_i\|_2 \leq 1
\end{align*}
$$

yield the same solutions for $u_j v_j^\top$ (Bach et al., 2008).
Efficient algorithms for sparse matrix factorization

Focus on previous formulation:

\[
\min_{U, V} \|X - UV^\top\|_F^2 + \lambda \sum_i \|v_j\|_1 \quad \text{s.t. } \|u_j\|_2 \leq 1
\]

- Problem is convex in \( U \) and \( V \) separately, but not jointly.
  - Alternating scheme: optimize \( U \) and \( V \) in turn.
- Even better: use simple column updates (Lee et al., 2007; Witten et al., 2009):

  With \( \tilde{X} = X - \sum_{j' \neq j} u_j v_j^\top \), we have

  either \( u_j \leftarrow \frac{\tilde{X} v_j}{\|\tilde{X} v_j\|} \) or \( v_j \leftarrow \arg\min_v \|X^\top u_j - v\|_2^2 + \lambda \|v\|_1 \)

  - requires no matrix inversion
  - can take advantage of efficient algorithms for Lasso
  - can use warm start + active sets
“Sparse projector” (Zou et al., 2006)

Find $\tilde{V} = [\tilde{v}_1, \ldots, \tilde{v}_m] \in \mathbb{R}^{p \times n}$ and $V = [v_1, \ldots, v_m] \in \mathbb{R}^{p \times n}$ such that

$$\min_{\tilde{V}, V} \sum_{i=1}^{n} \|x_i - \tilde{V} V^\top x_i\|_2^2 + \lambda_1 \|V\|_1 + \lambda_2 \|V\|_F^2$$

such that $\tilde{V}^\top \tilde{V} = I_p$

- The data should be reconstructed from sparse projections
- Non-convex formulation $\rightarrow$ alternating minimization
Sparse PCA vs Dictionary Learning a.k.a. Sparse Coding

In signal processing $X^T = V U^T = D \alpha$

Sparse PCA

- e.g. microarray data
- sparse dictionary
- (Witten et al., 2009; Bach et al., 2008)

Dictionary Learning

- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3 Conclusion
Dictionary Learning

\[
\min_{A \in \mathbb{R}^{k \times n}, \ D \in \mathbb{R}^{p \times k}} \sum_{i=1}^{n} \left( \| x_i - D \alpha_i \|_2^2 + \lambda \| \alpha_i \|_1 \right) \quad \text{s.t.} \quad \forall j, \| d_j \|_2 \leq 1.
\]

- As before not jointly convex but convex in each \( d_j \) and \( \alpha_j \)
- Alternating scheme becomes slow for large signal databases ...

\[
\rightarrow \quad \text{use \textit{Stochastic Optimization / Online learning} (Mairal et al., 2009a)}
\]

- can handle potentially infinite datasets
- can adapt to dynamic training sets
Inpainting a 12-Mpixel photograph
Inpainting a 12-Mpixel photograph
Inpainting a 12-Mpixel photograph

If a dread of morning came

big mountains from the valley

Inpainting a 12-Mpixel photograph
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
     - Sparsity patterns stable by union
     - Sparse Structured PCA
     - Hierarchical Dictionary Learning

3 Conclusion
Sparsity with Structure

Notion emerged very recently through the work of several authors: Yuan and Lin (2006), Zhao et al. (2009), Baraniuk et al. (2008), Bach (2008), Jacob et al. (2009), Jenatton et al. (2009), Jenatton et al. (2010b), He and Carin (2009), Huang et al. (2009).

The support is sparse but we have prior information about its structure.

- The variables should be selected in groups.
- The variables lie in a hierarchy.
- The variables lie on a graph or network and the support should be localized or densely connected on the graph.
- The variables are pixels of an image and form rectangles or convex shapes.
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
     - Sparsity patterns stable by union
     - Sparse Structured PCA
     - Hierarchical Dictionary Learning

3 Conclusion
## Biological markers for cancer

### Metastasis prognosis: Predict if a tumor will produce metastases.

<table>
<thead>
<tr>
<th>Gene expression in tumor</th>
<th>Metastasis?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Gene expression images]</td>
<td>![Checkmark]</td>
</tr>
<tr>
<td>![Gene expression images]</td>
<td>![Checkmark]</td>
</tr>
<tr>
<td>![Gene expression images]</td>
<td>![X]</td>
</tr>
<tr>
<td>![Gene expression images]</td>
<td>![Question mark]</td>
</tr>
</tbody>
</table>

Can we predict metastasis and identify few predictive genes?
Biological pathways as relevant groups of genes

Predictive genes are naturally grouped in *biological pathways*
- Correspond to genes participating in same biological mechanisms
- Contain often very correlated genes
- The pathways form overlapping groups
- Ultimately relevant to the biologist

⇒ Instead of selecting genes individually, select entire pathways.

The support is a **union of overlapping groups**.
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in G} \|w_g\|

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in G} \|w_g\|

The support obtained is an intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2009))
Not a union of groups
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in G} \|w_g\|$

The support obtained is an intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2009)) Not a union of groups
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

**Lasso:** $\Omega(w) = \sum_i |w_i|$

**Group Lasso (Yuan and Lin, 2006):** $\Omega(w) = \sum_{g \in G} \|w_g\|$

The support obtained is an intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2009)). Not a union of groups.
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in G} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in G} \|w_g\|
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$  

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in G} \|w_g\|$  

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in G} \|w_g\|$
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

**Lasso:** $\Omega(w) = \sum_i |w_i|$  

**Group Lasso (Yuan and Lin, 2006):** $\Omega(w) = \sum_{g \in G} \|w_g\|$  

**Group Lasso when groups overlap:** $\Omega(w) = \sum_{g \in G} \|w_g\|$
Sparsity patterns induced for $L(w) + \lambda \Omega(w)$

Lasso: $\Omega(w) = \sum_i |w_i|$

Group Lasso (Yuan and Lin, 2006): $\Omega(w) = \sum_{g \in G} \|w_g\|$

Group Lasso when groups overlap: $\Omega(w) = \sum_{g \in G} \|w_g\|$

The support obtained is

- An intersection of the complements of the groups set to 0 (cf. Jenatton et al. (2009))
- Not a union of groups
New formulation (Jacob, Obozinski and Vert (2009))

Introducing latent variables $v_g$:

\[
\begin{cases}
\min_{w,v} L(w) + \lambda \sum_{g \in G} \|v_g\|_2 \\
w = \sum_{g \in G} v_g \\
\text{supp}(v_g) \subseteq g.
\end{cases}
\]

Properties

- Resulting support is a union of groups in $G$.
- Possible to select one variable without selecting all the groups containing it.
A new “overlap” norm

Equivalent reformulation

\[
\min_{w,v} L(w) + \lambda \sum_{g \in G} \|v_g\|_2 = \min_w L(w) + \lambda \Omega_{\text{overlap}}(w)
\]

with

\[
\Omega_{\text{overlap}}(w) \triangleq \min_v \sum_{g \in G} \|v_g\|_2
\]

\[
= \min_{v} \sum_{g \in G} \|v_g\|_2
\]

\[
w = \sum_{g \in G} v_g
\]

\[
\text{supp}(v_g) \subseteq g.
\]

(*)
A new “overlap” norm

Equivalent reformulation

\[
\begin{align*}
\min_{w,v} & \quad L(w) + \lambda \sum_{g \in G} \| v_g \|_2 \\
\text{s.t.} & \quad w = \sum_{g \in G} v_g \\
& \quad \text{supp}(v_g) \subseteq g.
\end{align*}
\]

with

\[
\Omega_{overlap}(w) \triangleq \min_v \sum_{g \in G} \| v_g \|_2 \\
\text{s.t.} & \quad w = \sum_{g \in G} v_g \\
& \quad \text{supp}(v_g) \subseteq g.
\]

\(\Omega_{overlap}(w)\) is a norm of \(w\).
A new “overlap” norm

Equivalent reformulation

\[
\begin{align*}
\min_{w,v} L(w) + \lambda \sum_{g \in G} \|v_g\|_2 \\
w = \sum_{g \in G} v_g \\
\text{supp}(v_g) \subseteq g.
\end{align*}
\]

with

\[
\Omega_{\text{overlap}}(w) \triangleq \begin{cases} \\
\min_v \sum_{g \in G} \|v_g\|_2 \\
w = \sum_{g \in G} v_g \\
\text{supp}(v_g) \subseteq g.
\end{cases}
\]

\* \* \* \* \* \* \*

\(\Omega_{\text{overlap}}(w)\) is a norm of \(w\).
Overlap and group unity balls

Balls for $\Omega^G_{\text{group}}(\cdot)$ (middle) and $\Omega^G_{\text{overlap}}(\cdot)$ (right) for the groups $G = \{\{1, 2\}, \{2, 3\}\}$ where $w_2$ is represented as the vertical coordinate. Left: group-lasso ($G = \{\{1, 2\}, \{3\}\}$), for comparison.
Results

Breast cancer data

- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\ell_1$</th>
<th>$\Omega^G_{\text{overlap}}(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Misclassification error</td>
<td>0.38 ± 0.04</td>
<td>0.36 ± 0.03</td>
</tr>
<tr>
<td>Number of pathways involved</td>
<td>148, 58, 183</td>
<td>6, 5, 78</td>
</tr>
</tbody>
</table>
Outline

1 Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2 Structured sparsity
   - Overview
   - Sparsity patterns stable by union
     - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3 Conclusion
Sparse PCA / Dictionary Learning

Sparse PCA

\[ X^T = D \alpha \]

- e.g. microarray data
- sparse dictionary
- (Witten et al., 2009; Bach et al., 2008)

Dictionary Learning

\[ X^T = D \alpha \]

- e.g. overcomplete dictionaries for natural images
- sparse decomposition
- (Elad and Aharon, 2006)

Other constraints
Structured matrix factorizations - Many instances

- \( M = UV^\top, \quad U \in \mathbb{R}^{n \times m} \text{ and } V \in \mathbb{R}^{p \times m} \)

- **Structure on** \( U \text{ and/or } V 
  
  - Low-rank: \( U \text{ and } V \) have few columns
  - Dictionary learning / sparse PCA: \( U \text{ or } V \) has many zeros
  - Clustering (\( k \)-means): \( U \in \{0, 1\}^{n \times m}, \quad U1 = 1 \)
  - Pointwise positivity: non negative matrix factorization (NMF)
  - Specific patterns of zeros
  - etc.

- **Many applications**
  - e.g., source separation (Févotte et al., 2009), exploratory data analysis
From SPCA to SSPCA

Sparse PCA:

\[
\min_{A \in \mathbb{R}^{k \times n}, \, D \in \mathbb{R}^{p \times k}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|^2_2 + \lambda \sum_{j=1}^{k} \|d_j\|_1 \quad \text{s.t.} \quad \forall j, \|\alpha_j\|_2 \leq 1.
\]

Sparse structured PCA

\[
\min_{A \in \mathbb{R}^{k \times n}, \, D \in \mathbb{R}^{p \times k}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|^2_2 + \lambda \sum_{j=1}^{k} \Omega(d_j) \quad \text{s.t.} \quad \forall j, \|\alpha_j\|_2 \leq 1.
\]

- No orthogonality
- Not jointly convex but convex in each \(d_j\) and \(\alpha_j\)
- \(\Rightarrow\) efficient block-coordinate descent algorithms
A basis to decompose faces?  
Eigenfaces  
Find parts?  
Localized components  
NMF (Lee and Seung, 1999)
Faces

NMF
Rectangular supports

\[ \Omega(d) = \sum_{g \in G} \|d_g\|_2: \text{Selection of rectangles on the 2D-grid.} \]

- \( G \) 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
General “convex” supports

- $\Omega(\mathbf{d}) = \sum_{g \in G} \| \mathbf{d}_g \|_2$: Selection of “convex” patterns on a 2-D grids.

- It is possible to extend such settings to 3-D space, or more complex topologies.
Learning **sparse and structured dictionary elements:**

\[
\min_{A \in \mathbb{R}^{k \times n}, D \in \mathbb{R}^{p \times k}} \sum_{i=1}^{n} \|x_i - D\alpha_i\|_2^2 + \lambda \sum_{j=1}^{p} \Omega(d_j) \text{ s.t. } \forall i, \|\alpha_i\|_2 \leq 1
\]

- Structure of the dictionary elements determined by the choice of \(G\) (and thus \(\Omega\))
- Efficient learning procedures through *variational formulation*.

Reweighted \(\ell_2\):

\[
\sum_{g \in G} \|y_g\|_2 = \min_{\eta_g \geq 0, g \in G} \frac{1}{2} \sum_{g \in G} \left\{ \frac{\|y_g\|_2^2}{\eta_g} + \eta_g \right\}
\]
Faces

- **AR Face database**
- 100 individuals (50 W/50 M)
- For each
  - 14 non-occluded
  - 12 occluded
  - lateral illuminations
  - reduced resolution to $38 \times 27$ pixels
Decomposition of faces

SPCA

SSPCA
Decomposition of faces II

SPCA

SSPCA
k-NN classification based on decompositions
Outline

1. Matrix Sparsity
   - Learning on matrices
   - Forms of sparsity for matrices
   - Multivariate learning and row sparsity
   - Sparse spectrum
   - Sparse Principal Component Analysis
   - Dictionary learning, image denoising and inpainting

2. Structured sparsity
   - Overview
   - Sparsity patterns stable by union
   - Sparse Structured PCA
   - Hierarchical Dictionary Learning

3. Conclusion
Hierarchical Topic Models for text corpora

Flat Topic Model

Each document $x_j$ is modeled through word counts:

$$x_{ij} = \text{nb of occurrences of word } i \text{ in document } j, \quad x_j^\top 1 = n_j,$$

$\theta=$topic proportions, $D=$topic word frequencies

Model $x_j$ as:

$$x_j \sim M(D\theta, n_j)$$

- Low-rank matrix factorization of word-document matrix
- Multinomial PCA (Buntine and Perttu, 2003)
- Bayesian approach: Latent Dirichlet Allocation (Blei et al., 2003)

Hierarchical Model: Organise the topics in a tree?

- Previous approaches: non-parametric Bayesian methods
- Can we obtain a similar model with \textbf{structured} matrix factorization?
Hierarchical Norm

(Jenatton, Mairal, Obozinski and Bach, 2010)

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

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

Efficient Optimization

\[
\min_{A \in \mathbb{R}^{k \times n}, D \in \mathbb{R}^{p \times k}} \sum_{i=1}^{n} \left\| x_i - D \alpha_i \right\|_2^2 + \lambda \Omega(\alpha_i) \text{ s.t. } \forall j, \left\| d_j \right\|_2 \leq 1.
\]

- Proximal methods
- Requires solving \( \min_{\alpha \in \mathbb{R}^p} \frac{1}{2} \left\| y - \alpha \right\|_2^2 + \lambda \Omega(\alpha) \)
- Can we do this for tree-structured norms?
Tree-structured groups

Proposition (Jenatton et al., 2010a)

- If $\mathcal{G}$ is a tree-structured set of groups, i.e.,
  \[ g \cap g' \neq \emptyset \implies g \subset g' \text{ or } g' \subset g, \]

- If the groups are sorted from the leaves to the root,
- If $\Pi_g$ is
  - the proximal operator $w_g \mapsto \text{Prox}_{\mu \parallel \cdot \parallel_q} (w_g)$ on the subspace corresponding to group $g$ and
  - the identity on the orthogonal

Then the proximal operator for $\Omega$ is the composition of all operators from the leaves to the root.

\[ \text{Prox}_{\mu \Omega} = \Pi_{g_m} \circ \ldots \circ \Pi_{g_1}. \] (1)

→ Tree-structured regularization: Efficient linear time algorithm
Tree of Topics

NIPS abstracts
- 1714 documents
- 8274 words
Classification based on topics

Comparison on predicting newsgroup article subjects

- 20 newsgroup articles (1425 documents, 13312 words)
Hierarchical dictionary for image patches
Summary

Sparse linear estimation with $\ell_1$-regularization
- Convex optimization and algorithms
- Theoretical results

Group sparsity
- Block norm
- Multiple Kernel Learning

Matrix Sparsity
- Row sparsity for Multivariate Learning
- Low rank, SPCA and Dictionary Learning

Structured Sparsity
- Overlapping groups and supports stable by union or intersection
- SSPCA and Hierarchical Dictionary Learning
Conclusions

Sparse methods are not limited to regression

High-dimension

- Sparse methods perform well with very many predictors:
- Can algorithms tackle \( \log(p) = o(n) \) for \( n > 100 \)?

Performance

- Inducing sparsity does not always improve predictive performance
- Sparsity is a prior
- “Problems are sparse if you look at them the right way”

Capture structure

- Structured sparsity enhances interpretability
- Norm design: make the right norm for your problem
Acknowledgements

Rodolphe Jenatton
Julien Mairal
Laurent Jacob
Jean Ponce

Jean-Philippe Vert
Ben Taskar
Martin Wainwright
Michael Jordan
Acknowledgements

Thank you!
References I


References II


References III


