Lecture 4: Optimization in Unsupervised Learning

Laurent El Ghaoui

EECS and IEOR Departments
UC Berkeley

Spring seminar TRANSP-OR, Zinal, Jan. 16-19, 2012
Outline

Overview of Unsupervised Learning
  Unsupervised learning models
  Matrix facts

Principal Component Analysis
  Motivations
  Variance maximization
  Deflation
  Factor models
  Example

Sparse PCA
  Basics
  SAFE
  Relaxation
  Algorithms
  Examples
  Variants

Sparse Covariance Selection
  Sparse graphical models
  Penalized maximum-likelihood
  Example

References
Outline

Overview of Unsupervised Learning
Unsupervised learning models
Matrix facts

Principal Component Analysis
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Relaxation
Algorithms
Examples
Variants

Sparse Covariance Selection
Sparse graphical models
Penalized maximum-likelihood
Example

References
What is unsupervised learning?

In unsupervised learning, we are given a matrix of data points $X = [x_1, \ldots, x_m]$, with $x_i \in \mathbb{R}^n$; we wish to learn some condensed information from it.

**Examples:**

- Find one or several direction of maximal variance.
- Find a low-rank approximation or other structured approximation.
- Find correlations or some other statistical information (e.g., graphical model).
- Find clusters of data points.
The empirical covariance matrix

**Definition**

Given \( p \times n \) data matrix \( A = [a_1, \ldots, a_m] \) (each row representing say a log-return time-series over \( m \) time periods), the **empirical covariance matrix** is defined as the \( p \times p \) matrix

\[
S = \frac{1}{m} \sum_{i=1}^{m} (a_i - \hat{a})(a_i - \hat{a})^T, \quad \hat{a} := \frac{1}{m} \sum_{i=1}^{m} a_i.
\]

We can express \( S \) as

\[
S = \frac{1}{m} A_c A_c^T,
\]

where \( A_c \) is the **centered data matrix**, with \( p \) columns \( (a_i - \hat{a}) \), \( i = 1, \ldots, m \).
The empirical covariance matrix
Link with directional variance

The (empirical) variance along direction $x$ is

$$\text{var}(x) = \frac{1}{m} \sum_{i=1}^{m} [x^T (a_i - \hat{a})]^2 = x^T S x = \frac{1}{m} \|A_c x\|^2_2.$$ 

where $A_c$ is the centered data matrix.

Hence, covariance matrix gives information about variance along any direction.
Eigenvalue decomposition for symmetric matrices

Theorem (EVD of symmetric matrices)

We can decompose any symmetric $p \times p$ matrix $Q$ as

$$Q = \sum_{i=1}^{p} \lambda_i u_i u_i^T = U \Lambda U^T,$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$, with $\lambda_1 \geq \ldots \geq \lambda_n$ the eigenvalues, and $U = [u_1, \ldots, u_p]$ is a $p \times p$ orthogonal matrix ($U^T U = I_p$) that contains the eigenvectors of $Q$. That is:

$$Qu_i = \lambda_i u_i, \quad i = 1, \ldots, p.$$
Singular Value Decomposition (SVD)

Theorem (SVD of general matrices)

We can decompose any non-zero $p \times m$ matrix $A$ as

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = U \Sigma V^T, \quad \Sigma = \text{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$$

where $\sigma_1 \geq \ldots \geq \sigma_r > 0$ are the singular values, and $U = [u_1, \ldots, u_m], V = [v_1, \ldots, v_p]$ are square, orthogonal matrices ($U^T U = I_p, V^T V = I_m$). The first $r$ columns of $U, V$ contains the left-and right singular vectors of $A$, respectively, that is:

$$A v_i = \sigma_i u_i, \quad A^T u_i = \sigma_i v_i, \quad i = 1, \ldots, r.$$
Links between EVD and SVD

The SVD of a $p \times m$ matrix $A$ is related to the EVD of a (PSD) matrix related to $A$.

If $A = U \Sigma V^T$ is the SVD of $A$, then

- The EVD of $AA^T$ is $U \Lambda U^T$, with $\Lambda = \Sigma^2$.
- The EVD of $A^T A$ is $V \Lambda V^T$.

Hence the left (resp. right) singular vectors of $A$ are the eigenvectors of the PSD matrix $AA^T$ (resp. $A^T A$).
Variational characterizations
Largest and smallest eigenvalues and singular values

If $Q$ is square, symmetric:

$$\lambda_{\text{max}}(Q) = \max_{x : \|x\|_2 = 1} x^T Q x.$$ 

If $A$ is a general rectangular matrix:

$$\sigma_{\text{max}}(A) = \max_{x : \|x\|_2 = 1} \|Ax\|_2.$$ 

Similar formulae for minimum eigenvalues and singular values.
Variational characterizations
Other eigenvalues and singular values

If $Q$ is square, symmetric, the $k$-th largest eigenvalue satisfies

$$\lambda^k = \max_{x \in S^k, \|x\|_2 = 1} x^T Q x,$$

where $S^k$ is the subspace spanned by $\{u_k, \ldots, u_p\}$.

A similar result holds for singular values.
Low-rank approximation

For a given $p \times m$ matrix $A$, and integer $k \leq m, p$, the \textit{k-rank approximation} problem is

$$A^{(k)} := \arg \min_X \|X - A\|_F : \text{Rank}(X) \leq k,$$

where $\| \cdot \|_F$ is the Frobenius norm (Euclidean norm of the vector formed with all the entries of the matrix). The solution is

$$A^{(k)} = \sum_{i=1}^{k} \sigma_i u_i v_i^T,$$

where $A = U\Sigma V^T$ is an SVD of the matrix $A$. 
Low-rank approximation

Interpretation: rank-one case

Assume data matrix $A \in \mathbb{R}^{p \times m}$ represents time-series data (each row is a time-series). Assume also that $A$ is rank-one, that is, $A = uv^T \in \mathbb{R}^{p \times m}$, where $u, v$ are vectors. Then

$$A = \begin{pmatrix} a_1^T \\ \vdots \\ a_m^T \end{pmatrix}, \quad a_j(t) = u(j)v(t), \quad 1 \leq j \leq p, \quad 1 \leq t \leq m.$$

Thus, each time-series is a “scaled” copy of the time-series represented by $v$, with scaling factors given in $u$. We can think of $v$ as a “factor” that drives all the time-series.
Low-rank approximation

Interpretation: low-rank case

When $A$ is rank $k$, that is,

$$A = UV^T, \quad U \in \mathbb{R}^{p \times k}, \quad V \in \mathbb{R}^{m \times k}, \quad k << m, p,$$

we can express the $j$-th row of $A$ as

$$a_j(t) = \sum_{i=1}^{k} u_i(j) v_i(t), \quad 1 \leq j \leq p, \quad 1 \leq t \leq m.$$ 

Thus, each time-series is the sum of scaled copies of $k$ time-series represented by $v_1, \ldots, v_k$, with scaling factors given in $u_1, \ldots, u_k$. We can think of $v_i$’s as the few “factors” that drive all the time-series.
Overview of Unsupervised Learning
  Unsupervised learning models
  Matrix facts

Principal Component Analysis
  Motivations
  Variance maximization
  Deflation
  Factor models
  Example

Sparse PCA
  Basics
  SAFE
  Relaxation
  Algorithms
  Examples
  Variants

Sparse Covariance Selection
  Sparse graphical models
  Penalized maximum-likelihood
  Example

References
Motivation

Votes of US Senators, 2002-2004. The plot is impossible to read...

- Can we project data on a lower dimensional subspace?
- If so, how should we choose a projection?
Principal Component Analysis

Overview

Principal Component Analysis (PCA) originated in psychometrics in the 1930’s. It is now widely used in

- Exploratory data analysis.
- Simulation.
- Visualization.

Application fields include

- Finance, marketing, economics.
- Biology, medicine.
- Engineering design, signal compression and image processing.
- Search engines, data mining.

Principal Component Analysis (PCA) originated in psychometrics in the 1930’s. It is now widely used in

- Exploratory data analysis.
- Simulation.
- Visualization.

Application fields include

- Finance, marketing, economics.
- Biology, medicine.
- Engineering design, signal compression and image processing.
- Search engines, data mining.
Solution principles

PCA finds “principal components” (PCs), *i.e.* **orthogonal** directions of maximal variance.

- PCs are computed via EVD of covariance matrix.
- Can be interpreted as a “factor model” of original data matrix.
Variance maximization problem

Definition

Let us normalize the direction in a way that does not favor any direction.

\[ \max_x \text{var}(x) : \|x\|_2 = 1. \]

A non-convex problem!

Solution is easy to obtain via the eigenvalue decomposition (EVD) of \( S \), or via the SVD of centered data matrix \( A_c \).
Variance maximization problem

Solution

**Variance maximization problem:**

\[
\max_x x^T S x : \|x\|_2 = 1.
\]

Assume the EVD of \( S \) is given:

\[
S = \sum_{i=1}^{p} \lambda_i u_i u_i^T,
\]

with \( \lambda_1 \geq \ldots \lambda_p \), and \( U = [u_1, \ldots, u_p] \) is orthogonal (\( U^T U = I \)). Then

\[
\arg \max_{x : \|x\|_2=1} x^T S x = u_1,
\]

where \( u_1 \) is any eigenvector of \( S \) that corresponds to the largest eigenvalue \( \lambda_1 \) of \( S \).
Projection of US Senate voting data on random direction (left panel) and direction of maximal variance (right panel). The latter reveals party structure (party affiliations added after the fact). Note also the much higher range of values it provides.
Finding orthogonal directions

A deflation method

Once we’ve found a direction with high variance, can we repeat the process and find other ones?

**Deflation method:**

- Project data points on the subspace orthogonal to the direction we found.
- Find a direction of maximal variance for projected data.

The process stops after \( p \) steps (\( p \) is the dimension of the whole space), but can be stopped earlier (to find only \( k \) directions, with \( k << p \)).
Finding orthogonal directions

Result

It turns out that the direction that solves

$$\max_x \text{var}(x) : x^T u_1 = 0$$

is $u_2$, an eigenvector corresponding to the second-to-largest eigenvalue.

After $k$ steps of the deflation process, the directions returned are $u_1, \ldots, u_k$. 
Factor models

PCA allows to build a low-rank approximation to the data matrix:

\[ A = \sum_{i=1}^{k} \sigma_i u_i v_i^T \]

Each \( v_i \) is a particular factor, and \( u_i \)'s contain scalings.
Example
PCA of market data

Plot shows the eigenvalues of covariance matrix in decreasing order.
First ten components explain 80% of the variance.
Largest magnitude of eigenvector for 1st component correspond to financial sector (FABC, FTU, MER, AIG, MS).

Overview of Unsupervised Learning
Unsupervised learning models
Matrix facts

Principal Component Analysis
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Relaxation
Algorithms
Examples
Variants

Sparse Covariance Selection
Sparse graphical models
Penalized maximum-likelihood
Example

References
Motivation

One of the issues with PCA is that it does not yield principal directions that are easily interpretable:

- The principal directions are really combinations of all the relevant features (say, assets).
- Hence we cannot interpret them easily.
- The previous thresholding approach (select features with large components, zero out the others) can lead to much degraded explained variance.
Sparse PCA

Problem definition

Modify the variance maximization problem:

\[ \max_x x^T S x - \lambda \text{Card}(x) : \|x\|_2 = 1, \]

where penalty parameter \( \lambda \geq 0 \) is given, and \( \text{Card}(x) \) is the cardinality (number of non-zero elements) in \( x \).

The problem is hard but can be approximated via convex relaxation.
Safe feature elimination

Express $S$ as $S = R^T R$, with $R = [r_1, \ldots, r_p]$ (each $r_i$ corresponds to one feature).

**Theorem (Safe feature elimination [2])**

We have

$$\max_{x : \|x\|_2 = 1} x^T S x - \lambda \text{Card}(x) = \max_{z : \|z\|_2 = 1} \sum_{i=1}^{p} \max(0, (r_i^T z)^2 - \lambda).$$
Corollary

If $\lambda > \|r_i\|_2^2 = S_{ii}$, we can safely remove the $i$-th feature (row/column of $S$).

- The presence of the penalty parameter allows to prune out dimensions in the problem.
- In practice, we want $\lambda$ high as to allow better interpretability.
- Hence, interpretability requirement makes the problem easier in some sense!
Relaxation for sparse PCA

Step 1: $l_1$-norm bound

Sparse PCA problem:

$$
\phi(\lambda) := \max_x x^T S x - \lambda \text{Card}(x) : \|x\|_2 = 1,
$$

First recall Cauchy-Schwartz inequality:

$$
\|x\|_1 \leq \sqrt{\text{Card}(x)} \|x\|_2,
$$

hence we have the upper bound

$$
\phi(\lambda) \leq \overline{\phi}(\lambda) := \max_x x^T S x - \lambda \|x\|_1^2 : \|x\|_2 = 1.
$$
Relaxation for sparse PCA
Step 2: lifting and rank relaxation

Next we rewrite problem in terms of (PSD, rank-one) $X \triangleq xx^T$:

$$\overline{\phi} = \max_X \text{Tr} SX - \lambda \|X\|_1 : X \succeq 0, \quad \text{Tr} X = 1, \quad \text{Rank}(X) = 1.$$ 

Drop the rank constraint, and get the upper bound

$$\overline{\lambda} \leq \psi(\lambda) := \max_X \text{Tr} SX - \lambda \|X\|_1 : X \succeq 0, \quad \text{Tr} X = 1.$$ 

- Upper bound is a semidefinite program (SDP).
- In practice, $X$ is found to be (close to) rank-one at optimum.
Sparse PCA Algorithms

- The Sparse PCA problem remains challenging due to the huge number of variables.
- Second-order methods become quickly impractical as a result.
- SAFE technique often allows huge reduction in problem size.
- Dual block-coordinate methods are efficient in this case [7].
- Still area of active research. (Like SVD in the 70’s-90’s . . )
Example 1
Sparse PCA of New York Times headlines

*Data:* NYTtimes text collection contains 300,000 articles and has a dictionary of 102,660 unique words.

The variance of the features (words) decreases very fast:

![Sorted variances of 102,660 words in NYTimes data.](image)

With a target number of words less than 10, SAFE allows to reduce the number of features from $n \approx 100,000$ to $n = 500$. 
Example
Sparse PCA of New York Times headlines

<table>
<thead>
<tr>
<th>1st PC (6 words)</th>
<th>2nd PC (5 words)</th>
<th>3rd PC (5 words)</th>
<th>4th PC (4 words)</th>
<th>5th PC (4 words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>million</td>
<td>point</td>
<td>official</td>
<td>president</td>
<td>school</td>
</tr>
<tr>
<td>percent</td>
<td>play</td>
<td>government</td>
<td>campaign</td>
<td>program</td>
</tr>
<tr>
<td>business</td>
<td>team</td>
<td>united_states</td>
<td>bush</td>
<td>children</td>
</tr>
<tr>
<td>company</td>
<td>season</td>
<td>u.s</td>
<td>administration</td>
<td>student</td>
</tr>
<tr>
<td>market</td>
<td>game</td>
<td>attack</td>
<td></td>
<td></td>
</tr>
<tr>
<td>companies</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: the algorithm found those terms without any information on the subject headings of the corresponding articles (unsupervised problem).
NYT Dataset
Comparison with thresholded PCA

Thresholded PCA involves simply thresholding the principal components.

<table>
<thead>
<tr>
<th>( k = 2 )</th>
<th>( k = 3 )</th>
<th>( k = 9 )</th>
<th>( k = 14 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>even</td>
<td>even</td>
<td>even</td>
<td>would</td>
</tr>
<tr>
<td>like</td>
<td>like</td>
<td>we</td>
<td>new</td>
</tr>
<tr>
<td>states</td>
<td>like</td>
<td>even</td>
<td></td>
</tr>
<tr>
<td></td>
<td>now</td>
<td>we</td>
<td></td>
</tr>
<tr>
<td></td>
<td>this</td>
<td>like</td>
<td></td>
</tr>
<tr>
<td></td>
<td>will</td>
<td>now</td>
<td></td>
</tr>
<tr>
<td></td>
<td>united</td>
<td>this</td>
<td></td>
</tr>
<tr>
<td></td>
<td>states</td>
<td>will</td>
<td></td>
</tr>
<tr>
<td></td>
<td>if</td>
<td>united</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>states</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>world</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>so</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>some</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>if</td>
<td></td>
</tr>
</tbody>
</table>

1st PC from Thresholded PCA for various cardinality \( k \). The results contain a lot of non-informative words.
Robust Optimization & Machine Learning

4. Unsupervised Learning

Overview
Unsupervised learning
Matrix facts

PCA
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Algorithm
Examples
Variants

Sparse Covariance Selection
Sparsity
Penalized maximum-likelihood
Example

References

Robust PCA

PCA is based on the assumption that the data matrix can be (approximately) written as a low-rank matrix:

\[ A = LR^T, \]

with \( L \in \mathbb{R}^{p \times k} \), \( R \in \mathbb{R}^{m \times k} \), with \( k << m, p \).

Robust PCA [1] assumes that \( A \) has a “low-rank plus sparse” structure:

\[ A = N + LR^T \]

where “noise” matrix \( N \) is sparse (has many zero entries).

How do we discover \( N, L, R \) based on \( A \)?
Robust PCA model

In robust PCA, we solve the convex problem

$$\min_{N} \|A - N\|_* + \lambda \|N\|_1$$

where $\| \cdot \|_*$ is the so-called nuclear norm (sum of singular values) of its matrix argument. At optimum, $A - N$ has usually low-rank.

*Motivation:* the nuclear norm is akin to the $l_1$-norm of the vector of singular values, and $l_1$-norm minimization encourages sparsity of its argument.
CVX syntax

Here is a matlab snippet that solves a robust PCA problem via CVX, given integers \( n, m \), a \( n \times m \) matrix \( A \) and non-negative scalar \( \lambda \) exist in the workspace:

```matlab
cvx_begin
    variable X(n,m);
    minimize( norm_nuc(A-X) + lambda*norm(X(:),1))
cvx_end
```

Not the use of `norm_nuc`, which stands for the nuclear norm.
Overview of Unsupervised Learning
Unsupervised learning models
Matrix facts

Principal Component Analysis
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Relaxation
Algorithms
Examples
Variants

Sparse Covariance Selection
Sparse graphical models
Penalized maximum-likelihood
Example

References
Motivation

We’d like to draw a graph that describes the links between the features (e.g., words).

- Edges in the graph should exist when some strong, natural metric of similarity exist between features.
- For better interpretability, a \textit{sparse} graph is desirable.
- Various motivations: portfolio optimization (with sparse risk term), clustering, etc.

Here we focus on exploring \textit{conditional independence} within features.
Gaussian assumption

Let us assume that the data points are zero-mean, and follow a multi-variate Gaussian distribution: \( x \sim \mathcal{N}(0, \Sigma) \), with \( \Sigma \) a \( p \times p \) covariance matrix. Assume \( \Sigma \) is positive definite.

Gaussian probability density function:

\[
p(x) = \frac{1}{(2\pi \det \Sigma)^{p/2}} \exp((1/2)x^T \Sigma^{-1} x).
\]

where \( X := \Sigma^{-1} \) is the precision matrix.
Conditional independence

The pair of random variables $x_i, x_j$ are \textit{conditionally independent} if, for $x_k$ fixed ($k \neq i, j$), the density can be factored:

$$p(x) = p_i(x_i)p_j(x_j)$$

where $p_i, p_j$ depend also on the other variables.

\begin{itemize}
  \item \textit{Interpretation:} if all the other variables are fixed then $x_i, x_j$ are independent.
  \item \textit{Example:} Gray hair and shoe size are independent, conditioned on age.
\end{itemize}
Conditional independence
C.I. and the precision matrix

Theorem (C.I. for Gaussian RVs)
The variables $x_i, x_j$ are conditionally independent if and only if the $i, j$ element of the precision matrix is zero:

$$(\Sigma^{-1})_{ij} = 0.$$ 

Proof.
The coefficient of $x_i x_j$ in $\log p(x)$ is $(\Sigma^{-1})_{ij}$.  ■
Sparse precision matrix estimation

Let us encourage sparsity of the precision matrix in the maximum-likelihood problem:

$$\max_X \log \det X - \text{Tr} SX - \lambda \|X\|_1,$$

with $\|X\|_1 := \sum_{i,j} |X_{ij}|$, and $\lambda > 0$ a parameter.

- The above provides an invertible result, even if $S$ is not positive-definite.
- The problem is convex, and can be solved in a large-scale setting by optimizing over column/rows alternatively.
Dual

Sparse precision matrix estimation:

$$\max_X \log \det X - \text{Tr} SX - \lambda \|X\|_1.$$  

**Dual:**

$$\min_U - \log \det(S + U) : \|U\|_\infty \leq \lambda.$$  

**Block-coordinate descent:** Minimize over one column/row of $U$ cyclically. Each step is a QP.
Example
Data: Interest rates

Using covariance matrix ($\lambda = 0$).

The original precision matrix is dense, but the sparse version reveals the maturity structure.

Using $\lambda = 0.1$.
Again the sparse version reveals information, here political blocks within each party.
Overview of Unsupervised Learning
Unsupervised learning models
Matrix facts

Principal Component Analysis
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Relaxation
Algorithms
Examples
Variants

Sparse Covariance Selection
Sparse graphical models
Penalized maximum-likelihood
Example

References
Robust Optimization & Machine Learning

4. Unsupervised Learning

Overview
Unsupervised learning
Matrix facts

PCA
Motivations
Variance maximization
Deflation
Factor models
Example

Sparse PCA
Basics
SAFE
Relaxation
Algorithms
Examples
Variants

Sparse Covariance Selection
Sparsity
Penalized maximum-likelihood
Example

References

Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright.
Robust principal component analysis?
2009.

L. El Ghaoui.
On the quality of a semidefinite programming bound for sparse principal component analysis.

Olivier Ledoit and Michael Wolf.
A well-conditioned estimator for large-dimensional covariance matrices.

O.Banerjee, L. El Ghaoui, and A. d'Aspremont.
Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data.

S. Sra, S.J. Wright, and S. Nowozin.
Optimization for Machine Learning.

Sparse PCA: Convex relaxations, algorithms and applications.
To appear.

Y. Zhang and L. El Ghaoui.
Large-scale sparse principal component analysis and application to text data.
December 2011.