Expectation Maximization for Sparse and Non-Negative PCA

Christian D. Sigg   Joachim M. Buhmann

Institute of Computational Science, ETH Zurich, Switzerland

ICML 2008, Helsinki

Overview

Constrained PCA

Loadings $\mathbf{w}_{(1)}$ of first principal component (PC) are found by solving a convex maximization problem:

$$\arg\max_{\mathbf{w}} \mathbf{w}^\top \mathbf{C}\mathbf{w}, \text{ s.t.} \|\mathbf{w}\|_2 = 1, \tag{1}$$

where $\mathbf{C} \in \mathbb{R}^{D\times D}$ is the symmetric p.s.d. covariance matrix.

Loadings $\mathbf{w}_{(l)}$ of further PCs again maximize (1), subject to $\mathbf{w}_{(l)}^\top \mathbf{w}_{(k)} = 0$, $l > k$.

We consider (1) under one or two additional constraints:

1. Sparsity: $\|\mathbf{w}\|_0 \leq K$
2. Non-negativity: $w_d \geq 0$, $\forall d = 1, \ldots, D$

Motivation and Applications

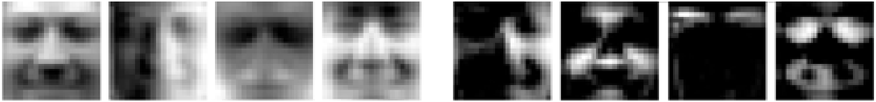
Constraints facilitate trade-off between:

1. statistical fidelity – maximization of variance
2. interpretability – feature selection omits irrelevant variables
3. applicability – constraints imposed by domain (e.g. physical process, transaction costs in finance)

Sparse PCA:

- Interpretation of PC loadings (Jolliffe et. al., 2003 JCGS)
- Gene selection (Zou et. al., 2004 JCGS) and ranking (d'Aspremont et. al., 2007 ICML).

Non-negative sparse PCA: Image parts extraction (Zass and Shashua, 2006 NIPS)



Left: Full loadings of first four PCs, Right: Corresponding non-negative sparse loadings

Combinatorial Solvers

Write (1) as

$$\mathbf{w}^\top \mathbf{C} \mathbf{w} = \sum_{i,j} C_{ij} w_i w_j \qquad (2)$$

For given sparsity pattern $\mathcal{S} = \{i | w_i \neq 0\}$, optimal solution is dominant eigenvector of corresponding submatrix $\mathbf{C}_{\mathcal{S}}$.

Algorithms:

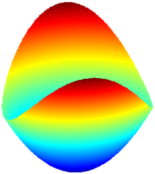
- Exact: Branch and bound (Moghaddam et. al., 2006 NIPS) for small D
- Greedy forward search (d'Aspremont et. al., 2007), improved to $O(D^2)$ per step

Optimality verifiable in $O(D^3)$ (d'Aspremont et. al., 2007).

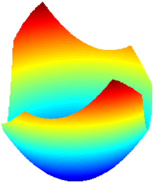
Continuous Approximation

L_1 relaxation: $\|\mathbf{w}\|_1 \leq B$

Adding constraint creates local minima (Jolliffe et. al., 2003)



$\mathbf{w}^\top \mathbf{C}\mathbf{w}$, subject to $\|\mathbf{w}\|_2 \leq 1$



$\mathbf{w}^\top \mathbf{C}\mathbf{w}$, subject to $\|\mathbf{w}\|_2 \leq 1 \wedge \|\mathbf{w}\|_1 \leq 1.2$

and makes the problem NP-hard.

Algorithms:

- Iterative L_1 regression (Zou et. al., 2004)
- Convex $O(D^4\sqrt{\log D})$ SDP approximation (d'Aspremont et. al., 2005 NIPS)
- d.c. minimization (Sriperumbudur et. al., 2007 ICML), $O(D^3)$ per iteration

Overview

Generative Model

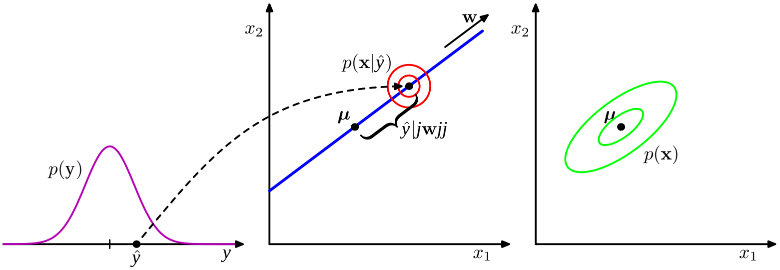
(Bishop and Tipping, 1997 TECR; Roweis, 1998 NIPS)

Latent variable $\mathbf{y} \in \mathbb{R}^L$ in PC subspace

$$\mathrm{p}(\mathbf{y}) = \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Observation $\mathbf{x} \in \mathbb{R}^D$ conditioned on $\mathbf{y}$

$$\mathrm{p}(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{W}\mathbf{y} + \boldsymbol{\mu}, \sigma^2\mathbf{I}).$$



Illustrations from (Bishop, 2006 PRML).

Limit-Case EM Algorithm

Three Simplifications: Take limit $\sigma^2 \to 0$, consider $L = 1$ subspace $\mathbf{w}$ and normalize $\|\mathbf{w}\| = 1$.

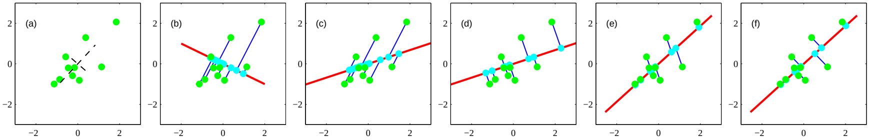
E-Step: Orthogonal projection on current estimate ($\mathbf{X} \in \mathbb{R}^{N \times D}$)

$$\mathbf{y} = \mathbf{X}\mathbf{w}_{(t)}. \tag{3}$$

M-Step: Minimization of L_2 reconstruction error

$$\mathbf{w}_{(t+1)} = \arg\min_{\mathbf{w}} \sum_{n=1}^{N} \left(\mathbf{x}_{(n)} - y_n \mathbf{w}\right)^2. \tag{4}$$

Renormalization: $\mathbf{w}_{(t+1)} = \frac{\mathbf{w}_{(t+1)}}{\|\mathbf{w}_{(t+1)}\|}$



Illustrations from (Bishop, 2006 PRML).

Rewrite M-step (4) as isotropic QP, favor sparsity using L_1 constraint and add lower bounds

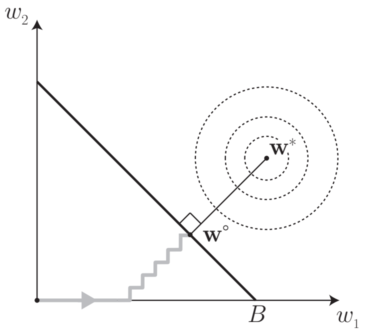
$$\begin{aligned} \mathbf{w}^\circ &= \arg\min_{\mathbf{w}} \left(h\mathbf{w}^\top\mathbf{w} - 2\mathbf{f}^\top\mathbf{w} \right) \\ \text{s.t.} & \quad \|\mathbf{w}\|_1 \le B \\ & \quad w_d \ge 0,\ \forall d \in \{1, \dots, D\} \end{aligned}$$

with $h = \sum_{n=1}^N y_n^2$ and $\mathbf{f} = \sum_{n=1}^N y_n \mathbf{x}_{(n)}$.

Minimize L_2 distance to unconstrained optimum $\mathbf{w}^* = \frac{\mathbf{f}}{h}$.

Axis-Aligned Gradient Descent

After transformation into non-negative orthant:



Express sparsity by K directly:

- L_1 bound B set implicitly due to monotonicity
- Regularization path obtained by sorting elements of $\mathbf{w}^*$, $O(D \log D)$

Multiple Principal Components

Iterative Deflation:

1. Compute sparse PC loadings $\mathbf{w}_{(l)}$.
2. Project data onto orthogonal subspace, using projector $\mathbf{P} = \mathbf{I} - \mathbf{w}_{(l)}\mathbf{w}_{(l)}^\top$.

Including nonnegativity constraints:

$$w_i^{(l)} > 0 \Rightarrow w_i^{(m)} = 0$$

for $m \neq l$, i.e. $\mathcal{S}_l$ and $\mathcal{S}_m$ are disjoint.

This might be a too strong requirement. Enforcing quasi-orthogonality: Add constraint

$$\mathbf{V}^\top \mathbf{w} \leq \alpha \tag{5}$$

where $\mathbf{V} = \left[\mathbf{w}_{(1)}\mathbf{w}_{(2)} \cdots \mathbf{w}_{(l-1)}\right]$.

Overview

Algorithms and Datasets

Algorithms considered:

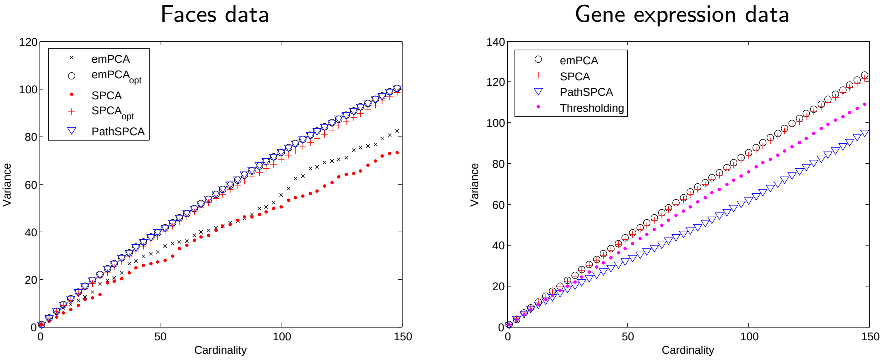
- SPCA (Zou et. al., 2004): iterative L1 regression, continuous, uses $\mathbf{X}$ instead of $\mathbf{C}$ and ranking
- PathSPCA (d'Aspremont et. al., 2007): combinatorial, greedy step is $O(D^2)$
- NSPCA (Zass and Shashua, 2006): non-negativity, quasi-orthogonality

Data sets:

1. $N > D$: MIT CBCL faces dataset (Sung, 1996), $N = 2429$, $D = 361$, referenced in NSPCA and NMF literature.
2. $D \gg N$: Gene expression data of three types of Leukemia (Armstrong et. al., 2002), $N = 72$, $D = 12582$

Standardized to zero mean, unit variance per dimension

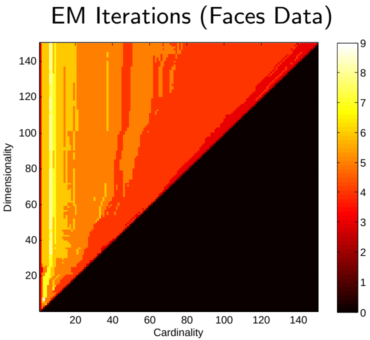
Variance vs. Cardinality

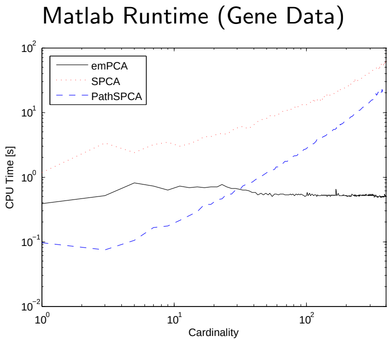


- "opt" denotes result after recomputing weights for given $\mathcal{S}$.

Computational Complexity

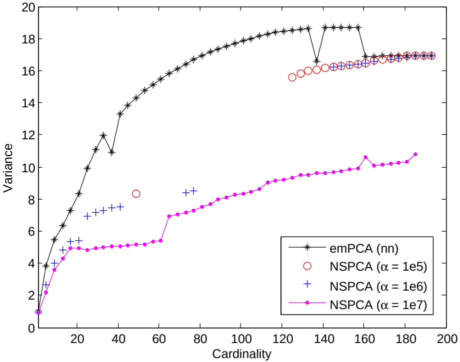
Parameter sensitivity of EM convergence and effective computational cost:





- Sublinear dependence of EM iterations on D
- Sparse emPCA solutions $(10 \leq K \leq 40)$ require more effort

Variance vs. Cardinality

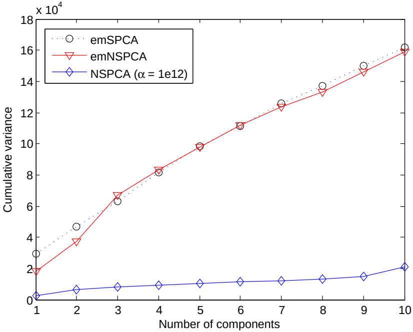


- Best result after 10 random restarts
- NSCPA sparsity parameter β determined by bisection search
- α is an orthonormality penalty

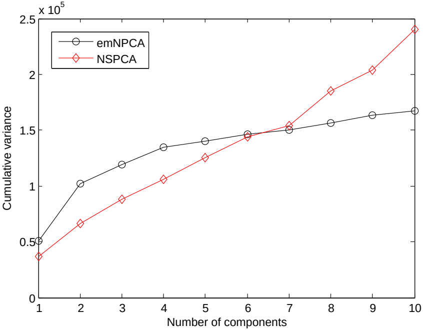
First PC loadings may lie in nonnegative orthant. Recover more than one component:

1. Keep orthogonality requirement, but constrain cardinality.
2. Require minimum angle between components.

Orthogonal Sparse Loadings



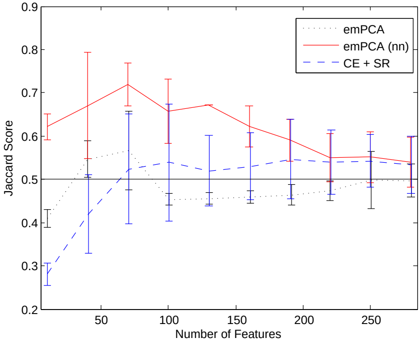
Full Quasi-Orthogonal Loadings



Unsupervised Gene Selection

Compare loadings of emPCA to CE+SR criterion (Varshavsky et al., 2006 BIOI) (LOO comparison of SV spectrum):

1. Choose gene subset
2. k-means clustering of samples ($k = 3$, 100 restarts)
3. Compare clustering assignment to label (AML, ALL, MLL)



Jaccard Score: $\frac{n_{11}}{n_{11}+n_{10}+n_{01}}$

n_{11}: pairs which have same assignment in true labeling and clustering solution.

Overview

Summary

We have presented an $O(D^2)$ constrained PCA algorithm:

- Applicability to wide range of problems:
  - sparse and nonnegative constraints
  - strict and quasi-orthogonality between components
  - $N > D$ and $D \gg N$
- Competitive variance per cardinality for sparse PCA, superior for non-negative sparse PCA
- Unmatched computational efficiency
- Direct specification of desired cardinality K, instead of bound B on L_1 norm)

Thank you for your attention.

We have machine learning PhD and postdoc positions available at ETH Zurich:

- Computational (systems) biology
- Robust (noisy) combinatorial optimization
- Learning dynamical systems

Contact: Prof. J.M. Buhmann, jbuhmann@inf.ethz.ch

http://www.ml.ethz.ch/open_positions

Joint Optimization (Work in Progress)

EM algorithm for simultaneous computation of L PCs:

E-Step:

$$\mathbf{Y} = \left(\mathbf{W}_{(t)}^{\top}\mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)}^{\top}\mathbf{X}$$

M-Step:

$$\begin{aligned} \mathbf{W}_{(t+1)} &= \arg\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\mathbf{Y}\|_F^2 \\ \text{s.t.} &\quad \|\mathbf{w}_{(l)}\|_1 \leq B,\ l = \{1, \ldots, L\} \\ &\quad w_{i,j} \geq 0 \end{aligned}$$

Tasks:

- Make this as fast as iterative deflation approach
- Avoid bad minima: enforce orthogonality?
- Specify K instead of B