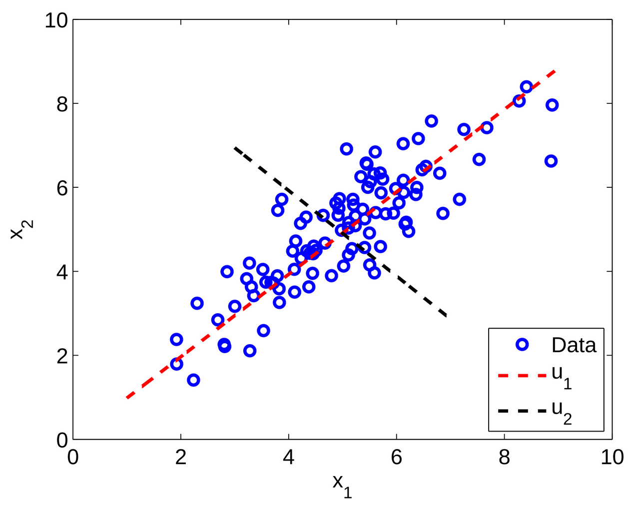
Orthogonal Sparse Eigenvectors: A Procrustes Problem

Konstantinos Benidis, Ying Sun, Prabhu Babu, Daniel P. Palomar

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong

Background & Motivation

- Principal Component Analysis (PCA) is a popular technique for data analysis and dimensionality reduction.
  - Captures directions of maximum variance of the data.
  - These directions (eigenvectors - PC loadings) form an orthonormal basis.
  - Principal components (PCs) are uncorrelated.
- Principal components are, in general, combinations of all the input variables.
  - PC loadings are dense vectors.
  - In many applications the variables have a physical meaning (e.g. gene expression).
  - A sparse basis would help significantly the interpretability of the result.
- Trade-offs:
  - Explained variance.
  - Orthogonality of the PC loadings.
  - Uncorrelatedness of the PCs.



Related Work

- Existing methods:
  - All the existing algorithms sacrifice orthogonality for a sparse result.
  - Benchmark: GPower (Journée et al. [2010]).
- Goal: Extract sparse eigenvectors that preserve the orthogonality property.

Problem Formulation

- The orthogonal sparse eigenvector extraction translates to the following optimization problem:

$$\begin{array}{ll}\underset{U}{\text{maximize}} & \text{Tr}\left(U^T S U D\right) - \sum_{i=1}^{q} \rho_i \|u_i\|_0 \\ \text{subject to} & U^T U = I_q,\end{array}$$

  where $U \in \mathbb{R}^{m\times q}$ denotes the eigenvectors, $S \in \mathbb{R}^{m\times m}$ the sample covariance matrix and $\|u_i\|_0$ the number of nonzero elements of the i-th eigenvector. $D = \text{Diag}(d) \in \mathbb{R}_+^{q\times q}$ and ρ_i are regularization parameters.
- Without the sparsity (red) term it is the typical eigenvector extraction problem.
- Discontinuous, non-differentiable, non-concave objective function.
- Non-convex set.

Approximate Smooth Formulation

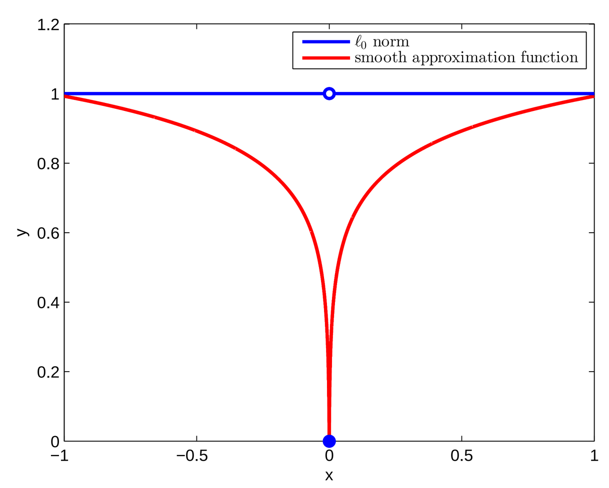
- We approximate the ℓ_0 norm with a smooth continuous and differentiable function (Song et al. [2015]):

$$\begin{array}{ll}\underset{U}{\text{maximize}} & \text{Tr}\left(U^T S U D\right) - \sum_{j=1}^{q} \rho_j \sum_{i=1}^{m} g_p^{\epsilon}(u_{ij}) \\ \text{subject to} & U^T U = I_q,\end{array} \tag{1}$$

  where

$$g_p^{\epsilon}(x) = \begin{cases} \frac{x^2}{2\epsilon(p+\epsilon)\log(1+1/p)}, & |x| \le \epsilon, \\ \frac{\log\left(\frac{p+|x|}{p+\epsilon}\right) + \frac{\epsilon}{2(p+\epsilon)}}{\log(1+1/p)}, & |x| > \epsilon, \end{cases}$$

  with $0 < p \le 1$ and $0 < \epsilon \ll 1$.
- The problem is still non-convex.
  - Use the MM framework.



Minorization-Maximization Framework

- Problem:

$$\begin{array}{ll}\underset{\mathbf{x}}{\text{maximize}} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \in \mathcal{X}.\end{array}$$

- Minorization-Maximization algorithm:

$$\mathbf{x}^{(k+1)} = \arg\max_{\mathbf{x}\in\mathcal{X}} g\left(\mathbf{x}|\mathbf{x}^{(k)}\right).$$

- Surrogate function $g(\mathbf{x}|\mathbf{x}_t)$ satisfies:

$$f\left(\mathbf{x}^{(k)}\right) = g\left(\mathbf{x}^{(k)}|\mathbf{x}^{(k)}\right),$$
$$f(\mathbf{x}) \ge g\left(\mathbf{x}|\mathbf{x}^{(k)}\right) \quad \forall \mathbf{x} \in \mathcal{X},$$
$$f'\left(\mathbf{x}^{(k)}; \mathbf{d}\right) = g'\left(\mathbf{x}^{(k)}; \mathbf{d}|\mathbf{x}^{(k)}\right),$$
$$\forall \mathbf{x}^{(k)} + \mathbf{d} \in \mathcal{X}.$$

- Iteratively maximize $g\left(\mathbf{x}|\mathbf{x}^{(k)}\right)$ instead of maximizing $f(\mathbf{x})$.

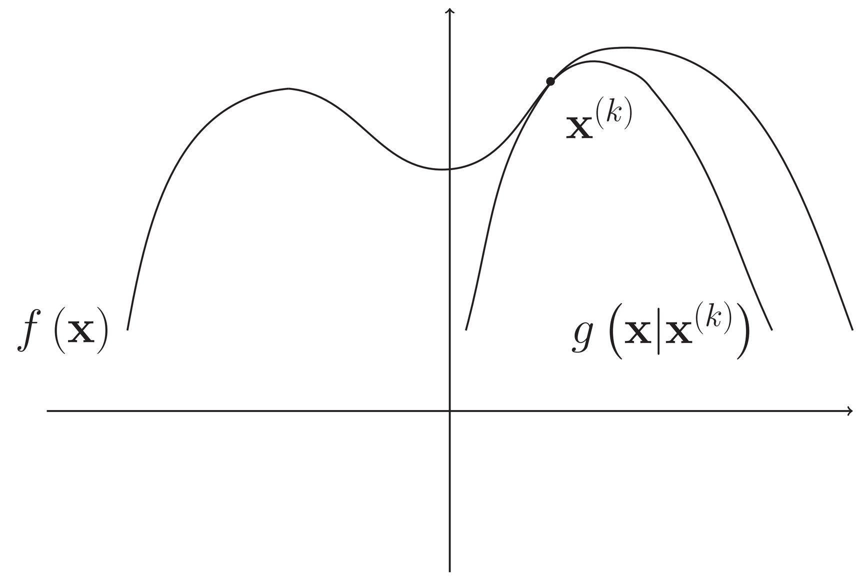


Figure: Minorization-Maximization Algorithm

Proposition

The objective function of (1) is lowerbounded by the surrogate function

$$g\left(U|U^{(k)}\right) = 2\text{Tr}\left(\left(G^{(k)} - H^{(k)}\right)^T U\right) + c,$$

where

$$G^{(k)} = S U^{(k)} D, \tag{2}$$
$$H^{(k)} = \left[\text{diag}\left(w^{(k)} - w_{\max}^{(k)} \otimes \mathbf{1}_m\right) \text{vec}\left(U^{(k)}\right)\right]_{m\times q}, \tag{3}$$

and c is an optimization irrelevant constants. The weights $w^{(k)} \in \mathbb{R}_+^{mq}$ are given by

$$w_i^{(k)} = \begin{cases} \frac{\rho_i}{2\epsilon(p+\epsilon)\log(1+1/p)}, & \left|u_i^{(k)}\right| \le \epsilon, \\ \frac{\rho_i}{2\log(1+1/p)\left|u_i^{(k)}\right|\left(\left|u_i^{(k)}\right|+p\right)}, & \left|u_i^{(k)}\right| > \epsilon, \end{cases}$$

where $u^{(k)} = \text{vec}\left(U^{(k)}\right)$, and $w_{\max}^{(k)} \in \mathbb{R}_+^q$, with $w_{\max,i}^{(k)}$ being the maximum weight that corresponds to the i-th eigenvector. Equality is achieved when $U = U^{(k)}$.

Procrustes Reformulation

- Observe that

$$\arg\max_{U\in V_{m,q}} \text{Tr}\left(\left(G^{(k)} - H^{(k)}\right)^T U\right) = \arg\min_{U\in V_{m,q}} \left\|U - \left(G^{(k)} - H^{(k)}\right)\right\|_F^2$$

  where $V_{m,q} = \{U \in \mathbf{R}m \times q | U^T U = I_q\}$ denotes the Stiefel manifold.
- The optimization problem of every MM iteration takes the following form:

$$\begin{array}{ll}\underset{U}{\text{minimize}} & \left\|U - \left(G^{(k)} - H^{(k)}\right)\right\|_F^2 \\ \text{subject to} & U^T U = I_q.\end{array} \tag{4}$$

- The optimization problem (4) is a rectangular Procrustes problem.
  - Closed-form solution.

Lemma: Rectangular Procrustes

An optimal solution of the optimization problem (4) is $U^\star = V_L V_R^T$, where V_L, V_R are the left and right singular vectors of the matrix $\left(G^{(k)} - H^{(k)}\right)$, respectively (Manton [2002]).

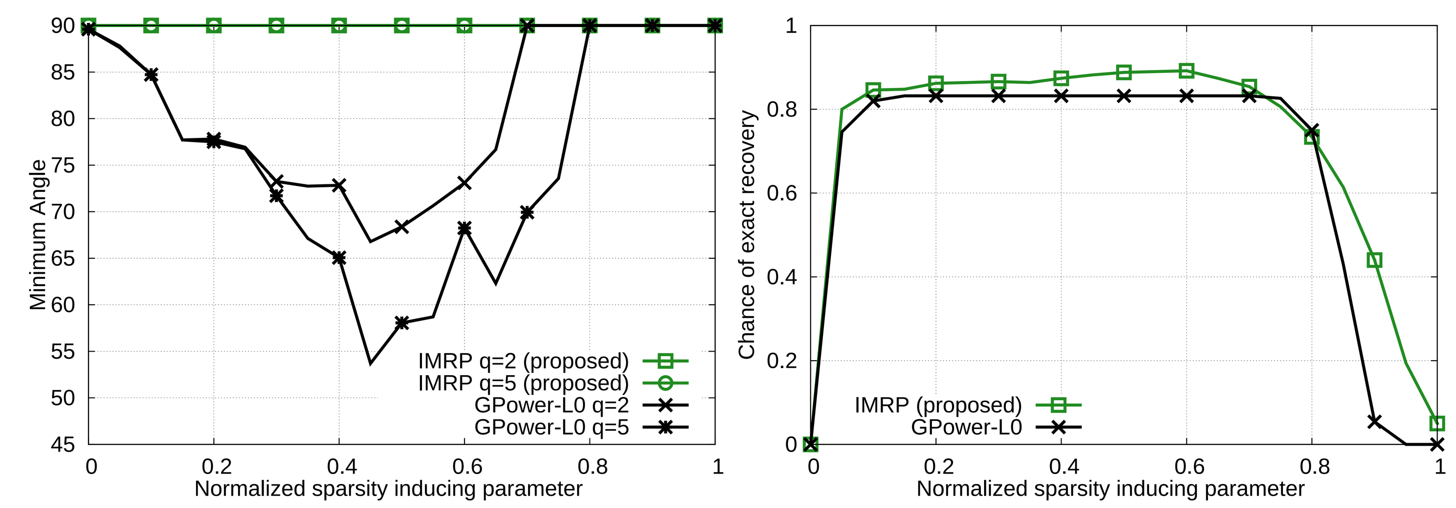
Algorithm

Algorithm 1 IMRP - Iterative Minimization of Rectangular Procrustes

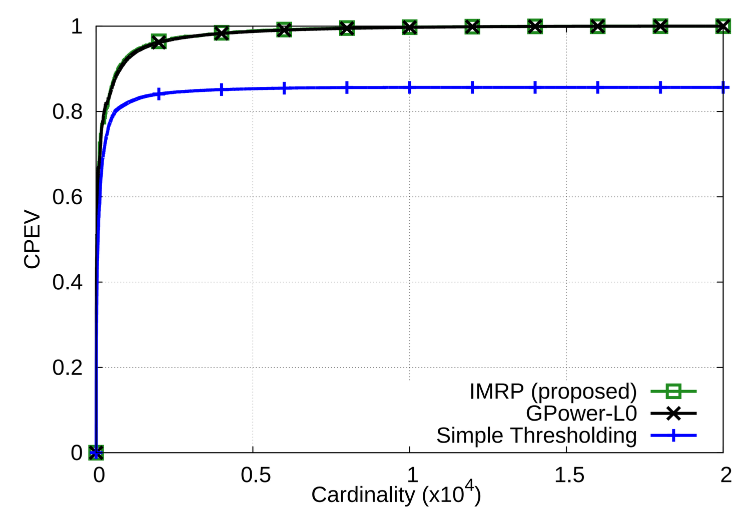
1: Set $k = 0$, choose $U^{(0)} \in \{U : U^T U = I_q\}$
2: **repeat**:
3: Compute $G^{(k)}, H^{(k)}$ with (2),(3), respectively
4: Compute V_L, V_R, the left and right singular vectors of $\left(G^{(k)} - H^{(k)}\right)$, respectively
5: $U^{(k+1)} = V_L V_R^T$
6: $k \leftarrow k+1$
7: **until** convergence
8: **return** $U^{(k)}$

Numerical Results

- Construct a covariance matrix Σ through the eigenvalue decomposition $\Sigma = V\text{diag}(\lambda)V^T$.
- The first q eigenvectors have a pre-specified sparse structure.
- We consider a setup with $m = 500$, $q = 5$.
- Generate 500 data matrices $A \in \mathbf{R}^{m\times n}$ by drawing $n = 50$ samples from a zero-mean normal distribution with covariance matrix Σ, i.e., $a_i \sim \mathcal{N}(0, \Sigma)$, for $i = 1, \ldots, n$.



- Gene expression dataset: We consider $m = 4000$ genes with the largest variance and we estimate $q = 5$ sparse eigenvectors.



Conclusion

- We have proposed a new algorithm (IMRP) for sparse eigenvalue extraction.
  - Unlike all the other existing methods, the resulting sparse eigenvectors preserve the orthogonality property.
  - IMRP improves the chance of exact recovery of the eigenvectors and matches the cumulative percentage of explained variance (CPEV).

References

Michel Journée, Yurii Nesterov, Peter Richtárik, and Rodolphe Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, March 2010.

Junxiao Song, Prabhu Babu, and Daniel P Palomar. Sparse generalized eigenvalue problem via smooth optimization. *IEEE Transactions on Signal Processing*, 63(7):1627–1642, April 2015.

Jonathan H Manton. Optimization algorithms exploiting unitary constraints. *IEEE Transactions on Signal Processing*, 50(3):635–650, March 2002.

Email: kbenidis@ust.hk