Robust Kernel Principal Component Analysis

Xinwei Deng
School of Industrial and Systems Engineering
Georgia Institute of Technology
(Joint work with Prof. Ming Yuan of Georgia Tech)

Outline

- Review of Kernel PCA.
- Robust Kernel PCA.
- Sparsity Consideration.
- Simulation Illustrations.
- Conclusions.

Kernel PCA

- Kernel PCA is to apply the PCA in the feature space F.
- F is from the nonlinear mapping φ

$$\varphi : R^p \mapsto F,\ x \mapsto y.$$

- Given the data $x_1, x_2, \ldots, x_n$, the sample covariance matrix in F

$$C = \frac{1}{n} \sum_{i=1}^{n} \varphi(x_i)\varphi(x_i)^T \triangleq \frac{1}{n} Y^T Y.$$

- From the eignvalue equation $Cv = \lambda v$,

$$v = \sum_{i=1}^{n} \alpha_i \varphi(x_i).$$

Kernel PCA

- YY^T and Y^TY have the same eignvalues, i.e., if

$$YY^T\alpha = \lambda\alpha,$$

then

$$Y^TY(Y^T\alpha) = \lambda(Y^T\alpha).$$

Hence, $v = Y^T\alpha = \sum_{i=1}^n \alpha_i\varphi(x_i)$.

- Given the kernel function $k(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$, $YY^T = K$, whose ij-th element is $k(x_i, x_j)$, and then

$$K\alpha = \lambda\alpha.$$

- The eignvector equation corresponds to K, which involves the kernel function.

Kernel PCA

- Given a test point x, the nonlinear principal component

$$v^T\varphi(x) = \sum_{i=1}^{n} \alpha_i(\varphi(x_i)^T\varphi(x)) = \sum_{i=1}^{n} \alpha_i k(x, x_i).$$

- From the angle of projection pursuit, the first kernel principal component maximizes the sample variance, i.e.,

$$\begin{aligned}\max_{\|v\|_2=1} \sum_{i=1}^{n}(v^T\varphi(x_i))^2 &= \max_{\|v\|_2=1} \|Yv\|_2^2 = \max_{\|\alpha\|_2=1} \|Y^T\alpha\|_2^2 \\ &= \max_{\|\alpha\|_2=1} \alpha^T YY^T\alpha \\ &= \max_{\|\alpha\|_2=1} \alpha^T K\alpha.\end{aligned}$$

- Not robust: the influence function is not bounded for the L_2 norm.

Robust Kernel PCA

- Classical robust approaches:
  - Projection pursuit.
  - Robust covariance estimation.
  - Robust loss function.
- Key issue: How to keep the kernel property for robust methods?
- Consider the robust L_1 norm,

$$\max_{v \in F,\ v^T v = 1} \sum_{i=1}^{n} |v^T \varphi(x_i)| = \max_{\|v\|_2 = 1} \|Yv\|_1 .$$

- Since L_1 is not differentiable, is it still possible to hold the kernel structure?

Robust Kernel PCA

- Matrix transposition invariant property
  - **Lemma.** Suppose $A \in \mathbf{R^{n \times n}}$ and $x \in \mathbf{R^n}$, define

$$\|A\|_{pr} = \max_{\|x\|_r = 1} \|Ax\|_p,$$

where $\|\cdot\|_p$ is a vector p-norm and $p,\ r > 0$, then

$$\|A\|_{pr} = \|A^T\|_{sq},$$

where p and q (respectively, r and s) are conjugate, i.e. $\frac{1}{p} + \frac{1}{q} = 1,\ \frac{1}{r} + \frac{1}{s} = 1.$

- The classical Kernel PCA is a special case by taking $p = 2$ and $r = 2$.

Robust Kernel PCA

- Choose $p = 1$ and $r = 2$ to address the robustness:
  - The L_1 projection pursuit in F is

$$\max_{v \in F,\, v^T v = 1} \sum_{i=1}^{n} |v^T \varphi(x_i)| = \max_{-1 \leq \alpha_i \leq 1} \sqrt{\alpha^T K \alpha} = \max_{\alpha \in B_n} \sqrt{\alpha^T K \alpha},$$

    where $B_n = \{\alpha = (\alpha_1, \ldots, \alpha_n)^T : \alpha_i \in \{-1, 1\},\ i = 1, \ldots, n\}$.
  - Denote $\hat{\alpha} = \arg\max_{\alpha \in B_n} \alpha^T K \alpha$, and the projection direction $\hat{v} = \arg\max_{v^T v = 1} \sum_{i=1}^{n} |v^T \varphi(x_i)|$, then

$$\hat{v} = \frac{1}{\sqrt{\hat{\alpha}^T K \hat{\alpha}}} \sum_{i=1}^{n} \hat{\alpha}_i \varphi(x_i).$$

  - $\hat{v}^T \varphi(x)$ is the kernel principal component.

Robust Kernel PCA

- The second robust kernel principal component:
  - Define $v_1 = \hat{v}$ and $\alpha_1 = \hat{\alpha}$, where $v_1 = Y^T \alpha_1 / \sqrt{\alpha_1^T K \alpha_1}$.
  - Orthogonal to v_1, it can obtained from v_1's complement space Y_2 as

$$Y_2 = Y - Y v_1 v_1^T.$$

  - The corresponding kernel matrix K_2

$$K_2 = Y_2 Y_2^T = K - \frac{1}{\alpha_1^T K \alpha_1} K \alpha_1 \alpha_1^T K.$$

  - The rest calculation is simliar as v_1.
- Other robust kernel principal components can be calculated in the same way.

Sparsity Consideration

- Not 'sparse': the kernel principal component is in terms of every training vector,

$$v^T\varphi(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x).$$

- Direct formulation on the sparseness

$$\begin{aligned} \max \quad & \alpha^T K \alpha \\ s.t. \quad & -1 \le \alpha_i \le 1, \\ & \mathbf{Card}(\alpha) \le m, \end{aligned}$$

  where m controls the level of sparsity.

- A non-convex constraint in quadratic programming.

Sparsity Consideration

- Since K is positive semi-definitive, the equivalent formulation by KKT condition is

$$\begin{aligned} \max \quad & \alpha^T K \alpha \\ s.t. \quad & -1 \leq \alpha_i \leq 1, \\ & \mathbf{1}^T |\alpha| \leq m, \end{aligned}$$

  where $\mathbf{1}^T |\alpha| = |\alpha_1| + \cdots + |\alpha_n|$.

- Viewed as penalizing the cardinality, it becomes

$$\begin{aligned} \max \quad & \alpha^T K \alpha - \rho \mathbf{Card}^2(\alpha) \\ s.t. \quad & -1 \leq \alpha_i \leq 1. \end{aligned}$$

Robust Interpretation for Sparsity Consideration

- Under the scheme of semidefinite programming (SDP), it is

$$\begin{aligned} \max \quad & \mathbf{Tr}(KA) - \rho \mathbf{Card}(A) \\ s.t. \quad & -\mathbf{1}\mathbf{1}^T \leq A \leq \mathbf{1}\mathbf{1}^T, \\ & \mathbf{Tr}(A) \leq m, \\ & A \succeq 0, \ \mathbf{Rank}(A) = 1. \end{aligned}$$

- A relaxation form,

$$\begin{aligned} \max \quad & \mathbf{Tr}(KA) - \rho \mathbf{1}^T |A| \mathbf{1} \\ s.t. \quad & -\mathbf{1}\mathbf{1}^T \leq A \leq \mathbf{1}\mathbf{1}^T, \\ & \mathbf{Tr}(A) \leq m, \\ & A \succeq 0. \end{aligned}$$

Robust Interpretation for Sparsity Consideration

- The Maxmin formulation,

$$\max_{-\mathbf{1}\mathbf{1}^T \leq A \leq \mathbf{1}\mathbf{1}^T,\ \mathbf{Tr}(A) \leq m,\ A \succeq 0} \quad \min_{|\Delta_{ij}| \leq \rho} \mathbf{Tr}(A(K + \Delta)).$$

- With the dual property, it is

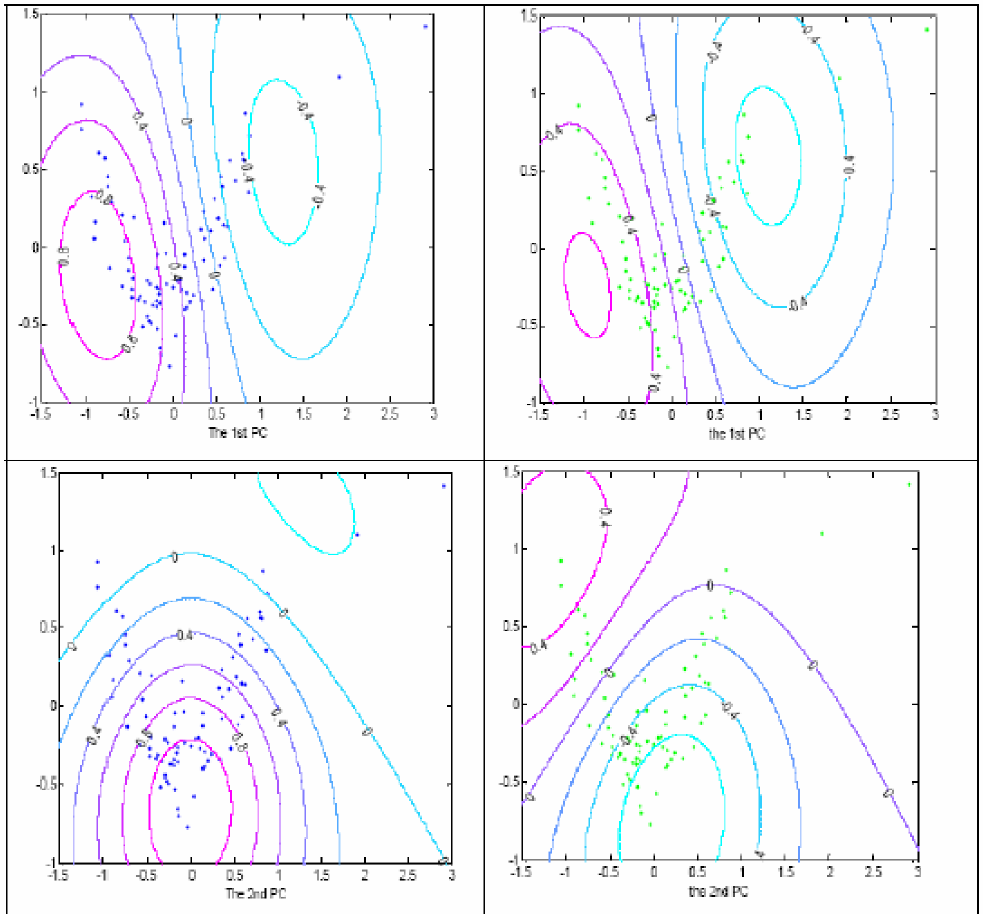
$$\begin{array}{ll} \min & \max_{-1 \leq \alpha_i \leq 1} \alpha^T (K + \Delta) \alpha \\ s.t. & |\Delta_{ij}| \leq \rho, \quad i, j = 1, \ldots, n. \end{array}$$

- It is the generalized maximum eignvalue problem with $\Delta \in \mathbf{R}^{n \times n}$.

- It corresponds to the worst-case formulation, with element-wise bounded disturbance of intensity ρ on the kernel matrix K.

Simulation Illustrations

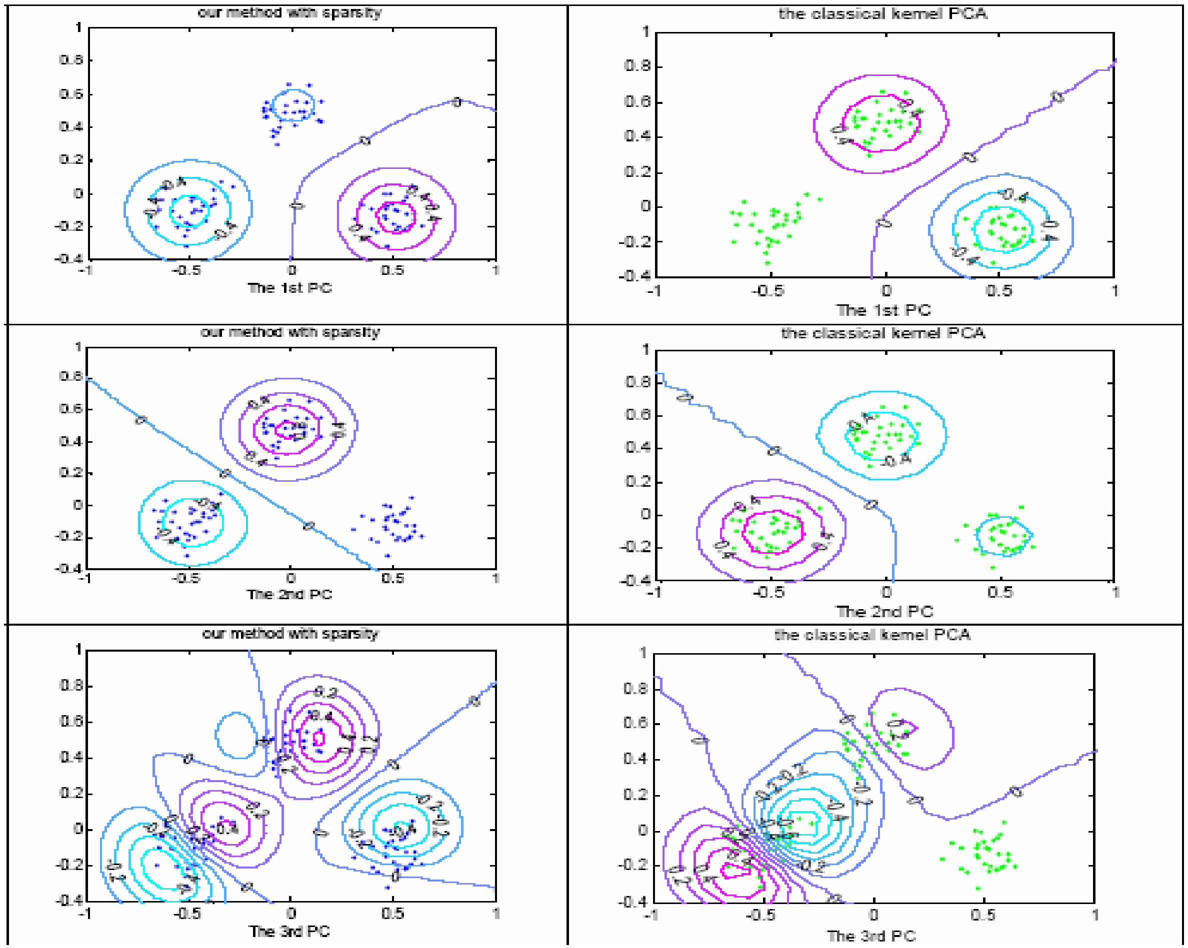
- Using the Gaussian kernel $k(x,z) = \exp(-\frac{\|x-z\|^2}{2\sigma^2})$ with $\sigma^2 = 1$, where σ^2 is the parameter for the bandwidth.



- Our robust kernel PCA approach is not affected by the 2 outliers.

Simulation Illustrations

- 3 Gaussian clusters, each having 30 vectors; applying the Gaussian kernel $\exp(-\frac{\|x-z\|^2}{r^2})$, with $r = 0.25$; $m = 0.5n$.



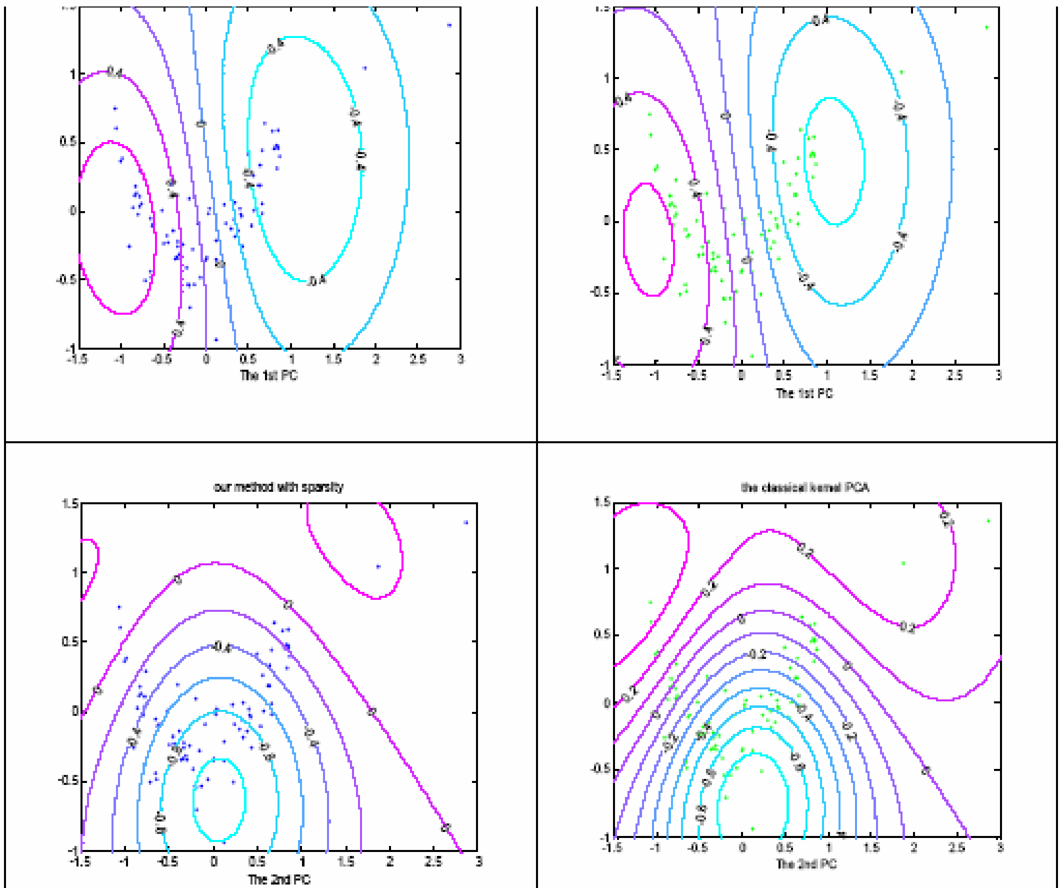
- The method with sparsity consideration performs as well as the classical kernel PCA.

Conclusions

- A robust kernel PCA approach is proposed.
- Sparsity consideration on the robust kernel PCA.
- Robust interpretation for the sparsity property.

Simulation Illustrations

- Sparsity control level: $m = 0.8n$, utilizing the Gaussian kernel $k(x, z) = \exp(-\frac{\|x-z\|^2}{2})$.



- The robust kernel PCA approach with sparsity consideration still has the robust performance.