Kernel Principal Components are maximum entropy projections

António R. C. Paiva, Jian-Wu Xu and José C. Príncipe

Computational NeuroEngineering Laboratory,
University of Florida

March 06, 2006

Outline

Outline

Introduction
- Motivation
- Cost function of Kernel PCA
- Information-Theoretic Learning concepts

Understanding Kernel PCA projections in input space

Conclusions

Motivation

- Kernel PCA provides an analytical solution to nonlinear PCA. But...

Motivation

- Kernel PCA provides an analytical solution to nonlinear PCA. But...
- What do the projections mean (in input space)?
- Is Kernel PCA better than (linear) PCA? If so, Why?

Cost function of PCA and Kernel PCA

- PCA cost function:

$$J(\mathbf{w}) = \mathbf{w}^T E\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1). \tag{1}$$

Cost function of PCA and Kernel PCA

- PCA cost function:

$$J(\mathbf{w}) = \mathbf{w}^T E\left\{\mathbf{x}\mathbf{x}^T\right\}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1). \tag{1}$$

- Kernel PCA cost function:

$$J(\mathbf{w}) = \mathbf{w}^T E\left\{\boldsymbol{\Phi}(\mathbf{x})\boldsymbol{\Phi}(\mathbf{x})^T\right\}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{w} - 1). \tag{2}$$

Kernel PCA in feature space

- $\mathbf{C} = E\left\{\boldsymbol{\Phi}(\mathbf{x})\boldsymbol{\Phi}(\mathbf{x})^T\right\}$ is the covariance matrix of the vectors in the feature space;
- Solutions are the same as the eigenvalue problem

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}. \tag{3}$$

- But solving (3) is complicated. Kernel PCA provides a workaround for this.

Kernel PCA in feature space

- $\mathbf{C} = E\left\{\mathbf{\Phi}(\mathbf{x})\mathbf{\Phi}(\mathbf{x})^T\right\}$ is the covariance matrix of the vectors in the feature space;
- Solutions are the same as the eigenvalue problem

$$\mathbf{C}\mathbf{w} = \lambda \mathbf{w}. \tag{3}$$

- But solving (3) is complicated. Kernel PCA provides a workaround for this.

Theorem

Kernel PCA is simply PCA applied in the feature space!

Rényi's quadratic entropy and Information Potential

- Rényi's quadratic entropy is defined, for a r.v. $\mathbf{x}$ with pdf $f(\mathbf{x})$, as

$$H_{R^2}(\mathbf{x}) = -\log \int_{-\infty}^{\infty} f^2(\mathbf{x})dx. \quad (4)$$

- For optimization purposes is simpler to work with the argument of the logarithm,

$$V(\mathbf{x}) = \int_{-\infty}^{\infty} f^2(\mathbf{x})dx = E\left\{f(\mathbf{x})\right\}, \quad (5)$$

named Information Potential.

Estimating the Information Potential directly from data

- First, use Parzen windowing to state a pdf estimate in terms of the data

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \kappa_{\sigma/\sqrt{2}}(\mathbf{x}, \mathbf{x}_i). \tag{6}$$

- Substituting (6) in (5) yields

$$\hat{V}(\mathbf{x}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{\sigma}(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

Estimating the Information Potential directly from data

- First, use Parzen windowing to state a pdf estimate in terms of the data

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \kappa_{\sigma/\sqrt{2}}(\mathbf{x}, \mathbf{x}_i). \tag{6}$$

- Substituting (6) in (5) yields

$$\hat{V}(\mathbf{x}) = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{\sigma}(\mathbf{x}_i, \mathbf{x}_j). \tag{7}$$

  - There is no approximation involved in this substitution!
  - No need to explicitly estimate the pdf.

Outline

Information potential in the feature space

- Using the kernel trick,

$$\kappa_\sigma(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \right\rangle ,$$

we can rewrite the information potential as

$$\begin{aligned} \hat{V}(\mathbf{x}) &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left\langle \mathbf{\Phi}(\mathbf{x}_i), \mathbf{\Phi}(\mathbf{x}_j) \right\rangle \\ &= \left\langle \frac{1}{N} \sum_{i=1}^{N} \mathbf{\Phi}(\mathbf{x}_i), \frac{1}{N} \sum_{j=1}^{N} \mathbf{\Phi}(\mathbf{x}_j) \right\rangle = \parallel \mu_{\mathbf{\Phi}} \parallel^2, \qquad (8) \end{aligned}$$

where $\mu_{\mathbf{\Phi}}$ is the mean of the feature vectors.

Variance measured in the feature space

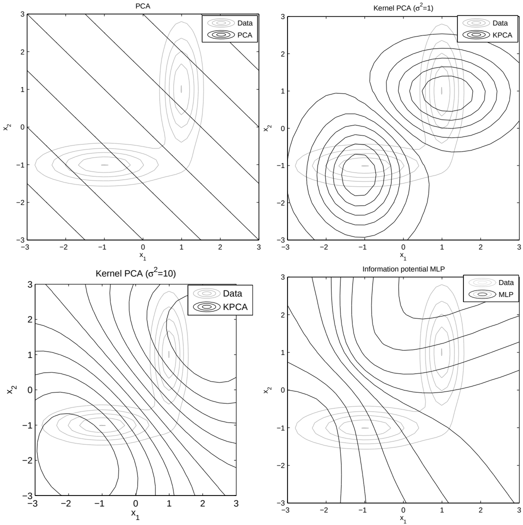
- The variance of the vectors in the feature space is

$$\begin{aligned}\operatorname{var}(\boldsymbol{\Phi}(\mathbf{x})) &= E\left\{\boldsymbol{\Phi}(\mathbf{x})^T\boldsymbol{\Phi}(\mathbf{x})\right\} - E\left\{\boldsymbol{\Phi}(\mathbf{x})\right\}^T E\left\{\boldsymbol{\Phi}(\mathbf{x})\right\} \\ &= E\left\{\kappa(\mathbf{x},\mathbf{x})\right\} - V(\mathbf{x}),\end{aligned} \tag{9}$$

where:

  - $E\left\{\kappa(\mathbf{x},\mathbf{x})\right\}$ is the maximum value of the information potential;
  - $E\left\{\boldsymbol{\Phi}(\mathbf{x})\right\}^T E\left\{\boldsymbol{\Phi}(\mathbf{x})\right\}$ is the information potential of x, $V(x)$, as shown previously.

Example



- Generated 100 samples from 2-D multimodal Gaussian distribution.
- Plot the contours of constant projection.
- Used MLP (2-4-1) for minimization of information potential.
- Notice dependence on the Kernel PCA solution on the kernel size, and different solution due to different basis functions.

Conclusions

- Reviewing:
  - Kernel PCA finds projections of maximum variance in feature space.

Conclusions

- Reviewing:
  - Kernel PCA finds projections of maximum variance in feature space.
  - From (9) (variance of the feature vectors), this implies the minimization of the information potential $V(x)$.

Conclusions

- Reviewing:
  - Kernel PCA finds projections of maximum variance in feature space.
  - From (9) (variance of the feature vectors), this implies the minimization of the information potential $V(x)$.
  - Information potential is inversely proportional to the entropy in input space.

Conclusions

- Reviewing:
  - Kernel PCA finds projections of maximum variance in feature space.
  - From (9) (variance of the feature vectors), this implies the minimization of the information potential $V(x)$.
  - Information potential is inversely proportional to the entropy in input space.

Theorem

Kernel PCA finds the principal components (basis) for projections with maximum entropy.