Probabilistic PCA and Extensions

Michael Montemurri[1] Ahmed Mhedhbi[2]

[1]McGill University
[2]Université de Montréal

December 2024

Fundamental Results of PPCA

Based on [Bishop and Tipping, 1999]

Goal of PPCA: To model high-dimensional data $\mathbf{x}_n \in \mathbb{R}^d$ using a lower-dimensional latent representation $\mathbf{z}_n \in \mathbb{R}^q$ with $q < d$, while accounting for Gaussian noise $\boldsymbol{\epsilon}$.

Generative Model:

$$\mathbf{x}_n = \mathbf{W}\mathbf{z}_n + \boldsymbol{\mu} + \boldsymbol{\epsilon}_n, \quad \boldsymbol{\epsilon}_n \sim \mathcal{N}(0, \sigma^2\mathbf{I}),$$

Where: $\mathbf{W} \in \mathbb{R}^{d \times q}$: matrix mapping latent space to data space.

Maximum Likelihood Estimation:

$$\mathbf{W}_{\text{ML}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I})^{1/2}\mathbf{R}, \quad \sigma^2_{\text{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d} \lambda_j,$$

The generative framework allows us to apply the EM Algorithm to find $\mathbf{W}$ and σ^2, which can be computationally efficient for large d

Mixture of PPCA Models

Based on [Tipping and Bishop, 1999]

Generative Model:

$$p(x_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_n | \mu_k, C_k) \quad C_k = W_k W_k^T + \sigma_k^2 I$$

Introduction of Posterior responsibilities:

$$r_{nk} = p(z_n = k | \mathbf{x}_n) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \mathbf{C}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \mathbf{C}_j)}.$$

Use the EM algorithm to update all parameters simultaneously.

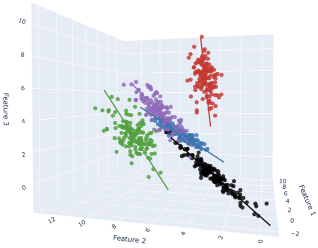

Figure 1: Mixture of PPCA Models

Probabilistic Kernel PCA (PKPCA)

Based on [Zhang et al., 1999]

Generative Model: In the feature space $\mathcal{F}$, we assume:

$$\mathbf{g} = \mathbf{B}\mathbf{w} + u\mathbf{1}_n + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V}),$$

where:

- $\mathbf{g} \in \mathbb{R}^n$: feature vector in the kernel-induced space.
- $\mathbf{B} \in \mathbb{R}^{n \times m}$: weight matrix mapping latent variables $\mathbf{w} \in \mathbb{R}^m$ to the feature space.
- $u\mathbf{1}_n$: scalar bias term for mean.
- $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V})$: noise term.

$\mathbf{V} = \mathbf{n}\sigma^{\mathbf{2}}\mathbf{I_n}/\mathbf{r}$ and $w \sim \mathcal{N}(0, n\mathbf{I_m}/\mathbf{r})$. but r, the dimensionality of the feature space, is unknown. We use the kernel trick to yield estimation procedure for $\mathbf{B}$ and σ^2. **Main Result:** The kernel

matrix K is a Wishart random matrix $W_n(r, \Sigma)$, allowing a probabilistic generative model in the kernel space. This provides probabilistic interpretation of KPCA.

Summary of PKPCA and Comparison to PPCA

Goal: Extend probabilistic PCA (PPCA) to nonlinear relationships by leveraging the kernel trick to model data in a high-dimensional feature space $\mathcal{F}$.

Generative Model in $\mathcal{F}$:

$$\mathbf{g} = \mathbf{B}\mathbf{w} + u\mathbf{1}_n + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{V}),$$

where $\mathbf{g}$ is the feature vector, $\mathbf{B}$ maps latent variables $\mathbf{w}$ to $\mathcal{F}$, and $\mathbf{K} \sim W_n(r, \Sigma)$ is a Wishart random matrix.

- Models feature vectors $\mathbf{g}$ in the kernel-induced feature space $\mathcal{F}$.
- Uses the kernel trick to enable nonlinear dimensionality reduction.
- Probabilistic interpretation of the kernel matrix $\mathbf{K}$ as a Wishart random matrix.

Original Contributions

Experimental Results

Comparison of Analytical and EM-Based PPCA: Mixture Models:

Mixture of PPCA models demonstrated superior performance in clustering multimodal datasets, capturing local linear structures effectively.

PKPCA:

Probabilistic Kernel PCA significantly outperformed linear PPCA in modeling complex, nonlinear datasets. Temporal kernels further enhanced performance in time-series applications.

More Results + Future Directions

References

- Tipping, Michael E., and Christopher M. Bishop. *"Mixtures of probabilistic principal component analyzers."* Neural Computation, 11.2 (1999): 443-482.
- Tipping, Michael E., and Christopher M. Bishop. *"Probabilistic principal component analysis."* Journal of the Royal Statistical Society Series B: Statistical Methodology, 61.3 (1999): 611-622.
- Zhang, Zhihua, et al. *"Probabilistic kernel principal component analysis."* Department of Computer Science, The Hong Kong University of Science and Technology, Technical Report (2004).

M-Step:

- Update parameters for each mixture component:

$$\pi_i = \frac{1}{N} \sum_{n=1}^{N} R_{ni},$$

$$\boldsymbol{\mu}_i = \frac{\sum_{n=1}^{N} R_{ni} \mathbf{x}_n}{\sum_{n=1}^{N} R_{ni}},$$

$$\mathbf{W}_i = \left(\sum_{n=1}^{N} R_{ni} (\mathbf{x}_n - \boldsymbol{\mu}_i) \mathbb{E}[\mathbf{z}_{n,i+1}]^T \right) \left(\sum_{n=1}^{N} R_{ni} \mathbb{E}[\mathbf{z}_{n,i+1} \mathbf{z}_{n,i+1}^T] \right)^{-1},$$

$$\sigma_i^2 = \frac{1}{d} \sum_{n=1}^{N} R_{ni} \left[\|\mathbf{x}_n - \boldsymbol{\mu}_i\|^2 - 2\mathbb{E}[\mathbf{z}_{n,i+1}]^T \mathbf{W}_i^T (\mathbf{x}_n - \boldsymbol{\mu}_i) + \mathrm{Tr}(\mathbf{W}_i^T \mathbf{W}\right.$$