Probabilistic PCA and CCA and Applications to Missing Data

Marc Boëlle Tess Breton

école normale supérieure paris—saclay

MATHÉMATIQUES VISION APPRENTISSAGE

Project Overview

We worked on the articles **"Probabilistic Principal Component Analysis"** by Michael Tipping and Christopher Bishop [1] and **"A Probabilistic Interpretation of Canonical Correlation Analysis"** by Francis Bach and Michael I. Jordan [2].

They provide probabilistic frameworks to PCA anc CCA, using latent variable models. Our work focused on the application of PPCA and PCCA to missing data handling. We implemented EM algorithms taking as input incomplete datasets, and introduced a novel EM algorithm for PCCA with missing data.

Background on PCA and CCA

PCA (Principal Component Analysis) is a dimensionality reduction technique. It finds linear combinations of features (*principal components*) that **maximize the variance** of the data projections along these directions.

CCA (Canonical Correlation Analysis) explores linear correlations between **two paired datasets $\mathbf{X}_A$ and $\mathbf{X}_B$**, often representing two sets of features of a single dataset. It finds pairs of directions (*canonical components*), one for each set, that **maximize the correlation** of the data projections.

Latent variable models for PPCA and PCCA

Graphical models corresponding to PPCA and PCCA:

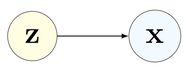


Figure 1. Graphical model for PPCA

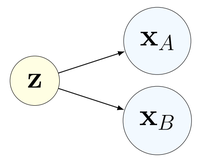


Figure 2. Graphical model for PCCA

Mathematically, the latent variable models are defined as follows:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$$
$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_d)$$

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$$
$$\mathbf{x}_A \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}_A\mathbf{z} + \boldsymbol{\mu}_A, \boldsymbol{\Psi}_A)$$
$$\mathbf{x}_B \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}_B\mathbf{z} + \boldsymbol{\mu}_B, \boldsymbol{\Psi}_B)$$

Handling Missing Data in Standard PCA and CCA

Common solutions:

- **Fill missing entries with the mean or median of the observed values** - inaccurate when the data involves complex structures such as clusters with different properties
- **Remove rows with missing entries** - impractical when many observations have one or more missing values.

Proposed approach: Use PPCA/PCCA and **estimate missing values along with latent variables in the E-step** of the EM algorithm.

EM algorithm for PPCA with missing values

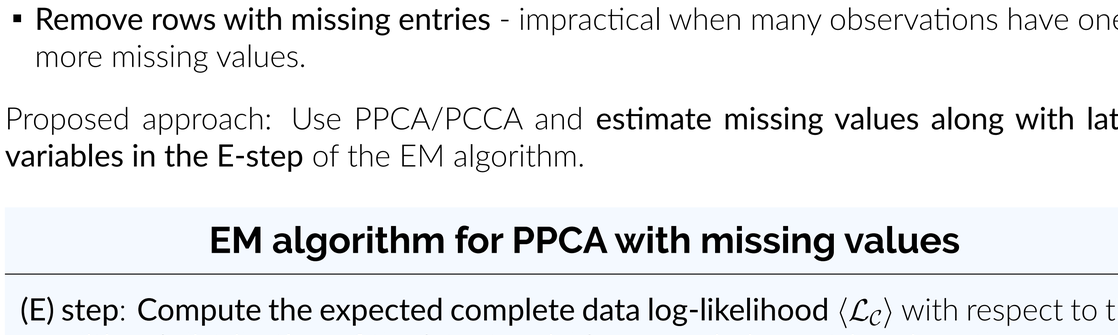
(E) step: **Compute the expected complete data log-likelihood** $\langle \mathcal{L}_\mathcal{C} \rangle$ with respect to the conditional distribution of the latent variables and missing entries given the observed values and current parameter estimates.

$$\langle \mathcal{L}_\mathcal{C} \rangle = -\sum_{i=1}^{n} \left\{ \frac{d}{2}\ln(\sigma^2) + \frac{1}{2}\mathrm{tr}(\langle \mathbf{z}_i \mathbf{z}_i^\top \rangle) + \frac{1}{2\sigma^2}(\langle \mathbf{x}_i \rangle - \boldsymbol{\mu})^\top (\langle \mathbf{x}_i \rangle - \boldsymbol{\mu}) \right.$$
$$\left. - \frac{1}{\sigma^2}\langle \mathbf{z}_i \rangle^\top \mathbf{W}^\top (\langle \mathbf{x}_i \rangle - \boldsymbol{\mu}) + \frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{W}^\top \mathbf{W} \langle \mathbf{z}_i \mathbf{z}_i^\top \rangle) \right\}$$

$$\langle \mathbf{z} \rangle = \mathbf{M}_{\text{obs}}^{-1} \mathbf{W}_{\text{obs}}^\top (\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}})$$
$$\langle \mathbf{z}\mathbf{z}^\top \rangle = \sigma^2 \mathbf{M}_{\text{obs}}^{-1} + \langle \mathbf{z} \rangle \langle \mathbf{z} \rangle^\top$$
$$\langle \mathbf{x} \rangle = (\mathbf{x}_{\text{obs}}, \langle \mathbf{x}_{\text{miss}} \rangle) = (\mathbf{x}_{\text{obs}}, \mathbf{W}_{\text{miss}} \langle \mathbf{z} \rangle)$$

(M) step: **Update model parameters** to maximize the complete data log-likelihood, using closed-form expressions.

Visual Comparison of PCA and PPCA



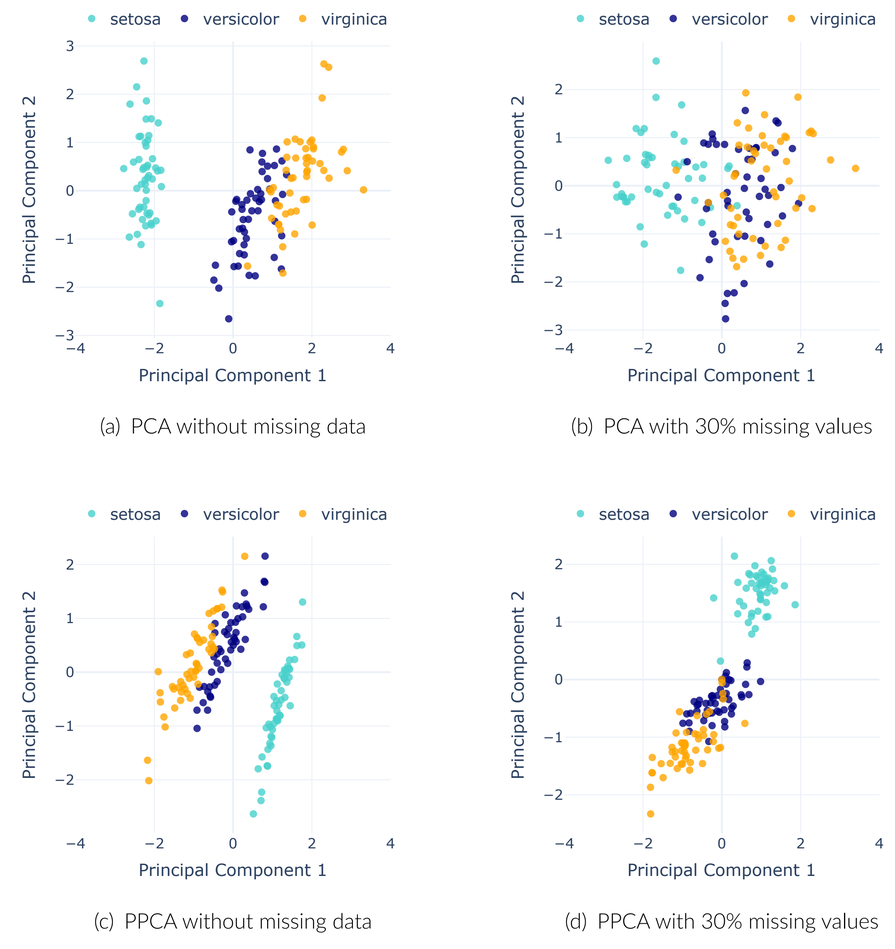
(a) Standard PCA (b) PPCA with fixed $\sigma^2 = 10$ (c) ML PPCA with $\hat{\sigma}^2_{\text{ML}} = 2.34$

Figure 3. PCA and PPCA for $n = 30$ samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{pmatrix} 10 \\ 6 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 10 & 7 \\ 7 & 10 \end{pmatrix}$

- Loss of orthogonal projection when using PPCA.
- A high σ^2 brings the projections closer to each other, and the observations are then **mostly explained by noise**.
- The optimal σ^2 finds a **balance** between data correlation and independent noise.

Experiment on Iris dataset - PCA and PPCA with missing values

Iris dataset : 150 samples from three flower species with four features : "sepal length", "sepal width", "petal length" and "petal width". For PCA and PPCA, we plot the projections onto the first two principal components with **0%, 15% and 30% random missing values**. For PCA, missing values were **filled with the mean** of the observed ones.



(a) PCA without missing data (b) PCA with 30% missing values

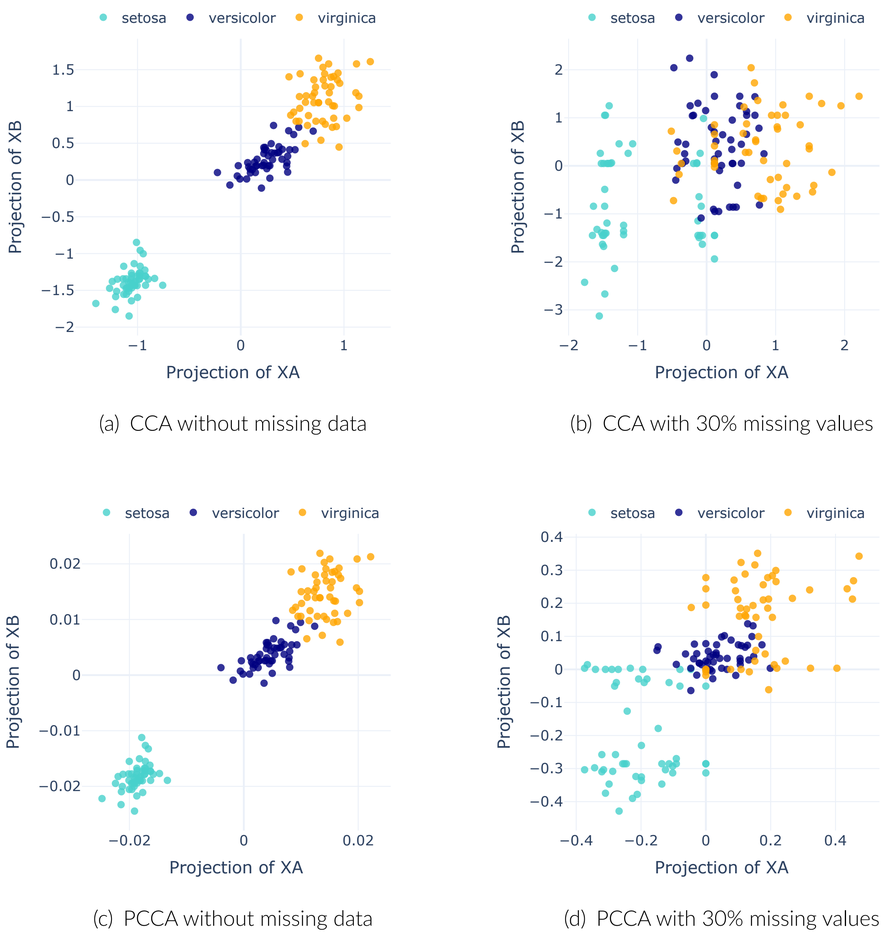
(c) PPCA without missing data (d) PPCA with 30% missing values

Figure 4. PCA and PPCA on Iris dataset with different levels of missing data

- Results quite similar for PCA and PPCA without missing data, with a slightly higher variance for PCA.
- **With missing data, PPCA outperforms PCA** by preserving the three clusters.

Experiment on Iris dataset - CCA and PCCA with missing values

We split the Iris dataset into two subsets containing respectively the features "petal length" and "sepal length" for $\mathbf{X}_A$, and "petal width" and "sepal width" for $\mathbf{X}_B$.



(a) CCA without missing data (b) CCA with 30% missing values

(c) PCCA without missing data (d) PCCA with 30% missing values

Figure 5. CCA and PCCA on Iris dataset with different levels of missing data

- Similar results for CCA and PCCA on complete data and strong correlation.
- With missing data, **PCCA demonstrates greater robustness**, maintaining better class separation and greater correlation.

	0% missing	15% missing	30% missing
CCA	0.97	0.58	0.41
PCCA	0.97	**0.85**	**0.70**

Table 1. Correlation of the linear projections obtained by CCA and PCCA shown in Figure 5

References

[1] M. E. Tipping and Christopher Bishop, "Probabilistic Principal Component Analysis", Journal of the Royal Statistical Society, Series B: 21.3 (January 1999), pages 611–622.

[2] Francis R. Bach and Michael I. Jordan, "A probabilistic interpretation of canonical correlation analysis", 2005.