Probabilistic PCA and CCA and Applications to Missing Data

MVA - Probabilistic Graphical Models and Deep Generative Models

Marc Boëlle and Tess Breton

1 Introduction and Contributions

For this project, we worked on the papers ”Probabilistic Principal Component Analysis” by Tipping and Bishop [3] and “A Probabilistic Interpretation of Canonical Correlation Analysis” by Bach and Jordan [1]. Tipping and Bishop provide a probabilistic framework to Principal Component Analysis (PCA) using a latent variable model to define Probabilistic PCA (PPCA). Bach and Jordan extend their work to Canonical Correlation Analysis (CCA) and define a new model, which we refer to as Probabilistic CCA (PCCA).

In Sections 2 and 3, we first go through the definitions of PPCA and PCCA, providing a unified framework with common notations. For both methods, we implement Expectation-Maximization (EM) algorithms to estimate parameters. In Section 4 we focus on a specific application introduced by Tipping and Bishop [3], which is handling missing data. We follow [3] to build a robust EM algorithm for PPCA, and adapt it to PCCA. We visualize results and assess the performance of the proposed approach on the Iris dataset. Our implementation is available in our GitHub repository.

Contributions: We jointly formulated the mathematical framework and implemented the EM algorithms. Regarding theory, Marc covered PPCA and Tess PCCA. Marc wrote the initial EM algorithm for PPCA and PCCA, and Tess adapted it to handle missing data. Tess led the Iris dataset experiments, while Marc worked on visualizing differences between PCA and PPCA.

2 Probabilistic PCA

2.1 Background on PCA

PCA is a well-known dimensionality reduction technique. It identifies directions, known as *principal components*, which are linear combinations of features maximizing the variance of the projected data. Mathematically, they correspond to the eigenvectors of the sample covariance matrix. In Appendix A, we provide further theoretical details on PCA and derive the first principal component.

PCA is a data-based method, which makes no assumption on how the observations were generated. This is one of the motivations of Tipping and Bishop [3] in designing PPCA.

2.2 Latent Variable Model for PPCA

The probabilistic framework of PPCA is a latent variable model. Such models aim at linking d-dimensional observation vectors $\mathbf{x}$ to q-dimensional latent variables $\mathbf{z}$, with $q < d$, capturing the underlying structure in a compact representation. The corresponding graphical model is provided in Figure 1.

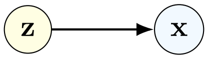


Figure 1: Graphical model for PPCA

As elaborated in Appendix B.1, one of the simplest latent variable models is Factor Analysis, where observations are modeled as noisy linear transformations of latent variables. PPCA builds upon this idea, further adding the assumption of isotropic Gaussian noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_d)$.

With a Gaussian prior on the latent variables, we obtain the following model:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{1}$$

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

Notably, these equations result in a Gaussian observation model $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with covariance matrix $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}_d$.

The Maximum Likelihood Estimators (MLEs) of $\boldsymbol{\mu}, \mathbf{W}$ and σ^2 can be derived in closed form, as detailed in Appendix B.2. Alternatively, $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 can be estimated using the EM algorithm described below. Note that $\hat{\boldsymbol{\mu}}_{\mathrm{ML}}$ is simply the sample mean, so it needs to be computed only once.

2.3 EM algorithm for PPCA

The EM approach maximizes the expected log-likelihood of the complete data, which includes both observed and latent variables. Each iteration of the algorithm involves the following steps:

- E-step: compute the expected complete data log-likelihood with respect to the conditional distribution of the latent variables given the observations and parameters $p\left(\{\mathbf{z}_i\}|\{\mathbf{x}_i\}, \mathbf{W}, \sigma^2\right)$.
- M-step: maximize the obtained expectation with respect to $\mathbf{W}$ and σ^2.

Letting $\langle\cdot\rangle$ denote the expected value with respect to the distribution mentioned above, Tipping and Bishop [3] prove that the parameter updates can be written as:

$$\mathbf{W}_{t+1} = \left\{\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})\langle\mathbf{z}_i\rangle_t^\top\right\}\left(\sum_{n=1}^{N}\langle\mathbf{z}_i\mathbf{z}_i^\top\rangle_t\right)^{-1} \tag{3}$$

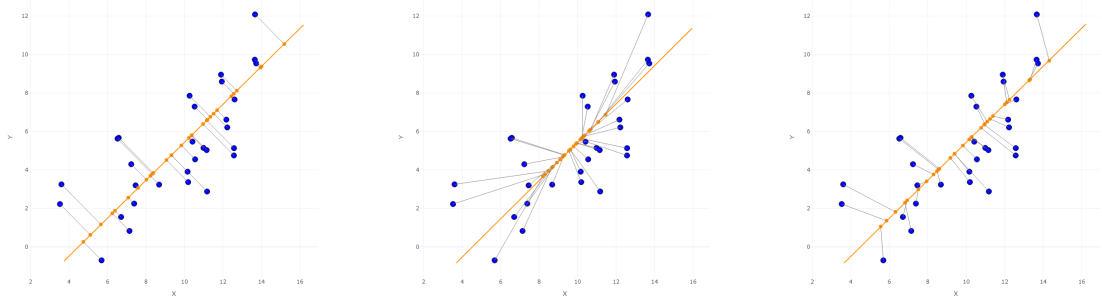
$$\sigma_{t+1}^2 = \frac{1}{nd}\sum_{i=1}^{n}\left\{\|\mathbf{x}_i - \boldsymbol{\mu}\|^2 - 2\langle\mathbf{z}_i\rangle_t^\top\mathbf{W}_{t+1}^\top(\mathbf{x}_i - \boldsymbol{\mu}) + \operatorname{tr}\left(\langle\mathbf{z}_i\mathbf{z}_i^\top\rangle_t\mathbf{W}_{t+1}^\top\mathbf{W}_{t+1}\right)\right\}. \tag{4}$$

They also show that among all stationary points of the log-likelihood, the only stable maximum is obtained when $\mathbf{W}$ contains the q principal eigenvectors. Thus, local maximization techniques such as the EM algorithm always converge to an optimal solution, up to an arbitrary rotation of $\mathbf{W}$.

2.4 Visual Comparison of PCA and PPCA

To visualize the difference between PCA and PPCA projections, and especially the role of σ^2 in PPCA, we ran an experiment on a synthetic dataset in $\mathbb{R}^2$ sampled from a Gaussian distribution. We applied PCA and PPCA with one component. For PPCA, we ran a first EM algorithm with fixed $\sigma^2 = 10$, deriving only the MLE of $\mathbf{W}$ as defined in Equation (21) in Appendix B.2. In a second run, we computed the MLEs of both σ^2 and $\mathbf{W}$. The samples and their reconstructions are shown in Figure 2.

As expected, standard PCA projections are orthogonal. PPCA projections are not orthogonal, so they do not minimize the least-squares reconstruction error. By enforcing a high σ^2, the projections get closer to each other, as the observations are then mostly explained by random noise. The MLE for σ^2 strikes an optimal balance between data correlation and independent noise.



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

Figure 2: PCA and PPCA for $n = 30$ samples from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{pmatrix} 10 \\ 6 \end{pmatrix}$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 10 & 7 \\ 7 & 10 \end{pmatrix}$

3 Probabilistic CCA

3.1 Canonical Correlation Analysis (CCA)

CCA is a statistical technique used to explore linear correlations between two paired datasets, often representing two sets of features of a single dataset such as socio-economic factors and academic performance. It identifies pairs of directions, one for each dataset, that maximize the correlation of the data projections. These directions, called *canonical components*, are obtained by solving a generalized eigenvalue problem involving sample covariance matrices. In Appendix C, we provide details on the theory behind CCA and the derivation of the first pair of canonical components.

3.2 Latent Variable Model

In PCCA, a pair of observed variables $\mathbf{x}_A \in \mathbb{R}^{d_A}$ and $\mathbf{x}_B \in \mathbb{R}^{d_B}$ is assumed to stem from a single latent variable $\mathbf{z} \in \mathbb{R}^q$. The corresponding graphical model is given in Figure 3.

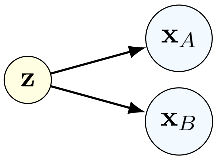


Figure 3: Graphical model for PCCA

Similarly as PPCA, PCCA assumes a Gaussian prior on the latent variables. The conditional distributions of the observed variables given the latent are then also Gaussian:

$$\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q) \tag{5}$$

$$\mathbf{x}_A \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}_A\mathbf{z} + \boldsymbol{\mu}_A, \boldsymbol{\Psi}_A) \tag{6}$$

$$\mathbf{x}_B \mid \mathbf{z} \sim \mathcal{N}(\mathbf{W}_B\mathbf{z} + \boldsymbol{\mu}_B, \boldsymbol{\Psi}_B) \tag{7}$$

3.3 EM algorithm for PCCA

Analytical expressions for the maximum-likelihood estimates of the parameters $\mathbf{W}_A$, $\mathbf{W}_B$, $\boldsymbol{\mu}_A$, $\boldsymbol{\mu}_B$, $\boldsymbol{\Psi_A}$ and $\boldsymbol{\Psi_B}$ can be derived. However, in this project, we focus on the EM approach to later handle missing data. We refer to the work of Bach and Jordan [1] for the detailed analytical expressions, and to Appendix D.1 for the EM algorithm with complete observations.

4 PPCA and PCCA to Handle Missing Data

4.1 Missing Data

Real-world datasets often contain missing values, making traditional methods like PCA or CCA unsuitable as they rely on the sample covariance matrix. A common solution is to fill missing values with the sample mean or median. But this can be inaccurate in data with complex structures such as clusters, where the mean or median may not reflect the underlying patterns accurately. Alternatively, removing rows with missing data is impractical if many observations are missing.

Notations: Each observation $\mathbf{x}$ is split into its observed and missing features $\mathbf{x} = (\mathbf{x}_{\text{obs}}, \mathbf{x}_{\text{miss}})$. The observed and missing features may vary across observations, and this notation is used for clarity purposes without changing the order of features. For an observation $\mathbf{x}$, the matrix $\mathbf{A}_{\text{obs}}$ contains the rows of $\mathbf{A}$ corresponding to the features observed in $\mathbf{x}$. Note that $\mathbf{A}_{\text{obs}}$ depends on $\mathbf{x}$.

4.2 PPCA with Missing Data

Tipping and Bishop [3] argue that PPCA is effective when some or all of the observations $\{\mathbf{x}_i\}$ have missing entries. They adapt their EM algorithm to handle missing data, the complete data now being made of the observed values $\{\mathbf{x}_{i,\text{obs}}\}$, the latent variables $\{\mathbf{z}_i\}$ and the missing entries $\{\mathbf{x}_{i,\text{miss}}\}$. However, they do not provide further detail nor mathematical formulas, which we present in the following. The idea is to estimate the complete data at every iteration of the EM, and then perform the standard steps. The pseudo-code of our approach is provided in Algorithm 1 in Appendix B.3.

4.2.1 Model

To impute missing values based on the observed ones and current parameter estimates, we first need the conditional distributions of $\mathbf{z}$ given $\mathbf{x}_{\text{obs}}$, and of $\mathbf{x}_{\text{miss}}$ given $\mathbf{z}$. With $\mathbf{M} = \mathbf{W}^\top \mathbf{W} + \sigma^2 \mathbf{I}_q$:

$$\mathbf{z} \mid (\mathbf{x}_{\text{obs}}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \sim \mathcal{N}\left(\mathbf{M}_{\text{obs}}^{-1} \mathbf{W}_{\text{obs}}^\top (\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}}),\ \sigma^2 \mathbf{M}_{\text{obs}}^{-1}\right) \tag{8}$$

$$\mathbf{x}_{\text{miss}} \mid (\mathbf{x}_{\text{obs}}, \mathbf{z}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) \ \sim\ \mathcal{N}\left(\mathbf{W}_{\text{miss}} \mathbf{z} + \boldsymbol{\mu}_{\text{miss}}, \sigma^2 \mathbf{I}\right) \tag{9}$$

In the E-step, we compute the complete data log-likelihood using the following MLEs of $\mathbf{z}$ and $\mathbf{x}_{\text{miss}}$ given $\mathbf{x}_{\text{obs}}$ and current parameter estimates $\mathbf{W}_t, \boldsymbol{\mu}_t$ and σ_t^2:

$$\langle \mathbf{z} \rangle_t = \mathbf{M}_{t,\text{obs}}^{-1} \mathbf{W}_{t,\text{obs}}^\top (\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_{t,\text{obs}}) \tag{10}$$

$$\langle \mathbf{x} \rangle_t = (\mathbf{x}_{\text{obs}}, \langle \mathbf{x}_{\text{miss}} \rangle_t) = (\mathbf{x}_{\text{obs}}, \mathbf{W}_{t,\text{miss}} \langle \mathbf{z} \rangle_t + \boldsymbol{\mu}_{t,\text{miss}}) \tag{11}$$

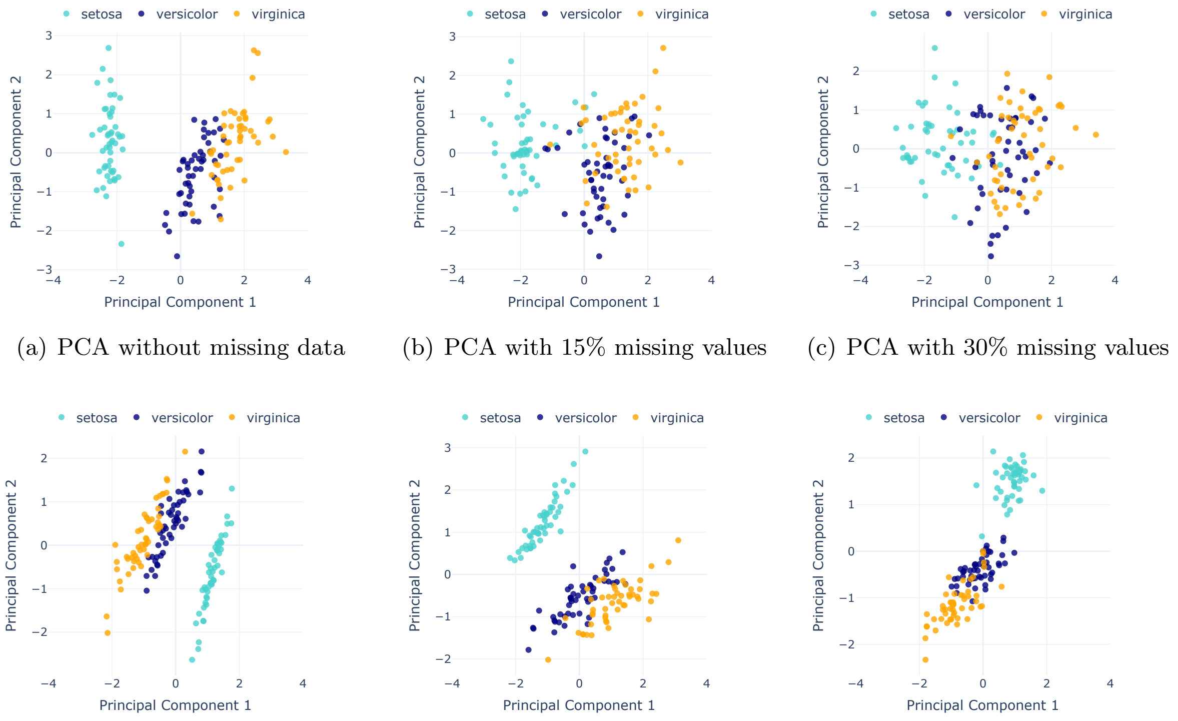
For the M-step, the updates are the same as in Equations (3) and (4), with $\langle \mathbf{x}_i \rangle_t$ replacing $\mathbf{x}_i$:

$$\boldsymbol{\mu}_{t+1} = \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{x}_i \rangle_t\,, \quad \mathbf{W}_{t+1} = \left\{ \sum_{i=1}^{n} (\langle \mathbf{x}_i \rangle_t - \boldsymbol{\mu}_{t+1}) \langle \mathbf{z}_i \rangle_t^\top \right\} \left(\sum_{n=1}^{N} \langle \mathbf{z}_i \mathbf{z}_i^\top \rangle_t \right)^{-1} \tag{12}$$

$$\sigma_{t+1}^2 = \frac{1}{nd} \sum_{i=1}^{n} \left\{ \| \langle \mathbf{x}_i \rangle_t - \boldsymbol{\mu}_{t+1} \|^2 - 2 \langle \mathbf{z}_i \rangle_t^\top \mathbf{W}_{t+1}^\top (\langle \mathbf{x}_i \rangle_t - \boldsymbol{\mu}_{t+1}) + \operatorname{tr}\left(\langle \mathbf{z}_i \mathbf{z}_i^\top \rangle_t \mathbf{W}_{t+1}^\top \mathbf{W}_{t+1} \right) \right\} \tag{13}$$

4.2.2 Experiment on the Iris Dataset to Evaluate PPCA

We tested the method on the Iris dataset, which includes 150 samples from three species with four features: sepal length, sepal width, petal length and petal width. Projections onto the first two principal components from PCA and PPCA were compared with 0%, 15% and 30% random missing values. For PCA, missing values were filled with the mean of the observed ones[1]. We used our implementation for PCA and for the PPCA-EM algorithm. Results are shown in Figure 4.



(a) PCA without missing data (b) PCA with 15% missing values (c) PCA with 30% missing values

(d) PPCA without missing data (e) PPCA with 15% missing values (f) PPCA with 30% missing values

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

As expected, without any missing values, PCA and PPCA provide similar results shown in Figures 4(a) and 4(d), PCA providing projections with a larger variance. But when some entries are missing, PPCA clearly outperforms standard PCA. In Figures 4(e) and 4(f), the three clusters remain visible even with many missing values. On the other hand, the data gets much more mixed up with standard PCA, as shown in Figures 4(b) and 4(e). This experiment shows that PPCA is able to detect a kind of structure in the dataset, and use it to impute missing values accordingly.

4.3 PCCA with missing data

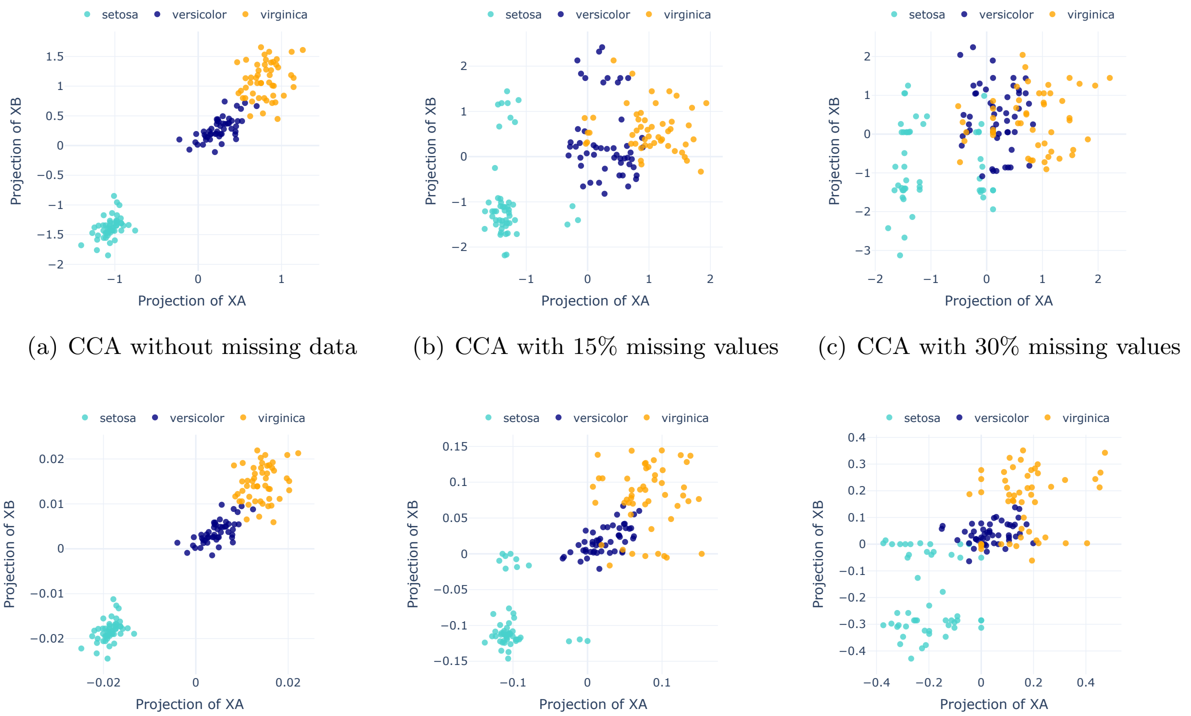
4.3.1 Method

The PCCA-EM method is similar to the one used for PPCA. We provide in Appendix D.2 the formulas used in the E-step, based on the work of Bach and Jordan [1]. The pseudo-code of the PCCA-EM algorithm is given in Algorithm 2 in Appendix D.2.

[1] NB: Class labels were not used for missing data imputation, but only for visualization purposes.

4.3.2 Experiment on the Iris Dataset to Evaluate PCCA

To evaluate our PCCA algorithm, we split the Iris dataset into two subsets: one with petal and sepal lengths, and the other with petal and sepal widths. We expected CCA and PCCA to reveal strong linear correlations. We used `scikit-learn` for CCA and our EM implementation for PCCA, after introducing 0%, 15%, and 30% missing values. The results are presented in Figure 5.



(a) CCA without missing data (b) CCA with 15% missing values (c) CCA with 30% missing values

(d) PCCA without missing data (e) PCCA with 15% missing values (f) PCCA with 30% missing values

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

Figures 5(a) and 5(d) show that with complete data, both methods yield the same output[2] with clear clusters and strong linear correlation. But as the level of missing values increases, CCA performance deteriorates significantly, with overlapping clusters and weaker correlation in Figures 5(b) and 5(c). PCCA demonstrates greater robustness to missing data in Figures 5(e) and 5(f), maintaining better class separation and correlation. Table 1 in Appendix D.3 provides the corresponding correlations.

5 Conclusion and Discussion

In this report, we covered PPCA and PCCA with a particular focus on handling missing values. We proposed a new EM algorithm for PCCA with missing data. Our experiments on the Iris dataset demonstrate that the probabilistic methods outperform their standard counterparts.

However, we acknowledge the limitations of the approach, in particular the strong assumptions of the model. Specifically, assuming a Gaussian observation model for the data and exploring only linear relationships between variables might limit its applicability.

[2] Bach and Jordan [1] proved that standard PCCA provides the same canonical directions as CCA.

References

[1] Francis R. Bach and Michael I. Jordan. “A Probabilistic Interpretation of Canonical Correlation Analysis”. **in**2005: URL: https://api.semanticscholar.org/CorpusID:17380611.

[2] Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data.* Wiley, 1987.

[3] M. E. Tipping and Christopher Bishop. “Probabilistic Principal Component Analysis”. **in***Journal of the Royal Statistical Society, Series B*: 21.3 (**january** 1999), **pages** 611–622. URL: https://www.microsoft.com/en-us/research/publication/probabilistic-principal-component-analysis/.

[4] L Yu, RR Snapp, T Ruiz and M Radermacher. “Probabilistic Principal Component Analysis using Expectation Maximization (PPCA-EM) for Analyzing 3D Volumes with Missing Data”. **in***Microscopy and Microanalysis*: 16.S2 (2010), **pages** 836–837. DOI: 10.1017/S143192761005734X.

Notations

Bold capital letters denote matrices, bold lower-case letters denote vectors and normal lower-case letters denote scalars.

$\mathbf{X} \in \mathbb{R}^{n \times d}$	Dataset with n observations and d features.
$\mathbf{z} \in \mathbb{R}^{q}$	Latent variables.
$\mathbf{w} \in \mathbb{R}^{d}$	1D vectors to project onto.
$\mathbf{y} \in \mathbb{R}^{n}$	1D projected data.
$\mathbf{W} \in \mathbb{R}^{d \times q}$	Loadings matrix.
$\boldsymbol{\varepsilon} \in \mathbb{R}^{d}$	Noise.
$\boldsymbol{\mu} \in \mathbb{R}^{d}$	Mean.
$\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$	Sample covariance matrix.
ρ	Lagrange multipliers.
λ	Eigenvalues.
r	Correlation coefficient.
k	Subspace dimension for projections in higher dimensions.
$\langle \cdot \rangle$	Expected value.

Appendix

Appendices A and C cover the theory behind PCA and CCA. Appendices B and D provide further details on PCCA and PCCA, along with the pseudo-code of our EM algorithm.

A Principal Component Analysis

A.1 Derivation of the First Principal Component

In the following, we provide a mathematical derivation of the first principal component. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a dataset with n observations and d features. For the sake of simplicity, we assume that the features of $\mathbf{X}$ are centered.

The goal of one-dimensional PCA is to find a vector $\mathbf{w} \in \mathbb{R}^d$ such that the linear projections $\mathbf{y} = \mathbf{X}\mathbf{w} \in \mathbb{R}^n$ have maximal empirical variance, defined as:

$$\mathrm{v}(\mathbf{y}) = \frac{1}{n}\|\mathbf{X}\mathbf{w}\|_2^2 = \frac{1}{n}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \tag{14}$$

with $\mathbf{\Sigma} = \frac{1}{n}\mathbf{X}^\top \mathbf{X}$ the sample covariance matrix.

To ensure that the problem is well posed, we impose a normalization constraint on $\mathbf{x}$. The resulting optimization problem becomes:

$$(\mathcal{P}_{\text{PCA}}) : \max_{\|\mathbf{w}\|_2 = 1} \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} \tag{15}$$

The Lagrangian $\mathcal{L}$ of $(\mathcal{P}_{\text{PCA}})$ is:

$$\mathcal{L}(\mathbf{w}, \rho) = \mathbf{w}^\top \mathbf{\Sigma} \mathbf{w} - \rho(\mathbf{w}^\top \mathbf{w} - 1) \tag{16}$$

Taking the gradient of $\mathcal{L}$ with respect to $\mathbf{w}$ and setting it to zero gives:

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \rho) = 2\mathbf{\Sigma}\mathbf{w} - 2\rho\mathbf{w} = 0 \implies \mathbf{\Sigma}\mathbf{w} = \rho\mathbf{w} \tag{17}$$

This is a standard eigenvalue problem. Any $\mathbf{w}$ such that $\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}, \rho) = 0$ verifies $\mathbf{\Sigma}\mathbf{w} = \rho\mathbf{w}$, and then $\mathbf{w}^\top \mathbf{\Sigma}\mathbf{w} = \rho\mathbf{w}^\top \mathbf{w} = \rho$ from the normalization constraint.

Hence the optimal $\mathbf{w}^*$ maximizing the variance of the projections is the eigenvector of $\mathbf{\Sigma}$ corresponding to its largest eigenvalue λ_1.

A.2 Higher-Dimensional PCA

To project onto an k-dimensional subspace, the goal is to identify k orthogonal vectors $\mathbf{w}_1, \ldots, \mathbf{w}_k$ that maximize variance. It can be shown that these vectors correspond to the eigenvectors of $\mathbf{\Sigma}$ associated with its top k eigenvalues. The proof is available at this **link**. The total variance explained by the first k principal components is $\sum_{i=1}^{k} \lambda_i$.

B Probabilistic PCA

B.1 Background on Factor Analysis

Factor Analysis is a simple latent variable model which assumes a linear relationship between the latent and observed variables:

$$\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon} \tag{18}$$

with latent $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_q)$ and noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\psi})$ with $\boldsymbol{\psi}$ diagonal.

Then the resulting observation model is Gaussian:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \boldsymbol{\psi}) \tag{19}$$

For a general $\boldsymbol{\psi}$, there is no closed-form solution for ML estimators. It is possible to estimate them via an EM algorithm.

Probabilistic PCA turns out to be a special case of Factor Analysis where $\boldsymbol{\psi} = \sigma^2 \boldsymbol{I_d}$ for $\sigma^2 > 0$. For this particular configuration, there exists an explicit formula for ML estimators.

B.2 Explicit Solution via Maximum Likelihood Estimation

For a set of n samples $(\mathbf{x}_i)_{i=1,\dots,n}$, we note Σ the sample covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top$$

Then, the log-likelihood of the observed variables following $\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ with $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I_d}$ is:

$$\mathcal{L}(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) = -\frac{n}{2}\left\{d \ln 2\pi + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\boldsymbol{\Sigma})\right\}$$

By considering partial derivatives of the log-likelihood, the MLE for $\boldsymbol{\mu}$ is the sample mean of the observed variables:

$$\hat{\boldsymbol{\mu}}_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \tag{20}$$

A main difference with factor analysis is that we can also obtain explicit MLEs for $\mathbf{W}$ and σ^2:

$$\hat{\mathbf{W}}_{\mathrm{ML}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \sigma^2\mathbf{I}_d)^{1/2}\mathbf{R} \tag{21}$$

where $\mathbf{U_q}$ contains the q principal components of $\boldsymbol{\Sigma}$ as columns, $\boldsymbol{\Lambda_q}$ is the diagonal matrix with corresponding eigenvalues and $\mathbf{R}$ is an arbitrary rotation matrix. In practice, we use $\mathbf{R} = \mathbf{I}_d$.

For the MLE of σ, by noting $\lambda_{q+1}, \dots, \lambda_d$ the $d - q$ non-selected eigenvalues of Σ:

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d}\lambda_j \tag{22}$$

Interestingly, $\hat{\sigma}_{\mathrm{ML}}$ can be interpreted as the variance lost in the projection, divided by the number of lost dimensions.

B.3 PPCA EM Algorithm with Missing Values

We propose an EM algorithm for PPCA when data $\mathbf{X} = (\mathbf{x}_i)_{i=1,\dots,n}$ have missing values. We follow the same approach as Little and Rubin [2].The pseudo-code is given in Algorithm 1.

Algorithm 1 PPCA EM Algorithm with Missing Values

Input: Data matrix $X \in \mathbb{R}^{n \times d}$ with missing values.
Parameters: Number of latent dimensions q, maximum iterations max_iter, tolerance tol.
Output: Parameters $\mathbf{W}$, σ^2 and $\boldsymbol{\mu}$.

```
Initialize $\mathbf{W} \in \mathbb{R}^{d \times k}$ randomly, $\sigma^2 > 0$, and $\boldsymbol{\mu} \leftarrow \text{mean}(X)$.
while not converged and iteration < max_iter do
    E-Step:
    for each sample $\mathbf{x} \in X$ do
        Estimate latent variable $\langle \mathbf{z} \rangle$: $\langle \mathbf{z} \rangle \leftarrow \left(\mathbf{W}_{\text{obs}}^\top \mathbf{W}_{\text{obs}} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{W}_{\text{obs}}^\top (\mathbf{x}_{\text{obs}} - \boldsymbol{\mu}_{\text{obs}})$
        Estimate covariance of latent variables: $\langle \mathbf{z}\mathbf{z}^\top \rangle \leftarrow \sigma^2 \left(\mathbf{W}_{\text{obs}}^\top \mathbf{W}_{\text{obs}} + \sigma^2 \mathbf{I}\right)^{-1} + \langle \mathbf{z} \rangle \langle \mathbf{z} \rangle^\top$
        Impute missing values: $\langle \mathbf{x}_{\text{miss}} \rangle \leftarrow \mathbf{W}_{\text{miss}} \langle \mathbf{z} \rangle + \boldsymbol{\mu}_{\text{miss}}$
    end for
    M-Step:
    Update mean: $\boldsymbol{\mu} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{x}_i \rangle$
    Update $\mathbf{W}$ using:$\mathbf{W} \leftarrow \left( \sum_{i=1}^{n} (\langle \mathbf{x}_i \rangle - \boldsymbol{\mu}) \langle \mathbf{z}_i \rangle^\top \right) \left( \sum_{i=1}^{n} \langle \mathbf{z}_i \mathbf{z}_i^\top \rangle \right)^{-1}$
    Update $\sigma^2$: $\sigma^2 \leftarrow \frac{1}{nd} \sum_{i=1}^{n} \left( \|\langle \mathbf{x}_i \rangle - \boldsymbol{\mu}\|^2 - 2 \langle \mathbf{z}_i \rangle^\top \mathbf{W}^\top (\langle \mathbf{x}_i \rangle - \boldsymbol{\mu}) + \text{tr}(\mathbf{W}^\top \mathbf{W} \langle \mathbf{z}_i \mathbf{z}_i^\top \rangle) \right)$
end while
Return: $\mathbf{W}$, $\sigma^2$, $\boldsymbol{\mu}$.
```

C Canonical Correlation Analysis

Let $\mathbf{X}_A \in \mathbb{R}^{n \times d_A}$ and $\mathbf{X}_B \in \mathbb{R}^{n \times d_B}$ two paired datasets, meaning that the i-th sample is the concatenation $\mathbf{x}_i = (\mathbf{x}_{A,i}, \mathbf{x}_{B,i}) \in \mathbb{R}^{d_A + d_B}$. We assume that all features are centered.

C.1 Derivation of the First Canonical Component

In the following, as for PCA, we provide a mathematical derivation of the canonical components in the case of 1D projections.

The goal of one-dimensional CCA is to find two vectors $\mathbf{w}_A \in \mathbb{R}^{d_A}$ and $\mathbf{w}_B \in \mathbb{R}^{d_B}$ such that the linear projections $\mathbf{y}_A = \mathbf{X}_A \mathbf{w}_A \in \mathbb{R}^n$ and $\mathbf{y}_B = \mathbf{X}_B \mathbf{w}_B \in \mathbb{R}^n$ are maximally correlated. The empirical correlation between these projections is given by Pearson's coefficient:

$$r_{\mathbf{y}_A \mathbf{y}_B} = \frac{\mathbf{y}_A^\top \mathbf{y}_B}{\|\mathbf{y}_A\|_2 \|\mathbf{y}_B\|_2} = \frac{\mathbf{w}_A^\top \boldsymbol{\Sigma}_{AB} \mathbf{w}_B}{\sqrt{\mathbf{w}_A^\top \boldsymbol{\Sigma}_{AA} \mathbf{w}_A} \sqrt{\mathbf{w}_B^\top \boldsymbol{\Sigma}_{BB} \mathbf{w}_B}} \tag{23}$$

with $\boldsymbol{\Sigma}_{ij} = \frac{1}{n} \mathbf{X}_i^\top \mathbf{X}_j$ for $i, j \in \{A, B\}$ the sample covariance matrices.

Imposing normalization constraints on $\mathbf{y}_A$ and $\mathbf{y}_B$, the resulting problem becomes:

$$(\mathcal{P}_{\text{CCA}}) : \max_{\|\mathbf{y}_A\|_2 = \|\mathbf{y}_B\|_2 = 1} \mathbf{w}_A^\top \mathbf{\Sigma}_{AB} \mathbf{w}_B \tag{24}$$

To solve this problem, we can again use duality and Lagrange multipliers:

$$\mathcal{L}(\mathbf{w}_A, \mathbf{w}_B, \rho_A, \rho_B) = \mathbf{w}_A^\top \mathbf{\Sigma}_{AB} \mathbf{w}_B - \rho_A \left(\mathbf{w}_A^\top \mathbf{\Sigma}_{AA} \mathbf{w}_A - 1 \right) - \rho_B \left(\mathbf{w}_B^\top \mathbf{\Sigma}_{BB} \mathbf{w}_B - 1 \right) \tag{25}$$

Taking the gradients w.r.t. $\mathbf{w}_A$ and $\mathbf{w}_B$ and setting them to 0, we get:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_A} = \mathbf{\Sigma}_{AB} \mathbf{w}_B - 2\rho_A \mathbf{\Sigma}_{AA} \mathbf{w}_A = 0 \tag{26}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}_B} = \mathbf{\Sigma}_{BA} \mathbf{w}_A - 2\rho_B \mathbf{\Sigma}_{BB} \mathbf{w}_B = 0 \tag{27}$$

By multiplying on the left Equation (26) by $\mathbf{w}_A^\top$ and Equation (27) by $\mathbf{w}_B^\top$, and using the normalization constraints, we first show that:

$$\rho_A = \rho_B = \frac{1}{2} \mathbf{w}_A^\top \mathbf{\Sigma}_{AB} \mathbf{w}_B \coloneqq \rho \tag{28}$$

Then, solving for $\mathbf{w}_A$ in Equation (26):

$$\mathbf{w}_A = \frac{\mathbf{\Sigma}_{AA}^{-1} \mathbf{\Sigma}_{AB} \mathbf{w}_B}{\rho}$$

And substituting into Equation (27), with $\tilde{\mathbf{\Sigma}} = \mathbf{\Sigma}_{BB}^{-1} \mathbf{\Sigma}_{BA} \mathbf{\Sigma}_{AA}^{-1} \mathbf{\Sigma}_{AB}$, we end up with a standard eigenvalue problem:

$$\tilde{\mathbf{\Sigma}} \mathbf{w}_B = \rho^2 \mathbf{w}_B \tag{29}$$

From Equations (28) and (29), we conclude that the canonical components correspond to the eigenvectors associated with the largest eigenvalues of $\tilde{\mathbf{\Sigma}}$.

NB: Seeing the objective of $(\mathcal{P}_{\text{CCA}})$ as $\mathbf{y}_A^\top \mathbf{y}_B$ provides a nice geometric interpretation to CCA: it aims at getting projected data vectors pointing in the same direction.

C.2 Higher-Dimensional CCA

In q-dimensional CCA, the goal is to find q pairs of directions $\{\mathbf{w}_{A,i}, \mathbf{w}_{B,i}\}_{i=1}^q$ such that the projections are maximally correlated and orthogonal to previously identified components. The maximum number of canonical components is then:

$$q_{\max} = \min(d_A, d_B) \tag{30}$$

See this **link** for mathematical details on this.

D Probabilistic CCA

D.1 EM algorithm for PCCA

We denote by $\mathbf{\Sigma}$ the covariance matrix of the paired-datasets under the model:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{W}_A\mathbf{W}_A^\top + \boldsymbol{\Psi}_A & \mathbf{W}_A\mathbf{W}_B^\top \\ \mathbf{W}_B\mathbf{W}_A^\top & \mathbf{W}_B\mathbf{W}_B^\top + \boldsymbol{\Psi}_B \end{pmatrix} \tag{31}$$

Bach and Jordan [1] show that an EM algorithm is obtained by repeating the following updates, with $\mathbf{M}_t = \big(\mathbf{I} + \mathbf{W}_t^\top \boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\big)^{-1}$:

$$\mathbf{W}_{t+1} = \boldsymbol{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\big(\mathbf{M}_t + \mathbf{M}_t\mathbf{W}_t^\top\boldsymbol{\Psi}_t^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\big)^{-1} \tag{32}$$

$$\boldsymbol{\Psi}_{t+1} = \begin{pmatrix} \big(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\mathbf{W}_t^\top\big)_{AA} & 0 \\ 0 & \big(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Psi}_t^{-1}\mathbf{W}_t\mathbf{M}_t\mathbf{W}_t^\top\big)_{BB} \end{pmatrix} \tag{33}$$

D.2 EM for PCCA with Missing Values

Based on the same idea as for PPCA, we propose an EM algorithm for PCCA with missing values. To this end, we introduce the following notations: $\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{AA} & \boldsymbol{\Sigma}_{AB} \\ \boldsymbol{\Sigma}_{BA} & \boldsymbol{\Sigma}_{BB} \end{pmatrix}$ is the $(d_A + d_B) \times (d_A + d_B)$ sample covariance matrix. By noting respectively $\mathbf{V}_A$ and $\mathbf{V}_B$ the left and right singular vectors (in columns) of $(\boldsymbol{\Sigma}_{AA})^{-1/2}\boldsymbol{\Sigma}_{AB}(\boldsymbol{\Sigma}_{BB})^{-1/2}$, $\mathbf{U}_A \in \mathbb{R}^{d_A \times q}$ and and $\mathbf{U}_B \in \mathbb{R}^{d_B \times q}$ contain respectively the first q columns of $(\boldsymbol{\Sigma}_{AA})^{-1/2}\mathbf{V}_A$ and $(\boldsymbol{\Sigma}_{BB})^{-1/2}\mathbf{V}_B$. Finally, we define:

$$\boldsymbol{\Psi} = \begin{pmatrix} \boldsymbol{\Psi}_A & 0 \\ 0 & \boldsymbol{\Psi}_B \end{pmatrix} \tag{34}$$

$$\mathbf{M} = \big(\mathbf{I}_q + \mathbf{W}^\top\boldsymbol{\Psi}^{-1}\mathbf{W}\big)^{-1} \tag{35}$$

$$\mathbf{M}_A = \big(\mathbf{I}_q + \mathbf{W}_A^\top\boldsymbol{\Psi}_A^{-1}\mathbf{W}_B\big)^{-1} \tag{36}$$

$$\mathbf{M}_B = \big(\mathbf{I}_q + \mathbf{W}_B^\top\boldsymbol{\Psi}_B^{-1}\mathbf{W}_B\big)^{-1} \tag{37}$$

Then, for each $\mathbf{x} = (\mathbf{x}_A, \mathbf{x}_B) = ((\mathbf{x}_{A:\text{obs}}, \mathbf{x}_{A:\text{miss}}), (\mathbf{x}_{B:\text{obs}}, \mathbf{x}_{B:\text{miss}}))$:

$$\langle \mathbf{z}_A \rangle = \mathbf{M}_A^\top\mathbf{U}_{A:\text{obs}}^\top(\mathbf{x}_{A:\text{obs}} - \boldsymbol{\mu}_{A:\text{obs}}) \tag{38}$$

$$\langle \mathbf{z}_B \rangle = \mathbf{M}_B^\top\mathbf{U}_{B:\text{obs}}^\top(\mathbf{x}_{B:\text{obs}} - \boldsymbol{\mu}_{B:\text{obs}}) \tag{39}$$

$$\langle \mathbf{x}_A \rangle = (\mathbf{x}_{A:\text{obs}}, \langle \mathbf{x}_{A:\text{miss}} \rangle) = (\mathbf{x}_{A:\text{obs}}, \mathbf{W}_{A:\text{miss}}\langle \mathbf{z} \rangle + \boldsymbol{\mu}_{A:\text{miss}}) \tag{40}$$

$$\langle \mathbf{x}_B \rangle = (\mathbf{x}_{B:\text{obs}}, \langle \mathbf{x}_{B:\text{miss}} \rangle) = (\mathbf{x}_{B:\text{obs}}, \mathbf{W}_{B:\text{miss}}\langle \mathbf{z} \rangle + \boldsymbol{\mu}_{B:\text{miss}}) \tag{41}$$

With these notations, the pseudo-code is provided in Algorithm 2.

D.3 Correlation Results

On the Iris dataset, we observe in Table 1 that PCCA is better than CCA at preserving the correlation of the linear projections when the proportion of missing data increases.

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

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

Algorithm 2 PCCA EM Algorithm with Missing Values

Input: Datasets $\mathbf{X}_A \in \mathbb{R}^{n\times d_A}$, $\mathbf{X}_B \in \mathbb{R}^{n\times d_B}$ with missing values.
Output: Parameters $\mathbf{W}$, $\boldsymbol{\Psi}$, $\mathbf{U}_A$, $\mathbf{U}_B$.

```
Initialize $\mathbf{W}$, $\boldsymbol{\Psi}$, $\mathbf{U}_A$, $\mathbf{U}_B$, and $\mathbf{M}$.
Compute initial means $\boldsymbol{\mu}_A, \boldsymbol{\mu}_B$ and covariance $\boldsymbol{\Sigma}$ from $X_A$, $X_B$.
while not converged and iteration < max_iter do
    E-Step:
    for each sample $\mathbf{x} \in [X_A, X_B]$ do
        Fill missing values in $\mathbf{x}$ using conditional expectations.
    end for
    Update $\boldsymbol{\Sigma}$ using the filled data.
    Update $\mathbf{U}_A$, $\mathbf{U}_B$ via SVD of $\boldsymbol{\Sigma}$ submatrices.
    M-Step:
    Update $\mathbf{M}$: $\mathbf{M} \leftarrow \big(\mathbf{I} + \mathbf{W}^\top \boldsymbol{\Psi}^{-1} \mathbf{W}\big)^{-1}$
    Update $\mathbf{W}$: $\mathbf{W} \leftarrow \boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{M}\big(\mathbf{M} + \mathbf{M}\mathbf{W}^\top \boldsymbol{\Psi}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{M}\big)^{-1}$
    Update $\boldsymbol{\Psi}$: $\boldsymbol{\Psi} \leftarrow \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Psi}^{-1}\mathbf{W}\mathbf{M}\mathbf{W}^\top$
    Set cross-covariance blocks of $\boldsymbol{\Psi}$ to zero.
end while
Return: $\mathbf{W}$, $\boldsymbol{\Psi}$, $\mathbf{U}_A$, $\mathbf{U}_B$.
```