Introduction to Probabilistic Principal Component Analysis

by Le Hoai Nam, Huynh Hoang Trung Nghia,
Huynh Huu Nhat, Tran Trung Viet

May 2020

Content

- What is PPCA?
- Difference between PPCA and PCA.
- Difference between PPCA and Factor Analysis.
- Derivation of PPCA
- Applications

Review PCA

- Primary uses:
  - Analyze data and extract variables with similar concepts (principal components)
  - Project the data onto a lower dimensional space
  - Principal components which explain a greater amount of the variance are considered to be more important
- Accomplishes this by:
  - Maximizing variance of the projected data $\mathbf{x}$
  - Represent matrix $\mathbf{x}$ in a different (q-dimensional) space using Weight matrix $\mathbf{W}$
    - $\mathbf{W}$ represents a re-mapping of original data $\mathbf{t}$ into its "ideal" principal subspace, represented by $\mathbf{x}$
    - $\{\mathbf{w}_1, ..., \mathbf{w}_q\}$ are the column vectors of $\mathbf{W}$ represents a separate principal component

From PCA to Probabilistic PCA

Limitations of PCA

- No probabilistic model for observed data
- Does not deal properly with missing data
- The outliers affects the model

Motivation behind Probabilistic PCA

- Addresses limitations of regular PCA
- Maximum-likelihood estimates can be computed for elements associated with principal components
- Comparison with other density-estimation techniques and facilitate statistical testing

Latent Variable Models

- Latent variable(s): unobserved variable (s)
  - offer a lower dimensional representation of the data and their dependencies
- Latent variable model:
  - $\mathbf{t}$: observed variables (d-dimensions)
  - $\mathbf{x}$: latent variables (q-dimensions)
  - $q < d$
- Purpose: offer a more parsimonious description of the data

Latent Variable models

$$\mathbf{t} = \mathbf{y}(\mathbf{x}; \mathbf{w}) + \epsilon \tag{1}$$

- $\mathbf{y}(\mathbf{x}; \mathbf{w})$ is a function of $\mathbf{x}$ with parameters $\mathbf{w}$
- ϵ is an x-independent noise process

Probabilistic PCA (PPCA)

- Latent variable model with linear relationship (factor analysis)
  - $\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$
  - Latent variables: $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$
  - $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$
  - mean of $\mathbf{t}$: $\boldsymbol{\mu}$
- PPCA: Noise variances constrained to be equal $\psi_i = \boldsymbol{\sigma}^2$ or $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ (isotropic noise model)
  - $\mathbf{t}|\mathbf{x} \sim N(\mathbf{W}\mathbf{x} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$
  - $\mathbf{t} \sim N(\boldsymbol{\mu}, \mathbf{C})$ where $\mathbf{C} = \mathbf{W}\mathbf{W}^{\mathsf{T}} + \sigma^2\mathbf{I}$ (the covariance matrix sample)
- PCA is a limiting case of PPCA, taken as the limit as the covariance of the noise becomes infinitesimally small

$$\boldsymbol{\Psi} = \lim_{\sigma^2 \to 0} \sigma^2\mathbf{I}$$

Probabilistic PCA - Maximum likelihood

- Log-likelihood for Gaussian noise model:

$$\mathcal{L} = -\frac{N}{2}\{d\ \ln(2\pi) + \ln|\mathbf{C}| + \mathrm{tr}(\mathbf{C}^{-1}\mathbf{S})\} \tag{2}$$

where $\mathbf{S}$ is sample covariance matrix of the observations $\mathbf{t}_n$ with mean vector $\boldsymbol{\mu}$

$$\mathbf{S} = \frac{1}{N}\sum_{k=1}^{N}(\mathbf{t}_n - \boldsymbol{\mu})(\mathbf{t}_n - \boldsymbol{\mu})^{\mathsf{T}} \tag{3}$$

- MLE's for $\mathbf{W}$ and σ^2 can be solved in two ways:
  - Closed-form (Tipping and Bishop)
  - Expectation Maximization (EM) algorithm (Roweis)

MLE for $\mathbf{W}$ and σ^2

Likelihood of $\mathcal{L}$ is maximized with respect to $\mathbf{W}$ and σ^2, MLE's can be obtained in closed form:

$$\sigma^2_{\text{ML}} = \frac{1}{d-q}\sum_{j=q+1}^{d}(\lambda_j), \text{ and } \mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma^2_{ML}\mathbf{I})^{\frac{1}{2}}\mathbf{R} \quad (4)$$

- σ^2_{ML} variance lost in the projection, averaged over dims decreased
- $\mathbf{W}_{ML}$ the mapping of the latent space (containing $\mathbf{x}$) to that of the principal subspace (containing $\mathbf{t}$)
- Columns $\{\mathbf{u}_1, ..., \mathbf{u}_q\}$ of $\mathbf{U}_q$ (d x q matrix): principal eigenvectors of $\mathbf{S}$
- $\mathbf{\Lambda}_q$ (q x q diagonal matrix): corresponding principal eigenvalues $\{\lambda_1, ..., \lambda_q\}$ of $\mathbf{S}$
- $\mathbf{R}$: q x q arbitrary rotation matrix (can be set to $\mathbf{R} = \mathbf{I}$)

Difference between PPCA and PCA

Dimensionality reduction and optimal reconstruction

- In conventional PCA, the reduced-dimensionality transformation of a data point of $\mathbf{t}_n$ is $\mathbf{x}_n = \mathbf{U}_q^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$ and its reconstruction is $\hat{\mathbf{t}}_n = \mathbf{U}_q \mathbf{x}_n + \boldsymbol{\mu}$.
- In PPCA, using Bayes rule, we can obtain a posterior estimate of the latent variables:
  - $\mathbf{x}|\mathbf{t} = \mathcal{N}(\mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t} - \boldsymbol{\mu}), \boldsymbol{\sigma}^2\mathbf{M}^{-1})$, where $\mathbf{M} = \mathbf{W}^{\mathrm{T}}\mathbf{W} + \sigma^2\mathbf{M}^{-1}$, $\mathbf{M}$ is a $q \times q$ matrix
  - Posterior mean: $\mathbf{E}[\mathbf{x}|\mathbf{t}] = \langle \mathbf{x}_n | \mathbf{t}_n \rangle = \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}(\mathbf{t}_n - \boldsymbol{\mu})$
- Reconstruction of the observed data with respect to the new subspace:
  - The latent projection of conventional PCA is skewed towards the origin (due to marginal distribution for $\mathbf{x}$), so the reconstruction
    - $\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}}\mathbf{x}_n + \boldsymbol{\mu}$ is not orthogonal and thus is not optimal
  - Optimal reconstruction of the observed data may still be obtained from conditional latent mean:
    - $\hat{\mathbf{t}}_n = \mathbf{W}_{\mathrm{ML}}(\mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\mathbf{W}_{\mathrm{ML}})^{-1}\mathbf{M}\langle \mathbf{x}_n | \mathbf{t}_n \rangle + \boldsymbol{\mu}$

Difference between PPCA and Factor Analysis

Factor Analysis

The Factor Analysis concept is very familiar to PPCA. Factor analysis is designed to factorize (described) the variation of observed variables by fewer number of unobserved variables. The observed variables is modelized by the linear combination of latent variables.

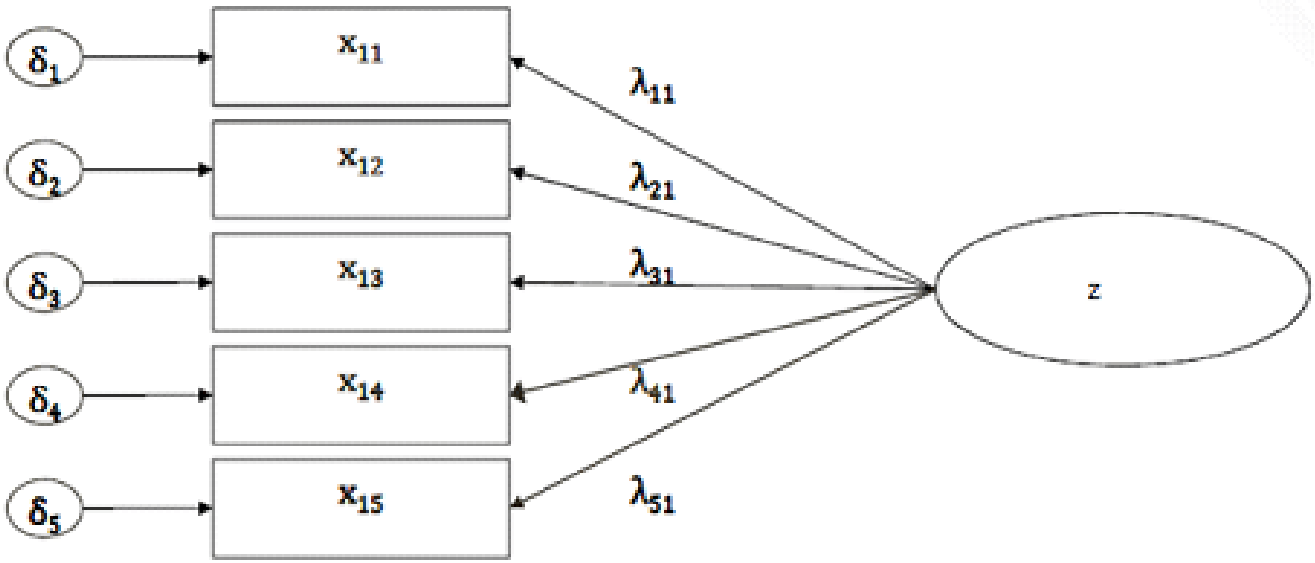

Figure: The basic model of Factor analysis

Difference between PPCA and Factor Analysis

Factor Analysis

The assumption of distribution of Factor Analysis:

$$\mathbf{t}|\mathbf{x} \sim \mathcal{N}(\mathbf{W}\mathbf{x}, \mathbf{D}), \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{5}$$

So we have:

$$\mathbf{t} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{W}\mathbf{W}^T + \mathbf{D}\right) \tag{6}$$

With $\mathbf{D}$ is the diagonal covariance matrix

Probabilistic PCA

On the other hands, the PPCA assumption distribution is:

$$\mathbf{t}|\mathbf{x} \sim \mathcal{N}\left(\mathbf{W}\mathbf{x}, \sigma^2\mathbf{I}\right), \quad \mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

Difference between PPCA and Factor Analysis

The difference

The difference due to the to difference assumption of ϵ terms in:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \tag{8}$$

Where PPCA we have $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right)$ and factor analysis $\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{D}\right)$

Difference between PPCA and Factor analysis

- Covariance
  - PPCA (and standard PCA) is covariant under rotation of the original data axes
  - Factor analysis is covariant under component-wise rescaling
- Principal components
  - In PPCA: different principal components (axes) can be found incrementally
  - Factor analysis: factors from a two-factor model may not correspond to those from a one-factor model

Difference between PPCA and Factor analysis

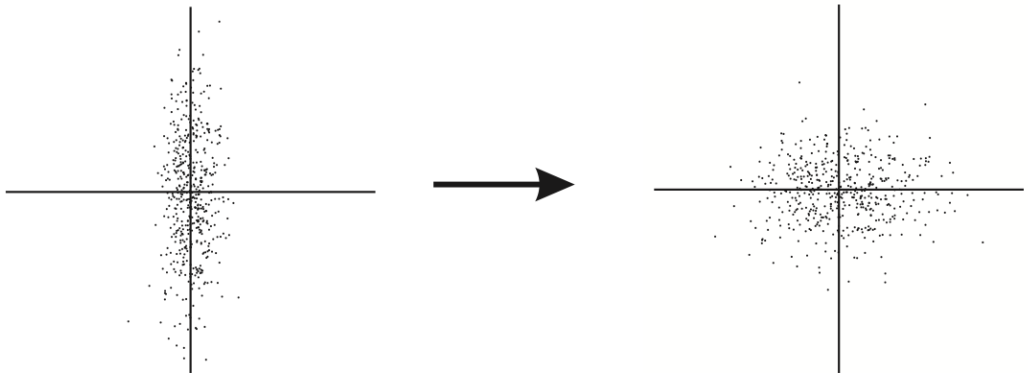

Figure 1: Factor analysis is covariant under a component-wise rescaling of the data variables

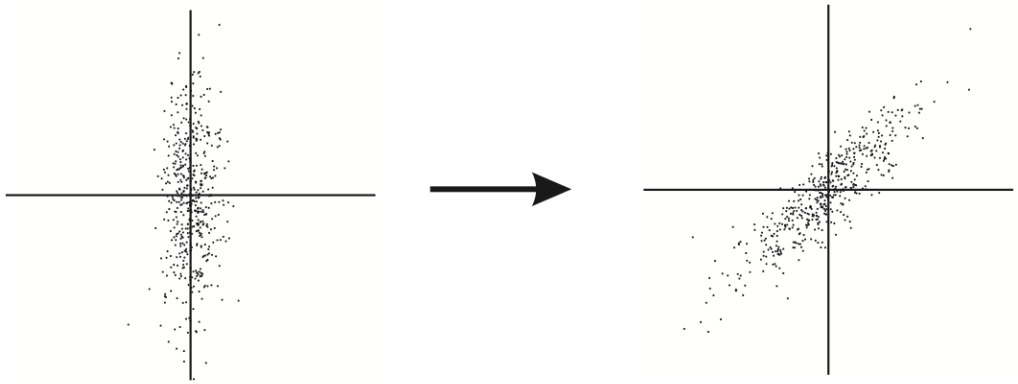

Figure 2: PCA and PPCA are covariant under rotations of the data space coordinates

EM concept

The Expectation maximization algorithm is an alternative method for estimating MLE. There often be such cases that it is very hard to optimize the likelihood function, such as in PPCA, we maximize the log-likelihood function of marginal conditional distribution of $P(t|x)$:

$$argmaxP(t|\mathbf{W},\sigma^2) = argmax\prod_{i=0}^{n}\int P(t|x,\mathbf{W},\sigma^2)dx \tag{9}$$

Assuming that we know or guess the distribution of x $q(x)$. The EM use the "coordinate ascent" to estimate the parameter

Derivation of PPCA

Expectation Maximization Method

EM concept

Form the "coordinate ascent", we can derive the EM algorithms which have the E-step and M-step.

- E-step: Calculate the complete data log-likelihood
- M-step: To maximize the result in E-step

EM of PPCA

In the PPCA paper, the E-step and M-step are encapsulated into the iteration of these two equations:

$$\tilde{\mathbf{W}} = \mathbf{S}\mathbf{W}\left(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{S}\mathbf{W}\right)^{-1} \tag{10}$$

$$\tilde{\sigma}^2 = \frac{1}{d}\,\mathrm{tr}\left(\mathbf{S} - \mathbf{S}\mathbf{W}\mathbf{M}^{-1}\widetilde{\mathbf{W}}^{\mathrm{T}}\right) \tag{11}$$

Application

PPCA Application

Due to the employment of EM into the PPCA, so there are serveral application of PPCA in practice.

- More effective dimension reduction with large dataset than the MLE closed form.
- Dimension reduction with missing data.
- Missing data imputation.
- Outlier detection.

Application

Dimension reduction with missing dataset

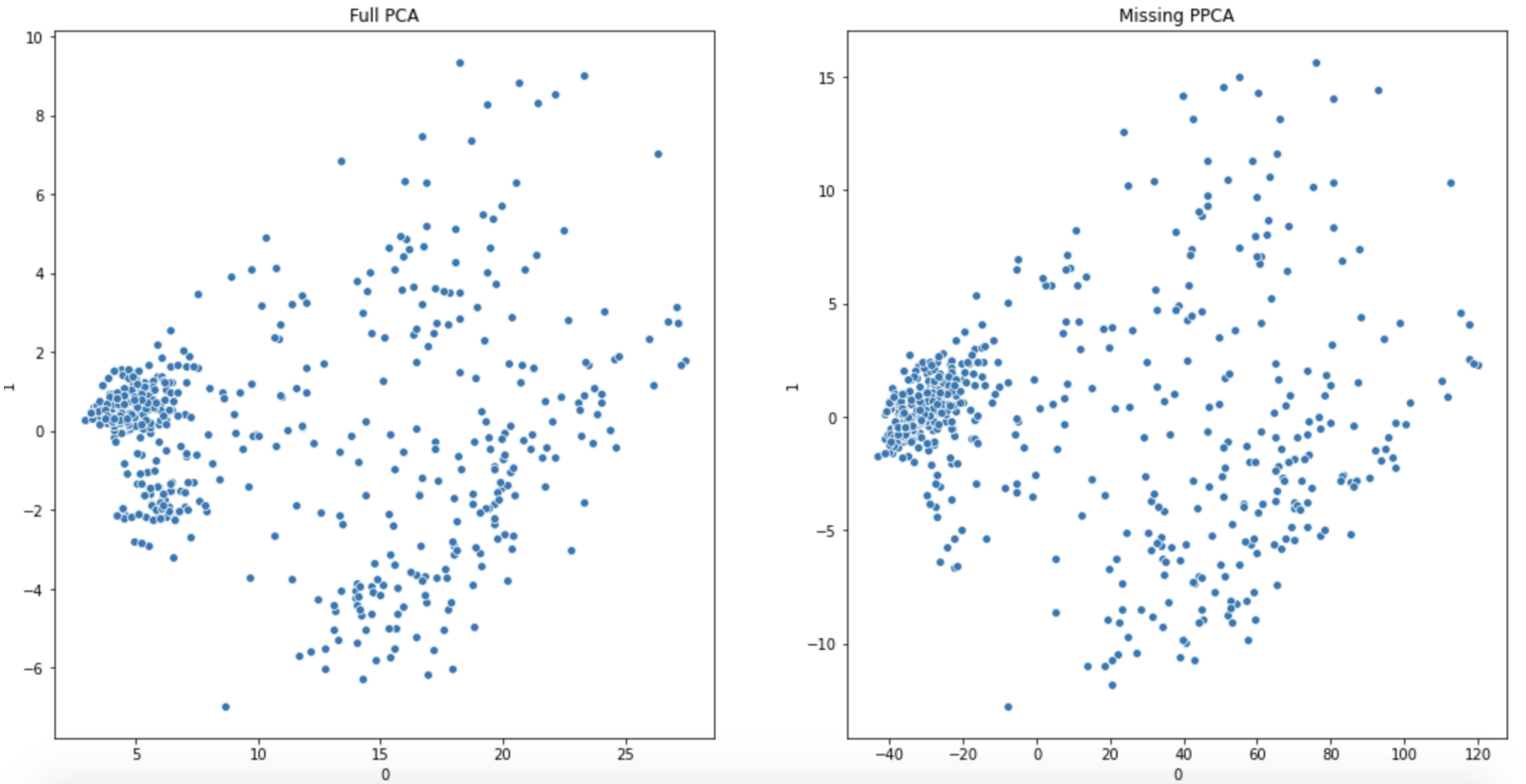

Figure: Comparison between Full data PCA vs Missing data PPCA

Application

Missing data imputation

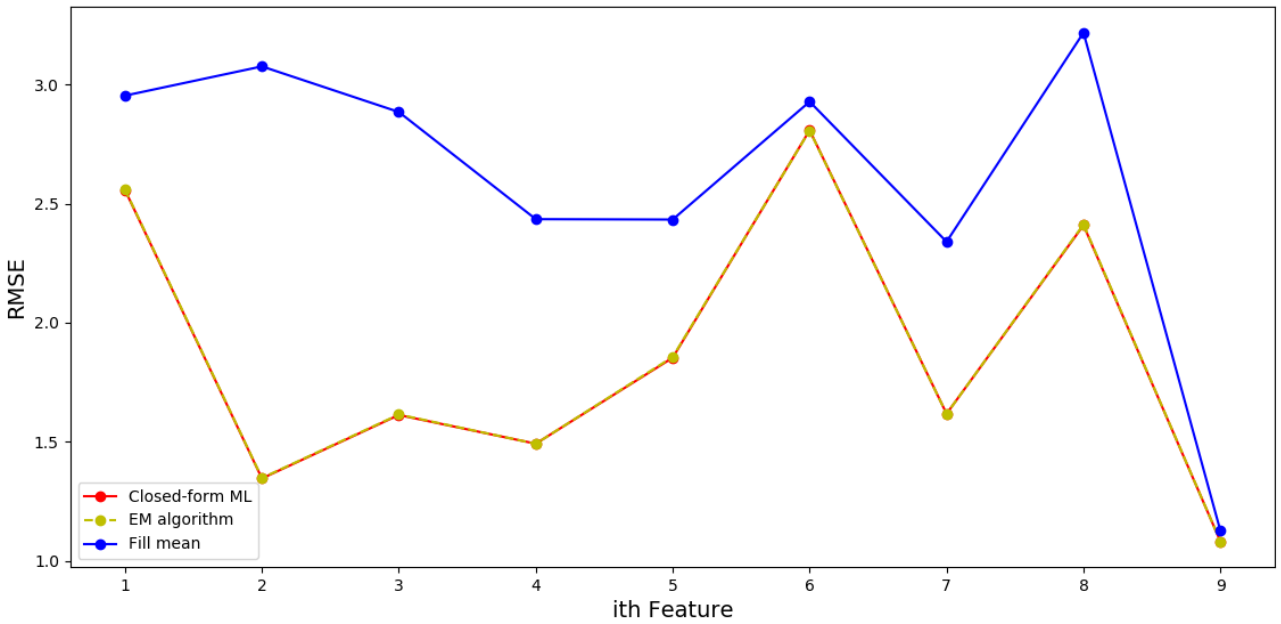

Figure: PPCA compared to fixed filled with Mean

Appendix A: Derivation of MLEs

- The derivative of $\mathcal{L} = -\frac{N}{2}\{d\ \ln(2\pi) + \ln|\mathbf{C}| + \text{tr}(\mathbf{C}^{-1}\mathbf{S})\}$ w.r.t $\mathbf{W}$:
  - $\partial\mathcal{L}\ \partial\mathbf{W} = N(\mathbf{C}^{-1}\mathbf{S}\mathbf{C}^{-1}\mathbf{W} - \mathbf{C}^{-1}\mathbf{W})$ where $\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^T = \mathbf{W}\mathbf{W}^T + \sigma^2$
- The stationary points are $\mathbf{S}\mathbf{C}^{-1}\mathbf{W} = \mathbf{W}$
  - Non-trivial case: $\mathbf{W} \neq \mathbf{0}$ and $\mathbf{C} \neq \mathbf{S}$
- SVD: $\mathbf{W} = \mathbf{U}\mathbf{L}\mathbf{V}^T$, $\mathbf{U}$: d x q orthonormal vectors, $\mathbf{L}$: q x q matrix of singular values, $\mathbf{V}$: q x q orthogonal matrix.
  - $\mathbf{C}^{-1}\mathbf{W} = (\sigma^2\mathbf{I} + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} = \mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^T$
- At the stationary points:
  - $\mathbf{S}\mathbf{U}\mathbf{L}(\sigma^2\mathbf{I} + \mathbf{L}^2)^{-1}\mathbf{V}^T = \mathbf{U}\mathbf{L}\mathbf{V}^T \Rightarrow \mathbf{S}\mathbf{U}\mathbf{L} = \mathbf{U}(\sigma^2\mathbf{I} + \mathbf{L}^2)\mathbf{L}$
- Column vectors of $\mathbf{U}, \mathbf{u}_j$, are eigenvectors of $\mathbf{S}$, with eigenvalue λ_j, such that $\sigma^2 + l_j^2 = \lambda_j$. So, $l_j = (\lambda_j^2 - \sigma^2)^{\frac{1}{2}}$
- Substitute into SVD, we have:

$$\mathbf{W}_{ML} = \mathbf{U}_q(\mathbf{\Lambda}_q - \sigma_{ML}^2\mathbf{I})^{\frac{1}{2}}\mathbf{R} \tag{12}$$

Appendix A: Derivation of MLEs

- Substitute equation (12) into original $\mathcal{L}$ expression

$$\mathcal{L} = -\frac{N}{2}\left\{ d ln(2\pi) + \sum_{j=1}^{q} ln(\lambda_j) + \frac{1}{\sigma^2} \sum_{j=q+1}^{d} \lambda_j + (d-q) ln(\sigma^2) + q \right\} \quad (13)$$

  - $\lambda_1, ..., \lambda_q$, are q non-zero eigenvalues of $\mathbf{u}_j$ and $\lambda_{q+1}, ..., q_d$, are zero
- Taking derivative of above with respect to σ^2 and solving for zero gives:

$$\sigma^2_{\text{ML}} = \frac{1}{d-q} \sum_{j=q+1}^{d} \lambda_j, \quad (14)$$

The End