Probabilistic Principal Component Analysis

Erik Bekkers

(Bishop 12.2.1)
Probabilistic PCA

- Probabilistic view of PCA:
  - Learn it via maximum likelihood
  - (Third) alternative view of PCA
  - Both latent and observed variables are Gaussian
Continuous latent variable model

- Data: \( \mathbf{X} = \{ \mathbf{x}_1, \ldots, \mathbf{x}_N \}, \mathbf{x}_n \in \mathbb{R}^D \)

- Goal: learn a \( M < D \) **continuous latent space** by maximizing the likelihood of the probabilistic model, \( M \) given

- Recall the continuous latent variable model

\[
p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}
\]

(Bishop 12.9)
PPCA modeling assumptions

- The **generative model** works as follows

\[ x = Wz + \mu + \epsilon \]

- With \( \mu \in \mathbb{R}^D \) and the **continuous latent variable** \( z \in \mathbb{R}^M \) with Gaussian prior

\[ p(z) = \mathcal{N}(z \mid 0, I) \]

- Matrix \( W \in \mathbb{R}^{D \times M} \) transforms the latent variables into observed variables.

- The independent noise is also Gaussian

\[ p(\epsilon) = \mathcal{N}(\epsilon \mid 0, \sigma^2 I) \]
It follows…

- The conditional distribution of the observed variable

\[ p(x | z) = \mathcal{N}(x | Wz + \mu, \sigma^2 I) \]

- The marginal \( p(x) \) is also a Gaussian (Bishop Ch. 2.3)

\[ p(x) = \int p(x, z)dz = \int p(x | z)p(z)dz = \mathcal{N}(x | \mu, \sigma^2 I) \]

- The expected value for \( x \sim p(x) \) is given by

\[ \mathbb{E}[x] = \mathbb{E}[Wz + \mu + \epsilon] = W\mathbb{E}[z] + \mu + \mathbb{E}[\epsilon] = \mu \]
It follows…

- The covariance of $x \sim p(x)$ is given by
  \[
  \text{Cov}[x] = \mathbb{E}[(x - \mu)(x - \mu)^T] \\
  = \mathbb{E}[(Wz + \epsilon)(Wz + \epsilon)^T] \\
  = \mathbb{E}[Wzz^T W^T + 2Wz \epsilon^T + \epsilon \epsilon^T] \\
  = W \mathbb{E}[zz^T] W^T + 2W \mathbb{E}[z \epsilon^T] + \mathbb{E}[\epsilon \epsilon^T] \\
  = WW^T + \sigma^2 I = C
  \]

- Therefore
  \[
  p(x) = \mathcal{N}(x | \mu, C)
  \]
Prior $p(z) = \mathcal{N}(z \mid 0, I)$

Likelihood $p(x \mid z) = \mathcal{N}(x \mid Wz + \mu, \sigma^2 I)$

Marginal $p(x) = \mathcal{N}(x \mid \mu, C)$
The log-likelihood

\[ \ln p(X | \mu, W, \sigma^2) = \sum_{n=1}^{N} \ln p(x_n | \mu, W, \sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}(x_n | \mu, C) \]

\[ = - \frac{ND}{2} \ln(2\pi) - \frac{N}{2} \ln |C| - \frac{1}{2} \sum_{n=1}^{N} (x_n - \mu)^T C^{-1} (x_n - \mu) \]

- Once again, take it’s derivative w.r.t. the parameter of interest, set it to zero, and solve it. For the mean:

\[ \sum_{n=1}^{N} C^{-1}(x_n - \mu) = 0 \quad \Rightarrow \quad \mu = \frac{1}{N} \sum_{n=1}^{N} x_n \]
PPCA has closed-form solutions

- Let \( \mathbf{S} \) be the sample covariance, as defined in PCA, and let its eigenvectors and eigenvalues be given.

- The optimal parameters for the maximal log-likelihood are

\[
\mu = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n
\]

\[
\sigma^2 = \frac{1}{D - M} \sum_{j=M+1}^{D} \lambda_j
\]

\[
\mathbf{W} = \mathbf{U}_M (\Lambda_M - \sigma^2 \mathbf{I})^{1/2}
\]
PPCA

- PPCA is the probabilistic **generative** version of PCA: we can also draw samples from it

- PPCA is a form of Gaussian distribution with number of parameters **restricted** (by the latent space)

- PPCA is the basis of **Bayesian PCA** (Bishop 12.2.3) in which the dimension of the latent space can be found from the data

- Like with the Gaussian Mixture Model, also the PPCA can be done via an **EM algorithm** (Bishop 12.2.2) - *Though not necessary because we have close form solutions*
PCA: Summary

- **Three views**
  - Max variance, min reconstruction error, probabilistic

- **Applications**
  - Dimensionality reduction
  - 2D/3D visualization
  - Compression
  - Whitening (de-correlating features)
  - (not mentioned) De-noising: discard the smallest variance features = the noise components (hopefully!)

- **Limitation**
  - Only linear transformations