Data Analysis and Manifold Learning
Lecture 6: Probabilistic PCA and Factor Analysis

Radu Horaud
INRIA Grenoble Rhone-Alpes, France
Radu.Horaud@inrialpes.fr
http://perception.inrialpes.fr/

Outline of Lecture 6

- A short reminder from Lecture 1
- Probabilistic formulation of PCA
- Maximum-likelihood PCA
- EM PCA
- What is Bayesian PCA?
- Factor Analysis

Material for This Lecture

- C. M. Bishop. Pattern Recognition and Machine Learning. 2006. (Chapter 12)
- More involved readings:
  - S. Roweis. EM algorithms of PCA and SPCA. NIPS 1998.
  - M. E. Tipping and C. M. Bishop. Pobabilistic Principal Component Analysis. J. R. Stat. Soc. B. 1999.
  - M. E. Tipping and C. M. Bishop. Mixtures of Probabilistic Principal Component Analysers. Neural Computation. 1999.

PCA at a Glance

- The input (observation) space: $\mathbf{X} = [\boldsymbol{x}_1 \dots \boldsymbol{x}_j \dots \boldsymbol{x}_n]$, $\boldsymbol{x}_j \in \mathbb{R}^D$
- The output (latent) space: $\mathbf{Y} = [\boldsymbol{y}_1 \dots \boldsymbol{y}_j \dots \boldsymbol{y}_n]$, $\boldsymbol{y}_j \in \mathbb{R}^d$
- **Projection:** $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$ with $\mathbf{W}^\top$ a $d \times D$ matrix.
- **Reconstruction:** $\mathbf{X} = \mathbf{W}\mathbf{Y}$ with $\mathbf{W}$ a $D \times d$ matrix.
- $\mathbf{W}^\top \mathbf{W} = \mathbf{I}_d$, i.e., $\mathbf{W}^\top$ is a row-orthonormal matrix when both data sets $\mathbf{X}$ and $\mathbf{Y}$ are represented in orthonormal bases: $\boldsymbol{y}_j = \widetilde{\mathbf{U}}^\top (\boldsymbol{x}_j - \overline{\boldsymbol{x}})$.
- $\mathbf{W}^\top \mathbf{W}^\top = \boldsymbol{\Lambda}_d^{-1}$, i.e., this corresponds to the case of *whitening*: $\boldsymbol{y}_j = \boldsymbol{\Lambda}_d^{-1/2} \widetilde{\mathbf{U}}^\top (\boldsymbol{x}_j - \overline{\boldsymbol{x}})$.
- Remember that $\mathbf{W}^\top$ was estimated from the d largest eigenvalue-eigenvector pairs of the data covariance matrix.

From Lecture #1: Data Projection on a Linear Subspace

- From $\mathbf{Y} = \mathbf{W}^\top \mathbf{X}$ we have

$$\mathbf{Y}\mathbf{Y}^\top = \mathbf{W}^\top \mathbf{X}\mathbf{X}^\top \mathbf{W} = \mathbf{W}^\top \widetilde{\mathbf{U}} \mathbf{\Lambda}_d \widetilde{\mathbf{U}}^\top \mathbf{W}$$

1. The projected data has a diagonal covariance matrix: $\mathbf{Y}\mathbf{Y}^\top = \mathbf{\Lambda}_d$, by identification we obtain

$$\mathbf{W}^\top = \widetilde{\mathbf{U}}^\top$$

2. The projected data has an identity covariance matrix, this is called *whitening the data*: $\mathbf{Y}\mathbf{Y}^\top = \mathbf{I}_d$

$$\mathbf{W}^\top = \mathbf{\Lambda}_d^{-\frac{1}{2}} \widetilde{\mathbf{U}}^\top$$

- In what follow, we will consider $\mathbf{W}$ (reconstruction) istead of $\mathbf{W}^\top$ (projection).

The Probabilistic Framework (I)

- Consider again the *reconstruction* of the observed variables from the latent variables. A point $\boldsymbol{x}$ is reconstructed from $\boldsymbol{y}$ with:

$$\boldsymbol{x} - \boldsymbol{\mu} = \mathbf{W}\boldsymbol{y} + \boldsymbol{\varepsilon}$$

- $\boldsymbol{\varepsilon} \in \mathbb{R}^D$ is the reconstruction error and let's suppose that it has a Gaussian distribution with zero mean and spherical covariance:

$$\boldsymbol{\varepsilon} = \mathcal{N}(\boldsymbol{\varepsilon}|0, \sigma^2\mathbf{I})$$

The Probabilistic Framework (II)

- We can now define the conditional distribution of the observed variable $\boldsymbol{x}$ conditioned on the value of the latent variable $\boldsymbol{y}$:

$$P(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\mathbf{W}\boldsymbol{y} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

- The prior distribution of the latent variable is a Gaussian with zero-mean and unit-covariance:

$$P(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}|0, \mathbf{I})$$

- The marginal distribution $P(\boldsymbol{x})$ can be obtained from the sum and product rules, supposing continuous latent variables:

$$P(\boldsymbol{x}) = \int_y P(\boldsymbol{x}|\boldsymbol{y})P(\boldsymbol{y})d\boldsymbol{y}$$

- This is a linear-Gaussian model, hence it is Gaussian as well:

$$P(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \mathbf{C})$$

The Probabilistic Framework (III)

- The mean and covariance of this *predictive distribution* can be formally derived from the expression of $\boldsymbol{x}$ and from the Gaussian distributions just defined:

$$\begin{aligned} E[\boldsymbol{x}] &= E[\mathbf{W}\boldsymbol{y} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}] = \mathbf{W}E[\boldsymbol{y}] + E[\boldsymbol{\mu}] + E[\boldsymbol{\varepsilon}] = \boldsymbol{\mu} \\ \mathbf{C} &= E[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top] = E[(\mathbf{W}\boldsymbol{y} + \boldsymbol{\varepsilon})(\mathbf{W}\boldsymbol{y} + \boldsymbol{\varepsilon})]^\top \\ &= \mathbf{W}E[\boldsymbol{y}\boldsymbol{y}^\top]\mathbf{W}^\top + E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top] = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I} \end{aligned}$$

- If assumed that $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ are independent. Gaussian distributions require the inverse of the covariance matrix:

$$\mathbf{C}^{-1} = \sigma^{-2}(\mathbf{I} - \mathbf{W}\mathbf{M}^{-1}\mathbf{W}^\top)$$

- Where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$ is a $d \times d$ matrix. *This is interesting when $d \ll D$.*

Maximum-likelihood PCA (I)

- The observed-data log-likelihood writes:

$$\ln P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n | \boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{j=1}^{n} \ln P(\boldsymbol{x}_j | \boldsymbol{\mu}, \mathbf{W}, \sigma^2)$$

- This expression can be developed using the previous equations, to obtain:

$$\ln P(\mathbf{X} | \boldsymbol{\mu}, \mathbf{C}) = -\frac{n}{2}(D \ln(2\pi) + \ln|\mathbf{C}|) - \frac{1}{2} \sum_{j=1}^{n} (\boldsymbol{x}_j - \boldsymbol{\mu})^\top \mathbf{C}^{-1} (\boldsymbol{x}_j - \boldsymbol{\mu})$$

Maximum-likelihood PCA (II)

- The log-likelihood is quadratic in $\boldsymbol{\mu}$, by setting the derivative with respect to $\boldsymbol{\mu}$ equal to zero, we obtain the expected result:

$$\boldsymbol{\mu}_{ML} = \sum_{j=1}^{n} \boldsymbol{x}_j = \overline{\boldsymbol{x}}$$

- Maximization with respect to $\mathbf{W}$ and σ^2, while is more complex, has a closed-form solution:

$$\begin{aligned} \mathbf{W}_{ML} &= \widetilde{\mathbf{U}}(\mathbf{\Lambda_d} - \sigma^2_{ML}\mathbf{I}_d)^{1/2}\mathbf{R} \\ \sigma^2_{ML} &= \frac{1}{D-d}\sum_{i=d+1}^{D} \lambda_i \end{aligned}$$

- With $\mathbf{\Sigma}_X = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, $d < D$, and $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$ (a $d \times d$ matrix).

Maximum-likelihood PCA (Discussion)

- The covariance of the predictive density, $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$, is not affected by the arbitrary orthogonal transformation $\mathbf{R}$ of the latent space:

$$\mathbf{C} = \widetilde{\mathbf{U}}\mathbf{\Lambda_d}\widetilde{\mathbf{U}}^\top - \sigma^2(\mathbf{I} - \widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^\top)$$

- The covariance projected onto a unit vector is $\boldsymbol{v}^\top\mathbf{C}\boldsymbol{v}$. We obtain the following cases:
  - $\boldsymbol{v}$ is orthogonal to $\widetilde{\mathbf{U}}$, then $\boldsymbol{v}^\top\mathbf{C}\boldsymbol{v} = \sigma^2\mathbf{I}$ or the average variance associated with the discarded dimensions.
  - $\boldsymbol{v}$ is one of the column vectors of $\widetilde{\mathbf{U}}$, then $\boldsymbol{u}_i^\top\mathbf{C}\boldsymbol{u}_i = \lambda_i$
- Matrix $\mathbf{R}$ introduces an arbitrary orthogonal transformation of the latent space.

From Probabilistic to Standard PCA

- The maximum-likelihood solution allows to estimate the *reconstruction* matrix $\mathbf{W}$ and the variance σ. The *projection* can be estimated from the pseudo-inverse of the reconstruction. We obtain:

$$(\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{W}^\top = (\mathbf{\Lambda}_d - \sigma^2 \mathbf{I}_d)^{-1/2}\widetilde{\mathbf{U}}^\top$$

- When $\sigma^2 = 0$ this corresponds to the standard PCA solution – rotating, projecting and whitening the data.

EM for PCA

- We can derive an EM algorithm for PCA, by following the EM framework: derive the complete-data log-likelihood conditioned by the observed data, and take its expectation:

$$\ln P(\mathbf{X}, \mathbf{Y} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2) = \sum_{j=1}^{n} (\ln P(\boldsymbol{x}_j | \boldsymbol{y}_j) + \ln P(\boldsymbol{y}_j))$$

- Then we take the expectation with respect to the posterior distribution of the latent variables, $E[\ln P(\mathbf{X}, \mathbf{Y} | \boldsymbol{\mu}, \mathbf{W}, \sigma^2)]$, which depends on the current model parameters $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$, $\mathbf{W}$, and σ^2, as well as on (these are the posterior statistics):

$$\begin{aligned} E[\boldsymbol{y}_j] &= \mathbf{M}^{-1}\mathbf{W}^\top(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) \\ E[\boldsymbol{y}_j \boldsymbol{y}_j^\top] &= \sigma^2 \mathbf{M}^{-1} + E[\boldsymbol{y}_j]E[\boldsymbol{y}_j]^\top \end{aligned}$$

The EM Algorithm

- *Initialize* the parameter values $\mathbf{W}$ and σ^2.
- *E-step:* Estimate the posterior statistics $E[\boldsymbol{y}_j]$ and $E[\boldsymbol{y}_j\boldsymbol{y}_j^\top]$ using the current parameter values.
- *M-step:* Update the parameter values from the current ones to new ones:

$$\mathbf{W}_{new} = \left(\sum_{j=1}^{n}(\boldsymbol{x}_j - \overline{\boldsymbol{x}})E[\boldsymbol{y}_j]^\top\right)\left(\sum_{j=1}^{n}E[\boldsymbol{y}_j\boldsymbol{y}_j^\top]\right)^{-1}$$

$$\sigma^2_{new} = \frac{1}{nD}\sum_{j=1}^{n}(\|\boldsymbol{x}_j - \overline{\boldsymbol{x}}\|^2 - 2E[\boldsymbol{y}_j]^\top\mathbf{W}_{new}^\top(\boldsymbol{x}_j - \overline{\boldsymbol{x}}) + \mathrm{tr}(E[\boldsymbol{y}_j\boldsymbol{y}_j^\top]\mathbf{W}_{new}^\top\mathbf{W}_{new}))$$

EM for PCA (Discussion)

- Computational efficiency for high-dimensional spaces. EM is iterative, but each iteration can be quite efficient. The covariance matrix is never estimated explicitly.
- The case of $\sigma^2 = 0$ corresponds to a valid EM algorithm: *S. Roweis. EM algorithms of PCA and SPCA. NIPS 1998.*
- The case of EM in the presence of missing data can be found in *M. E. Tipping and C. M. Bishop. Pobabilistic Principal Component Analysis. J. R. Stat. Soc. B. 1999*

Bayesian PCA (I)

- Select the dimension d of the latent space.
- The generative model just introduced (well defined likelihood function) allows to address the problem in a principled way.
- The idea is to consider each column in $\mathbf{W}$ as having an independent Gaussian prior:

$$P(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{d} \left(\frac{\alpha_i}{2\pi}\right)^{D/2} \exp\left(-\frac{1}{2}\alpha_i \boldsymbol{w}_i^\top \boldsymbol{w}\right)$$

- where $\alpha_i = 1/\sigma_i^2$ is called the precision parameter. The objective is to estimate these parameters, one for each principal direction, and select only a subset of these directions.
- We need to select directions of maximum variance, hence directions with *infinite precision* will be disregarded.

Bayesian PCA (II)

- The approach is based on *evidence approximation* or *empirical Bayes*.
- The marginal likelihood function (the latent space $\mathbf{W}$ is *integrated out*):

$$P(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma_2) = \int \underbrace{P(\mathbf{X}|\boldsymbol{\mu}, \mathbf{W}, \sigma^2)}_{\text{ML PCA}} P(\mathbf{W}|\boldsymbol{\alpha}) d\mathbf{W}$$

- The formal derivation is quite involved. The maximization with respect to the precision parameters yields a simple form:

$$\alpha_i^{new} = \frac{D}{\boldsymbol{w}_i^\top \boldsymbol{w}}$$

- This estimation is interleaved with the EM updates for estimating $\mathbf{W}$ and σ^2.

Factor Analysis

- Probabilistic PCA so far (the predictive covariance is isotropic):
$$P(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\mathbf{W}\boldsymbol{y} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$
- In factor analysis, the covariance is diagonal rather than isotropic:
$$P(\boldsymbol{x}|\boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}|\mathbf{W}\boldsymbol{y} + \boldsymbol{\mu}, \boldsymbol{\Psi})$$
- the columns of $\mathbf{W}$ are called *factor loadings* and the diagonal entries of $\boldsymbol{\Psi}$ are called *uniquenesses*.
- The factor analysis point of view: one form of latent-variable density model, the form of the latent space is of interest but not the particular choice of coordinates (up to an orthogonal transformation).
- The factor analysis parameters, $\mathbf{W}$, and $\boldsymbol{\Psi}$ are estimated via the maximum likelihood and EM frameworks.