Data Analysis and Manifold Learning
Lecture 5: Minimum-error Formulation of PCA and Fisher's Discriminant Analysis

Radu Horaud

INRIA Grenoble Rhone-Alpes, France

Radu.Horaud@inrialpes.fr

http://perception.inrialpes.fr/

Outline of Lecture 5

- Minimum-error formulation of PCA
- PCA for high-dimensional spaces
- Fischer's discriminant analysis for two classes and generalization to K classes

Material for This Lecture

- C. M. Bishop. Pattern Recognition and Machine Learning. 2006. (Chapters 4 and 12)
- `http://en.wikipedia.org/wiki/Linear_discriminant_analysis`
- Numerous textbooks treat PCA and LDA

Projecting the Data

- Let $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_n) \subset \mathbb{R}^D$,
- Consider an orthonormal basis vector, e.g., the columns of a $D \times D$ orthonormal matrix $\mathbf{U}$, namely $\boldsymbol{u}_i^\top \boldsymbol{u}_j = \delta_{ij}$.
- We can write:

$$\boldsymbol{x}_j = \sum_{i=1}^{D} \alpha_{ji} \boldsymbol{u}_i \text{ with } \alpha_{ji} = \boldsymbol{x}_j^\top \boldsymbol{u}_i$$

- Moreover, consider a lower-dimensional subspace of dimension $d < D$. We approximate each data point with:

$$\widetilde{\boldsymbol{x}}_j = \sum_{i=1}^{d} z_{ji} \boldsymbol{u}_i + \sum_{i=d+1}^{D} b_i \boldsymbol{u}_i$$

Minimizing the Distortion

- Choose the vectors $\{\boldsymbol{u}_j\}$ and the scalars $\{z_{ji}\}$ and $\{b_i\}$ that minimize the following distortion error:

$$J = \frac{1}{n} \sum_{j=1}^{n} \|\boldsymbol{x}_j - \widetilde{\boldsymbol{x}}_j\|^2$$

- By substitution of $\widetilde{\boldsymbol{x}}_j$ and by setting the derivatives to 0, $\partial J / \partial z_{ji} = 0$, $\partial J / \partial b_i = 0$ we obtain:

$$\begin{aligned} z_{ji} &= \boldsymbol{x}_j^\top \boldsymbol{u}_i, \;\; i = 1 \dots d \\ b_i &= \overline{\boldsymbol{x}}^\top \boldsymbol{u}_i, \;\; i = d+1 \dots D \end{aligned}$$

Closed-form Expression of the Distorsion

- By substitution we obtain the following distorsion between each data point and its projection onto the *principal subspace*:

$$\boldsymbol{x}_j - \widetilde{\boldsymbol{x}}_j = \sum_{i=d+1}^{D} \left((\boldsymbol{x}_j - \overline{\boldsymbol{x}})^\top \boldsymbol{u}_i \right) \boldsymbol{u}_i$$

  This *error-vector* lies in a space perpendicular to the principal space. The distorsion becomes:

$$J = \frac{1}{n} \sum_{j=1}^{n} \sum_{i=d+1}^{D} (\boldsymbol{x}_j^\top \boldsymbol{u}_i - \overline{\boldsymbol{x}}^\top \boldsymbol{u}_i)^2 = \sum_{i=d+1}^{D} \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i$$

Minimizing the Distortion (I)

- Note that in the previous equation,
  $1/n \sum_{j=1}^{n} (\boldsymbol{x}_j^\top \boldsymbol{u}_i - \overline{\boldsymbol{x}}^\top \boldsymbol{u}_i)^2$
  corresponds to the variance of the projected data onto $\boldsymbol{u}_i$. Minimizing the distortion is equivalent to minimizing the variances along the directions perpendicular to the principal directions $\boldsymbol{u}_1 \dots \boldsymbol{u}_d$. This can be done by minimizing $\tilde{J}$ with respect to $\boldsymbol{u}_{d+1} \dots \boldsymbol{u}_D$:

$$\tilde{J} = \sum_{i=d+1}^{D} \boldsymbol{u}_i^\top \boldsymbol{\Sigma} \boldsymbol{u}_i + \sum_{i=d+1}^{D} \lambda_i (1 - \boldsymbol{u}_i^\top \boldsymbol{u}_i)$$

Minimizing the Distortion (II)

- By setting the derivatives to 0 we obtain:

$$\frac{\partial \tilde{J}}{\partial \boldsymbol{u}_i} = 0 \leftrightarrow \boldsymbol{\Sigma} \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \; i = d+1 \dots D$$

- The distortion becomes:

$$\tilde{J} = \lambda_{d+1} + \ldots + \lambda_D$$

- The principal directions correspond to the largest d eigenvalue-eigenvector pairs of the covariance matrix $\boldsymbol{\Sigma}$

Choosing the Dimension of the Principal Subspace

- The covariance matrix can be written as $\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$. The trace of the diagonal matrix $\mathbf{\Lambda}$ can be interpreted as the *total variance*.
- One way to choose the principal subspace is to choose the largest d eigenvalue-eigenvector pairs such that:

$$\alpha(d) = \frac{\lambda_1 + \ldots + \lambda_d}{\lambda_1 + \ldots + \lambda_D} = \frac{\lambda_1 + \ldots + \lambda_d}{\mathrm{tr}(\mathbf{\Sigma})} \approx 0.95$$

High-dimensional Data

- When D is very large, the number of data points n may be smaller than the dimension. In this case it is better to use the $n \times n$ Gram matrix instead of the $D \times D$ covariance matrix.
- For centred data we have:

$$\begin{aligned} \mathbf{\Sigma} &= \frac{1}{n}\mathbf{X}\mathbf{X}^\top \text{ with } (\lambda_i, \boldsymbol{u}_i) \\ \mathbf{G} &= \mathbf{X}^\top\mathbf{X} \text{ with } (\mu_i, \boldsymbol{v}_i) \end{aligned}$$

- By premultiplication of $\mathbf{X}\mathbf{X}^\top\boldsymbol{u} = n\lambda\boldsymbol{u}$ with $\mathbf{X}^\top$ we obtain:

$$\boldsymbol{v} = \mathbf{X}^\top\boldsymbol{u} \text{ and } \mu = n\lambda$$

- From which we obtain: $\boldsymbol{u} = \frac{1}{\mu}\mathbf{X}\boldsymbol{v}$
- Assuming that the eigenvectors of the Gram matrix are normalized, we obtain:

$$\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} = \frac{1}{\sqrt{\mu}}\mathbf{X}\boldsymbol{v}$$

Discriminant Analysis

- Project the high-dimensional input vector to one dimension, i.e., along the direction of $\boldsymbol{w}$:

$$y = \boldsymbol{w}^\top \boldsymbol{x}$$

- This results in a loss of information and well-separated clusters in the initial space may overlap in one dimension.
- With a proper choice of of $\mathbf{w}$ one can select a projection that maximizes the class separation.

Two-Class Problem

- Let's assume that the data points belong to two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$ and that the mean vectors of these two clusters are $\overline{\boldsymbol{x}}_1 = 1/n_1 \sum_{j \in \mathcal{C}_1} \boldsymbol{x}_j$ and $\overline{\boldsymbol{x}}_2 = 1/n_2 \sum_{j \in \mathcal{C}_2} \boldsymbol{x}_j$
- One can choose $\boldsymbol{w}$ to maximize the distance between the projected means: $\overline{y}_1 - \overline{y}_2 = \boldsymbol{w}^\top (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$
- We can enforce the constraint $\boldsymbol{w}^\top \boldsymbol{w} = 1$ using a Lagrance multiplier and obtain the following solution:

$$\boldsymbol{w} = \overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2$$

- This solution is optimal when the two clusters are spherical.

Fisher's Linear Discriminant

- The solution consists in enforce small variances within each class. The criterion to be maximized becomes:

$$J(\boldsymbol{w}) = \frac{(\overline{x}_1 - \overline{x}_2)^2}{\sigma_1^2 + \sigma_2^2}$$

- Where the within cluster projected variance is:

$$\sigma_k^2 = \frac{1}{n_k} \sum_{j \in \mathcal{C}_k} (y_j - \overline{y}_k)^2$$

Maximizing Fisher's Criterion

- The criterion can be rewriten in the form:

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top \boldsymbol{\Sigma}_B \boldsymbol{w}}{\boldsymbol{w}^\top \boldsymbol{\Sigma}_W \boldsymbol{w}}$$

- Where $\boldsymbol{\Sigma}_B$ is the *between-cluster* covariance and $\boldsymbol{\Sigma}_W$ the total *within-cluster* covariance. They are given by:

$$\begin{aligned}
\boldsymbol{\Sigma}_B &= (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^\top \\
\boldsymbol{\Sigma}_W &= \frac{1}{n_1} \sum_{j \in \mathcal{C}_1} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_1)(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_1)^\top \\
&+ \frac{1}{n_2} \sum_{j \in \mathcal{C}_2} (\boldsymbol{x}_j - \overline{\boldsymbol{x}}_2)(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_2)^\top
\end{aligned}$$

- Optimal solution: $\boldsymbol{w} = \boldsymbol{\Sigma}_W^{-1}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)$

Discriminant Analysis

- The method just described can be applied to a training data set, where each data point belongs to one of the two clusters.
- Once the choice of an optimal direction of projection was performded, the projected data can be used to construct a discriminant.
- The projected training data belongs to a 1D Gaussian mixture with two clusters, $\mathcal{C}_1$ and $\mathcal{C}_2$. The parameters of eachone of these two clusters can be computed with $(k = 1, 2)$:

$$\pi_k = n_k/n, \;\; \overline{y}_k = \boldsymbol{w}^\top \overline{\boldsymbol{x}}_k, \;\; \text{and} \;\; \sigma_k^2 = \frac{1}{n_k} \sum_{j \in \mathcal{C}_k} (y_j - \overline{y}_k)^2$$

- Classification of a *new data point* y can be done using the class-posterior probabilities:

$$\mathcal{C} = \arg\max_k \pi_k \mathcal{N}(y|\overline{y}_k, \sigma_k^2)$$

Fisher's Discriminant for Multiple Classes

- The two-class discriminant analysis can be extended to $K > 2$ classes.
- The idea is to consider several linear projections, i.e., $y_k = \boldsymbol{w}_k^\top \boldsymbol{x}$ with $k = 1 \dots K-1$.
- A formal derivation can be found in: C. M. Bishop. Pattern Recognition and Machine Learning. 2006. (Chapter 4, pp 191-192).