Notes on PCA in Pattern Classification

Dan Beatty, Dr. Mitra

April 16, 2007

- Combine features in order to reduce the dimension of the feature space
- Linear combinations are simple to compute and tractable
- Project high dimensional onto a lower dimensional space
- Two classical approaches for finding "optimal" linear transformation
  - Principal Component Analysis "Projection that best **represents** the data in a least-square sense."
  - Multiple Discriminant Analysis "Projection that bests **separates** the data in a least squares sense"

1 Principle Component Analysis

Let us have a set of d dimensional vectors $\vec{x_1}, ..., \vec{x_n}$. We want to represent the set by a single vector $\vec{x_0}$ in such a way that the squared error criterion function:

$$J_0(\vec{x_0}) = \sum_{k=1}^{n} ||\vec{x_0} - \vec{x_k}||^2 \tag{1}$$

$$\vec{m} = \frac{1}{n} \sum_{k=1}^{n} \vec{x_k} \tag{2}$$

$\vec{x_k}$ is a zero dimensional representation of the data set.

For a one-dimensional representation of the data set let us look at a projection of the data onto a line passing through the sample mean.

$$\vec{x} = \vec{m} + a\vec{e} \tag{3}$$

where $\vec{e}$ is a unit vector in the direction of the line.

$$\vec{x_k} = m + a_k\vec{e} \tag{4}$$

then an optimal set of a_k can be found by minimizing

$$J_i(a_1, ..., a_n, e) = \sum_{k=1}^{n} ||(\vec{m} + a_k\vec{e}) - \vec{x_k}||^2 \tag{5}$$

$$J_i(a_1, ..., a_n, e) = \sum_{k=1}^{n} ||a_k\vec{e} - (\vec{x_k} - \vec{m})||^2 \tag{6}$$

$$= \sum_{k=1}^{n} a_k^2||\vec{e}||^2 - 2\sum_{k=1}^{n} a_k\vec{e}^T(\vec{x}_k - \vec{m}) + \sum_{k=1}^{n} ||\vec{x_k} - \vec{m}||^2 \tag{7}$$

To minimize J_1 we take $\frac{dJ_1}{da_k} = 0$ and we obtain:

$$a_k = \vec{e}^T(\vec{x_k} - \vec{m}) \tag{8}$$

which is the least squares solution by projecting $\vec{x_k}$ into a line passing through $\vec{m}$ in the direction of $\vec{e}$.

A scatter matrix S is defined by

$$\mathbf{S} = \sum_{k=1}^{n}(\vec{x_k} - \vec{m})(\vec{x_k} - \vec{m})^T \tag{9}$$

which happens to be the sample covariance $n - 1$ times.

We use it in

$$J_1(\vec{e}) = \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} ||\vec{x_k} - \vec{m}||^2 \tag{10}$$

$$= -\sum_{k=1}^{n} |\vec{e}^T(\vec{x_k} - \vec{m})|^2 + \sum_{k=1}^{n} ||\vec{x_k} - \vec{m}||^2 \tag{11}$$

$$= -\sum_{k=1}^{n} \vec{e}^T(\vec{x_k} - \vec{m})(\vec{x_T} - \vec{m})^T\vec{e} + \sum_{k=1}^{n} ||\vec{x_k} - \vec{m}||^2 \tag{12}$$

$$= -\vec{e}^T\mathbf{S}\vec{e} + \sum_{k=1}^{n} ||\vec{x_k} - \vec{m}||^2 \tag{13}$$

In order to satisfy the minimal case of J_1 using $\vec{e}$, we need to maximize the term $\vec{e}^T\mathbf{S}\vec{e}$.

Let us use the Lagrange multiplier λ subject to the contruct $||e|| = 1$,

$$\vec{u} = \vec{e}^T\mathbf{S}\vec{e} - \lambda(\vec{e}^T\vec{e} - 1) \tag{14}$$

$$\frac{\partial\vec{u}}{\partial\vec{e}} = 2\mathbf{S}\vec{e} - 2\lambda\vec{e} \tag{15}$$

$$\Rightarrow \mathbf{S}\vec{e} = \lambda\vec{e} \tag{16}$$

Applying to d' - dimensional projection such that $d' \leq d$

$$\vec{x} = \vec{m} + \sum_{i=1}^{d'} a_i \vec{e_i} \tag{17}$$

$$J_{d'} = \sum_{k=1}^{n} ||(\sum_{i=1}^{d'} a_i \vec{e_i}) - \vec{x_k}||^2 \tag{18}$$

needs to be minimized when the vectors $e_1, ..., e_{d'}$ are the d' eigenvectors of the scatter matrix $\mathbf{S}$ with the largest eigenvalues. a_i are the principle components of $\vec{x}$ in that basis.

2 Fisher's Linear Discriminant

Discriminant analysis, we need to find projected directions of the data that can discriminate the embedded patterns.

We have a set of n d-dimensional samples $(\vec{x_1}, ..., \vec{x_n})$ having two subsets D_1 and D_2, with n_1 and n_2 samples respectively.

$$y = \vec{w}^T \vec{x} \tag{19}$$

such that y is a linear combination of the components of $\vec{x}$.

We define corresponding subsets by Y_1 and Y_2. If $||\vec{w}|| = 1$ then each y_i is a projection of x_i onto a line in the direction of $\vec{w}$.

$$\vec{m_i} = \frac{1}{n_i} \sum_{\vec{x} \in D_i} \vec{x} \tag{20}$$

$$\tilde{m_i} = \frac{1}{n_i} \sum_{y \in Y_i} y \tag{21}$$

$$= \frac{1}{n_i} \frac{\vec{x} \in D_i}{\vec{w}^T \vec{x}} \tag{22}$$

$$= \vec{w}^T \vec{m_i} \Rightarrow |\tilde{m_1} - \tilde{m_2}| = |\vec{w}^T (\vec{m_1} - \vec{m_2})| \tag{23}$$

Equation **??** is the projected mean, which is a projection on $\vec{m_i}$ [**?**, 118].

$$\tilde{s_i}^2 = \sum_{y \in Y_i} (y - \tilde{m_i})^2 \tag{24}$$

$$\frac{1}{n}(\tilde{s_1}^2 + \tilde{s_2}^2) \tag{25}$$

$$\tilde{s_1}^2 + \tilde{s_2}^2 \tag{26}$$

$$J(\vec{w}) = \frac{|\tilde{m_1} - \tilde{m_2}|^2}{\tilde{s_1}^2 + \tilde{s_2}^2} \tag{27}$$

- Equation **??** is the scatter for projected samples.
- Equation **??** is an estimate of the variance of the pooled data and
- equation **??** is the total within-class scatter.

The Fisher Linear discriminant uses the criterion function (equation **??**).

$$\mathbf{S_i} = \sum_{\vec{x} \in D_i} (\vec{x} - \vec{m_i})(\vec{x} - \vec{m_i})^T \tag{28}$$

$$\mathbf{S_W} = \mathbf{S_1} + \mathbf{S_2} \tag{29}$$

$$\tilde{s_i}^2 = \sum_{\vec{x} \in D_i} (\vec{w}^T \vec{x} - \vec{w}^T \vec{m_i})^2 \tag{30}$$

$$= \sum_{\vec{x} \in D_i} \vec{w}^T (\vec{x} - \vec{m_i})(\vec{x} - \vec{m_i})^T \vec{w} \tag{31}$$

$$= \vec{w}^T \mathbf{S_i} \vec{w} \tag{32}$$

$$\therefore \tilde{s_1}^2 + \tilde{s_2}^2 = \vec{w}^T \mathbf{S_W} \vec{w} \tag{33}$$

Separation of projected means has its own scatter matrix for which it obeys:

$$(\tilde{m_1} - \tilde{m_2})^2 = (\vec{w}^T \vec{m_1} - \vec{w}^T \vec{m_2})^2 \tag{34}$$

$$= \vec{w}^T (\vec{m_1} - \vec{m_2})^2 \vec{w} \tag{35}$$

$$= \vec{w}^T (\vec{m_1} - \vec{m_2})(\vec{m_1} - \vec{m_2})^T \vec{w} \tag{36}$$

$$= \vec{w}^T \mathbf{S_B} \vec{w} \tag{37}$$

$$\because \mathbf{S_B} = (\vec{m_1} - \vec{m_2})(\vec{m_1} - \vec{m_2})^T \tag{38}$$

In terms of $\mathbf{S_B}$ and $\mathbf{S_W}$, the criterion function $J(\cdot)$ can be written as:

$$J(\vec{w}) = \frac{\vec{w}^T \mathbf{S_B} \vec{w}}{\vec{w}^T \mathbf{S_W} \vec{w}} \tag{39}$$

[**?**, 120]

Equation **??** is well known as the Rayleigh quotient. A $\vec{w}$ that minimizing of $J(\vec{w})$ must satisfiy equation **??** such that λ is a generalized eigenvalue.

$$\mathbf{S_B} \vec{w} = \lambda \mathbf{S_W} \vec{w} \tag{40}$$

If $\mathbf{S_W}$ is non-singular, then equation **??** is Fisher's Linear Discriminant.

$$\vec{w} = \mathbf{S_W}^{-1} (\vec{m_1} - \vec{m_2}) \tag{41}$$

Equation **??** is a mapping from d dimensional to one dimensional classification problem.

To find the threshold of the point along the mapped one-dimensional subspace separated the projected points, let us assume that the conditional densities $p(x|\omega_i)$ are multivariate normal with equal covariance matrices Σ then the optimal decision boundary to given by

$$\vec{w}^T\vec{x} + w_0 = 0 \tag{42}$$

where

$$\vec{w} = \mathbf{\Sigma}^{-1}(\vec{\mu_1} - \vec{\mu_2}) \tag{43}$$

By estimating $\mu_i + \Sigma$ from the sample means and covariances, we can get the direction of w that maximizes $J(\cdot)$. The computational complexity of this approach is mainly due to computing the within-class total scatter and its inverse + involves $O(\alpha^2 n)$ operations.

3 MDA

For c- classes problem, we consider the projection for a d-dimensional space to $(c-1)$ dimensional space assuming $d \geq c$

$$\therefore \mathbf{S_w} = \sum_{i=1}^{c} \mathbf{S_i} \tag{44}$$

$$\mathbf{S_i} = \sum_{\vec{x}\in D_i} (\vec{x} - \vec{m}_i)(\vec{x} - \vec{m}_i)^T \tag{45}$$

$$\vec{m}_i = \frac{1}{n_i}\sum_{x\in D_i} \vec{x} \tag{46}$$

The generalization $\mathbf{S_B}$ is not as direct. Define a total mean vector $\vec{m}$ and a total scatter matrix S_T by

$$\vec{m} = \frac{1}{n}\sum_{\vec{x}} \vec{x} \tag{47}$$

$$= \frac{1}{n}\sum_{i=1}^{c} n_i\vec{m}_i \tag{48}$$

$$\mathbf{S_T} = \sum_{x}(x-m)(x-m)^T \tag{49}$$

$$\because \mathbf{S_B} = \sum_{i=1}^{c} n_i(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T \tag{50}$$

$$\therefore \mathbf{S_T} = \mathbf{S_w} + \sum_{i=1}^{c} n_i(\vec{m}_i - \vec{m})(\vec{m}_i - \vec{m})^T \tag{51}$$

$$= \mathbf{S_w} + \mathbf{S_B} \tag{52}$$

The $(c-1)$ discriminant function are given by

$$y_i = \vec{w_i}^T \vec{x}, i = 1, ..., c-1 \tag{53}$$
$$\Rightarrow \vec{y} = \mathbf{W}^T \vec{x}, \tag{54}$$

where y is vector with y_i components and w is a matrix $[dx(c-1)]$ with w_i are the column.

Now

$$\tilde{m_i} = \frac{1}{n_i} \sum_{y \in Y_i} y \tag{55}$$
$$\tilde{m} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{m_i} \tag{56}$$
$$\tilde{\mathbf{S_w}} = \sum_{i=1}^{c} \sum_{y \in Y_i} (y - \tilde{m_i})(y - \tilde{m_i})^T \tag{57}$$
$$\tilde{\mathbf{S_B}} = \sum_{i=1}^{c} n_i (\tilde{m_i} - \tilde{m})(\tilde{m_i} - \tilde{m})^T \tag{58}$$
$$\therefore \tilde{\mathbf{S_w}} = \mathbf{W}^T \mathbf{S_W} \mathbf{W} \tag{59}$$
$$\tilde{\mathbf{S_B}} = \mathbf{W}^T \mathbf{S_B} \mathbf{W} \tag{60}$$
$$J(\mathbf{W}) = \frac{|\tilde{S_B}|}{|\tilde{S_w}|} = \frac{|w^T S_B w|}{|w^T S_B w|} \tag{61}$$

Now $\mathbf{S_B w_i} = \lambda_i \mathbf{S_W w_i}$, since the columns of an optimal $\mathbf{W}$ are the generalized eigenvectors corresponding to the largest eigenvalues. Now we can find the eigenvalues as the roots of the characteristic polynomial

$$|\mathbf{S_B} - \lambda_i \mathbf{S_w}| = 0 \tag{62}$$

and solve

$$(\mathbf{S_B} - \lambda_i \mathbf{S_W}) \vec{w_i} = 0 \tag{63}$$

for the eigenvectors $\vec{w_i}$.

3.1 Example problem 3-40 [?, 152]

Problem statement as read from [**?**, 152].

> If S_B and S_w are two real, symmetric, d by d matrices, it is well known that there exists a set of n eigenvalues $\lambda_1, ..., \lambda_n$ satisfying $|\mathbf{S_B} - \lambda \mathbf{S_w}| = 0$, with a corresponding set of n eigenvectors, $\vec{e_1}, ..., \vec{e_n}$ satisfying $\mathbf{S_B} \vec{e_i} = \lambda_i \mathbf{S_w} \vec{e_i}$. Furthermore, if $\mathbf{S_w}$ is positive definite, the eigenvectors can always be normalized so that $\vec{e_i}^T \mathbf{S_w} \vec{e_i} = \delta_{ij}$ and $\vec{e_i}^T \mathbf{S_B} \vec{e_i} = \delta_{ij}$. Let $\tilde{\mathbf{S_w}} = \mathbf{W}^T \mathbf{S_w} \mathbf{W}$ and

Algorithm 1 Multiple Discriminant Analysis

Determine $\vec{m_t}$
for all Classes D_i in Discriminant Set D **do**
  Compute $\vec{m_i}$
  Determine n_i
  Determine $\hat{m_i} = \vec{m_i} - \vec{m_t}$
  Compute $S_i = \sum_{\vec{x_i} \in D_i} (\vec{x_i} - \vec{m_i})(\vec{x_i} - \vec{m_i})^T$
end for
$S_w = \sum_{S_i \in D} S_i$
Compute $S_B = \sum_{\hat{m_i} \in D} n_i \hat{m_i}$
Compute Top eigenvectors for equation:

$$\mathbf{S_B w_i} = \lambda_i \mathbf{S_W w_i}$$

return $\mathbf{W}, \mathbf{\Lambda}$

$\tilde{\mathbf{S_B}} = \mathbf{W}^T \mathbf{S_B W}$, where $\mathbf{W}$ is a d-by-n matrix whose columns correspond to n distinct eigenvectors.

1. Show that $\tilde{\mathbf{S_w}}$ is the n-by-n identify matrix $\mathbf{I}$ and that $\tilde{\mathbf{S_B}}$ is a diagonal matrix whose elements are the corresponding eigenvalues. (This show that the discriminant functions in multiple discriminant analysis analysis are uncorrelated.)
2. What is the value of $J = \frac{|\tilde{\mathbf{S_B}}|}{|\tilde{\mathbf{S_W}}|}$
3. Let $\vec{y} = \mathbf{W}^T \vec{x}$ be transformed by scaling the axes with a nonsingular n-by-n diagonal matrix $\mathbf{D}$ and by rotating this result with an orthogonal matrix $\mathbf{Q}$ where $\vec{y'} = \mathbf{QD}\vec{y}$. Show that J is invariant to this transformation.

S_B and $S_w \rightarrow$ two real, symmetric, $d \times d$ matrices. Therefore $|S_B - \lambda S_W| = 0$ for a set of n λ's and the corresponding n eigenvectors $e_1, ..., e_n$, satisfying

$$S_B e_1 = \lambda_i S_w e_i \tag{64}$$

If S_w is positive definite the eigenvectors can be normalized so that $e_i^T S_w e_i = \delta_{ij}$ and $e_i^T S_B e_j = \lambda_i \delta_{ij}$.

Let $\tilde{\mathbf{S_w}} = \mathbf{W}^T \mathbf{S_w W}$, and $\tilde{\mathbf{S_B}} \mathbf{W}^T \mathbf{S_B W}$ where $\mathbf{W}$ is a $d \times n$ matrix whose columns correspond to n distinct eigenvectors.

1. Show that $\tilde{\mathbf{S_w}} = \mathbf{I}$ (of size $n \times n$) and $\tilde{\mathbf{S_B}} \rightarrow$ a diagonal matrix with eigenvalues as diagonal elements. The discriminant functions in MDA analysis are uncorrelated.
2. What is the value of $J \frac{\tilde{\mathbf{S_B}}}{\tilde{\mathbf{S_W}}}$?

3. Let $\vec{y} = \vec{w}^T \vec{x}$ be transformed by scaling the axes with a non-singular $n \times n$ diagonal matrix D and by rotating the result with an orthogonal matrix $\mathbf{Q}$, where $\vec{y'} = \mathbf{QD}\vec{y}$. Show that J is invariant to this transformation.

Answer **??**, let the set {} are normalized eigenvectors, then $\vec{e_i}^T \mathbf{S_B} \vec{w_i} = \lambda_i \delta_{ij}$ $\vec{e_i}^T \mathbf{S_w} \vec{e_j} = \lambda_i \delta_{ij}$ and the matrix

$$\mathbf{W} = [\vec{e_1}, ..., \vec{e_n}] \tag{65}$$

Then the within scatter matrix ins: the now representation is

$$\tilde{S_w} = \mathbf{W}^T \mathbf{S_W} \mathbf{W} = \begin{pmatrix} \vec{e_1}^T \\ \vdots \\ \vec{e_n}^T \end{pmatrix} S_w(\vec{e_1}, ..., \vec{e_n}) \tag{66}$$

$$= \begin{pmatrix} \vec{e_1}^T \mathbf{S_W} \vec{e_1} & ... & \vec{e_1}^T \mathbf{S_W} \vec{e_n} \\ \vdots & & \\ \vec{e_n}^T \mathbf{S_W} \vec{e_1} & ... & \vec{e_n}^T \mathbf{S_W} \vec{e_n} \end{pmatrix} = I \tag{67}$$

Similar the between scatter matrix S_B is estimated as

$$\tilde{S_B} = \mathbf{W}^T \mathbf{S_B} \mathbf{W} = \begin{pmatrix} \vec{e_1}^T \mathbf{S_B} \vec{e_1} & ... & \vec{e_1}^T \mathbf{S_B} \vec{e_n} \\ \vdots & & \\ \vec{e_n}^T \mathbf{S_B} \vec{e_1} & ... & \vec{e_n}^T \mathbf{S_B} \vec{e_n} \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ \vdots & \lambda_2 & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_n \end{pmatrix} \tag{68}$$

$$\therefore \tilde{\mathbf{S_w}} = I \ (n \times n) \tag{69}$$

$$\tilde{\mathbf{S_B}} \text{ is diagonal containing } \lambda_i \tag{70}$$

Answer **??**,

$$|\tilde{\mathbf{S_B}}| = \lambda_1 \lambda_2 ... \lambda_n \tag{71}$$

$$|\tilde{\mathbf{S_W}}| = 1 \tag{72}$$

$$\therefore J = \lambda_1 \lambda_2 ... \lambda_n \tag{73}$$

Answer **??**, Let

$$\tilde{\mathbf{W}} = \mathbf{QDW}^T \tag{74}$$

$$\tilde{\mathbf{S_W}} = \tilde{\mathbf{W}}^T \mathbf{S_W} \tilde{\mathbf{W}} \tag{75}$$

$$= \mathbf{QDW}^T \mathbf{S_W} \mathbf{WDQ}^T \tag{76}$$

Then:

$$|\tilde{S_W}| = |D|^2 \tag{77}$$

$$\tilde{\mathbf{S_B}} = \tilde{\mathbf{W}}^T \mathbf{S_B} \tilde{\mathbf{W}} \tag{78}$$

$$= \mathbf{QDW}^T\mathbf{S_B}\mathbf{WDQ}^T \tag{79}$$

$$\therefore |\tilde{\mathbf{S_B}}| = |D|^2\lambda_1\lambda_2...\lambda_n \tag{80}$$

$$J = \frac{|\tilde{\mathbf{S_B}}|}{|\tilde{\mathbf{S_W}}|} \tag{81}$$

Therefore J is invariant to this transformation.

4 Computer Problems

Section 3.9 - 3.10 Expectation Maximization / Hidden Markov Models
Chapter 4 - Non parametric approach —
Parsen's Windows —– 164
Chapter 5 Linear Discriminant Functions —-
Chapter 6
Chapter 7 Stochastic methods
Chapter 8 — Cart algorithm
Chapter 10 —