Some Thoughts on Kernel PCA and LDA

Ramesh A. Gopinath
IBM T.J. Watson Research Center
email: rameshg@us.ibm.com

April, 2000

Abstract
This note describes a viewpoint for Kernel PCA and LDA.

1 Preliminaries

Given labeled data \( \{(x_i, y_i)\}, \ i \in \{1, \ldots, l\}, x_i \in \mathbb{R}^n, y_i \in \{0,1,\ldots,J-1\} \), let \( X = (x_1, x_2, \ldots, x_l) \) be the \( n \times l \) data matrix. Let \( e_j \in \mathbb{R}^l \) be the indicator vector for class \( j \) and let \( e = \sum_j e_j \). Each class has \( l_j = e_j' e_j \) samples and \( P_j = \frac{1}{l_j} e_j e_j' \) is the orthogonal projection along the mean of class \( j \). The overall mean and class means are

\[ \mu = \frac{1}{l} X e; \quad \mu_j = \frac{1}{l_j} X e_j. \]  

(1)

Let \( T, W \) and \( B \) be respectively the total, within-class and between-class covariances respectively:

\[ lT = [\sum_i x_i x_i'] - l \mu \mu' = XX' - \frac{1}{l} X e e' X' = X(I - P)X' = (XQ_T)(XQ_T)'; \]  

(2)

\[ lW = [\sum_i x_i x_i'] - \sum_j l_j \mu_j \mu_j' = X(I - \sum_j P_j)X' = XQ_W X' = (XQ_W)(XQ_W)'; \]  

(3)

\[ lB = I(l - W) = X(\sum_j P_j - P)X' = XQ_B X' = (XQ_B)(XQ_B)', \]  

(4)

\( Q_T, Q_W \) and \( Q_B \) are appropriate orthogonal projections - recall for any orthogonal projection \( R \), that \( R^2 = R = R' \). One can also interpret \( lQ_T, lQ_W \) and \( lQ_B \) as the total, within and between-class covariances of the \( l \times l \) “identity” data matrix \( I \); indeed \( Q_T = Q_W + Q_B \). Moreover, Eqn. 2-Eqn. 4 then can be interpreted as the same matrices after the data \( I \) has been linearly transformed by \( X \), i.e., \( I \mapsto XI = X \).

Since \( I = (I - \sum_j P_j) + (\sum_j P_j - P) + P = Q_W + Q_B + P \) with \( Q_W, Q_B \) and \( P \) mutually orthogonal projections, \( \mathcal{R}^l = Q_W \mathcal{R}^l \oplus Q_B \mathcal{R}^l \oplus P \mathcal{R}^l \) is a direct decomposition of \( \mathcal{R}^l \) from which the eigenspaces of \( Q_B, Q_W \) and \( Q_T \) are trivially inferred. For example, \( Q_T \) has one eigenvector \( \frac{1}{\|e\|} e \) with eigenvalue 0, and any vector on the orthogonal complement is an eigenvector with eigenvalue 1; recall projections have eigenvalues in \( \{0,1\} \).

Consider the following generalized eigenvalue problem: \( Q_T u = \lambda Q_W u \). By direct verification it is clear that the generalized eigenspaces are \( Q_W \mathcal{R}^l, Q_B \mathcal{R}^l \) and \( P \mathcal{R}^l \).
1. Any vector \( u \in Q_W R^l \) is a generalized eigenvector with eigenvalue \( 1 \) - clear since \( Q_T u = u, Q_W u = u \).

2. Any vector \( u \in Q_B R^l \) is a generalized eigenvector with \( \lambda = \infty \) - clear since \( Q_W u = 0 \times u \) and \( Q_T u = 1 \times u \).

3. \( u = \frac{1}{\| u \|} e \) is a generalized eigenvector and \( \lambda \) is indeterminate \( (0/0) \) - clear since \( Q_T u = 0 \times u \) and \( Q_W u = 0 \times u \).

If one is interested in eigenspace corresponding to large values of \( \lambda \) then one has to focus on \( Q_B R^l \).

For an arbitrary matrix \( A \), let \( A = U D V' \) be the SVD of \( A - U, V \) orthogonal and \( D \) diagonal. Columns of \( V \) are the eigenvectors of \( A^T A \) (with eigenvalues \( diag(DD') \)) and columns of \( U \) are the eigenvectors of \( A A^T \) (with eigenvalues \( diag(DD') \)). Furthermore, \( AV = UD \) (i.e., columns of \( U \) - with non-zero singular values - are in the column space of \( A \) with coefficients the corresponding column of \( V \)) and \( A' U = V D' \) (i.e., columns of \( V \) - with non-zero singular values - are in the rowspace of \( U \) with coefficients the corresponding column of \( U \)). Assuming \( rank(A) = r \), SVD can be interpreted as an expansion of \( A \)

\[
A = \sum_{i=1}^{r} d_i u_i v_i'.
\]  

(5)

2 PCA

PCA directions of the data correspond to eigenvectors of the total covariance. Let \( XQ_T = UDV' \) be the SVD of \( XQ_T \). Then from Eqn. 2 \( IT = UDD'U' \) and therefore the PCA directions are the columns \( U \) with eigenvalues \( diag(DD') \). The columns of \( V \) are the eigenvectors of the Gram-matrix of \( XQ_T \) i.e., \( Q_T X' XQ_T \). Recall that \( XQ_T \) is an \( n \times l \) matrix. If \( n \leq l \) - which is typically the case - one obtains \( U \) directly. If \( n > l \) it may be more convenient to solve for \( V \) and obtain scaled columns of \( U \) as an expansion in the basis \( XQ_T \) with coefficients \( V \) - i.e., \( XQ_T V = UD \). In any case, state-of-the-art numerical algorithms for SVD can be used to compute \( U \) and/or \( V \). It is typically more stable to compute SVD of \( XQ_T \) instead of computing the eigenvectors of \( XQ_T X' \) because \( c(XQ_T X') = c^2(XQ_T) \), where \( c(A) \) denotes the condition number of matrix \( A \). Note that \( rank(XQ_T) \leq rank(X) = r \leq \min(n,l) \), where \( r \) is the number of linearly independent data vectors. In typical applications \( l >> n \) and \( r = n \) and all the eigen-values are non-zero. However, if \( l = n \) and \( r = n \), then because \( Q_T \) is a projection, \( rank(XQ_T X') = n - 1 \) and there is an eigenvector with eigenvalue \( 0 \).

3 Kernel PCA

Let \( k(x, y) \) be a reproducing kernel over \( \mathcal{R}^n \times \mathcal{R}^n \). That is

\[
\int k(x, z)k(y, z)dz = k(x, y).
\]  

(6)

Consider the mapping \( x_i \mapsto k(x_i, x) \) taking \( \mathcal{R}^n \) to \( \mathcal{R}^n \). Let \( X_k \) denote transformed (or feature-space) data matrix - \( X_k = (k(x_1, x), \ldots, k(x_l, x)) \). In general \( X_k \) will not be centered, even if \( X \) is. Eqn. 6 shows how the standard inner product \( (L^2) \) in feature space
between data-vectors is determined exclusively by the kernel. Formally the total covariance in feature-space is \( X_k Q_T X'_k \) and therefore the PCA directions in feature-space are the eigenvectors of \( X_k Q_T X'_k \).

\[
\ell T(x, y) = \sum_i k(x_i, x)k(x_i, y) - l\mu(x)\mu(y) = X_k Q_T X'_k = (X_k Q_T)(X_k Q_T)'.
\]

(7)

Note that \( T(x, y) \) is a finite-rank operator (\( rank(T) \leq l \)) and therefore has an expansion in terms of normalized eigenfunctions:

\[
T(x, y) = \sum_{i=1}^l d_i^2 u_i(x)u_i(y),
\]

Indeed if \( rank(X_k) = r \) (i.e., there are \( r \) linearly independent data vectors in feature space) then one can consider the following “SVD” of \( X_k Q_T \):

\[
X_k Q_T = UDV' = \sum_{i=1}^{r \leq l} d_i u_i(x)v_i',
\]

where \( u_i \in \mathcal{R}^\mathcal{R}^n, v_i \in \mathcal{R}^n, d_i \in \mathcal{R}^+, .\) As we have seen \( U \) is the set of eigenvectors of \( X_k Q_T X'_k \) and \( V \) is the matrix of eigenvectors of the Gram matrix of \( X_k Q_T \), i.e., \( Q_T X'_k X_k Q_T \).

\[
Q_T X'_k X_k Q_T = Q_T \left( \int k(x_i, x)k(x_j, x)dx \right) Q_T = Q_T \left( k(x_i, x_j) \right) Q_T = Q_T K Q_T; Q_T K Q_T v_i = d_i^2 v_i.
\]

For Kernel PCA since the Gram matrix is finite dimensional it is more convenient to compute \( V \) (eigenvectors of \( Q_T K Q_T \)) and then compute \( U \). The PCA projection of data sample \( y \) (\( k(y, x) \) in feature-space) along \( u_i(x) \) can be computed as follows:

\[
\int k(y, x)u_i(x)dx = \frac{1}{d_i} \int k(y, x)X_k(x)Q_T v_i dx = \frac{1}{d_i} \left[ \int k(y, x)X_k(x)dx \right] Q_T v_i = \frac{1}{d_i} X_k(y)Q_T v_i.
\]

Sometimes one is interested in the eigenvectors of \( W \) or \( B \). Equivalently, in feature-space the eigenvectors of \( X_k Q_W X'_k \) or \( X_k Q_B X'_k \). The same analysis as above can be carried out with \( Q_T \) replaced by \( Q_W \) or \( Q_B \).

4 LDA

The basic idea in LDA is to find a \( d \)-dimensional projection of the data corresponding to the \( d \) generalized eigenvectors of \( (T, W) \) associated with the \( d \) largest generalized eigenvalues. From Eqn. 2-Eqn. 3 one has to solve the generalized eigenvalue problem for \( (X Q_T X', X Q_W X') \).

\[
X Q_T X' u = l T u = l W u \iff \lambda X Q_W X' u.
\]

(8)

Notice wlog that \( u \) is in the span of the data (if it is in the orthogonal complement then it is in the null-space of \( X' \) i.e., in the null-space of both \( T \) and \( W \), in which case \( \lambda \) is indeterminate (0/0) and not interesting). In other words \( u = Xv \) for some vector \( v \in \mathcal{R}^i \). Multiplying both sides by \( X' \) we get

\[
X Q_T X' X v = \lambda X Q_W X' X v; \quad X' X Q_T X' X v = \lambda X' X Q_W X' X v.
\]

(9)
One can therefore either solve for $u$ in Eqn. 8 or for $v$ in Eqn. 9 and choose those $v$'s that are not in the null-space of $X$. An interesting situation happens (that hardly ever occurs in practice for standard LDA since $n << l$) when $\text{rank}(X) = l$. In this case $X'X$ is invertible (i.e., the data samples are linearly independent - which typically happens when $l \leq n$) and therefore Eqn. 9 becomes

$$Q_T s = \lambda Q_W s,$$

where $s = X'Xv$ and $u = X(X'X)^{-1}s$. In this case,

1. Any vector $u \in X(X'X)^{-1}Q_W R_l$ is a generalized eigenvector with eigenvalue 1.
2. Any vector $u \in X(X'X)^{-1}Q_B R_l$ is a generalized eigenvector with $\lambda = \infty$
3. $u = X(X'X)^{-1}\frac{1}{\|e\|}e$ is a generalized eigenvector and $\lambda$ is indeterminate ($0/0$).

If $d \leq J - 1$, then the $d$ LDA directions can be chosen to be $X(X'X)^{-1}(\frac{1}{l}e_j - \frac{1}{l}e)$. Clearly there is an arbitrariness in this choice. The key result is that any ON basis for the span of $\mu_j - \mu$ will suffice from the point of view of LDA. In practice we can do better. Our proposal is to choose the top $d$ PCA directions of the between class covariance matrix i.e., $XQ_BX'$. The basic interpretation of this result is that in this situation that $W$ (and in particular all the class covariances) is singular. As such trying to compute the LDA directions by “sphering” the data (i.e., inverting $W$) leads to unbounded generalized eigenvalues. The solution is to ignore $W$ and focus on the PCA directions of $B$.

5 Kernel LDA

As before consider the mapping $x_i \mapsto k(x_i, x)$ taking $\mathcal{R}^n$ to $\mathcal{R}^{n'}$ and let $X_k$ denote the data-matrix. In this case LDA direction $X_k u$ can be obtained by solving for the eigenvalue problem:

$$X_k Q_T X_k u = \lambda X_k Q_W X_k u,$$

or

$$X_k' X_k Q_T X_k' X_k v = \lambda X_k' X_k Q_W X_k' X_k v.$$

In the latter case the solution is $u = X_k v$. Unlike in standard LDA, in kernel LDA the situation Gram matrix $X_k'X_k$ is typically invertible (because the mapping is into an infinite dimensional space e.g., RBF kernels). Hence kernel LDA for different choices of the kernel the LDA directions are given by $X_k(X_k'X_k)^{-1}s$, where $s$ solves $Q_T s = \lambda Q_W s$. This situation is described in earlier section. Any direction in the span of $XQ_B$ will suffice as an LDA direction - and all such directions are equally preferred. In this situation our proposed solution is to choose the top $d$ PCA directions of $XQ_BX'$ which can be computed using the techniques in Section 3.