Linear Discriminant Analysis

Lecturer:
Jiří Matas

Authors:
Jiří Matas, Jana Kostlivá

Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

8.6.2014
LDA alias Fisher Linear Discriminant (FLD)

**Goal:**

- performs dimensionality reduction by mapping high-dimensional data to a low-dimensional space
- finds optimal subspace such that the separability of classes is maximized
- achieved by minimizing the within-class distance and maximizing the between-class distance simultaneously
LDA Formulation

• given:
  • data set $X = \{x_1, ..., x_N\}, x_i \in \mathbb{R}^D$.
  • each data point belongs exactly to one of $C$ object classes $\{L_1, ..., L_C\}$.
  • $X_i$ is the data matrix for class $L_i$.
  • $N_i$ is the number of vectors in class $L_i$, thus $N = \sum N_i$.
  • $m_i$ is the class mean and $m$ is the global mean of $X$.

• define:
  • within-class scatter matrix $S_w$:
    $$S_w = \sum_{i=1}^{C} \sum_{x_j \in L_i} (x_j - m_i)(x_j - m_i)^T = H_w H_w^T,$$
    $$H_w = \begin{bmatrix} X_1 - m_1 \cdot 1_1^T, \ldots, X_C - m_C \cdot 1_C^T \end{bmatrix} \in \mathbb{R}^{D \times N}$$
  • between-class scatter matrix $S_b$:
    $$S_b = \sum_{i=1}^{C} N_i (m_i - m)(m_i - m)^T = H_b H_b^T,$$
    $$H_b = \begin{bmatrix} \sqrt{N_1}(m_1 - m), \ldots, \sqrt{N_C}(m_C - m) \end{bmatrix} \in \mathbb{R}^{D \times C}$$
  • total scatter matrix $S_t$:
    $$S_t = S_b + S_w = \sum_{i=1}^{N} (x_i - m)(x_i - m)^T = H_t H_t^T,$$
    $$H_t = \begin{bmatrix} x_1 - m, \ldots, x_N - m \end{bmatrix} \in \mathbb{R}^{D \times N}$$
LDA Formulation

- the Fisher criterion:

\[ J_F(\Phi) = \text{trace}\{ (\Phi^T S_w \Phi)^{-1} (\Phi^T S_b \Phi) \} , \]

\( (\text{trace}\{A\} = \sum_{i=1}^{N} a_{ii} \Rightarrow \text{sum of the diagonal elements}) \)

where \( \Phi \) is a linear transformation matrix.

- solution maximizing \( J_F \):

  - \( \Phi^* = \arg \max [\frac{\Phi^T S_b \Phi}{\Phi^T S_w \Phi}] \)
  - the set of the first eigenvectors \( \{\phi_i\} \) that satisfies

\[ S_b \phi = \lambda S_w \phi \]

(generalized eigenvalue problem)

\( \Rightarrow J_F \) is maximized by optimal linear transformation \( \Phi^* \) such that the projected data are most linearly separable

[derivation]
Input 2D dataset:
\[ \mathbf{X}_1 = \{(4, 2), (2, 4), (2, 3), (3, 6), (4, 4)\} \]
\[ \mathbf{X}_2 = \{(9, 10), (6, 8), (9, 5), (8, 7), (10, 8)\} \]

Solution:

- **class means:**
  \[ \mathbf{m}_1 = (3.0, 3.8)^T \]
  \[ \mathbf{m}_2 = (8.4, 7.6)^T \]

- **class covariances:**
  \[ \text{cov}(\mathbf{X}_1) = \begin{pmatrix} 1.00 & -0.25 \\ -0.25 & 2.20 \end{pmatrix} \]
  \[ \text{cov}(\mathbf{X}_2) = \begin{pmatrix} 2.30 & -0.05 \\ -0.05 & 3.30 \end{pmatrix} \]

- **within- and between-class scatter matrices:**
  \[ \mathbf{S}_w = \text{cov}(\mathbf{X}_1) + \text{cov}(\mathbf{X}_2) = \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix} \]
  \[ \mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T = \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} \]
We solve the generalized eigenvalue problem:

\[ S_w^{-1}S_b \phi = \lambda \phi \]

\[ \Rightarrow |S_w^{-1}S_b - \lambda I| = 0 \]

\[ \Rightarrow \left| \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \]

\[ \Rightarrow \left| \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0 \]

\[ \Rightarrow \left| \begin{pmatrix} 9.2213 - \lambda & 6.489 \\ 4.2339 & 2.9794 - \lambda \end{pmatrix} \right| = (9.2213 - \lambda)(2.9794 - \lambda) - (6.489 \cdot 4.2339) = 0 \]

\[ \Rightarrow \lambda^2 - 12.2007\lambda = 0 \]

\[ \Rightarrow \lambda(\lambda - 12.2007) = 0 \]

\[ \Rightarrow \lambda_1 = 0, \quad \lambda_2 = 12.2007 \]
Hence

\[
\begin{pmatrix}
9.2213 & 6.489 \\
4.2339 & 2.9794
\end{pmatrix}
\phi_1 = \lambda_1 \begin{pmatrix}
\phi_1 \\
\phi_2
\end{pmatrix}
\]

\[
\Rightarrow \begin{pmatrix}
9.2213 & 6.489 \\
4.2339 & 2.9794
\end{pmatrix}
\phi_1 = 0 \begin{pmatrix}
\phi_1 \\
\phi_2
\end{pmatrix}
\]

\[
\begin{pmatrix}
9.2213 & 6.489 \\
4.2339 & 2.9794
\end{pmatrix}
\phi_2 = \lambda_2 \begin{pmatrix}
\phi_1 \\
\phi_2
\end{pmatrix}
\]

\[
\Rightarrow \begin{pmatrix}
9.2213 & 6.489 \\
4.2339 & 2.9794
\end{pmatrix}
\phi_2 = 12.2007 \begin{pmatrix}
\phi_1 \\
\phi_2
\end{pmatrix}
\]

Thus

\[
\phi_1 = \begin{pmatrix}
-0.5755 \\
0.8178
\end{pmatrix}
\]

\[
\phi_2 = \begin{pmatrix}
0.9088 \\
0.4173
\end{pmatrix} = \phi^*
\]
The projection vector corresponding to the **smallest** eigen value

Using this vector leads to **bad separability** between the two classes.
LDA Example [courtesy of A. Farag & S. Elhabian: CVIP Lab]

The projection vector corresponding to the highest eigen value

LDA projection vector with the highest eigen value = 12.2007

Classes PDF: using the LDA projection vector with highest eigen value = 12.2007

Using this vector leads to good separability between the two classes
Fukunaga-Koontz Transform (FKT) and LDA

- set

\[ S_t = S_w + S_b = [U, U_\perp] \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^T \\ U_\perp^T \end{bmatrix}. \]

- \( S_t \) may be singular, \( r = \text{rank}(S_t) < D \)
- \( D = \text{diag}\{\lambda_1, \ldots, \lambda_r\}, \lambda_1 \geq \ldots \geq \lambda_r > 0 \)
- \( U \in \mathbb{R}^{D \times r} \) is the set of eigenvectors corresponding to nonzero eigenvalues
- \( U_\perp \in \mathbb{R}^{D \times (D-r)} \) is the orthogonal complement of \( U \)
- rewrite \( S_t \) using transformation operator \( P = UD^{-1/2} \):

\[ P^T S_t P = P^T (S_w + S_b) P = \tilde{S}_w + \tilde{S}_b = I \]

- \( \tilde{S}_w = P^T S_w P = V \Lambda_w V^T \)
- \( \tilde{S}_b = P^T S_b P = V \Lambda_b V^T \)
- \( I = \Lambda_w + \Lambda_b \)
- \( V \in \mathbb{R}^{r \times r} \) is the orthogonal eigenvector matrix
- \( \Lambda_w, \Lambda_b \in \mathbb{R}^{r \times r} \) are diagonal eigenvalue matrices
- classification performed by nearest neighbor in the transformed subspace
Fukunaga-Koontz Transform (FKT) and LDA

The data space can be decomposed into 4 subspaces:

1. **S1**: \( \text{span}(S_b) \cap \text{null}(S_w) \), eigenvectors \( \{v_i\} \) corresp. to \( \lambda_w = 0 \) and \( \lambda_b = 1 \).
   Hence, \( \frac{\lambda_b}{\lambda_w} = \infty \).

2. **S2**: \( \text{span}(S_b) \cap \text{span}(S_w) \), eigenvectors \( \{v_i\} \) corresp. to \( 0 < \lambda_w < 1 \) and \( 0 < \lambda_b < 1 \).
   Hence, \( 0 < \frac{\lambda_b}{\lambda_w} < \infty \).

3. **S3**: \( \text{null}(S_b) \cap \text{span}(S_w) \), eigenvectors \( \{v_i\} \) corresp. to \( \lambda_w = 1 \) and \( \lambda_b = 0 \).
   Hence, \( \frac{\lambda_b}{\lambda_w} = 0 \).

4. **S4**: \( \text{null}(S_b) \cap \text{null}(S_w) \), eigenvectors corresp. to the zero eigenvalues of \( S_t \).

![Diagram](image)

*Fig. 1. The whole data space is decomposed into four subspaces via FKT. In \( U_\perp \), the null space of \( S_t \), there is no discriminant information. \( \lambda_b + \lambda_w = 1 \). Note that we represent all possible subspaces, but, in real cases, some of these subspaces may not be available.*
LDA/FKT algorithm

**Input:** Data matrix \( X \)

**Output:** Projection matrix \( \Phi_F \) such that \( J_F \) is maximized

1. Compute \( H_b, H_t \) from data matrix \( X \)

2. Apply QR decomposition on \( H_t = QR \), where \( Q \in \mathbb{R}^{D \times r_t}, R \in \mathbb{R}^{r_t \times N} \) and \( r_t = \text{rank}(H_t), N = |X|, D \)...dimensionality of \( x_i \)

3. Let \( \tilde{S}_t = RR^T \), since \( \tilde{S}_t = Q^T S_t Q = Q^T H_t H_t^T Q = RR^T \)

4. Let \( Z = Q^T H_b \)

5. Let \( \tilde{S}_b = ZZ^T \), since \( \tilde{S}_b = Q^T S_b Q = Q^T H_b H_b^T Q = ZZ^T \)

6. Compute the eigenvectors \( \{v_i\} \) and eigenvalues \( \{\lambda_i\} \) of \( \tilde{S}_t^{-1}\tilde{S}_b \)

7. Sort the eigenvectors \( v_i \) according to \( \lambda_i \) in decreasing order

8. The final projection matrix \( \Phi_F = QV \), where \( V = \{v_i\} \). Note that \( QV \) is the union of Subspaces 1,2, and 3. If only Subspaces 1 and 2 are needed, \( \Phi_F = QV_k \) (the first \( k \) columns of \( V \))

- Time complexity: \( O(DN^2) \)
- Space complexity: \( O(DN) \)
LDA Limitations

- optimal only for two Gaussian distributions with equal covariances
- fails when classes have the same mean and differ only in variance

⇒ transform the multiclass problem into a binary classification problem and define:

- intraclass space: \( \Omega_I = \{(x_i - x_j) \mid L(x_i) = L(x_j)\} \), \( L(x_i) \) is the label of \( x_i \)
- number of samples in \( \Omega_I \): \( N_I = \frac{1}{2} \sum n_i(n_i - 1) \)
- extraclass space: \( \Omega_E = \{(x_i - x_j) \mid L(x_i) \neq L(x_j)\} \)
- number of samples in \( \Omega_E \): \( N_E = \sum_{L_i \neq L_j} n_i n_j \)
- \( m_I = m_E = 0 \)
- \( S_I = H_I H_I^T = \frac{1}{N_I} \sum_{L(x_i) = L(x_j)} (x_i - x_j)(x_i - x_j)^T \)
- \( H_I = \frac{1}{\sqrt{N_I}} [..., (x_i - x_j), ...], \forall i > j \text{ such that } L(x_i) = L(x_j) \)
- \( S_E = H_E H_E^T = \frac{1}{N_E} \sum_{L(x_i) \neq L(x_j)} (x_i - x_j)(x_i - x_j)^T \)
- \( H_E = \frac{1}{\sqrt{N_E}} [..., (x_i - x_j), ...], \forall i > j \text{ such that } L(x_i) \neq L(x_j) \)
- \( S_t = \frac{N_I}{2N} S_I + \frac{N_E}{2N} S_E = S'_I + S'_E \)
Multiple Discriminant Analysis (MDA)

- Goal: find a subspace $\Phi$ in which $\Omega_I$ and $\Omega_E$ are the most separable.
- Based on Bhattacharyya distance:
  - Measures the overlap of any two probability density functions.
  - Error bound of the Bayes classifier.
  - For Gaussian pdfs:
    \[
    D_{bh} = \frac{1}{8}(m_E - m_I)^T \left( \frac{S_E + S_I}{2} \right)^{-1} (m_E - m_I) + \frac{1}{2} \ln \frac{|S_E + S_I|}{\sqrt{|S_E| \sqrt{|S_I|}}}
    \]
    - Since $m_I = m_E$:
      \[
      D_{bh} = \frac{1}{2} \ln \frac{|S_E + S_I|}{\sqrt{|S_E| \sqrt{|S_I|}}}
      = \frac{1}{4} \left\{ \ln |S_E^{-1} S_I + S_I^{-1} S_E + 2I| - D \ln 4 \right\}.
      \]
- Define new criterion based on Bhattacharyya distance:
  \[
  J_{MDA} = \ln \left| (\Phi^T S_E \Phi)^{-1} (\Phi^T S_I \Phi) + (\Phi^T S_I \Phi)^{-1} (\Phi^T S_E \Phi) + 2I_d \right|
  \]
- Find optimal subspace $\Phi^*$ maximizing class separability:
  \[
  \Phi^* = \arg \max \left( \ln \left| \frac{(\Phi^T S_I \Phi)}{\Phi^T S_E \Phi} + \frac{(\Phi^T S_E \Phi)}{\Phi^T S_I \Phi} + 2I_d \right| \right)
  \]
MDA/FKT Algorithm

**Input:** Data matrix $X$

**Output:** Projection matrix $\Phi_{MDA}$ such that $J_{MDA}$ is maximized

1. Compute $H_I, H_t$ from data matrix $X$

2. Apply QR decomposition on $H_t = QR$, where $Q \in \mathbb{R}^{D \times r_t}, R \in \mathbb{R}^{r_t \times N}$ and $r_t = \text{rank}(H_t)$

3. Let $\tilde{S}_t = RR^T$, since $\tilde{S}_t = Q^T S_t Q = Q^T H_t H_t^T Q = RR^T$

4. Let $Z = Q^T H_I$

5. Let $\tilde{S}_I' = \frac{N_I}{2N} ZZ^T$, since $\tilde{S}_I' = Q^T S_I' Q = \frac{N_I}{2N} Q^T H_I H_I^T Q = \frac{N_I}{2N} ZZ^T$

6. Compute the eigenvectors $\{v_i\}$ and eigenvalues $\{\sigma_i\}$ of $\tilde{S}_t^{-1}\tilde{S}_I'$

7. Compute the generalized eigenvalues $\{\lambda_i\}$ of $(S_I, S_E)$ using $\lambda_i = \frac{N_I \sigma_i}{N_E (1 - \sigma_i)}$

8. Sort the eigenvectors $v_i$ according to $\lambda_i + \frac{1}{\lambda_i}$ in decreasing order

9. The final projection matrix $\Phi_{MDA} = QV_k$ (the first $k$ columns of $V$), where $V = \{v_i\}$. Note that $k$ could be greater than $C - 1$

- Time complexity: $O(DN^2)$
- Space complexity: $O(DN)$
Experiments: LDA vs. MDA

1. Three Gaussian classes: the same mean, different covariance matrices

(a) Original 3D data. (b) Two-dimensional projection by MDA/FKT. Note, that LDA-based methods fail since $S_b = 0$.

2. Two classes: Gaussian mixture

Fig. 5. (a) Original 3D data. (b) Two-dimensional projection by MDA/FKT. (c) One-dimensional projection by LDA/FKT. The projection of MDA/FKT is more separable than that of LDA/FKT because the former can provide a larger discriminant subspace.
Questions?
Derivation of the solution of $J_F$

To find the maximum of $J_F$ we derive and equate to zero:

\[
\frac{d}{d\Phi} J_F(\Phi) = \frac{d}{d\Phi} \left( \frac{\Phi^T S_b \Phi}{\Phi^T S_w \Phi} \right) = 0
\]

\[\Rightarrow (\Phi^T S_w \Phi) \frac{d}{d\Phi} (\Phi^T S_b \Phi) - (\Phi^T S_b \Phi) \frac{d}{d\Phi} (\Phi^T S_w \Phi) = 0\]

\[\Rightarrow (\Phi^T S_w \Phi) 2S_b \Phi - (\Phi^T S_b \Phi) 2S_w \Phi = 0 \quad / : 2\Phi^T S_w \Phi\]

\[\Rightarrow \left( \frac{\Phi^T S_w \Phi}{\Phi^T S_w \Phi} \right) S_b \Phi - \left( \frac{\Phi^T S_b \Phi}{\Phi^T S_w \Phi} \right) S_w \Phi = 0\]

\[\Rightarrow S_b \Phi - J_F(\Phi) S_w \Phi = 0\]

\[\Rightarrow S_w^{-1} S_b \Phi - J_F(\Phi) \Phi = 0\]

Solving the generalized eigenvalue problem

\[S_w^{-1} S_b \phi = \lambda \phi \quad \text{where} \quad \lambda = J_F(\phi) = \text{scalar}\]

yields

\[\Phi^* = \arg \max J_F(\Phi) = \arg \max \left( \frac{\Phi^T S_b \Phi}{\Phi^T S_w \Phi} \right)\]
The projection vector corresponding to the **smallest** eigen value

LDA projection vector with the other eigen value = 8.8818e-016

Using this vector leads to **bad separability** between the two classes
The projection vector corresponding to the highest eigen value

LDA projection vector with the highest eigen value = 12.207

Classes PDF: using the LDA projection vector with highest eigen value = 12.207

Using this vector leads to good separability between the two classes.
Fig. 1. The whole data space is decomposed into four subspaces via FKT. In $U_\perp$, the null space of $S_f$, there is no discriminant information. $\lambda_b + \lambda_w = 1$. Note that we represent all possible subspaces, but, in real cases, some of these subspaces may not be available.
Fig. 5. (a) Original 3D data. (b) Two-dimensional projection by MDA/FKT. (c) One-dimensional projection by LDA/FKT. The projection of MDA/FKT is more separable than that of LDA/FKT because the former can provide a larger discriminant subspace.