Linear Discriminant Analysis

Lecturer:
Jiří Matas

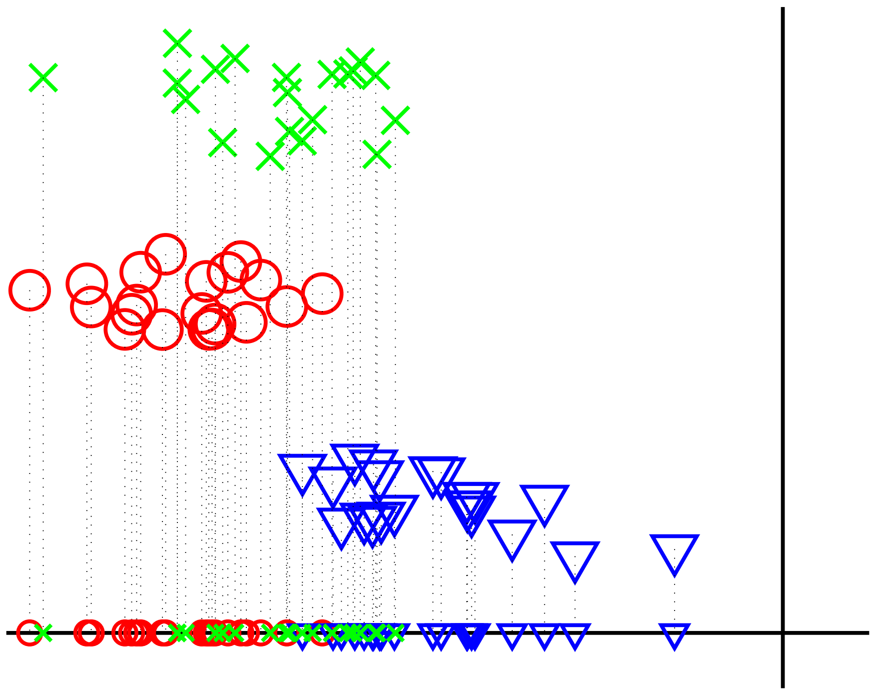
Authors:
Jiří Matas, Jana Kostlivá

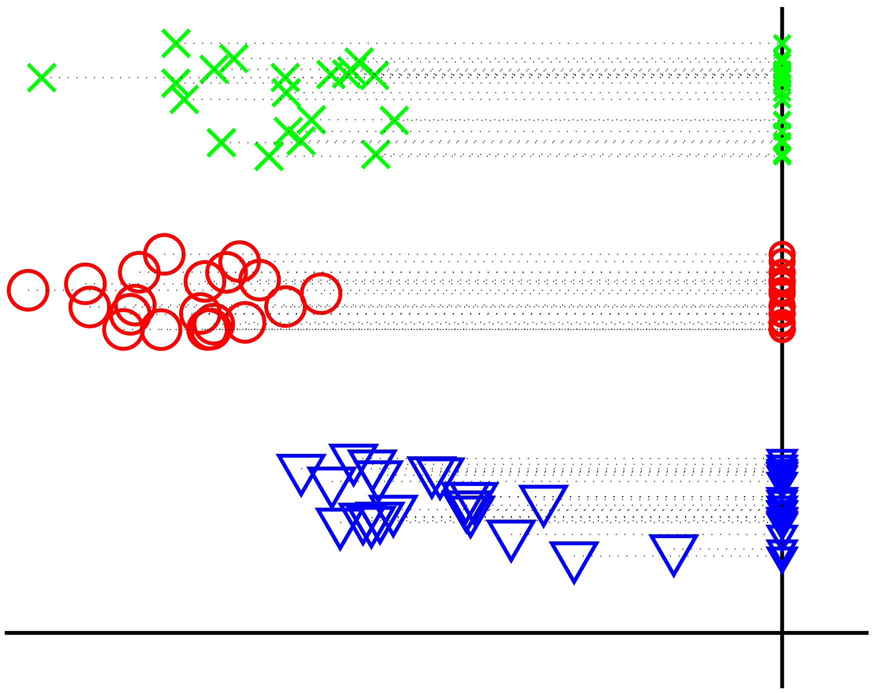
Centre for Machine Perception
Czech Technical University, Prague
http://cmp.felk.cvut.cz

8.6.2014

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 $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}, \mathbf{x}_i \in \mathbb{R}^D$.
  - each data point belongs exactly to one of C object classes $\{L_1, ..., L_C\}$.
  - $\mathbf{X}_i$ is the data matrix for class L_i.
  - N_i is the number of vectors in class L_i, thus $N = \sum N_i$.
  - $\mathbf{m}_i$ is the class mean and $\mathbf{m}$ is the global mean of $\mathbf{X}$.
- define:
  - within-class scatter matrix $\mathbf{S}_w$:
    $$\mathbf{S}_w = \sum_{i=1}^{C} \sum_{\mathbf{x}_j \in L_i} (\mathbf{x}_j - \mathbf{m}_i)(\mathbf{x}_j - \mathbf{m}_i)^T = \mathbf{H}_w \mathbf{H}_w^T,$$
    $$\mathbf{H}_w = [\mathbf{X}_1 - \mathbf{m}_1 \cdot \mathbf{1}_1^T, ..., \mathbf{X}_C - \mathbf{m}_C \cdot \mathbf{1}_C^T] \in \mathbb{R}^{D \times N}$$
  - between-class scatter matrix $\mathbf{S}_b$:
    $$\mathbf{S}_b = \sum_{i=1}^{C} N_i (\mathbf{m}_i - \mathbf{m})(\mathbf{m}_i - \mathbf{m})^T = \mathbf{H}_b \mathbf{H}_b^T,$$
    $$\mathbf{H}_b = [\sqrt{N_1}(\mathbf{m}_1 - \mathbf{m}), ..., \sqrt{N_C}(\mathbf{m}_C - \mathbf{m})] \in \mathbb{R}^{D \times C}$$
  - total scatter matrix $\mathbf{S}_t$:
    $$\mathbf{S}_t = \mathbf{S}_b + \mathbf{S}_w = \sum_{i=1}^{N} (\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T = \mathbf{H}_t \mathbf{H}_t^T,$$
    $$\mathbf{H}_t = [\mathbf{x}_1 - \mathbf{m}, ..., \mathbf{x}_N - \mathbf{m}] \in \mathbb{R}^{D \times N}$$

- the Fisher criterion:

$$J_F(\Phi) = trace\{(\Phi^T \mathbf{S}_w \Phi)^{-1}(\Phi^T \mathbf{S}_b \Phi)\},$$

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

where Φ is a linear transformation matrix.

- solution maximizing J_F:
  - $\Phi^* = \arg\max[\frac{\Phi^T \mathbf{S}_b \Phi}{\Phi^T \mathbf{S}_w \Phi}]$
  - the set of the first eigenvectors $\{\phi_i\}$ that satisfies

$$\mathbf{S}_b \phi = \lambda \mathbf{S}_w \phi$$

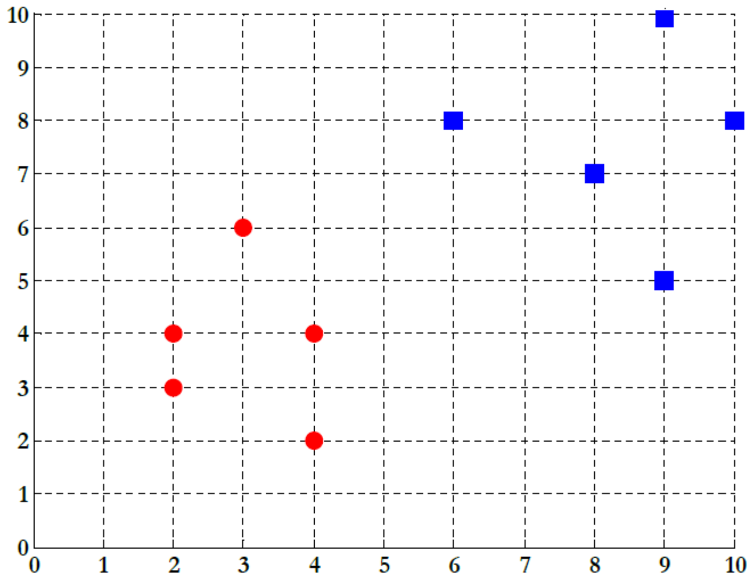
(generalized eigenvalue problem)

$\Rightarrow$ J_F is maximized by optimal linear transformation Φ^* 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:
  $$cov(\mathbf{X}_1) = \begin{pmatrix} 1.00 & -0.25 \\ -0.25 & 2.20 \end{pmatrix}$$
  $$cov(\mathbf{X}_2) = \begin{pmatrix} 2.30 & -0.05 \\ -0.05 & 3.30 \end{pmatrix}$$
- within- and between-class scater matrices:
  $$\mathbf{S}_w = cov(\mathbf{X}_1) + cov(\mathbf{X}_2) = \begin{bmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{bmatrix}$$
  $$\mathbf{S}_b = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^T = \begin{bmatrix} 29.16 & 20.52 \\ 20.52 & 14.44 \end{bmatrix}$$

We solve the generalized eigenvalue problem:

$$\mathbf{S}_w^{-1}\mathbf{S}_b\phi = \lambda\phi$$

$$\Rightarrow |\mathbf{S}_w^{-1}\mathbf{S}_b - \lambda\mathbf{I}| = 0$$

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

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

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

$$\Rightarrow \lambda^2 - 12.2007\lambda = 0$$

$$\Rightarrow \lambda(\lambda - 12.2007) = 0$$

$$\Rightarrow \lambda_1 = 0, \quad \lambda_2 = 12.2007$$

Hence

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

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

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

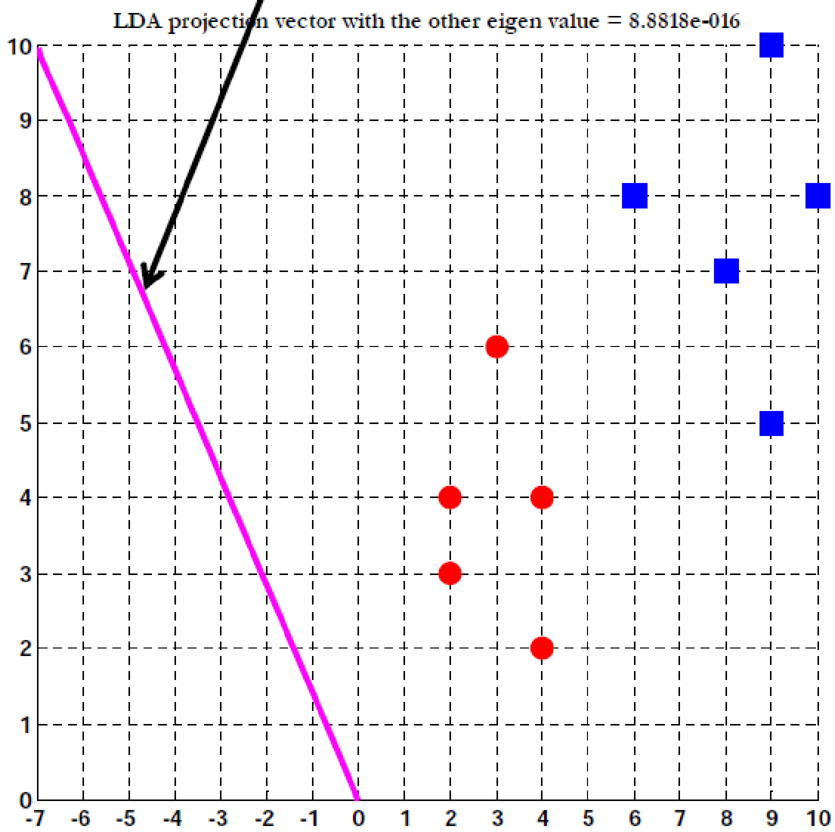
$$\Rightarrow \begin{pmatrix} 9.2213 & 6.489 \\ 4.2339 & 2.9794 \end{pmatrix} \phi_2 = 12.2007 \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix}$$

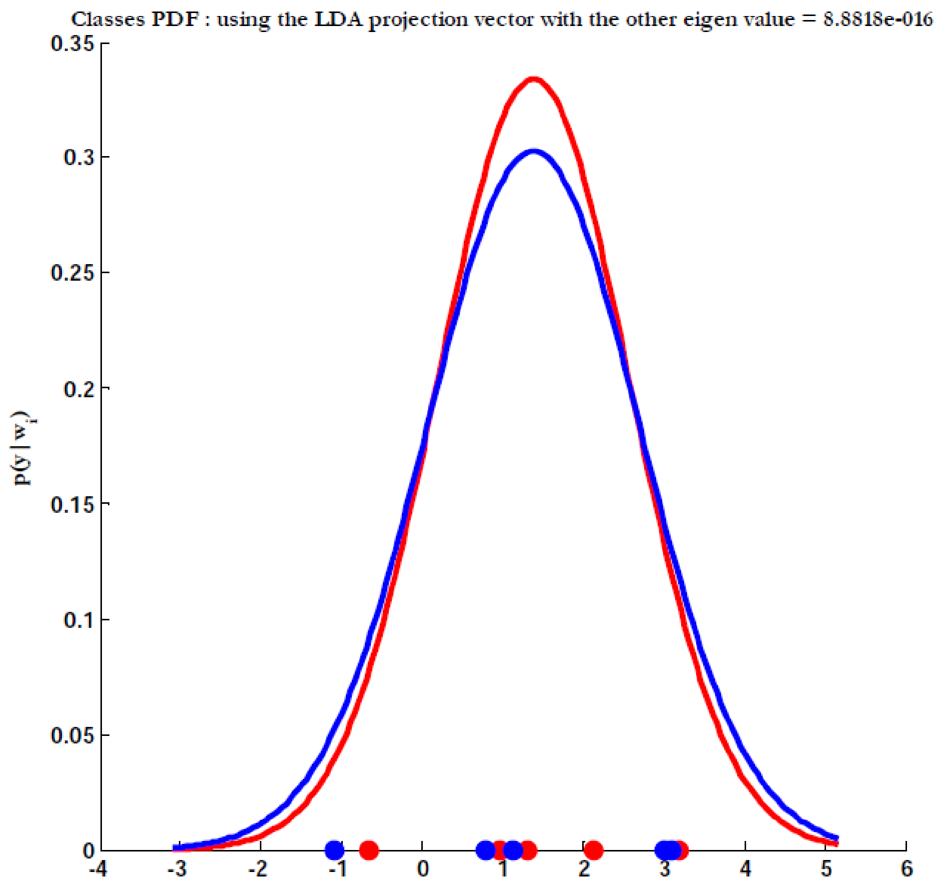
Thus

$$\phi_1 = \begin{pmatrix} -0.5755 \\ 0.8178 \end{pmatrix}$$

$$\phi_2 = \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix} = \phi^*$$

The projection vector corresponding to the **smallest** eigen value

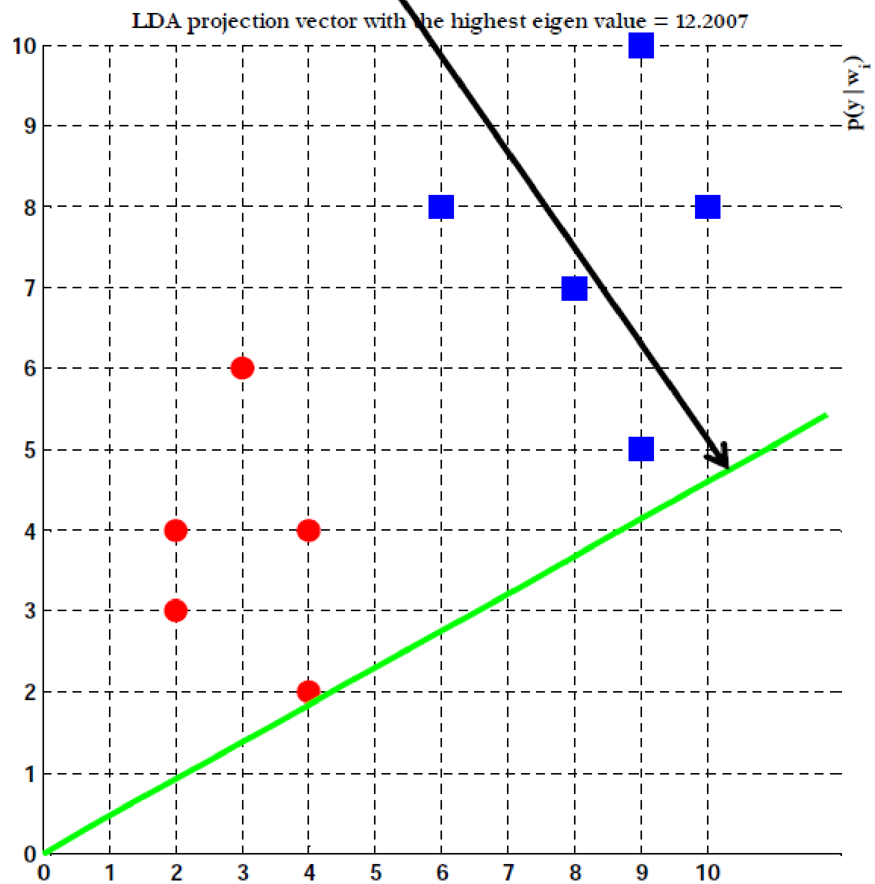


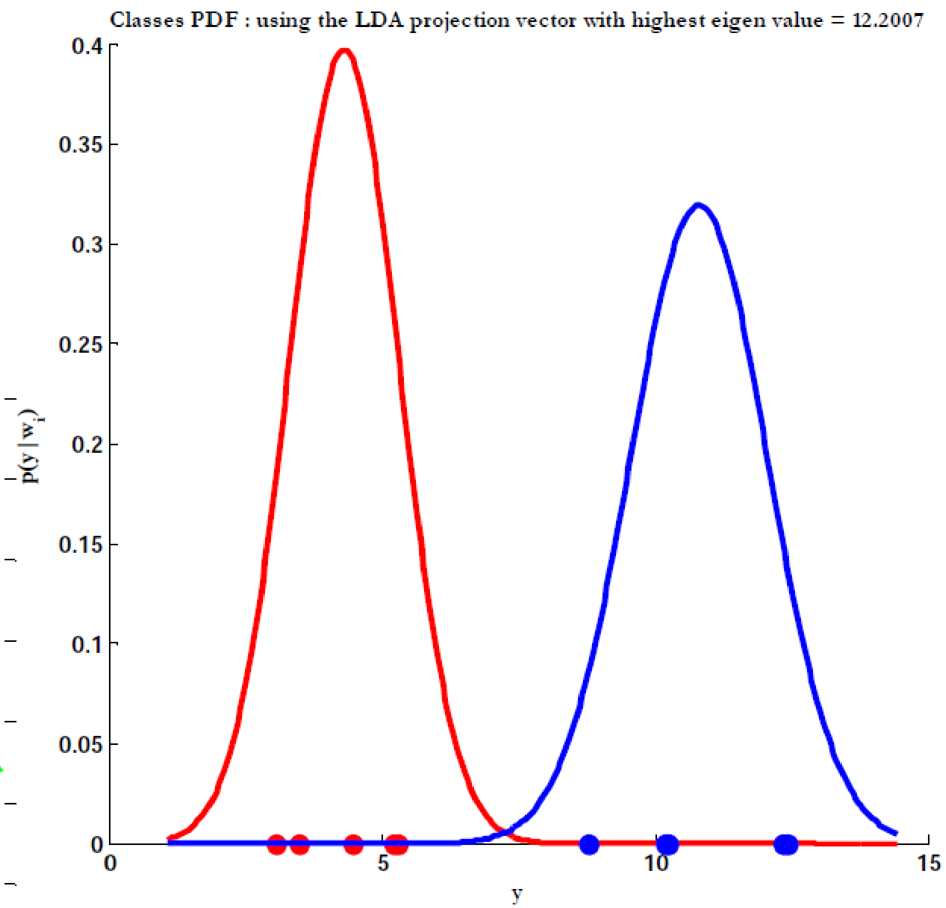


Using this vector leads to **bad separability** between the two classes

LDA Example [courtesy of A. Farag & S. Elhabian: CVIP Lab]

The projection vector corresponding to the **highest** eigen value





Using this vector leads to **good separability** between the two classes

- set

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = [\mathbf{U}, \mathbf{U}_\perp] \begin{bmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^T \\ \mathbf{U}_\perp^T \end{bmatrix}.$$

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

$$\mathbf{P}^T\mathbf{S}_t\mathbf{P} = \mathbf{P}^T(\mathbf{S}_w + \mathbf{S}_b)\mathbf{P} = \widetilde{\mathbf{S}}_w + \widetilde{\mathbf{S}}_b = \mathbf{I}$$

  - $\widetilde{\mathbf{S}}_w = \mathbf{P}^T\mathbf{S}_w\mathbf{P} = \mathbf{V}\Lambda_w\mathbf{V}^T$
  - $\widetilde{\mathbf{S}}_b = \mathbf{P}^T\mathbf{S}_b\mathbf{P} = \mathbf{V}\Lambda_b\mathbf{V}^T$
  - $\mathbf{I} = \Lambda_w + \Lambda_b$
  - $\mathbf{V} \in \mathbb{R}^{r \times r}$ 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

- data space can be decomposed into 4 subspaces:
  1. **S1:** $\mathrm{span}(\mathbf{S}_b) \bigcap \mathrm{null}(\mathbf{S}_w)$, eigenvectors $\{\mathbf{v}_i\}$ corresp. to $\lambda_w = 0$ and $\lambda_b = 1$. Hence, $\frac{\lambda_b}{\lambda_w} = \infty$.
  2. **S2:** $\mathrm{span}(\mathbf{S}_b) \bigcap \mathrm{span}(\mathbf{S}_w)$, eigenvectors $\{\mathbf{v}_i\}$ corresp. to $0 < \lambda_w < 1$ and $0 < \lambda_b < 1$. Hence, $0 < \frac{\lambda_b}{\lambda_w} < \infty$.
  3. **S3:** $\mathrm{null}(\mathbf{S}_b) \bigcap \mathrm{span}(\mathbf{S}_w)$, eigenvectors $\{\mathbf{v}_i\}$ corresp. to $\lambda_w = 1$ and $\lambda_b = 0$. Hence, $\frac{\lambda_b}{\lambda_w} = 0$.
  4. **S4:** $\mathrm{null}(\mathbf{S}_b) \bigcap \mathrm{null}(\mathbf{S}_w)$, eigenvectors corresp. to the zero eigenvalues of $\mathbf{S}_t$

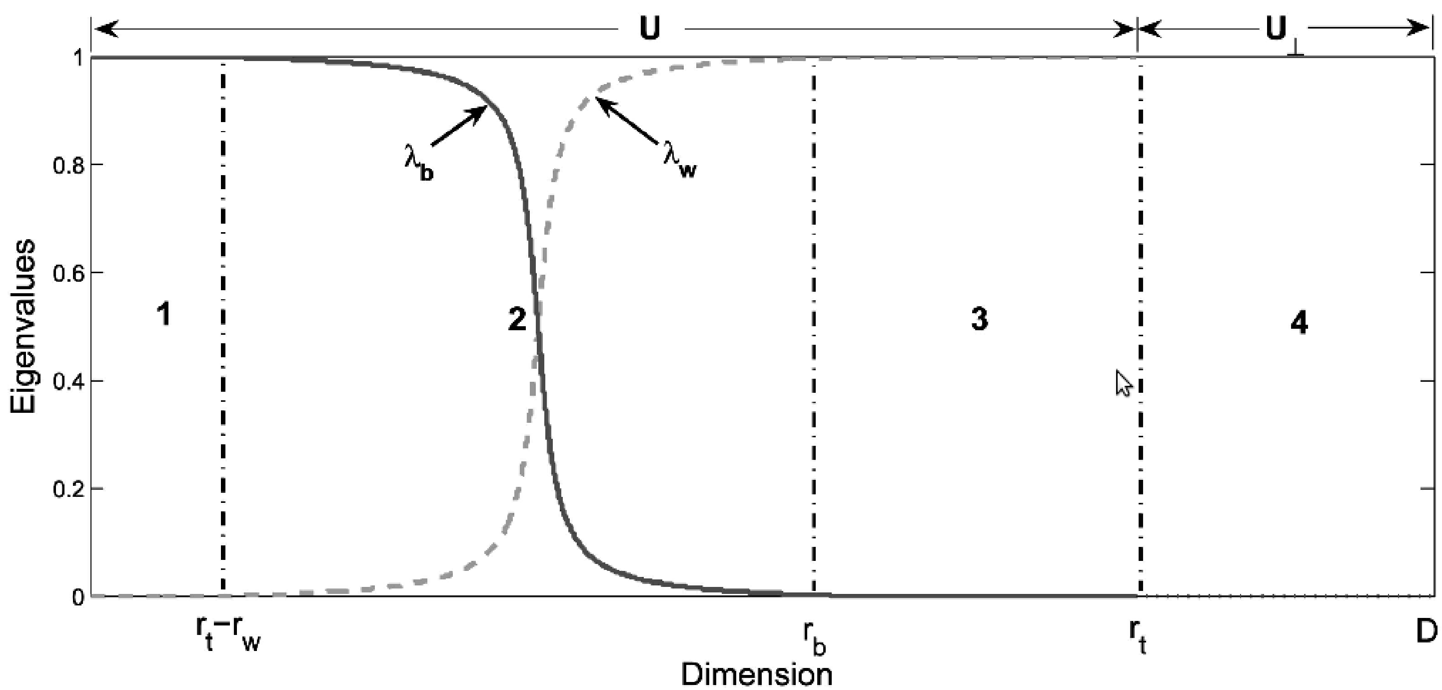


Fig. 1. The whole data space is decomposed into four subspaces via FKT. In $\mathbf{U}_\perp$, the null space of $\mathbf{S}_t$, there is no discriminant information. $\lambda_b + \lambda_w = 1$. Note that we represent all possible subspaces, but, in real cases, some of these subspaces may not be available.

LDA/FKT algorithm

Input: Data matrix $\mathbf{X}$
Output: Projection matrix Φ_F such that J_F is maximized

1. Compute $\mathbf{H}_b, \mathbf{H}_t$ from data matrix $\mathbf{X}$
2. Apply QR decomposition on $\mathbf{H}_t = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{D \times r_t}, \mathbf{R} \in \mathbb{R}^{r_t \times N}$ and $r_t = rank(\mathbf{H}_t),\ N = |\mathbf{X}|,\ D\text{...dimensionality of } \mathbf{x}_i$
3. Let $\widetilde{\mathbf{S}}_t = \mathbf{RR}^T$, since $\widetilde{\mathbf{S}}_t = \mathbf{Q}^T\mathbf{S}_t\mathbf{Q} = \mathbf{Q}^T\mathbf{H}_t\mathbf{H}_t^T\mathbf{Q} = \mathbf{RR}^T$
4. Let $\mathbf{Z} = \mathbf{Q}^T\mathbf{H}_b$
5. Let $\widetilde{S}_b = \mathbf{ZZ}^T$, since $\widetilde{\mathbf{S}}_b = \mathbf{Q}^T\mathbf{S}_b\mathbf{Q} = \mathbf{Q}^T\mathbf{H}_b\mathbf{H}_b^T\mathbf{Q} = \mathbf{ZZ}^T$
6. Compute the eigenvectors $\{\mathbf{v}_i\}$ and eigenvalues $\{\lambda_i\}$ of $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\mathbf{S}}_b$
7. Sort the eigenvectors $\mathbf{v}_i$ according to λ_i in decreasing order
8. The final projection matrix $\Phi_F = \mathbf{QV}$, where $\mathbf{V} = \{\mathbf{v}_i\}$. Note that $\mathbf{QV}$ is the union of Subspaces 1,2, and 3. If only Subspaces 1 and 2 are needed, $\Phi_F = \mathbf{QV}_k$ (the first k columns of $\mathbf{V}$)

- 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

$\Rightarrow$ transform the multiclass problem into a binary classification problem and define:

- intraclass space: $\Omega_I = \{(\mathbf{x}_i - \mathbf{x}_j) \mid L(\mathbf{x}_i) = L(\mathbf{x}_j)\}$, $L(\mathbf{x}_i)$ is the label of $\mathbf{x}_i$
- number of samples in Ω_I: $N_I = \frac{1}{2}\sum n_i(n_i - 1)$
- extraclass space: $\Omega_E = \{(\mathbf{x}_i - \mathbf{x}_j) \mid L(\mathbf{x}_i) \neq L(\mathbf{x}_j)\}$
- number of samples in Ω_E: $N_E = \sum_{L_i \neq L_j} n_i n_j$
- $\mathbf{m}_I = \mathbf{m}_E = 0$
- $\mathbf{S}_I = \mathbf{H}_I \mathbf{H}_I^T = \frac{1}{N_I} \sum_{L(\mathbf{x}_i) = L(\mathbf{x}_j)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$
- $\mathbf{H}_I = \frac{1}{\sqrt{N_I}} [..., (\mathbf{x}_i - \mathbf{x}_j), ...], \forall i > j$ such that $L(\mathbf{x}_i) = L(\mathbf{x}_j)$
- $\mathbf{S}_E = \mathbf{H}_E \mathbf{H}_E^T = \frac{1}{N_E} \sum_{L(\mathbf{x}_i) \neq L(\mathbf{x}_j)} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$
- $\mathbf{H}_E = \frac{1}{\sqrt{N_E}} [..., (\mathbf{x}_i - \mathbf{x}_j), ...], \forall i > j$ such that $L(\mathbf{x}_i) \neq L(\mathbf{x}_j)$
- $\mathbf{S}_t = \frac{N_I}{2N} \mathbf{S}_I + \frac{N_E}{2N} \mathbf{S}_E = \mathbf{S}_I' + \mathbf{S}_E'$

- Goal: find a subspace Φ in which Ω_I ans Ω_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:

$$D_{bh} = \frac{1}{8}(\mathbf{m}_E - \mathbf{m}_I)^T \left(\frac{\mathbf{S}_E+\mathbf{S}_I}{2}\right)^{-1} (\mathbf{m}_E - \mathbf{m}_I) + \frac{1}{2}\ln \frac{\left|\frac{\mathbf{S}_E+\mathbf{S}_I}{2}\right|}{\sqrt{|\mathbf{S}_E|}\sqrt{|\mathbf{S}_I|}}$$

  - since $\mathbf{m}_I = \mathbf{m}_E$:

$$\begin{aligned} D_{bh} &= \frac{1}{2}\ln \frac{\left|\frac{\mathbf{S}_E+\mathbf{S}_I}{2}\right|}{\sqrt{|\mathbf{S}_E|}\sqrt{|\mathbf{S}_I|}} \\ &= \frac{1}{4}\{\ln|\mathbf{S}_E^{-1}\mathbf{S}_I + \mathbf{S}_I^{-1}\mathbf{S}_E + 2\mathbf{I}| - D\ln 4\}. \end{aligned}$$

- define new criterion based on Bhattacharyya distance:

$$J_{MDA} = \ln|(\Phi^T\mathbf{S}_E\Phi)^{-1}(\Phi^T\mathbf{S}_I\Phi) + (\Phi^T\mathbf{S}_I\Phi)^{-1}(\Phi^T\mathbf{S}_E\Phi) + 2\mathbf{I}_d|$$

- find optimal subspace Φ^* maximizing class separability:

$$\Phi^* = \arg\max\left(\ln\left|(\frac{(\Phi^T\mathbf{S}_I\Phi)}{\Phi^T\mathbf{S}_E\Phi)} + \frac{(\Phi^T\mathbf{S}_E\Phi)}{(\Phi^T\mathbf{S}_I\Phi)} + 2\mathbf{I}_d\right|\right)$$

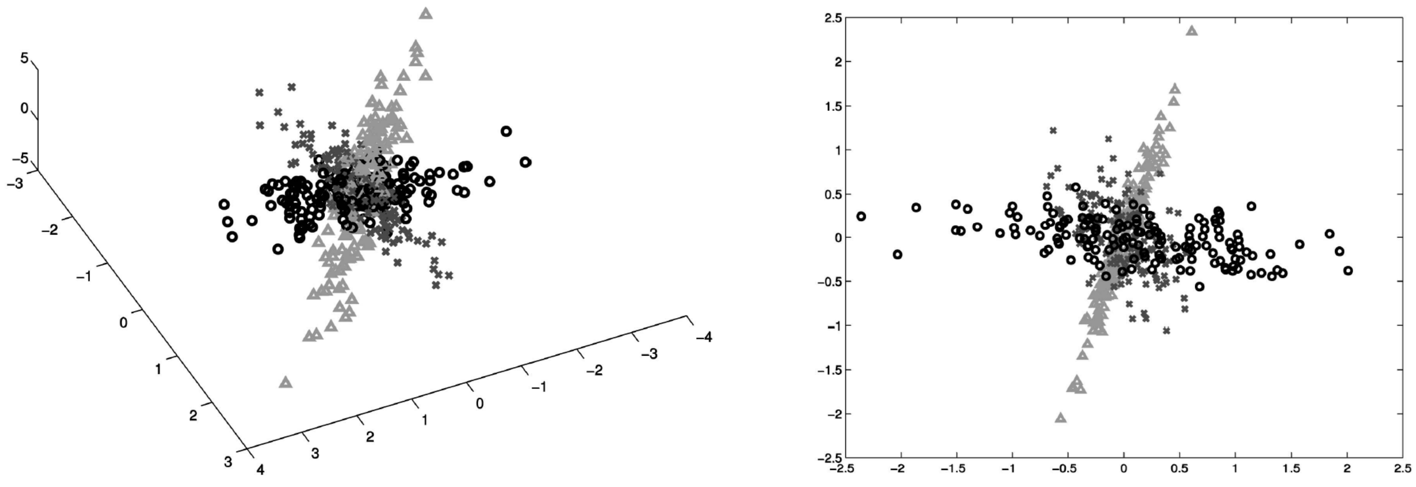
Input: Data matrix $\mathbf{X}$
Output: Projection matrix Φ_{MDA} such that J_{MDA} is maximized

1. Compute $\mathbf{H}_I, \mathbf{H}_t$ from data matrix $\mathbf{X}$
2. Apply QR decomposition on $\mathbf{H}_t = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{D \times r_t}, \mathbf{R} \in \mathbb{R}^{r_t \times N}$ and $r_t = rank(\mathbf{H}_t)$
3. Let $\widetilde{\mathbf{S}}_t = \mathbf{RR}^T$, since $\widetilde{\mathbf{S}}_t = \mathbf{Q}^T\mathbf{S}_t\mathbf{Q} = \mathbf{Q}^T\mathbf{H}_t\mathbf{H}_t^T\mathbf{Q} = \mathbf{RR}^T$
4. Let $\mathbf{Z} = \mathbf{Q}^T\mathbf{H}_I$
5. Let $\widetilde{S}_I' = \frac{N_I}{2N}\mathbf{ZZ}^T$, since $\widetilde{\mathbf{S}}_I' = \mathbf{Q}^T\mathbf{S}_I'\mathbf{Q} = \frac{N_I}{2N}\mathbf{Q}^T\mathbf{H}_I\mathbf{H}_I^T\mathbf{Q} = \frac{N_I}{2N}\mathbf{ZZ}^T$
6. Compute the eigenvectors $\{\mathbf{v}_i\}$ and eigenvalues $\{\sigma_i\}$ of $\widetilde{\mathbf{S}}_t^{-1}\widetilde{\mathbf{S}}_I'$
7. Compute the generalized eigenvalues $\{\lambda_i\}$ of $(\mathbf{S}_I, \mathbf{S}_E)$ using $\lambda_i = \frac{N_I\sigma_i}{N_E(1-\sigma_i)}$
8. Sort the eigenvectors $\mathbf{v}_i$ according to $\lambda_i + \frac{1}{\lambda_i}$ in decreasing order
9. The final projection matrix $\Phi_{MDA} = \mathbf{QV}_k$ (the first k columns of $\mathbf{V}$), where $\mathbf{V} = \{\mathbf{v}_i\}$. Note that k could be greater than $C-1$

- Time complexity: $O(DN^2)$
- Space complexity: $O(DN)$

Experiments: LDA vs. MDA

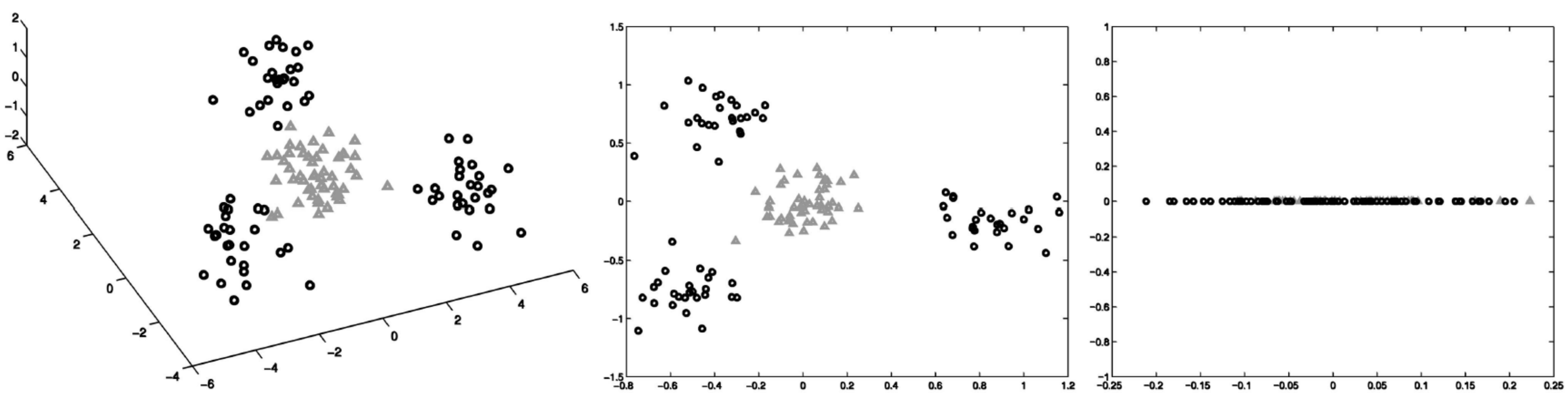
1. Three Gaussian classes: the same mean, different covariance matrices



(a) (b)

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

2. Two classes: Gaussian mixture



(a) (b) (c)

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:

$$\begin{aligned}
\frac{d}{d\Phi}J_F(\Phi) &= \frac{d}{d\Phi}\left(\frac{\Phi^T\mathbf{S}_b\Phi}{\Phi^T\mathbf{S}_w\Phi}\right) = 0 \\
&\Rightarrow (\Phi^T\mathbf{S}_w\Phi)\frac{d}{d\Phi}(\Phi^T\mathbf{S}_b\Phi) - (\Phi^T\mathbf{S}_b\Phi)\frac{d}{d\Phi}(\Phi^T\mathbf{S}_w\Phi) = 0 \\
&\Rightarrow (\Phi^T\mathbf{S}_w\Phi)2\mathbf{S}_b\Phi - (\Phi^T\mathbf{S}_b\Phi)2\mathbf{S}_w\Phi = 0 \qquad / : 2\Phi^T\mathbf{S}_w\Phi \\
&\Rightarrow \left(\frac{\Phi^T\mathbf{S}_w\Phi}{\Phi^T\mathbf{S}_w\Phi}\right)\mathbf{S}_b\Phi - \left(\frac{\Phi^T\mathbf{S}_b\Phi}{\Phi^T\mathbf{S}_w\Phi}\right)\mathbf{S}_w\Phi = 0 \\
&\Rightarrow \mathbf{S}_b\Phi - J_F(\Phi)\mathbf{S}_w\Phi = 0 \\
&\Rightarrow \mathbf{S}_w^{-1}\mathbf{S}_b\Phi - J_F(\Phi)\Phi = 0
\end{aligned}$$

Solving the generalized eigenvalue problem

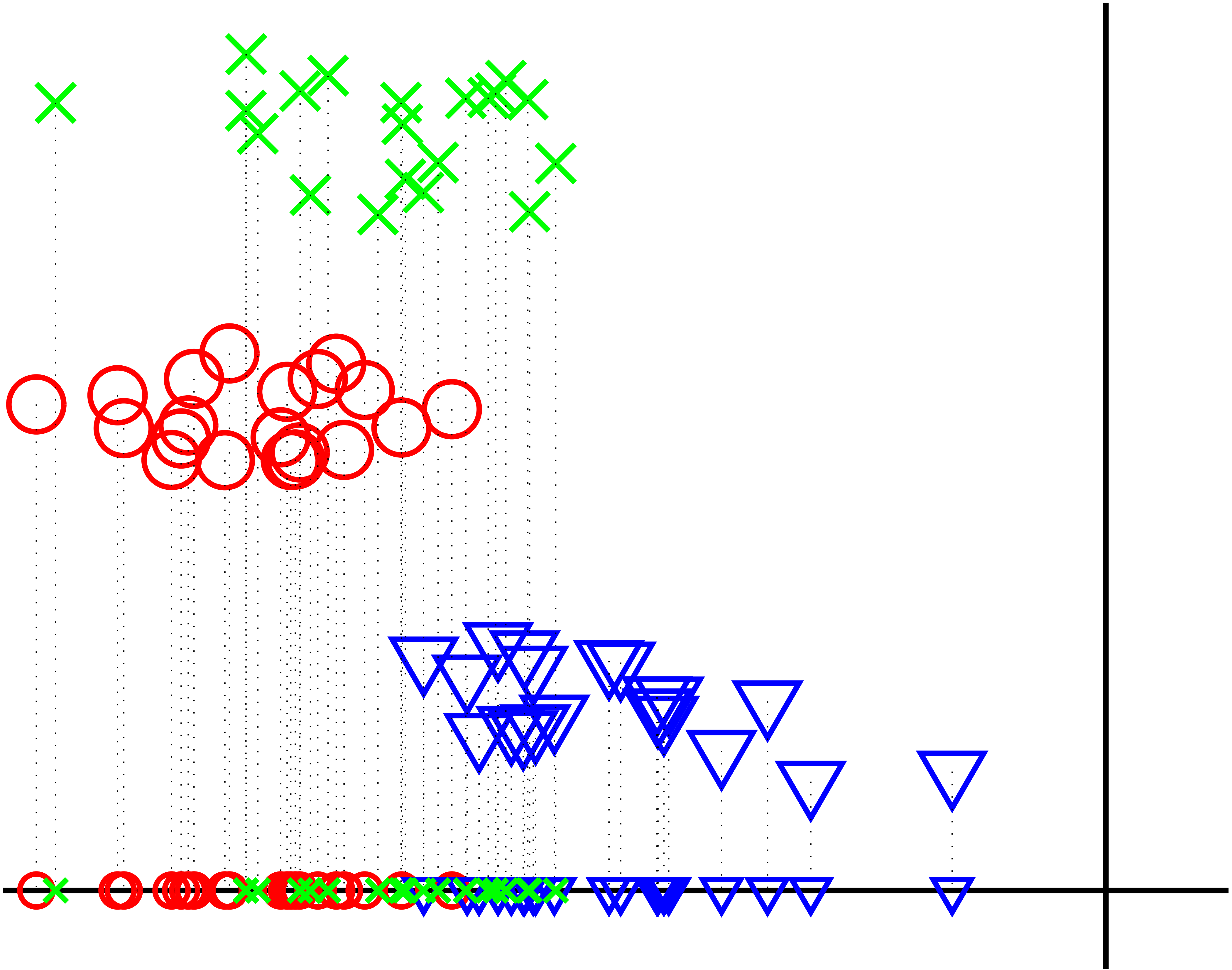
$$\mathbf{S}_w^{-1}\mathbf{S}_b\phi = \lambda\phi \quad \text{where} \quad \lambda = J_F(\phi) = scalar$$

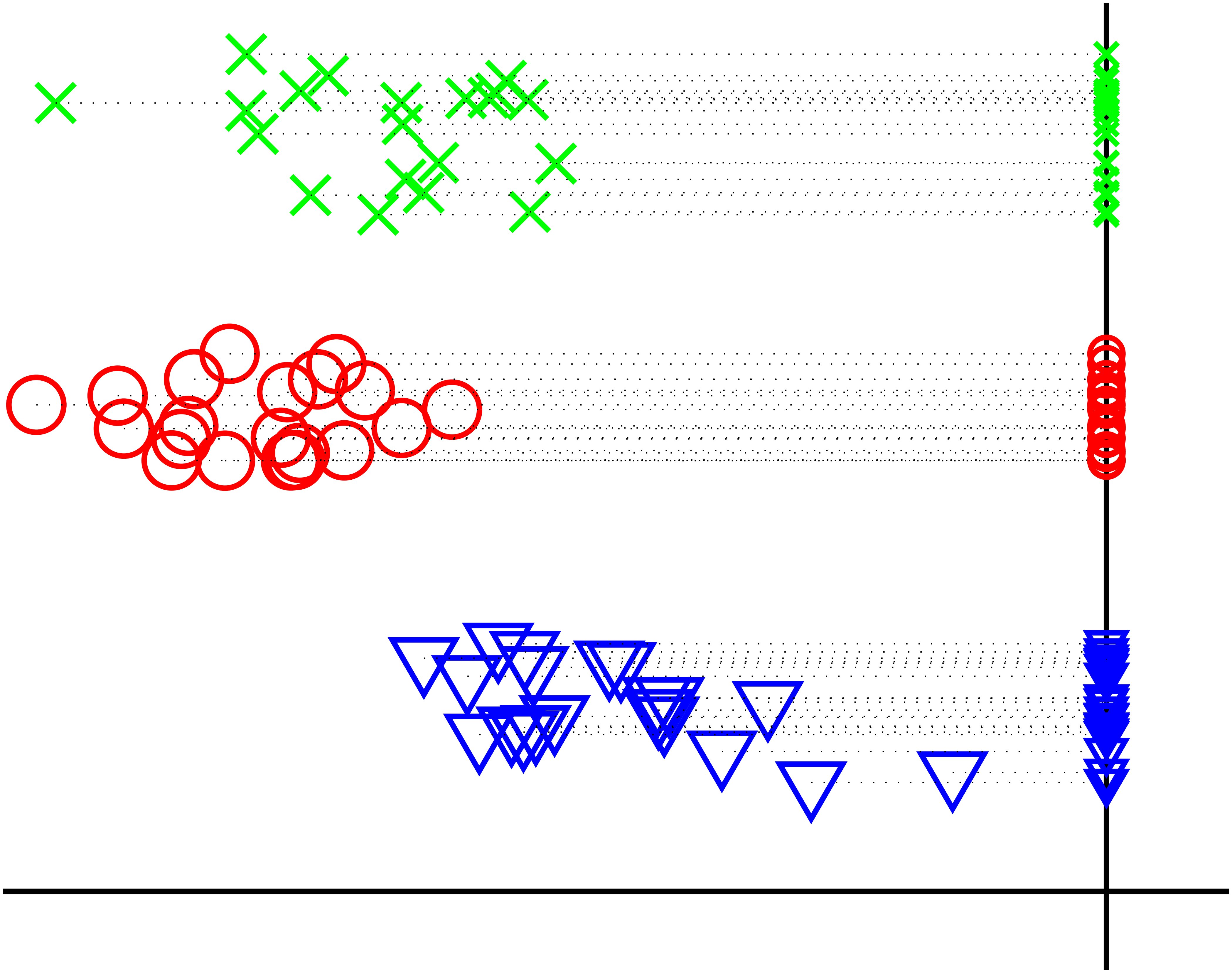
yields

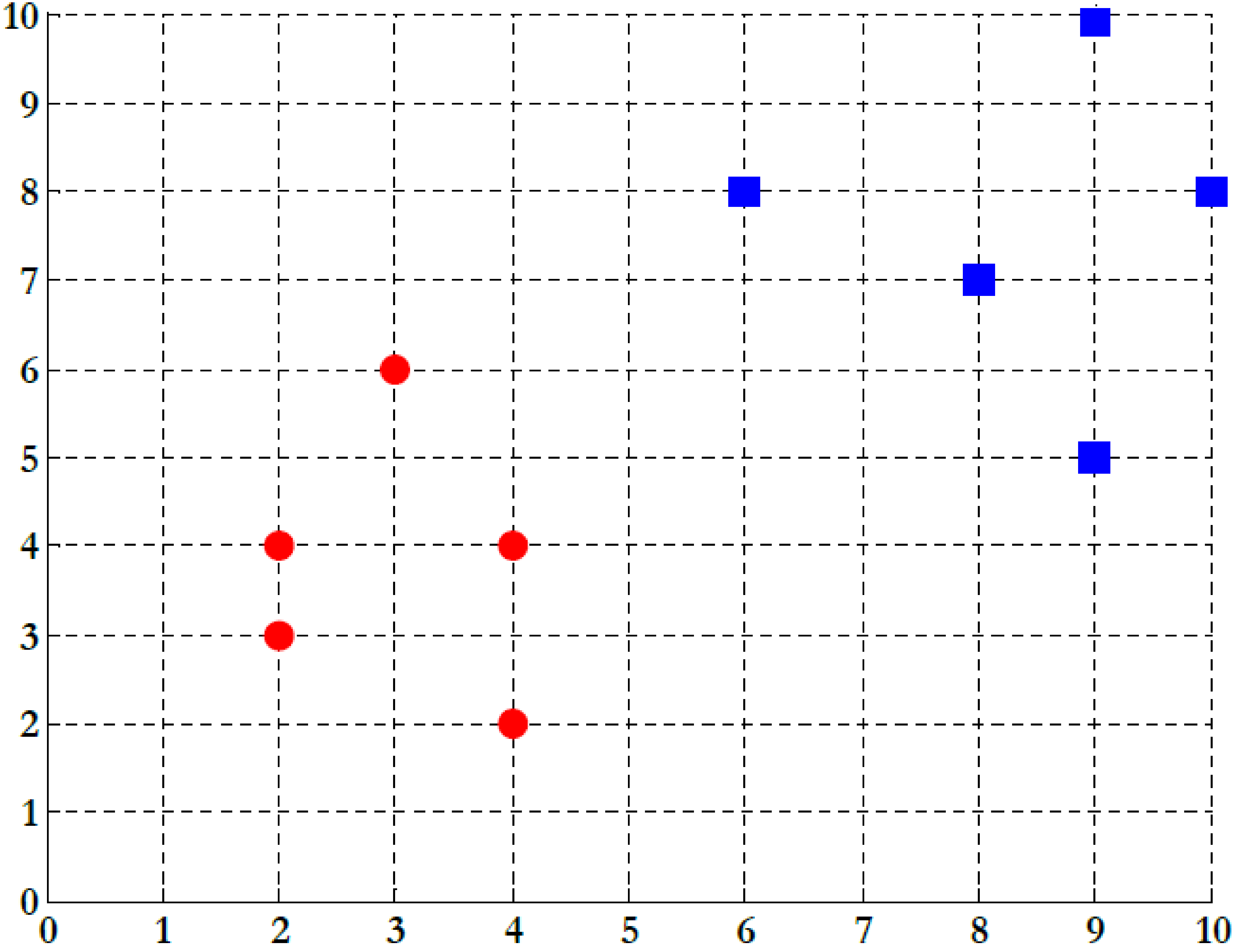
$$\Phi^* = \arg\max J_F(\Phi) = \arg\max\left(\frac{\Phi^T\mathbf{S}_b\Phi}{\Phi^T\mathbf{S}_w\Phi}\right)$$

[back]

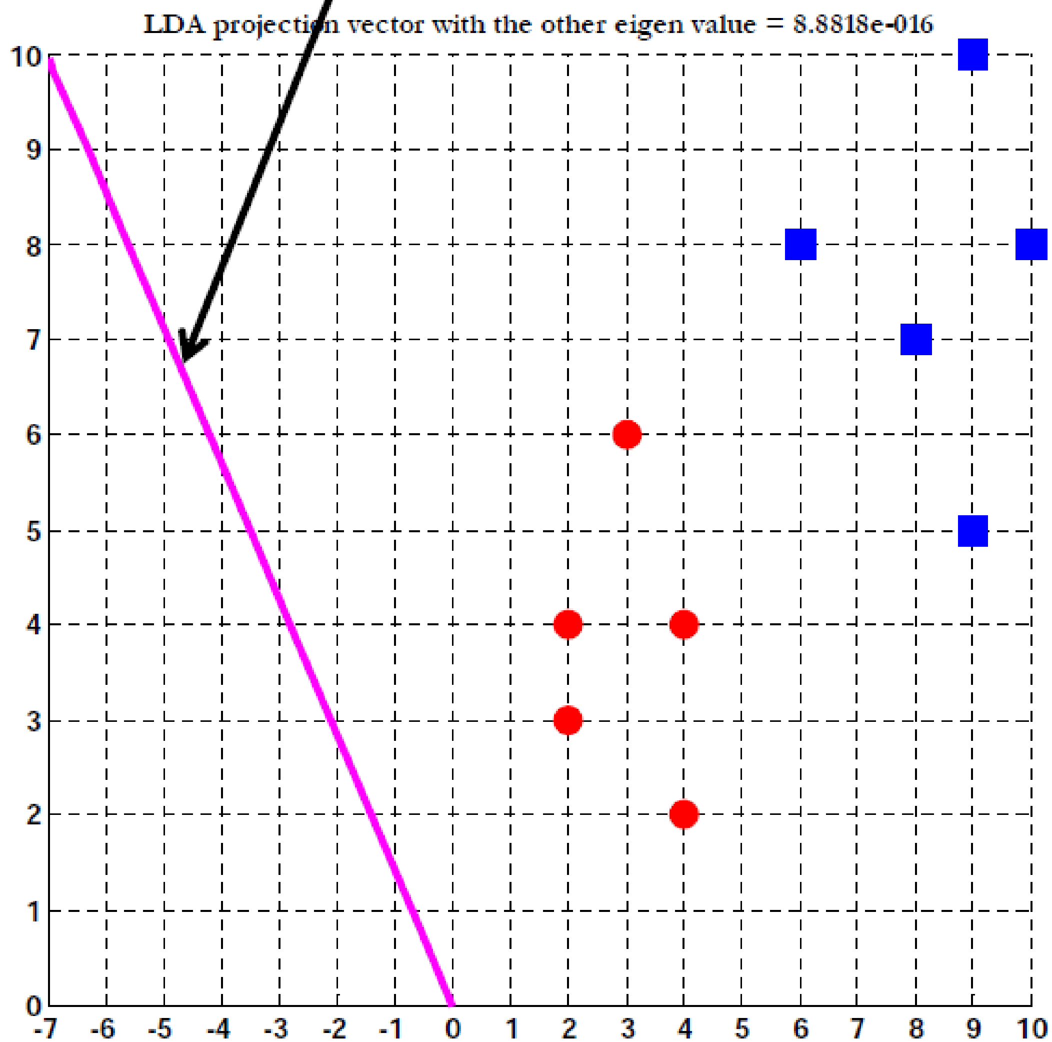
m p

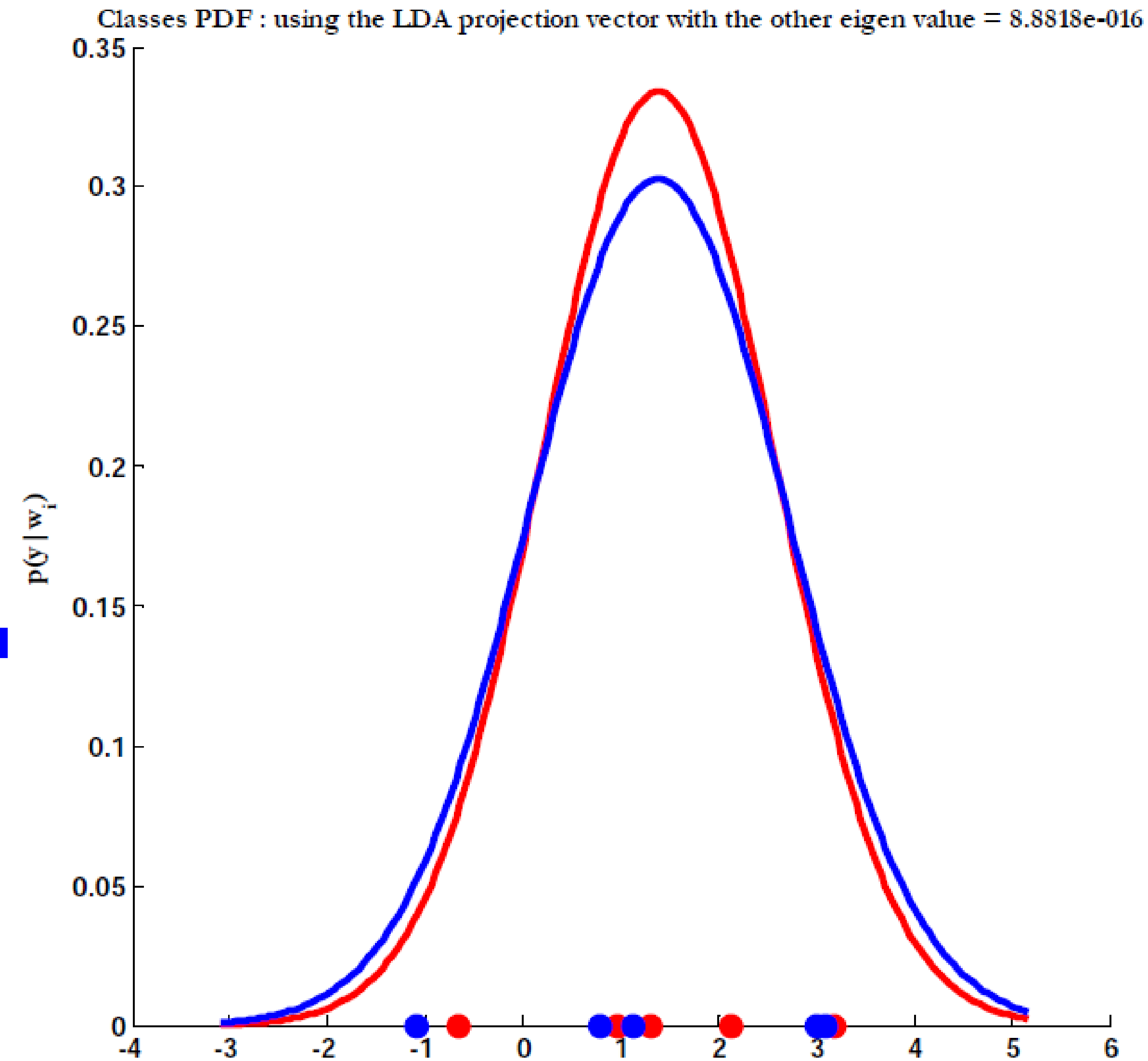





0
1
2
3
4
5
6
7
8
9
10
0
1
2
3
4
5
6
7
8
9
10

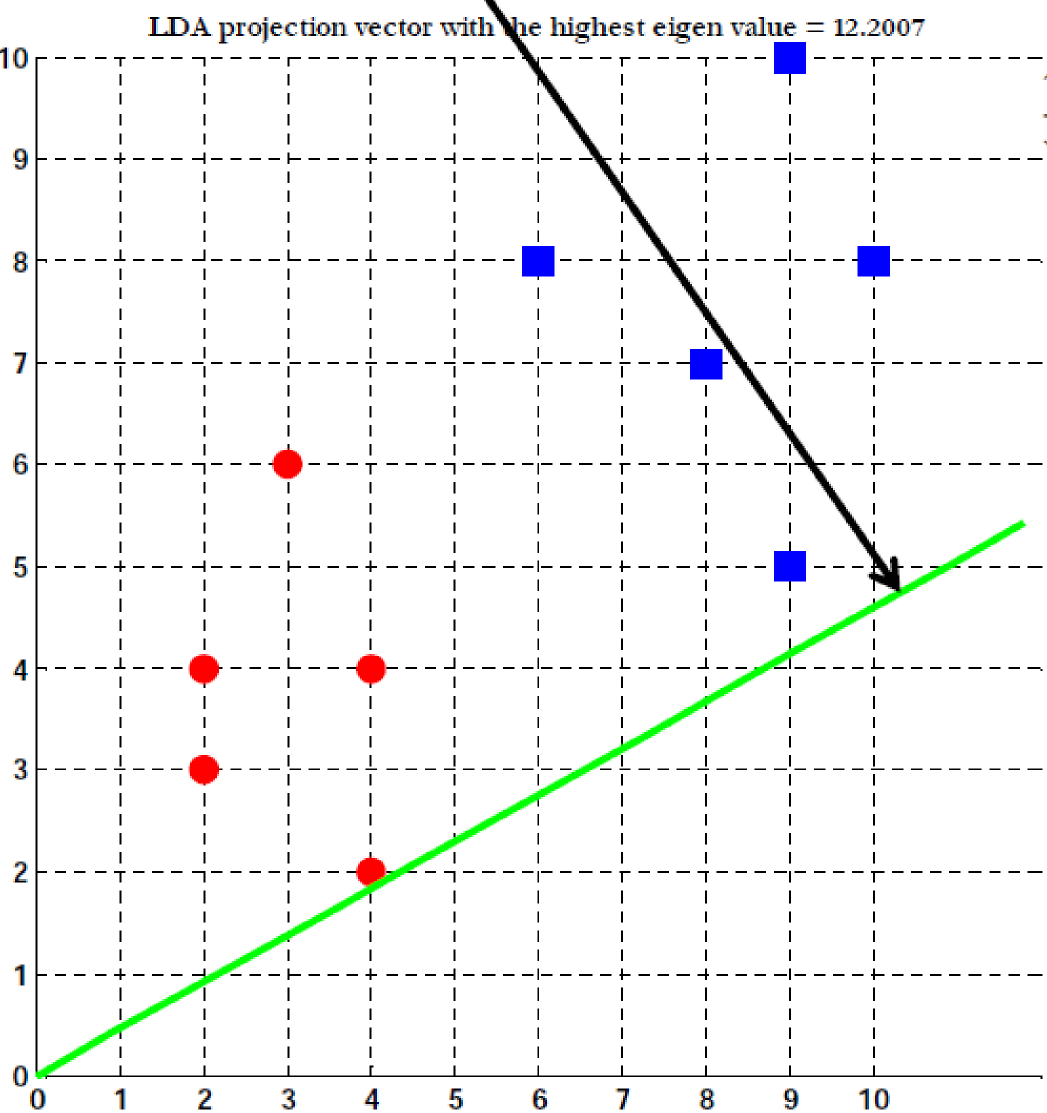
The projection vector corresponding to the **smallest** eigen value

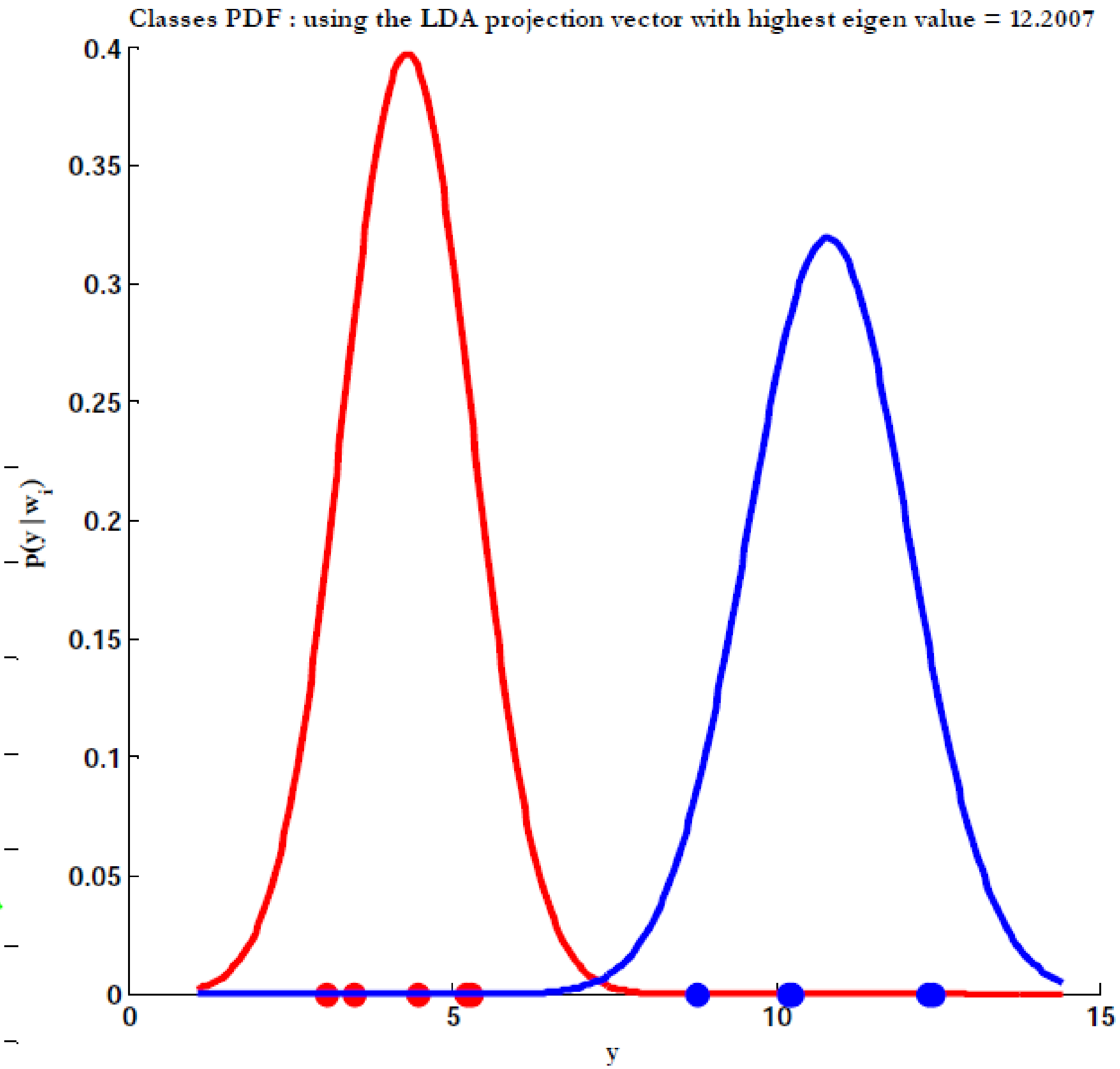




Using this vector leads to **bad separability** between the two classes

The projection vector corresponding to the **highest** eigen value





Using this vector leads to **good separability** between the two classes

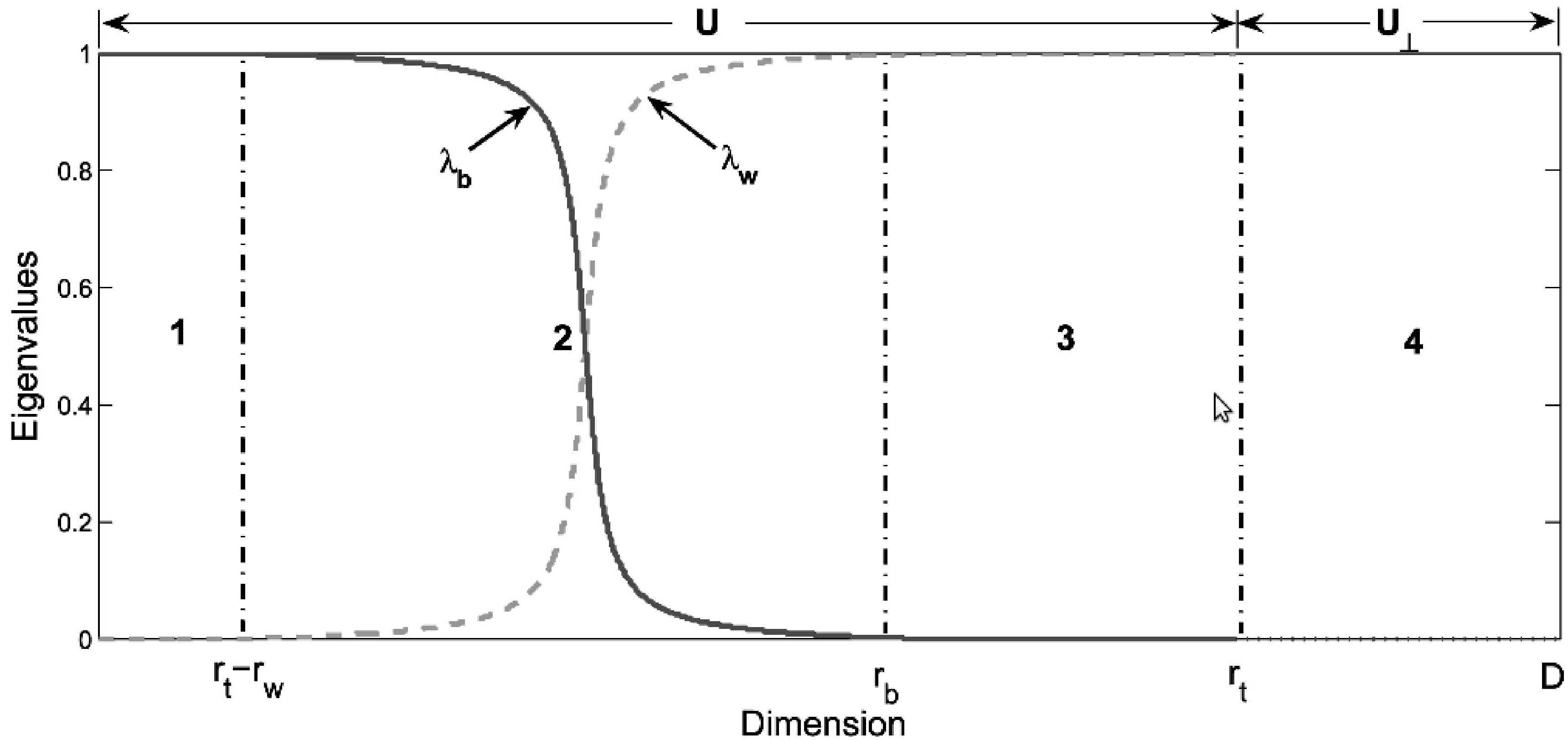
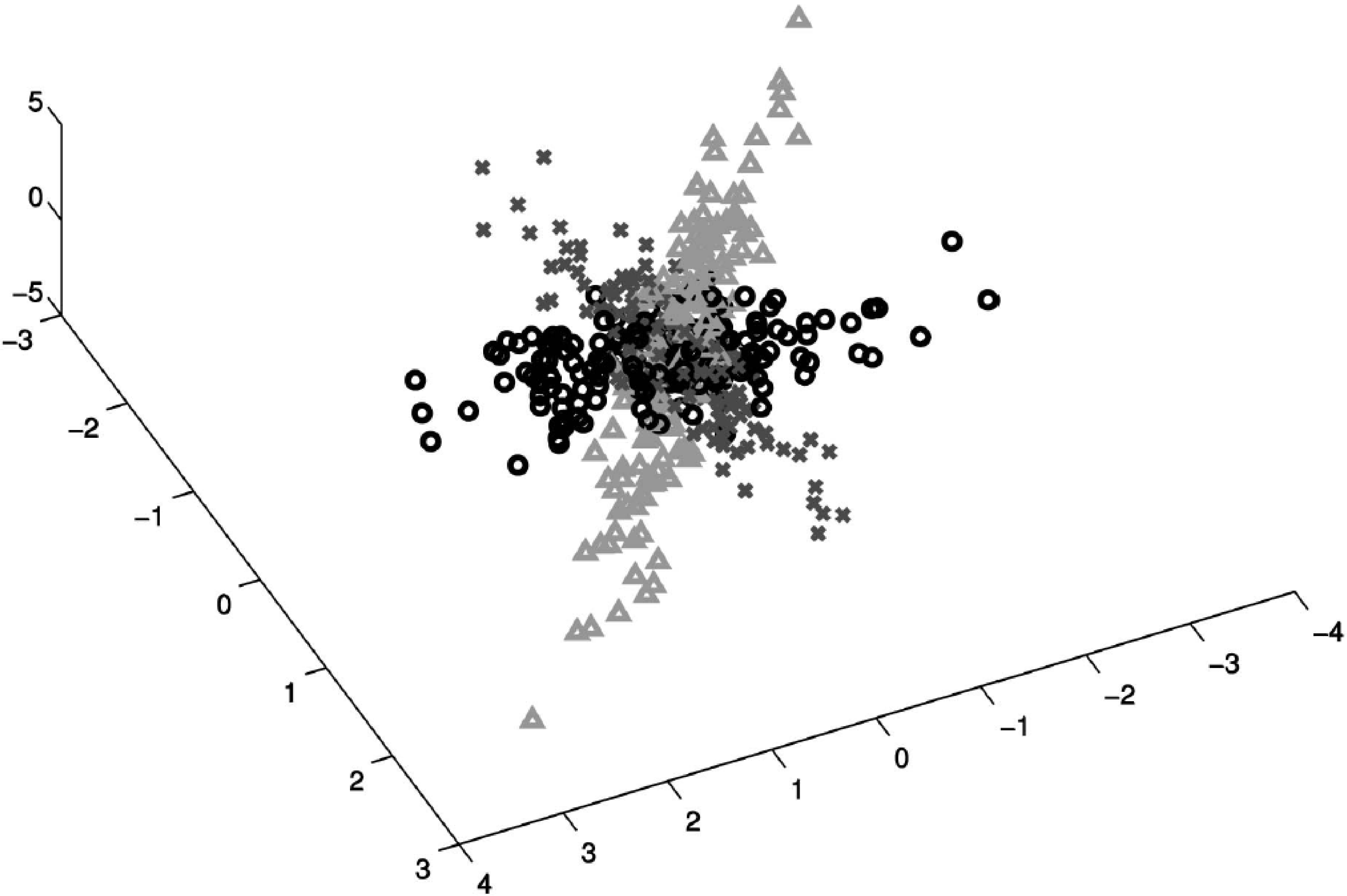
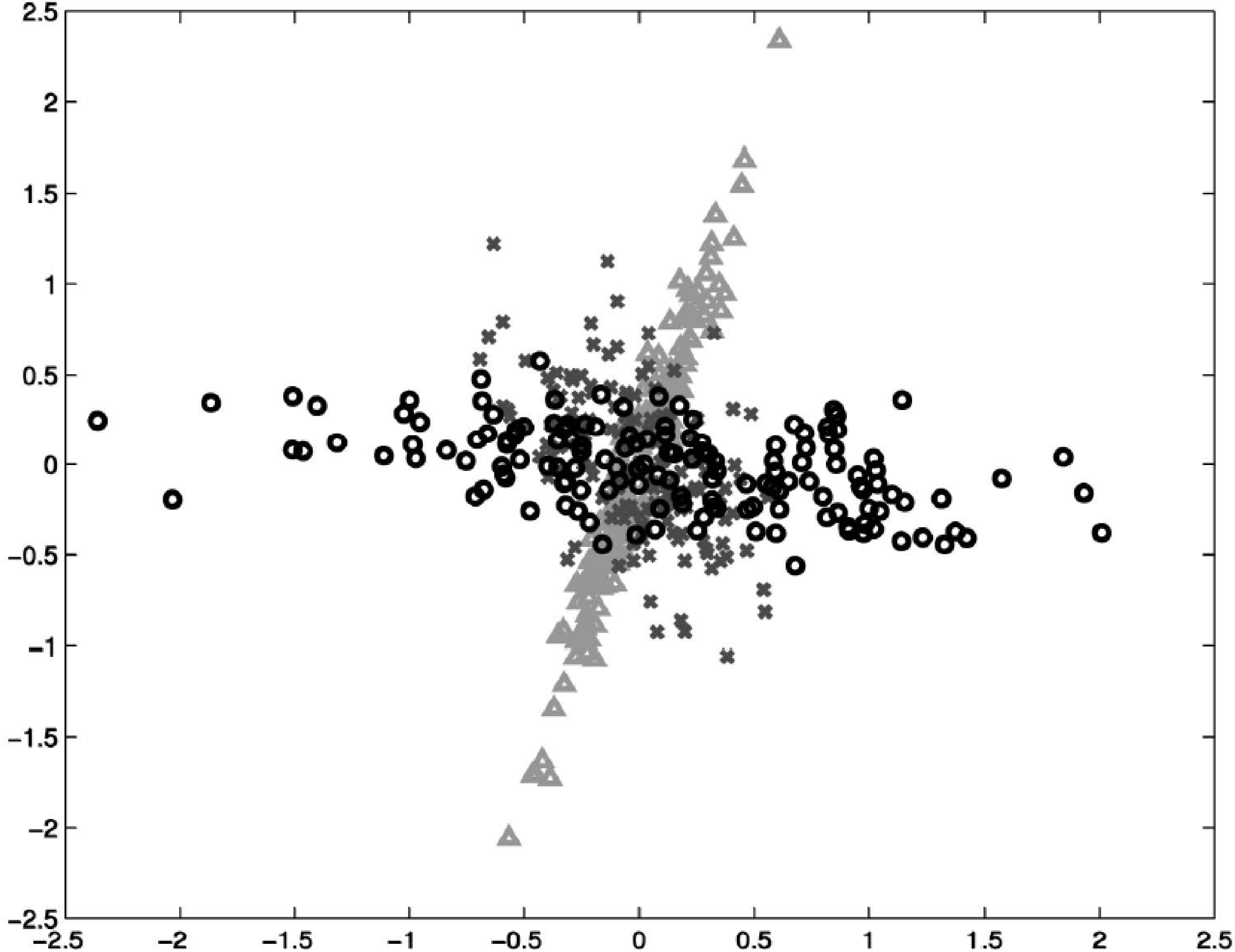


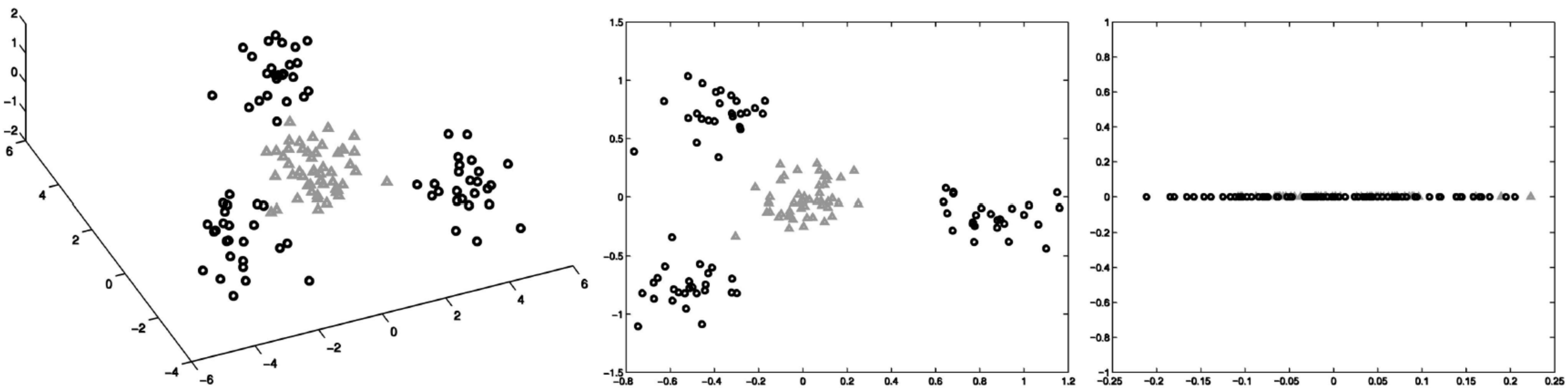
Fig. 1. The whole data space is decomposed into four subspaces via FKT. In $\mathbf{U}_\perp$, the null space of $\mathbf{S}_t$, there is no discriminant information. $\lambda_b + \lambda_w = 1$. Note that we represent all possible subspaces, but, in real cases, some of these subspaces may not be available.



(a)



(b)



(a) (b) (c)

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.