Implementational Aspects of Linear Discriminant Analysis for Classification Tasks

Jurjen Duintjer Tebbens

Institute of Computer Science

Academy of Sciences of the Czech Republic

joint work with

Pavel Schlesinger

Institute of Formal and Applied Linguistics

Charles University, Prague

Summer School on Numerical Linear Algebra in Signals and Systems

Monopoli, September 11, 2006.

Outline

1. Classification tasks: Introduction
2. Fisher’s Linear Discriminant Analysis (FLDA)
3. Fisher’s criterion for the $p \gg n$ problem
4. Efficient implementation
5. Experiments
6. Conclusions

1. Classification tasks: Introduction

Classification task:

- Assign a given sample, based on its properties, to one of the pre-defined classes
- Different classifiers utilize different decision rules, e.g. centroid based, nearest neighborhood and support vector machines methods
- Decision rules are derived from information gained from training samples, ’learning process’

1. Classification tasks: Introduction

Examples of modern applications:

Text document classification - information retrieval systems: Documents are represented by p-dimensional vectors, every variable corresponds to a keyword in the text, its value indicates the number of times the keyword occurs in the document.

Protein fold prediction: Protein strings are represented by p-dimensional vectors, every variable indicates the number of times a pair of amino-acids occurs in the string.

Biomedical signal processing - (our software is currently being integrated in the BioSig open source software library (TU Graz))

1. Classification tasks: Introduction

In applications like text document classification or protein fold class prediction, the samples may contain a very large number of variables.

To enhance efficiency one frequently performs a preprocessing step known as dimension reduction of the space of variables.

One of the simplest and most popular methods that incorporate dimension reduction is Fisher's Linear Discriminant Analysis (FLDA).

2. Fisher's Linear Discriminant Analysis (FLDA)

FLDA idea: project the variables on a space of small dimension such that class information is maximally preserved.

This is achieved with between-class- and within-class-covariance: With

- n: Number of training samples
- p: Number of variables
- $x_i \in \mathbb{R}^p$: The ith sample
- $\bar{x} \in \mathbb{R}^p$: The grand mean of all samples, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

the *total covariance matrix* **T** $\in \mathbb{R}^{p\times p}$ is defined as

$$\mathbf{T} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T.$$

2. Fisher’s Linear Discriminant Analysis (FLDA)

With

- g: The number of classes
- n_j: The number of samples in class j
- N_j: The set of indices i for which x_i is in class j
- $\bar{x}_j \in \mathbb{R}^p$: The mean vector in class j, $\bar{x}_j = \frac{1}{n_j} \sum_{i \in N_j} x_i$

the *between-class-covariance matrix* **B** $\in \mathbb{R}^{p \times p}$ is defined as

$$\mathbf{B} = \frac{1}{g-1} \sum_{j=1}^{g} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T$$

and the *within-class-covariance matrix* **W** $\in \mathbb{R}^{p \times p}$ is defined as

$$\mathbf{W} = \frac{1}{n-g} \sum_{j=1}^{g} \sum_{i \in N_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T.$$

2. Fisher’s Linear Discriminant Analysis (FLDA)

Clearly, **T**, **B** and **W** are symmetric positive semi-definite.
One can easily prove that

$$(n-1)\mathbf{T} = (g-1)\mathbf{B} + (n-g)\mathbf{W}.$$

For a sample $v \in \mathbb{R}^p$ we call:

- $v^T\mathbf{T}v$: Total covariance
- $v^T\mathbf{B}v$: Between class covariance
- $v^T\mathbf{W}v$: Within class covariance

FLDA seeks transformation vectors $c_i \in \mathbb{R}^p$, $i < g$, such that the transformed samples

$$(c_1, \ldots, c_i)^T x_j$$

have maximal between-class-covariance and minimal within-class-covariance. This leads to *Fisher’s criterion*:

2. Fisher’s Linear Discriminant Analysis (FLDA)

Fisher’s criterion: A transformation vector c_i must satisfy

$$\frac{c_i^T \mathbf{B} c_i}{c_i^T \mathbf{W} c_i} = \max_{c \in \mathbb{R}^p,\ c \neq 0} \frac{c^T \mathbf{B} c}{c^T \mathbf{W} c}.$$

The criterion is formulated for nonsingular **W**. Then it is equivalent to finding the largest eigenpairs of the generalized eigenproblem

$$(\mathbf{B} - \lambda \mathbf{W}) c = 0, \tag{1}$$

which can be transformed to a standard eigenproblem, e.g.

$$(\mathbf{W}^{-1} \mathbf{B} - \lambda \mathbf{I}) c = 0. \tag{2}$$

The FLDA-transformation space of dimension i, $i < g$, is spanned by the eigenvectors corresponding to the i largest eigenvalues.

3. Fisher's criterion for the $p \gg n$ problem

In many modern applications (text document classification, protein fold prediction), the number of variables is often so high that one cannot afford to work with the same number of samples (the so-called small sample size problem, also '$p \gg n$ problem').

As a sum of n rank one matrices,

$$\mathbf{W} = \frac{1}{n-g} \sum_{j=1}^{g} \sum_{i \in N_j} (x_i - \bar{x}_j)(x_i - \bar{x}_j)^T \in \mathbb{R}^{p \times p}$$

has rank$(\mathbf{W}) \leq n$.

When $p > n$, then **W** is singular.

3. Fisher's criterion for the $p \gg n$ problem

This makes Fisher's maximization problem

$$\max_{c \in \mathbb{R}^p,\ c \neq 0} \quad \frac{c^T \mathbf{B} c}{c^T \mathbf{W} c}$$

challenging:

- Coping with a generalized eigenproblem $(\mathbf{B} - \lambda \mathbf{W})c = 0$ where **B** and **W** have a common null space. Computing the Kronecker canonical form with GUPTRI?
- Meaning of Fisher's criterion ?

We first present five methods used in statistics for the $p \gg n$ case. In the next section we address their implementation.

3. Fisher's criterion for the $p \gg n$ problem

Some methods solve a modified generalized eigenproblem

$$(\mathbf{B} - \lambda \tilde{\mathbf{W}})c = 0,$$

with a nonsingular matrix satisfying in some sense

$$\tilde{\mathbf{W}} \approx \mathbf{W}.$$

1. Perturbation methods compute the SVD of **W** and modify the singular values in order to make them all nonzero ([Hong and Yang - 1991], [Cheng, Zhuang and Yang - 1992], [Krzanowski et al. - 1995]). For instance if

$$\mathbf{W} = \mathbf{Q}\,\text{diag}(s_1, \ldots, s_p)\mathbf{Q}^T,$$

and $s_{r+1}, \ldots, s_p = 0$, then for some small ε,

$$\tilde{\mathbf{W}} = \mathbf{Q}\,\text{diag}(s_1, \ldots, s_r, s_{r+1} + \varepsilon, s_p + \varepsilon)\mathbf{Q}^T.$$

Perturbation methods compute an expensive eigenproblem of dimension p. They also ask for determination of suitable perturbation parameters.

3. Fisher's criterion for the $p \gg n$ problem

2. Moore-Penrose methods compute the truncated SVD of **W** ([Hong and Yang - 1991], [Cheng, Zhuang and Yang - 1992], [Krzanowski et al. - 1995], R-environment):

$$\tilde{\mathbf{W}} = \mathbf{Q}_r \,\text{diag}(s_1, \ldots, s_r)\mathbf{Q}_r^T,$$

where $\mathbf{Q}_r$ contains the first r singular vectors.

Transformation to a standard eigenproblem is achieved through multiplication with the Moore-Penrose pseudo-inverse of $\tilde{\mathbf{W}}$.

In comparison with perturbation methods, Moore-Penrose methods avoid determination of parameters (except the truncation parameter) and need only the first r eigenvectors of **W** instead of all p. The Moore-Penrose method is implemented in R-environment by the lda-function.

3. Fisher's criterion for the $p \gg n$ problem

In perturbation and Moore-Penrose methods one solves a modified maximization problem corresponding to $\tilde{\mathbf{W}}$ instead of $\mathbf{W}$. Deterioration from the original problem may be strong (see [DT, Schlesinger - 2006 ?])

Others methods address Fisher's criterion directly [Cheng, Liao, Ko, Lin and Yu - 2000], [Howland, Park et al. - 2003, 2004, 2005]:
They argue that

$$\frac{c^T \mathbf{B} c}{c^T \mathbf{W} c} \tag{3}$$

is maximized for $c \in \text{null}(\mathbf{W})$. Hence the best transformation vectors are to be chosen from $\text{null}(\mathbf{W})$. Indeed, a vector c from $\text{null}(\mathbf{W})$ trivially has minimal within-class covariance

$$c^T \mathbf{W} c = 0.$$

3. Fisher’s criterion for the $p \gg n$ problem

3. The null space method ([Krzanowski et al. - 1995], [Cheng, Liao, Ko, Lin and Yu - 2000]) simply solves the maximization problem

$$\max_{c \in \mathbb{R}^p,\ \mathbf{W}c=0} \quad c^T \mathbf{B} c.$$

The null space of **W** has dimension at least $p - n$, only little less than p. Hence finding the null space is expensive and so is solving the eigenproblem of dimension at least $p - n$.

4. The LDA/GSVD method [Howland, Jeon, Park - 2003] computes

$$\mathbf{B} = \mathbf{C}^{-T} \begin{pmatrix} \mathbf{S}_\alpha & 0 \\ 0 & 0 \end{pmatrix} \mathbf{C}^{-1}, \qquad \mathbf{W} = \mathbf{C}^{-T} \begin{pmatrix} \mathbf{S}_\beta & 0 \\ 0 & 0 \end{pmatrix} \mathbf{C}^{-1},$$

for a nonsingular $\mathbf{C} \in \mathbb{R}^{p \times p}$ and diagonal matrices with nonnegative entries $\mathbf{S}_\alpha =$ diag$(\alpha_1, \ldots, \alpha_t)$ and $\mathbf{S}_\beta =$ diag$(\beta_1, \ldots, \beta_t)$ such that $\mathbf{S}_\alpha + \mathbf{S}_\beta = \mathbf{I}_t$ and $t \leq n + g$.

3. Fisher's criterion for the $p \gg n$ problem

This implies the first t columns c_i of **C** are eigenvectors for (1) and satisfy

$$\beta_i \mathbf{B} c_i = \alpha_i \mathbf{W} c_i .$$

The LDA/GSVD method selects as transformation vectors

- first the c_i with $\beta_i = 0$. They lie in the null space of **W**.
- if necessary further c_i with maximal ratio

$$\alpha_i / \beta_i = \frac{c_i^T \mathbf{B} c_i}{c_i^T \mathbf{W} c_i} .$$

The LDA/GSVD method can be implemented such that computational costs are of order $\mathcal{O}(pn^2) + \mathcal{O}(n^3)$.

3. Fisher’s criterion for the $p \gg n$ problem

5. The ’closest to original criterion’ method (see, e.g. [Yang, Yang - 2003]) combines the transformation criteria of the two previous methods as follows: Choose transformation vectors c_i satisfying

$$\max_{c \in \mathbb{R}^p,\ \mathbf{W}c=0} \quad c^T \mathbf{B} c, \tag{4}$$

with proceeding to the original criterion

$$\max_{c \in \mathbb{R}^p,\ \mathbf{W}c\neq 0} \quad \frac{c^T \mathbf{B} c}{c^T \mathbf{W} c} \tag{5}$$

as soon as the maximum in (4) becomes zero.
This criterion is closest to Fisher’s original criterion for the regular case.
Experiments seem to indicate that optimal linear discriminant analysis leads to the most powerful classification.

At first sight, it seems this method asks for computation of all eigenvectors **W**, i.e. a p-dimensional eigenproblem.

4. Efficient implementation

All mentioned methods must solve full or partial eigenproblems of dimension p. In practice often p is so large this becomes unfeasible (e.g. $p > 10.000$ is very common).

We now list and combine as many advantageous implementation tricks we know of to circumvent too large computations. Some are already used in some methods but can be applied to others, some have been used in other statistical tasks and some are new.

1. Factorization of the covariance matrices
For instance **T** can be written as a product of rectangular matrices:

$$\mathbf{T} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{n-1}(\mathbf{X} - \bar{x}\mathbf{1_n})(\mathbf{X} - \bar{x}\mathbf{1_n})^T.$$

Here, **X** $\in \mathbb{R}^{p\times n}$ is the sample matrix whose ith column contains the ith sample and $\mathbf{1_n} = (1, 1, \ldots, 1)$.

4. Efficient implementation

Similarly, **B** and **W** can be written as products of rectangular matrices $\mathbf{R}_B, \mathbf{R}_W \in \mathbb{R}^{p\times n}$

$$\mathbf{B} = \mathbf{R}_B \mathbf{R}_B^T, \qquad \mathbf{W} = \mathbf{R}_W \mathbf{R}_W^T.$$

Advantages are:

- We need to store only $\mathbf{X} - \bar{x}\mathbf{1_n}, \mathbf{R}_B, \mathbf{R}_W \in \mathbb{R}^{p\times n}$ instead of $\mathbf{T}, \mathbf{B}, \mathbf{W} \in \mathbb{R}^{p\times p}$
- Multiplication with a covariance matrix may be split into 2 cheaper multiplications with its rectangular factors.

This technique is already used in the LDA/GSVD method and in the R-environment implementation of the Moore-Penrose method. The technique can (and should) be used in all other methods, too.

4. Efficient implementation

2. Elimination of the common null space of **B** and **W**

This strategy is justified by the fact that vectors c in the common null space do not contribute to discrimination because $c^T\mathbf{B}c = 0 = c^T\mathbf{W}c$.

As rank(**B**) $\leq g$ and rank(**W**) $\leq n$, the complement of the common null space is of dimension $n + g \ll p$ at most, hence the gain is considerable.

Elimination can be done very efficiently with the following 2 lemma's:

Lemma 1: *The common null space of* **B** *and* **W** *is the null space of* **T**.
Proof: Follows from $(n-1)\mathbf{T} = (g-1)\mathbf{B} + (n-g)\mathbf{W}$. □

Hence elimination of the common null space equals restriction to the eigenspace of **T** corresponding to nonzero eigenvalues. This is nothing but performing a 100% Principal Components Analysis. It is best done by factorization of **T** into rectangular factors and the following lemma:

4. Efficient implementation

Lemma 2: *Let* $\mathbf{Z} \in \mathbb{R}^{n\times p}$ *with* $n < p$, *let the diagonal matrix* $\mathbf{D}_1$ *contain the nonzero eigenvalues of* $\mathbf{Z}\mathbf{Z}^T \in \mathbb{R}^{n\times n}$ *and let the columns of* $\mathbf{V}_1$ *contain the corresponding eigenvectors. Then the normalized eigenvectors for nonzero eigenvalues of* $\mathbf{Z}^T\mathbf{Z} \in \mathbb{R}^{p\times p}$ *are given by the columns of* $\mathbf{Z}^T\mathbf{V}_1\mathbf{D}_1^{-1/2}$.

Proof: See [Johnson, Wichern - 1998]. □

Hence we can extract eigenvectors of the p-dimensional matrix $\mathbf{T} = (\mathbf{X}^T - \bar{x}\mathbf{1_n}^T)(\mathbf{X} - \mathbf{1_n}\bar{x}^T)$ by forming the n-dimensional spectral decomposition of $(\mathbf{X} - \mathbf{1_n}\bar{x}^T)(\mathbf{X}^T - \bar{x}\mathbf{1_n}^T)$ and multiplying the obtained (scaled) eigenvectors with

$$(\mathbf{X}^T - \bar{x}\mathbf{1_n}^T) \in \mathbb{R}^{p\times n}.$$

4. Efficient implementation

Thus p-dimensional computations are restricted to this last multiplication with $(\mathbf{X}^T - \bar{x}\mathbf{1_n}^T)$, which can exploit the fact $\bar{x}\mathbf{1_n}^T$ is rank one, and possible sparsity of **X**.

Another important advantage of elimination of the common null space of **B** and **W** is that it enhances stability of the generalized eigenproblem

$$(\mathbf{B} - \lambda\mathbf{W})c = 0.$$

Elimination of the common null space makes sense in the Moore-Penrose, the LDA/GSVD and the ’closest to original criterion’ method. It is done only in our implementation of the last method (and in case of PCA+LDA strategies).

4. Efficient implementation

3. Efficient computations in the complement of the common null space

We denote the projections of the matrices **B**, **W** and **T** onto the complement of the common null space by $\overline{\mathbf{B}}$, $\overline{\mathbf{W}}$ and $\overline{\mathbf{T}}$, respectively. To solve the projected generalized eigenproblem

$$(\overline{\mathbf{B}} - \lambda \overline{\mathbf{W}})c = 0,$$

we propose to use the simple

Lemma 3: *An eigenvector c for* $\mathbf{Y}c = \mu(\mathbf{Y} + \mathbf{Z})c$ satisfies

$$\mathbf{Y}c = \frac{\mu}{1-\mu}\mathbf{Z}\,c.$$

Hence with $\overline{\mathbf{T}} = \overline{\mathbf{B}} + \overline{\mathbf{W}}$, any eigenvector c with $(\overline{\mathbf{B}} - \mu\overline{\mathbf{T}})c = 0$ is also an eigenvector for

$$(\overline{\mathbf{B}} - \lambda \overline{\mathbf{W}})c = 0, \qquad \lambda = \frac{\mu}{1-\mu}.$$

4. Efficient implementation

We can as well solve

$$(\overline{\mathbf{B}} - \mu\overline{\mathbf{T}})c = 0. \tag{6}$$

if we select the eigenvectors correctly.

Two important advantages are:

- The matrix $\overline{\mathbf{T}}$, as the restriction of **T** to the complement of its own null space, is non-singular. Hence, we can transform (6) to a standard eigenproblem. With $\overline{\mathbf{W}}$ this is in general not possible.
- In addition, as a simple computation shows, the matrix $\overline{\mathbf{T}}$ is diagonal.

Interchanging **W** with **T** has been presented in the literature as a modification of Fisher's criterion (in fact it is not), ([Cheng, Zhuang, Yang - 1992], [Hong, Yang - 1991]).
But the implementational advantages appear only in the complement of the common null space.

4. Efficient implementation

4. Using sparse methods

We are looking for $g - 1 \ll p$ transformation vectors at most. In addition, they correspond to leading eigenvalues.

Hence in all methods, the involved (generalized or standard) eigenproblems are often best solved with a sparse method (Lanczos, Arnoldi, etc...).

Unfortunately, in the literature on classification we didn’t find a word on the usage of sparse methods.

5. Experiments

We implemented all 5 methods from Section 3 with as many of the just mentioned implementation strategies as possible:

- In the perturbation method („Perturb"): Factorization of covariance matrices, sparse method, $\mathcal{O}(p^2n)$ comp. costs
- In the Moore-Penrose method („MP"): Factorization of covariance matrices, elimination of common null space, sparse method, $\mathcal{O}(pn^2)$ comp. costs
- In the null space method („Null space"): Factorization of covariance matrices, sparse method, $\mathcal{O}(p^2n)$ comp. costs
- In the LDA/GSVD method („GSVD"): Factorization of covariance matrices, GSVD (original implementation), $\mathcal{O}(pn^2) + \mathcal{O}(n^3)$ comp. costs
- In the 'closest to original criterion' method („COC"): All implementation strategies, $\mathcal{O}(pn^2) + \mathcal{O}(n^3)$ comp. costs

All methods were implemented in MATLAB.

5. Experiments

First a small example to compare all five methods:

Gene expression profile data: Investigation of DNA microarrays for multiple cancer types diagnosis. It consists of 63 measurements of

$p = 2\,308$ genes belonging to $g = 4$ groups.

We divided the objects by choosing randomly from every group one half as training and one half as test set. This gave a training sample matrix of dimension $32 \times 2\,308$, i.e.

$n = 32$.

The individual methods satisfy Fisher's criterion as follows:

5. Experiments

Dimension	trace($c^T Bc$)				
	Perturb.	MP	GSVD	Nullspace	COC
1	794	183	602	794	794
2	1 405	331	1 148	1 405	1 405
3	1 829	391	1 715	1 829	1 829

Dimension	trace($c^T Wc$)				
	Perturb.	MP	GSVD	Null space	COC
1	0	15	0	0	0
2	0	32	0	0	0
3	0	43	0	0	0

5. Experiments

Dimension	Succesful classification rate				
	Perturb.	MP	GSVD	Null space	COC
1	74.2%	51.6%	51.6%	74.2%	74.2%
2	93.6%	77.4%	96.8%	93.6%	83.9%
3	96.8%	83.9%	96.8%	96.8%	96.8%

Timing (s)				
Perturb.	MP	GSVD	Null space	COC
2.9	0.025	0.024	8.6	0.024

5. Experiments

The second example is much larger:

The MEDLINE data studies the classification of medical documents into $g = 5$ groups. After applying a preprocessing technique we obtain

$p = 22\,095$ distinct terms as explanatory variables. As Perturb. and Null space have storage costs of order $\mathcal{O}(p^2)$ we were not able to apply them!

We use a training set and test set with the same number of examples $n = 1\,250$;

In this example the resulting sample matrix is sparse: The number of non-zeroes of the $1\,250 \times 22\,095$ training sample matrix is $\mathsf{nnz} = 99\,765$. Consequently,

- In MP: $\mathcal{O}(\mathsf{nnz}\, n) + \mathcal{O}(n^3)$ comp. costs
- In GSVD: $\mathcal{O}(pn^2) + \mathcal{O}(n^3)$ comp. costs
- In COC: $\mathcal{O}(\mathsf{nnz}\, n) + \mathcal{O}(n^3)$ comp. costs

5. Experiments

Dimension	trace($c^T Bc$)		
	MP	GSVD	COC
1	0.58	0.53	0.74
2	0.66	0.91	0.91
3	0.70	1.08	1.08
4	0.78	1.12	1.12

Dimension	trace($c^T Wc$)		
	MP	GSVD	COC
1	4.72e-06	0	0
2	1.07e-05	0	0
3	4.13e-04	4.72e-06	4.72e-06
4	7.61e-04	1.06e-05	1.06e-05

5. Experiments

Dimension	Succesful classification rate		
	MP	GSVD	COC
1	41.0%	31.9%	48.6%
2	50.2%	54.6%	55.0%
3	63.4%	74.6%	74.6%
4	86.7%	87.5%	87.5%

Timing (s)			
MP	GSVD	COC	COC, direct method
81	150.5	33	60.5

6. Conclusions

FLDA-based classification in the $p \gg n$ case seems most powerful with the ’closest to original criterion’ method.

The method can be implemented by combining a variety of attractive strategies, enabling among others its application to problems with high-dimensional sparse data matrices.

For more details see ‚Improving Implementation of Linear Discriminant Analysis for the Small Sample Size Problem’, DT, Schlesinger - 2006 ?, submitted to CSDA.

6. Conclusions

Thank you for your attention.

Supported by the project “Information Society” of the Academy of Sciences of the Czech Republic under No. 1ET400300415 and the MSMT CR Project LC536.