Journal of Machine Learning Research 8 (2007) 2121-2123 Submitted 11/06; Revised 5/07; Published 9/07

A Complete Characterization of a Family of Solutions to a Generalized Fisher Criterion

Marco Loog[*] LOOG@DIKU.DK
Datalogical Institute
University of Copenhagen
Universitetsparken 1
DK-2100 Copenhagen Ø, Denmark

Editor: Marina Meila

Abstract

Recently, Ye (2005) suggested yet another optimization criterion for discriminant analysis and proposed a characterization of the family of solutions to this objective. The characterization, however, merely describes a part of the full solution set, that is, it is not *complete* and therefore not at all a characterization. This correspondence first gives the correct characterization and afterwards compares it to Ye's.

Keywords: linear discriminant analysis, Fisher criterion, small sample, characterization

1. Classical and New Criteria

Given N feature vectors of dimensionality n, a linear reduction of dimensionality, based on classical Fisher LDA, determines an $n \times d$ transformation matrix $\mathbf{L}$ that, for a given $d < K$, K the number of classes, maximizes the so-called Fisher criterion: $F(\mathbf{A}) = \mathrm{tr}((\mathbf{A}^t\mathbf{S}_W\mathbf{A})^{-1}(\mathbf{A}^t\mathbf{S}_B\mathbf{A}))$ or, equivalently, $F_0(\mathbf{A}) = \mathrm{tr}((\mathbf{A}^t\mathbf{S}_T\mathbf{A})^{-1}(\mathbf{A}^t\mathbf{S}_B\mathbf{A}))$. Here, $\mathbf{S}_B := \sum_{i=1}^K p_i(\mathbf{m}_i - \overline{\mathbf{m}})(\mathbf{m}_i - \overline{\mathbf{m}})^t$, $\mathbf{S}_W := \sum_{i=1}^K p_i\mathbf{S}_i$, and $\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W$. The matrices $\mathbf{S}_B$, $\mathbf{S}_W$, and $\mathbf{S}_T$ are the so-called between-class, pooled within-class, and total covariance matrices. In addition, $\mathbf{m}_i$ is the mean vector of class i, p_i is the prior of class i, and the overall mean $\overline{\mathbf{m}}$ equals $\sum_{i=1}^k p_i\mathbf{m}_i$. Finally, $\mathbf{S}_i$ is the covariance matrix of class i.

A solution to these optimization problems can be obtained by means of a generalized eigenvalue decomposition, which Fukunaga (1990) relates to a simultaneous diagonalization of the two matrices involved (see also Campbell and Atchley, 1981). More common is it to apply a standard eigenvalue decomposition to $\mathbf{S}_T^{-1}\mathbf{S}_B$ (or $\mathbf{S}_W^{-1}\mathbf{S}_B$), resulting in an equivalent set of eigenvectors. The d columns of the optimal solution $\mathbf{L}$ are simply taken to equal the d eigenvectors corresponding to the d largest eigenvalues. It is known that this solution is not unique and the full class can be obtained by multiplying $\mathbf{L}$ to the right with nonsingular $d \times d$ matrices (see Fukunaga, 1990).

Clearly, if the total covariance $\mathbf{S}_T$ is singular, neither the generalized nor the standard eigenvalue decomposition can be readily employed. Directly or indirectly, the problem is that the matrix inverse $\mathbf{S}_T^{-1}$ does not exist, which is the typical situation when dealing with small samples. In an attempt to overcome this problem, Ye (2005) introduced a different criterion that is defined as

$$F_1(\mathbf{A}) = \mathrm{tr}((\mathbf{A}^t\mathbf{S}_T\mathbf{A})^+(\mathbf{A}^t\mathbf{S}_B\mathbf{A}))\,, \tag{1}$$

∗. Also at Nordic Bioscience Imaging, Hovegade 207, DK-2730 Herlev, Denmark.

where $^+$ denotes taking the Moore-Penrose generalized inverse of a matrix. Like for F_0, an optimal transform $\mathbf{L}$ is one that maximizes the objective F_1. Again, this solution is not unique.

2. Correct Characterization

For the full characterization of the set of solutions to Equation (1), initially the problem is looked at from a geometrical point of view (cf., Campbell and Atchley, 1981). It is assumed that the number of samples N is smaller than or equal to the feature dimensionality n. In the undersampled case, it is clear that all data variation occurs in an $N-1$-dimensional subspace of the original space.

To start with, a PCA of the data is carried out and the first $N-1$ principal components are rotated to the first $N-1$ axes of the n-dimensional space by means of a rotation matrix $\mathbf{R}$. This matrix consists of all (normalized) eigenvectors of $\mathbf{S}_T$ taken as its columns. After this rotation, new total and between-class covariance matrices, $\mathbf{S}'_T = \mathbf{R}^{\mathrm{t}}\mathbf{S}_T\mathbf{R}$ and $\mathbf{S}'_B = \mathbf{R}^{\mathrm{t}}\mathbf{S}_B\mathbf{R}$, are obtained. These matrices can be partitioned as follows: $\mathbf{S}'_T = \left(\begin{smallmatrix}\Sigma_T & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{smallmatrix}\right)$ and $\mathbf{S}'_B = \left(\begin{smallmatrix}\Sigma_B & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{smallmatrix}\right)$, where Σ_T and Σ_B are $N-1\times N-1$ covariance matrices and Σ_T is nonsingular and diagonal by construction. The between-class variation is obviously restricted to the $N-1$-dimensional subspace in which the total data variation takes place, therefore a same partitioning of $\mathbf{S}'_B$ is possible. However, the covariance submatrix Σ_B is not necessarily diagonal, neither does it have to be nonsingular. Basically, the PCA-based rotation $\mathbf{R}$ converts the initial problem into a more convenient one, splitting up the original space in an $N-1$-dimensional one in which "everything interesting" takes place and a remaining $n-N+1$-dimensional subspace in which "nothing happens at all".

Now consider F_1 in this transformed space and take a general $n\times d$ transformation matrix $\mathbf{A}$, which is partitioned in a way similar to the covariance matrices, that is,

$$\mathbf{A} = \begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}. \tag{2}$$

Here, $\mathbf{X}$ is an $N-1\times d$-matrix and $\mathbf{Y}$ is of size $n-N+1\times d$. Taking this definition, the following holds (cf., Ye, 2005):

$$\begin{aligned}
F_1(\mathbf{A}) &= \mathrm{tr}((\mathbf{A}^{\mathrm{t}}\mathbf{S}'_T\mathbf{A})^+(\mathbf{A}^{\mathrm{t}}\mathbf{S}'_B\mathbf{A})) = \mathrm{tr}\left(\left(\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}^{\mathrm{t}}\begin{pmatrix}\Sigma_T & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right)^+\left(\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}^{\mathrm{t}}\begin{pmatrix}\Sigma_B & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{X}\\ \mathbf{Y}\end{pmatrix}\right)\right)\\
&= \mathrm{tr}\left(\begin{pmatrix}\mathbf{X}^{\mathrm{t}}\Sigma_T\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}^+\begin{pmatrix}\mathbf{X}^{\mathrm{t}}\Sigma_B\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right) = \mathrm{tr}\left(\begin{pmatrix}(\mathbf{X}^{\mathrm{t}}\Sigma_T\mathbf{X})^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\begin{pmatrix}\mathbf{X}^{\mathrm{t}}\Sigma_B\mathbf{X} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right)\\
&= \mathrm{tr}((\mathbf{X}^{\mathrm{t}}\Sigma_T\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{t}}\Sigma_B\mathbf{X})) = F_0(\mathbf{X}).
\end{aligned}$$

From this it is immediate that a matrix $\mathbf{A}$ maximizes F_1 if and only if the submatrix $\mathbf{X}$ maximizes the original Fisher criterion in the lower-dimensional subspace. Moreover, if $\mathbf{L}$ is such a matrix maximizing F_1 in the PCA-transformed space, it is easy to check that $\mathbf{R}^{-1}\mathbf{L} = \mathbf{R}^{\mathrm{t}}\mathbf{L}$ provides a solution to the original, general problem that has not been preprocessed by means of a PCA (see also Fukunaga, 1990). A characterization of the complete family of solutions can now be given.

Let $\Lambda\in\mathbb{R}^{N-1\times d}$ be an optimal solution of $F_0(\mathbf{X}) = \mathrm{tr}((\mathbf{X}^{\mathrm{t}}\Sigma_T\mathbf{X})^{-1}(\mathbf{X}^{\mathrm{t}}\Sigma_B\mathbf{X}))$. As already noted in Section 1, the full set of solutions is given by $\mathscr{F} = \{\Lambda\mathbf{Z}\in\mathbb{R}^{N-1\times d}\,|\,\mathbf{Z}\in\mathrm{GL}_d(\mathbb{R})\}$, where $\mathrm{GL}_d(\mathbb{R})$ denotes the general linear group of $d\times d$ invertible matrices. The previous paragraph essentially demonstrates that if $\mathbf{X}\in\mathscr{F}$, $\mathbf{A}$ in Equation (2) maximizes F_1. The matrix $\mathbf{Y}$ can be chosen ad

libitum. Now, the latter provides the solution in the PCA-transformed space and to solve the general problem we need to take the rotation back to the original space into account. All in all, this leads to the following complete family of solutions $\mathscr{L}$, maximizing F_1 in the original space:

$$\mathscr{L} = \left\{ \mathbf{R}^{\mathrm{t}} \begin{pmatrix} \Lambda \mathbf{Z} \\ \mathbf{Y} \end{pmatrix} \in \mathbb{R}^{n \times d} \,\middle|\, \mathbf{Z} \in \mathrm{GL}_d(\mathbb{R}), \mathbf{Y} \in \mathbb{R}^{n-N+1 \times d} \right\}, \tag{3}$$

where $\Lambda = \mathrm{argmax}_{\mathbf{X}} \mathrm{tr}((\mathbf{X}^{\mathrm{t}} \Sigma_T \mathbf{X})^{-1} (\mathbf{X}^{\mathrm{t}} \Sigma_B \mathbf{X}))$ and $\mathbf{R}^{\mathrm{t}}$ takes care of the rotation back.

3. Original Characterization

Though not noted by Ye (2005), his attempt to characterize the full set of solutions of Equation (1) is based on a simultaneous diagonalization of the three covariance matrices $\mathbf{S}_B$, $\mathbf{S}_W$, and $\mathbf{S}_T$ that is similar to the ideas described by Campbell and Atchley (1981) and Fukunaga (1990). Moreover, Golub and Van Loan (Theorem 8.7.1. 1996) can be readily applied to demonstrate that such simultaneous diagonalization is possible in the small sample setting. After the diagonalization step, partitioned between-class, pooled within-class, and total covariance matrices are considered. This partitioning is similar to the one employed in the previous section, which does not enforce matrices to be diagonal however.

In the subsequent optimization step, the classical Fisher criterion is maximized basically in the appropriate subspace, comparable to the approach described above, but in a mildly more involved and concealed way. For this, matrices of the form $\mathbf{R}^{\mathrm{t}} \left(\begin{smallmatrix} \mathbf{X} \\ \mathbf{Y} \end{smallmatrix}\right)$ are considered, consider Equations (2) and (3). However, $\mathbf{Y}$ is simply the null matrix and the family of solutions $\mathscr{L}'$ provided is limited to

$$\mathscr{L}' = \left\{ \mathbf{R}^{\mathrm{t}} \begin{pmatrix} \Lambda \mathbf{Z} \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n \times d} \,\middle|\, \mathbf{Z} \in \mathrm{GL}_d(\mathbb{R}) \right\}.$$

Obviously, this is far from a complete characterization, especially when $N - 1 \ll n$ which is, for instance, typically the case for the data sets considered by Ye (2005).

Generally, the utility of a dimensionality reduction criterion, without additional constrains, depends on the efficiency over the full set of solutions. As Ye (2005) only considers two very specific instances from the large class of possibilities, it is unclear to what extent the new criterion really provides an efficient way of performing a reduction of dimensionality.

References

N. A. Campbell and W. R. Atchley. The geometry of canonical variate analysis. *Systematic Zoology*, 30(3):268–280, 1981.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990.

G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, third edition, 1996.

J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.