Classification of high dimensional data: High Dimensional Discriminant Analysis

Charles Bouveyron, Stephane Girard, Cordelia Schmid

Classification of high dimensional data: High Dimensional Discriminant Analysis *

Charles Bouveyron[1,2], Stéphane Girard[1], and Cordelia Schmid[2]

[1] LMC – IMAG, BP 53, Université Grenoble 1,
38041 Grenoble cedex 9 – France
charles.bouveyron@imag.fr, stephane.girard@imag.fr
[2] LEAR – INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot,
38334 Saint-Ismier Cedex – France
Cordelia.Schmid@inrialpes.fr

Abstract. We propose a new method of discriminant analysis, called High Dimensional Discriminant Analysis (HHDA). Our approach is based on the assumption that high dimensional data live in different subspaces with low dimensionality. Thus, HDDA reduces the dimension for each class independently and regularizes class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. This regularization is achieved by assuming that classes are spherical in their eigenspace. HDDA is applied to recognize objects in real images and its performances are compared to classical classification methods.

Keywords: Discriminant analysis, dimension reduction, regularization.

1 Introduction

Over the past few years, statistical learning has become a specific discipline. Indeed, many scientific domains need to analyze data which are increasingly complex. For example, medical research, financial analysis and computer vision provide very high dimensional. Classifying such data is a very challenging problem. In high dimensional feature spaces, the performances of learning methods suffer from the curse of dimensionality, which degrades both classification accuracy and efficiency. To address this issue, we present in this paper a new method of discriminant analysis, called High Dimensional Discriminant analysis (HDDA), which classifies high dimensional data. Our method assumes that high dimensional data live in different subspaces with low dimensionality. Thus, HDDA reduces the dimension for each class independently and regularizes class conditional covariance matrices in order to adapt the Gaussian framework to high dimensional data. This regularization is based on the assumption that classes are spherical in their eigenspace. It is also possible to make additional assumptions to reduce the number of parameters to estimate. This paper is organized as

* This paper was supported by the French department of Research through the *ACI Masse de données* (MoViStaR project).

follows. Section 2 presents the discrimination problem and existing regularized discriminant analysis methods. Section 3 introduces the theoretical framework of HDDA. Section 4 is devoted to the inference aspects. Our method is then compared to reference methods on a real images dataset in section 5.

2 Discriminant analysis framework

In this section, we remind the general framework of the discrimination problem and present the main regularization methods of discriminant analysis.

2.1 Discrimination problem

The goal of discriminant analysis is to assign an observation $x \in \mathbb{R}^p$ with unknown class membership to one of k classes $C_1, ..., C_k$ known *a priori*. To this end, we have a learning dataset $A = \{(x_1, c_1), ..., (x_n, c_n)/x_j \in \mathbb{R}^p \text{ et } c_j \in \{1, ..., k\}\}$, where the vector x_j contains p explanatory variables and c_j indicates the index of the class of x_i. The optimal decision rule, called *Bayes decision rule*, affects the observation x to the class C_{i^*} which has the *maximum a posteriori* probability which is equivalent, in view of the Bayes formula, to minimize a cost function $K_i(x)$ *i.e.* $i^* = \text{argmin}_{i=1,...,k} K_i(x)$, with $K_i(x) = -2\log(\pi_i f_i(x))$, where π_i is the *a priori* probability of class C_i and $f_i(x)$ denotes the class conditional density of x, $\forall i = 1, ..., k$.

2.2 Dimension reduction and regularization

Classical discriminant analysis methods (QDA and LDA) have disappointing behavior when the size n of the training dataset is small compared to the number p of variables. In such cases, a dimension reduction step and/or a regularization of the discriminant analysis are introduced.

Fisher discriminant analysis (FDA) This approach combines a dimension reduction step and a discriminant analysis procedure and is in general efficient on high dimensional data. FDA provides the $(k-1)$ discriminant axes maximizing the ratio between the inter class variance and the intra class variance. It is then possible to perform one of the previous methods on the projected data (usually LDA).

Regularized discriminant analysis (RDA) In [4] a regularization technique of discriminant analysis is proposed. RDA uses two regularization parameters to design an intermediate classifier between LDA and QDA. The estimation of the covariance matrices depends on a complexity parameter and on a shrinkage parameter. The complexity parameter controls the ratio between Σ_i and the common covariance matrix Σ. The other parameter controls shrinkage of the class conditional covariance matrix toward a specified multiple of the identity matrix.

Eigenvalue decomposition discriminant analysis (EDDA) This other regularization method [1] is based on the re-parametrization of the covariance matrices: $\Sigma_i = \lambda_i D_i A_i D_i^t$, where D_i is the matrix of eigenvectors of Σ_i, A_i is a diagonal matrix containing standardized and ordered eigenvalues of Σ_i and $\lambda_i = |\Sigma_i|^{1/p}$. Parameters λ_i, D_i and A_i respectively control the volume, the orientation and the shape of the density contours of class C_i. By allowing some but not all of these quantities to vary, the authors obtain geometrical interpreted discriminant models including QDA, QDAs, LDA and LDAs.

3 High Dimensional Discriminant Analysis

The *empty space* phenomena enables us to assume that high-dimensional data live in subspaces with dimensionality lower than p. In order to adapt discriminant analysis to high dimensional data and to limit the number of parameters to estimate, we propose to work in class subspaces with lower dimensionality. In addition, we assume that classes are spherical in these subspaces, *i.e.* class conditional covariance matrices have only two different eigenvalues.

3.1 Definitions and assumptions

Similarly to classical discriminant analysis, we assume that class conditional densities are Gaussian $\mathcal{N}(\mu_i, \Sigma_i)\ \forall i = 1, ..., k$. Let Q_i be the orthogonal matrix of eigenvectors of the covariance matrix Σ_i and $\mathcal{B}_i$ be the eigenspace of Σ_i, *i.e.* the basis made of eigenvectors of Σ_i. The class conditional covariance matrix Δ_i is defined in the basis $\mathcal{B}_i$ by $\Delta_i = Q_i^t\, \Sigma_i\, Q_i$. Thus, Δ_i is diagonal and made of eigenvalues of Σ_i. We assume in addition that Δ_i has only two different eigenvalues $a_i > b_i$. Let $\mathbb{E}_i$ be the affine space generated by the eigenvectors associated to the eigenvalue a_i with $\mu_i \in \mathbb{E}_i$, and let $\mathbb{E}_i^\perp$ be $\mathbb{E}_i \oplus \mathbb{E}_i^\perp = \mathbb{R}^p$ with $\mu_i \in \mathbb{E}_i^\perp$. Thus, the class C_i is both spherical in $\mathbb{E}_i$ and in $\mathbb{E}_i^\perp$. Let $P_i(x) = \tilde{Q}_i \tilde{Q}_i^{\ t}(x - \mu_i) + \mu_i$ be the projection of x on $\mathbb{E}_i$, where $\tilde{Q}_i$ is made of the d_i first raws of Q_i and supplemented by zeros. Similarly, let $P_i^\perp(x) = (Q_i - \tilde{Q}_i)(Q_i - \tilde{Q}_i)^t(x - \mu_i) + \mu_i$ be the projection of x on $\mathbb{E}_i^\perp$.

3.2 Decision rule

The preceding assumptions lead to the cost function (*cf.* [2] for the proof):

$$K_i(x) = \frac{\|\mu_i - P_i(x)\|^2}{a_i} + \frac{\|x - P_i(x)\|^2}{b_i} + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i).$$

In order to interpret the decision rule the following notations are needed: $\forall i = 1, ..., k$, $a_i = \frac{\sigma_i^2}{\alpha_i}$ and $b_i = \frac{\sigma_i^2}{(1-\alpha_i)}$ with $\alpha_i \in]0, 1[$ and $\sigma_i > 0$. The cost function can be rewritten:

$$\begin{aligned} K_i(x) = \frac{1}{\sigma_i^2}\left(\alpha_i \|\mu_i - P_i(x)\|^2 + (1 - \alpha_i)\|x - P_i(x)\|^2\right) \\ + 2p\log(\sigma_i) + d_i \log\left(\frac{1 - \alpha_i}{\alpha_i}\right) - p\log(1 - \alpha_i) - 2\log(\pi_i). \end{aligned}$$

The Bayes formula allows to compute the classification error risk based on the *a posteriori* probability

$$p(C_i|x) = \exp\left(-\frac{1}{2}K_i(x)\right) \Bigg/ \sum_{j=1}^{k} \exp\left(-\frac{1}{2}K_j(x)\right).$$

Note that particular cases of HDDA reduce to classical discriminant analysis. If $\forall i = 1, ..., k,\ \alpha_i = 1/2$: HDDA reduces to QDAs. If moreover $\forall i = 1, ..., k$, $\sigma_i = \sigma$: HDDA reduces to LDAs.

3.3 Particular rules

By allowing some but not all of HDDA parameters to vary between classes, we obtain 23 particular rules which are easily geometrically interpretable and correspond to different types of regularization [2]. Due to space restrictions, we present only two methods: HDDAi and HDDAh.

Isometric decision rule (HDDAi) The following additional assumptions are made: $\forall i = 1, ..., k,\ \alpha_i = \alpha,\ \sigma_i = \sigma,\ d_i = d$ and $\pi_i = \pi_*$, leading to the cost function

$$K_i(x) = \alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2.$$

Case $\alpha = 0$: HDDAi affects x to the class C_{i^*} if $\forall i = 1, ..., k,\ d(x, \mathbb{E}_{i^*}) < d(x, \mathbb{E}_i)$. From a geometrical point of view, the decision rule affects x to the class associated to the closest subspace $\mathbb{E}_i$.
Case $\alpha = 1$: HDDAi affects x to the class C_{i^*} if $\forall i = 1, ..., k,\ d(\mu_{i^*}, P_{i^*}(x)) < d(\mu_i, P_i(x))$. It means that the decision rule affects x to the class for which the mean is closest to the projection of x on the subspace.
Case $0 < \alpha < 1$: the decision rule affects x to the class realizing a compromise between the two previous cases. The estimation of α is discussed in the following section.

Homothetic decision rule (HDDAh) This method differs from the previous one by removing the constraint $\sigma_i = \sigma$. The corresponding cost function is:

$$K_i(x) = \frac{1}{\sigma_i^2}(\alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2) + 2p\log(\sigma_i)$$

It generally favours classes with large variance. Indeed, if the point x is equidistant to two classes, it is natural to affect x to the class with the larger variance.

Removing constraints on d_i and π_i The two previous methods assume that d_i and π_i are fixed. However, these assumptions can be too restrictive. If these constraints are removed, it is necessary to add the corresponding terms in $K_i(x)$: if d_i are free, then add $d_i \log(\frac{1-\alpha}{\alpha})$ and if π_i are free, then add $-2\log(\pi_i)$.

4 Estimators

The methods HDDA, HDDAi and HDDAh require the estimation of some parameters. These estimators are computed through maximum likelihood (ML) estimation based on the learning dataset A. In the following, the *a priori* probability π_i of the class C_i is estimated by $\hat{\pi}_i = n_i/n$, where $n_i = card(C_i)$ and the class covariance matrix Σ_i is estimated by $\hat{\Sigma}_i = \frac{1}{n_i}\sum_{x_j \in C_i}(x_j - \hat{\mu}_i)^t(x_j - \hat{\mu}_i)$ where $\hat{\mu}_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j$.

4.1 HDDA estimators

Starting from the log-likelihood expression found in [3, eq. (2.5)], and assuming for the moment that the d_i are known, we obtain the following ML estimates:

$$\hat{a}_i = \frac{1}{d_i}\sum_{j=1}^{d_i}\lambda_{ij} \quad \text{and } \hat{b}_i = \frac{1}{(p-d_i)}\sum_{j=d_i+1}^{p}\lambda_{ij},$$

where $\lambda_{i1} \geq \ldots \geq \lambda_{ip}$ are the eigenvalues of $\hat{\Sigma}_i$. Moreover, the jth column of Q_i is estimated by the unit eigenvector of $\hat{\Sigma}_i$ associated to the eigenvalue λ_{ij}. Note that parameters a_i and b_i are estimated by the empirical variances of C_i respectively in $\hat{\mathbb{E}}_i$ and in $\hat{\mathbb{E}}_i^{\perp}$. The previous result allows to deduce the maximum likelihood estimators of α_i and σ_i^2:

$$\hat{\alpha}_i = \hat{b}_i/(\hat{a}_i + \hat{b}_i) \quad \text{and } \hat{\sigma}_i^2 = \hat{a}_i\hat{b}_i/(\hat{a}_i + \hat{b}_i).$$

4.2 Estimation of the intrinsic dimension

Estimation of the dataset intrinsic dimension is a difficult problem which does not have an explicit solution. Our approach is based on the eigenvalues of the class conditional covariance matrix Σ_i. The jth eigenvalue of Σ_i corresponds to the fraction of the full variance carried by the jth eigenvector of Σ_i. Consequently, we propose to estimate dimensions d_i, $i = 1, ..., k$, by a common thresholding on the cumulative class conditional variance:

$$\hat{d}_i = \underset{d=1,\ldots,p-1}{\operatorname{argmin}} \left\{ \sum_{j=1}^{d}\lambda_{ij} \Big/ \sum_{j=1}^{p}\lambda_{ij} \geq s \right\},$$

where $s \in]0, 1[$ is the threshold determined by maximizing the correct classification rate on the learning dataset A.

4.3 Particular rule estimators

Among the 23 particular rules, 8 benefit from explicit ML estimators (see [2]). The computation of the ML estimates associated to the 15 other particular rules requires iterative algorithms. We do not reproduce them here by lack of space.

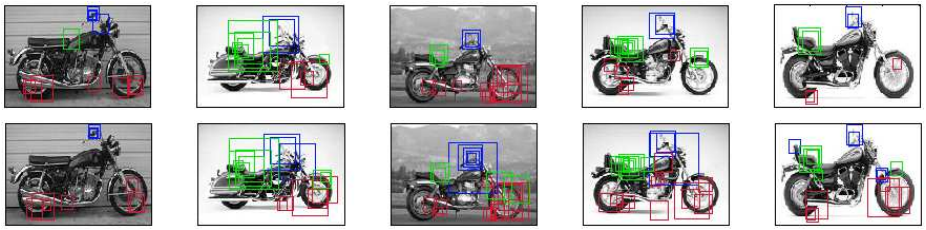

Fig. 1. Recognition of the class "motorbike" using HDDA (top) and SVM (bottom). The colors blue, red and green are respectively associated to handlebars, wheels and seat.

5 Results

Object recognition is one of the most challenging problems in computer vision. Many successful object recognition approaches use local images descriptors. However, local descriptors are high-dimensional and this penalizes classification methods and consequently recognition. Thus, HDDA seems well adapted to this problem. For our experiments, small scale-invariant regions are detected on the training set and they are characterized by the local descriptor SIFT [5]. The object is recognized in a test image if a sufficient number of matches with the training set is found. Recognition uses supervised classification methods like LDA or SVM. Figure 1 presents recognition results obtained for 5 motorbike images. These results show that HDDA ($s = 0.78$) combined with error probability thresholding (see [2]) gives better recognition results than SVM. Indeed, the classification errors are significantly lower for HDDA compared to SVM. For example, on the 5th image, HDDA recognizes the motorbike parts without error whereas SVM makes five errors.

A natural extension to this work is to use the statistical model of HDDA to adapt the method to the context of unsupervised classification.

References

1. H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996.
2. C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. Technical Report 5470, INRIA, January 2005.
3. B. W. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–897, 1984.
4. J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.
5. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.