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
(e-mail: charles.bouveyron@imag.fr, stephane.girard@imag.fr)

[2] LEAR – INRIA Rhône-Alpes, 655 avenue de l'Europe, Montbonnot,
38334 Saint-Ismier Cedex – France (e-mail: 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 object in real images and its performances are compared to classical classification methods.

Keywords: Discriminant analysis, Dimension reduction, Regularization.

1 Introduction

In this paper, we introduce a new method of discriminant analysis, called High Dimensional Discriminant analysis (HDDA) to classify high dimensional data, as occur for example in visual object recognition. We assume 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 follows. We first remind in section 2 the discrimination problem and classical discriminant analysis methods. Section 3 presents 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 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 $A = \{(x_1, c_1), ..., (x_n, c_n)/x_j \in \mathbb{R}^p$ and $c_j \in \{1, ..., k\}\}$, where the vector x_j contains p explanatory variables and c_j indicates the index of the class of x_i. It is a statistical decision problem and the learning dataset allows to construct a decision rule which associates a new vector $x \in \mathbb{R}^p$ to one of the k classes. 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.* $i^* = \operatorname{argmin}_{i=1,...,k} K_i(x)$, 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 Classical discriminant analysis methods

Some classical discriminant analysis methods can be obtained by combining additional assumptions with the Bayes decision rule. We refer to [Celeux, 2003] and [Saporta, 1990, chap. 18] for further informations on this topic. For instance, Quadratic discriminant analysis (QDA) assumes that, $\forall i = 1, ..., k$, the class conditional density f_i for the class C_i is Gaussian $\mathcal{N}(\mu_i, \varSigma_i)$ which leads to the cost function

$$K_i(x) = (x - \mu_i)^t \varSigma_i^{-1}(x - \mu_i) + \log(\det \varSigma_i) - 2\log(\pi_i).$$

This decision rule makes quadratic separations between the classes. In practice, this method is penalized in high-dimensional spaces since it requires the estimation of many parameters. For this reason, particular rules of QDA exist in order to regularize the estimation of $\varSigma_i$. As an example, it can be assumed that covariance matrices are proportional to the identity matrix, *i.e.* $\varSigma_i = \sigma_i^2 Id$. In this case, classes are spherical and this method is referred to as QDAs. One can also assume that covariance matrices are equal, *i.e.* $\varSigma_i = \varSigma$, which yields the framework of the linear discriminant analysis (LDA). This method makes linear separations between the classes. If, in addition, covariance matrices are assumed equal and proportional to the identity matrix, we obtain the so-called LDAs method.

2.3 Dimension reduction and regularization

Classical discriminant analysis methods 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 [Friedman, 1989] 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 [Bensmail and Celeux, 1996] 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* [Scott and Thompson, 1983] 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, in other words 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:

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

(*cf.* [Bouveyron *et al.*, 2005] for the proof). 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

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

Note that some 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 24 particular models which some ones have easily geometrically interpretable rules and correspond to different types of regularization (see [Bouveyron *et al.*, 2005]). 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 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 [Flury, 1984, eq. (2.5)], and assuming for the moment that the d_i are known, we obtain the following ML estimates:

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

where $\lambda_{i1} \geq \cdots \geq \lambda_{ip}$ 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:

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

4.2 Estimation of the intrinsic dimension

Estimation of the dataset intrinsic dimension is a difficult problem which we can find for example in the choice of the factor number in PCA. 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 the empirical method of the scree-test of Cattell [Cattell, 1966] which analyses the differences between eigenvalues in order to find a break in the scree. The selected dimension is the dimension for which the following differences are very small compared to the maximum of differences.

4.3 Particular model estimators

Among the 24 particular models, 9 benefit from explicit ML estimators (see [Bouveyron *et al.*, 2005]). 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.

5 Application to object recognition

Object recognition is one of the most challenging problems in computer vision. In the last few years, many successful object recognition approaches use local images descriptors. However, local descriptors are high-dimensional and this penalizes classification methods and consequently recognition. For this reason, HDDA seems well adapted to this application. In the following, we show that HDDA outperform existing techniques in this context.

5.1 Framework of the object recognition

In our framework, small scale-invariant regions are detected on a learning image set and they are then characterized by the local descriptor SIFT [Lowe, 2004]. The object is recognized in a test image if a sufficient number of matches with the learning set is found. The recognition step is done using supervised classification methods. Frequently used methods are LDA and, more recently, kernel methods (SVM) [Hastie *et al.*, 2001, chap. 12]. In our approach, the object is represented as a set of object parts. For the motorbike, we consider three parts: wheels, seat and handlebars.

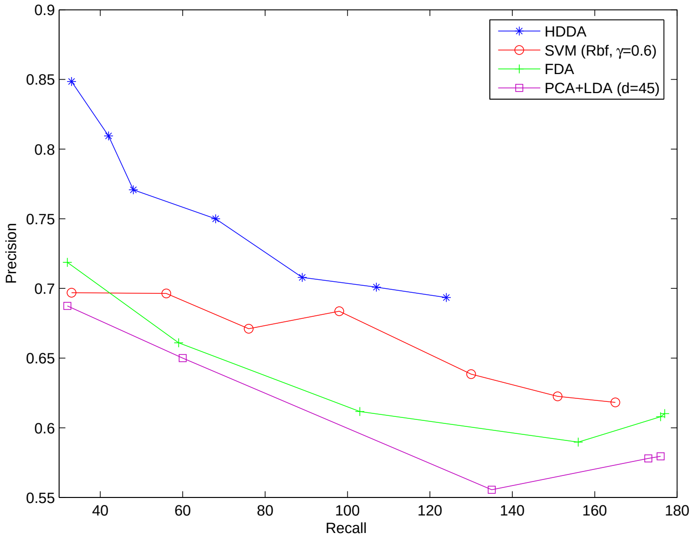

Fig. 1. Comparison of classification results between HDDA method and reference methods.

5.2 Data and protocol

Sift descriptors are computed on 200 motorbike images and 1000 descriptors of motorbike features and of the background were preserved. Consequently, the dataset is made of descriptors in 128 dimensions divided into 4 classes: wheels, seat, handlebars and background. The learning and test dataset are respectively made of 500 and 500 descriptors. Class proportions are respectively: $\forall i = 1, ..., 3$, $\pi_i = 1/6$ and $\pi_4 = 1/2$.

5.3 Results

Figure 1 presents classification results obtained on test data. In order to synthesize the results, only two classes were considered to plot recall-precision curve: motorbike (positive) and background (negative). We remind that the *precision* is the ratio between the number of true positives and the number of detected positives, and the *recall* is the number of detected positives. The different values for each method corresponds to different classifiers. For SVM, the parameter γ is fixed to the best value (0.6) while the parameter C varies. For the other methods, the decision rule varies according to the *a posteriori* probability. In addition, for LDA, we reduced the dimension of data to 45 using PCA in order to obtain the best results for this method. It appears that HDDA outperforms the other methods. In addition, HDDA method

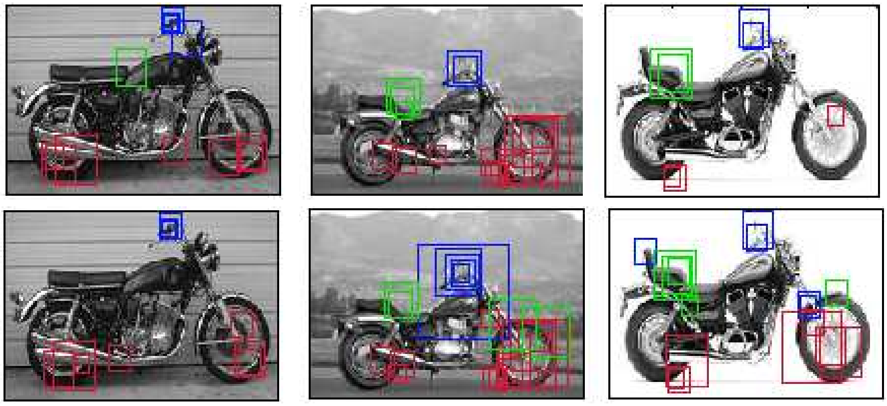

Fig. 2. Recognition of the class "motorbike" using HDDA (top) and SVM (bottom) classifiers. Only descriptors classified as motorbike are displayed. The colors blue, red and green are respectively associated to handlebars, wheels and seat.

is as fast as classical discriminant analysis (computation time $\simeq$ 1 sec. for 1000 descriptors) and much faster than SVM ($\simeq$ 7 sec.). Figure 2 presents recognition results obtained on 5 motorbike images. These results show that HDDA gives better recognition results than SVM. Indeed, the classification errors are significantly lower for HDDA compared to SVM. For example, on the 3th image, HDDA recognizes the motorbike parts without error whereas SVM makes five errors.

6 Conclusion and further work

We presented in this paper a new generative model to classify high-dimensional data in the Gaussian framework. This new model estimates the intrinsic dimension of each class and uses this information to reduce the number of parameters to estimate. In addition, classes are assumed spherical in both subspaces in order to reduce again the number of parameters to estimate and to obtain easily geometrically interpretable rules. In the supervised framework, this model gives very good results without dimension reduction of the data and with a small learning set. Another advantage of this generative model is that it can be used either in supervised or in unsupervised classification. In unsupervised classification, the model presented here arises to a new clustering method based on the EM algorithm. In addition, it is possible to combine unsupervised and supervised classification to recognize an object in a natural image without human interaction. Indeed, the clustering method associated to our model can be used to learn automatically the discriminant part of the object, and then HDDA can be used to recognize the object on a new natural image. First results obtained using this approach are very promising.

Acknowledgments

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

References

[Bensmail and Celeux, 1996]H. Bensmail and G. Celeux. Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996.

[Bouveyron *et al.*, 2005]C. Bouveyron, S. Girard, and C. Schmid. High dimensional discriminant analysis. Technical Report 5470, INRIA, January 2005.

[Cattell, 1966]R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):140–161, 1966.

[Celeux, 2003]G. Celeux. Analyse discriminante. In G. Govaert, editor, *Analyse de Données*, pages 201–233. Hermes Science, Paris, France, 2003.

[Flury, 1984]B. W. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79:892–897, 1984.

[Friedman, 1989]J.H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.

[Hastie *et al.*, 2001]T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2001.

[Lowe, 2004]D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Saporta, 1990]G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, Paris, France, 1990.

[Scott and Thompson, 1983]D. Scott and J. Thompson. Probability density estimation in higher dimensions. In *Proceedings of the Fifteenth Symposium on the Interface, North Holland-Elsevier Science Publishers*, pages 173–179, 1983.