Class-Specific Subspace Discriminant Analysis for High-Dimensional Data

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 for discriminant analysis, called High Dimensional Discriminant Analysis (HDDA). Our approach is based on the assumption that high dimensional data live in different subspaces with low dimensionality. We therefore propose a new parameterization of the Gaussian model to classify high-dimensional data. This parameterization takes into account the specific subspace and the intrinsic dimension of each class to limit the number of parameters to estimate. HDDA is applied to recognize object parts in real images and its performance is compared to classical methods.

Keywords: Discriminant analysis, class-specific subspaces, dimension reduction, regularization.

1 Introduction

Many scientific domains need to analyze data which are increasingly complex. For example, visual descriptors used in object recognition are often high-dimensional and this penalizes classification methods and consequently recognition. In high-dimensional feature spaces, the performance of learning methods suffers from the *curse of dimensionality* [1] which deteriorates both classification accuracy and efficiency. Popular classification methods are based on a Gaussian model and show a disappointing behavior when the size of the training dataset is too small compared to the number of parameters to estimate. To avoid overfitting, it is therefore necessary to find a balance between the number of parameters to estimate and the generality of the model. In this paper we propose a Gaussian model which determines the specific subspace in which each class is located and therefore limits the number of parameters to estimate. The maximum likelihood method is used for parameter estimation and the intrinsic dimension of each class is determined automatically with the scree-test of Cattell. This allows to derive a robust discriminant analysis method in high-dimensional spaces, called High Dimensional Discriminant Analysis (HDDA). It is possible to make additional assumptions on the model to further limit the number of parameters. We can

C. Saunders et al. (Eds.): SLSFS 2005, LNCS 3940, pp. 139–150, 2006.

assume that classes are spherical in their subspaces and it is possible to fix some parameters to be common between classes. A comparison with standard discriminant analysis methods on a recently proposed dataset [4] shows that HDDA outperforms them.

This paper is organized as follows. Section 2 presents the discrimination problem and existing methods to regularize discriminant analysis in high-dimensional spaces. Section 3 introduces the theoretical framework of HDDA and, in section 4, some particular cases are studied. Section 5 is devoted to the inference aspects. Our method is then compared to classical methods on a real image dataset in section 6.

2 Discriminant Analysis Framework

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*. For this, 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_j. The optimal decision rule, called *Bayes decision rule*, assigns the observation x to the class C_{i^*} which has the *maximum a posteriori* probability. This is equivalent 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, $\forall i = 1, ..., k$. For instance, assuming that $f_i(x)$ is a Gaussian density leads to the well known Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) methods.

2.2 Dimension Reduction and Parsimonious Models

In high dimensional spaces, the majority of classification methods shows a disappointing behavior when the size of the training dataset is too small compared to the number of parameters to estimate. To avoid overfitting, it is therefore necessary to reduce the number of parameters. This is possible by either reducing the dimension of the data or by using a parsimonious model with additional assumptions on the model.

Dimension reduction. Many methods use global dimension reduction techniques to overcome problems due to high-dimensionality. A widely used solution is to reduce the dimensionality of the data before using a classical discriminant analysis method. The dimension reduction can be done using Principal Components Analysis (PCA) or a feature selection technique. It is also possible to reduce the data dimension with classification as a goal by using Fisher Discriminant Analysis (FDA) which projects the data on the $(k-1)$ discriminant axes and then classifies the projected data. The dimension reduction is often advantageous in terms of performance but loses information which could be discriminant due to the fact that most approaches are global and not designed for classification.

Parsimonious models. Another solution is to use a model which requires the estimation of fewer parameters. The parsimonious models used most often involve an identical covariance matrix for all classes (used in LDA), *i.e.,* $\forall i,\ \Sigma_i = \Sigma$, or a diagonal covariance matrix, *i.e.,* $\Sigma_i = \mathrm{diag}(\sigma_{i1}, ..., \sigma_{ip})$. Other approaches propose new parameterizations of the Gaussian model in order to find different parsimonious models. For example, Regularized Discriminant Analysis [6] uses two regularization parameters to design an intermediate classifier between QDA and LDA. The Eigenvalue Decomposition Discriminant Analysis [2] proposes to re-parameterize the covariance matrices of the classes in their eigenspace. These methods do not allow to efficiently solve the problem of the high-dimensionality, as they do not determine the specific subspaces in which the data of each class live.

3 High Dimensional Discriminant Analysis

The *empty space* phenomenon [9] allows us to assume that high-dimensional data live in different low-dimensional subspaces hidden in the original space. We therefore propose in this section a new parameterization of the Gaussian model which combines a local subspace approach and a parsimonious model.

3.1 The Gaussian Mixture Model

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 of eigenvectors of Σ_i. The class conditional covariance matrix Δ_i is then 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 $a_i > b_i$:

$$\Delta_i = \begin{pmatrix} \boxed{\begin{matrix} a_i & & 0 \\ & \ddots & \\ 0 & & a_i \end{matrix}} & & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} d_i \\ & \boxed{\begin{matrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{matrix}} & \left.\begin{matrix} \\ \\ \\ \end{matrix}\right\} (p - d_i) \end{pmatrix}$$

Let $\mathbb{E}_i$ be the affine space generated by the eigenvectors associated with 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 columns 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 $\mathbb{E}_i^\perp$. Figure 1 summarizes these notations.

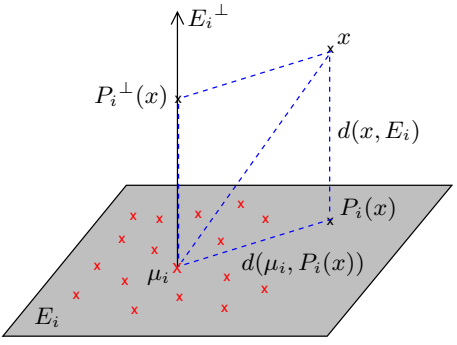


Fig. 1. The subspaces $\mathbb{E}_i$ and $\mathbb{E}_i^{\perp}$ of the class C_i

3.2 Decision Rule and *a Posteriori* Probability

Deriving the Bayes decision rule with the model described in the previous section yields the decision rule of High Dimensional Discriminant Analysis (HDDA).

Theorem 1. *Bayes decision rule yields the decision rule* δ^+ *which classifies* x *as the class* C_{i^*} *such that* $i^* = \text{argmin}_{i=1,\ldots,k}\{K_i(x)\}$ *where* K_i *is defined by:*

$$K_i(x) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i).$$

Proof. We derive Bayes decision rule for the Gaussian model presented in section 3.1. Writing f_i with the class conditional covariance matrix Δ_i gives:

$$-2\log(f_i(x)) = (x - \mu_i)^t (Q_i \Delta_i Q_i^t)^{-1}(x - \mu_i) + \log(\det \Delta_i) + p\log(2\pi).$$

Moreover, $Q_i^t Q_i = Id$ and consequently:

$$-2\log(f_i(x)) = \left[Q_i^t(x - \mu_i)\right]^t \Delta_i^{-1} \left[Q_i^t(x - \mu_i)\right] + \log(\det \Delta_i) + p\log(2\pi).$$

Given the structure of Δ_i, we obtain:

$$-2\log(f_i(x)) = \frac{1}{a_i}\|\tilde{Q_i}^t(x - \mu_i)\|^2 + \frac{1}{b_i}\|(Q_i - \tilde{Q_i})^t(x - \mu_i)\|^2 + \log(\det \Delta_i) + p\log(2\pi).$$

Using definitions of P_i and $P_i^{\perp}$ and in view of figure 1, we obtain:

$$-2\log(f_i(x)) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + \log(\det \Delta_i) + p\log(2\pi).$$

The relation $\log(\det \Delta_i) = d_i \log(a_i) + (p - d_i)\log(b_i)$ concludes the proof. □

The *a posteriori* probability $P(x \in C_i|x)$ measures the probability that x belongs to C_i and allows to identify dubiously classified points. Basing on Bayes' formula, we can write: $P(x \in C_i|x) = 1/\sum_{l=1}^{k} \exp\left(\frac{1}{2}(K_i(x) - K_l(x))\right)$.

4 Particular Rules

By allowing some of the HDDA parameters to be common between classes, we obtain particular rules which correspond to different types of regularization, some of which are easily geometrically interpretable. Due to space restrictions, we present only the two most important particular cases: HDDAi and HDDAh. In order to interpret these particular decision rules, the following notations are useful: $\forall i = 1, ..., k$, $a_i = \frac{\sigma_i^2}{\alpha_i}$ and $b_i = \frac{\sigma_i^2}{(1-\alpha_i)}$ with $\alpha_i \in]0, 1[$ and $\sigma_i > 0$.

4.1 Isometric Decision Rule (HDDAi)

Here, the following additional assumptions are made: $\forall i = 1, ..., k$, $\alpha_i = \alpha$, $\sigma_i = \sigma$, $d_i = d$ and $\pi_i = \pi_*$. In this case, the classes are isometric.

Proposition 1. *Under these assumptions, the decision rule classifies x as the class C_{i^*} such that $i^* = \text{argmin}_{i=1,...,k}\{K_i(x)\}$ where K_i is defined by:*

$$K_i(x) = \frac{1}{\sigma^2}\left(\alpha\|\mu_i - P_i(x)\|^2 + (1-\alpha)\|x - P_i(x)\|^2\right).$$

For particular values of α, HDDAi has simple geometrical interpretations:

- Case $\alpha = 0$: HDDAi assigns 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 assigns x to the class associated with the closest subspace $\mathbb{E}_i$.
- Case $\alpha = 1$: HDDAi assigns 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 assigns 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 assigns x to the class realizing a compromise between the two previous cases. The estimation of α is discussed in the following section.

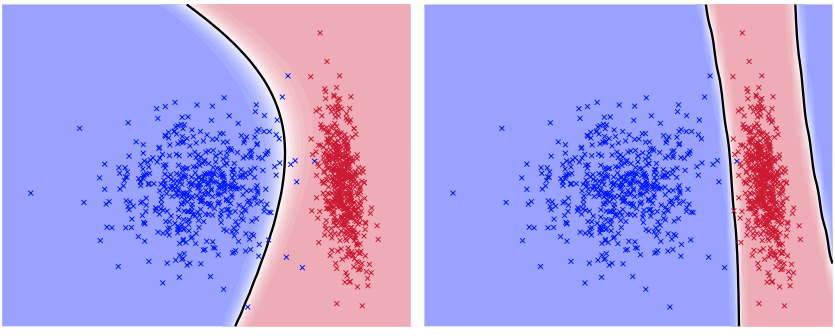
4.2 Homothetic Decision Rule (HDDAh)

This method differs from the previous one by removing the constraint $\sigma_i = \sigma$, and classes are thus homothetic.

Proposition 2. *In this case, the decision rule classifies x as the class C_{i^*} such that $i^* = \text{argmin}_{i=1,...,k}\{K_i(x)\}$ where K_i is defined by:*

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

HDDAh generally favours classes with a large variance. We can observe on Figure 2 that HDDAh favours the blue class which has the largest variance whereas HDDAi gives the same importance to both classes. It assigns to the blue class a point which is far from the axis of the red class, *i.e.,* which does not live in the specific subspace of the red class.



(a) HDDAi (b) HDDAh

Fig. 2. Decision rules obtained with HDDAi and HDDAh on simulated data

4.3 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 $d_i \log(\frac{1-\alpha}{\alpha})$ must be added and if π_i are free, then $-2\log(\pi_i)$ must be added.

5 Estimators

The estimators are obtained by Maximum Likelihood (ML) estimation based on the learning dataset. In the following, parameters π_i, μ_i and Σ_i of the class C_i are estimated by their empirical counterparts:

$$\hat{\pi_i} = \frac{n_i}{n}, \; \hat{\mu_i} = \frac{1}{n_i} \sum_{x_j \in C_i} x_j, \; \hat{\Sigma_i} = \frac{1}{n_i} \sum_{x_j \in C_i} (x_j - \hat{\mu_i})^t (x_j - \hat{\mu_i}),$$

where n_i is the cardinality of the class C_i.

5.1 HDDA Estimators

Assuming for the moment that the d_i are known, we obtain the following ML estimates.

Proposition 3. *The ML estimators of matrices $\tilde{Q}_i$ and parameters a_i and b_i exist and are unique, $\forall i = 1, ..., k$:*

(i) The d_i first columns of $\tilde{Q}_i$ are estimated by the eigenvectors associated to the d_i largest eigenvalues of $\hat{\Sigma_i}$,
(ii) $\hat{a_i}$ is the mean of the d_i largest eigenvalues of $\hat{\Sigma_i}$:

$$\hat{a_i} = \frac{1}{d_i} \sum_{l=1}^{d_i} \lambda_{il},$$

where λ_{il} is the lth largest eigenvalue of $\hat{\Sigma_i}$,

(iii) $\hat{b_i}$ *is the mean of the* $(p-d_i)$ *smallest eigenvalues of* $\hat{\Sigma_i}$ *and can be written:*

$$\hat{b_i} = \frac{1}{(p-d_i)}\left(\mathrm{Tr}(\hat{\Sigma_i}) - \sum_{l=1}^{d_i}\lambda_{il}\right).$$

Proof. Equation (2.5) of [5] provides the following log-likelihood expression:

$$-2\log(L_i(x_j \in C_i, \mu_i, \Sigma_i)) = n_i \sum_{l=1}^{p}\left(\log\delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma_i} q_{il}\right) + C^{te},$$

with $\delta_{il} = a_i$ if $l \leq d_i$ and $\delta_{il} = b_i$ otherwise. This quantity is to be minimized under the constraint $q_{il}^t q_{il} = 1$, which is equivalent to find a saddle point of the Lagrange function:

$$\mathcal{L}_i = n_i \sum_{l=1}^{p}\left(\log\delta_{il} + \frac{1}{\delta_{il}} q_{il}^t \hat{\Sigma_i} q_{il}\right) - \sum_{l=1}^{p}\theta_{il}(q_{il}^t q_{il} - 1),$$

where θ_{il} are the Lagrange multipliers. The derivative with respect to a_i is:

$$\frac{\partial \mathcal{L}_i}{\partial a_i} = \frac{n_i d_i}{a_i} - \frac{n_i}{a_i^2}\sum_{l=1}^{d_i} q_{il}^t \hat{\Sigma_i} q_{il},$$

and the condition $\frac{\partial \mathcal{L}_i}{\partial a_i} = 0$ implies that:

$$\hat{a_i} = \frac{1}{d_i}\sum_{l=1}^{d_i} q_{il}^t \hat{\Sigma_i} q_{il}. \tag{1}$$

In the same manner, the partial derivative of $\mathcal{L}_i$ with respect to b_i is:

$$\frac{\partial \mathcal{L}_i}{\partial b_i} = \frac{n_i(p-d_i)}{b_i} - \frac{n_i}{b_i^2}\sum_{l=d_i+1}^{p} q_{il}^t \hat{\Sigma_i} q_{il},$$

and the condition $\frac{\partial \mathcal{L}_i}{\partial b_i} = 0$ implies that:

$$\hat{b_i} = \frac{1}{(p-d_i)}\sum_{l=d_i+1}^{p} q_{il}^t \hat{\Sigma_i} q_{il} = \frac{1}{(p-d_i)}\left(\mathrm{Tr}(\hat{\Sigma_i}) - \sum_{l=1}^{d_i} q_{il}^t \hat{\Sigma_i} q_{il}\right). \tag{2}$$

In addition, the gradient of $\mathcal{L}_i$ with respect to q_{il} is $\forall l \leq d_i$:

$$\nabla_{q_{il}}\mathcal{L}_i = 2\frac{n_i}{\delta_{il}}\hat{\Sigma_i} q_{il} - 2\theta_{il} q_{il},$$

and by multiplying this quantity on the left by q_{il}^t, we obtain:

$$q_{il}^t \nabla_{q_{il}}\mathcal{L}_i = 0 \Leftrightarrow \theta_{il} = \frac{n_i}{\delta_{il}} q_{il}^t \hat{\Sigma_i} q_{il}.$$

Consequently, $\hat{\Sigma}_i q_{il} = \frac{\theta_{il}\delta_{il}}{n_i} q_{il}$ which means that q_{il} is the eigenvector of $\hat{\Sigma}_i$ associated to the eigenvalue $\lambda_{il} = \frac{\theta_{il}\delta_{il}}{n_i}$. Replacing in (1) and (2), we obtain the ML estimators for a_i and b_i. Vectors q_{il} being eigenvectors of $\hat{\Sigma}_i$ which is a symmetric matrix, this implies that $q_{il}^t q_{ih} = 0$ if $h \neq l$. In order to minimize the quantity $-2\log L_i$ at the *optimum*, $\hat{a}_i$ must be as large as possible. Thus, the d_i first columns of Q_i must be the eigenvectors associated to the d_i largest eigenvalues of $\hat{\Sigma}_i$. □

Note that the decision rule of HDDA requires only the estimation of the matrix $\tilde{Q}_i$ instead of the entire Q_i and this reduces significantly the number of parameters to estimate. For example, if we consider 100-dimensional data, 4 classes and common intrinsic dimensions d_i equal to 10, HDDA estimates only 4 323 parameters whereas QDA estimates 20 603 parameters.

5.2 HDDAi Estimators

Proposition 4. *The ML estimators of parameters α and σ exist and are unique:*

$$\hat{\alpha} = \frac{\hat{b}}{\hat{a} + \hat{b}}, \ \hat{\sigma^2} = \frac{\hat{a}\hat{b}}{\hat{a} + \hat{b}},$$

with $\hat{a} = \frac{\sum_{i=1}^k n_i \sum_{l=1}^{d_i} \lambda_{il}}{np\gamma}, \hat{b} = \frac{\sum_{i=1}^k n_i \left(\mathrm{Tr}(\hat{\Sigma}_i) - \sum_{l=1}^{d_i} \lambda_{il}\right)}{np(1-\gamma)}$ *where* $\gamma = \frac{1}{np}\sum_{i=1}^k n_i d_i$ *and λ_{il} is the lth largest eigenvalue of $\hat{\Sigma}_i$.*

Proof. In this case, the log-likelihood expression is:

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{l=1}^{p} \left(\log \delta_{il} + \frac{1}{\delta_{il}}\lambda_{il}\right) + C^{te},$$

where $\delta_{il} = a$ if $l \leq d_i$ and b otherwise. One can write:

$$-2\frac{\partial}{\partial a}\log(L) = 0 \Leftrightarrow \sum_{i=1}^{k} n_i \sum_{l=1}^{d_i} \left(\frac{1}{a} - \frac{1}{a^2}\lambda_{il}\right) = 0 \Leftrightarrow \hat{a} = \frac{\sum_{i=1}^k n_i \sum_{l=1}^{d_i} \lambda_{il}}{np\gamma},$$

with $\gamma = \frac{1}{np}\sum_{i=1}^k n_i d_i$. Similarly,

$$-2\frac{\partial}{\partial b}\log(L) = 0 \Leftrightarrow \hat{b} = \frac{\sum_{i=1}^k n_i \left(\mathrm{Tr}(\hat{\Sigma}_i) - \sum_{l=1}^{d_i} \lambda_{il}\right)}{np(1-\gamma)}.$$

Replacing these estimates in expressions of α and σ concludes the proof. □

5.3 HDDAh Estimators

Proposition 5. *The ML estimate of α has the following formulation according to σ_i, $\forall i = 1, ..., k$:*

$$\hat{\alpha}(\sigma_1, ..., \sigma_k) = \frac{(\Lambda + 1) - \sqrt{\Delta}}{2\Lambda},$$

with the notations:

$$\Delta = (\Lambda + 1)^2 - 4\Lambda\gamma,\ \Lambda = \frac{1}{np}\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\left(2\sum_{l=1}^{d_i}\lambda_{il} - \mathrm{Tr}(\hat{\Sigma}_i)\right),$$

and the ML estimate of σ_i^2 *has the following formulation according to* α*:*

$$\forall i = 1, ..., k,\ \hat{\sigma_i^2}(\alpha) = \frac{1}{p}\left((2\alpha - 1)\sum_{l=1}^{d_i}\lambda_{il} + (1-\alpha)\,\mathrm{Tr}(\hat{\Sigma}_i)\right).$$

Proof. In this case, one can write:

$$-2\log(L) = \sum_{i=1}^{k} n_i \left[2p\log\sigma_i - d_i\log\alpha - (p-d_i)\log(1-\alpha) + \frac{1}{\sigma_i^2}\left((2\alpha-1)\sum_{l=1}^{d_i}\lambda_{il} + (1-\alpha)\,\mathrm{Tr}(\hat{\Sigma}_i)\right)\right].$$

Therefore,

$$\frac{\partial}{\partial\alpha}\log(L) = 0 \Leftrightarrow \sum_{i=1}^{k} n_i\left(-\frac{d_i}{\alpha} + \frac{(p-d_i)}{(1-\alpha)} + \frac{2\sum_{l=1}^{d_i}\lambda_{il}}{\sigma_i^2} - \frac{\mathrm{Tr}(\hat{\Sigma}_i)}{\sigma_i^2}\right) = 0,$$
$$\Leftrightarrow np\left(-\frac{\gamma}{\alpha} + \frac{(1-\gamma)}{(1-\alpha)} + \Lambda\right) = 0,$$

where $\gamma = \frac{1}{np}\sum_{i=1}^{k} n_i d_i$ and $\Lambda = \frac{1}{np}\sum_{i=1}^{k}\frac{n_i}{\sigma_i^2}\left(2\sum_{l=1}^{d_i}\lambda_{il} - \mathrm{Tr}(\hat{\Sigma}_i)\right)$. Thus,

$$\frac{\partial}{\partial\alpha}\log(L) = 0 \Leftrightarrow \psi(\alpha) = \Lambda\alpha^2 - (\Lambda+1)\alpha + \gamma = 0.$$

The discriminant of the previous equation is $\Delta = (\Lambda + 1 - 2\gamma)^2 + 4\gamma(1-\gamma)$ with $\gamma < 1$ and consequently $\Delta > 0$. By remarking that $\psi(0) = \gamma > 0$ and $\psi(1) = \gamma - 1 < 0$, one can conclude that the solution is in $[0, 1]$ and is the smallest of both solutions of $\frac{\partial}{\partial\alpha}\log(L) = 0$. In addition,

$$\frac{\partial}{\partial\sigma_i}\log(L) = 0 \Leftrightarrow \sigma_i^2 = \frac{1}{p}\left((2\alpha-1)\sum_{l=1}^{d_i}\lambda_{il} + (1-\alpha)\,\mathrm{Tr}(\hat{\Sigma}_i)\right),$$

and thus provides the expression of σ_i^2 according to α. □

Note that the estimators of both α and σ_i are not explicit and thus they should be computed using an iterative procedure.

5.4 Estimation of the Intrinsic Dimension

The estimation of the dataset intrinsic dimension is a difficult problem which does not have an explicit solution. If the dimensions d_i are common between classes, *i.e.,* $\forall i = 1, ..., k$, $d_i = d$, we determine by cross-validation the dimension d which maximizes the correct classification rate on the learning dataset. Otherwise, we use an approach based on the eigenvalues of the class conditional covariance matrix $\hat{\Sigma}_i$. The jth eigenvalue of $\hat{\Sigma}_i$ corresponds to the fraction of the full variance carried by the jth eigenvector of $\hat{\Sigma}_i$. We therefore estimate the class specific dimension d_i, $i = 1, ..., k$, with the empirical method scree-test of Cattell [3] which analyzes the differences between eigenvalues in order to find a break in the scree. The selected dimension is the one for which the subsequent differences are smaller than a threshold t. In our experiments, the threshold t is chosen by cross-validation. We also used the probabilistic criterion BIC [8] which gave very similar dimension choices. In our experiments, the estimated intrinsic dimensions are on average 10 whereas the dimension of the original space is 128.

6 Experimental 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. HDDA seems therefore well adapted to this problem. In this section, we use HDDA to recognize object parts in images.

6.1 Protocol and Data

For our experiments, small scale-invariant regions are detected on each image and they are characterized by the local Sift descriptor [7]. We extracted Sift descriptors of dimension 128 for 100 motorbike images from a recently proposed visual recognition database [4]. For these local descriptors, we selected 2000 descriptors representing 4 classes: wheels, seat, handlebars and background. The learning and the test dataset contain respectively 500 and 1500 descriptors. The pre-processed data are available at `http://lear.inrialpes.fr/~bouveyron/data/data_swc.tgz`. We compared HDDA to the following classical discriminant analysis methods: Linear Discriminant Analysis (LDA), Fisher Discriminant Analysis (FDA) and Support Vector Machines with a RBF kernel (SVM). The parameters t of HDDA and γ of SVM are estimated by cross-validation on the learning dataset.

6.2 Recognition Results

Figure 3 shows recognition results obtained using HDDA methods and state-of-the-art methods with respect to the number of descriptors classified as positive.To obtain the plots we vary the decision boundary between object classes and background, *i.e.*, we change the posterior probabilities provided by generative methods and, for SVM, we vary the parameter C. On the left, only the

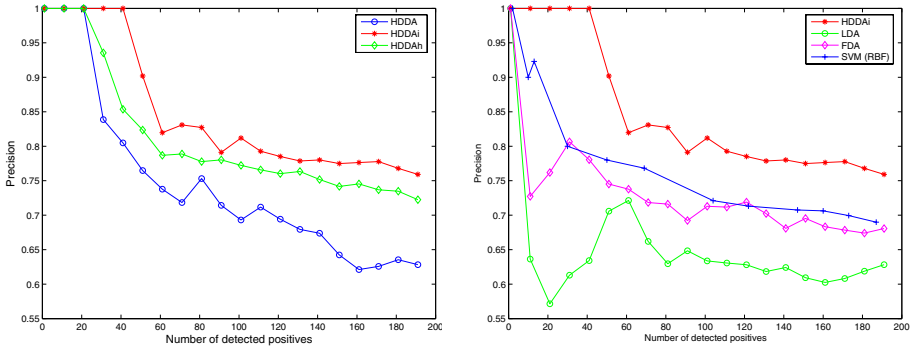


Fig. 3. Classification results for the "motorbike" recognition: comparison between HDDA methods (left) and between HDDAi and state-of-the-art methods (right)

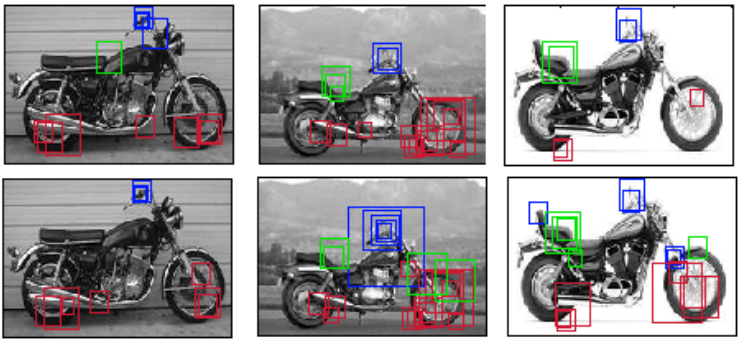


Fig. 4. Recognition of "motorbike" parts using HDDAi (top) and SVM (bottom). The colors blue, red and green are respectively associated to handlebars, wheels and seat.

descriptors with the highest probabilities to belong to the object are used. As a results only a small number of descriptors are classified as positive and their precision (number of correct over total number) is high.

The left plot shows that HDDAi is more efficient than other HDDA methods for this application. This is due to the fact that parameters b_i are common in HDDAi, *i.e.,* the noise is common between classes. More extensive experiments have confirmed that HDDA with common b_i performs in general well for our data. The right plot compares HDDAi to SVM, LDA and FDA. First of all, we observe that the results for LDA (without dimension reduction) are unsatisfying. The results for FDA show that global dimension reduction improves the results. Furthermore, HDDAi obtains better results than SVM and FDA and this demonstrates that our class-specific subspace approach is a good way to classify high-dimensional data. Note that HDDA and HDDAh are also more precise than these three methods when the number of detected positives is small. A comparison with a SVM with a quadratic kernel did not improve the results over the RBF kernel.

Figure 4 presents recognition results for a few test images. These results confirm that HDDAi gives better recognition results than SVM, *i.e.*, the classification errors are significantly lower for HDDAi than for SVM. For example, in the 3rd column of Figure 4, HDDA recognizes the motorbike parts without error whereas SVM makes five errors. In addition, training time for HDDA is as fast as other generative methods and 7 times faster than SVM. Note that recognition time is negligible for all methods.

7 Conclusion

We presented a parameterization of the Gaussian model to classify high-dimen–sional data in a supervised framework. This model results in a robust and fast discriminant analysis method. We successfully used this method to recognize object parts in natural images. An extension of this work is to use the statistical model of HDDA to adapt the method to unsupervised classification.

Acknowledgment

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

References

1. R. Bellman. *Dynamic Programing*. Princeton University Press, 1957.
2. H. Bensmail and G. Celeux. Regularized Gaussian Discriminant Analysis Through Eigenvalue Decomposition. *Journal of the American Statistical Association*, 91: 1743–1748, 1996.
3. R. B. Catell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1:140–161, 1966.
4. R. Fergus, P. Perona, and A. Zisserman. Object Class Recognition by Unsupervised Scale-Invariant Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 264–271, 2003.
5. B. W. Flury. Common Principal Components in k groups. *Journal of the American Statistical Association*, 79:892–897, 1984.
6. J.H. Friedman. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84:165–175, 1989.
7. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
8. G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6: 461–464, 1978.
9. 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.