Class-specific Variable Selection in High-Dimensional Discriminant Analysis through Bayesian Sparsity

F Orlhac, P.-A Mattei, C. Bouveyron, N Ayache

Class-specific Variable Selection in High-Dimensional Discriminant Analysis through Bayesian Sparsity

F. Orlhac[1], P.-A. Mattei[3], C. Bouveyron[1,2], N. Ayache[1]

[1]Epione team, Inria Sophia-Antipolis & Université Côte d'Azur - France

[2]Laboratoire J.A. Dieudonné, UMR CNRS 7135, Université Côte d'Azur - France

[3]Department of Computer Science, IT University of Copenhagen - Denmark

Abstract

Although the ongoing digital revolution in fields such as chemometrics, genomics or personalized medicine gives hope for considerable progress in these areas, it also provides more and more high-dimensional data to analyze and interpret. A common usual task in those fields is discriminant analysis, which however may suffer from the high dimensionality of the data. The recent advances, through subspace classification or variable selection methods, allowed to reach either excellent classification performances or useful visualizations and interpretations. Obviously, it is of great interest to have both excellent classification accuracies and a meaningful variable selection for interpretation. This work addresses this issue by introducing a subspace discriminant analysis method which performs a class-specific variable selection through Bayesian sparsity. The resulting classification methodology is called sparse high-dimensional discriminant analysis (sHDDA). Contrary to most sparse methods which are based on the Lasso, sHDDA relies on a Bayesian modeling of the sparsity pattern and avoids the painstaking and sensitive cross-validation of the sparsity level. The main features of sHDDA are illustrated on simulated and real-world data. In particular, we propose an exemplar application to cancer characterization based on medical imaging using radiomic feature extraction is in particular proposed.

1 Introduction

In many applications, such as chemometrics, genomics or personalized medicine, the observed data are frequently high-dimensional and the classification of such data remains a challenging problem. In particular, when considering the generative (model-based) framework, the corresponding classification (or discriminant analysis) methods show a disappointing behavior in high-dimensional spaces. They suffer from the well-known *curse of dimensionality* (Bellman, 1957) which is mainly due to the fact that model-based methods are dramatically over-parametrized in high-dimensional spaces. Moreover, even though many variables are measured to describe the studied phenomenon, only a small subset of these original variables are in fact relevant for both modeling and classification.

In recent years, several works tried to reduce the data dimensionality or select relevant variables while building a generative classifier. In this context, there are two main approaches. On the one hand, some works assume that the data of each class live in different low-dimensional subspaces. On the other hand, some other works assume that the classes differ only with respect to some of the original features. Therefore, the classification task aims to discriminate the data on a subset of relevant features. Both approaches present two practical advantages: classification results are improved by the removing of non informative features and the interpretation of the obtained classification is eased by the visualization in the subspaces or the meaning of retained variables. We detail hereafter some key works in both approaches. For further reading, we may recommend to refer to Bouveyron and Brunet-Saumard (2013) and Bouveyron (2013).

1.1 Subspace classification

The earliest work in the context of subspace classification is due to R. Fisher who introduced in 1936 the famous-to-be (Fisher) linear discriminant analysis (LDA, Fisher (1936)) method. LDA aims to find a linear subspace that best separates the classes (see Duda et al. (2000) for more details). For this, LDA looks for a linear transformation which projects the observations in a discriminative and low dimensional subspace. The optimal linear transformation is the one maximizing a criterion which is large when the between-class covariance matrix (S_B) is large and when the within-covariance matrix (S_W) is small. Four different criteria can be found in the literature which satisfy such a constraint (see Fukunaga (1990) for a review). The criterion which is traditionally used is $J(U) = \mathrm{trace}((U^t S_W U)^{-1} U^t S_B U)$ where S_W and S_B are respectively the within and the between empirical covariance matrices. Although LDA suffers from some limitations in high-dimensional spaces, it remains a baseline method which usually offers satisfying classification results in most situation, like in this radiomic study (Kirienko et al., 2018), while providing a useful visualization of the data.

Among the subspace classification methods, partial least square discriminant analysis (PLS-DA, Barker and Rayens (2003)) is probably the most popular and used method in chemometrics and genomics. PLS-DA is built on partial least square regression and is adapted to discriminant analysis by replacing the categorical response variable (encoding class memberships) by a multi-dimensional binary one (of the size of the number of classes). Let us recall that PLS aims to find latent representations of both the explanatory and response variables such that the covariance of the corresponding latent variables is maximum. Thus, PLS-DA realizes a supervised dimension reduction which, most of the time, allows an efficient classification of high-dimensional data. In addition, the latent representation of the observed data is usually meaningful for practitioners interested in having a look at their data.

On the model-based side, high-dimensional discriminant analysis (HDDA, Bouveyron et al. (2007)) has also become a standard tool for discriminant analysis, in particular for multi-class classification. HDDA is based on a parsimonious Gaussian mixture model which translates the assumption that the data of each class live in specific low-dimensional subspaces. To do so, the HDDA model assumes a low-rank covariance structure for the classes by imposing that the covariance matrices of each class have only $d_k + 1$ different eigenvalues (d_k being the intrinsic dimensionality of the kth class). In particular, the intrinsic dimensions of the classes control the complexity of the modeling. Interestingly, HDDA allows to recover the usual Gaussian mixture model when assuming that the intrinsic dimensions are all equal to the observed data dimension.

1.2 Variable selection methods

Another way of doing classification in high-dimensional spaces is to perform a selection of the original variables that are relevant for discriminating the classes. Early variable selection methods for classification relied on scores, such as the Fisher score (Duda et al., 2000), to evaluate the ability of a set of variables to discriminate. However, those approaches had to face the combinatorial problem of exploring all possible sets of the original variables. We refer to McLachlan (1992) for further details on early approaches. The most recent approaches rely on greedy algorithms to explore the possible sets of variables. In particular, Murphy et al. (2010) and Maugis et al. (2011) use a forward-backward algorithm to add or remove variables in the selection based on a model selection criterion. The main idea is here to consider a Gaussian mixture model and to choose using Bayesian Information Criterion (BIC) between a model with a specific variable and the same model without it. Among similar variable selection approaches for classification, we can mention Pacheco et al. (2006), Chiang and Pell (2004) and Indahl and Næs (2004).

In the last decade, a new approach for variable selection has emerged which is more computationally tenable for very high-dimensional: variable selection through sparsity penalization. Sparse methods for classification usually intend to find a low-dimensional modeling of the classes under some sparsity constraints. In particular, they impose that the building of the low-dimensional representation of the data relies on only a few of the original variables. Such a constraint can be

imposed using ℓ_0 or ℓ_1 penalties, as commonly done in the context of regression by Lasso (Tibshirani, 1996). A first sparse version of Fisher's linear discriminant analysis was introduced by Trendafilov and Jolliffe (2007). In this seminal work, they introduce DALASS which looks for the solution of the Fisher discrimination problem under sparsity constraints, through ℓ_1 penalization via the Lasso. Thus, the Fisher discriminant subspace that is built only depends on the relevant original variables, helping in turn the interpretation of the classification results. A few years later, Witten and Tibshirani (2011) and Clemmensen et al. (2011) proposed two other formulations for this problem. More specifically, Witten and Tibshirani (2011) considered the Fisher discrimination problem under ℓ_1 penalization and recast it as a biconvex problem. They make use of a minorization-maximization algorithm to optimize the resulting objective function. Conversely, Clemmensen et al. (2011) used optimal scoring, which involves the reformulation of the classification problem as a regression one, which is solved under sparsity constraints. An accelerated optimization procedure for this problem was proposed in Atkins et al. (2017) using accelerated proximal gradient.

Sparse versions of PLS-DA have also been proposed. A first sparse PLS-DA method was proposed by Chung and Keles (2010) which solves the PLS-DA problem with Lasso constraints through a two-step optimization procedure. In the line of sparse PLS (Lê Cao et al., 2008, 2009), Lê Cao et al. (2011) introduced a one-step technique by reformulating the PLS-DA problem as a regression one, on which the Lasso penalty is then added.

1.3 Contributions and organization of the paper

Although many methods have been proposed to perform variable selection in discriminant analysis, all approaches select variables that globally discriminate the classes. Conversely, it would be of interest for practitioners to make a selection of relevant variables for describing each class. In this work, we introduce a modeling which extends the HDDA model of Bouveyron et al. (2007) in order to perform class-specific variable selection. The HDDA model is extended by incorporating a sparsity pattern for each class, allowing in turn an improvement of the classification performance and an easier interpretation of the modeling. The resulting classification methodology is called sparse HDDA (sHDDA). Contrary to most sparse discriminant analysis methods which are based on the Lasso, sHDDA relies on a Bayesian modeling of the sparsity pattern. This allows in particular to avoid the painstaking and sensitive cross-validation of the sparsity level of Lasso-based approaches.

Section 2 introduces the probabilistic model of sHDDA. The inference algorithm for estimating the intrinsic dimensionalities, the sparsity patterns and the model parameters are presented in Section 3. Section 4 is devoted to numerical experiments. These numerical experiments aim to highlight the main features of sHDDA, in particular regarding the variable selection stability and the classification performance. Section 5 proposes an application to cancer characterization based on medical imaging (using radiomic feature extraction). Some concluding remarks are given in Section 6.

It is worth noticing that a package for the R software is in development and it is currently available upon request. The package will incorporate a graphical interface allowing for the users to perform the analysis and visualization of their data in a facilitated way.

2 Sparse HDDA

This section introduces the model for sparse discriminant analysis and discusses model inference. This model extends the HDDA model of Bouveyron et al. (2007) by incorporating a sparsity pattern for each class, allowing in turn an easiest interpretation of the modeling.

2.1 The classification model

Let us consider a complete training dataset $\{(x_1, z_1), ..., (x_n, z_n)\}$ where $z_i \in \{1, \ldots, K\}$ indicates the class label of the observation $x_i \in \mathbb{R}^p$. Let us assume that $\{x_1, \ldots, x_n\}$ and $\{z_1, \ldots, z_n\}$ are respectively independent observed realizations of a random vector $X \in \mathbb{R}^p$ and a random variable $Z \in \{1, \ldots, K\}$. Let us first assume that distribution of Z is a multinomial distribution:

$$Z \sim \mathcal{M}(\pi),$$

with $\pi = (\pi_1, ..., \pi_K)$ and where π_k is therefore the prior probability of the kth class, for $k = 1, ..., K$.

A low-rank decomposition Let us also assume that, conditionally to Z, the observed variable X can be expressed as follows:

$$X_{|Z=k} = Q_k Y + \mu_k + \varepsilon, \tag{1}$$

where $Y \in \mathbb{R}^{d_k}$ is a low-dimensional latent vector, Q_k is a $p \times d_k$ projection matrix, $\mu_k \in \mathbb{R}^p$ is a centering term and ε is a noise term, for $k = 1, ..., K$. From a geometric point of view, Q_k is the orthogonal orientation matrix of the subspace where the data of the kth class are supposed to live. The d_k column vectors of Q_k span the subspace of the kth class and the intrinsic dimensionality of the class is d_k. Let us further assume that, conditionally to Z, the continuous latent variable Y is normally distributed

$$Y|Z = k \sim \mathcal{N}(0, \Delta_k),$$

with variance-covariance matrix $\Delta_k = \mathrm{diag}(\lambda_{k1}, ..., \lambda_{kd_k})$. Conditionally to Z, the noise term ε is also assumed to be normally distributed

$$\varepsilon|Z = k \sim \mathcal{N}(0, b_k I_p).$$

Incorporating sparsity In order to incorporate sparsity in a generative way, we assume that the orientation matrices Q_k of the class-specific subspaces, for $k = 1, ..., K$, can be decomposed as

$$Q_k = V_k W_k,$$

where $V_k = \mathrm{diag}(v_k)$ with $v_k = (v_{k1}, ..., v_{kp}) \in \{0, 1\}^p$ and W_k is a $p \times d_k$ matrix. Thus, (1) becomes:

$$X_{|Z=k} = V_k W_k Y + \mu_k + \varepsilon. \tag{2}$$

Note that each binary parameter v_{kj} indicates wether the jth variable is relevant for modelling the kth class. For each class k, $q_k = \sum_{j=1}^{p} v_{kj}$ represents the number of relevant variables.

Marginal distribution With all these assumptions, the conditional distribution of X is:

$$X|Y,\, Z = k \sim \mathcal{N}(V_k W_k Y + \mu_k, b_k I_p),$$

and its marginal distribution is therefore a mixture of Gaussians:

$$p(x) = \sum_{k=1}^{K} \pi_k \phi(x; \mu_k, \Sigma_k),$$

where $\phi(.)$ denotes the multivariate Gaussian density function parametrized, for the kth component, by its mean vector μ_k and its covariance matrix $\Sigma_k = V_k W_k \Delta_k W_k^t V_k^t + b_k I_p$. Notice that

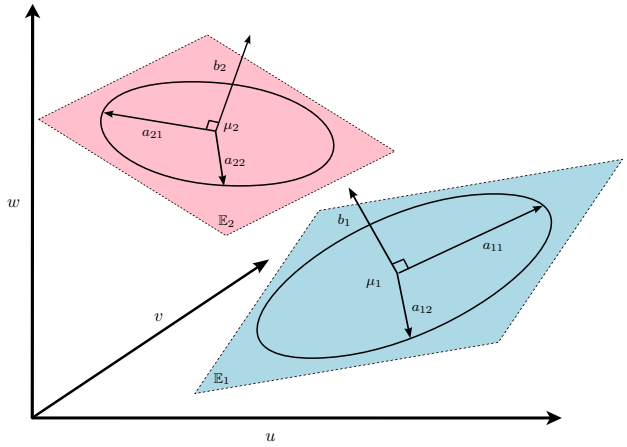

Figure 1: Illustration of the proposed modeling for two classes in their specific subspaces.

Σ_k has therefore a specific covariance srtucture since $\tilde{Q}_k^t \Sigma_k \tilde{Q}_k$ is such that:

$$\tilde{Q}_k^t \Sigma_k \tilde{Q}_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ & \ddots & \\ 0 & & b_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} p - d_k \end{matrix}$$

where $\tilde{Q} = [Q_k, \bar{Q}_k]$, $\bar{Q}_k$ being an arbitrary complementary matrix, and $a_{kj} = \lambda_{kj} + b_k$, for $j = 1, ..., d_k$. Figure 1 illustrates the modeling introduced here for the case of two classes.

2.2 Classification of new observations

In the discriminant analysis framework, new observations are usually assigned to a class using the *maximum a posteriori* (MAP) rule which assigns a new observation $x \in \mathbb{R}^p$ to the class for which x has the highest posterior probability $P(Z = k|X = x)$, *i.e.*:

$$\hat{z} = \underset{k=1,...,K}{\operatorname{argmax}} P(Z = k|X = x).$$

Therefore, the classification step mainly consists in calculating this posterior probability $P(Z = k|X = x)$ for each class $k = 1, ..., K$. Maximizing the posterior probability over k is equivalent to minimizing the classification function $\Gamma_k(y) = -2\log(\pi_k \phi(y; \mu_k, \Sigma_k)$ which is, for our model, equal to:

$$\Gamma_k(x) = \|\mu_k - P_k(x)\|^2_{\mathcal{A}_k} + \frac{1}{b_k}\|x - P_k(x)\|^2 + \sum_{j=1}^{d_k} \log(a_{kj}) + (p - d_k)\log(b_k) - 2\log(\pi_k). \quad (3)$$

where $||x||^2_{\mathcal{A}_k} = y^t \mathcal{A}_k y$, $\mathcal{A}_k = Q_k \Delta_k^{-1} Q_k^t$ and $P_k(x) = Q_k Q_k^t (x - \mu_k) + \mu_k$. Proof of this result is provided in Bouveyron et al. (2007).

Besides its computational interest, the above formula provides as well a comprehensive interpretation of the classification function Γ_k which mainly governs the computation of $P(Z = k|Y = y)$. Indeed, it appears that Γ_k mainly depends on two distances: on the one hand, the

distance between the projections on the discriminant subspace $\mathbb{E}$ of the observation y_i and the mean m_k and, on the other hand, the distance between the projections on the complementary subspace of x and μ_k. Therefore, the posterior probability $P(Z = k|X = x)$ is close to 1 if both distances are small which seems quite natural. Obviously, these distances are also balanced by the variances in the class-specific subspace and its complementary and by the mixture proportions.

3 Model inference

We now present a strategy for inferring the model described above. Due to the complex nature of the modeling, we address the inference in three steps: intrinsic dimensionality estimation, sparsity pattern estimation and model parameter estimation.

3.1 Intrinsic dimensionality estimation

Before to go further in the inference, it is first necessary to estimate the intrinsic dimensionality of each class, *i.e.* estimate the discrete parameters $d_1, ..., d_K$. The estimation of the intrinsic dimension of a dataset is a difficult problem which occurs frequently in data analysis, such as in principal component analysis (PCA). A classical solution in PCA is to look for a break in the eigenvalue scree of the covariance matrix. This strategy relies on the fact that the jth eigenvalue of the covariance matrix corresponds to the fraction of the full variance carried by the jth eigenvector of this matrix. Following Bouveyron et al. (2007), we propose to make use of the scree-test of Cattell (1966) for estimating the intrinsic dimensions d_k, $k = 1, ..., K$. For each class, the selected dimension is the one for which the subsequent eigenvalues differences are smaller than a threshold. The threshold can be provided by the user (we recommend 10% of the largest difference), selected through cross-validation or using model selection tools, such as BIC (Schwarz, 1978).

3.2 Sparsity pattern estimation

In order to recover the sparsity pattern for each of the K classes, we leverage the strategy proposed by Bouveyron et al. (2016) for performing Bayesian variable selection in probabilistic PCA. To this end, we complete the previous modeling by considering priors for both the sparsity patterns $p(v_1), ..., p(v_K)$, and the projection parameters $p(W_1), ..., p(W_K)$. This allows us to select the sparsity patterns that are the sparsity patterns with the highest posterior probability. Specifically, following the parametric empirical Bayes framework leads to choose, for each $k = 1, ..., K$, the following sparsity pattern

$$\begin{aligned}
\hat{v}_k &= \text{argmax}_{v_k \in \{0,1\}^p} p(v_k|X_k) \\
&= \text{argmax}_{v_k \in \{0,1\}^p} p(v_k) p(X_k|v_k) \\
&= \text{argmax}_{v_k \in \{0,1\}^p} p(v_k) \prod_{i=1}^{n_k} \int_{\mathbb{R}^{p \times d_k}} p(x_{ki}|W_k, v_k) p(W_k) dW_k,
\end{aligned}$$

where X_k is the set of $n_k = \sum_i^n \mathbf{1}\{z_i = k\}$ observations belonging to the kth class, *i.e.* $X_k = \{x_i|z_i = k\}$.

Regarding the projection parameters following Bouveyron et al. (2016), we derive a closed-form expression of the marginal likelihood $p(X_k|v_k)$. To this end, we consider a Gaussian prior for the matrix W. Specifically, we assume that its rows are a priori i.i.d. following

$$\forall j \in \{1, .., p\},\ w_i \sim \mathcal{N}(0, \Delta_k^{-1}/\alpha_k^2).$$

The hyperparameter $\alpha_k > 0$ controls the width of the prior variance, and will be chosen via parametric empirical Bayes. In that setting, the marginal likelihood becomes

$$p(X_k|v_k,\alpha_k) = \prod_{i=1}^{n_k} p(x_{ki}|v_k,\alpha_k) = \prod_{i=1}^{n_k} \int_{\mathbb{R}^{p\times d}} p(x_{ki}|W_k,v_k)p(W_k|\alpha_k)dW_k.$$

In order to obtain a closed-form formulation of the evidence, we dissociate the modelling of the noise of active and inactive variables. To do so, for each class $k = 1, ..., K$, the noise ε is assumed to decompose as follows:

$$\varepsilon_{|Z=k} = v_k\varepsilon_1 + \bar{v}_k\varepsilon_2,$$

where $\varepsilon_1|Z = k \sim \mathcal{N}(0, \sigma_{k1}^2\mathbf{I}_p)$ is the noise of the inactive variables and $\varepsilon_2|Z = k \sim \mathcal{N}(0, \sigma_{k2}^2\mathbf{I}_p)$ is the noise of the active variables, both for the kth class. Making use of Theorem 2 of Bouveyron et al. (2016) and assuming that $\sigma_{k1}^2 \to 0$, we end up with a closed-form formulation for the evidence:

$$\begin{aligned}\log p(X_k|v_k,\alpha_k) = &-\frac{\sum_{i=1}^{n_k} ||x_{ki}^{\bar{v}_k} - \mu_k^{\bar{v}_k}||^2}{2\sigma_{k2}^2} - n_k(p-q_k)\log\sigma_{k2} + \frac{n_k q_k}{2}\log\alpha_k \\ &+ \sum_{i=1}^{n_k}\left(\log K_{(q_k-d_k)/2}(\alpha_k||x_{ki}^{v_k} - \mu_k^{v_k}||) - q_k\log||x_{ki}^{v_k} - \mu_k^{v_k}||\right).\end{aligned} \tag{4}$$

where K_ν denotes the modified Bessel function of the second kind (Abramowitz and Stegun, 1965). The hyperparameter α_k can be found by solving a univariate convex optimization problem, as in Bouveyron et al. (2016).

Regarding the prior on the sparsity pattern, while Bouveyron et al. (2016) only used the uniform noninformative prior $p(v_k) \propto 1$, we extend their approach by also considering sparsity inducing priors. Specifically, we consider a product of Bernoulli distributions with the same parameter. This shared parameter is chosen so that the prior probability of selecting more than n variables is at most 5%. This rationale comes from the fact that singular value decomposition, which is the cornerstone of our inference strategy, works best when $n > p$ or $n \approx p$ (Johnstone and Lu, 2009). A similar reasoning was applied to linear regression by Narisetty and He (2014).

The main advantage of the chosen framework is that it exhibits a closed-form formulation of the posterior probabilities of all sparsity patterns. However, directly maximizing this quantity over the binary vector v_k is not tractable because 2^p models need to be evaluated. It is however possible to reduce the complexity of the optimization problem by ranking the candidate variables using the fast variational expectation-maximization (VEM) algorithm proposed in Bouveyron et al. (2016) on a relaxed version of the above modeling. Then, once the variable are ranked, it is possible to optimize the evidence over the path of models (4). This procedure, first introduced by Latouche et al. (2016) in a linear regression context, provides both the number q_k and the list of active variables for the kth class.

3.3 Model parameter estimation

Finally, the estimation of the model parameters $\theta_k = \{\pi_k, \mu_k, Q_k, a_{kj}, b_k\}$ is performed through maximum likelihood. The log-likelihood function $\mathcal{L}$ takes the following form in the case of the statistical model described above:

$$\mathcal{L}(X;\theta) = \sum_{i=1}^{n}\log p(x_i, z_i) = \sum_{i=1}^{n}\log p(z_i) + \sum_{i=1}^{n}\log p(x_i|z_i). \tag{5}$$

Only the first of those two terms depend on π, leading to the estimate $\hat{\pi}_k = n_k/n$ for all $k \in \{1, ..., K\}$. The second terms depend on all the other parameters, we can rewrite it:

$$\sum_{i=1}^{n} \log p(x_i|z_i) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log p(x_{ki}|z_i = k) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log(p(x_{ki}^{v_k}|z_i = k)p(x_{ki}^{\bar{v}_k}|z_i = k)) \quad (6)$$

$$= \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log \phi(x_{ki}^{v_k}; \mu_k^{v_k}, Q_k^{v_k}\Delta_k(Q_k^{v_k})^t + b_k I_{q_k}) + \sum_{k=1}^{K}\sum_{i=1}^{n_k} \log \phi(x_{ki}^{\bar{v}_k}; \mu_k^{\bar{v}_k}, b_k I_{p-q_k}). \quad (7)$$

Therefore, the maximum likelihood estimates of the class means is

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_{ik}.$$

The first term of Equation (7) exactly corresponds to a sum of likelihoods of probabilistic PCA models (Tipping and Bishop, 1999). As detailed in Lemma 1 (see Appendix A), maximum likelihood in such models is linked to the eigendecomposition of the class-specific covariances defined as

$$S_k = \frac{1}{n_k}\sum_{i=1}^{n_k}(x_{ki}^{v_k} - \mu_k^{v_k})(x_{ki}^{v_k} - \mu_k^{v_k})^t. \quad (8)$$

Therefore, maximum likelihood estimates of the projection parameters are given by

- the columns of $\hat{Q}_k^{v_k}$ are the d_k eigenvectors of S_k that correspond to its d_k largest eigenvalues $\nu_1, ..., \nu_{d_k}$ (recall also that $\hat{Q}_k^{\bar{v}_k} = 0$ by definition of the model),
- for all $k \in \{1, ..., K\}$ and $l \in \{1, ..., d_k\}$, $\hat{\lambda}_{kl} = \nu_{kl} - b_k$.

The last remaining parameters, $b_1, ..., b_K$, appear in both terms of Equation (7). Using Lemma 1, up to constants that do not depend on any of the $b_1, ..., b_K$, this leads, at a maximum of the likelihood, to

$$\sum_{i=1}^{n} \log p(x_i|z_i) = -\sum_{k=1}^{K}\left(\frac{n_k(q_k - d_k)}{2}\log b_k + \frac{n_k}{2b_k}\sum_{l=d_k+1}^{q_k}\lambda_{kl} + \frac{n_k(p-q_k)}{2}\log b_k + \frac{1}{b_k}\sum_{i=1}^{n_k}||x_{ki}^{\bar{v}_k} - \mu_k^{\bar{v}_k}||^2\right), \quad (9)$$

which leads to

$$\hat{b}_k = \frac{1}{p - d_k}\left(\sum_{l=d_k+1}^{q_k}\lambda_{kl} + \frac{1}{n_k}\sum_{i=1}^{n_k}||x_{ki}^{\bar{v}_k} - \mu_k^{\bar{v}_k}||^2\right).$$

3.4 The sHDDA algorithm

We provide in Algorithm 3.4 a summary of the whole process for performing sparse HDDA on a set of data. As any supervised classification algorithm, sHDDA is split into a learning phase and a prediction phase. The learning phase gathers the three inference steps described above: estimation of intrinsic dimensionalities, estimation of active variable sets and estimation of model parameters. Then, the prediction of labels $\{\hat{z}_1^*, ..., \hat{z}_m^*\}$ for new observations $\{x_1^*, ..., x_m^*\}$ is done through the MAP rule. Notice that the prediction step is not mandatory for users only interested in the class-specific variable selection.

4 Numerical experiments

In this section, our goal is to present the specificity of the proposed sHDDA method on a simulated data set and some real data sets, and to compare the performances with other statistical methods.

Algorithm 1 The sparse HDDA (sHDDA) algorithm.

Input: a learning dataset $\{(x_1, z_1), ..., (x_n, z_n)\}$ and a set of new observations $\{x_1^*, ..., x_m^*\}$ to label

Output: predicted labels $\{\hat{z}_1^*, ..., \hat{z}_m^*\}$ for the new observations and estimated model parameters $\hat{\theta}$.

// Learning
For each class $k = 1, ..., K$ do

- estimate the intrinsic dimensionality d_k of the class using Cattell's scree-test,
- estimate the set of active variables v_k for the class by maximizing (4),
- estimate model parameters θ_k through maximizing the likelihood function (5).

// Classification
For each new observation x^* do

- compute for each class $k = 1, ..., K$ the posterior probability $P(Z = k|X = x, \hat{\theta})$ through (3),
- assign the observation x^* to the class with the highest posterior probability:

$$z^* = \text{argmax}_k P(Z = k|X = x, \hat{\theta}).$$

// Return results
Return the predicted labels $\{\hat{z}_1^*, ..., \hat{z}_m^*\}$ and estimated model parameters $\hat{\theta}$.

4.1 Implementation of statistical methods

We used the R package `MASS` for `LDA` function (Fisher, 1936) and `sparseLDA` for `sLDA` function (Clemmensen et al., 2011) with default settings. We used the function of the `penalizedLDA` R package with a cross-validation of parameters for `PenalizedLDA` function (Witten and Tibshirani, 2011). Based on `accSDA` package, we used `ASDA` function with default parameters for the accelerated sparse discriminant analysis of Atkins et al. (2017). For PLS-DA and sPLS-DA (Lê Cao et al., 2011), we used the `mixOmics` R package. For sPLS-DA, we used the function `tune.splsda` to select the optimal number of components and the optimal number of variables to choose in the learning subset with 5-fold cross validation and by using the maximal distance to estimate the classification error rate. For HDDA (Bouveyron et al., 2007), we used the `HDclassif` R package (Bergé et al., 2012). The performances of each statistical method were evaluated by taking into account the percentage of observations correctly identified in the test subsets.

4.2 An introductory example

We simulated here a data set to illustrate the incorporation of a sparsity pattern for each class in the model. For this experiment, $n = 150$ observations were simulated according to (2) with $p = 100$ variables, $k = 3$ classes (50 observations by class), $d = (10, 5, 2)$ for the dimensions of the latent space for each class and $q = 20$ true number of relevant variables in each class. The sparsity pattern was variables 1 to 20 for class 1, variables 16 to 35 for class 2 and variables 31 to 50 for class 3. Therefore, the variables 16 to 20 were common to class 1 and 2 and the variables 31 to 35 were common to class 2 and 3. The variables 51 to 100 were not related to any class. In this simulation, the signal to noise ratio was equal to 0.89 for class 1, 1.00 for class 2 and 1.11 for class 3.

Based on only one run for illustration, sHDDA estimated the dimensions of the latent space

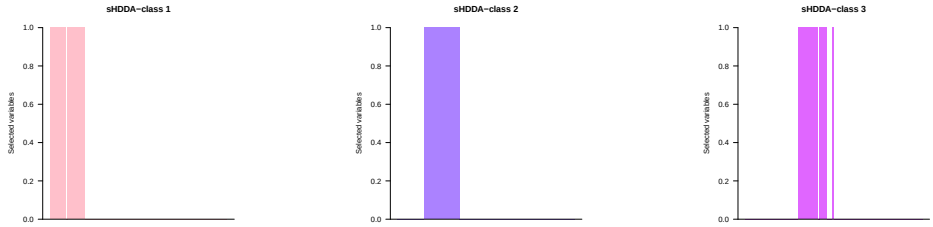

Figure 2: Selected variables for the numerical simulation in each class by sHDDA for one run: class 1 (left), class 2 (center) and class 3 (right).

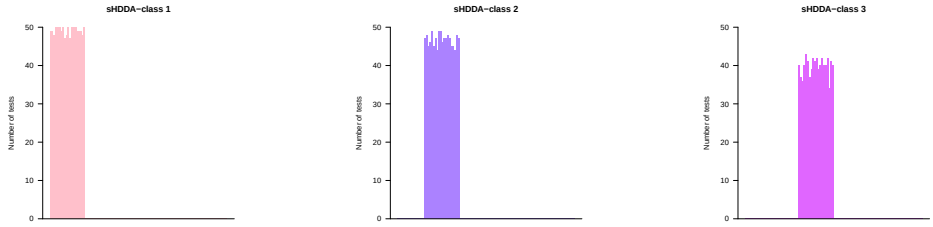

Figure 3: Selected variables for the numerical simulation in each class by sHDDA for 50 runs: class 1 (left), class 2 (center) and class 3 (right).

equal to (7, 5, 2), very close to the simulated parameters. Concerning the variable selection, sHDDA selected 19 variables out of 20 for class 1, 20/20 variables for class 2 and 16/20 variables for class 3. Figure 2 shows that sHDDA selected variables corresponding to the sparsity pattern for each class, did not select variable other than those which were specific of each class and not the variables 51 to 100. Similarly, sHDDA estimated the noise of each class subspace equal to 2.33 for class 1 (versus 2.25 for the simulation), 0.98 for class 2 (versus 1.00) and 0.36 for class 3 (versus 0.36).

4.3 Stability of variable selection

To evaluate the stability of variable selection, we generated 50 data sets repeating the simulation of the previous section, and we compared the selected variables by sHDDA and those by other sparse statistical methods described in 4.1. Figure 3 shows that sHDDA identified all twenty true variables for classes 1 and 2, and 16/20 true variables for class 3 in more than 75% of runs. In addition, sHDDA never selected variables other than those which were specific of each class. Inversely, as shown in Figure 4, other sparse statistical methods selected all variables either for all runs (sLDA, pLDA and ASDA), or for approximately half of the runs (sPLS-DA). As expected, sHDDA leaded to the best performances as the sparsity pattern for each class in the model is advantageous for this method which performs a variable selection by class. In contrast, other sparse methods perform a globally variable selection, but it is surprising that they selected irrelevant variables (variables 51 to 100).

4.4 Benchmark

In this section, the results of sHDDA were compared to other classification approaches based on three data sets. The first data set corresponds to the simulation example described in the previous sections, while the two other data sets represent realistic situations to evaluate the classification power of sHDDA and compared with those obtained for other statistical methods.

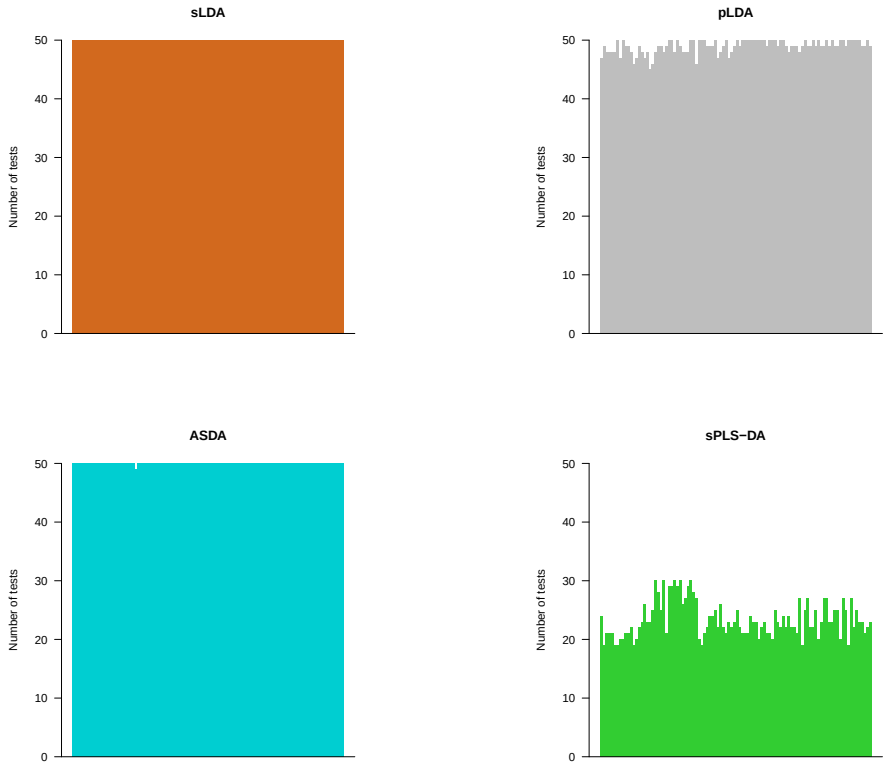

Figure 4: Selected variables for the numerical simulation with 50 runs for: sLDA (top left), pLDA (top right), ASDA (bottom left) and sPLS-DA (bottom right).

Data description

- Data set 1: the first data set corresponds to the simulation setup described in the section 4.2. The simulation was repeated 50 times with $n = 150$ observations, $p = 100$ variables and $k = 3$ classes with each 50 observations. The observations were randomly separated in a learning subset (100 observations) and a test subset (50 observations).
- Data set 2: the second data set illustrates a problem encountered in chemometrics. It comes from the 3-class near-infrared spectroscopy data set previously described and published in Devos et al. (2009). The data set contains $n = 202$ near-infrared spectra of manufactured textiles quantified by $p = 2800$ variables. The objective is to classify samples according to a physical property summarized in $k = 3$ classes (48 samples in class 1, 108 in class 2 and 46 in class 3). The samples were randomly divided 50 times in a learning subset (135 observations) and a test subset (67 observations).
- Data set 3: the third data set corresponds to the study of breast subtype lesions based on gene expression patterns derived from cDNA microarrays. It is presented in the `datamicroarray` R package and initially published in Sørlie et al. (2001). The data set consists of $n = 85$ experimental samples with $p = 456$ cDNA clones in breast cancer carcinoma. The authors divided the data set in $k = 5$ sub-types (14 samples in class 1, 11 in class 2, 13 in class 3, 15 in class 4 and 32 in class 5). The population was randomly divided 50 times in a learning subset (57 observations) and a test subset (28 observations).

Results The performances of classification obtained on the three data sets with the eight statistical methods are presented in Figure 5.

Based on data set 1, we observed that the performances of sLDA were not better than those of LDA for the distinction of 3 classes. pLDA showed a large variability of performances compared to other methods. ASDA, PLS-DA and sPLS-DA provided better results than LDA but the performances were statistically lower than those of HDDA and sHDDA (p-values of Wilcoxon test < 0.001). This result was planned since the simulated sparsity pattern is in favour of these two methods. Compared to HDDA, sHDDA yielded performances slightly below (p-value $= 0.016$) but enables to interpret model more easily with the selection of variables.

For the data set 2, the results is divided into two groups. sLDA, pLDA and PLS-DA yielded similar performances to LDA. These results suggested that the variable selection by sLDA and pLDA compared to LDA was not effective. In contrast, ASDA, sPLS-DA, HDDA and sHDDA provided better results.

Finally, based on the data set 3, ASDA and PLS-DA achieved poorer results compared to LDA and sPLS-DA. Best performances were obtained with pLDA, HDDA and sHDDA to distinguish the 5 classes.

In summary, based on the study of these three data sets, we observed that HDDA and sHDDA yielded good and stable performances compared to the other statistical methods.

5 Application to radiomics

In this section, we focus on a possible application of sHDDA for the selection and combination of radiomic features (Gillies et al., 2016). Radiomics is an emerging discipline in medical studies that consists to extract a large number of image derived phenotypes from medical images. In oncology, the aim is to quantify the tumor heterogeneity based on histogram, shape and texture features in order to improve diagnosis, patient management and treatment monitoring. As is common in medical context, most studies include a hundred patients (or even less), while it is possible to extract from tens to hundreds radiomic features according to the technique used. To illustrate the potential interest of sHDDA in this field, we considered a data set extracted from Grossmann et al. (2017).

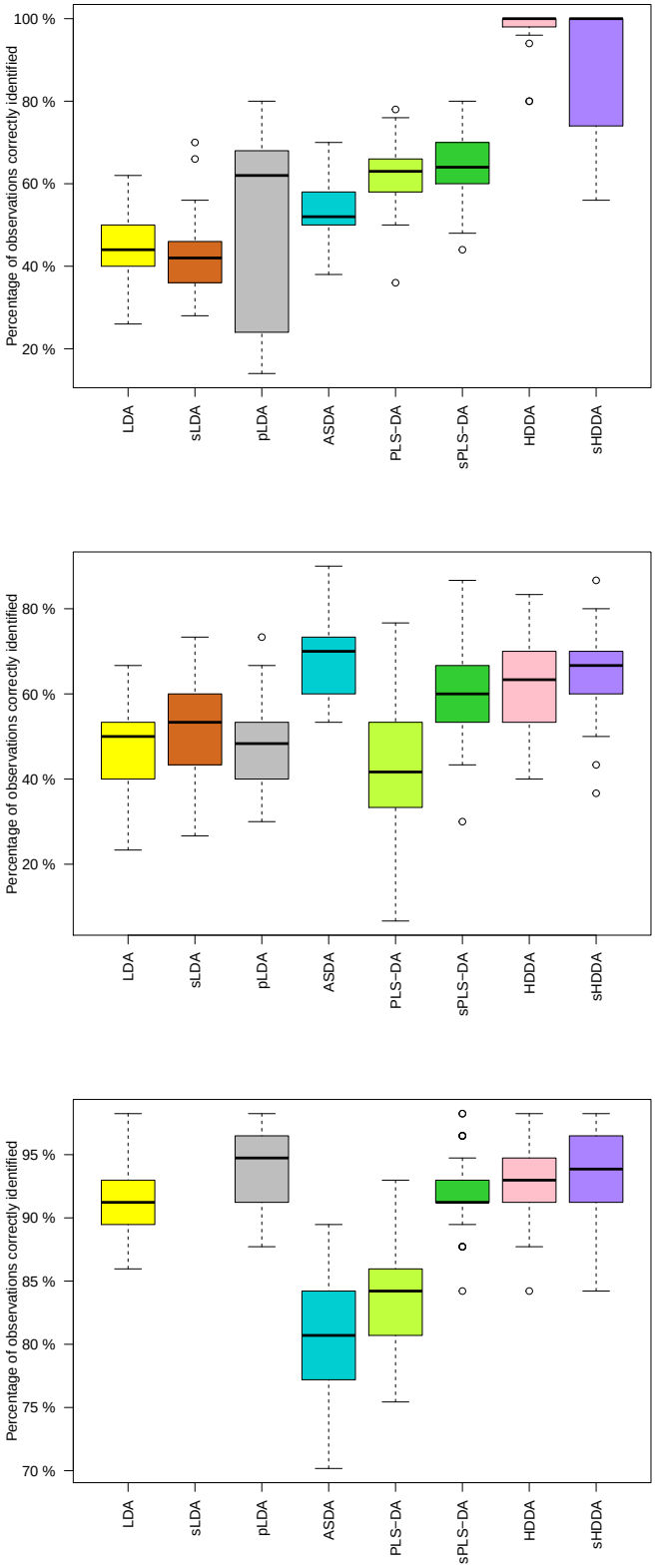

Figure 5: Classification accuracy of the eight statistical methods on data set 1 (top), data set 2 (center) and data set 3 (bottom). Note that sLDA is not applicable on data set 3 because some variables are co-linear.

Method	Any run	At least 25%	At least 50%	At least 75%	All runs
sLDA	35	84	75	53	21
pLDA	0	636	636	609	0
ASDA	0	598	487	265	8
sPLS-DA	0	372	77	2	0
sHDDA - class 1	618	17	14	11	4
sHDDA - class 2	615	17	17	15	8
sHDDA - class 3	617	18	18	14	10

Table 1: Number of features no selected, selected in at least 25% or 50% or 75% of runs and in all runs by each sparse statistical method.

Data description The available data set consists of $n = 87$ patients with a lung cancer. Based on Computed Tomography images, $p = 636$ radiomic features are extracted by the authors including tumor intensity, shape, texture and wavelet features and are made available in supplemental data of Grossmann et al. (2017). Our purpose is to identify the sub-type of lesions according to $k = 3$ classes: 42 adenocarcinoma lesions, 33 squamous carcinoma lesions and 12 tumors of another sub-type. The cohort was randomly divided 50 times in a learning subset (57 patients) and a test subset (30 patients). On test subsets, we compared the classification accuracy of sHDDA with those of other statistical methods and we studied the selected radiomic features.

Results To distinguish lung lesions in three sub-types, sLDA, pLDA and ASDA performed poorer results than LDA, in contrast to PLS-DA, sPLS-DA, HDDA and sHDDA (Figure 6). In particular, HDDA and sHDDA yielded better performances than other methods (p-values of Wilcoxon test < 0.05) except for sPLS-DA.
With sHDDA, 615 features were not selected in any run compared with 35 for sLDA (Table 1). Other sparse methods (pLDA, ASDA and sPLS-DA) selected all features in at least one run. Inversely, sHDDA selected 4, 8 and 10 features respectively for class 1, 2 and 3 in all runs (Figure 7). Twenty-one features with sLDA and 8 features with ASDA were selected in all runs, versus any with pLDA and sPLS-DA (Figure 8). These results show that the model built by sHDDA was more sparse and stable compared to the other statistical methods. Based on at least 75% of runs, sHDDA selected 11 features for the class 1, 15 features for the class 2 and 14 features for the class 3, as the variable selection by sHDDA is specific for each class compared to other sparse methods. In particular, as shown in Table 2, some of selected variables were specified to one class like *Wavelet_HHH_stats_energy* for class 3. Other variables were specified to two classes like *Wavelet_LHL_stats_energy* for classes 1 an 2. The features identified by sHDDA were also selected in at least 94% of runs by sLDA, 72% for pLDA and 56% for ASDA, but in less than 56% of tests for sPLS-DA.
Even though the performances were close for sHDDA and sPLS-DA, the selections of features were completely different: only two features (*Shape_compactness* and *Shape_spherDisprop*) were selected in at least 75% of tests by sPLS-DA and that were not identified by sHDDA in any run. Furthermore, sHDDA retained none of the shape features, only histogram, texture and wavelet features (Table 2).
In the context of radiomics, it is very important to identify a small number of features of interest in order to study each of them and to understand the link between these features and the biological characteristics of lesions (Buvat et al., 2015; Orlhac et al., 2016). This biological validation is necessary for the acceptance of radiomics in clinical practice (Lubner et al., 2017).

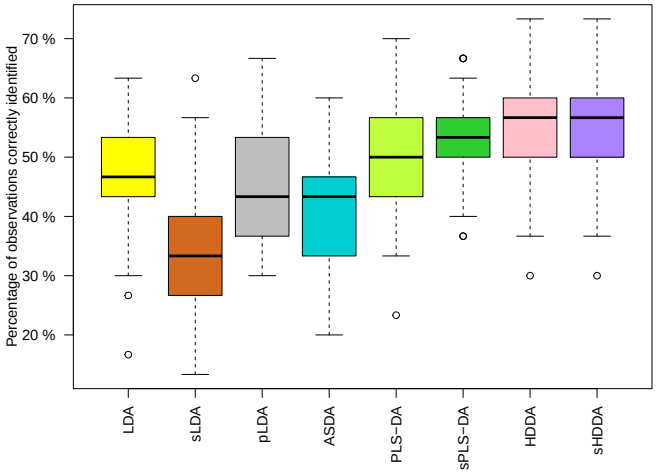

Figure 6: Classification accuracy of the eight statistical methods on the radiomic data set.

Radiomic features	sHDDA class 1	sHDDA class 2	sHDDA class 3	sLDA	pLDA	ASDA	sPLS-DA
GLCM_clusProm		88%	98%	100%	84%	62%	16%
GLSZM_highIntensityLarteAreaEmp	100%	100%	100%	100%	92%	82%	56%
GLSZM_largeAreaEmphasis		94%	96%	100%	72%	98%	20%
Stats_energy	100%	100%	100%	94%	86%	90%	22%
Stats_totalenergy	84%	90%	100%	94%	68%	98%	14%
Wavelet_HHH_stats_energy			82%	100%	90%	90%	50%
Wavelet_HHL_stats_energy	100%	100%	100%	100%	92%	92%	26%
Wavelet_HLH_stats_energy	96%	100%	100%	100%	88%	94%	14%
Wavelet_HLL_stats_energy	98%	100%	100%	100%	90%	98%	28%
Wavelet_LHH_stats_energy	80%	100%	98%	100%	82%	96%	10%
Wavelet_LHL_stats_energy	98%	96%		100%	88%	72%	14%
Wavelet_LLH_stats_energy	98%	92%		98%	72%	86%	10%
Wavelet_LLL_glcm_clusProm		88%	100%	100%	76%	62%	14%
Wavelet_LLL_stats_energy	100%	100%	100%	98%	86%	90%	22%
Wavelet_LLL_stats_totalenergy	90%	100%	100%	100%	80%	98%	12%
Wavelet_LLL_stats_var		82%	100%	100%	82%	56%	16%

Table 2: List of selected features by sHDDA in at least 75% of runs for the radiomic data set and for each class (class 1: other sub-type, class 2: adenocarcinoma and class 3: squamous cell carcinoma) and associated percentage. The four last columns correspond to the percentage of runs for which each feature is selected for each sparse statistical method.

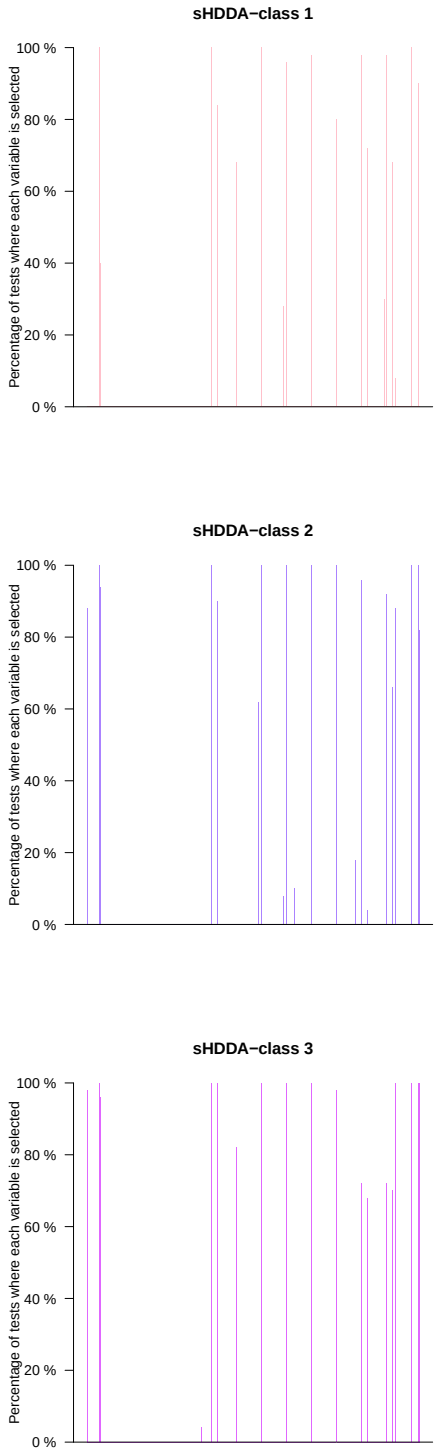

Figure 7: Selected variables on 50 runs for the radiomic data set by sHDDA: other sub-type (top), adenocarcinoma (center) and squamous carcinoma (bottom).

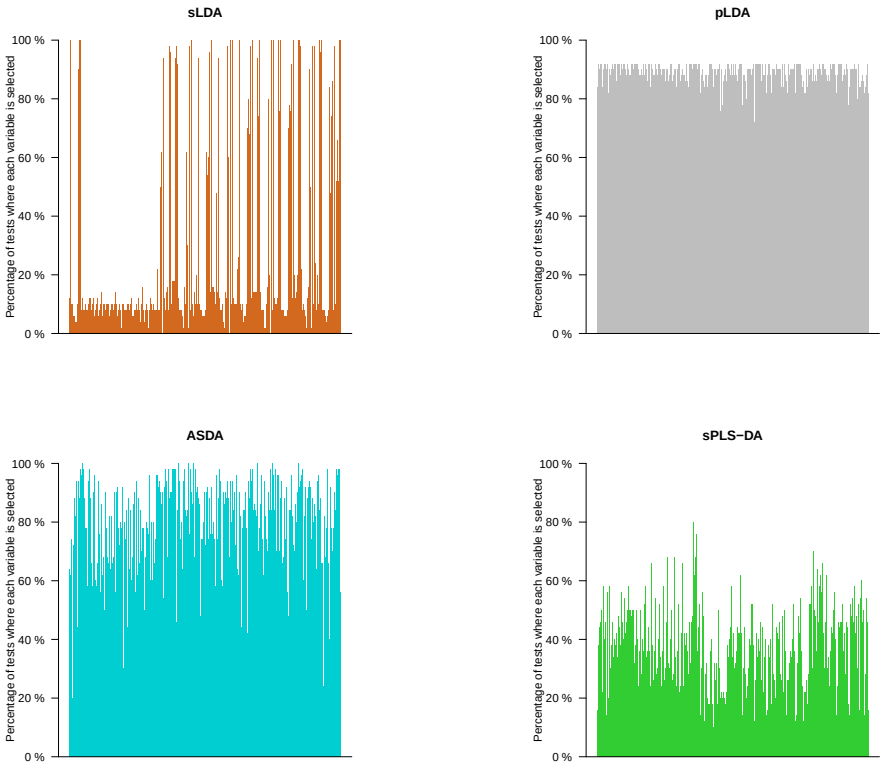

Figure 8: Selected variables for the radiomic data set with 50 runs for: sLDA (top left), pLDA (top right), ASDA (bottom left) and sPLS-DA (bottom right).

6 Conclusion

Contrary to other sparse statistical methods that make a globally selection of relevant variables, this article presents a modeling which extends the HDDA model for allowing to perform class-specific variable selection. sHDDA includes a subspace discriminant analysis method which leds to a class-specific variable selection through Bayesian sparsity.The method allows an improvement of the classification performance and an easiest interpretation of the modeling thanks to the meaningful variable selection by class. Experimental results demonstrated the high performances for the classification and the great stability for the variable selection. These advantages make it an efficient tool for all applications with high-dimensional data and for which an interpretation of the model is expected such as in all "omics" sciences (genomics, proteomics, metabolomics, radiomics, ...) for instance.

A Maximum likelihood estimates

The following lemma gives us the technical tools to derive the estimates of Section 3.3. Notice that the model considered in this lemma is non-identifiable, and that infinitely many other maximum likelihood estimates exist. The ones we consider can be conveniently computed by simply performing a singular value decomposition.

Lemma 1 *Consider some i.i.d. data* $(x_1, ..., x_n)$ *coming from the model*

$$X \sim \mathcal{N}(\mu, Q\Delta Q^t + b),$$

with $\mu \in \mathbb{R}^p$, $Q \in \mathbb{R}^{p\times d}$, *and* $\Delta = diag(a_1, ..., a_d)$.

Let $S = (1/n)\sum_{i=1}^{n}(x_i - \mu)(x_i - \mu)^t$, *and let* $S = U diag(\lambda_1, ..., \lambda_p)U^T$ *be the eigenvalue decomposition of* S. *Maximum likelihood estimates of* μ, Q *and* Δ *are given by:*

- $\hat{\mu}$ *is the empirical mean of* $x_1, ..., x_n$,
- *the columns of* $\hat{Q}$ *are the first* d *columns of* U,
- *for all* $l \in \{1, ..., d\}$, $a_l = \lambda_l - b$.

  Moreover, at its maximum, the log-likelihood can be written

$$\sum_{i=1}^{n} \log p(x_i) = -\frac{n(p-d)}{2}\log b - \frac{n}{2b}\sum_{l=d+1}^{p} \lambda_l + K,$$

  where K *is a constant that does not depend on* b.

A proof of this lemma can be found in Tipping and Bishop (1999, Appendices A-1,A-2).

References

M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

S. Atkins, G. Einarsson, B. Ames, and L. Clemmensen. Proximal methods for sparse optimal scoring and discriminant analysis. *arXiv preprint arXiv:1705.07194*, 2017.

M. Barker and W. Rayens. Partial least squares for discrimination. *Journal of chemometrics*, 17 (3):166–173, 2003.

R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

L. Bergé, C. Bouveyron, and S. Girard. Hdclassif: An r package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6):1–29, 2012.

C. Bouveyron. Probabilistic model-based discriminant analysis and clustering methods in chemometrics. *Journal of Chemometrics*, 27(12):433–446, 2013.

C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics and Data Analysis*, 71:52–78, 2013.

C. Bouveyron, S. Girard, and C. Schmid. High-dimensional discriminant analysis. *Communication in Statistics: Theory and Methods*, 36:2607–2623, 2007.

C. Bouveyron, P. Latouche, and P.-A. Mattei. Bayesian variable selection for globally sparse probabilistic PCA. *Technical report, HAL-01310409, Université Paris Descartes*, 2016.

I. Buvat, F. Orlhac, and M. Soussan. Tumor texture analysis in pet: Where do we stand? *Journal of Nuclear Medicine*, 56(11):1642–1644, 2015.

R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2): 245–276, 1966.

L. Chiang and R. Pell. Genetic algorithms combined with discriminant analysis for key variable identification. *Journal of Process Control*, 14(2):143–155, 2004.

D. Chung and S. Keles. Sparse partial least squares classification for high dimensional data. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

L. Clemmensen, T. Hastie, D. Witten, and B. Ersbøll. Sparse discriminant analysis. *Technometrics*, 53(4):406–413, 2011.

O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, and J.-P. Huvenne. Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems*, 96(1):27 – 33, 2009.

R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley & Sons, 2000.

R.A. Fisher. The use of multiple measurements in taxonomic problems. *The Annals of Eugenics*, 7:179–188, 1936.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic. Press, San Diego, 1990.

R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

P. Grossmann, O. Stringfield, N. El-Hachem, M. M. Bui, E. Rios Velazquez, C. Parmar, R.TH. Leijenaar, B. Haibe-Kains, P. Lambin, R. J. Gillies, and H. JWL Aerts. Defining the biological basis of radiomic phenotypes in lung cancer. *eLife*, 6:e23421, 2017.

U. Indahl and T. Næs. A variable selection strategy for supervised classification with continuous spectroscopic data. *Journal of Chemometrics*, 18(2):53–61, 2004.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

M. Kirienko, L. Cozzi, A. Rossi, E. Voulaz, L. Antunovic, A. Fogliata, A. Chiti, and M. Sollini. Ability of fdg pet and ct radiomics features to differentiate between primary and metastatic lung lesions. *European Journal of Nuclear Medicine and Molecular Imaging*, Apr 2018.

P. Latouche, P.-A. Mattei, C. Bouveyron, and J. Chiquet. Combining a relaxed EM algorithm with Occam's razor for Bayesian variable selection in high-dimensional regression. *Journal of Multivariate Analysis*, 146:177–190, 2016.

K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

K.-A. Lê Cao, P. Martin, C. Robert-Granié, and P. Besse. Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC bioinformatics*, 10(1):34, 2009.

K.-A. Lê Cao, S. Boitard, and P. Besse. Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC bioinformatics*, 12(1): 253, 2011.

M. G. Lubner, A. D. Smith, K. Sandrasegaran, D. V. Sahani, and P. J. Pickhardt. Ct texture analysis: Definitions, applications, biologic correlates, and challenges. *RadioGraphics*, 37(5): 1483–1503, 2017.

C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based discriminant analysis. *Journal of Multivariate Analysis*, 102(10):1374–1387, 2011.

G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

T.B. Murphy, N. Dean, and A.E. Raftery. Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications. *The Annals of Applied Statistics*, 4(1):219–223, 2010.

N. N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817, 2014.

F. Orlhac, B. Thézé, M. Soussan, R. Boisgard, and I. Buvat. Multiscale texture analysis: From 18f-fdg pet images to histologic images. *Journal of Nuclear Medicine*, 57(11):1823–1828, 2016.

J. Pacheco, S. Casado, L. Núñez, and O. Gómez. Analysis of new variable selection methods for discriminant analysis. *Computational Statistics & Data Analysis*, 51(3):1463–1478, 2006.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, 98:10869–10874, September 2001.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. (Statistical Methodology)*, 58(1):267–288, 1996.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

N. Trendafilov and I. Jolliffe. Dalass: Variable selection in discriminant analysis via the lasso. *Computational Statistics & Data Analysis*, 51(8):3718–3736, 2007.

D. Witten and R. Tibshirani. Penalized classification using fisher's linear discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):753–772, 2011.