Stability of feature selection in classification issues for high-dimensional correlated data

Emeline Perthame, Chloé Friguet, David Causeur

Stat Comput (2016) 26:783–796
DOI 10.1007/s11222-015-9569-2

Stability of feature selection in classification issues for high-dimensional correlated data

Émeline Perthame[1] · **Chloé Friguet**[2] · **David Causeur**[1]

Received: 25 August 2014 / Accepted: 9 April 2015 / Published online: 31 May 2015

Abstract Handling dependence or not in feature selection is still an open question in supervised classification issues where the number of covariates exceeds the number of observations. Some recent papers surprisingly show the superiority of naive Bayes approaches based on an obviously erroneous assumption of independence, whereas others recommend to infer on the dependence structure in order to decorrelate the selection statistics. In the classical linear discriminant analysis (LDA) framework, the present paper first highlights the impact of dependence in terms of instability of feature selection. A second objective is to revisit the above issue using a flexible factor modeling for the covariance. This framework introduces latent components of dependence, conditionally on which a new Bayes consistency is defined. A procedure is then proposed for the joint estimation of the expectation and variance parameters of the model. The present method is compared to recent regularized diagonal discriminant analysis approaches, assuming independence among features, and regularized LDA procedures, both in terms of classification performance and stability of feature selection. The proposed method is implemented in the `R` package `FADA`, freely available from the `R` repository `CRAN`.

Keywords Variable selection · High dimension · Stability · Classification · Discriminant Analysis

✉ David Causeur
david.causeur@agrocampus-ouest.fr

[1] Institut de Recherche Mathématique de Rennes (IRMAR), UMR 6625 du Centre National de la Recherche Scientifique (CNRS), Agrocampus Ouest, 65 Rue de Saint-Brieuc, 35042 Rennes, France

[2] Laboratoire de Mathématiques de Bretagne Atlantique (LMBA), UMR 6205 du Centre National de la Recherche Scientifique (CNRS), University of South Brittany, Bât. Y. Coppens, Campus de Tohannic, 56000 Vannes, France

1 Introduction

High-throughput technologies, increasingly used in diverse contexts such as brain activity modeling, astronomy, or gene expression analysis, share the common property to generate a huge volume of data, which makes possible the large-scale analysis of complex systems. Such data are generally characterized by their high dimension, as the number of features can reach several thousands, whereas the sample size is usually about some tens. More and more authors also point out their heterogeneity, as the true signal and some confusing factors (uncontrolled and unobserved) are often observed at the same time. For example, in Omics data used for quantitative issues in systems biology, both these factors and the joint contribution of subsets of features to common biological pathways generate a biologically meaningful dependence structure among features. The impact of such a dependence on the performance of the supervised classification procedures which are used to predict the class of a biological sample from its genomic profile is still questioning.

Recent advances on the impact of dependence on the performance of supervised classification methods in situations where the number of covariates is much larger than the number of sampling items have led to apparently contradictory conclusions. Indeed, the superiority of approaches based on an erroneous independence assumption is reported (Dudoit et al. 2002; Levina 2002; Bickel and Levina 2004), whereas more and more methods account for the covariance structure (Guo et al. 2007; Dabney and Storey 2007; Xu et al. 2009; Zuber and Strimmer 2009). More recently, Ahdesmäki and Strimmer (2010) gives more insight to this issue by revisiting the naive Bayes approach of Efron (2008) called diagonal discriminant analysis (DDA), using decorrelated individual scores. In the DDA framework in which independence among features is assumed, finding the support of the sig-

nal, namely the subset of truly informative covariates, shows some similarity with large-scale significance study since it consists in ranking individual scores. However, as recalled by Ahdesmäki and Strimmer (2010), whereas multiple testing procedures aim at controlling the number of false discoveries, this is usually more relevant for selection issues to control the number of erroneously non-selected features. Interestingly, in this multiple testing context, some papers (Leek and Storey 2007, 2008; Efron 2007; Friguet et al. 2009; Sun et al. 2012) have also reported the negative impact of large correlation among scores on the consistency of the ranking of p values. The authors propose to handle this correlation in a joint modeling of the relationships between features and covariates and residual variance–covariance using a flexible model which assumes that latent effects can linearly affect the dependence among features. The present paper introduces a specific procedure for the supervised classification issue.

The first objective of the present paper is to illustrate the instability of variable selection in the classical linear discriminant analysis (LDA) context, when the number of covariates exceeds the number of observations. For such high-dimensional issues, regularized procedures based on ℓ_1 or ℓ_2 penalization of usual loss functions are now well established to handle efficiently a bias-variance trade-off for the estimation of the discriminant scores [see for example Tibshirani et al. (2002, 2003) for a regularized DDA, Hastie et al. (1995) for a penalized LDA or Friedman et al. (2010) for an elastic net penalization of deviance-based estimation]. The stability and classification performance of some of these usual procedures are investigated and the impact of dependence on their repeatability properties is studied.

Section 2 introduces the context of feature selection for high-dimensional supervised classification in a normal framework, focusing on the two-class issue. A regression factor model is proposed to identify a low-dimensional linear kernel which captures data dependence. Some analytical properties are derived and a new strategy for model selection is deduced. This approach is described in Sect. 3. Sections 4 and 5 investigate the properties of variable selection procedures for high-dimensional data, considering different structures for dependence and real data. The improvements brought by the proposed approach in terms of stability and classification performance are highlighted.

2 High-dimensional variable selection for classification

In order to highlight the selection stability issue, we intentionally focus hereafter on two-class prediction in a normal setting with equal covariance in both groups. However, the general principles of our approach are applicable in the wider framework of more than two classes or unequal covariance structures.

2.1 Notation

Let $x \in \mathbb{R}^m$ denote a vector of explanatory variables. The response is a two-class variable denoted Y, with prior probabilities $p_1 = \mathbb{P}(Y = 1)$ and $p_0 = \mathbb{P}(Y = 0) = 1 - p_1$. It is assumed that x is normally distributed with mean μ_1 if $Y = 1$, and μ_0 otherwise. For both group, the positive within-group variance–covariance matrix is Σ.

$$x = \mu_y + e; \quad \text{with } y = 1 \quad \text{if } Y = 1 \text{ and} \\ y = 0 \quad \text{otherwise}, \tag{1}$$

where e is a random error normally distributed with mean 0 and covariance Σ given Y.

The sample consists of n independent joint observations (x_i', Y_i), $i = 1, \ldots, n$, of the explanatory and response variables. In the present high-dimensional framework, n can be much smaller than m. Hereafter, n_1 (resp. $n_0 = n - n_1$) denotes the number of observations in the sample for which $Y = 1$ (resp. $Y = 0$).

2.2 Bayes consistency and usual estimation procedures

In the present multivariate normal situation, it is well known that the log-ratio $\mathrm{LR}(x)$ of posterior class probabilities given x is a linear function of the explanatory profiles:

$$\mathrm{LR}(x) = \log \frac{\mathbb{P}_x(Y = 1)}{\mathbb{P}_x(Y = 0)} = \beta_0^* + x'\beta^* \tag{2}$$

where $\beta^* \in \mathbb{R}^m$ and $\beta_0^* \in \mathbb{R}$ are closed-form functions of the conditional moments of x given Y:

$$\beta_0^* = \log \frac{p_1}{p_0} - \frac{1}{2}\left(\mu_1' \Sigma^{-1} \mu_1 - \mu_0' \Sigma^{-1} \mu_0\right) \tag{3}$$

$$\beta^* = \Sigma^{-1}\left(\mu_1 - \mu_0\right). \tag{4}$$

The above settings therefore provide a natural framework in which linear classifiers, namely functions $x \mapsto \beta_0 + x'\beta$, can be used to predict the group variable Y given the explanatory profile x. The classification rule consists, for a given x, in predicting Y by $\hat{Y} = 1$ if $x'\beta$ exceeds a threshold c and by $\hat{Y} = 0$ otherwise. Deduced from the decision theory and firstly considered as a heuristic classification rule, the Bayes classifier is the linear predictor with minimal classification error $\pi(\beta; c)$:

$$\pi(\beta; c) = \mathbb{P}(\hat{Y} \neq Y)$$

$$
\begin{aligned}
&= \mathbb{P}(x'\beta \le c | Y = 1) \times p_1 \\
&\quad + \mathbb{P}(x'\beta > c | Y = 0) \times p_0 \\
&= \left[1 - \Phi\left(\frac{\mu_1'\beta - c}{(\beta'\Sigma\beta)^{1/2}}\right)\right] p_1 \\
&\quad + \Phi\left(\frac{\mu_0'\beta - c}{(\beta'\Sigma\beta)^{1/2}}\right) p_0. \qquad (5)
\end{aligned}
$$

It is straightforward checked that the slope and threshold of the linear Bayes classification rule are given by $\beta = \beta^*$ and $c = -\beta_0^*$. Let γ denote the following function:

$$
\begin{aligned}
\gamma : \ \Delta \mapsto \gamma(\Delta) = &\left[1 - \Phi\left(\frac{1}{\Delta}\log\frac{p_1}{p_0} + \frac{\Delta}{2}\right)\right] p_1 \\
&+ \Phi\left(\frac{1}{\Delta}\log\frac{p_1}{p_0} - \frac{\Delta}{2}\right) p_0, \qquad (6)
\end{aligned}
$$

where Φ is the cumulative Gaussian distribution function. Function γ represents the classification error for Mahalanobis distance Δ between 2 groups. One can notice that the minimal probability of misclassification for the linear Bayes classifier π^* can be written as $\pi^* = \gamma(\Delta_\Sigma)$, where Δ_Σ stands for the Mahalanobis distance between μ_1 and μ_0 with metric Σ: $\Delta_\Sigma^2 = (\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)$. Bayes consistency is defined as the asymptotic achievement of this optimal classification performance.

Apart from the choice of c which generally aims at a compromise between false discovery and false non-discovery rates, deriving a linear classification procedure can be viewed as an estimation issue for β. Among the most two famous methods, the so-called Fisher linear discriminant analysis is obtained by minimizing the least-squares criterion:

$$
(\hat{\beta}_0, \hat{\beta})_{\text{LDA}} = \underset{\beta_0,\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[V_i - (\beta_0 + x_i'\beta)\right]^2, \qquad (7)
$$

where V is defined as a symmetric recoding of Y:

$$
V = \begin{cases} 1 & \text{if } Y = 1 \\ -1 & \text{if } Y = 0. \end{cases}
$$

The above optimization issue has a closed-form solution which coincides with the moment estimator of (β_0^*, β^*). In particular, provided the sample within-group covariance matrix S of the explanatory variables is not singular:

$$
\hat{\beta}_{\text{LDA}} = S^{-1}(\bar{x}_1 - \bar{x}_0), \qquad (8)
$$

where $\bar{x}_0$ and $\bar{x}_1$ are the sample means in each group.

Another famous method is logistic regression which provides an alternative maximum likelihood estimation procedure:

$$
\begin{aligned}
(\hat{\beta}_0, \hat{\beta})_{\text{ML}} &= \underset{\beta_0,\beta}{\operatorname{argmin}} -2\sum_{i=1}^{n} \log\Big[1 + \exp\big(-V_i(\beta_0 + x_i'\beta)\big)\Big] \\
&= \underset{\beta_0,\beta}{\operatorname{argmin}} \mathcal{D}(\beta), \qquad (9)
\end{aligned}
$$

where $\mathcal{D}(\beta) = -2\sum_{i=1}^{n} \log\Big[1 + \exp\big(-V_i(\beta_0 + x_i'\beta)\big)\Big]$ is the deviance.

However, the invertibility of the sample covariance matrix S is also required to minimize $\mathcal{D}(\beta)$. This invertibility condition does not hold in a high-dimensional framework.

This issue can be addressed by assuming that the support $I = \{j,\ \beta_j \neq 0\} \subset [\![1; m]\!]$ of the classification model is small regarding the number m of features. Under this assumption of a sparse model, feature selection procedures, which aim at identifying the non-zero coefficients in β, are needed to reduce the explanatory profile to the most group predictive variables.

2.3 Feature selection

There is an abounding statistical literature dealing with the issue of feature selection in regression and classification. Among many other methods, minimization of the Akaike or Bayesian information criteria (AIC, BIC), which are based on a ℓ_0-penalization of the deviance, are frequently used. Indeed, minimization of BIC leads to consistent estimators of the support I and minimization of the AIC to minimax rate optimal rules for estimating the regression function (Yang 2005). The main cause of concern of these procedures in high dimension is of a computational nature, as an exhaustive search through all possible models (2^m) is needed. Stepwise exploration of the whole family of models provides an alternative, but this strategy can be unstable in high dimension because the number of fitted candidate models, at most $m(m+1)/2$, is extremely small regarding the number of possible models.

Alternatively, one can handle the fitting and selection issues at the same time by relaxing the ℓ_0-penalization by the ℓ_1-penalization. This leads to the LASSO estimator $\beta(\lambda)$ of logistic regression parameters (Tibshirani 1996):

$$
\beta(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\mathcal{D}(\beta) + \lambda \sum_{j=1}^{m} |\beta_j|\right) \qquad (10)
$$

where the tuning parameter λ is chosen to control the sparsity of the estimator: larger values of λ lead to more zero components in $\beta(\lambda)$. The choice of the tuning parameter can be achieved by minimization of the cross-validated residual deviance or misclassification rate, as implemented in the `R` package `glmnet` (Friedman et al. 2010). LASSO is computationally feasible for large m as the optimization problem

in (10) is convex. For variable selection or prediction purpose, two-stage procedures such as adaptive LASSO (Zou 2006) can be applied, generally improving the control of the number of false positives, but at the cost of a lack of power.

As mentioned by Van de Geer (2010), LASSO makes strong assumptions on the covariance matrix, mainly that correlations between variables are weak. Consequently, a major and still open question remains the application of the procedure while coping with large correlations between variables.

Let us illustrate the impact of dependence on the stability of a standard variable selection procedure (LASSO) by a simulation study, comparing the dependent and independent cases.

In the dependent case, let us consider a two-group variable Y, taking the value 0 for $n_0 = 30$ sampling items and 1 for the $n_1 = 30$ other items. A $(n_0 + n_1) \times m$ dataset, with $m = 500$, of normal $m-$profiles x is generated, with mean $\mu_0 = 0$ for the sampling items with $Y = 0$, and the components of μ_1 are also zero except for the $m_1 = 100$ last variables of the profile, for which the mean is $\delta = 0.74$. The former value of δ guarantees a reasonable power of 0.8 for the t test of mean comparison in the two groups. The within-group standard deviations of the explanatory variables are set to 1 and the within-group correlation matrix is a five-factor model $\Sigma = \Psi + BB'$, where Ψ is a diagonal matrix of specific variances and B a $m \times q$-matrix of loadings ($q = 5$). The values in B and Ψ are chosen so that the resulting correlations are strong, as shown by the histogram of correlations in Fig. 1a. The slope coefficients $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$ displayed in Fig. 1b are straightforward deduced from the above settings. Note that $|\beta|$ defines a natural rank among features: it is indeed expected that the features with largest coefficients are selected more often.

The same simulation setting is used for the independent case, except that the within-group correlation matrix is here $\mathbb{I}_m$. Besides, $\mu_1 = \beta$ to keep the same β as in the dependent case.

For each case, 1000 datasets are simulated. The same LASSO selection procedure is implemented using `glmnet`, where the penalty parameter is selected by minimization of a ten fold cross-validation residual deviance. Histograms of the numbers of selected features in both scenarios of dependence are reproduced in Fig. 2. The rank in $|\beta|$ of each selected feature is also deduced and the accuracy of the selection is assessed by the mean rank of the subset of selected features. Histograms of these mean ranks statistics are also provided in Fig. 3.

The first striking impact of dependence is related to the number of selected features (Fig. 2), which is much larger when the features are correlated. Moreover, whereas in the independent case, no erroneous selection of null features is

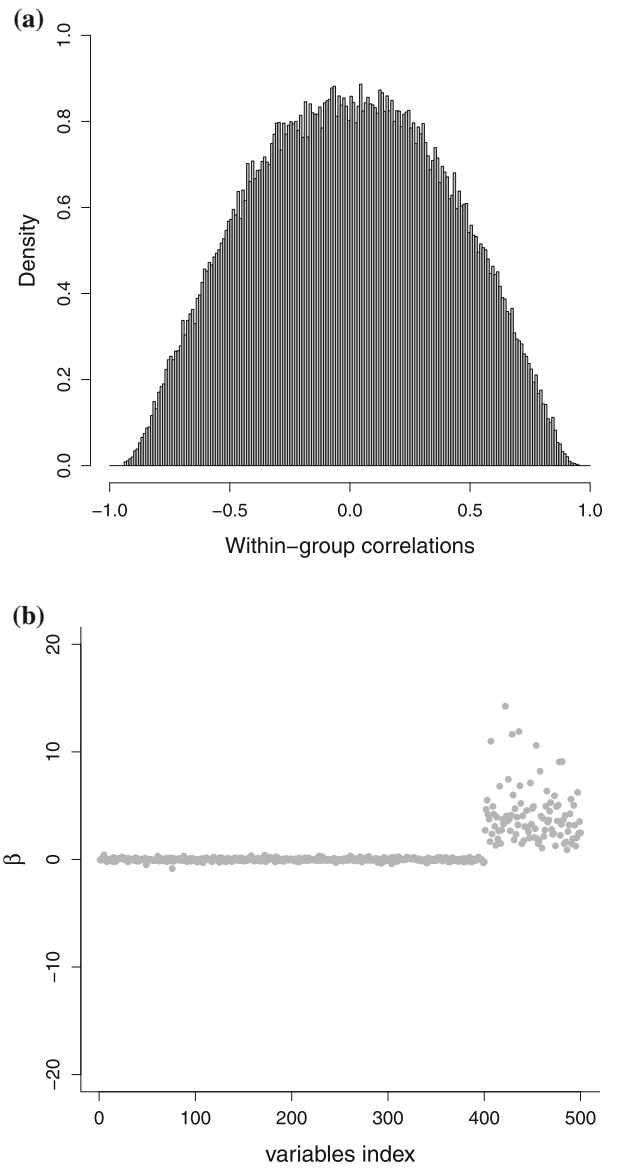


Fig. 1 Simulation settings. **a** Within-group correlations for the dependent case; **b** slope coefficients of the classification model

reported in the simulations, in 12.1 % of the simulations under dependence, the FDP is non-zero. Accuracy of selection is also clearly affected by dependence: the mean ranks in the independent case are consistent with the expected means if the most group predictive variables are selected (Fig. 2), namely half the number of selected features, whereas these mean ranks are much larger in the dependent case (Fig. 3a).

3 Factor-adjusted variable selection

We propose a framework in which dependence is tractable at the level of the original data, which allows a direct adjustment of the data for that dependence. This dependence adjustment

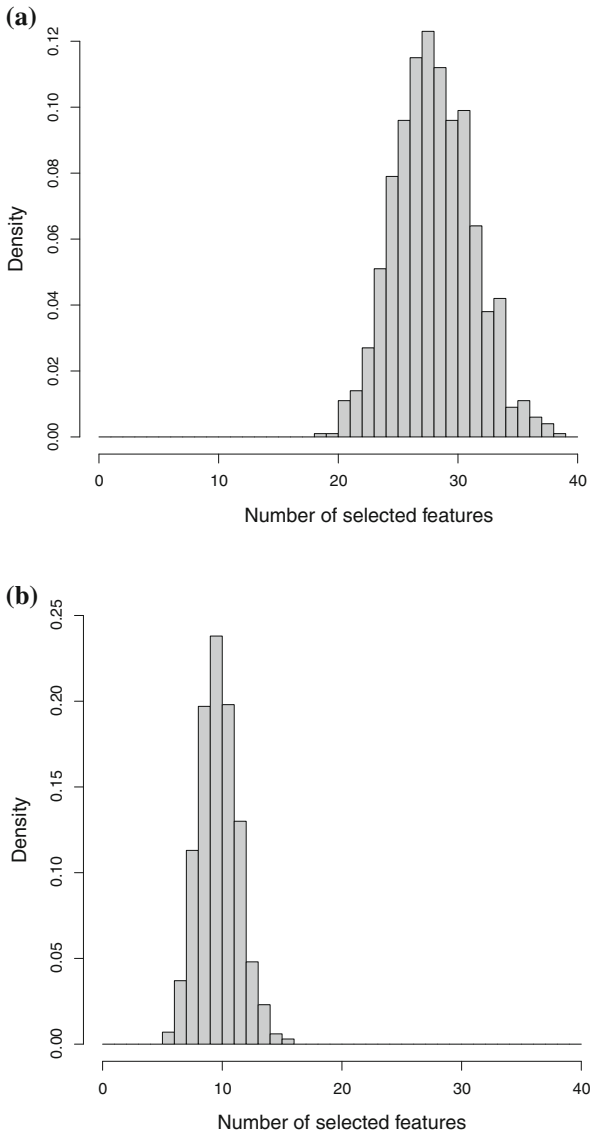


Fig. 2 Simulation study—number of selected features. **a** Under dependence; **b** under independence

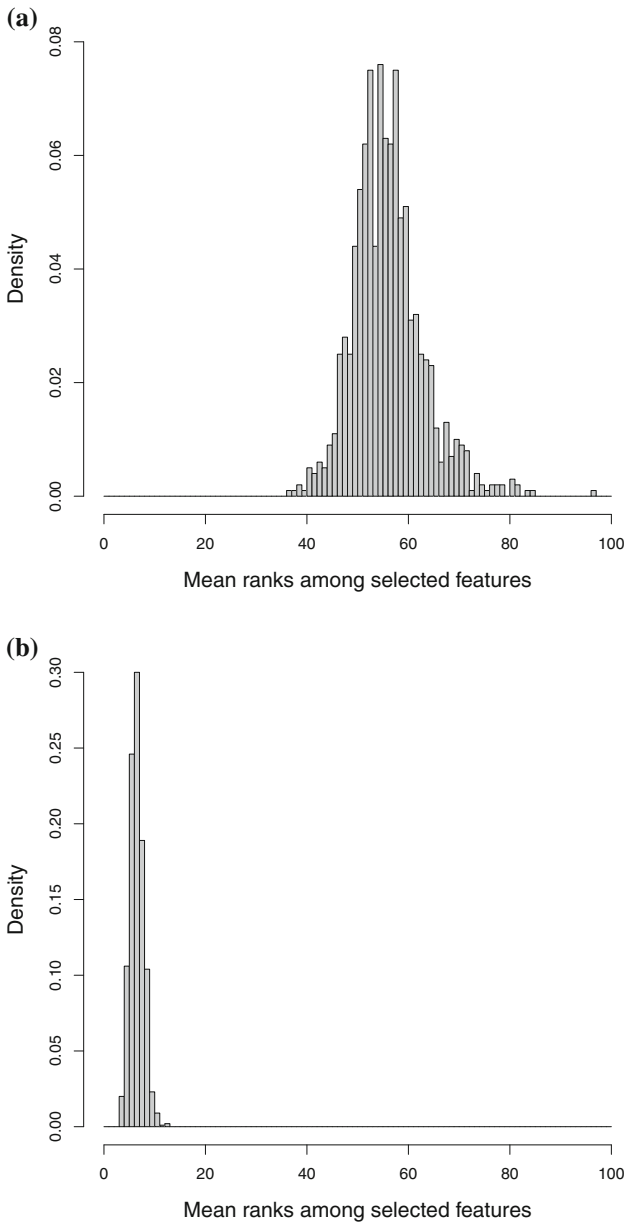


Fig. 3 Simulation study—mean ranks of the selected features in $|\beta|$. **a** Under dependence; **b** under independence

step can be combined to any selection procedures, as proposed in the comparative studies of Sect. 5.

3.1 Factor adjustment

In many areas, and particularly in the analysis of gene expression data (Kustra et al. 2006; Leek and Storey 2008; Carvalho et al. 2008; Friguet et al. 2009; Teschendorff et al. 2011; Sun et al. 2012), it has become frequent to cope with dependence by assuming the existence of a moderate number of latent factors conditionally on which it is assumed that features are independent. The main advantage of such an approach is that dependence is captured into a low-dimensional linear space. Then the statistical procedures initially designed for the independent (or weak correlation) case can be applied to the decorrelated data, obtained after adjustment for the latent effects. Several methods have been proposed to model the latent factors, such as (Independent) surrogate variable analysis (Leek and Storey 2007; Teschendorff et al. 2011), independent component analysis (Lee and Batzoglou 2003), latent-effect adjustment after primary projection (Sun et al. 2012) or factor analysis (Friguet et al. 2009) for example.

Hereafter, we introduce a supervised Factor Analysis model for classification. Based on this model, the conditional

linear Bayes classifier is defined and the conditional Bayes consistency of the factor-adjusted approach is proved.

3.2 A flexible framework for dependence

Latent effects models are used for many years in economics, social sciences, and psychometrics, originally in the field of intelligence research (Spearman 1904) and has appeared recently in the study of the dependence structure of high-dimensional data, such as those provided by microarray technology (Pournara and Wernisch 2007; Kustra et al. 2006; Blum et al. 2010). The model defined in (1) can indeed take advantage of a flexible parameterization of the within-group covariance matrix Σ. In practice, and especially in gene expression data for example, unmodeled and/or uncontrolled factors can interfere with the true signal, which introduces heterogeneity in the data and generates dependence across the variables. Residual e in model (1) is then split into two terms, one associated to heterogeneity components through latent variables Z, and independent residuals ε:

$$x = \mu_y + BZ + \varepsilon; \quad \begin{array}{ll} \text{with } y = 1 & \text{if } Y = 1 \\ \text{and } y = 0 & \text{otherwise,} \end{array} \tag{11}$$

where ε is a random vector with independent normal components $\varepsilon_j \sim \mathcal{N}(0, \psi_j^2)$ and B is a $m \times q$ matrix of loadings. Hence, $\mathbb{V}(\varepsilon) = \Psi = diag(\psi_j^2, 1 \leq j \leq m)$.

Model (12) establishes the existence of q latent variables $Z = [Z_1, \ldots, Z_q]'$ which capture the dependence among the m variables in a $q-$dimensional linear space, with $q \ll m$. Such model is called a regression Factor Analysis model (Carvalho et al. 2008) and the latent variables Z are hereafter called (common) factors. Without loss of generality, in the following, it is assumed that Z is normally distributed with mean 0 and variance $\mathbb{I}_q$. The mixed-effects regression model (12) is equivalently defined as a fixed-effects regression model as in (1) but the residual variance Σ is decomposed into the sum of two components, the diagonal matrix Ψ of specific variances and the common variance component BB':

$$\Sigma = BB' + \Psi. \tag{12}$$

Note that, under the above assumptions, the joint distribution of the factors and the explanatory variables, given Y, is normal:

$$\begin{pmatrix} X \\ Z \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \mu_y \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma & B \\ B' & \mathbb{I}_q \end{pmatrix} \right]. \tag{13}$$

The following linear Bayes classifier, which is optimal conditionally on the explanatory variables and the factors, is straightforward derived from the inversion of the partitioned variance matrix in expression (13):

$$\begin{aligned} \mathrm{LR}(x, z) &= \log \frac{p_1}{p_0} - \frac{1}{2}\left(\mu_1' \Psi^{-1} \mu_1 - \mu_0' \Psi^{-1} \mu_0\right) \\ &\quad + (x - Bz)' \Psi^{-1} \left(\mu_1 - \mu_0\right). \end{aligned} \tag{14}$$

It turns out that the conditional linear Bayes classifier (14) depends on x and z through the factor-adjusted explanatory variables $x - Bz$, which confirms that, assuming the factor structure is known, the best linear classifier is just the usual linear Bayes classifier based on the factor-adjusted explanatory profiles.

The minimal probability of misclassification for the conditional linear Bayes classifier is $\pi_z^* = \gamma(\Delta_\Psi)$, where γ is defined in (6) and Δ_Ψ stands for the Mahalanobis distance between μ_1 and μ_0 with metric Ψ: $\Delta_\Psi^2 = (\mu_1 - \mu_0)' \Psi^{-1} (\mu_1 - \mu_0)$. If $B^\star = \Psi^{-1/2} B$ stands for the normalized loadings of the factor model, the following inequalities hold:

$$\frac{1}{1 + \rho_{\max}^2} \leq \frac{\Delta_\Sigma^2}{\Delta_\Psi^2} \leq 1, \tag{15}$$

where $\rho_{\max}$ is the largest singular value of $B^\star$. As γ is a decreasing function of Δ, it is deduced from the right inequality in (15) that $\pi_z^* \leq \pi^*$. Moreover, the left inequality shows that the gain which can be expected by the conditional approach is increasing with $\rho_{\max}^2$, which is also the largest eigenvalue of $B' \Psi^{-1} B$. In other words, this expected gain is larger in situations of strong dependence, in which the loadings take large values with respect to the corresponding specific variances.

Note that the Bayes classifier general optimality, which is established without any assumption on Σ, is not questioned here. However, under the assumption of a factor model for Σ, the above result establishes the theoretical superiority of a conditional approach based on the factor-adjusted explanatory variables $x - Bz$. Consequently, we propose hereafter an estimation procedure for the regression factor model (12).

3.3 An iterative estimation procedure for the supervised factor model

We propose an iterative method, which alternates the estimation of μ_0, μ_1, B and Ψ, and the derivation of the latent factors Z.

3.3.1 Initialization

The algorithm starts with $\hat{\mu}_0 = \bar{x}_0$, $\hat{\mu}_1 = \bar{x}_1$. Based on these estimates of the group means, the centered profiles $x - \hat{\mu}_y$ are used to estimate B and Ψ, using the EM algorithm detailed

in Friguet et al. (2009). The corresponding estimators are hereafter denoted $\hat{B}$ and $\hat{\Psi}$.

3.3.2 Step 1: factors extraction (Z)

Thompson's method to derive the factors is adapted to the present regression factor model. It is indeed deduced from the joint multivariate normal distribution of the explanatory variables and the factors (see expression (13)) that the conditional expectation of the factors, given x, is given by:

$$\mathbb{E}_x(Z) = (I_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}\Big(x - \big[\mu_0\mathbb{P}_x(Y=0) + \mu_1\mathbb{P}_x(Y=1)\big]\Big), \quad (16)$$

where

$$\mathbb{P}_x(Y=1) = 1 - \mathbb{P}_x(Y=0) = \frac{1}{1+\exp(-\beta_0^* - \beta^{*'}x)}.$$

3.3.3 Remarks about the implementation

1. Note that the calculation of β_0^* and β^*, using expressions (3) and (4), only involves the inversion of a $q \times q$-matrix according to the Woodbury's identity:

$$\Sigma^{-1} = \Psi^{-1} - \Psi^{-1}B(I_q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}.$$

2. Besides, the plug-in estimator of $\mathbb{P}_x(Y=1)$ can be affected if the factor model is over-fitted, which penalizes the classification performance. Alternative estimation procedures can therefore be preferred to estimate $\mathbb{P}_x(Y=1)$, such as ℓ_1-penalized logistic regression, which introduces sparsity to reduce the effects of over-fitting.

Therefore, estimated factors $\hat{Z}$ are derived by plugging-in $\hat{\mu}_0$, $\hat{\mu}_1$, $\hat{B}$ and $\hat{\Psi}$ into expression (16):

$$\hat{Z} = (I_q + \hat{B}'\hat{\Psi}^{-1}\hat{B})^{-1}\hat{B}'\hat{\Psi}^{-1}\Big(x - \big[\hat{\mu}_0\hat{\mathbb{P}}_x(Y=0) + \hat{\mu}_1\hat{\mathbb{P}}_x(Y=1)\big]\Big). \quad (17)$$

3.3.4 Step 2: model parameters estimation (μ_0, μ_1, B and Ψ)

The estimations of μ_0 and μ_1 are updated by the least-squares fitting of the multivariate regression model (12), where Z is replaced by $\hat{Z}$. The factor decomposition of the centered profiles $(x - \hat{\mu}_y)$ covariance provides updated estimates of B and Ψ.

3.3.5 Iterations and stop criterion

Steps 1 and 2 are iterated, updating alternatively factors and model parameters estimations. The algorithm stops when two successive estimates of the factor model parameters are similar.

Therefore, the proposed strategy consists in defining factor-adjusted versions of usual classification methods by applying these methods on the factor-adjusted data $x - \hat{B}\hat{Z}$.

A crucial point in the present feature selection context is the choice of the proper number of factors. Indeed, an over-estimation of q would artificially reduce the estimation of the residual specific variances $\hat{\Psi}$, which could generate false positives. In a multiple testing context, Friguet et al. (2009) notice that the variance of the number of false positives is an increasing function of the amount of dependence among the test statistics and give a closed-form expression for the variance inflation $\mathcal{V}_k$ due to the k-factor model for this dependence. Consequently, they suggest an ad hoc procedure which consists, for each k-factor model (Ψ_k, B_k), to estimate the variance of the number of false positives when the tests are calculated with the k-factor-adjusted residuals: $\hat{e} - \hat{Z}_k\hat{B}_k$.

The algorithm described in this section is implemented in the `R` package `FADA` available from the `R` repository `CRAN`, providing functions for decorrelation, feature selection, and estimation of a classification model.

In the following two sections, we illustrate, on real data and by simulations, that this new factor adjustment algorithm improves variable selection, both in terms of classification or prediction performance and reproducibility of the selected variables.

4 Stability of variable selection in high dimension

4.1 DNA microarray data

In genomics, microarrays let biologists measure expression levels for thousands of genes in a single sample all at once. The level of measured gene expressions is influenced both by a biological trait of interest and by unwanted technical and/or biological factors, referred to as heterogeneity factors (Leek and Storey 2007, 2008). Moreover, it is now widely considered that groups of genes contributing to some few biological processes can show co-expression patterns: some genes are activators or inhibitors of others. This motivates the emerging issue of gene co-expression network inference from microarray data. In such context, dealing with dependence is a major concern in carrying statistical analyses.

Feature selection is increasingly common in genomic data analysis to identify genes which expression patterns have meaningful biological links with a phenotypic trait.

Therefore, as an illustration of selection issues in high dimension, let us consider the microarray experiment detailed in Hedenfalk et al. (2001), which is commonly used in the statistical literature for comparative studies of high-dimensional statistical procedures.

4.1.1 Data: breast cancer study

The data were primarily analyzed in order to compare expressions of three types of breast cancer tumor tissues: `BRCA1`, `BRCA2`, and `Sporadic`. The raw expression data, downloaded from http://research.nhgri.nih.gov/microarray/NEJM_Supplement/, initially consist of 3226 genes in 22 arrays; seven arrays from the `BRCA1` group, eight from the `BRCA2` group, and six from the `Sporadic` group. The label of one sample being unclear, it has been removed from the study. 196 genes presenting some suspicious levels of expression (larger than 10 or lower than 0.1) are removed and the data are finally $\log_2$ transformed. In the following, we focus on the selection of gene expressions among the $m = 3030$ included in the study that best predict the two types of tumors `BRCA1` and `BRCA2`. The sample size is then $n = 15$.

4.1.2 Methods

Variable selection is performed using the `R` package `glmnet` (Friedman et al. 2010) which provides a function to fit a two-group logistic regression model via ℓ_1-regularized maximum likelihood (Tibshirani 1996). The sample being small, the choice of the tuning parameter is done thanks to Leave-One-Out cross-validation. LASSO is known to be non consistent when performed on correlated data (Bach 2008). However, the following example aims to illustrate how a lack of stability can be observed on real data and how factor adjustment can stabilize a usual selection procedure.

The procedure is first applied on the complete dataset (with $n = 15$ observations). The performance of the procedure is evaluated through the number of selected variables and the cross-validation error.

Then to illustrate instability of variable selection, the same procedure is applied, removing successively one of the observations. The aim is to evaluate the sensibility of the procedure to changes in the data. The results of the selection procedure are compared to those obtained on the complete data considering the number of selected variables and the overlap with the subset of variables initially selected using the complete data.

Finally, the same procedure is applied on the factor-adjusted data. Factor adjustment is performed with the method presented in Sect. 3.3. The minimization of the variance inflation criterion suggests to keep $q = 1$ common factor for the complete data and for each incomplete dataset.

Table 1 Selection procedure on the complete dataset

Data	Features	Prediction error
Raw data	11	0.400
Factor-adjusted data	8	0.267

4.1.3 Results

Selection procedure on the complete dataset The results of the selection procedure on the complete dataset (for raw and factor-adjusted data) are reported in Table 1. The number of features selected by the LASSO procedure is lower when considering the factor-adjusted data. Moreover, the decorrelation step of factor adjustment leads to a better performance of the selection procedure, *i.e.* a lower prediction error. In the following, I_{raw} (resp. I_{FA}) denotes the subset of selected features when the selection procedure is applied on the complete raw (resp. complete factor-adjusted) data.

Selection procedure on the incomplete datasets The selection procedure is then applied on the 15 sub-datasets, removing successively one of the observations from the complete raw data (resp. complete factor-adjusted data). Table 2a (resp. 2b) reports the number of selected features, the number and proportion of selected variables which belongs to I_{raw} (resp. I_{FA}) and cross-validated prediction error of the selection procedure for each sub-dataset. For each criterion, the tables report the results after the removal of the first four and last four observations as an overview of results, as well as the mean and standard deviation in the last column. Results for all observations are not presented to avoid overloading.

A wide range of situations are reported in Table 2a, regarding both the number and the set of selected features, depending on which observation has been removed. Each observation has therefore a strong influence on the stability of the selection procedure.

For instance, the LASSO procedure seems to be very sensitive to removing the first observation as only one feature is selected instead of 11 for the complete data. Among the six variables selected after removing observation 14, only three are part of I_{raw}. This phenomenon becomes less pronounced when the procedure is applied on the factor-adjusted data (Table 2b). In this case, there is a much higher proportion of selected variables included in I_{FA} (38.2 vs. 70.8 %) and cross-validation errors are smaller.

Conclusion This illustrative situation shows that the usual statistical approaches for variable selection, such as LASSO selection here, are questioned for dependent high-dimensional data analysis. A small change in the data, just considering the removal of one observation, induces variability in the performance of the procedure and leads to different sets of selected

Table 2 Selection procedure after having removed one observation

Removed id.	1	2	3	4	…	12	13	14	15	Mean (SD)
(a) Raw data										
Features	1	10	7	8	…	6	12	6	6	6.4 (3.6)
Included (N)	1	9	3	6	…	6	6	3	5	4.2 (2.5)
Included (%)	9.1	81.8	27.3	54.5	…	54.5	54.5	27.3	45.5	38.2 (22.3)
Prediction error	0.571	0.214	0.286	0.214	…	0.214	0.357	0.214	0.357	0.3 (0.138)
(b) Factor-adjusted data										
Features	9	7	9	10	…	9	7	12	8	7.9 (2.5)
Included (N)	5	7	6	8	…	7	6	8	7	5.7 (2)
Included (%)	62.5	87.5	75.0	100.0	…	87.5	75.0	100.0	87.5	70.8 (24.9)
Prediction error	0.357	0.214	0.286	0.071	…	0.357	0.357	0.214	0.214	0.229 (0.115)

Features number of selected features; *included* (*N*) number of stable inclusions:number of selected variables which belongs to I_{raw} or I_{FA}; *included* (%) proportion of stable inclusions; *prediction error* cross-validated prediction error

variables. Factor adjustment helps to block such effects of heterogeneity and improves both the stability of the set of selected variables and the prediction error.

4.2 DNA methylation data

Recently, DNA methylation data have focused the attention of biologists because new biological processes can be identified from the analysis of such data. In this section, a study is conducted to highlight the contribution of factor adjustment for the analysis of data generated by such experiments.

4.2.1 Data: gastric tumors study

The data were primarily published in Zouridis (2012) and initially consist of 27578 DNA methylation measures and 297 observations. 2573 variables were removed because of missing data so that the studied dataset has 25,005 columns. The binary response variable codes for gastric tumors (203 cases) and gastric non-malignant samples (94 cases).

4.2.2 Methods

According to the simulation study in Sect. 5, shrinkage discriminant analysis (SDA) appears to be the most efficient method regarding the prediction error and the precision of selection. Thus, SDA is conducted on the whole dataset using the `R` package `sda`. The results are compared to the following three-step procedure. (1) A decorrelation step is performed on the whole dataset using `FADA` `R` package then, (2) to decrease the dimension of the dataset and to avoid high computation time in step (3), a rough selection is performed through standard *t* tests on decorrelated data and the first 3000 CpG sites are selected for the next step. (3) Variable selection and classification model are finally performed by SDA on the factor-adjusted sub-dataset. Prediction errors are computed through a tenfold cross-validation with 20 repetitions so that the model is estimated on 200 splits of the data.

Table 3 Nb. of selected features and estimated prediction errors for gastric tumors data

Method	Nb. features	Error rate
SDA	2638	0.0301
Factor-adjusted SDA	305	0.0217

4.2.3 Results

Ten factors are extracted from the whole dataset for factor adjustment at step (1). On the sub-dataset composed of the first 3000 CpG sites, one factor is extracted [step (2)]. Table 3 reports the prediction error and the precision of selection for the two compared procedures. When applied on factor-adjusted data, SDA leads to slightly lower prediction error rate than standard SDA but, most importantly, less variables are selected to achieve this precision.

5 Impact of the dependence design: a simulation study

In order to study the performance of factor adjustment for classification and variable selection, we propose a more intensive simulation study. Considering several scenarios of dependence between variables [independence, block dependence, factor structure, and Toeplitz design, in the manner of Meinshausen and Bühlmann (2010)], some well-known classification methods are applied on simulated datasets. The stability of original procedures is compared to their factor-adjusted versions.

5.1 Simulation design

Let us consider datasets simulated according to a multivariate normal distribution, each dataset being composed of $m = 1000$ variables and $n = 30$ observations. Besides, let us consider a binary variable Y such that the observations are split into two arbitrary groups of size $n_0 = n_1 = n/2$. The m-dimensional profiles X are normally distributed with mean $\mu_0 = 0_m$ in the first group ($Y = 0$), where $0_m \in \mathbb{R}^m$ is the zero vector, and μ_1 in the second group ($Y = 1$). A subset I of 50 variables is randomly chosen to be group predictive. For these variables, μ_1 has non-zero components: $\mu_{1j} = \delta$ for $j \in I$ and $\mu_{1j} = 0$ otherwise. The value of δ is set to 0.55 or 0.47, which matches with high and moderate signal strength, as introduced by Donoho and Jin (2008).

Thousand datasets are simulated considering each of the four following scenarios for the covariance matrix Σ:

(A) The m variables are normally and independently distributed with variance 1 so that Σ is the m-diagonal matrix $\mathbb{I}_m$. This scenario is used as a control situation to check that the proposed method does not falsely catch dependence;

(B) Σ is a two-blocks matrix. Correlation between the first 100 variables is set to 0.7 and correlation between the remaining 900 variables is equal to 0.3. This correlation matrix is used to study impact of dependence in multiple testing in the context of gene expression analysis in Zuber and Strimmer (2009);

(C) Σ is decomposed into a specific and a common part as in a factor model (see Sect. 3.2): $\Sigma = BB' + \Psi$. Ψ is a diagonal matrix of specific variances and B is a $m \times q$-matrix of coefficients, chosen so that the proportion $\text{trace}(BB')/\text{trace}(\Sigma)$ of dependence among variables is high (78 %). In the present simulation study, the number of common factors is $q = 5$. Note that the signal is here set to a weaker value $\delta = 0.47$ because generating dependence through a factor structure is a favored scenario.

(D) Σ is a Toeplitz matrix. This kind of design corresponds to auto-regressive time dependence such that the covariance between two variables i and j is equal to $\sigma\rho^{|i-j|}$. In this simulation study, $\sigma = 1$ and $\rho = 0.99$.

5.2 Methods

The following selection procedures are applied on each simulated dataset:

(LASSO) ℓ_1-regularized logistic regression using the `R` package `glmnet` (Friedman et al. 2010);

(SLDA) Sparse linear discriminant analysis, which is an ℓ_1-penalized LDA using the `R` package `SparseLDA` (Clemmensen et al. 2011), the stop parameter was set to 10;

(SDA) Shrinkage discriminant analysis, which is a James–Stein regularized version of LDA, using the `R` package `sda` (Ahdesmäki and Strimmer 2010). Note that SDA consists finally in a correlation adjustment of the scores used for feature selection in DDA;

(DDA) Shrinkage diagonal discriminant analysis, which assumes within-group independence among features, using the `R` package `sda` (Ahdesmäki and Strimmer 2010). Estimation of the DDA model is here regularized using a ridge approach.

Several cutoffs are implemented in the `R` package `sda` to conduct DDA and SDA such as the False Non-Discovery Rate (FNDR) or Higher Criticism (Donoho and Jin 2008). Both lead to similar results in this simulation study and the results reported here concern the FNDR cutoff.

Each procedure is applied both on raw data and on factor-adjusted data, using the decorrelation method presented in Sect. 3.3: for each simulated dataset, covariance parameters Ψ and B and latent factors Z are estimated on training datasets and factor-adjusted training data (decorrelation step) are computed using formula $x - Bz$ introduced in expression (14). Estimates $\hat{\Psi}$ and $\hat{B}$ are used to estimate latent factors of testing data and factor-adjusted testing data are computed in the same way. Classification methods are finally trained on decorrelated training samples and assessed on decorrelated testing sample.

Prediction errors are calculated on an independent balanced test dataset consisting of 10,000 sampling items, generated according to each structure of dependence. Performances of methods are assessed by calculating, for each simulated dataset, the prediction error calculated on the test dataset, the number of selected features, and the proportion of truly selected variables (or positive predictive value, reported hereafter as "precision").

5.3 Results

5.3.1 Cross-validation

Table 4 reports the prediction errors for a no-signal simulation study ($\mu_0 = \mu_1 = 0_m$, covariance pattern set here to two-blocks structure). Results are not overoptimistic as prediction errors are close to 0.5. This insures that all parameters are estimated independently of the test dataset and that selection and parameters estimation are newly performed for each simulated dataset.

Table 4 Check of cross-validated error rates (prediction errors) for a no-signal design

	Raw data	Factor-adjusted data
LASSO	0.4989	0.4990
SLDA	0.4989	0.4992
SDA	0.5000	0.5004
DDA	0.4999	0.4996

Table 5 No factor found for independence design (*A*)

	Prediction error	Features	Precision (%) mean (SD)
LASSO	0.3858	12.82	40.32 (20.96)
SLDA	0.3873	10.00	39.50 (15.33)
SDA	0.3868	35.09	35.52 (21.77)
DDA	0.3489	32.90	38.44 (23.68)

5.3.2 Independence design

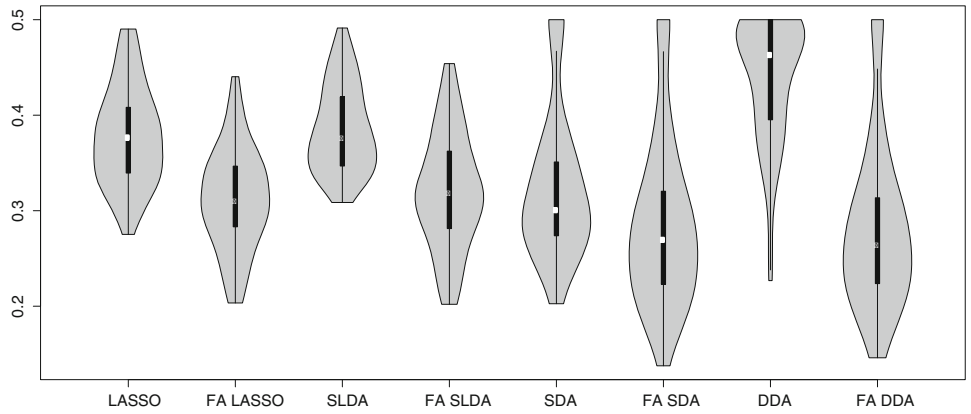
Scenario (*A*) confirms that the factor adjustment is not overoptimistic and does not wrongly locate correlation for an independent design. Indeed, no factor is extracted for the 1000 independently simulated datasets: factor-adjusted methods are therefore similar to their original versions (see Table 5).

5.3.3 Structures with correlations

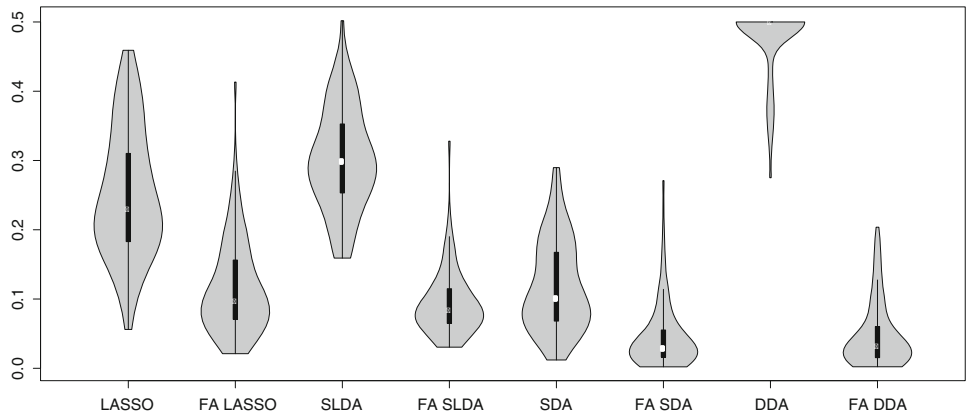
Considering the three scenarios of dependence (*B*), (*C*), and (*D*), Table 6 and Fig. 4 show that the four tested selection procedures (LASSO, Sparse LDA, DDA, and SDA) are improved overall while considering the factor adjustment: error rates are smaller and precisions are greatly improved.

Table 6 Simulation results for several designs of dependence

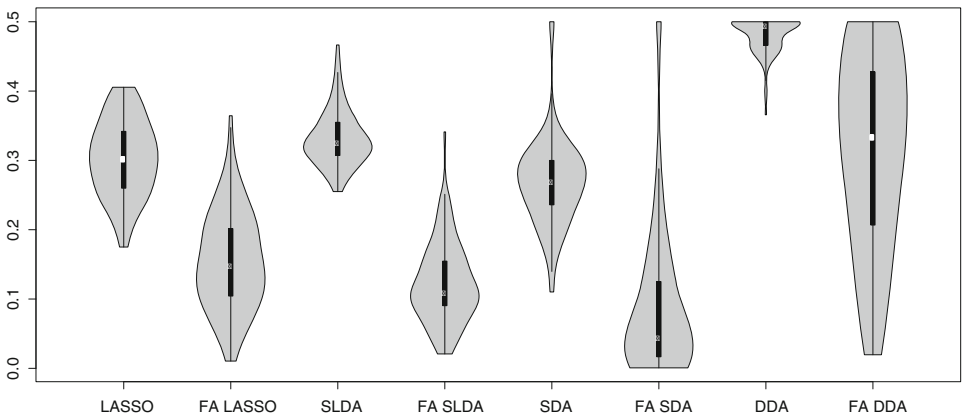
Method	Prediction error	Features	Precision (%) mean (SD)
Block structure (*B*)			
LASSO	0.3780	12.64	40.05 (23.85)
Factor-adjusted LASSO	0.3118	15.44	49.16 (21.30)
SLDA	0.3872	10.00	39.80 (15.50)
Factor-adjusted SLDA	0.3426	10.00	50.80 (16.00)
SDA	0.3244	41.63	42.12 (17.77)
Factor-adjusted SDA	0.2863	44.19	42.46 (18.08)
DDA	0.4393	165.10	28.31 (24.46)
Factor-adjusted DDA	0.2820	48.44	42.13 (19.14)
Factor structure (*C*)			
LASSO	0.2660	14.025	62.67 (14.94)
Factor-adjusted LASSO	0.1038	8.477	90.43 (12.35)
SLDA	0.3000	10.00	68.80 (17.25)
Factor-adjusted SLDA	0.0926	10.00	87.50 (11.67)
SDA	0.1258	70.00	50.29 (14.84)
Factor-adjusted SDA	0.0452	53.17	65.17 (19.00)
DDA	0.4772	4.18	69.75 (18.30)
Factor-adjusted DDA	0.0474	55.26	65.04 (20.61)
Temporal dependence (*D*)			
LASSO	0.3020	13.10	62.36 (20.63)
Factor-adjusted LASSO	0.1510	8.03	93.02 (9.69)
SLDA	0.3314	10.00	62.50 (17.08)
Facto-adjusted SLDA	0.1222	10.00	90.90 (10.83)
SDA	0.2695	57.20	75.07 (23.94)
Factor-adjusted SDA	0.0893	68.22	67.93 (25.66)
DDA	0.4813	149.42	15.58 (15.27)
Factor-adjusted DDA	0.3146	97.65	48.76 (29.91)



(a) 2-blocks structure **(B)**



(b) Factor structure **(C)**



(c) Temporal dependence **(D)**

Fig. 4 Violin plots of error rates. **a** Two-blocks structure (*B*); **b** factor structure (*C*); **c** temporal dependence (*D*)

Considering the block structure (*B*), errors rates are reduced for each classification method and relevant features are more often selected except for SDA.

As expected, scenario (*C*) leads to the most significant results mainly because this scenario is favored by the factor model used for the covariance matrix.

When applied on raw data, DDA always leads to the highest error rates. In scenario (*C*), the selection step is very unstable as no variable was selected in 15 % of simulations, which explains that the average number of selected features only rates 4.18 % variables. For the two other scenarios, the number of selected features is high, but without catching rel-

evant ones. As expected, DDA, which assumes independence between covariates, is more suitable on factor-adjusted data and performances are better both in terms of prediction ability and in selection.

LASSO and Sparse LDA are considerably improved by factor adjustment. Interestingly, the former methods give similar results, probably because they are both based on ℓ_1-regularization. However, the benefit of factor adjustment on SDA is lesser than on the other classification methods. SDA is indeed a competing method to factor adjustment as it is also based on decorrelation. Nevertheless, SDA seems to be improved by a factor adjustment, which could be explained by the better ability of the factor model to catch a complex dependence than the James–Stein approach.

6 Discussion and conclusion

The analysis of high-dimensional data has markedly renewed the statistical methodology for feature selection in classification issues. Such data are characterized by their heterogeneity, as confusing factors can interfere with the signal of interest. A common and notorious difficulty in large-scale data analysis is therefore the handling of these confounding factors, which may induce bias in significance studies, cause unreliable feature selection and high error rates.

The present article illustrates that data heterogeneity affects the ranking and the stability of supervised classification model selection. Most of the usual procedures in supervised classification assume a weak correlation structure between variables and heterogeneity of the data violates this assumption. This article describes an innovative methodology based on an explicit modeling of the data heterogeneity, which provides a general framework to deal with dependence in variable selection. A supervised factor model is used to capture data dependence into a linear low-dimensional space and a conditional Bayes consistency is defined in this framework. This paper provides an algorithm which takes advantage of the correlation structure to estimate at the same time the correlation structure, the signal and individual probabilities in order to decorrelate data. Furthermore, we show that the conditional optimality of the linear Bayes classifier is achieved by the usual Bayes classifier applied to the factor-adjusted data.

Factor adjustment is shown to improve stability of some usual procedures of selection and classification. One very important implication of the factor-adjusted approach is that, in situations where a strong dependence can be approximated using a factor decomposition, the performance for classification is markedly improved.

Our simulation study shows nice operating characteristics considering dependence structures that fit well to genomics, according to several authors, which is one of our scientific area of interest. We believe that this approach can also be convenient for other scientific areas. As an illustration, we have considered a Toeplitz design, which can be used to model simple auto-regressive time dependence structures.

In this paper, it is assumed that the covariance structures in both groups are the same, which is consistent with the homoscedasticity assumption of Linear Discriminant Analysis. Extraction of factors Z depending on the response variable Y is possible by considering a different factor model in each group. In such case, two models are independently estimated from the two sets of observations where $Y = 0$ or $Y = 1$. However, in high-dimensional data analysis, where the total number of observation is often small, it could reduce the power to detect the biological signal (different means in each group).

References

Ahdesmäki, M., Strimmer, K.: Feature selection in omics prediction problems using cat scores and false non-discovery rate control. Ann. Appl. Stat. **4**, 503–519 (2010)

Bach, F.: Bolasso: model consistent lasso estimation through the bootstrap. Proceedings of the twenty-fifth International Conference on Machine Learning (ICML) (2008)

Bickel, P., Levina, E.: Some theory for Fisher's linear discriminant function, naive Bayes, and some alternatives when there are many more variables than observations. Bernoulli **10**(6), 989–1010 (2004)

Blum, Y., LeMignon, G., Lagarrigue, S., Causeur, D.: A factor model to analyze heterogeneity in gene expression. BMC Bioinform. **11**, 368 (2010)

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., West, M.: High-dimensional sparse factor modeling: applications in gene expression genomics. J. Am. Stat. Assoc. Appl. Case Stud. **103**, 484 (2008)

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

Dabney, A., Storey, J.: Optimality driven nearest centroid classification from genomic data. PLoS ONE **2**(10), e1002 (2007)

Donoho, D., Jin, J.: Higher criticism thresholding: optimal feature selection when useful features are rare and weak. Proc. Natl. Acad. Sci. **105**(39), 14790–14795 (2008)

Dudoit, S., Fridlyand, J., Speed, T.: Comparison of discrimination methods for the classification of tumors using gene expression data. J. Am. Stat. Assoc. **97**, 77–87 (2002)

Efron, B.: Empirical Bayes estimates for large-scale prediction problems. Technical report, Department of Statistics, Stanford University (2008)

Efron, B.: Correlation and large-scale simultaneous testing. J. Am. Stat. Assoc. **102**, 93–103 (2007)

Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. **33**, 1–22 (2010)

Friguet, C., Kloareg, M., Causeur, D.: A factor model approach to multiple testing under dependence. J. Am. Stat. Assoc. **104**(488), 1406–1415 (2009)

Guo, Y., Hastie, T., Tibshirani, R.: Regularized discriminant analysis and its application in microarrays. Biostatistics **8**, 86–100 (2007)

Hastie, T., Buja, A., Tibshirani, R.: Penalized discriminant analysis. Ann. Stat. **23**(1), 73–102 (1995)

Hedenfalk, I., Duggan, D., Chen, Y.D., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O.P., Wilfond, B., Borg, A., Trent, J.: Gene expression profiles in hereditary breast cancer. New Engl. J. Med. **344**, 539–548 (2001)

Kustra, R., Shioda, R., Zhu, M.: A factor analysis model for functional genomics. BMC Inform. **7**, 216–229 (2006)

Lee, S., Batzoglou, S.: Application of independent component analysis to microarrays. Genome Biol. **4**(11), R76 (2003)

Leek, J.T., Storey, J.: Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. **3**(9), e161 (2007)

Leek, J.T., Storey, J.: A general framework for multiple testing dependence. Proc. Natl. Acad. Sci. **105**, 18718–18723 (2008)

Levina, E.: Statistical issues in texture analysis. PhD thesis, University of California, Berkeley (2002)

Meinshausen, N., Bühlmann, P.: Stability selection. J. R. Stat. Soc. B **72**(4), 417–473 (2010)

Pournara, I., Wernisch, L.: Factor analysis for gene regulatory networks and transcription factor activity profiles. BMC Bioinform. **8**, 61 (2007)

Spearman, C.: General intelligence, objectively determined and measured. Am. J. Psychol. **15**, 201–293 (1904)

Sun, Y., Zhang, N., Owen, A.: Multiple hypothesis testing adjusted for latent variables, with an application to the AGEMAP gene expression data. Ann. Appl. Stat. **6**(4), 1664–1688 (2012)

Teschendorff, A., Zhuang, J., Widschwendter, M.: Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics **27**(11), 1496–1505 (2011)

Tibshirani, R.: Regression shrinkage and selection via LASSO. J. R. Stat. Soc. B **58**, 267–288 (1996)

Tibshirani, R., Hastie, T., Narasimhan, B., Chu, G.: Diagnosis of multiple cancer type by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA **99**, 6567–6572 (2002)

Tibshirani, R., Hastie, T., Narsimhan, B., Chu, G.: Class prediction by nearest shrunken centroids, with applications to DNA microarrays. Stat. Sci. **18**, 104–117 (2003)

Van de Geer, S.: L1-regularization in high-dimensional statistical models. Proceedings of the International Congress of Mathematicians (2010)

Xu, P., Brock, G., Parrish, R.S.: Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. Comput. Stat. Data Anal. **53**, 1674–1687 (2009)

Yang, Y.: Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. Biometrika **92**(4), 937–950 (2005)

Zou, H.: The adaptive LASSO and its oracle properties. J. Am. Stat. Assoc. **101**(476), 1418–1429 (2006)

Zouridis, H., et al.: Methylation subtypes and large-scale epigenetic alterations in gastric cancer. Sci. Transl. Med. **4**(156), 156-140 (2012)

Zuber, V., Strimmer, K.: Gene ranking and biomarker discovery under correlation. Bioinformatics **25**, 2700–2707 (2009)