Asymptotic frameworks for high-dimensional two-group classification

Duarte Silva, A. Pedro
Universidade Catolica Portuguesa at Porto, Faculdade de Economia e Gestão
Rua Diogo Botelho 1327
Porto 4169-005, Portugal
E-mail: psilva@porto.ucp.pt

ABSTRACT

Asymptotic properties of two-group supervised classification rules designed for problems with much more variables than observations are discussed. Two types of asymptotic bounds on expected error rates are considered: (i) bounds that assume consistent mean estimators and focus on the impact of the covariance matrix estimation. (ii) bounds that consider the errors in mean and covariance estimation. Known results for independence-based classification rules are generalized to correlation-adjusted linear rules.

1. Introduction

The classical theory of Linear Discriminant Analysis assumes the existence of a non-singular empirical covariance matrix. However, nowadays many applications work with data bases where the number of variables (p) is larger than the number of observations (n). Although there are classification rules specifically designed to tackle these problems (*e.g.* Tibshirani, Hastie, Narismhan, and Chu (2003), Fan and Fan (2008), Duarte Silva (forthcoming)), their theoretical properties are often not well known. In this paper, assuming that both n and p go to infinity at appropriate rates, asymptotic bounds on expected error rates for some linear classifiers will be reviewed and extended.

In one of the first attempts to study theoretical properties of classification rules in the large p, smaller n setting, Bickel and Levina (2004) proposed an asymptotic framework that allows the number of variables to grow faster than the number of observations. Under the assumption that the vector of mean differences can be estimated consistently as the number of variables grows without limit, these authors have shown that the expected error of the Naive rule that ignores all sample correlations can approach a constant close to the expected error of the optimal Bayes rule. These results were extended by Duarte Silva (forthcoming) for the case of classification rules based on well-conditioned covariance matrix estimators derived from factor models. In this paper, a framework will be proposed that allows the identification of general conditions in which results of this type can be generalized, and tighter asymptotic bounds can be derived.

The analysis discussed in the previous paragraph focus on the impact of estimating covariances by estimators belonging to some restricted class of well-conditioned matrices, while conveniently assuming that the estimation error in the vector of mean differences is vanishingly small. However, this property does not hold for the vector of sample mean differences without any form or regularization of variable selection. Without making the former, somehow restrictive, assumption, Fan and Fan (2008) derived new asymptotic bounds for the error rate of linear classification rules employing diagonal covariance matrix estimators. Here, it will be shown how Fan and Fan results can be generalized to classification rules based on well-conditioned, but not necessarily diagonal, covariance estimators.

The remainder of this paper is organized as follows. Section 2 reviews and generalizes asymptotic error bounds of the type studied by Bickel and Levina. Section 3 generalizes error bounds similar to those considered by Fan and Fan. Section 4 concludes the paper.

2. Asymptotic error bounds of the first kind

Consider the two-group homoscedastic Gaussian model where entities are represented by binary pairs $(X, Y); X \in \Re^p; Y \in \{0, 1\}$ and the distribution of X conditioned on Y is the multivariate normal $N_p(\mu_{(Y)}, \Sigma)$. The classical discriminant problem deals with the development of rules capable of predicting unknown Y values (class labels) given X observations. When the parameters $\mu_{(0)}, \mu_{(1)}, \Sigma$ are known, and the a-priori probabilities $P(Y = 0), P(Y = 1)$ are equal, the classification rule that minimizes the expected misclassification error is the theoretical Bayes rule, given by

$$(1) \quad Y = \delta_B(X) = \mathbf{1}(\Delta^T \Sigma^{-1}(X - \mu_.) > 0),$$

where $\Delta = \mu_{(1)} - \mu_{(0)}$, $\mu_. = \frac{1}{2}(\mu_{(0)} + \mu_{(1)})$, and $\mathbf{1}(.)$ is the indicator function.

In this section we will discuss the asymptotic performance of empirical rules that try to approximate δ_B when $n, p \to \infty$, and $n/p \to d < \infty$. In particular, we will be concerned with the conditions for convergence, and the limit, of the worst case expected misclassification error

$$(2) \quad \overline{W}_{\Gamma_1}(\delta_E) = max_{\theta \in \Gamma_1} \left[P_\theta(\delta_E(X) = 1 | Y = 0) \right],$$

where $\theta = (\mu_., \Delta, \Sigma)$, δ_E is an empirical rule given by

$$(3) \quad Y = \delta_E(X) = \mathbf{1}(\hat{\Delta}^T \hat{\Sigma}^{-1}(X - \hat{\mu}_.) > 0,$$

and Γ_1 is some parameter space statisfying

$$\Gamma_1(c, k_1, k_2) = \left\{ \begin{array}{c} \theta : \\ \Delta^T \Sigma^{-1} \Delta \geq c^2 \\ k_1 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq k_2 \\ \Delta \in B \end{array} \right\}$$

with $\lambda_{min}(\Sigma)$ and $\lambda_{max}(\Sigma)$ being the smallest and largest eigenvalues of Σ, and B a compact subset of $l_2(N)$, the set of real number sequences with convergent square sums.

The condition $\Delta^T \Sigma^{-1} \Delta \geq c^2$ establishes the minimum degree of group separation on Γ_1. It is well known that for fixed θ, the expected error rate of the optimal Bayes rule equals $1 - \Phi(\frac{1}{2}\sqrt{\Delta^T \Sigma^{-1} \Delta})$ with $\Phi(.)$ being the cumulative probability of a standardized Gaussian random variable. Therefore, this condition implies that for all $\theta \in \Gamma_1$, the optimal misclassification rate is bounded from above by $1 - \Phi(c/2)$, which then becomes a useful benchmark against which the asymptotic rate of δ_E can be assessed. Condition $k_1 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq k_2$ ensures that Σ is always non-singular and well-conditioned. Condition $\Delta \in B$ is a technical requirement necessary to allow the possibility of estimating Δ consistently. Known results in the theory of countable Gaussian sequences (see Johnstone (2002) and Lemma 1 in Bickel and Levina (2004)) show that an estimator such that $E_\theta ||\hat{\Delta} - \Delta||^2 = o_p(1)$ exists if and only if Δ is restricted to lie on a compact subset of $l_2(N)$.

Furthermore, let

$$\Upsilon_1(k_0) = \left\{ \begin{array}{c} \Sigma_R : \\ \Sigma_R \text{ is a square symmetric matrix} \\ \lambda_{min}(\Sigma_R) \geq k_0 > 0 \\ \cdots \end{array} \right\}$$

be a class of restricted covariance matrices that, by construction, are forced to be non-singular. The non-singularity of the covariance estimator used in (3) is central for all the results presented in this paper. We note that for $p \geq n$ this restriction is not satisfied for the unbiased sample covariance, $S = \frac{1}{n-2}\left[\sum_{Y_i=0}(X_i - \bar{X}_0)(X_i - \bar{X}_0)^T + \sum_{Y_i=1}(X_i - \bar{X}_1)(X_i - \bar{X}_1)^T\right]$, that being one of the main reasons why traditional linear discriminant analysis breaks down in large p, smaller n, classification problems.

The main result of this section is presented in Theorem 1, below

Theorem 1

Assume that $P(Y=0) = P(Y=1) = \frac{1}{2}$, and $(\theta_1, \Sigma_{R1}, \delta_{E1})$, $(\theta_2, \Sigma_{R2}, \delta_{E2})$, $\ldots$, $(\theta_p, \Sigma_{Rp}, \delta_{Ep})$, $\ldots$ is a sequence of true parameters, restricted covariance matrices, and empirical classification rules, indexed by dimensionality, satisfying

(i) $\exists p_0 \in N : p > p_0 \Rightarrow \theta_p \in \Gamma_1$; $\Sigma_{Rp} \in \Upsilon_1$

(ii) $max_{j,l}\left|\hat{\Sigma}_p^{-1}(j,l) - \Sigma_{Rp}^{-1}(j,l)\right| \xrightarrow{P} 0$ uniformly over Γ_1 when $p \to \infty$

(iii) $E_\theta||\hat{\Delta} - \Delta||^2 = o_p(1)$

(iv) $log(p) = o(n)$

Then

$$lim_{p\to\infty}\ sup\ \overline{W}_{\Gamma_1}(\delta_{Ep}) \leq 1 - \Phi\left(\frac{\sqrt{K_{0\Upsilon_1}}}{1 + K_{0\Upsilon_1}}\ c\right)$$

where

$$K_{0\Upsilon_1} = lim_{p\to\infty}\ max_{\Upsilon_1}\frac{\lambda_{max}(\Sigma_{0Rp})}{\lambda_{min}(\Sigma_{0Rp})}\ ;\ \Sigma_{0Rp} = \Sigma_{Rp}^{-\frac{1}{2}}\ \Sigma_p\ \left(\Sigma_{Rp}^{-\frac{1}{2}}\right)^T$$

Theorem 1 generalizes the asymptotic error bounds derived in Bickel and Levina (2004) and Duarte Silva (forthcoming), respectively for the Naive, and correlation factor model, classification rules. The proof follows from a direct adaptation of Duarte Silva (forthcoming, 2011) proof of the equivalent result for factor model classification rules. When Υ_1 is the class of diagonal positive-definite matrices, $K_{0\Upsilon_1}$ becomes a majorant on ratios of correlation eigenvalues, and Theorem 1 replicates the classical result originally proved by Bickel and Levina. Likewise, when Υ_1 is a class of covariance matrices derived from q-factor models with strictly positive specific variances, $K_{0\Upsilon_1}$ becomes a constant that measures the distance between the true data generating process and the postulated model that, as shown by Duarte Silva (forthcoming), for some data conditions can lead to much lower asymptotic error bounds. In theorem 1 above, even more general classes of well-conditioned covariance matrices are allowed.

3. Asymptotic error bounds of the second kind

The results presented in the previous section rely on the assumption that rule (3) employs an Δ estimator, $\hat{\Delta}$, such that

$$(4) \quad E_\theta||\hat{\Delta} - \Delta||^2 = o_p(1).$$

While such estimators are known to exist whenever $\Delta \in l_2(N)$, property (4) requires some form of regularization and is not satisfied for the simple vector of sample mean differences, $\bar{X}_1 - \bar{X}_0 = \frac{1}{n_1}\sum_{Y_i=1} X_i - \frac{1}{n_0}\sum_{Y_i=0} X_i$ when $n = O(p)$, because of error accumulation in the components of $\bar{X}_1 - \bar{X}_0$ (see Johnstone 2002). The most common form of regularization is simple truncation, which leads to the estimators

$$\hat{\Delta} = [\hat{\Delta}(1), \hat{\Delta}(2), \cdots, \hat{\Delta}(j), \cdots] \;\; ; \;\; \hat{\Delta}(j) = \begin{cases} \bar{X}_1(j) - \bar{X}_0(j), & j \leq m \\ 0, & j > m \end{cases} \quad \text{for some}\, m < p.$$

This is equivalent of performing a preliminary step of variable selection, keeping only the m variables considered to be the most important. In his approach, the relative importance of the different variables is almost always assessed by their standardized difference of univariate differences in sample group means. However the choice the choice of m does not have one unique established solution. One possibility is to minimize an estimate of an upper bound on the expected error rate of δ_E. Fan and Fan (2008) followed this route for the case on the Naive classification rule, and derived the following asymptotic bound.

$$(5) \quad lim_{p\to\infty}\, sup\, \overline{W}_{\Gamma_2}(\delta_{Ep}) \leq 1 - \Phi\left(\frac{\sqrt{\frac{n_0\, n_1}{p\, n}}\Delta^T D^{-1}\Delta + \sqrt{\frac{p}{n\, n_0\, n_1}}(n_1 - n_0)}{2\sqrt{\lambda_{max}(R)(1 + \frac{n_0\, n_1}{p\, n})\Delta^T D^{-1}\Delta}}\right)$$

where n_0, n_1 are the number of training observations in each group, $D = diag(\Sigma)$, and $R = D^{-1/2}\Sigma D^{-1/2}$ is the true correlation matrix.
The bound (5) holds when the following conditions are satisfied

$$\theta \in \Gamma_2(b_0) \;\; = \;\; \left\{ \begin{array}{c} \theta : \\ \\ \Delta^T D^{-1}\Delta \geq C_p \\ \lambda_{max}(R) \leq b_0 \\ min_j\, \sigma_j^2 > 0 \end{array} \right\} \qquad log(p) = o(n)$$

where σ_j^2 is the variance of the j^{th} variable and C_p is a sequence of constants satisfying $n\, C_p \to \infty$ when $n, p \to \infty$.
Here, we will be interested in deriving and minimizing similar bounds for classification rules based on well-conditioned, but not necessarily diagonal, covariance estimators. In particular, a careful inspection of Fan and Fan proof reveals it can be readily adapted to show that for rules of the form (3), with $\hat{\Sigma}_R \in \Upsilon_2(k_0)$ such that

$$\Upsilon_2(k_0) \;\; = \;\; \left\{ \begin{array}{c} \Sigma_R : \\ \\ \Sigma_R \,\text{is a square symmetric matrix} \\ \lambda_{min}(\Sigma_R) \geq k_0 > 0 \\ ||\Sigma_R - \hat{\Sigma}_R|| \to 0 \qquad ||\Sigma_R^{-1} - \hat{\Sigma}_R^{-1}|| \to 0 \qquad \text{when } n, p \to \infty \\ \cdots \end{array} \right\}$$

it follows that when such a covariance estimator is used and

$$\theta \in \Gamma_3(k_1, k_2, b_0) \;=\; \left\{ \begin{array}{c} \theta : \\ \Delta^T \Sigma_R^{-1} \Delta \geq C_p \\ \lambda_{max}(\Sigma_{0R}) \leq b_0 \\ k_1 \leq \lambda_{min}(\Sigma_R) \leq \lambda_{max}(\Sigma_R) \leq k_2 \end{array} \right\} \text{ with } \Sigma_{0R} = \Sigma_R^{-\frac{1}{2}} \; \Sigma \; \Sigma_R^{-\frac{1}{2}}$$

and $n\, C_p \to \infty \; ; \; log(p) = o(n)$

Then

$$(6) \quad lim_{p\to\infty} \; sup \; \overline{W}_{\Gamma_3}(\delta_{Ep}) \leq 1 - \Phi \left(\frac{\sqrt{\frac{n_0\, n_1}{\gamma\, n}} \Delta^T \Sigma_R^{-1} \Delta + \sqrt{\frac{\gamma}{n\, n_0\, n_1}} (n_1 - n_0)}{2\sqrt{\lambda_{max}(\Sigma_{0R})(1 + \frac{n_0\, n_1}{\gamma\, n}) \Delta^T \Sigma_R^{-1} \Delta}} \right)$$

where $\gamma = tr\Sigma_{0R}$

4. Conclusions and perspectives

We have generalized known large p asymptotic bounds on expected error rates of diagonal two-group classification rules, to rules that take correlation information into account. Two types of bounds were considered, the first kind concentrates of the lost of accuracy that is uniquely attributed by restricting the class of covariance estimators, while the second also considers the error in mean estimation and may be useful to find an appropriate number of predictors. In future research we intend to evaluate numerically the proposed bounds for typical problems of high-dimensional supervised classification, and compare them with estimates of the true error rates.

REFERENCES

BICKEL, P.J. and LEVINA, E. (2004). Some theory for Fisher's linear discriminant function, naive Bayes and some alternatives when there are more variables than observations. *Bernoulli*, **10**, 989-1010.

DUARTE SILVA, A.P. (Forthcoming). Two-group classification with high-dimensional correlated data: A factor model approach. *Computational Statistic and Data Analysis*. DOI:10.1016/j.csda.2011.05.02

DUARTE SILVA, A.P. (2011). Supplement to Two-group classification with high-dimensional correlated data: A factor model approach. http://www.porto.ucp.pt/feg/docentes/psilva.

FAN J. and FAN, Y. (2008). High dimensional classification using Features Annealed Independence Rules. *The Annals of Statistics*, **38**. 2605-2637.

JOHNSTONE., I.M. (2002). Function estimation and Gaussian noise sequence models. Unpublished monograph, http://www-stat.stanford.edu/˜imj.

TIBSHIRANI, R., HASTIE, B., NARISMHAN, B. and CHU, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays, *Statistical Science*, **18**, 104-117.