Statistical Applications in Genetics and Molecular Biology manuscript No.
(will be inserted by the editor)

Estimating the true error rate of classification rules built with additional information. An application to a cancer trial

David Conde · Bonifacio Salvador · Cristina Rueda · Miguel A. Fernández

Received: date / Accepted: date

Abstract Classification rules that incorporate additional information usually present in discrimination problems are receiving certain attention during the last years as they perform better than the usual rules in poor discrimination problems. Fernández et al (2006) proved that these rules have a lower unconditional misclassification probability than the usual Fisher's rule but they did not consider the estimation of the conditional error probability when a training sample is given (the so-called true error rate) which is a very interesting parameter in practice.

In this paper we consider the problem of estimating the true error rate of these classification rules in the classical topic of discrimination among two normal populations. We prove theoretical results on the apparent error rate of the rules that expose the need of new estimators of their true error rate. Our proposal is to also consider the additional information in the definition of the true error rate estimators. We propose four such new estimators. Two of them are defined incorporating the additional information into the leave-one-out-bootstrap. The other two are the corresponding cross-validation after bootstrap versions. We compare these new estimators with the usual ones in a simulation study and in a cancer trial application, showing the very good behavior, in terms of mean square error, of the leave-one-out bootstrap estimators that incorporate the available additional information.

Keywords Discriminant Analysis · True error rate · Order Restrictions · Restricted Estimation · Bootstrap

Universidad de Valladolid
Departamento de Estadística e I.O., 47005 Valladolid, Spain
Corresponding author: M. A. Fernández.
Tel.: +34-983-423945
Fax: +34-983-423111
E-mail: miguelaf@eio.uva.es

1 Introduction

Consider the classical discrimination problem with two populations Π_1 and Π_2. Denote the training sample from which the rule is built as $M_n = \{(X_i, Y_i), i = 1, ..., n\}$, where X is the $p-$dimensional vector of classifiers and $Y = 1, 2$ is the binary variable identifying the population. Denote also as P_{XY} the joint distribution of the vector (X, Y). With this scheme, a classification rule is an application $R_n : \{\mathbb{R}^p \times \{1,2\}\}^n \times \mathbb{R}^p \longrightarrow \{1,2\}$ that classifies a new observation $u \in \mathbb{R}^p$ into one of the two available populations, $R_n(M_n, u) \in \{1,2\}$.

In applications it is usual that some additional information is available. Recent papers considering additional information issues are, for example, Beran and Dümbgen (2010) and Oh and Shin (2011). It is frequent that this information tells us that the observations from one of the populations, for example Π_1, take higher (or lower) values than those coming from the other, i.e. Π_2. The incorporation of this kind of information into the classification rules has been shown to improve the performance of the rule. To our best knowledge, the first paper in this line was Long and Gupta (1998). More recently, Fernández et al. (2006) generalized and improved the results in that paper and proposed rules that take into account this additional information and have lower total misclassification probabilities (TMP) than the classical rules that do not consider this information. A good example of this situation appears in Section **??** where bladder cancer patients are known to usually take higher values in some variables (and lower in others) than healthy people. This information is then used to build a classification rule that outperforms Fisher's rule.

There are other important issues in a classification rule. One of them is the robustness of the rule with respect to its theoretical assumptions. The robustness properties of the rules that incorporate additional information have been studied in Salvador et al. (2008). Another issue, at least as important as the robustness of a rule, is the proper evaluation of the error of the rule in practice. This should be done estimating the true error rate E_n of the rule R_n, which is the misclassification probability of the rule conditioned to the available training sample. Namely, $E_n = Error(R_n) = P_{XY}\left(R_n(M_n, X) \neq Y \,/ M_n\right)$. In Fernández et al. (2006) the behavior of the 'restricted' rules is evaluated using the TMP which is the expected, or unconditional, true error rate $E(E_n)$. This allows the study of global properties of the rule but not the evaluation of E_n for a given sample M_n.

It is well known that the best way of estimating the true classification error of a classification rule is the use of an independent sample, usually called test sample. However, in practice it is common that the sample size is not large enough to split it into a training and a test sample as that would decrease the efficiency of the rule. For this reason, the estimation of E_n for the usual rules such as Fisher's linear rule, the quadratic discriminant rule, the nearest neighbors rules or random forest rules, is a topic widely studied in the literature. Parametric and non-parametric estimators of E_n have been proposed and non-parametric estimators based on resampling have shown a good performance for the above mentioned rules. Schiavo and Hand (2000) summarizes the work made on this topic until that date. More recent references are, for example, Steele and Patterson (2000), Wehberg and Schumacher (2004), Fu et al. (2005), Molinaro et al. (2005), Kim and Cha (2006) or Kim (2009).

In this paper, we deal with the important issue of estimating E_n for the restricted rules. In Section **??**, we start reviewing the restricted rules defined in Fernández et al. (2006). In Section **??**, we prove some interesting results on the apparent error rate (also known as resubstitution error) of the restricted rules that point to the need for new estimators of E_n. Then, in Section **??**, we briefly review the usual estimators of E_n based on cross-validation and bootstrap and we propose new estimators of E_n specific for the restricted rules. The main idea under these new estimators is to use the additional information also on the definition of the estimators. The good behavior of the new estimators of E_n is evaluated in a simulation study appearing in Section **??** and in a real data case dealing with bladder cancer presented in Section **??**. Finally, in Section **??** we discuss the results and summarize the conclusions.

2 Classification rules with additional information

From now on, we assume two p-dimensional normal populations Π_1 and Π_2 with means μ_1 and μ_2 and common covariance matrix Σ. Using the notation given in the Introduction we have that, $X\big/Y=j \sim N_p(\mu_j,\Sigma)$. Let us denote as $\overline{X}_j, j=1,2$, the sample mean vector of the observations coming from population j (i.e. $\overline{X}_j = \left(\Sigma_{i=1}^n X_i I_{(Y_i=j)}\right)/\left(\Sigma_{i=1}^n I_{(Y_i=j)}\right)$ for $j=1,2$) and S the pooled sample covariance matrix. We also assume that we have additional information on the mean vectors that can be translated as $\delta=\mu_1-\mu_2\in C$, where C is a closed, convex, polyhedral cone in $\mathbb{R}^p$.

Let us further assume, without loss of generality, equal a priori probabilities π_j and misclassification costs. If we denote as $u\in\mathbb{R}^p$ a new observation to be classified, the optimal (theoretical) Bayes rule is:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u-\frac{\mu_1+\mu_2}{2}\right)' \Sigma^{-1}\delta \geq 0. \tag{1}$$

The usual linear classification rule, also known as Fisher's discriminant rule, is obtained replacing the unknown parameters μ_1, μ_2 and Σ by their estimators $\overline{X}_1$, $\overline{X}_2$ and S:

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u-\frac{\overline{X}_1+\overline{X}_2}{2}\right)' S^{-1}\overline{\delta} \geq 0, \text{ where } \overline{\delta}=\overline{X}_1-\overline{X}_2.$$

In order to obtain a classification rule that incorporates the additional information available for the problem, Fernández et al (2006) start rewriting rule (**??**) as

$$\text{Classify } U \text{ in } \Pi_1 \text{ iff } (U-(c_1\mu_1+c_2\mu_2)+c\delta)'\Sigma^{-1}\delta \geq 0,$$

where $c_i=n_i/(n_1+n_2)$, $i=1,2$ and $c=(c_1-c_2)/2$. The new classification rule is then obtained replacing Σ by S, $c_1\mu_1+c_2\mu_2$ by $c_1\overline{X}_1+c_2\overline{X}_2$ and the restricted parameter δ by an estimator that incorporates the additional information. To be more precise, δ is replaced by a member of the family δ_γ^*, with $\gamma\in[0,1]$, defined as the limit of the following iterative procedure that Fernández et al. (2006) show to be convergent. Let $\widehat{\delta}_\gamma^{(0)}=\overline{X}_1-\overline{X}_2$, and $\widehat{\delta}_\gamma^{(i)}=p_{S^{-1}}\left(\widehat{\delta}_\gamma^{(i-1)}/C\right)-\gamma p_{S^{-1}}\left(\widehat{\delta}_\gamma^{(i-1)}/C^p\right)$

for $i = 1,2,...$ where $p_{S^{-1}}(Z/C)$ is the projection of Z onto cone C with the metric given by S^{-1} and $C^p = \{z \in \mathbb{R}^p : z'x \leq 0, x \in C\}$ the polar cone of C. In this way the following family of new classification rules $R_n(\gamma) = R_n(\gamma, M_n)$ with $\gamma \in [0,1]$ is obtained

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)' S^{-1}\delta_\gamma^* \geq 0.$$

For more details on these rules and their properties the reader is referred to Fernández et al (2006).

3 Apparent Error Rate

The resubstitution estimator or apparent error rate, *APP*, estimates the true error rate as the proportion of observations in the training sample that are wrongly classified by the rule. It is well known, see, for example, McLachlan (1976) or Efron (1986), that *APP* is a biased estimator that underestimates the true error rate because the training sample data are used twice, both to build the rule and to check its accuracy.

In this section we obtain some properties of *APP*. In particular, in proposition **??** we prove that *APP* is less optimistic for the rules $R_n(\gamma)$, $\gamma \in [0,1]$ than for Fisher's rule. Consequently, we can expect that the usual estimators of the true error rate do not work well for these rules and new estimators of this parameter, specific for these rules, should be defined.

We can assume that $\Sigma = I$ without loss of generality. If the a priori probabilities for each population are equal the apparent error rate of rule $R_n(\gamma)$, $\gamma \in [0,1]$ is $APP(\gamma) = (APP_1(\gamma) + APP_2(\gamma))/2$, where

$$APP_1(\gamma) = \frac{1}{n_1}\sum_{i=1}^{n} I_{\left[\left(X_i - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)'\delta_\gamma^* < 0\right]}\, I_{[Y_i=1]}$$

$$APP_2(\gamma) = \frac{1}{n_2}\sum_{i=1}^{n} I_{\left[\left(X_i - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)'\delta_\gamma^* > 0\right]}\, I_{[Y_i=2]},$$

are the apparent error rates of populations Π_1 and Π_2 respectively.

Now, the expected apparent error rate for Π_1 is

$$E\left(APP_1(\gamma)\right) = P\left(\left(X_1 - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)'\delta_\gamma^* < 0, Y_1 = 1\right)$$
$$= E\left[P\left(\left(X_1 - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)'\delta_\gamma^* < 0, Y_1 = 1 \,/\overline{X}_1, \overline{X}_2\right)\right].$$

Proposition 1

$$P\left(\left(X_1 - \left(c_1\overline{X}_1 + c_2\overline{X}_2\right) + c\delta_\gamma^*\right)'\delta_\gamma^* < 0, Y_1 = 1 \,/\overline{X}_1, \overline{X}_2\right)$$
$$= \Phi\left(-\sqrt{\frac{n_1}{n_1 - 1}}\frac{\left(c_2\overline{\delta} + c\delta_\gamma^*\right)'\delta_\gamma^*}{\sqrt{\delta_\gamma^{*\prime}\delta_\gamma^*}}\right).$$

Proof In order to make the proof clearer and to remark the dependence of δ_γ^* on $\overline{X}_1$ and $\overline{X}_2$ during the proof we will write δ_γ^* as $\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)$. It is easy to check that

$$\begin{aligned}&\left(X_1-\left(c_1\overline{X}_1+c_2\overline{X}_2\right)+c\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\\&=\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)+\left(c_2\overline{\delta}+c\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\end{aligned}$$

so that

$$\begin{aligned}&P\left(\left(X_1-\left(c_1\overline{X}_1+c_2\overline{X}_2\right)+c\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)<0,Y_1=1\,/\overline{X}_1=t_1,\overline{X}_2=t_2\right)\\&=P\left(\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)<-\left(c_2\overline{\delta}+c\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right)\right)'\delta_\gamma^*\left(\overline{X}_1,\overline{X}_2\right),Y_1=1\,/\overline{X}_1=t_1,\overline{X}_2=t_2\right)\\&=P\left(\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(t_1,t_2\right)<-\left(c_2\left(t_1-t_2\right)+c\delta_\gamma^*\left(t_1,t_2\right)\right)'\delta_\gamma^*\left(t_1,t_2\right),Y_1=1\,/\overline{X}_1=t_1,\overline{X}_2=t_2\right).\end{aligned}$$

Now, $\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(t_1,t_2\right)\sim N\left(0,\frac{n_1-1}{n_1}\delta_\gamma^*\left(t_1,t_2\right)'\delta_\gamma^*\left(t_1,t_2\right)\right)$ is an ancillary statistic as its distribution does not depend on μ_1 or μ_2. As $\left(\overline{X}_1,\overline{X}_2\right)$ is sufficient and complete, from Basu's theorem we have that $\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(t_1,t_2\right)$ and $\left(\overline{X}_1,\overline{X}_2\right)$ are independent. From this fact we have that

$$\begin{aligned}&P\left(\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(t_1,t_2\right)<-\left(c_2\left(t_1-t_2\right)+c\delta_\gamma^*\left(t_1,t_2\right)\right)'\delta_\gamma^*\left(t_1,t_2\right),Y_1=1\,/\overline{X}_1=t_1,\overline{X}_2=t_2\right)\\&=P\left(\left(X_1-\overline{X}_1\right)'\delta_\gamma^*\left(t_1,t_2\right)<-\left(c_2\left(t_1-t_2\right)+c\delta_\gamma^*\left(t_1,t_2\right)\right)'\delta_\gamma^*\left(t_1,t_2\right),Y_1=1\right)\\&=\Phi\left(-\sqrt{\frac{n_1}{n_1-1}}\frac{\left(c_2\left(t_1-t_2\right)+c\delta_\gamma^*\left(t_1,t_2\right)\right)'\delta_\gamma^*\left(t_1,t_2\right)}{\sqrt{\delta_\gamma^*\left(t_1,t_2\right)'\delta_\gamma^*\left(t_1,t_2\right)}}\right).\end{aligned}$$

See Lehmann and Casella (1998) page 93 for the same argument in a similar situation.

In a similar way, for Π_2 we have

$$\begin{aligned}E\left(APP_2(\gamma)\right)&=P\left(\left(X_1-\left(c_1\overline{X}_1+c_2\overline{X}_2\right)+c\delta_\gamma^*\right)'\delta_\gamma^*>0,Y_1=2\right)\\&=E\left[P\left(\left(X_1-\left(c_1\overline{X}_1+c_2\overline{X}_2\right)+c\delta_\gamma^*\right)'\delta_\gamma^*>0,Y_1=2\,/\overline{X}_1,\overline{X}_2\right)\right]\end{aligned}$$

and

$$\begin{aligned}&P\left(\left(X_1-\left(c_1\overline{X}_1+c_2\overline{X}_2\right)+c\delta_\gamma^*\right)'\delta_\gamma^*>0,\;\;Y_1=2\,/\overline{X}_1,\overline{X}_2\right)\\&=\Phi\left(-\sqrt{\frac{n_2}{n_2-1}}\frac{\left(c_1\overline{\delta}-c\delta_\gamma^*\right)'\delta_\gamma^*}{\sqrt{\delta_\gamma^{*\prime}\delta_\gamma^*}}\right).\end{aligned}$$

And following the same lines we can also prove that for Fisher's rule

$$\begin{aligned}E(APP_1\left(Fisher\right))&=E\left[P\left(\left(X_1-\tfrac{1}{2}\left(\overline{X}_1+\overline{X}_2\right)\right)'\overline{\delta}<0,\;\;Y_1=1\,/\overline{X}_1,\overline{X}_2\right)\right]\\E(APP_2\left(Fisher\right))&=E\left[P\left(\left(X_1-\tfrac{1}{2}\left(\overline{X}_1+\overline{X}_2\right)\right)'\overline{\delta}>0,\;\;Y_1=2\,/\overline{X}_1,\overline{X}_2\right)\right]\end{aligned}$$

and

$$P\left(\left(X_1-\frac{1}{2}\left(\overline{X}_1+\overline{X}_2\right)\right)'\overline{\delta}<0,\ \ Y_1=1\,/\overline{X}_1,\overline{X}_2\right)=\Phi\left(-\tfrac{1}{2}\sqrt{\tfrac{n_1}{n_1-1}}\left\|\overline{\delta}\right\|\right)$$
$$P\left(\left(X_1-\frac{1}{2}\left(\overline{X}_1+\overline{X}_2\right)\right)'\overline{\delta}>0,\ \ Y_1=2\,/\overline{X}_1,\overline{X}_2\right)=\Phi\left(-\tfrac{1}{2}\sqrt{\tfrac{n_2}{n_2-1}}\left\|\overline{\delta}\right\|\right).$$

Proposition 2 *If $n_1=n_2$ then, for any $\gamma\in[0,1]$,*

$$E\left(APP(\gamma)\right)\geq E\left(APP(0)\right)\geq E\left(APP(Fisher)\right).$$

Proof As $n_1=n_2$ we have that $c=0$ and $c_1=c_2=\frac{1}{2}$. Now, $\delta_0^*=p(\overline{\delta}/C)$ and $\delta_\gamma^*\in C$ so taking into account Theorem 1.3.2 in Robertson et al. (1988), $\left(\overline{\delta}-\delta_0^*\right)'\delta_0^*=0$ and $\left(\overline{\delta}-\delta_0^*\right)'\delta_\gamma^*\leq 0$. From this,

$$\frac{\overline{\delta}'\delta_\gamma^*}{\sqrt{\delta_\gamma^{*\prime}\delta_\gamma^*}}\leq\frac{\delta_0^{*\prime}\delta_\gamma^*}{\sqrt{\delta_\gamma^{*\prime}\delta_\gamma^*}}=\|\delta_0^*\|\cos\left(\delta_0^*,\delta_\gamma^*\right)\leq\|\delta_0^*\|=\frac{\overline{\delta}'\delta_0^*}{\sqrt{\delta_0^{*\prime}\delta_0^*}}\leq\left\|\overline{\delta}\right\|,$$

and the result follows from Proposition **??**.

Remark 1 In Fernández et al. (2006) it is proved that, if $n_1=n_2$, the true error rate of rules $R_n(\gamma)$, $\gamma\in[0,1]$, is lower than that of Fisher's rule. Moreover, in all simulations performed the true error rate of rules $R_n(\gamma)$ is higher than their expected apparent error rates. As from Proposition **??**, $E\left(APP(\gamma)\right)\geq E\left(APP(Fisher)\right)$, this suggests that if $n_1=n_2$ the bias of *APP* for rules $R_n(\gamma)$, $\gamma\in[0,1]$ is lower than that for Fisher's rule.

A possible explanation for this is that the restricted rules are less dependent from the training sample values than Fisher's rule, as they are built not only using the training sample but also the additional information available for the problem.

4 Resampling based estimators

There are many non parametric estimators for the true error rate of a classification rule based on resampling techniques. In this section we describe the most usual ones and propose new estimators based on resampling techniques especially designed to cope with the inclusion of additional information in the classification rule.

4.1 Usual estimators

The cross-validation, or leave-one-out, method was proposed in Lachenbruch and Mickey (1968). With this method one of the observations in the training sample is left out, then the classification rule is computed and the excluded observation os classified. This is repeated with each of the observations in the training sample. Then the

cross-validation error CV is just the proportion of observations misclassified using this procedure. It is well known that this estimator has lower bias than APP.

Efron (1983) shows that the CV has a low bias but a not so low variability and proposes estimators based on the bootstrap methodology. Let us denote as $M_n = \{(X_i, Y_i), i = 1, ..., n\}$ the original training sample. A bootstrap training sample $M_n^* = \{(X_i^*, Y_i^*), i = 1, ..., n\}$ is a size n sample obtained randomly (with replacement) from the original training sample (i.e. $P((X_i^*, Y_i^*) = (X_s, Y_s)) = \frac{1}{n}$ with $s, i \in \{1, ..., n\}$). The probability that an observation from the original training sample is not included in the bootstrap training sample depends on n and is approximately 0.368. The bootstrap version of the classification rule is the rule based on the bootstrap training sample. From this methodology Efron proposes several ways of estimating the classification error. We consider two of them, the leave-one-out bootstrap ($LOOBT$) and the bootstrap 632 ($BT632$).

For the $LOOBT$ estimator, B bootstrap training samples are considered and B bootstrap versions of the classification rule are obtained. Then each of these rules is used for classifying the original observations not belonging to the corresponding bootstrap training sample. Finally, $LOOBT$ is the proportion of observations not correctly classified using this procedure. Efron notices that $LOOBT$ tends to overestimate the true error rate and then proposes $BT632 = 0.368APP + 0.632LOOBT$. In certain cases the value of APP is close to 0 (overfitting) so that $BT632$ is close to $0.632LOOBT$ and the true error is underestimated. For these situations with high overfitting, Efron and Tibshirani (1997) propose the bootstrap 632+, defined as $BT632+ = (1-\alpha)APP + \alpha LOOBT$, with $\alpha > 0.632$. In Section **??** we have proved that APP for rules $R_n(\gamma)$, $\gamma \in [0,1]$, is higher than APP for Fisher's rule. Consequently, the overfitting problem for these rules is less important and we do not consider $BT632+$ in our study.

More recently, Fu et al. (2005) propose a method based on cross-validation after bootstrap (BCV) that has a lower relative mean squared error than $LOOBT$ and $BT632$ for small training samples. In this procedure B bootstrap samples $M_b^*, b = 1, ..., B$ are obtained from M_n. Let CV_b be the true error rate estimator obtained using the cross-validation method on sample M_b^*. The final true error rate estimator is now $BCV = B^{-1}\sum_{b=1}^{B} CV_b$.

4.2 New proposals

In this paper, we propose new true error rate estimation procedures for the rules that take into account the additional information. These methods modify the $LOOBT$ and the BCV to make them able to cope properly with the information included in the rule. We will denote as $BT2$ and $BT3$ the methods generated from the $LOOBT$ method and $BT2CV$ and $BT3CV$ the ones coming from the BCV procedure.

The additional information we are considering can be written as $\delta = \mu_1 - \mu_2 \in C$, where $C = \left\{ z \in \mathbb{R}^p : a_j' z \geq 0, j = 1, ..., m \right\}$ is the appropriate cone of restrictions. Let

us denote as $\overline{C}$ the following random cone generated by $\overline{\delta}$

$$\overline{C} = \left\{ z \in \mathbb{R}^p : \begin{array}{l} a_j' z \geq 0, \text{ if } a_j' \overline{\delta} \geq 0 \\ a_j' z \leq 0, \text{ if } a_j' \overline{\delta} < 0 \end{array}, \ j = 1, ..., m \right\}.$$

The true error rate estimator $BT2$ is computed in a way similar to $LOOBT$ but considering bootstrap classification rules generated using projections onto cone $\overline{C}$ instead of C for each bootstrap training sample. In other words, for each bootstrap sample $M_b^* = \left\{ \left(X_i^{*b}, Y_i^{*b} \right), i = 1, 2, ..., n \right\}$ we compute the bootstrap version of the estimator of δ that we denote as δ_γ^{*b} (with $\gamma \in [0,1]$) defined as the limit of the following iterative procedure similar to the one considered in Section **??**. Let $\widehat{\delta}_\gamma^{(0)b} = \overline{X}_1^{*b} - \overline{X}_2^{*b}$ and $\widehat{\delta}_\gamma^{(i)b} = p_{S^{-1}} \left(\widehat{\delta}_\gamma^{(i-1)b} / \overline{C} \right) - \gamma p_{S^{-1}} \left(\widehat{\delta}_\gamma^{(i-1)b} / \overline{C}^p \right)$ for $i = 1, 2, \ldots$. Now we denote as $R_n^{*b}(\gamma)$ the bootstrap versions of the classification rules $R_n(\gamma)$ defined as,

$$\text{Classify } u \text{ in } \Pi_1 \text{ iff } \left(u - \left(c_1 \overline{X}_1^{*b} + c_2 \overline{X}_2^{*b} \right) + c \delta_\gamma^{*b} \right)' S^{-1} \delta_\gamma^{*b} \geq 0.$$

For each rule $R_n^{*b}(\gamma)$, $b = 1, 2, ...B$, we classify the observations in the original training sample that do not belong to the bootstrap sample M_b^*. The true error rate estimator $BT2$ is the proportion of observations wrongly classified.

The heuristic under $BT2$ is that the "bootstrap world" should mirror the "real world". In the "real world" the original training sample M_n is obtained from the populations Π_j, $j = 1, 2$, that verify $\delta = \mu_1 - \mu_2 \in C$. In the "bootstrap world" the population is M_n, which verifies $\overline{\delta} = \overline{X}_1 - \overline{X}_2 \in \overline{C}$. Therefore, the bootstrap versions of the rules should be obtained replacing the cone C by $\overline{C}$.

Our second proposal to use the additional information in a way that the "bootstrap world" imitates the "real world", is to adapt the original training sample, instead of modifying the cone, as follows.

Assume that the original training sample M_n does not verify the restrictions, i.e. $\overline{\delta} = \overline{X}_1 - \overline{X}_2 \notin C$. For any $\gamma \in [0,1]$, we can use δ_γ^*, the restricted estimator of δ, to obtain estimators for μ_i $i = 1, 2$. As $\mu_1 = (\mu_1 + \mu_2 + \delta)/2$ and $\mu_2 = (\mu_1 + \mu_2 - \delta)/2$, we can consider $\mu_{\gamma 1}^* = \left(\overline{X}_1 + \overline{X}_2 + \delta_\gamma^* \right)/2$ and $\mu_{\gamma 2}^* = \left(\overline{X}_1 + \overline{X}_2 - \delta_\gamma^* \right)/2$ as estimators for μ_1 and μ_2 respectively. Now, we transform the original training sample in such a way that the difference of the new sample means belongs to C. The transformed training sample is $\{(W_i, Y_i), i = 1, 2, ..., n\}$ where

$$W_i = X_i - \overline{X}_j + \mu_{\gamma j}^* \text{ if } Y_i = j, \ j = 1, 2.$$

In this way $\overline{W}_1 - \overline{W}_2 = \mu_{\gamma 1}^* - \mu_{\gamma 2}^* = \delta_\gamma^* \in C$. Now, the proposed estimator of the true error rate that we denote as $BT3$ is the $LOOBT$ replacing the original training sample by the transformed one. In this way, the bootstrap samples are extracted from populations that verify the same property that is fulfilled by the populations from which the original training sample is extracted.

We also consider the cross-validation after bootstrap versions of $BT2$ and $BT3$. They are denoted as $BT2CV$ and $BT3CV$ respectively.

5 Estimators behavior. Simulation study

The behavior of an estimator $\widehat{E}$ of the true error rate E_n is analyzed through the distribution of the random variable $\widehat{E}-E_n$. This distribution has been called deviation distribution of the error estimator by Braga-Neto and Dougherty (2004). As a global measure of the behavior of $\widehat{E}$ we will use $E\left[(\widehat{E}-E_n)^2\right]$. As usual, this measure can be decomposed in a variance and a bias component since $E\left[(\widehat{E}-E_n)^2\right] = Var(\widehat{E}-E_n)+\left[E(\widehat{E}-E_n)\right]^2$.

In this section, we conduct a simulation study to compare the behavior of the estimators, $APP(\gamma), CV(\gamma), LOOBT(\gamma), BT632(\gamma), BCV(\gamma), BT2(\gamma), BT3(\gamma), BT2CV(\gamma)$ and $BT3CV(\gamma)$ of the true error rate, $E_n(\gamma)$, of the restricted classification rules $R_n(\gamma)$.

The purpose of this study is to propose a reasonable estimator of $E_n(\gamma)$ when the training sample does not fulfill the restrictions. For simplicity we consider $p=3$ and identity covariance matrix and study the positive orthant restrictions case, i.e. $\delta \in O_3^+ = \left\{x \in \mathbb{R}^3 : x_i \geq 0, i=1,2,3\right\}$. We generate training samples of size $n_1 = n_2 = 10$, from populations Π_1, $N_3(\delta,\Sigma)$, and Π_2, $N_3(0,\Sigma)$, for different values of δ and Σ. Since in practice the sample sizes are usually larger than these values and the covariances are also larger, we have also run the simulations with bigger sample sizes ($n_1 = n_2 = 50$), rescaling the covariance matrix accordingly, obtaining similar results. The simulations were performed for many values of δ both in the interior of the cone and on the frontier of O_3^+ and for several values of Σ. However, since there was no significative variation in the results, in order to save space, we only present here the results obtained for values of $\Sigma = I$ and when δ is the vertex of the cone $(0,0,0)$ or is in the interior of O_3^+. To be more precise we show the results for values of δ in the diagonal direction of the cone, i.e. $\delta = \lambda\,(1,1,1)$ with $\lambda \geq 0$. The values of λ have been chosen so that $\|\delta\|^2 = 0, 0.25, 0.5, ..., 2.5$. Notice that the 11 values considered cover the situations where discrimination is easy since the distance between the means $\|\delta\|^2$ is large and others where the samples from the populations are much more likely to overlap. Larger values of $\|\delta\|^2$ are not given since for those values the restrictions are almost always fulfilled and therefore the true error estimation procedures are equivalent.

For each scenario, we generate 1000 training samples for which the rules $R_n(\gamma)$, with $\gamma \in \{0, 0.5, 1\}$, are determined. For each of these three rules we compute APP, CV, $LOOBT$, $BT632$, BCV, $BT2$, $BT3$, $BT2CV$ and $BT3CV$. The number of bootstrap replicas considered for the estimators involving bootstrap was $B = 100$. The true error rate E_n for each training sample was computed using a test sample with 1000 observations from each of the two populations. Using this procedure we have 1000 values of the deviation distribution of each of the 9 error estimators. With these values we approximate the values of $(E[(\widehat{E}-E_n)^2])^{\frac{1}{2}}$ and $E(\widehat{E}-E_n)$ that we will denote as $A(\widehat{E})$ and $B(\widehat{E})$ respectively. For example, if we denote as $BT2^i(0.5)$ and $E_n^i(0.5)$ the values of $BT2$ and E_n obtained from the i-th training sample for rule $R_n(0.5)$ we can estimate $A\,(BT2(0.5)) = (E[(BT2(0.5)-E_n(0.5))^2])^{\frac{1}{2}}$ and $B\,(BT2(0.5)) =$

$E(BT2(0.5)-E_n(0.5))$ by $(\frac{1}{1000}\sum_{i=1}^{1000}[BT2^i(0.5)-E_n^i(0.5)]^2)^{\frac{1}{2}}$ and $\frac{1}{1000}\sum_{i=1}^{1000}[BT2^i(0.5)-E_n^i(0.5)]$ respectively.

Again, in order to save space, as the results obtained for the three classification rules were similar, in Table **??** we only present the values for $\gamma = 1$. For each value of $\|\delta\|^2$, the two lowest values of $A(\widehat{E})$ appear in bold. Notice that the lowest values are the ones for $BT2$ and $BT3$ for almost all values of $\|\delta\|^2$.

Table 1 Simulations results for the 9 estimators under $\Sigma = I$ and $\gamma = 1$.

		$\|\delta\|^2$										
Estimator		0	0.25	0.5	0.75	1	1.25	1.5	1.75	2	2.25	2.5
APP	*A*	0.144	0.141	0.141	0.137	0.133	0.133	0.129	0.131	0.121	0.122	0.115
	B	-0.015	-0.054	-0.071	-0.076	-0.076	-0.074	-0.080	-0.080	-0.075	-0.081	-0.072
CV	*A*	0.130	0.126	0.126	0.123	0.118	0.125	0.114	0.119	0.112	0.104	0.104
	B	0.001	0.008	0.001	0.006	0.003	0.009	0.000	0.006	0.008	-0.006	0.005
LOOBT	*A*	**0.085**	0.090	0.091	0.092	0.090	0.097	0.094	0.098	**0.094**	0.092	**0.091**
	B	-0.005	0.008	0.006	0.014	0.015	0.018	0.014	0.019	0.022	0.014	0.024
*BT*632	*A*	0.104	0.102	0.102	0.098	0.095	0.100	0.095	0.098	0.091	**0.091**	**0.087**
	B	-0.009	-0.015	-0.022	-0.019	-0.018	-0.016	-0.021	-0.017	-0.014	-0.021	-0.011
BCV	*A*	0.114	0.114	0.112	0.108	0.105	0.107	0.103	0.103	0.097	0.097	0.092
	B	-0.011	-0.033	-0.045	-0.046	-0.049	-0.048	-0.052	-0.049	-0.046	-0.053	-0.043
*BT*2	*A*	0.094	**0.079**	**0.086**	**0.088**	**0.088**	**0.093**	**0.093**	**0.096**	0.095	**0.091**	0.092
	B	-0.069	-0.010	-0.002	0.010	0.013	0.015	0.013	0.018	0.021	0.014	0.024
*BT*3	*A*	**0.093**	**0.078**	**0.084**	**0.087**	**0.087**	**0.094**	**0.092**	**0.096**	**0.094**	**0.091**	**0.091**
	B	-0.069	-0.012	-0.002	0.010	0.013	0.016	0.013	0.018	0.021	0.014	0.024
BT2CV	*A*	0.146	0.109	0.109	0.105	0.102	0.103	0.102	0.100	0.097	0.096	**0.091**
	B	-0.129	-0.074	-0.068	-0.059	-0.057	-0.055	-0.055	-0.052	-0.048	-0.054	-0.044
BT3CV	*A*	0.145	0.110	0.108	0.104	0.102	0.103	0.102	0.100	0.097	0.097	**0.091**
	B	-0.128	-0.076	-0.067	-0.058	-0.057	-0.055	-0.056	-0.053	-0.048	-0.054	-0.044

In Figure **??**, we represent the values of $A(\widehat{E})$ and $B(\widehat{E})$ depending on $\|\delta\|^2$ for the 9 estimators of the true error rate that we are considering. As in other simulation studies *APP* generally has the largest negative bias, *CV* has the lowest bias but is the one with highest variance and *LOOBT* shows a positive bias. Estimators *BCV*, *BT2CV*, *BT3CV* and *BT*632 exhibit a negative bias. The new estimators proposed in this paper *BT*2 and *BT*3, which modify the bootstrap in order to cope with the additional information incorporated to the rules, have similar behavior. This is somehow surprising since they are based on very different ideas. They are also the best estimators of the true error rate for the smallest values of $\|\delta\|^2$. These are obviously the most interesting situations in practice, as they correspond to scenarios where the discrimination is more difficult and where the additional information is more likely to play a key role in the rule.

In order to have a more thorough idea of their behavior, we also obtained kernel estimators of the density of the deviation distribution for each of the 9 estimators of E_n. The kernel density estimators corresponding to scenario $\|\delta\|^2 = 0.3$ for the new estimators proposed in this paper, namely *BT*2, *BT*3, *BT*632, *BT2CV* and *BT3CV*, are represented in Figure **??**. From this Figure it is clear that the kernel estimators for *BT*2 and *BT*3 have the lowest values of bias and variance among the 5 represented

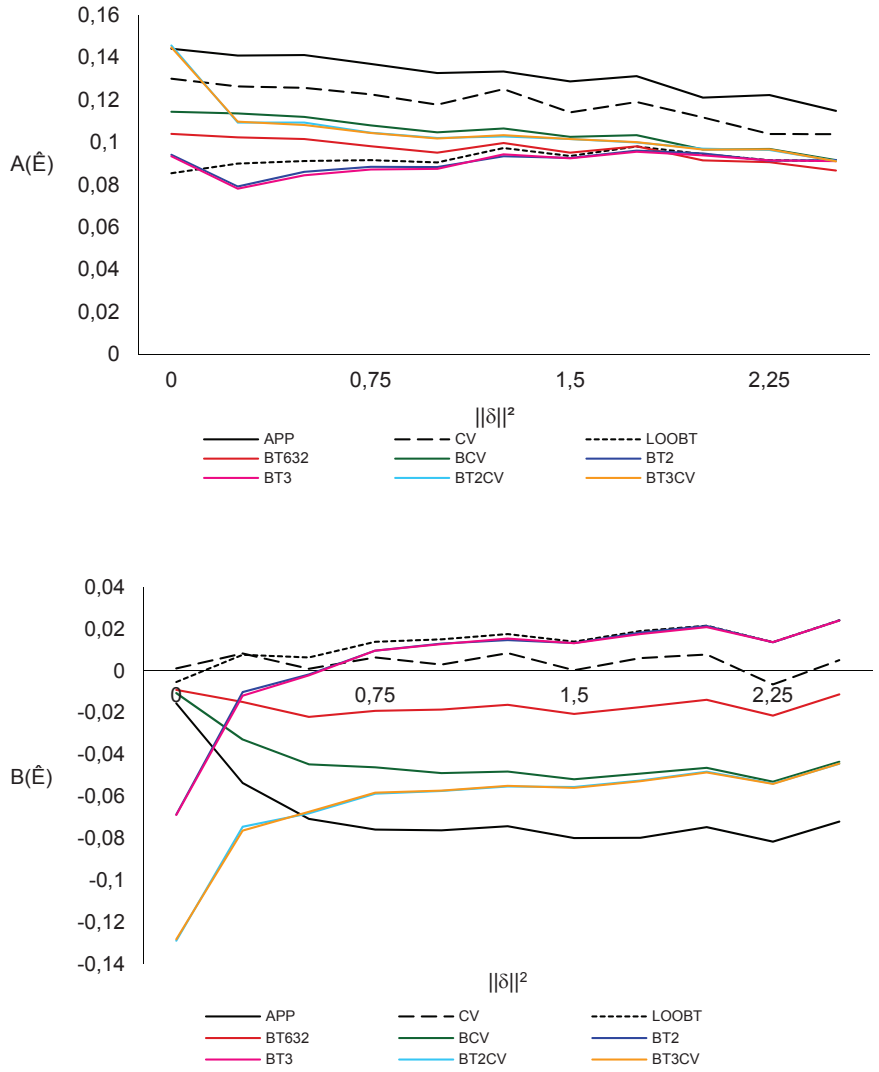

Fig. 1 $A(\widehat{E})$ and $B(\widehat{E})$ for the true error rate estimators for $\Sigma = I$ and $\gamma = 1$

in the graph. The estimators $BT2CV$ and $BT3CV$ have a similar variance component but are much more biased, while the $BT632$ has a higher variance component than the rest.

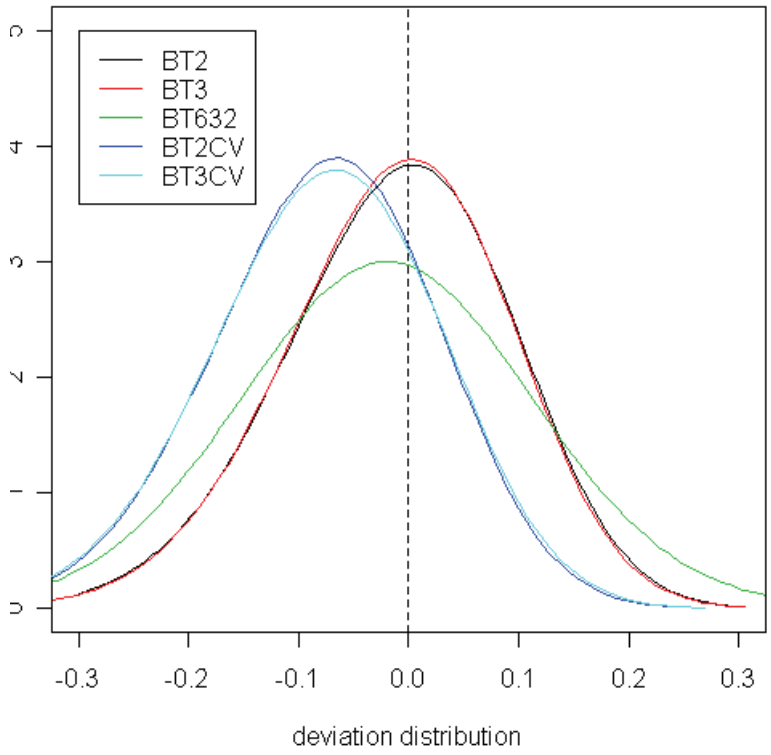

Fig. 2 Kernel estimators for the density function of $\widehat{E} - E_n$ for several estimators for $\|\delta\|^2 = 0.3$

6 Application. Bladder cancer data

The data considered in this application come from a bladder cancer project aimed to select classifiers in the context of an in vitro diagnostic tool for the disease. Our industrial and pharmaceutical partners in this research are Proteomika S.L. and Laboratorios SALVAT, S.A. For intellectual property reasons, the names of the proteins used in this study are not disclosed in the paper.

Patients were classified in five levels based on cytoscopy. First level is control level (i.e. negative result of cytoscopy, therefore considered as absence of bladder cancer) and the other four are denoted as Ta, T1G1, T1G3 and T2, each of them corresponding to increasingly advanced levels of cancer. This combines the TNM grading (see UICC (2009)) and the grading. To be more precise, stage T describes the size of the tumor and whether it has spread and grade G refers to the appearance of the cells under the microscope. For this example, and in order to keep the populations balanced we will consider the control level as population Π_1 and levels T1G3 and T2 as population Π_2.

As usual in this kind of research, an initial database was provided. The purpose of this pilot study was to confirm or discard the associations among the proteins and the illness in order to establish a larger multicenter study. The data set, D_1, contained information on 41 patients from Π_1 and 32 from Π_2 and 11 proteins together with the real stage of the illness the patients belonged to. This is the initial data set and it is the one we will use to build the rules. In the usual statistical terminology this is the training set.

For this example, we only consider 4 of the 11 available proteins. We will denote these proteins as P_1, P_2, P_3 and P_4. For each of these 4 proteins it was expected that higher values on the proteins were related to more advanced stages of the illness. As usual the values of the proteins levels have been transformed logarithmically in order to approach the variables to normality. The mean values in each of the populations and the pooled covariance matrix obtained from this data set appear in Table **??**. From this table it is obvious that the additional information was not fulfilled by the training set so the classifications rules $R_n(\gamma)$ are relevant in this problem. Table **??** contains the values for the restricted estimator δ^*_γ appearing in these rules for $\gamma \in \{0, 0.5, 1\}$.

Table 2 Mean for each group and pooled covariance matrix from D_1

		Means			
	N	$\log(P_1)$	$\log(P_2)$	$\log(P_3)$	$\log(P_4)$
Π_1	41	1.416	1.356	3.879	1.417
Π_2	32	1.409	0.976	4.348	1.578

$$S = \begin{pmatrix} 1.065 & 0.455 & -0.154 & 0.106 \\ 0.455 & 0.515 & -0.052 & -0.053 \\ -0.154 & -0.052 & 0.544 & 0.148 \\ 0.106 & -0.053 & 0.148 & 0.450 \end{pmatrix}$$

Table 3 Values of δ^*_γ for the $R_n(\gamma)$ rules built from D_1

γ	δ^*_γ
0	$(0.328, 0, 0.430, 0.123)$
0.5	$(0.496, 0.190, 0.411, 0.103)$
1	$(0.664, 0.380, 0.392, 0.084)$

Moreover, in this bladder cancer research, a second data set D_2, containing measures on the same 11 proteins and the real illness stage for a different set of 118 patients was received in a later stage. We use this second set as a test set in order to obtain an estimator of the true error rate of the rules. In this way we will be able to compare the estimators of the true error rate previously defined with another value obtained from an independent sample and evaluate the behavior of the true error rate estimators in this example.

Table **??** contains the results obtained with the 9 estimators of the true error rate considered in the paper and the independent estimation obtained from D_2. The bootstrap values have been obtained generating $B = 100$ bootstrap samples as in the sim-

Table 4 Estimations of the true error rate of the rules

Estimator	Fisher	$\gamma = 0$	$\gamma = 0.5$	$\gamma = 1$
APP	30.14%	32.88%	41.10%	50.68%
CV	36.99%	36.99%	43.84%	49.32%
LOOBT	36.49%	39.05%	45.52%	48.93%
*BT*632	34.15%	36.78%	43.89%	49.57%
BCV	29.37%	32.95%	41.07%	45.15%
*BT*2	–	35.53%	35.07%	34.76%
*BT*3	–	41.77%	43.45%	42.80%
*BT*2*CV*	–	29.05%	29.05%	29.33%
*BT*3*CV*	–	34.67%	36.71%	37.03%
Estimation from D_2	39.83%	36.44%	35.59%	33.90%

ulations section. There are several questions that are worth noticing. One of them is the fact that, as mentioned in Section **??**, *APP* increases with γ, which is logical since the rules with higher values of γ are less dependent from the original training sample. Another interesting fact is that, for the data in the example, *APP* is higher than the independent estimation of the true error obtained from D_2. This is not usual for Fisher's rule although it may happen more frequently for the new rules since *APP* usually increases and the true error decreases with γ. Notice, however, that from the results obtained in the simulations section *APP* still has a negative bias as estimator of the true error rate. We can see that, even in this not standard case, the *BT*2 estimator, which had the second best behavior in the estimations, has a very good performance, for all the values of γ considered.

7 Discussion

Fernández et al. (2006) defined new classification rules that take into account the additional information that is frequently available in classification problems. They showed that these rules have lower misclassification error than the usual Fisher's rule. However, the question of the estimation of the true error rate, i.e. the error rate of a given training sample, for those rules is a very important problem that has not been considered so far.

In this paper we check that the true error rate of the new rules has a different behavior than that of Fisher's rule. Namely, in Proposition **??** we prove that the expected apparent error of these rules is higher than that of Fisher's rule. As the true error rate of Fisher's rule is higher than that of the new rules, this means that these new rules do not suffer so much overfitting as Fisher's rule. Consequently, the usual procedures for estimating the true error rate such as *CV*, *LOOBT*, *BT*632 or *BCV* do not work as well as they should and new estimators for the true error rate of these new rules are needed. We consider 2 methods based on different bootstrap procedures that take into account the additional information available on the problem. The first one, that we denote as *BT*2, adjusts the cone of restrictions to the training sample while the second, denoted as *BT*3, adjusts the training sample to the cone of restrictions. The

corresponding cross-validation after bootstrap versions of these procedures, $BT2CV$ and $BT3CV$, are also considered.

Based on a simulation study and on the results obtained with a real data application we check that the new procedures $BT2$ and $BT3$ generally perform better as estimators of the true error rate, E_n, than the usual estimators designed for rules that do not account for additional information. Their performance is especially good for situations where the populations are not too separated. This is the scenario where the new rules are more interesting since it is the case where training samples not fulfilling the restrictions are more likely to appear.

We can also notice that for these rules it is not necessary to perform cross-validation after bootstrap, since $BT2CV$ and $BT3CV$ do not behave better than $BT2$ or $BT3$. Therefore, we conclude with the recommendation of estimators $BT2$ and $BT3$ to evaluate the true error rate of the discrimination rules defined in Fernández et al. (2006).

Acknowledgements This research was partially supported by Spanish DGES grant MTM2012-37129.

References

1. Beran, R., Dümbgen, L. (2010). Least squares and shrinkage estimation under bimonotonicity constraints. Statistics and Computing 20(2), 177-189.
2. Braga-Neto, U.M., Dougherty, E.R. (2004). Is cross-validation valid for small-sample microarray classification? Bioinformatics 20, 374-380.
3. Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association 78, 316-331.
4. Efron, B., Tibshirani, R. (1997). Improvement on cross-validation: the 632+bootstrap method. Journal of the American Statistical Association 92, 548-560.
5. Fernández, M.A., Rueda, C., Salvador, B. (2006). Incorporating additional information to normal linear discriminant rules. Journal of the American Statistical Association 101, 569-577.
6. Fu, W.J., Carroll, R.J., Wang, S. (2005). Estimating misclassification error with small samples via bootstrap cross-validation. Bioinformatics 21, 1979-1986.
7. Kim, J., Cha, E., (2006). Estimating prediction errors in binary classification problem: Cross-validation versus bootstrap. Korean Communications in Statistics 13, 151-165.
8. Kim, J–H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. Computational Statistics and Data Analysis 53(11), 3735-3745.
9. Lachenbruch, P., Mickey, M. (1968). Estimation of error rates in discriminant analysis. Technometrics 10, 167-178.
10. Lehmann, E.L., Casella, G. (1998). Theory of Point Estimation, 2nd edition. Springer-Verlag, New York.
11. Long, T., Gupta, R D. (1998). Alternative Linear Classification Rules Under Order Restrictions. Communications in Statistics, Part A- Theory and Methods, 27, 559-575.
12. McLachlan, G.J. (1976). The bias of the apparent error rate in discriminant analysis. Biometrika 63, 239-244.
13. Molinaro, A.M., Simon, R., Pfeiffer, R.M. (2005). Prediction error estimation: a comparison of resampling methods. Bioinformatics 15, 3301-3307.
14. Oh, M.-S., Shin, D.W. (2011). A unified Bayesian inference on treatment means with order constraints. Computational Statistics and Data Analysis 55(1) 924-934.
15. Robertson, T., Wright, F. T., Dykstra, R. L. (1988). Order Restricted Statistical Inference. Wiley, New York.
16. Salvador, B., Fernández, M.A., Martín, I., Rueda, C. (2008). Robustness of Classification Rules that Incorporate Additional Information. Computational Statistics and Data Analysis. 52(5), 2489-2495.

17. Schiavo, R. A., Hand, D. J. (2000). Ten More Years of Error Rate Research. International Statistical Review 68, 295–310.
18. Steele, B.M., Patterson, D.A., (2000). Ideal bootstrap estimation of expected prediction error for k-nearest neighbor classifiers: Applications for classification and error assessment. Statistics and Computing 10(4), 349-355.
19. UICC (2009). TNM Classification of Malignant Tumours, 7th edition. New Jersey: Wiley-Blackwell.
20. Wehberg, S., Schumacher, M. (2004). A comparison of nonparametric error rate estimation methods in classification problems. Biometrical Journal. 46, 35-47.