Dimitrios Ververidis and Constantine Kotropoulos, "Information loss of the Mahalanobis distance in high dimensions: Application to feature selection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 12, pp. 2275-2281, 2009.

Information loss of the Mahalanobis distance in high dimensions: Application to feature selection

Dimitrios Ververidis* and Constantine Kotropoulos, *Senior Member, IEEE*

***Abstract*— The Mahalanobis distance between a pattern measurement vector of dimensionality D and the center of the class it belongs to is distributed as a χ^2 with D degrees of freedom, when an infinite training set is used. However, the distribution of Mahalanobis distance becomes either Fisher or Beta depending on whether cross-validation or re-substitution is used for parameter estimation in finite training sets. The total variation between χ^2 and Fisher as well as between χ^2 and Beta allows us to measure the information loss in high dimensions. The information loss is exploited then to set a lower limit for the correct classification rate achieved by the Bayes classifier that is used in subset feature selection.**

***Index Terms*— Bayes classifier, Gaussian distribution, Mahalanobis distance, feature selection, cross-validation.**

I. Introduction

The accurate prediction of the error committed by a classifier, P_e, enables to measure the probability k out of $N_{\mathcal{T}}$ test patterns are misclassified. The latter probability is given by $P(k) = \binom{N_{\mathcal{T}}}{k} P_e^k (1-P_e)^{N_{\mathcal{T}}-k}$, because the random variable (r.v.) $\mathbf{k}$, that models the number of misclassified patterns, follows the binomial distribution. Accordingly, confidence limits for $P(k)$ can be easily set [1]. For a two-class pattern recognition problem, an upper limit for P_e is $P_{e,s} = P^s(\Omega_1)P^{1-s}(\Omega_2)\sum_i p^s(\underline{\mathrm{x}}_i|\Omega_1)p^{1-s}(\underline{\mathrm{x}}_i|\Omega_2)$, where $s \in [0,1]$, Ω_c denotes the cth class with $c = 1, 2$, $P(\Omega_c)$ is the a priori probability of the cth class, $\underline{\mathrm{x}}_i$ is a pattern measurement vector, and $p(\underline{\mathrm{x}}|\Omega_c)$ is the class conditional probability density function (pdf) [2]. The accuracy of $P_{e,s}$ depends on how well $p(\underline{\mathrm{x}}|\Omega_c)$ is estimated. Whenever $p(\underline{\mathrm{x}}_i|\Omega_c)$ is modeled as a Gaussian pdf, as is frequently assumed, the accurate estimation of the Mahalanobis distance between a measurement vector and a class center becomes significant. In this paper, the prediction error of the Bayes classifier is studied, when each class pdf is modeled by a multivariate Gaussian.

Several expressions relate the accuracy of the prediction error estimate with the number of measurement vectors per Ω_c in the design set $\mathcal{D}$ (denoted as $N_{\mathcal{D}c}$) and the dimensionality of the measurement vectors (denoted as D) [3]–[8]. For example, it is proposed that the ratio $N_{\mathcal{D}c}/D$ should be greater than 3 in order to obtain an accurate prediction error estimate [3]. In [4] and [6], experiments have demonstrated that this ratio should be at least 10. In [9]–[11], it has been found that as $N_{\mathcal{D}c}/D \to 1$, the prediction error estimated by cross-validation approaches that of the random choice. This effect is often called *curse of dimensionality*, and it is attributed to the sparseness of the measurement vectors in high dimensional spaces, which impedes the accurate estimation of the class-conditional pdfs [10]. If re-substitution is used to estimate the prediction error of the Bayes classifier, then it has been found by experiments that the prediction error tends to zero as $N_{\mathcal{D}c}/D \to 1$. In this case, although the training and test sets also contain sparse measurement vectors and their cardinality is of the same order of magnitude as in cross-validation, sparseness does not explain convincingly the curse of dimensionality.

In this paper, we study the behavior of the Mahalanobis distance pdf as D increases, while $N_{\mathcal{D}c}$ is kept constant. As D approaches $N_{\mathcal{D}c} - 1$, the class conditional dispersion matrix becomes singular. To avoid singularity, the class conditional dispersion matrix can be weighted by the *gross dispersion matrix* (i.e., the sample dispersion matrix of all training measurement vectors ignoring the class information) [12]. Alternatively, instead of the sample dispersion matrix, the *first-order tree-type* representation of the covariance matrix can be used [8]. However, the aforementioned proposals are only remedies. As $D \to N_{\mathcal{D}c} - 1$, only confidence limits of the correct classification rate (CCR=1-P_e) can be set, because there is no sufficient information to estimate accurately the covariance matrices. This is why we focus on the derivation of a lower limit for the CCR as a function of D and $N_{\mathcal{D}c}$.

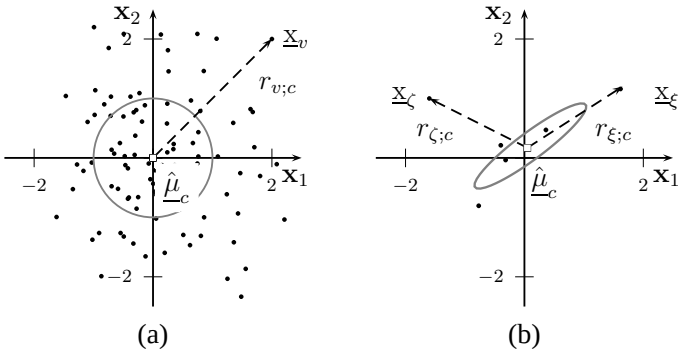
Let $r_{\underline{\mathrm{x}};c}$ be the Mahalanobis distance of measurement vector $\underline{\mathrm{x}}$ from the center of class Ω_c. It is proved that $r_{\underline{\mathrm{x}};c}$ is distributed as: a) a χ^2_D r.v., when an infinite training set is used; b) a Fisher-Snedecor r.v., when the training set is finite and cross-validation is used for parameter estimation; or c) a Beta distribution, when the training set is finite and re-substitution is used for parameter estimation. The difference between χ^2_D and either Fisher-Snedecor or Beta is small, when D is small and $N_{\mathcal{D}c}$ is large. However, as $D \to N_{\mathcal{D}c} - 1$, both Fisher-Snedecor and Beta distributions deviate significantly from χ^2_D. The *total variation* between two-distributions, which is a special form of an f-divergence [13], is used to define a quantity termed information loss for the distribution pairs χ^2_D and Fisher-Snedecor as well as χ^2_D and Beta. As a by product of the aforementioned analysis, a lower limit for CCR is set, that is tested for subset feature selection.

The outline of the paper is as follows. In Section II, the distributional properties of the Mahalanobis distance in high dimensions are studied and the proposed information loss is analytically derived. These findings are exploited for feature selection in Section III, where experimental results on real data-sets are demonstrated. Finally, conclusions are drawn in Section IV.

II. The Mahalanobis distance of a measurement vector from a class center

Let $\mathcal{U}_{\mathcal{A}} = \{u_i\}_{i=1}^N$ be the available set of patterns. Each $u_i = (\underline{\mathrm{x}}_i, c_i)$ consists of a measurement vector denoted as

* Corresponding author. Dr. Dimitrios Ververidis email: jimver04@yahoo.com, jimver@otenet.gr, jimver@aiia.csd.auth.gr Dept. of Informatics, Aristotle University of Thessaloniki Univ. Campus, Biology Dept. Building Box 451, GR-541 24 Thessaloniki, Greece Tel: +30 2310 996361 Fax: +30 2310 998453 Associate Professor Constantine Kotropoulos email: costas@aiia.csd.auth.gr, Dept. of Informatics, Aristotle University of Thessaloniki Univ. Campus, New Building of the Faculty of Applied Sciences Box 451, GR-541 24 Thessaloniki, Greece Tel.,Fax: +30 2310 998225



(a) (b)

Fig. 1. Gaussian models with the contour of unit Mahalanobis distance overlaid when: (a) $\hat{\underline{\mu}}_c$ and $\hat{\underline{\Sigma}}_c$ are estimated from $N_{\mathcal{D}c} = \infty$ design measurement vectors; (b) $\hat{\underline{\mu}}_c$ and $\hat{\underline{\Sigma}}_c$ are estimated from $N_{\mathcal{D}c}$=5 measurement vectors and $\underline{\mathrm{x}}_\xi$ is used in the estimation of $\underline{\Sigma}_c$ (re-substitution method), whereas $\underline{\mathrm{x}}_\zeta$ is *not* used in estimating $\underline{\Sigma}_c$ (cross-validation method).

$\underline{\mathrm{x}}_i = [\mathrm{x}_{i1}\ \mathrm{x}_{i2}\ \ldots\ \mathrm{x}_{id}\ \ldots\ \mathrm{x}_{iD}]^T$, and a label $c_i \in \{\Omega_1, \ldots, \Omega_c, \ldots, \Omega_C\}$. Let the superscript $\mathcal{W}$ in the notation $\mathcal{U}_\mathcal{A}^\mathcal{W}$ explicitly indicate the features extracted from $\mathcal{U}_\mathcal{A}$. In *s-fold cross-validation*, the data are split into s folds and $N_\mathcal{D} = \frac{s-1}{s}N$ patterns are randomly selected without re-substitution from $\mathcal{U}_\mathcal{A}^\mathcal{W}$ to build the design set $\mathcal{D}$, while the remaining N/s patterns form the test set $\mathcal{T}$. In the experiments, s equals 10. The design set is used to estimate the parameters of the class-conditional pdf, while the test set is used in classifier performance assessment. In practice, B cross-validation repetitions are made in order to collect enough test samples during classifier performance assessment. This cross-validation variant can be considered as a 10-fold cross-validation repeated many times. For example, when $B = 60$, this cross-validation variant is the 10-fold cross-validation repeated 6 times. Details for the estimation of B can be found in Section III. Let us denote the covariance matrix and the center vector of Ω_c, that are estimated from design set, as $\hat{\underline{\mu}}_{\mathcal{D}bc}$ and $\hat{\underline{\Sigma}}_{\mathcal{D}bc}$, respectively, where the additional subscript b indicates the cross-validation repetition. Throughout the paper, it is assumed that the class-conditional pdf is given by

$$p_b(\underline{\mathrm{x}}|\Omega_c) = f_{\mathcal{MVN}_D}(\underline{\mathrm{x}}|\hat{\underline{\mu}}_{\mathcal{D}bc}, \hat{\underline{\Sigma}}_{\mathcal{D}bc}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\hat{\underline{\Sigma}}_{\mathcal{D}bc}|^{\frac{1}{2}}} \times \exp\{-\frac{1}{2}\underbrace{(\underline{\mathrm{x}} - \hat{\underline{\mu}}_{\mathcal{D}bc})^T \hat{\underline{\Sigma}}_{bc}^{-1}(\underline{\mathrm{x}} - \hat{\underline{\mu}}_{\mathcal{D}bc})}_{r_{\underline{\mathrm{x}};c}}\}, \quad (1)$$

where $f_{\mathcal{MVN}_D}(\underline{\mathrm{x}}|\underline{\mu}, \underline{\Sigma})$ is the multivariate Gaussian pdf and D is the cardinality of $\mathcal{W}$.

In *re-substitution*, the whole set $\mathcal{U}_\mathcal{A}^\mathcal{W}$ is used to estimate the covariance matrix and the center vector of each class in a single repetition of the experiment. Then, $\hat{\underline{\mu}}_c$ and $\hat{\underline{\Sigma}}_c$ are estimated from $\mathcal{U}_{\mathcal{A}c}^\mathcal{W} = \{u_i \in \mathcal{U}_\mathcal{A}^\mathcal{W} | c_i \in \Omega_c\}$.

We examine the distributional properties of $r_{\underline{\mathrm{x}};c}$ for infinite many training patterns as well as for a finite number of training patterns, when the class-conditional pdf parameters are estimated by either cross-validation or re-substitution. To stimulate the reader to assess the analytical findings, first an example for a single class is discussed, where $D = 2$, $\underline{\Sigma}_c = I$, and $\underline{\mu}_c = \underline{0}$. The example highlights the problem that is addressed next for arbitrary many classes, D, $\underline{\Sigma}_c$, and $\underline{\mu}_c$.

Infinite case: The estimated Gaussian model for $N_{\mathcal{D}c} = \infty$ is plotted in Figure 1(a). Let $\underline{\mathrm{x}}_v$ be a measurement vector that stems from Ω_c. Let also $r_{v;c}$ be the Mahalanobis distance of $\underline{\mathrm{x}}_v$ from $\hat{\underline{\mu}}_c$. From the overlaid contour, it is seen that $r_{v;c}$ is accurately estimated, because $\hat{\underline{\Sigma}}_c$ can be estimated accurately. Accordingly, the CCR predicted by any classifier that employs this class-conditional pdf is expected to be accurate.

Finite case with the re-substitution method: The estimated Gaussian model for $N_{\mathcal{D}c} = 5$ is shown in Figure 1(b). From the inspection of Figure 1(b), it is inferred that 5 measurement vectors are not enough to accurately estimate the covariance matrix. Let $\underline{\mathrm{x}}_\xi$ be one among the 5 measurement vectors that are taken into account in the in the derivation of $\hat{\underline{\Sigma}}_c$. The Mahalanobis distance between $\underline{\mathrm{x}}_\xi$ and $\hat{\underline{\mu}}_c$ is denoted as $r_{\xi;c}$. Obviously $\hat{\underline{\Sigma}}_c$ bears information about $\underline{\mathrm{x}}_\xi$, as it is manifested by the eigenvector associated to its largest eigenvalue that is in the direction of $\underline{\mathrm{x}}_\xi$. So, $\underline{\mathrm{x}}_\xi$ is found to be too close to $\underline{\mu}_c$ with respect to the Mahalanobis distance. Therefore, $\underline{\mathrm{x}}_\xi$ is likely to be classified into Ω_c. Since all measurement vectors are used to estimate the covariance matrix, the CCR will tend to 1.

Finite case with the cross-validation method: Let $\underline{\mathrm{x}}_\zeta \in \Omega_c$. However, $\underline{\mathrm{x}}_\zeta$ is not exploited to derive $\hat{\underline{\Sigma}}_c$, as is depicted in Figure 1(b). The Mahalanobis distance between $\underline{\mathrm{x}}_\zeta$ and $\hat{\underline{\mu}}_c$ is denoted as $r_{\zeta;c}$. Since $\underline{\mathrm{x}}_\zeta$ has been ignored in $\hat{\underline{\Sigma}}_c$, the mode of variation in the direction of $\underline{\mathrm{x}}_\zeta$ is not captured adequately by $\hat{\underline{\Sigma}}_c$. The eigenvalue corresponding to the closest eigenvector of $\hat{\underline{\Sigma}}_c$ to $\underline{\mathrm{x}}_\zeta$ is small, i.e. the Mahalanobis distance of $\underline{\mathrm{x}}_\zeta$ from the center of the class is large, and $\underline{\mathrm{x}}_\zeta$ will probably be misclassified.

In the following, the pdfs of the r.vs. $\mathbf{r}_{v;c}, \mathbf{r}_{\xi;c}$, and $\mathbf{r}_{\zeta;c}$, are derived and the information loss is measured for finite training sets.

***Theorem** 1:* The total variation between the pdfs of $\mathbf{r}_{v;c}$ and $\mathbf{r}_{\zeta;c}$ causes the information loss in cross-validation given by

$$\mathrm{LOSS}_{cross}(N_{\mathcal{D}c}, D) = F_{\chi^2_D}(t_1) - I_{\frac{1}{1+\frac{N_{\mathcal{D}c}^2-1}{N_{\mathcal{D}c}}\frac{1}{t_1}}}\Big(\frac{D}{2}, \frac{N_{\mathcal{D}c}-D}{2}\Big), \quad (2)$$

where

$$t_1 = -N_c W_{-1}\Big(-\big[\frac{\Gamma(\frac{N_{\mathcal{D}c}}{2})}{\Gamma(\frac{N_{\mathcal{D}c}-D}{2})}\big]^{\frac{2}{N_{\mathcal{D}c}}} \frac{2^{\frac{D}{N_{\mathcal{D}c}}} N_{\mathcal{D}c}^{\frac{D}{N_{\mathcal{D}c}}-2}}{(N_{\mathcal{D}c}^2-1)^{\frac{D}{N_{\mathcal{D}c}}-1}} \times \exp\big(\frac{1-N_{\mathcal{D}c}^2}{N_{\mathcal{D}c}^2}\big)\Big) - N_{\mathcal{D}c} + \frac{1}{N_{\mathcal{D}c}^2}, \quad (3)$$

$F_{\chi^2_D}(x)$ is the cdf of χ^2_D, $I_x(a,b)$ is the incomplete Beta function with parameters a and b, and $W_k(x)$ is the kth branch of Lambert's W function [14].

Proof: According to Theorem 3 in Appendix I, $\mathbf{r}_{v;c}$ is distributed as

$$f_{\mathbf{r}_{v;c}}(r) = f_{\chi^2_D}(r) = \frac{(\frac{1}{2})^{\frac{D}{2}}}{\Gamma(\frac{D}{2})} r^{\frac{D}{2}-1} e^{-\frac{r}{2}} \quad (4)$$

with $f_{\chi^2_D}(x)$ being the pdf of χ^2_D. Theorem 4 in Appendix I dictates that $\mathbf{r}_{\zeta;c}$ has pdf

$$f_{\mathbf{r}_{\zeta;c}}(r) = \frac{N_{\mathcal{D}c}(N_{\mathcal{D}c}-D)}{(N_{\mathcal{D}c}^2-1)D} f_{\mathcal{F}isher}\Big(\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c}-D)}{(N_{\mathcal{D}c}^2-1)D} r | D, N_{\mathcal{D}c}-D\Big) \quad (5)$$

where $f_{\mathcal{F}isher}(x|a,b)$ is the pdf of the Fisher-Snedecor distribution with parameters a and b. Both $f_{\mathbf{r}_{v;c}}(r)$ and $f_{\mathbf{r}_{\zeta;c}}(r)$ are plotted in Figure 2, when $D = 6$. It is seen that for $N_{\mathcal{D}c} = 10$,

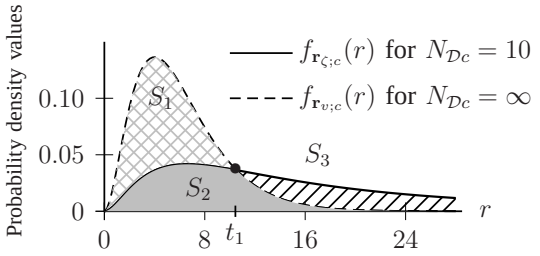


Fig. 2. The distribution of the Mahalanobis distance for $D = 6$ when a) $N_{\mathcal{D}c} = \infty$; and b) $N_{\mathcal{D}c} = 10$ and cross-validation is used.

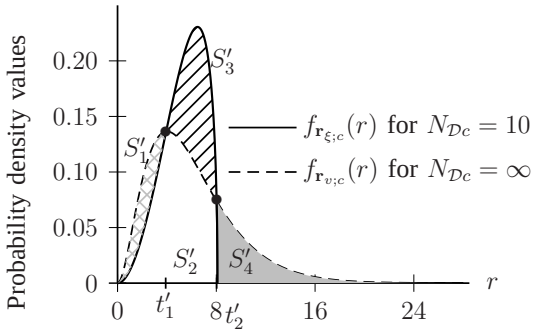


Fig. 3. The distribution of the Mahalanobis distance for $D = 6$ when a) $N_{\mathcal{D}c} = \infty$ and b) $N_{\mathcal{D}c} = 10$ and re-substitution is used.

$f_{\mathbf{r}_{v;c}}(r)$ intersects $f_{\mathbf{r}_{\zeta;c}}(r)$ at $r = t_1$ given by (3). The derivation of (3) can be found in Appendix II. Let us examine the area under each pdf. Since $S_1 + S_2 = S_3 + S_2 = 1$, we have $S_1 = S_3$. Let $\mathrm{LOSS}_{cross}(N_{\mathcal{D}c}, D) \equiv \int_0^{t_1} (f_{\mathbf{r}_{v;c}}(r) - f_{\mathbf{r}_{\zeta;c}}(r))dr$ be termed as the information loss in cross-validation. The information loss (i.e., the area S_3) is simply one half of the total variation between the aforementioned pdfs, which equals $S_1 + S_3 = 2\,S_3$. The area S_3 is given by (2). ■

***Theorem** 2:* The total variation between the distributions of $\mathbf{r}_{v;c}$ and $\mathbf{r}_{\xi;c}$ causes the information loss in re-substitution method given by

$$\mathrm{LOSS}_{resub}(N_{\mathcal{D}c}, D) = I_{\frac{N_{\mathcal{D}c}t'_2}{(N_{\mathcal{D}c}-1)^2}}\Big(\frac{D}{2}, \frac{N_{\mathcal{D}c}-D-1}{2}\Big) - I_{\frac{N_{\mathcal{D}c}t'_1}{(N_{\mathcal{D}c}-1)^2}}\Big(\frac{D}{2}, \frac{N_{\mathcal{D}c}-D-1}{2}\Big) - F_{\chi^2_D}(t'_2) + F_{\chi^2_D}(t'_1), \quad (6)$$

where for $\ell = 1, 2$

$$t'_\ell = (N_{\mathcal{D}c} - D - 3)W_{1-\ell}\Big(\frac{(N_{\mathcal{D}c}-1)^{\frac{2N_{\mathcal{D}c}-6}{N_{\mathcal{D}c}-D-3}}(2N_{\mathcal{D}c})^{\frac{D}{3+D-N_{\mathcal{D}c}}}}{(3+D-N_{\mathcal{D}c})N_{\mathcal{D}c}} \times \Big[\frac{\Gamma(\frac{(N_{\mathcal{D}c}-1)}{2})}{\Gamma(\frac{N_{\mathcal{D}c}-D-1}{2})}\Big]^{\frac{2}{3+D-N_{\mathcal{D}c}}} e^{\frac{(N_{\mathcal{D}c}-1)^2}{N_{\mathcal{D}c}(3+D-N_{\mathcal{D}c})}}\Big) + \frac{(N_{\mathcal{D}c}-1)^2}{N_{\mathcal{D}c}}. \quad (7)$$

Proof: According to Theorem 5 in Appendix I, the density of $\mathbf{r}_{\xi;c}$ is given by

$$f_{\mathbf{r}_{\xi;c}}(r) = \frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2}\, f_{\mathcal{B}eta}\Big(\frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2}r \Big| \frac{D}{2}, \frac{N_{\mathcal{D}c}-D-1}{2}\Big). \quad (8)$$

where $f_{\mathcal{B}eta}(x|a, b)$ is the pdf of the beta distribution with parameters a and b. The distribution of $\mathbf{r}_{v;c}$ is given by (4). The total variation between the distributions $f_{\mathbf{r}_{v;c}}(r)$ and $f_{\mathbf{r}_{\xi;c}}(r)$ equals the area $S'_1 + S'_3 + S'_4$. However, by examining the areas below the pdfs in Figure 3, one finds that $S'_1 + S'_2 + S'_4 = S'_2 + S'_3 = 1 \Rightarrow S'_3 = S'_1 + S'_4$. That is, the total variation is simply twice S'_3. Let us define the information loss in re-substitution as one-half of the total variation. Then, $\mathrm{LOSS}_{resub}(N_{\mathcal{D}c}, D) = S'_3 = \int_{t'_1}^{t'_2} (f_{\mathbf{r}_{\xi;c}}(r) - f_{\mathbf{r}_{v;c}}(r))dr$, which is given by (6), where t'_1 and t'_2 are the abscissas of the points, where $f_{\mathbf{r}_{v;c}}(r)$ intersects $f_{\mathbf{r}_{\xi;c}}(r)$. In Appendix II, it is proved that t'_ℓ, $\ell = 1, 2$ are given by (7). ■

The functions $\mathrm{LOSS}_{cross}(N_{\mathcal{D}c}, D)$ and $\mathrm{LOSS}_{resub}(N_{\mathcal{D}c}, D)$ for $N_{\mathcal{D}c} = 50, 200, 500, 10^3$ and $D = 1, 2, \ldots, N_{\mathcal{D}c}-1$ are plotted in Figures 4 and 5, respectively. The information loss in cross-validation is more severe than in re-substitution. For example, let us examine the information loss in both cases for $N_{\mathcal{D}c} = 10^3$ and $D = 200$. In cross-validation, the information loss is found to be 0.75, whereas in re-substitution is only 0.05.

III. Application to feature selection

In this section, we set a lower limit for CCR, that is estimated by either cross-validation or re-substitution based on the analytical results of Section II. Throughout the section, the Bayes classifier is used and the CCR of this classifier is studied.

Let $\mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}})$ be the cross-validation estimate of CCR , when B cross-validation repetitions are employed. $\mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}})$ is actually the average over $b = \{1, 2, \ldots, B\}$ CCRs, that are measured as follows. Let $\mathcal{L}[c_i, \hat{c}_i]$ denote the *zero-one loss function* between the ground truth label c_i and the predicted class label $\hat{c}_i$ determined by the Bayes classifier for u_i, i.e.

$$\mathcal{L}[c_i, \hat{c}_i] = \begin{cases} 1 & \text{if } c_i = \hat{c}_i, \\ 0 & \text{if } c_i \neq \hat{c}_i. \end{cases} \text{ where } \hat{c}_i = \underset{c=1}{\overset{C}{\operatorname{argmax}}}\{p_b(\underline{\mathbf{x}}_i|\Omega_c)P(\Omega_c)\}. \quad (9)$$

The cross-validation estimate of CCR is given by

$$\mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CCR}_b(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) = \frac{1}{B}\frac{1}{N_{\mathcal{T}}}\sum_{b=1}^{B}\sum_{u_i \in \mathcal{U}_{\mathcal{T}b}^{\mathcal{W}}} \mathcal{L}[c_i, \hat{c}_i]. \quad (10)$$

where $\mathcal{U}_{\mathcal{T}b}^{\mathcal{W}}$ is the test set during the bth iteration.B is estimated as in [15]. The higher the B, the less varies $\mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}})$. The estimate of CCR in re-substitution is obviously given by

$$\mathrm{CCR}_{resub}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) = \frac{1}{N}\sum_{u_i \in \mathcal{U}_{\mathcal{A}}^{\mathcal{W}}} \mathcal{L}[c_i, \hat{c}_i]. \quad (11)$$

The information loss implies that accurate estimates of CCRs can not be obtained. Therefore, we propose a lower limit for CCR in either cross-validation or re-substitution, that is expressed as function of the information loss of the Mahalanobis distance. In particular, the information loss is subtracted from CCR and a normalization term is added to guarantee a lower limit above

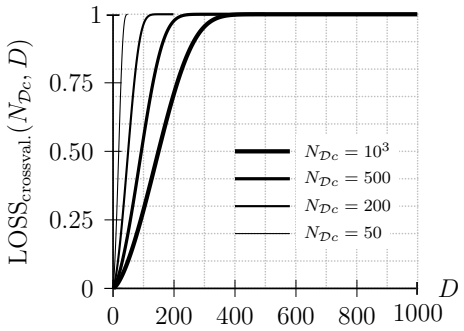


Fig. 4. The information loss in cross-validation.

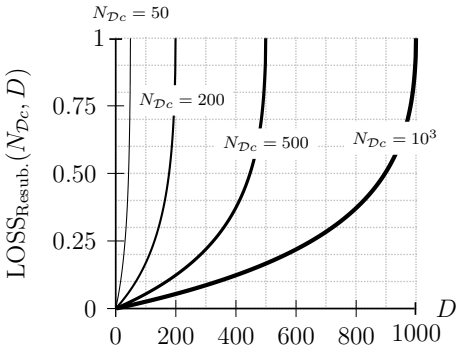


Fig. 5. The information loss in re-substitution.

$1/C$:

$$\mathrm{CCR}_{resub}^{Lower}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) = \mathrm{CCR}_{resub}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) - \mathrm{LOSS}_{resub}(N_{\mathcal{D}c}, D) \times [\mathrm{CCR}_{resub}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) - \frac{1}{C}], \quad (12)$$

$$\mathrm{CCR}_{B,cross}^{Lower}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) = \mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) - \mathrm{LOSS}_{cross}(N_{\mathcal{D}c}, D) \times [\mathrm{CCR}_{B,cross}(\mathcal{U}_{\mathcal{A}}^{\mathcal{W}}) - \frac{1}{C}]. \quad (13)$$

Such a lower limit is exploited to select the optimum feature subset in both cases. Three feature selection algorithms are tested[1], namely: a) the Sequential Forward Selection (SFS) [16]; b) the Sequential Floating Forward Selection (SFFS) [16]; and c) the ReliefF algorithm [17]. SFS starts from an empty feature set and includes one feature at a time. This feature maximizes the CCR. SFFS performs similarly to SFS except a conditional exclusion step that is tested after an inclusion step. In this test, it is tested whether the removal of a previously selected feature increases the CCR. In ReliefF, a stepwise weighting of all features in [-1,1] is performed. At each step, the weights are updated according to two distances, namely the distance between a randomly chosen pattern and the nearest pattern in the same class and that between itself and the nearest pattern of a different class. To evaluate the CCR at each step, only the features with positive weights are retained. Feature selection experiments are conducted on three data-sets:

SUSAS data-set: The Speech Under Simulated and Actual Stress data-set consists of 35 words expressed under several speech styles. The 35 words are related to aircraft environment such as break, change, degree, destination, etc. Each word is repeated twice under 4 speech styles namely neutral, anger, clear, and Lombard. The number of available utterances is thus $N = 2521$. The experiment with SUSAS data-set aims at recognizing the speech style by extracting 90 prosodic features from each utterance, such as the maximum intensity, the maximum pitch frequency to mention a few [18], [19].

Colon cancer data-set: Micro-array snapshots taken on colon cells are used in order to monitor the expression level of thousands of genes at the same time. Cancer cells can be thus separated from normal ones. 62 pattern snapshots that stem from 40 cancer and 22 normal cells are included in the experiments. The extracted features are the 2000 genes that have shown the highest minimal intensity across patterns [20].

Sonar data-set: It contains impinging pulse returns collected from a metal cylinder (simulating a mine) and a cylindrically shaped rock positioned on a sandy ocean floor at several aspect angles. The returns, which are temporal signals, are filtered by 60 sub-band filters in order to extract the spectral magnitude in each band as feature. The data-set consists of 208 returns, that are divided into 111 cylinder returns and 97 rock returns [21].

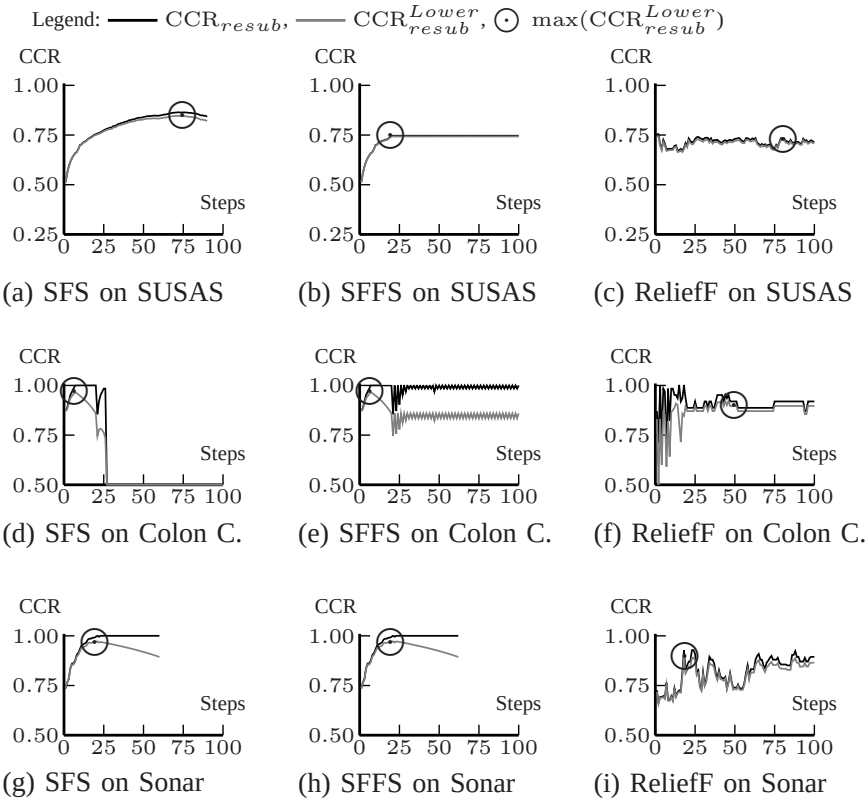
The experiments aim at demonstrating that the maximum value admitted by $\mathrm{CCR}_{B,cross}^{Lower}$ across feature selection steps is the most accurate criterion for determining the optimum feature subset. In particular, the following alternative criteria are considered: $\mathrm{CCR}_{B,cross}$, $\mathrm{CCR}_{resub}^{Lower}$, and CCR_{resub} and comparisons are also made between the CCR achieved when the just-mentioned criteria are employed in subset feature selection and the state of the art CCR reported for each data-set in the literature. More specifically, a 58.57% CCR has been achieved for the SUSAS data-set using hidden Markov models with 6 auto-correlation coefficients [19]. A 88.71% CCR has been reported for the colon cancer data-set using the naive Bayes classifier with 30 genes (features) [22]. Finally, a 84.7% CCR has been measured for the sonar data-set by a neural network using 10 spectral features [21].

The CCR is plotted versus the feature selection steps for both re-substitution and cross-validation in Figures 6 and 7, respectively. The lower limit of CCR predicted by either (12) or (13) is plotted with a gray line. The maximum lower limit marked by a $\odot$ indicates the step when the optimum selected feature subset is derived. From the inspection of Figure 6, it is seen that CCR_{resub} and $\mathrm{CCR}_{resub}^{Lower}$ are significantly higher than the state of art CCR. For example, CCR_{resub} may approach 100% as in Figures 6(d), 6(e), 6(f), 6(g), and 6(h). The same applies for $\mathrm{CCR}_{resub}^{Lower}$. As it is seen in Figure 7, $\mathrm{CCR}_{B,cross}$ is closer to the state of the art. However, $\mathrm{CCR}_{B,cross}$ curve does not often exhibit a clear peak (Figures 7(a), 7(b), and 7(f)). Accordingly, the optimum feature subset can be arbitrarily long. $\mathrm{CCR}_{B,cross}$ can also be untruthfully high approaching 100% sometimes (e.g. Figures 7(d) and 7(e)). On the contrary, the most reliable criterion is $\mathrm{CCR}_{B,cross}^{Lower}$ that exhibits a clear peak (Figures 7(a), 7(b), 7(d), 7(e), 7(g), and 7(h)) close to the state of the art CCR.

IV. Conclusions

In this paper, we have studied the collapse of the correct classification rate estimated by cross-validation as the dimensionality of measurement vectors increases. We have attributed

[1]An implementation of the feature selection algorithm with a graphical user interface that uses the proposed lower limits can be found at `http://www.mathworks.com/matlabcentral/fileexchange/` under 'Feature Selection DEMO in Matlab'.

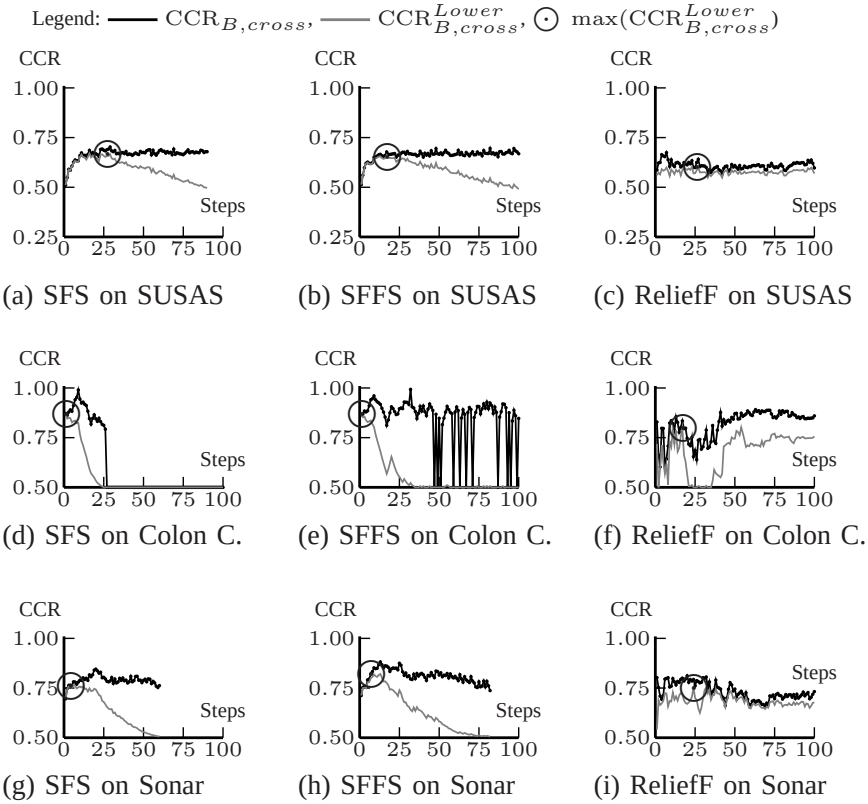


(a) SFS on SUSAS (b) SFFS on SUSAS (c) ReliefF on SUSAS

(d) SFS on Colon C. (e) SFFS on Colon C. (f) ReliefF on Colon C.

(g) SFS on Sonar (h) SFFS on Sonar (i) ReliefF on Sonar

Fig. 6. CCR vs. the feature selection step, when CCR is estimated by re-substitution.



(a) SFS on SUSAS (b) SFFS on SUSAS (c) ReliefF on SUSAS

(d) SFS on Colon C. (e) SFFS on Colon C. (f) ReliefF on Colon C.

(g) SFS on Sonar (h) SFFS on Sonar (i) ReliefF on Sonar

Fig. 7. CCR vs. the feature selection step, when CCR is estimated by cross-validation.

this phenomenon to the inaccurate estimation of the Mahalanobis distance of each measurement vector from the center of a class that causes the inaccurate estimation of the covariance matrix of each class. Furthermore, we have proved that the increase of correct classification rate in re-substitution as dimensionality increases is also due to the inaccurate estimation of the Mahalanobis distance of each measurement vector from the center of the class. To quantify the inaccurate estimation of the Mahalanobis distance, we have derived analytically the information loss with respect to the number of measurement vectors per class and the dimensionality of the measurement vectors for both cross-validation and re-substitution. The information loss has been exploited in setting a lower limit of the correct classification rate that was used in subset feature selection with the Bayes classifier.

Although, class-conditional pdfs were assumed to be multivariate Gaussian for the sake of analytical derivations, the results of the paper can be extended to Gaussian mixtures. Moreover, they can be applied to any feature subset selection method that employs as a criterion the correct classification rate achieved by classifiers, which resort to the Mahalanobis distance, e.g. the k-means. As the best method for feature selection, we propose the SFS where the criterion is as in (13), i.e. the lower limit of CCR found with cross-validation.

APPENDIX I
THEOREMS FOR THE DISTRIBUTION OF THE MAHALANOBIS DISTANCE

Let us assume that $\underline{\mathbf{x}} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_D]^T$ is a D-dimensional random vector of a pattern that belongs to Ω_c distributed according to the multivariate (MV) normal distribution $\mathcal{MVN}_D(\underline{\mu}_c, \underline{\Sigma}_c)$. The sample mean vector $\hat{\underline{\mu}}_c$ and the sample dispersion matrix $\hat{\underline{\Sigma}}_c$ of a set of measurement vectors $\mathcal{X}_{\mathcal{D}c} = \{\underline{\mathrm{x}}_i \in \mathcal{X}_{\mathcal{D}} | c_i \in \Omega_c\}$ of cardinality $N_{\mathcal{D}c}$ are used as estimates of $\underline{\mu}_c$ $\underline{\Sigma}_c$, respectively. Our interest is in the distribution of the Mahalanobis distances $\mathbf{r}_{v;c}$ for infinite many training measurement vectors (case A), $\mathbf{r}_{\zeta;c}$ for finite training measurement vectors when cross-validation is used for parameter estimation (case B), and $\mathbf{r}_{\xi;c}$ for finite training measurement vectors when re-substitution is employed for parameter estimation (case C).

***Theorem** 3:* (Case A) The Mahalanobis distance $\mathbf{r}_{v;c} = (\underline{\mathbf{x}}_v - \underline{\mu}_c)^T \underline{\Sigma}_c^{-1} (\underline{\mathbf{x}}_v - \underline{\mu}_c)$ for the ideal case of infinite many training measurement vectors ($N_{\mathcal{D}c} \to \infty$) is distributed according to χ^2_D [23].

Proof: Let $\underline{\Phi}_c$ be the matrix with columns the eigenvectors of $\underline{\Sigma}_c$ and $\underline{\Lambda}_c$ be the diagonal matrix of eigenvalues of $\underline{\Sigma}_c$. Then, $\underline{\Sigma}_c = \underline{\Phi}_c \underline{\Lambda}_c \underline{\Phi}_c^T$. So,

$$\mathbf{r}_{v;c} = (\underline{\mathbf{x}}_v - \underline{\mu}_c)^T \underline{\Phi}_c \underline{\Lambda}_c^{-1} \underline{\Phi}_c^T (\underline{\mathbf{x}}_v - \underline{\mu}_c) = [(\underline{\mathbf{x}}_v - \underline{\mu}_c)^T \underline{\Phi}_c \underline{\Lambda}_c^{-1/2}] \times [\underline{\Lambda}_c^{-1/2} \underline{\Phi}_c^T (\underline{\mathbf{x}}_v - \underline{\mu}_c)] = \underline{\mathbf{z}}_{v;c}^T \ \underline{\mathbf{z}}_{v;c}, \quad (14)$$

where $\underline{\mathbf{z}}_{v;c} = \underline{\Lambda}_c^{-1/2} \underline{\Phi}_c^T (\underline{\mathbf{x}}_v - \underline{\mu}_c)$ is a random vector consisting of univariate independent normal random variables with zero mean and unit variance. Hence, $\mathbf{r}_{v;c}$ follows the χ^2_D distribution [23], [24]. ■

***Theorem** 4:* (Case B) Let $\underline{\mathbf{x}}_\zeta \notin \mathcal{X}_{\mathcal{D}c}$. The distribution of $\mathbf{r}_{\zeta;c} = (\underline{\mathbf{x}}_\zeta - \hat{\underline{\mu}}_c)^T \hat{\underline{\Sigma}}_c^{-1} (\underline{\mathbf{x}}_\zeta - \hat{\underline{\mu}}_c)$, when $\underline{\mathbf{x}}_\zeta$ is not involved in the estimation of $\hat{\underline{\Sigma}}_c$, and accordingly, $\underline{\mathbf{x}}_\zeta$ and $\hat{\underline{\Sigma}}_c$ as well as $\underline{\mathbf{x}}_\zeta$ and $\hat{\underline{\mu}}_c$ are independent, is

$$f(r_{\zeta;c}) = \frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} f_{\mathcal{F}isher}\Big(\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} r_{\zeta;c} | D, N_{\mathcal{D}c} - D\Big). \quad (15)$$

Proof: Let $\underline{\mathbf{d}}_\zeta = \underline{\mathbf{x}}_\zeta - \hat{\underline{\boldsymbol{\mu}}}_c$, then $\underline{\mathbf{d}}_\zeta = \underline{\mathbf{x}}_\zeta - \frac{\underline{\mathbf{x}}_1 + \underline{\mathbf{x}}_2 + \ldots + \underline{\mathbf{x}}_{N_{\mathcal{D}c}}}{N_{\mathcal{D}c}} \sim \mathcal{MVN}_D(\underline{0}, \frac{N_{\mathcal{D}c}+1}{N_{\mathcal{D}c}} \underline{\Sigma}_c)$. Let

$$\tilde{\underline{\mathbf{d}}}_\zeta = \sqrt{\frac{N_{\mathcal{D}c}}{N_{\mathcal{D}c}+1}} (\underline{\mathbf{x}}_\zeta - \hat{\underline{\boldsymbol{\mu}}}_c) \sim \mathcal{MVN}_D(\underline{0}, \underline{\Sigma}_c). \quad (16)$$

Since $\underline{\mathbf{x}}_\zeta$ is not involved in the estimation of $\hat{\underline{\Sigma}}_c$, we can consider the distance $\tilde{\underline{\mathbf{d}}}_\zeta$ independent of $\hat{\underline{\Sigma}}_c$. So according to Hotelling's Theorem [25], $\boldsymbol{\tau} = \tilde{\underline{\mathbf{d}}}_\zeta^T \hat{\underline{\boldsymbol{\Sigma}}}_c^{-1} \tilde{\underline{\mathbf{d}}}_\zeta$ follows Hotelling distribution, i.e.

$$\frac{\boldsymbol{\tau}}{N_{\mathcal{D}c} - 1} \frac{N_{\mathcal{D}c} - D}{D} \sim f_{\mathcal{F}isher}(\frac{\tau}{N_{\mathcal{D}c} - 1} \frac{N_{\mathcal{D}c} - D}{D} | D, N_{\mathcal{D}c} - D). \quad (17)$$

However,

$$\boldsymbol{\tau} = \sqrt{\frac{N_{\mathcal{D}c}}{N_{\mathcal{D}c}+1}} (\underline{\mathbf{x}}_\zeta - \hat{\underline{\mu}}_c)^T \hat{\underline{\Sigma}}_c^{-1} \sqrt{\frac{N_{\mathcal{D}c}}{N_{\mathcal{D}c}+1}} (\underline{\mathbf{x}}_\zeta - \hat{\underline{\mu}}_c) = \frac{N_{\mathcal{D}c} \mathbf{r}_{\zeta;c}}{N_{\mathcal{D}c}+1}. \quad (18)$$

From (17) and (18), it is inferred that

$$\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} \mathbf{r}_{\zeta;c} \sim f_{\mathcal{F}isher}\Big(\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} r_{\zeta;c} | D, N_{\mathcal{D}c} - D\Big). \quad (19)$$

Given that $\mathbf{x} \sim f(x) \ \Rightarrow \ \mathbf{x}/a \sim a f(x)$ [26], then the distribution of $\mathbf{r}_{\zeta;c}$ is obtained as

$$f(r_{\zeta;c}) = \frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} f_{\mathcal{F}isher}\Big(\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c} - D)}{(N_{\mathcal{D}c}^2 - 1)D} r_{\zeta;c} | D, N_{\mathcal{D}c} - D\Big). \quad (20)$$

The cdf of $\mathbf{r}_{\zeta;c}$ can be found by integrating (20), which according to [24, eq. 26.6.2] yields

$$F(r_{\zeta;c}) = I_{\frac{1}{1 + \frac{N_{\mathcal{D}c}^2 - 1}{N_{\mathcal{D}c}} \frac{1}{r_{\zeta;c}}}} \Big(\frac{D}{2}, \frac{N_{\mathcal{D}c} - D}{2}\Big). \quad (21)$$

■

To find the pdf of $\mathbf{r}_{\xi;c}$ for finite training measurement vectors when re-substitution is employed in parameter estimation (case C), we resort to Lemmata 1 and 2 which are exploited in the proof of Theorem 5.

***Lemma** 1:* If $\sum_{i=1(\xi)}^{N_{\mathcal{D}c}}$ denotes the sum from 1 to $N_{\mathcal{D}c}$ except ξ, $\hat{\underline{\mathrm{A}}}_c = (N_{\mathcal{D}c} - 1)\hat{\underline{\Sigma}}_c$, and

$$\hat{\underline{\mathrm{A}}}_{c(\xi)} = \sum_{1(\xi)}^{N_{\mathcal{D}c}} (\underline{\mathrm{x}}_i - \hat{\underline{\mu}}_{c(\xi)})(\underline{\mathrm{x}}_i - \hat{\underline{\mu}}_{c(\xi)})^T, \ \text{where}$$

$$\hat{\underline{\mu}}_{c(\xi)} = \frac{1}{N_{\mathcal{D}c} - 1} \sum_{1(\xi)}^{N_{\mathcal{D}c}} \underline{\mathrm{x}}_i, \quad (22)$$

then

$$\hat{\underline{\mathrm{A}}}_{c(\xi)} = \hat{\underline{\mathrm{A}}}_c - \frac{N_{\mathcal{D}c}}{N_{\mathcal{D}c} - 1} (\underline{\mathrm{x}}_\xi - \hat{\underline{\mu}}_c)(\underline{\mathrm{x}}_\xi - \hat{\underline{\mu}}_c)^T. \quad (23)$$

Proof: See [27]. ■

***Lemma** 2:* Let $R_\xi = \frac{|\hat{\underline{\mathrm{A}}}_{c(\xi)}|}{|\hat{\underline{\mathrm{A}}}_c|}$ be called as one-outlier scatter ratio of measurement vector $\underline{\mathrm{x}}_\xi$, i.e. it denotes how much differs

the dispersion of the whole set from the same set when $\underline{\mathbf{x}}_\xi$ is excluded, then $\mathbf{R}_\xi \sim f_{\mathcal{B}eta}(R_\xi|\frac{N_{\mathcal{D}c}-D-1}{2}, \frac{D}{2})$, where $f_{\mathcal{B}eta}(x|a,b)$ is the pdf of the beta distribution with parameters a and b.

Proof: See [27]. ■

***Theorem** 5:* If $\mathbf{R}_\xi \sim f_{\mathcal{B}eta}(R_\xi|\frac{N_{\mathcal{D}c}-D-1}{2}, \frac{D}{2})$ then

$$\mathbf{r}_{\xi;c} \sim \frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2} f_{\mathcal{B}eta}\Big(\frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2} r_{\xi;c}|\frac{D}{2}, \frac{N_{\mathcal{D}c}-D-1}{2}\Big). \tag{24}$$

Proof: See [27]. ■

APPENDIX II
ROOTS OF EQUATIONS $f_{\mathbf{r}_{v;c}}(t) = f_{\mathbf{r}_{\zeta;c}}(t)$ AND $f_{\mathbf{r}_{v;c}}(t) = f_{\mathbf{r}_{\xi;c}}(t)$.

The roots of $f_{\mathbf{r}_{v;c}}(t) = f_{\mathbf{r}_{\zeta;c}}(t)$ can be found as follows.

$$f_{\chi^2_D}(t) = \frac{N_{\mathcal{D}c}(N_{\mathcal{D}c}-D)}{(N^2_{\mathcal{D}c}-1)D} f_{\mathcal{F}isher}\Big(\frac{N_{\mathcal{D}c}(N_{\mathcal{D}c}-D)}{(N^2_{\mathcal{D}c}-1)D} t|D, N_{\mathcal{D}c}-D\Big) \quad \Rightarrow \tag{25}$$

$$\frac{(\frac{1}{2})^{\frac{D}{2}}}{\Gamma(\frac{D}{2})} t^{\frac{D}{2}-1} e^{-\frac{t}{2}} = \frac{\Gamma(\frac{N_{\mathcal{D}c}}{2})}{\Gamma(\frac{D}{2})\Gamma(\frac{N_{\mathcal{D}c}-D}{2})} \Big(\frac{N_{\mathcal{D}c}}{N^2_{\mathcal{D}c}-1}\Big)^{\frac{D}{2}} t^{\frac{D}{2}-1} \Big[1+\frac{N_{\mathcal{D}c}}{N^2_{\mathcal{D}c}-1}t\Big]^{-\frac{N_{\mathcal{D}c}}{2}} \Rightarrow \tag{26}$$

$$e^{-\frac{t}{2}} = \underbrace{\frac{\Gamma(\frac{N_{\mathcal{D}c}}{2})}{\Gamma(\frac{N_{\mathcal{D}c}-D}{2})}\Big(\frac{2N_{\mathcal{D}c}}{N^2_{\mathcal{D}c}-1}\Big)^{\frac{D}{2}}}_{a} \Big[1+\underbrace{\frac{N_{\mathcal{D}c}}{N^2_{\mathcal{D}c}-1}}_{b} t\Big]^{-\frac{N_{\mathcal{D}c}}{2}} \Rightarrow$$
$$e^{-\frac{t}{2}} = a(1+bt)^{-N_{\mathcal{D}c}/2} \Rightarrow \tag{27}$$

$$e^{t/N_{\mathcal{D}c}} = \underbrace{ba^{-2/N_{\mathcal{D}c}}}_{m} t + \underbrace{a^{-2/N_{\mathcal{D}c}}}_{d} \Rightarrow e^{t/N_{\mathcal{D}c}} = \underbrace{mt+d}_{\lambda} \Rightarrow$$
$$\underbrace{-\frac{\lambda}{mN_{\mathcal{D}c}}}_{W_k(z)} e^{-\frac{\lambda}{mN_{\mathcal{D}c}}} = \underbrace{-\frac{1}{mN_{\mathcal{D}c}} e^{-\frac{d}{mN_{\mathcal{D}c}}}}_{z}, \tag{28}$$

or $W_k(z)e^{W_k(z)} = z$, where $W_k(z)$ is the kth branch of Lambert's W function [14]. So

$$W_k(-\frac{1}{mN_{\mathcal{D}c}} e^{-\frac{d}{mN_{\mathcal{D}c}}}) = -\frac{\lambda}{mN_{\mathcal{D}c}} \underbrace{\Rightarrow}_{\lambda=mt+d}$$
$$t = -N_{\mathcal{D}c} W_k(-\frac{1}{mN_{\mathcal{D}c}} e^{-\frac{d}{mN_{\mathcal{D}c}}}) - \frac{d}{m}. \tag{29}$$

We are interested for $t \in \mathbb{R}$, thus k is 0 or -1. According to Figure 2, χ^2_D intersects the Fisher-Snedecor distribution for one positive value of t. It is found experimentally that the positive t is given by (29) when $k = -1$. So (3) results.

Following similar lines, the roots of $f_{\mathbf{r}_{v;c}}(t) = f_{\mathbf{r}_{\xi;c}}(t)$ can be found. According to Figure 3, χ^2_D intersects the Beta distribution at two positive values of t'_1 and t'_2. Both branches $k = 1-\ell = 0, -1$ of Lambert's function result in a positive t'. The roots t'_ℓ, $\ell = 1, 2$ can be derived in a similar manner from

$$f_{\chi^2_D}(t') = \frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2} f_{\mathcal{B}eta}(\frac{N_{\mathcal{D}c}}{(N_{\mathcal{D}c}-1)^2} t'|\frac{D}{2}, \frac{N_{\mathcal{D}c}-D-1}{2}). \tag{30}$$

REFERENCES

[1] W. Highleyman, "The design and analysis of pattern recognition experiments," *Bell Syst. Techn. J.*, vol. 41, p. 723, 1962.

[2] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*. N.Y.: Wiley, 1973.

[3] D. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. 18, no. 5, pp. 618–626, 1972.

[4] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error and complexity of classification agorithm in pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 2, no. 3, pp. 242–252, 1980.

[5] K. Fukunaga and R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 8, pp. 873–885, 1989.

[6] S. Raudys and A. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, no. 3, pp. 252–264, 1991.

[7] S. Raudys, "On dimensionality, sample size, and classification error of nonparametric linear classification algorithms," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 19, no. 6, pp. 667–671, 1997.

[8] ——, "First-order tree-type dependence between variables and classification performance," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 2, pp. 233–239, 2001.

[9] P. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. N.J.: Prentice Hall, 1982.

[10] V. Vapnik, *Statistical Learning Theory*. N.Y.: Wiley, 1998.

[11] F. van der Heijden, R. Duin, D. de Ridder, and D. M. J. Tax, *Classification, Parameter Estimation and State Estimation - An Engineering Approach using Matlab*. London: Wiley, 2004.

[12] J. Hoffbeck and D. Landgrebe, "Covariance matrix estimation and classification with limited training data," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, no. 7, pp. 763–767, 1996.

[13] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4394–4412, 2006.

[14] R. Corless, G. Gonnet, D. Hare, D. Jeffrey, and D. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, pp. 329–359, 1996.

[15] D. Ververidis and C. Kotropoulos, "Fast and accurate feature subset selection applied to speech emotion recognition," *Elsevier Signal Processing*, vol. 88, no. 12, pp. 2956–2970, 2008.

[16] P. Pudil, J. Novovicova, and J. Kittler, "Floating search methods in feature selection," *Pattern Rec. Lett.*, vol. 15, pp. 1119–1125, 1994.

[17] I. Kononenko, E. Simec, and M. Sikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, pp. 39–55, 1997.

[18] D. Ververidis and C. Kotropoulos, "Fast sequential floating forward selection applied to emotional speech features estimated on DES and SUSAS data collections," in *Proc. European Signal Processing Conf. (EUSIPCO '06)*, 2006.

[19] B. Womack and J. Hansen, "N-Channel hidden Markov models for combined stressed speech classification and recognition," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 6, pp. 668–667, 1999.

[20] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, and A. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array," *Proc. Natl. Acad. Sci. USA*, vol. 96, no. 12, pp. 6745–6750, 1999.

[21] R. Gorman and T. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, pp. 75–89, 1988.

[22] C. Ding and H. Peng, "Minimum redudancy feature selection from microarray gene expression data," *J. Bioinformatics & Computational Biology*, vol. 3, no. 2, pp. 185–205, 2005.

[23] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correated system variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Mag.*, vol. 50, pp. 157–175, 1900.

[24] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. N.Y.: Dover, 1972.

[25] T. Anderson, *An Introduction to Multivariate Statistics*. N.Y.: Wiley, 1984.

[26] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. N.Y.: McGraw-Hill, 2002.

[27] D. Ververidis and C. Kotropoulos, "Gaussian mixture modeling by exploiting the Mahalanobis distance," *IEEE Trans. Signal Processing*, vol. 56, no. 7B, pp. 2797–2811, 2008.