



Data Adaptive Simultaneous Parameter and Kernel Selection in Kernel Discriminant Analysis (KDA) Using Information Complexity

Caterina Liberati

caterina.liberati@unibo.it

*Dipartimento di Scienze Statistiche "F. Fortunati", Università di Bologna
Bologna, Italy*

J. Andrew Howe

jhowe4@utk.edu

*University of Tennessee, Department of Statistics, Operations, and Management Science
Knoxville, Tennessee 37996, USA*

Hamparsum Bozdogan

bozdogan@utk.edu

*University of Tennessee, Department of Statistics, Operations, and Management Science
Knoxville, Tennessee 37996, USA*

Abstract

Kernel Discriminant Analysis (KDA) is the usual extension of Fisher Linear Discriminant Analysis (FLDA) in a high dimensional feature space via kernel mapping. KDA recently has become a popular classification technique in machine learning and in data mining. The performance of KDA depends very heavily on the choice of the best kernel function for a given data set and the optimal choice of the kernel parameters. In this paper, we develop a novel data adaptive simultaneous parameter and kernel selection approach in KDA using information complexity (ICOMP) type criteria. We achieve this by reducing the multivariate input data into one dimension in order to find a range of the possible values to tune the parameters of the kernel mapping directly from the data rather than using trial-and-error. We tune the parameters of the kernel functions by utilizing the Mahalanobis distance of each point from the multivariate mean (centroid), Jackknife Mahalanobis distance Data Depth (JMDD), and the Smoothed Complexity Mahalanobis distance (SCMD). Such an approach provides the researcher a new and novel method to simultaneously choose optimal tuning parameters of the kernel functions; how to choose the optimal kernel function; and their effect on the KDA classifier using ICOMP. We show numerical examples on real benchmark data sets to illustrate the efficiency and the performance of our new approach in terms of reducing the misclassification error rate.

Keywords: Kernel Discriminant Analysis, Information Criteria, Parameter Tuning; Choice of Kernel Function; and Robust Covariance Estimation.

1. Introduction

In recent years, kernel based methods have been shown to be an excellent choice for solving supervised classification problems. Fisher Linear Discriminant Analysis may also be applied to the same task but it is not as good at capturing the nonlinearly clustered structure present in the data. To overcome such limitations, FLDA has been reformulated in the nonlinear space framework induced by kernel machines, while avoiding the explicit knowledge of the nonlinear mapping [3, 27].

It is well known that the selection of a suitable kernel function is a critical issue when nonlinear hyperplane-based methods such as Kernel Discriminant Analysis (KDA) are used for classification. In the recent literature this critical issue is typically solved by choosing a parameterized family of kernels, (e.g., Gaussian or Polynomial), and tuning the kernel parameters via cross-validation (CV) [11] or using a trial-and-error method. Both of these

types of approaches are too time consuming and computationally very expensive due to the high dimensionality of the feature space. Also, it is arbitrary to use the trial-and-error method to tune the parameters which are not optimal but subjective. Moreover, there is no guarantee that the selected kernel is the optimal choice for a classification task. So far, several authors [20], [23], [4, 2] have proposed the linear combination of kernels formed by a family of different kernel functions and parameters. This transforms the problem of choosing a kernel model into one of finding an optimal linear combination of the members of the kernel family [18]. Using this approach inflates the hypothesis space and forces us to solve the resulting overly-complex optimization problem. The kernel selection problem has been studied by [18], in which an iterative method for fitting KDA using different kernels was developed. [22] reformulated this problem as a convex optimization, but they used the misclassification rate as a model selection criteria. Of course, this is a heuristic index that does not produce any estimates about the complexity (correlational structures) of the model employed. Recently, [14] has proposed a quantitative measure for capturing the degree of agreement between the kernel space and the target variable which is a more efficient method for learning with kernels, but it has its own drawbacks.

In this paper, we introduce and develop a new and novel data adaptive approach to choose the best parameter for the kernel mapping according to the structure of the data. We achieve this by reducing a multivariate input data to a univariate index using the Mahalanobis distance (MD), Jackknife Mahalanobis distance data depth (JMDD), and the Smoothed Complexity Mahalanobis distance (SCMD). Manifestation of singular within-group covariance matrix in KDA is a major problem which has not been addressed in the literature of machine learning. Therefore, in this paper, we also introduce and develop a hybridization of a covariance stabilization method [30] along with robust or smoothed covariance estimation of the *within-group covariance matrices* in KDA in order to avoid singular solutions while performing kernel selection. Finally, a novel criterion using *Information Complexity Theory* is proposed for simultaneous evaluation and selection of different nonlinear mappings and parameters.

The rest of this paper is organized as follows. Section 2 provides a brief outline of the Regularized KDA algorithm and the hybrid method for overcoming ill-posed covariance estimation problems in the feature space. In Section 3, we introduce and explore several dimension reduction techniques. Section 4 contains the analytic formulation of the information complexity criteria and its derivation in the kernel domain as our fitness function. Results from numerical experiments based on real benchmark data sets are shown in Section 5. Finally, conclusions and considerations for further work and developments are discussed in Section 6.

2. Regularized Kernel Discriminant Analysis

Assume that we are given the input data set $\mathcal{I}_{XY} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ of training vectors $\mathbf{x}_i \in \mathcal{X} \subseteq \mathcal{R}^d$. Let the corresponding values of $y_i \in \mathcal{Y} = \{1, -1\}$ be the class indices of the training vectors such that an observation is either in one group or the other. Consider $\Phi : \mathcal{R}^d \rightarrow F$ satisfying Mercer's conditions ([26] and [15]), which corresponds to a nonlinear mapping of the data from the *input space* to a higher dimensional *feature space* defined by a kernel function $k(x, y) = \phi(x)\phi(y)$. The basic idea of KDA is to find the direction in the feature space in which the projections of the two sets are such that the ratio of the scatter

matrices (between and within classes), or the so-called *Rayleigh quotient*

$$J_F = \frac{\alpha' S_B^\Phi \alpha}{\alpha' S_W^\Phi \alpha} \quad (1)$$

is maximized, where S_B^Φ and S_W^Φ are respectively the *between and within covariance matrices* in the feature space. That is,

$$\begin{aligned} S_B^\Phi &= (m_1^\Phi - m_2^\Phi)(m_1^\Phi - m_2^\Phi)' \\ &= (\bar{\kappa}_1 - \bar{\kappa}_2)(\bar{\kappa}_1 - \bar{\kappa}_2)', \end{aligned} \quad (2)$$

$$\begin{aligned} S_W^\Phi &= \sum_{i=1,2} \sum_{x \in X_i} (\Phi(x) - m_i^\Phi)(\Phi(x) - m_i^\Phi)' \\ &= \mathcal{K}\mathcal{K}' - \sum_{k=1}^2 n_k \bar{\kappa}_k \bar{\kappa}_k', \end{aligned} \quad (3)$$

where $\mathcal{K} = \kappa(x_i, x_j)_{(n \times n)}$, $\bar{\kappa}_k = \frac{1}{n_k} \sum_{j \in I_k} \mathcal{K}_j$, and \mathcal{K}_j is the j th column of \mathcal{K} and I_k is the index set of group k .

The kernel discriminant function $f(x)$ for the binary classifier can be written as

$$f(x) = \alpha' \mathcal{K} \quad (4)$$

$$= (m_1^\Phi - m_2^\Phi)' (S_W^\Phi)^{-1} \Phi(x). \quad (5)$$

The classification assignment is defined as follows.

$$\begin{cases} \text{if } |\alpha'(\Phi(x) - \frac{1}{2}(m_1^\Phi + m_2^\Phi))| > 0, \text{ group 1} \\ \text{if } |\alpha'(\Phi(x) - \frac{1}{2}(m_1^\Phi + m_2^\Phi))| < 0, \text{ group 2.} \end{cases} \quad (6)$$

This means that we allocate a given observation to the group to which it is closest. This is equivalent to computing the Mahalanobis squared distance between each observation and the centroids of each of the groups given by

$$D_{jk} = (\Phi(x_j) - m_k^\Phi)' (S_W^\Phi)^{-1} (\Phi(x_j) - m_k^\Phi)'. \quad (7)$$

Since the matrix S_W^Φ is at most of rank $n-1$ the proposed setting is ill-posed, and numerical problems frequently cause the matrix S_W^Φ to not even be positive semi-definite. As such, regularization methods to overcome the singularity and instability are widely applied in the statistical domain [17, 32]. Instead of employing ridge type estimators for the within-group covariance matrix S_W^Φ , as is usually done in the literature, here we first apply the stabilization method proposed by [30]:

1. Find the eigenvectors (V) and eigenvalues (Λ) of $S_p = \frac{1}{n-k} S_W^\Phi$;
2. Calculate the average eigenvalue $\bar{\lambda}$
3. Form a new matrix of eigenvalues based on the following largest dispersion values:
 $\Lambda^* = \text{diag}[\max(\lambda_1, \bar{\lambda}), \dots, \max(\lambda_n, \bar{\lambda})]$
4. Reform the modified within-group scatter matrix: $S_W^* = (V\Lambda^*V^T)/(n-k)$.

In a subsequent step, we smooth the stabilized covariance matrix using the Convex Sum Covariance Estimator (CSE) based on the quadratic loss function used by [28] and [12] who proposed the CSE given by

$$\hat{\Sigma}_{CSE} = \frac{n}{n+m} \hat{\Sigma}_W + (1 - \frac{n}{n+m}) \hat{D}_W, \quad (8)$$

where n is the sample size and $\hat{D}_W = (\frac{1}{p}tr\hat{\Sigma}_W)\mathbf{I}_p$ with p number of variables. For $p \geq 2$, m is chosen to be

$$0 < m < \frac{2[p(1 + \beta) - 2]}{p - \beta}, \quad (9)$$

where

$$\beta = \frac{(tr\hat{\Sigma}_W)^2}{tr(\hat{\Sigma}_W^2)}. \quad (10)$$

This estimator improves upon the $\hat{\Sigma}_W$ by shrinking all the estimated eigenvalues of $\hat{\Sigma}_W$ toward their common mean. Moreover, rather than just naively picking a ridge parameter, we let the data pick it for us. In the literature, there are other smoothed, or robust, covariance matrix estimators which improve the efficiency of the kernel solution in terms of stability and misclassification rate. For example these include:

- Maximum Likelihood/Empirical Bayes
- Stipulated Ridge [29]
- Stipulated Diagonal [29], and others.

3. Dimension Reduction Techniques

Transforming data from a high-dimensional space to a lower dimensional space without losing critical information is not a trivial task. In the statistical literature, there are several techniques for learning grouping structures and reducing each multivariate measurement to a univariate index. Based on the work of [8], in this paper we will utilize the Mahalanobis squared distance (MSD) ([16]) of each point from the centroid of the data distribution,

$$MSD = (x_i - \bar{x})S^{-1}(x_i - \bar{x}), i = 1, 2, ..n, \quad (11)$$

where S^{-1} is the inverse of the sample covariance matrix. The MSD can identify hyperellipsoidal clusters because it takes into account the covariance structure of the sample data. We can also use the notion of data depth discussed in [24]. We define and compute what is called the Mahalanobis distance data depth (MDD), given by

$$MDD_G = \frac{1}{1 + \sqrt{(x_i - \bar{x})S^{-1}(x_i - \bar{x})}}, i = 1, 2, ..n. \quad (12)$$

MDD_G measures how “deep” a point is with respect to a given distribution G ; it induces a center-outward ordering of the sample points, if depth values for all points are computed and compared.

Using the arithmetic mean and the usual sample covariance matrix for these computations can present critical issues concerning outliers as masking or swamping the true data structure - a small cluster of outliers might attract the sample mean and inflate the covariance [19]. To guard against the presence of outliers, we can apply the Jackknife Mahalanobis distance data depth introduced by [8] is defined by

$$JMDD_G = \frac{1}{1 + \sqrt{(x_i - \bar{x})S_{(-i)}^{-1}(x_i - \bar{x})}}, \quad (13)$$

where $S_{(-i)}^{-1}$ is the inverse of the sample covariance matrix computed from the data set deleting the i th observation x_i . This is given by

$$S_{(-i)}^{-1} = (n - 2)\left[\frac{1}{n - 1}S^{-1} + \frac{\gamma}{(n - 1)}S^{-1}(x_i - \bar{x})(x_i - \bar{x})^T S^{-1}\right], \quad (14)$$

and where

$$\gamma = \frac{n}{n-1} \left[1 - \left(\frac{n}{(n-1)^2} \right)^2 (x_i - \bar{x}) S^{-1} (x_i - \bar{x}) \right]. \quad (15)$$

In his paper, [31] showed that the direct application of the Mahalanobis distance does not consistently identify hyperellipsoidal clusters. To combat this problem, they introduced a regularization term which takes into account the minimization of each cluster variance in each direction. Another regularization form for the Mahalanobis distance is introduced by [25], but the parameter choice of the regularized covariance matrix is still an open problem. To overcome this, in this paper we propose the Complexity Smoothed Mahalanobis distance (CSMD), a hybridized version of the regularized Mahalanobis distance given by:

$$CSMD_G = C_{1F}(S^*) ((x - \bar{x})(S^*)^{-1}(x - \bar{x})), \quad (16)$$

where $C_{1F}(S^*)$ is the entropic complexity of the regularized covariance matrix (CSE) based on the Frobenius norm. Expressed in terms of eigenvalues λ_j of the covariance matrix S^* it is given by

$$C_{1F}(S^*) = \frac{1}{4\lambda} \sum_{j=1}^p (\lambda_j - \bar{\lambda})^2. \quad (17)$$

CSMD distance overcomes the singularity problems caused by multicollinearity structures present in the data and at the same time it takes into account the complexity of the fitted models.

4. Information Complexity Index - ICOMP

The choice of the best mapping function is not as simple and automatic as one may think. A useful method for selecting the appropriate kernel function is currently a major research issue in the literature. There are many kernel functions to choose from. All kernel based methods such as KDA and SVM classification and their performance depend upon the choice of the kernel function. It has been shown that there is no unique kernel which is a panacea and performs best for all problems. Here, we introduce and propose the use of an information complexity index developed by [8, 9] as our model selection criteria. The general formulation of the information complexity index is defined by

$$ICOMP = \text{Lack of fit} + \text{Complexity term} \quad (18)$$

$$= -2 \log L(\hat{\theta}) + 2C(\hat{\Sigma}_{model}), \quad (19)$$

where $L(\hat{\theta})$ is the maximized likelihood function of the model (or the lack-of-fit) and $C(\hat{\Sigma}_{model})$ is the complexity of the covariance matrix of the estimated parameters of the model. The concept of complexity involves notions such as connectivity patterns of how the models are related and how the components of models are dependent upon each other. In other words, here the complexity is measured by the dependency structure of the model. Therefore, the measure of “overall” model complexity allows us to predict model behavior and choose the best fit, for a given finite sample, among a class of competing models with varying parameter cardinality.

The derivation of ICOMP for the supervised classification problem could be reformulated in terms of a multivariate analysis of variance (MANOVA) model. Let

$$y_{gi} = \mu_g + \varepsilon_g, i = 1, 2, \dots, n_g; g = 1, 2, \dots, k, \quad (20)$$

where y_{gi} is a $(p \times 1)$ response pattern in the g th group for the i th observation, μ_g are the mean vectors, and $\varepsilon_g \sim \mathcal{N}_p(0, \Sigma_g)$ are (i.i.d.) random error vectors. Under this assumption we can easily derive ICOMP given by

$$ICOMP(IFIM) = -2 \log L(\hat{\theta}) + 2C_1(\hat{\mathcal{F}}^{-1}(\hat{\theta})) \quad (21)$$

$$= np \log 2\pi + n \log |\hat{S}_W| + np + 2C_1(\hat{\mathcal{F}}^{-1}), \quad (22)$$

where \hat{S}_W is the estimated within-group covariance matrix, C_1 denotes the maximal information-theoretic measure of complexity, and $\hat{\mathcal{F}}^{-1}$ is the estimated inverse Fisher information matrix (IFIM) of the model. C_1 of the MANOVA model in open form is given by

$$C_1(\hat{\mathcal{F}}^{-1}) = \frac{s}{2} \log \left[\frac{\text{tr}(\hat{\mathcal{F}}^{-1})}{2} \right] - \frac{1}{2} \log |\hat{\mathcal{F}}^{-1}| \quad (23)$$

$$= \frac{kp + kp(p+1)/2}{2} \times \log \left[\frac{\sum_{g=1}^k \left\{ \frac{n}{n_g} \text{tr}(\hat{S}_W) + \frac{1}{2} \text{tr}(\hat{S}_W)^2 + \frac{1}{2} (\text{tr}(\hat{S}_W))^2 + \sum_{i=1}^{n_g} \sigma_{gii}^2 \right\}}{kp + kp(p+1)/2} \right] - \frac{p}{2} \log \left(\prod_{g=1}^k \frac{n}{n_g} \right) - \frac{(k+p+1)}{2} \log |\hat{S}_W| - \frac{p(p-1)}{4} \log(2). \quad (24)$$

In the supervised classification problem of discriminant analysis, the goal is to find a rule which allocates the most observations into the correct group. Intuitively, then, the penalty term for a poorly fitting model would be based on the *classification error rate*. In the usual case of regression, the error variance, σ^2 is estimated by the *mean squared (MS) difference between actual response (group labels) values and predicted response (group labels) values*. For *KDA*, we have error variance given by

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2, \quad (25)$$

where y_i are the original labels and \hat{y}_i are the predicted labels. Thus, we express ICOMP in terms of model performance measure and relate it to the misclassification error rate as follows

$$ICOMP_{Pf} = n \log(2\pi) + n \log(\sigma^2) + n + 2C_{1F}(\hat{S}_W). \quad (26)$$

We also have the Frobenius norm version of the complexity, which was already defined previously. Finally, we can also express ICOMP in terms of the Alignment measure defined by [14]. This metric cannot be employed as a model selection criteria because it does not have a penalty term in its original expression and as such it cannot distinguish among different models with the same number of parameters. By adding the complexity penalty term to the Alignment measure, we define a new $ICOMP_A$ as

$$ICOMP_A = -n \log(1 - A^2) + 2C_{1F}(\hat{S}_W). \quad (27)$$

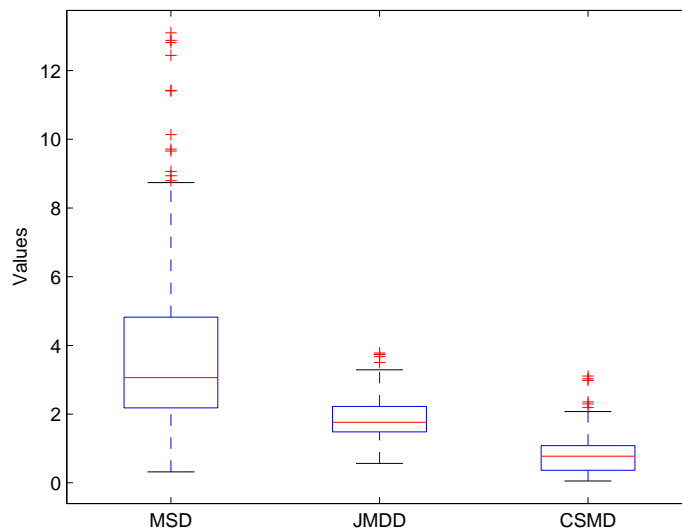


Fig. 1: Distances boxplots Iris data.

In the literature, cross-validation (CV) based criteria have been extensively used in KDA. CV criteria are too time consuming to execute due to the high dimensionality of the feature space and furthermore they are computationally very expensive. As was shown in [10], ICOMP is a clever model selection criterion to choose the best kernel function among a portfolio of kernel functions and it provides a just-in-time decision procedure by shortening the model selection time.

5. Numerical Examples and Results

Here we present numerical examples that demonstrate the application of information complexity criteria for finding the best kernel mapping in a multivariate dataset. We also illustrate the employment of stabilization and regularization techniques that give us stable and robust discriminant solutions. As a first toy example we consider the well-known Fisher Iris data. This dataset presents 3 different species of Irises: Setosa, Versicolor, Virginica. The sample is composed of $n = 150$, with 50 observations for each species on 4 variables.

In order to reduce the multivariate data to a univariate problem, we use the entire sample. We apply the dimension reduction technique we discussed in (Section 3). Using boxplots, the results from our analyses are shown in Figure 1. We notice that the different scales of the distances are quite evident between the the Mahalanobis squared distance (MSD), Jackknife Mahalanobis distance data depth (JMDD), and Complexity Smoothed Mahalanobis distance (CSMD) in the plot. The distribution of JMDD and CSMD certainly has shrunk towards their medians which lay in the interval $[0, 2]$. On the other hand, MSD presents more dispersion in the data and it has a much larger variance than the other two distance measures. Next, we study the shape of these distributions via the χ^2 quantile plots shown in Figure 2. By looking at the plots, we see that all three measures follow a χ_{p-2}^2 ([13]) distribution. In the histogram plots, while JMDD and CSMD exhibit a bimodal

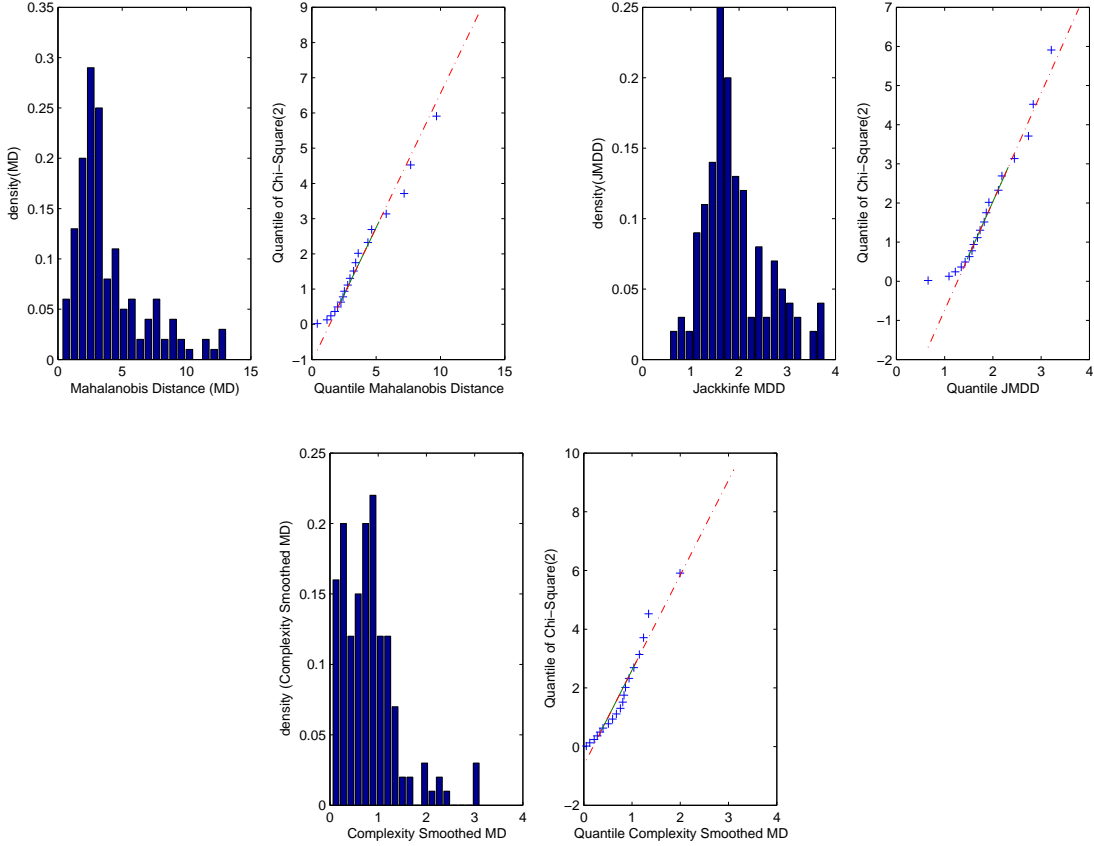


Fig. 2: Histogram graphs and Q-Q plot for different Distances.

structure which indicates different groups (or populations), the histogram of MSD shows much less evidence of a bimodal group structure. In order to discover the best parameter for each kernel mapping, we employed a grid search in the interquartile range $[Q_3 - Q_1]$ of each distance distribution. Thus we obtained variance estimates by leaving out extreme observations. After mapping the data from the input space onto the feature space using the kernel trick, we regularized the within-group covariance matrix S_W^Φ by using the hybrid stabilization and regularization technique presented in Section 2. The allocations of the observations $\Phi(x_i)$ to one of the 3 iris groups is carried out by computing the minimum Mahalanobis distance between the observations and the centroid of each of the groups.

We considered several kernel functions given in Table 1. For the Power Exponential (PE) kernel we decided to fix the degree of the mapping at $d = 2$, and we only compute c , the scale parameter.

Table 1: Kernel Functions.

Kernel Mapping	$k(\mathbf{x}, \mathbf{z})$
Cauchy	$\frac{1}{1 + \frac{c}{\ \mathbf{x} - \mathbf{z}\ ^2}}, c \in \mathbb{R}$
Gaussian (RBF)	$\exp\left(\frac{-\ \mathbf{x} - \mathbf{z}\ ^2}{2c^2}\right), c \in \mathbb{R}$
Exp. Gaussian (ERBF)	$\exp\left(\frac{-\ \mathbf{x} - \mathbf{z}\ }{2c^2}\right), c \in \mathbb{R}$
Laplace (LAP)	$\exp\left(-\sqrt{\frac{\ \mathbf{x} - \mathbf{z}\ ^2}{c^2}}\right), c \in \mathbb{R}$
Multi-quadric (MULTQ)	$\sqrt{\ \mathbf{x} - \mathbf{z}\ ^2 + c^2}, c \in \mathbb{R}_+$
Power Exponential (PE)	$\exp\left(\left(\frac{-\ \mathbf{x} - \mathbf{z}\ ^2}{c^2}\right)^d\right), c \in \mathbb{R}, d \in \mathbb{N}$
Sigmoidal (SIG)	$\tanh[c(\mathbf{x}_i \cdot \mathbf{z}_j) + 1], c \in \mathbb{R}_+$
Linear (LINE)	$(\mathbf{x} \cdot \mathbf{z}),$

The steps of the process of our analysis are as follows:

1. Get data and compute MSD, JMDD, and CSMD on the input variables (raw matrix)
2. Obtain the interquartile range $[Q_3 - Q_1]$ for each distance distribution
3. Standardize the input variables and compute the kernel mapping of the data using parameters from step 2
4. Stabilize and regularize the kernel within-group covariance matrix S_W^Φ
5. Obtain the KDA solution
6. Apply the Euclidean distance in the feature space to allocate observations to groups
7. Compute the error rate and ICOMP_{Pf}
8. Repeat steps 1 to 7 for different kernel mappings.

Table 2 displays parameter values, ICOMP_{Pf} , and misclassification rates for each distance metric. For each kernel we show only the optimal parameter values in terms of the minimum value of ICOMP_{Pf} . As can be seen, the kernel mapping gave a mild improvement in terms of misclassification rate when we used the squared Mahalanobis distance to tune the kernel parameter. However, ICOMP_{Pf} scores were the largest (among all those shown in this table) for all the kernel functions except sigmoid. This is because the MSD conceals the group structure as is shown in Figure 2. The widths we selected are too large for separating data suitably in the Feature Space.

The shrinkage we applied in the CSMD tuned the kernel parameter to a good set of values, as evidenced by the fact that ICOMP_{Pf} decreases considerably, reaching its minimum values across all the kernel mappings. The best kernel function for the Iris data according to ICOMP_{Pf} score is ERBF with $c = 0.3781$ which gives an error rate of 0.0067. This translates to a gain of over 66% in misclassification error rate over the Fisher Linear Discriminant Analysis (FLDA)¹.

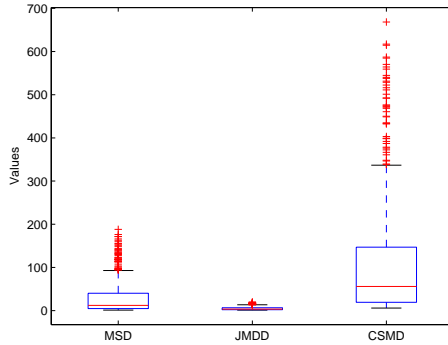
To further demonstrate the performance of our approach as a second numerical example, we consider the benchmark Ionosphere data². This data set contains $n = 351$ instances of $p = 34$ continuous predictors and a binary target variable which classifies the radar signal into two groups (Good/Bad). On the entire data set we compute MSD, JMDD, and CSMD

¹ The misclassification rate in the Iris case using FLDA without employing kernel machines is 0.02, this rate increases up to 0.03 using the linear kernel. For these results, we regularized the covariance matrix which can weight on solutions found.

² The ionosphere dataset is available from the UCI repository.

Table 2: KDA results for Iris data.

MSD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	3.2601	2.1838	2.1838	2.6681	2.1838	2.1838	3.2601	1
ICOMP _{Pf}	346.7591	699.5103	528.3516	332.2371	864.6732	562.7542	998.2209	2186.3
Error	0.0133	0.02	0.0133	0.0133	0.0267	0.0267	3.2601	0.0333
JMDD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	1.4989	1.4838	1.4838	1.7711	1.7711	1.5443	1.5594	1
ICOMP _{Pf}	300.8676	463.5590	432.8070	232.1917	851.1599	356.6593	977.9573	2109.6
Error	0.0267	0.0200	0.0133	0.0133	0.0267	0.0200	0.0267	0.02
CSMD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	0.3634	0.3928	0.3781	0.3634	0.3781	0.8191	1.0691	1
ICOMP _{Pf}	-12.3160	69.7644	-294.1176	-259.5427	710.3241	130.7500	1079.1	2109.6
Error	0.0133	0.0200	0.0067	0.0067	0.0200	0.0200	0.0267	0.02

**Fig. 3:** Distances boxplots Ionosphere data.

in order to reduce the dimension of the problem. As can be clearly seen in Figure 3, this time the CSMD does not produce the shrinkage we obtained in the previous example. For the Ionosphere dataset JMDD gives us better performance. This is due to the fact that there is multicollinearity in this data which impacts the $C_{1F}(S^*)$ term. In order to get results comparable with the work of [14], we partitioned the data into training (80%) and testing sets (20%) by employing a simple random sampling. Three different indices for model selection were computed: ICOMP_{Pf}³, the Alignment between the kernel matrix of the training sample and its corresponding labels vector [14], and ICOMP_A. The process was repeated for 100 times in order to compute the empirical confidence interval of the error rates. Table 3 shows the results obtained from KDA. As before, we only show results for c values corresponding to the optimum score for each kernel mapping across all the three distances. The criteria values and misclassification rates are the arithmetic means of the 100 replications. Visual inspection of the table reveals that the two measures based on information complexity give us the same best three kernels. Using these criteria, the best kernel mapping is Cauchy with $c = 2.1634$. Although we obtained the same result by

³ ICOMP_{Pf} is computed for the testing sample; thus σ refers to the misclassification rate of the test data.

Table 3: KDA results for Ionosphere data.

SMD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	4.6314	19.3737	32.8874	4.6314	18.1452	4.6314	18.1452	1
ICOMP _{Pf}	3596.1	3935.8	3929.6	3875.4	3916.6	9741.5	4153.8	6362.1
ICOMP _A	3968.9	4246.6	4310.0	4277.1	4225.0	9462.2	4340.7	6568.4
AL	0.2574	0.0840	0.0803	0.1819	0.0759	0.0801	0.1041	0.2133
Test Error	0.0717	0.0836	0.0667	0.0636	0.0841	0.6414	0.1281	0.1233
JMDD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	2.1634	2.1634	6.7363	2.1634	6.2632	2.1634	6.4209	1
ICOMP _{Pf}	2929.2	3890.5	3938.1	3144.3	3970.7	16973	4184.5	6361.1
ICOMP _A	3300.3	4249.3	4322.5	3520.2	4309.9	16695	4374.4	6565.6
AL	0.2877	0.2859	0.0855	0.2684	0.0538	0.0805	0.1047	0.2137
Test Error	0.0736	0.0767	0.0660	0.0687	0.0761	0.6400	0.1267	0.1240
CSMD	CAUCHY	RBF	ERBF	LAP	MULTQ	PE	SIG	LINEAR
c Value	146.849	19.326	89.6837	146.8492	19.326	19.326	54.505	1
ICOMP _{Pf}	4040.2	3936.9	3922.7	3935.8	3919.3	16973	4152.0	6.3621
ICOMP _A	4457.5	4247.4	4308.1	4318.4	4223.9	16695	4341.1	6568.4
AL	0.0986	0.0844	0.0805	0.0836	0.0767	0.0805	0.1042	0.2133
Test Error	0.0599	0.0836	0.0661	0.0664	0.0850	0.6400	0.1270	0.1233

Table 4: Confidence Intervals for misclassification errors from the top 5 kernels selected with ICOMP_{Pf}.

Kernel	Distance	Misc. Error Train	Misc. Error Test	ICOMP _{Pf}	ICOMP _A
Cauchy, 2.1634	JMMD	0.0375 - 0.0441	0.0681 - 0.0790	2929.2	3300.3
Laplace, 2.1634	JMMD	0.0308 - 0.0361	0.0632 - 0.0742	3144.3	3520.5
Cauchy, 4.6314	MSD	0.0454 - 0.0487	0.0653 - 0.0753	3596.1	3961.4
Laplace, 4.6314	MSD	0.0412 - 0.0442	0.0594 - 0.0689	3875.4	4274.2
RBF, 2.1634	JMMD	0.0569 - 0.0598	0.0713 - 0.0822	3890.5	4249.3

applying Alignment, which reaches its maximum value for the same kernel and parameter, we have to highlight the fact that our index is more general and robust. In fact, ICOMP_{Pf} cannot only be obtained without rigid codification for the labels but it can even be easily computed for the multi-class tasks as we showed in the first numerical example above on the Iris data. Moreover, the classification improvement gained with respect to cited work is notable: with our approach, the mean test error rate is 0.0736 - we have decreased the published error rate in the literature by more than half. In Figure 4, we see that the solution found is stable, as the confidence intervals of the error rates are close to their means.

In Table 4, we show the confidence intervals of the top 5 kernels selected by ICOMP_{Pf}, across all the results collected. We notice that no matter which kernel we pick, we obtain classification rates much superior to published SVM error rates.

6. Conclusions and Discussion

The problem of choosing the best kernel function and tuning the optimal kernel parameters are major critical issues in the machine learning literature that have daunted researchers for some time. Being able to identify the mapping into the feature space that optimizes

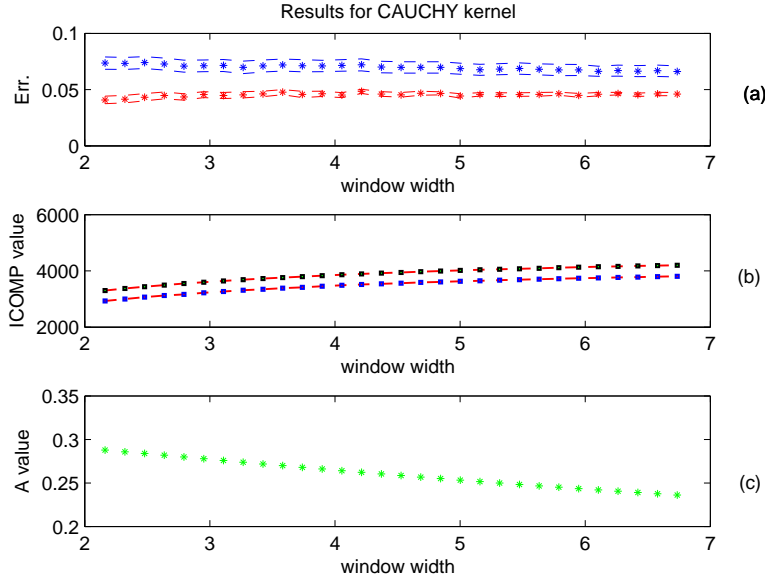


Fig. 4: Plots for Ionosphere data: (a) test and train misclassification error rates with confidence intervals, (b) ICOMP_{Pf} and ICOMP_A values, (c) Alignment trend for Cauchy kernel.

the separation, without any subjective choice, remains the key research issue in directing kernel-based modeling to achieve effective solutions. In the Alignment metric [14, 21], we recognize a first step for defining an index to overcome heuristic model selection methods (as cross-validation). At the same time, we do believe that the agreement between the embedding and learning task is not enough to identify the “true” model. Alignment presents neither a measure of the multicollinearity among variables nor overall complexity, which is a fundamental part of the penalty term used in well-known information-theoretic model selection criteria such as AIC [1], CAIC [5], ICOMP [6], and others. Additionally, this index is currently defined only for binary classification; it cannot be used to solve the multi-class problem.

In this paper, we have proposed the use of information complexity criterion as a fitness function for carrying out kernel selection and choosing the optimal kernel parameters by taking into account performance and parsimony of all models evaluated in one criterion function. As we have shown, ICOMP_{Pf} has several theoretical properties which make it a desirable index for quantifying the fit among competing discriminant functions using different kernel mappings. We showed that our approach can be used in binary classification problems, and it can be easily extended also to multi-class classification problems. We can estimate models using training and testing sets to guard against overfitting. Therefore, ICOMP_{Pf} criterion is particularly flexible and easily derived for other learning tasks. We have set forth an algorithm for performing KDA using an intelligent method for tuning the kernel parameter and a robust estimate of the within-group covariance matrix in order to obtain stabilization in DA. We acknowledge that the usage of the JMMD is computationally time consuming - it requires time proportional to sample size. However, as we have shown, the Complexity Smoothed Mahalanobis distance (CSMD) is also able to select parameter

values which lead to good separation in the feature space. Moreover, we also gave the definition of the Mahalanobis squared distance in Reproducing Kernel Hilbert Space (RKHS). This distance is superior to the Euclidian distance that is used in the Alignment method, since it takes account the sample correlations.

The limitations of a nonprobabilistic classification obtained with the SVM approach has been overcome within the KDA framework. With KDA, it is possible to specify different prior probabilities and obtain probabilistic group membership results for each observation.

A valuable extension of this work would be to derive a method for identifying which variables in the input space are most responsible for the separation in the feature space. To this end, we plan to introduce the genetic algorithm to carry out the best subset selection of the original variable which are doing the separation in the feature space using ICOMP within probabilistic KDA. The results of this work will be published elsewhere.

Acknowledgments

The first two authors gratefully acknowledge the continued support given by Professor Bozdogan throughout this study. For the first author this study was funded by the Marco Polo Visiting Scholarship funds from Italy for six months while visiting the Department of Statistics, Operations, and Management Science at the University of Tennessee in Knoxville, Tennessee as a Post-Doctoral Researcher. The first author extends her great appreciation for the warm hospitality shown by Prof. Bozdogan. It was a pleasure to work under his guidance and supervision in a conducive atmosphere. Finally, the authors also wish to express their appreciation to the anonymous referees, who provided constructive comments. Their comments helped to improve this research paper.

References

- [1] Akaike, H. "Information theory and an extension of the maximum likelihood principle." In Petrov, B. N. and (Eds.), B. F. C. (eds.), *Second International Symposium on Information Theory* (1973).
- [2] Bach, F. R., Lanckriet, G. R. G., and Jordan, M. I. "Fast kernel learning using sequential minimal optimization." Technical report, Division of Computer Science, University of California, Berkeley (2004).
- [3] Baudat, G. and Anouar, F. "Generalized discriminant analysis using kernel approach." *Neural Computation* (2000).
- [4] Bennett, K. P., Momma, M., and Embrechts, M. J. "MARK: A Boosting algorithm for heterogeneous kernel models." (2002).
- [5] Bozdogan, H. "Model selection and Akaike's information criterion (AIC): the General theory and its analytical extensions." *PSYCHOMETRIKA* (1987).
- [6] "On the Information-based measure of covariance complexity and its application to the evaluation of multivariate linear model." *Communications in Statistics - Theory and Methods* (1990).
- [7] "Akaike's information criterion and recent developments in informational complexity." *Journal of Mathematical Psychology* (2000).
- [8] "Exploring multivariate modality by unsupervised mixture of cubic B-splines in 1-D using model selection criteria." In *Data Analysis: Scientific Model and Pratical Application*, 105–119. Springer (2000).
- [9] "Intelligent statistical data mining with information complexity and genetic algorithms." In Bozdogan, H. (ed.), *Statistical Data Mining and Knowledge Discovery*, 15–56. Florida: Chapman and Hall/CRC, Boca Raton (2004).

- [10] Bozdogan, H., Camillo, F., and Liberati, C. “On the choice of the kernel function in kernel discriminant analysis using information complexity.” In *Data Analysis, Classification and the Forward Search*, 11–21. Springer Berlin Heidelberg (2006).
- [11] Cawley, G. and Talbot, N. L. C. “Efficient leave-one-out cross-validation of kernel Fisher discriminant classifiers.” *Pattern Recognition*, 36(11):2585–2592 (2003).
- [12] Chen, M. C.-F. “Estimation of covariance matrices under a quadratic loss function.” Research report s-46, Department of Mathematics SUNY at Albany, Albany, N.Y. (1976).
- [13] Cox, D. “Notes on some aspects of regression analysis.” *Journal of Royal Statistical Society A* (1968).
- [14] Cristianini, N., Kandola, J., Shawe-Taylor, J., and Elisseeff, A. “On kernel-target alignment.” *Journal of Machine Learning Research* (2002).
- [15] Cristianini, N. and Shawe-Taylor, J. *An introduction to Support Vector Machines*. Cambridge University Press (2000).
- [16] Fisher, R. A. “The use of multiple measurements in taxonomic problems.” *Annals of Eugenics* (1936).
- [17] Friedman, J. H. “Regularized discriminant analysis.” *Journal of the American Statistical Association*, 84(405):165–175 (1989).
- [18] Fung, G., Dundar, M., Bi, J., and Rao, R. “A fast iterative algorithm for fisher discriminant using heterogeneous kernels.” In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, 40. New York, NY, USA: ACM (2004).
- [19] Hadi, A. S. “Identifying multiple outliers in multivariate data.” *Journal of the Royal Statistical Society. Series B*, 54(3) (1992).
- [20] Hamers, B., Suykens, J., Leemans, V., and De Moor, B. “Ensemble learning of coupled parameterized kernel models.” (2003).
- [21] Kandola, J., Shawe-Taylor, J., and Cristianini, N. “On the extensions of kernel alignment.” (2002). <http://eprints.ecs.soton.ac.uk/9745/>
- [22] Kim, S.-J., Magnani, A., and Boyd, S. “Optimal kernel selection in kernel Fisher discriminant analysis.” In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, 465–472. New York, NY, USA: ACM (2006).
- [23] Lanckriet, G. R. G., Cristianini, N., Bartlett, P., El Ghaoui, L., and Jordan, M. I. “Learning the kernel matrix with semidefinite programming.” *Journal of Machine Learning Research*, 5:27–72 (2004).
- [24] Liu, R. Y. “Control chart and multivariate processes.” *Journal of the American Statistical Association*, 90(432) (1995).
- [25] Mao, J. and Jain, A. “A self-organizing network for hyperellipsoidal clustering (HEC).” In *Intelligent Processing Systems, 1997. ICIPS '97*. IEEE International (1996).
- [26] Mercer, J. “Functions of positive and negative type and their connection with the theory of integral equations.” *Philosophical Transactions Royal Society London* (1909).
- [27] Mika, S., Smola, A., and Schölkopf, B. “An improved training algorithm for kernel Fisher discriminants.” In Jaakkola, I. T. and Richardson, T. (eds.), *Proceedings AISTATS 2001*. San Francisco, CA (2001).
- [28] Press, S. *Estimation of a normal covariance matrix*. Santa Monica Rand Corporation (1975).
- [29] Shurygin, A. M. “The linear combination of the simplest discriminator and Fishers one.” *Applied Statistics* (1983).
- [30] Thomaz, C., Boardman, J., Hill, D., Hajnal, J., Edwards, D., Rutherford, M., Gillies, D., and Rueckert, D. “Using a maximum uncertainty LDA-based approach to classify and analyse MR brain images.” In *Medical Image Computing and Computer-Assisted Intervention MICCAI 2004*, 291–300. Springer Berlin / Heidelberg (2004).
- [31] Wang, S., Feng, M., Wei, S., and Shaowei, X. “The hyperellipsoidal clustering using genetic algorithm.” In *Intelligent Processing Systems, 1997. ICIPS '97*. IEEE International (1997).
- [32] Ye, J., Xiong, T., Li, Q., Janardan, R., Bi, J., Cherkassky, V., and Kambhamettu, C. “Efficient model selection for regularized linear discriminant analysis.” In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, 532–539. New York, NY, USA: ACM (2006).