To appear in the *Journal of Statistical Computation and Simulation*
Vol. 00, No. 00, Month 20XX, 1–17

The mRMR variable selection method: a comparative study for functional data

J.R. Berrendero, A. Cuevas and J.L. Torrecilla [*]

Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain

(*April 2015*)

The use of variable selection methods is particularly appealing in statistical problems with functional data. The obvious general criterion for variable selection is to choose the 'most representative' or 'most relevant' variables. However, it is also clear that a purely relevance-oriented criterion could lead to select many redundant variables. The mRMR (minimum Redundance Maximum Relevance) procedure, proposed by Ding and Peng (2005) and Peng *et al.* (2005) is an algorithm to systematically perform variable selection, achieving a reasonable trade-off between relevance and redundancy. In its original form, this procedure is based on the use of the so-called *mutual information criterion* to assess relevance and redundancy. Keeping the focus on functional data problems, we propose here a modified version of the mRMR method, obtained by replacing the mutual information by the new association measure (called *distance correlation*) suggested by Székely *et al.* (2007). We have also performed an extensive simulation study, including 1600 functional experiments (100 functional models × 4 sample sizes × 4 classifiers) and three real-data examples aimed at comparing the different versions of the mRMR methodology. The results are quite conclusive in favor of the new proposed alternative.

Keywords: functional data analysis ; supervised classification ; distance correlation ; variable selection

AMS Subject Classification: Primary: 62H30; Secondary: 62H20

1. Introduction

The use of high-dimensional or functional data entails some important practical issues. Besides the problems associated with computation time and storage costs, high-dimensionality introduces noise and redundancy. Thus, there is a strong case for using different techniques of dimensionality reduction.

We will consider here dimensionality reduction via variable selection techniques. The general aim of these techniques is to replace the original high-dimensional (perhaps functional) data by lower dimensional projections obtained by just selecting a small subset of the original variables in each observation. In the case of functional data, this amounts to replace each observation $\{x(t),\ t \in [0,1]\}$ with a low-dimensional vector $(x(t_1), \dots, x(t_k))$. Then, the chosen statistical methodology (supervised classification, clustering, regression,...) is performed with the 'reduced', low-dimensional data. Usually the values $t_1, \dots, t_k$ identifying the selected variables are the same for all considered data. A first advantage of variable selection (when compared with other dimension reduction

[*]Corresponding author. Email: joseluis.torrecilla@uam.es

methods, as Partial Least Squares) is the ease of interpretability, since the dimension reduction is made in terms of the original variables. In a way, variable selection appears as the most natural dimension reduction procedure in order to keep in touch, as much as possible, with the original data: see for instance [1, 2] among many other examples in experimental sciences or engineering. In [1] the authors note that 50 genes (among almost 7000) are enough for cancer subtype classification. Likewise, Lindquist and McKeague [2] point out that in some functional data regression (or classification) problems, as functional magnetic resonance imaging or gene expression, 'the influence is concentrated at sensitive time points'.

We refer to [3] for an account of different variable selection methods in the multivariate (non-functional) case. A partial comparative study, together with some new proposals for the functional framework, can be found in [4].

Throughout this work we will consider variable selection in the setting of functional supervised classification (the extension to more general regression problems is also possible with some obvious changes). Thus, the available sample information is a data set of type $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ of n independent observations drawn from a random pair (X, Y). Here Y denotes a binary random variable, with values in $\{0, 1\}$, indicating the membership to one of the populations P_0 or P_1 and X_i are iid trajectories (in the space $\mathcal{C}[0, 1]$ of real continuous functions on $[0, 1]$), drawn from a stochastic process $X = X(t)$. The supervised classification problem aims at predicting the membership class Y of a new observation for which only the variable X is known. Any function $g_n(x) = g_n(x; \mathcal{D}_n)$ with values in $\{0, 1\}$ is called a classifier.

Several functional classifiers have been considered in the literature; see, e.g., [5] for a survey. Among them maybe the simplest one is the so-called k-nearest neighbours (k-NN) rule, according to which an observation x is assigned to P_1 if and only if the majority among their k nearest sample observations X_i in the training sample fulfil $Y_i = 1$. Here $k = k_n \in \mathbb{N}$ is a sequence of smoothing parameters which must satisfy $k_n \to \infty$ and $k_n/n \to 0$ in order to achieve consistency. In general, k-NN could be considered (from the limited experience so far available; see e.g., [6]) a sort of benchmark, reference method for functional supervised classification. Simplicity, ease of motivation and general good performance (it typically does not lead to gross classification errors) are perhaps the most attractive features of this method. Besides k-NN, we have also considered (inspired in the paper by Ding and Peng [7] where a similar study is carried out) three additional classifiers: the popular *Fisher's linear classifier* (LDA) used often in classical discriminant analysis, the so-called *Naïve Bayes method* (NB) and the (linear) *Support Vector Machine classifier* (SVM). Note that, in our empirical studies, all the mentioned classifiers (k-NN, LDA, NB and SVM) are used *after the variable selection step*, on the 'reduced data' resulting from the variable selection process.

In fact, as we will point out below, the main goal of our study is not to compare different classifiers. We are rather concerned with the comparison of different methods for variable selection (often referred to as *feature selection*). A relevant procedure for variable selection, especially popular in the machine learning community, is the so-called *minimum Redundancy Maximum Relevance* (mRMR) method. It was proposed by Ding and Peng [7] and Peng *et al.* [8] as a tool to select the most discriminant subset of variables in the context of some relevant bioinformatics problems. See also [9–11] for closely related ideas.

The purpose of this paper. Overall, we believe the mRMR procedure is a very natural way to tackle the variable selection problem if one wants to make completely explicit the trade-off relevance/redundancy. The method relies on the use of an association measure to assess the relevance and redundancy of the considered variables. In the original papers the so-called 'mutual information' measure was used for this purpose. The aim of the present paper is to propose other alternatives for the association measure, still keeping the main idea behind the mRMR procedure. In fact, most mRMR researchers admit that there is considerable room for improvement. We quote from the discussion in [8]: *'The mRMR paradigm can be better viewed as a general framework to effectively select features and allow all possibilities for more sophisticated or more powerful implementation schemes'*. In this vein, we consider several versions of the mRMR and compare them by an extensive empirical study. Two of these versions are new: they are based on the 'distance covariance' and 'distance correlation' association measures proposed by Székely *et al.* [12]. Our results suggest (and this is the main conclusion of our study) that the new version based on the distance correlation measure represents a clear improvement of the mRMR methodology.

The rest of the paper is organized as follows. Section 2 contains a brief summary and some remarks about the mRMR algorithm. The different association measures under study (which are used to define the different versions of the mRMR method) are explained in Section 3, with especial attention to the *correlation of distances*. [12, 13] The empirical study, consisting of 1600 simulation experiments and some representative real data sets, is explained in Section 4. Finally, some conclusions are given.

2. The trade-off relevance/redundancy. The mRMR criterion

When faced with the problem of variable selection methods in high-dimensional (or functional) data sets, a natural idea arises at once: obviously, one should select the variables according to their relevance (representativeness). However, at the same time, one should avoid the redundancy which appears when two highly relevant variables are closely associated to each other. In that case, one might expect that both variables essentially carry the same information, so that to choose just one of them should suffice.

The mRMR variable selection method, as proposed in [7, 8], provides a formal implementation of a variable selection procedure which explicitly takes into account this trade-off relevance/redundancy.

In our functional binary classification problem, the description of the mRMR method is as follows: the functional explanatory variable $X(t)$, $t \in [0,1]$ will be used in a discretized version $(X(t_1), \ldots, X(t_N))$. When convenient, the notations X_t and $X(t)$ will be used indistinctly. For any subset S of $\{t_1, \ldots, t_N\}$, the *relevance* and the *redundancy* of S are defined, respectively, by

$$\mathrm{Rel}(S) = \frac{1}{\mathrm{card}(S)} \sum_{t \in S} I(X_t, Y), \tag{1}$$

and

$$\mathrm{Red}(S) = \frac{1}{\mathrm{card}^2(S)} \sum_{s,t \in S} I(X_t, X_s), \tag{2}$$

where $\mathrm{card}(S)$ denotes the cardinality of S and $I(\cdot,\cdot)$ is an 'association measure'. This function I measures how much related are two variables. So, it is natural to think that the relevance of X_t is measured by how much related it is with the response variable Y, that is $I(X_t, Y)$, whereas the redundancy between X_t and X_s is given by $I(X_s, X_t)$.

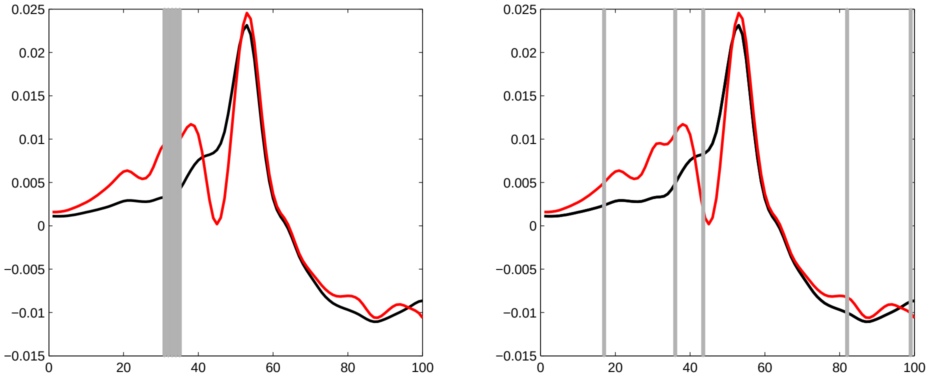

Figure 1. Mean functions for both classes considered in the Tecator data set (first derivative). Left panel shows the five variables selected by Maximum Relevance. Right panel corresponds to the variables selected by mRMR.

Now, in summary, the mRMR algorithm aims at maximizing the relevance avoiding an excess of redundancy.

The choice of the association measure I is a critical aspect in the mRMR methodology. In fact, this is the central point of the present work so that we will consider it in more detail later. By now, in order to explain how the mRMR method works, let us assume that the measure I is given:

(a) The procedure starts by selecting the most relevant variable, given by the value t_i such that the set $S_i = \{t_i\}$ maximizes $\text{Rel}(S)$ among all the singleton sets of type $S_j = \{t_j\}$.
(b) Then, the variables are sequentially incorporated to the set S of previously selected variables, with the criterion of maximizing the difference $\text{Rel}(S) - \text{Red}(S)$ (or alternatively the quotient $\text{Rel}(S)/\text{Red}(S)$).
(c) Finally, different stopping rules can be considered. We set the number of variables through a validation step (additional details can be found in Sections 4 and 5).

In practice, the use of the mRMR methodology is especially important in the functional data problems, where those variables which are very close together are often strongly associated.

The following example shows to what extent the mRMR makes a critical difference in the variable selection procedure. It concerns the well-known *Tecator data set* (a benchmark example very popular in the literature on functional data; see Section 5 for details). To be more specific, we use the first derivative of the curves in the Tecator data set, which is divided into two classes. We first use a simple 'ranking procedure', where the variables are sequentially selected according to their relevance (thus avoiding any notion of redundancy). The result is shown in the left panel of Figure 1 (the selected variables are marked with grey vertical lines). It can be seen that in this case, all the five selected variables provide essentially the same information. On the right panel we see the variables selected from mRMR procedure which are clearly better placed to provide useful information. This visual impression is confirmed by comparing the error percentages obtained from a supervised classification method using only the variables selected by both methods. While the classification error obtained with the mRMR selected variables is 1.86%, the corresponding error obtained with those of the ranking method is 4.09%.

3. Association measures

As indicated in the previous section, the mRMR criterion relies on the use of an association measure $I(X, Y)$ between random variables. The choice of appropriate association measures is a classical issue in mathematical statistics. Many different proposals are available and, in several aspects, this topic is still open for further research, especially in connection with the use of high-dimensional data sets (arising, e.g., in genetic microarray examples,[14, 15]).

A complete review of the main association measures for random variables is clearly beyond the scope of this paper. So, we will limit ourselves to present here the measures $I(X, Y)$ we have used in this work:

(1) *The ordinary correlation coefficient between X and Y (in absolute value).* This is the first obvious choice for the association measure $I(X, Y)$. It clearly presents some drawbacks (it does not characterize independence and it is unsuitable to capture non-linear association) but still, it does a good job in many practical situations.

(2) The *Mutual Information Measure*, $MI(X, Y)$ is defined by

$$MI(X,Y) = \int \log \frac{p(x,y)}{p_1(x)p_2(y)} p(x,y) d\mu(x,y), \tag{3}$$

where X, Y are two random variables with respective μ-densities p_1 and p_2; in the standard, absolutely continuous case, μ would be the product Lebesgue measure. In the discrete case, μ would be a counting measure on a countable support. The joint density of (X, Y) is denoted by $p(x, y)$.

This is the association measure used in the original version of the mRMR procedure.[7, 8].

It is clear that $MI(X, Y)$ measures how far is $p(x, y)$ from the independence situation $p(x, y) = p_1(x)p_2(y)$. It is easily seen that $MI(X, Y) = MI(Y, X)$ and $MI(X, Y) = 0$ if and only if X and Y are independent.

In practice, $MI(X, Y)$ must be approximated by considering, if necessary, 'discretized versions' of X and Y, obtained by grouping their values on intervals represented by suitable label marks, a_i, b_j. This leads to approximate expressions of type

$$\widehat{MI}(X,Y) = \sum_{i,j} \log \frac{\mathbb{P}(X = a_i, Y = b_j)}{\mathbb{P}(X = a_i)\mathbb{P}(X = b_j)} \mathbb{P}(X = a_i, Y = b_j), \tag{4}$$

where, in turn, the probabilities can be empirically estimated by the corresponding relative frequencies. In [7] the authors suggest a threefold discretization pattern, i.e., the range of values of the variable is discretized in three classes. The limits of the discretization intervals are defined by the mean of the corresponding variable $\pm\sigma/2$ (where σ is the standard deviation). We will explore this criterion in our empirical study below.

(3) *The Fisher-Correlation (FC) criterion*: It is a combination of the F-statistic,

$$F(X,Y) = \frac{\sum_k n_k(\bar{X}_k - \bar{X})^2/(K-1)}{\sum_k (n_k - 1)\sigma_k^2/(n-K)}, \tag{5}$$

used in the relevance measure (1), and the ordinary correlation, C, used in the redundancy measure (2). In the expression (5), K denotes the number of classes (so $K = 2$ in our binary classification problem), $\bar{X}$ denotes the mean of X, $\bar{X}_k$ is the mean value of X of the elements belonging the k-th class, for $k = 0, 1$, and n_k and σ_k^2 are the sample size and the variance of the k-th class, respectively.

Ding and Peng [7] suggest that, in principle, this criterion might look more useful than $\widehat{MI}$ when dealing with continuous variables but their empirical results do not support that idea. Such results are confirmed by our study so that, in general terms, we conclude that the mutual information (4) is a better choice even in the continuous setting.

(4) *Distance covariance*: this is an association measure recently proposed by Székely *et al.* [12]. Denote by $\varphi_{X,Y}$, φ_X, φ_Y the characteristic functions of (X, Y), X and Y, respectively. Here X and Y denote multivariate random variables taking values in $\mathbb{R}^p$ and $\mathbb{R}^q$, respectively (note that the assumption $p = q$ is not needed). Let us suppose that the components of X and Y have finite first-order moments. The distance covariance between X and Y is the non-negative value $\mathcal{V}(X, Y)$ defined by

$$\mathcal{V}^2(X,Y) = \int_{\mathbb{R}^{p+q}} \mid \varphi_{X,Y}(u,v) - \varphi_X(u)\varphi_Y(v) \mid^2 w(u,v)dudv, \qquad (6)$$

with $w(u, v) = (c_p c_q |u|_p^{1+p} |v|_q^{1+q})^{-1}$, where $c_d = \frac{\pi^{(1+d)/2}}{\Gamma((1+d)/2)}$ is half the surface area of the unit sphere in $\mathbb{R}^{d+1}$ and $|\cdot|_d$ stands for the Euclidean norm in $\mathbb{R}^d$.

While definition (6) has a rather technical appearance, the resulting association measure has a number of interesting properties. Apart from the fact that (6) allows for the case where X and Y have different dimensions, we have $\mathcal{V}^2(X, Y) = 0$ if and only if X and Y are independent. Moreover, the indicated choice for the weights $w(u, v)$ provides valuable equivariance properties for $\mathcal{V}^2(X, Y)$ and the quantity can be consistently estimated from the mutual pairwise distances $|X_i - X_j|_p$ and $|Y_i - Y_j|_q$ between the sample values X_i and Y_j (no discretization is needed).

We refer to [12, 13, 16, 17] for a detailed study of this increasingly popular association measure. We refer also to [4] for an alternative use (not related to mRMR) of $\mathcal{V}^2(X, Y)$ in variable selection.

(5) *Distance correlation*: this is just a sort of standardized version of the distance covariance. If we denote $\mathcal{V}^2(X) = \mathcal{V}^2(X, X)$, the (square) distance correlation between X and Y is defined by $\mathcal{R}^2(X, Y) = \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}$ if $\mathcal{V}^2(X)\mathcal{V}^2(Y) > 0$, $\mathcal{R}^2(X, Y) = 0$ otherwise.

Of course, other association measures might be considered. However, in order to get an affordable comparative study, we have limited our study to the main association measures previously used in the mRMR literature. We have only added the new measures $\mathcal{V}^2$ and $\mathcal{R}^2$, which we have tested as possible improvements of the method.

Also, alternative versions of the mRMR procedure have been proposed in literature. In particular, the Mutual Information measure could be estimated by kernel density estimation,[18]. Regarding the kernel-based estimation of the MI measure, the crucial issue [19] of the optimal selection of the smoothing parameter has not been, to our knowledge, explicitly addressed; note that here 'optimal' should refer to the estimation of MI. Likewise, other weighting factors might be used instead of just $card(S)$ in equation (2),[20]. However, still the 'original' version of mRMR (with discretization-based MI estimation) seems to be the most popular standard; see [21, 22] for very recent examples.

Let us finally note that all the association measures we are considering take positive values. So, the phenomena associated with the the negative association values analyzed in [23] do not apply in this case.

Notation. The association measures defined above will we denoted in the tables of our empirical study by **C**, **MI**, **FC**, **V** and **R**, respectively.

4. The simulation study

We have checked five different versions of the mRMR variable selection methodology. They have been obtained by using different association measures (as indicated in the previous section) to assess relevance and redundancy.

In all cases, the comparisons have been made in the context of problems of binary supervised classification, using 100 different models to generate the data (X, Y). These models are defined either by

(i) specifying the distributions of $X|Y=0$ and $X|Y=1$; in all cases, we take $p = \mathbb{P}(Y = 0) = 1/2$.
(ii) specifying both the marginal distribution of X and the conditional distribution $\eta(x) = \mathbb{P}(Y = 1|X = x)$.

Our experiments essentially consist of performing variable selection for each model using the different versions of mRMR and evaluating the results in terms of the respective probabilities of correct classification when different classifiers are used on the selected variables. The full list of considered models is available at the *Supplemental material* document. All these models have been chosen in such a way that the optimal (Bayes) classification rule depends on just a finite number of variables. The processes considered include Brownian motion (with different mean functions), Brownian bridge and several other Gaussian models, in particular the Ornstein-Uhlenbeck process. Other mixture models based on them are also considered. All these models are generated according to the pattern (i) above. In addition, we have considered several 'logistic-type' models, generated by using pattern (ii).

For each considered model all the variable selection methods (**C, MI**, etc.) are checked for four sample sizes, $n = 30,\ 50,\ 100,\ 200$ and four classification methods (**k-NN**, **LDA**, **NB** and **SVM**). So, we have in total $100 \times 4 \times 4 = 1600$ simulation experiments.

4.1. *Classification methods*

We have used the four classifiers considered in the paper by Ding and Peng [7], except that we have replaced the logistic regression classifier (which is closely related to the standard linear classifier) with the non-parametric k-NN method. All of them are widely known and details can be found, e.g. in [24].

- **Naïve Bayes classifier (NB).** This method relies on the assumption that the selected variables are Gaussian and conditionally independent in each class. So a new observation is assigned according to its posterior probability calculated from the Bayes rule. Of course the independence assumption will often fail (especially in the case of functional data). However, as shown in [7], this rule works as an heuristics which offers sometimes a surprisingly good practical performance.
- **The k-Nearest Neighbors classifier (k-NN).** According to this method (already

commented in the introduction of the paper) a new observation is assigned to the class of the majority of its k closest neighbors. We use the usual Euclidean distance (or L^2-distance when the method is used with the complete curves) to define the neighbors. The parameter k is fitted through the validation step, as explained below.

- **Linear Discriminant Analysis (LDA).** The classic Fisher's linear discriminant is, still today, the most popular classification method among practitioners. It is know to be optimal under gaussianity and homoscedasticity of the distributions in both populations but, even when these conditions are not fulfilled, LDA tends to show a good practical performance in many real data sets. See, e.g., [25].
- **Support Vector Machine (SVM).** This is one of the most popular classification methodologies in the last two decades. The basic idea is to look for the 'best hyperplane' in order to maximize the separation margin between the two classes. The use of different kernels (to send the observations to higher dimensional spaces where the separation is best achieved) is the most distinctive feature of this procedure. As in [7] we have used linear kernels.

As an objective reference, our simulation outputs include also the percentages of correct classification obtained with those classifiers based on the complete curves, i.e., when no variable selection is done at all (except for LDA whose functional version is not feasible; see [5]). This reference method is called **Base**. A somewhat surprising conclusion of our study is that this **Base** method is often outperformed by the variable selection procedures. This could be due to the fact that the whole curves are globally more affected by noise than the selected variables. Thus, variable selection is beneficial not only in terms of simplicity but also in terms of accuracy.

4.2. *Computational details*

All codes have been implemented in MATLAB and are available from the authors upon request. We have used our own code for k-NN and LDA (which is a faster implementation of the MATLAB function *classify*). The Naïve Bayes classifier is based on the MATLAB functions *NaiveBayes.fit* and *predict*. The linear SVM has been performed with the MATLAB version of the LIBLINEAR library (see [26]) with bias and solver type 2, which obtains (with our data) very similar results to those of the default solver type 1 but faster. The mRMR method has been implemented in such a way that different association measures can be used to define it. An online implementation of the original mRMR method can be found in `http://penglab.janelia.org/proj/mRMR/` .

Following Ding and Peng [7], the criteria (1) and (2) to assess relevance and redundancy, respectively, are in fact replaced by approximate expressions, numbered (6) and (7) in [7]: as these authors point out, their expression (6) is equivalent to the relevance criterion (1) while (7) provides an approximation for the minimum redundancy criterion (2). The empirical estimation of the distance covariance (and distance correlation) implemented is the one proposed in Székely *et al.* [12] expression (2.8).

All the functional simulated data are **discretized** to $(x(t_1),\ldots,x(t_{100}))$, where t_i are equi-spaced points in $[0,1]$. There is a partial exception in the case of the Brownian-like model, where (to avoid the degeneracy $x(t_0)=0$) we take $t_1 = 5/105$. Also (for a similar reason), a truncation is done at the end of the interval $[0,1]$ in those models including the Brownian Bridge.

The number k of nearest neighbours in the k-NN classifier, the cost parameter C of the SVM classifier and the number of selected variables are chosen by standard validation procedures.[3]. To this end, in the simulation study, we have generated independent vali-

Table 1. Performance outputs for the considered methods, using NB and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (NB)	Sample size	MID	FCD	RD	VD	CD	Base
Average accuracy	$n = 30$	78.08	78.42	$\boxed{79.56}$	79.24	$\boxed{79.28}$	77.28
	$n = 50$	79.64	79.34	$\boxed{80.92}$	80.45	$\boxed{80.46}$	78.29
	$n = 100$	80.76	80.06	$\boxed{81.90}$	81.34	$\boxed{81.41}$	78.84
	$n = 200$	81.46	80.44	$\boxed{82.55}$	81.90	$\boxed{82.05}$	79.13
Average dim. red	$n = 30$	8.7	9.3	$\boxed{7.2}$	$\boxed{7.1}$	7.8	100
	$n = 50$	7.9	9.0	$\boxed{6.8}$	$\boxed{6.7}$	7.4	100
	$n = 100$	7.2	8.5	$\boxed{6.3}$	$\boxed{6.2}$	6.8	100
	$n = 200$	6.6	8.1	$\boxed{5.8}$	$\boxed{5.7}$	6.4	100
Victories over Base	$n = 30$	57	61	$\boxed{77}$	$\boxed{71}$	69	-
	$n = 50$	66	61	$\boxed{79}$	$\boxed{74}$	70	-
	$n = 100$	77	61	$\boxed{88}$	81	$\boxed{85}$	-
	$n = 200$	84	62	$\boxed{93}$	85	$\boxed{91}$	-

dation and test samples of size 200. Each simulation output is based on 200 independent runs.

4.3. *A few numerical outputs from the simulations*

We present here just a small sample of the entire simulation outputs, which can be downloaded from `www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx` . Some additional results, including a complete list of the considered models, can be found in the *Supplemental material* file.

Tables 1 - 4 contain the results obtained with NB, k-NN, LDA and SVM respectively. The boxed outputs in these tables correspond to the winner and second best method in each row. The columns headings (MID, FCD, etc.) correspond to the different mRMR methods based on different association measures, as defined in Section 3 (see the respective notations at the end of that section). The added letter 'D' refers to the fact that global criterion to be maximized is just the difference between the measures (1) and (2) of relevance and redundancy, respectively. There are other possibilities to combine (1) and (2). One could take for instance the quotient. The corresponding outputs methods are denoted MIQ, FCQ, etc. in our supplementary material files. However, the outputs are not given here for the sake of brevity. In any case, our results suggest that the difference-based methods are globally (although not uniformly) better than those based on quotients. The column 'Base' gives the results when no variable selection method is used (that is, the entire curves are considered). This column does not appear when the LDA method is used, since LDA cannot directly work on functional data.

The row entries 'Average accuracy' provide the average percentage of correct classification over the 100 considered model outputs; recall that every output is in turn obtained as an average over 200 independent runs. The rows 'Average dim. red.' provide the average numbers of selected variables. The average number of times that every method beats the 'Base' benchmark procedure is given in 'Victories over Base'.

It can be seen from these results that the global winner is the R-based mRMR method, with a especially good performance for small sample sizes. Note that the number of variables required by this method is also smaller, in general, than that of the remaining

Table 2. Performance outputs for the considered methods, using k-NN and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (k-NN)	Sample size	MID	FCD	RD	VD	CD	Base
Avgerage accuracy	$n = 30$	80.09	79.26	$\boxed{81.30}$	$\boxed{80.54}$	80.40	78.98
	$n = 50$	81.43	79.91	$\boxed{82.44}$	$\boxed{81.47}$	81.33	80.34
	$n = 100$	$\boxed{83.01}$	80.76	$\boxed{83.82}$	82.54	82.32	81.99
	$n = 200$	$\boxed{84.28}$	81.34	$\boxed{84.89}$	83.37	83.15	83.38
Average dim. red	$n = 30$	9.2	9.8	$\boxed{7.7}$	8.3	$\boxed{8.0}$	100
	$n = 50$	9.3	9.9	$\boxed{7.9}$	8.5	$\boxed{8.1}$	100
	$n = 100$	9.6	10.2	$\boxed{8.2}$	8.7	$\boxed{8.3}$	100
	$n = 200$	9.8	10.4	$\boxed{8.5}$	8.8	$\boxed{8.7}$	100
Victories over Base	$n = 30$	71	51	$\boxed{83}$	$\boxed{72}$	69	-
	$n = 50$	$\boxed{71}$	45	$\boxed{81}$	70	68	-
	$n = 100$	$\boxed{71}$	38	$\boxed{78}$	60	65	-
	$n = 200$	$\boxed{73}$	33	$\boxed{82}$	56	58	-

Table 3. Performance outputs for the considered methods, using LDA and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (LDA)	Sample size	MID	FCD	RD	VD	CD	Base
Avgerage accuracy	$n = 30$	$\boxed{78.72}$	76.87	$\boxed{79.35}$	78.23	78.37	-
	$n = 50$	$\boxed{80.28}$	77.84	$\boxed{80.59}$	79.15	79.36	-
	$n = 100$	$\boxed{81.85}$	78.97	$\boxed{81.88}$	80.22	80.47	-
	$n = 200$	$\boxed{82.96}$	79.83	$\boxed{82.87}$	81.02	81.30	-
Average dim. red	$n = 30$	5.6	$\boxed{4.9}$	5.0	$\boxed{4.6}$	5.2	-
	$n = 50$	6.5	5.9	$\boxed{5.9}$	$\boxed{5.5}$	6.1	-
	$n = 100$	7.9	7.5	$\boxed{7.1}$	$\boxed{6.8}$	7.4	-
	$n = 200$	9.0	8.9	$\boxed{8.0}$	$\boxed{8.0}$	8.3	-

Table 4. Performance outputs for the considered methods, using SVM and the difference criterion, with different sample sizes. Each output is the result of the 100 different models for each sample size.

Output (SVM)	Sample size	MID	FCD	RD	VD	CD	Base
Avgerage accuracy	$n = 30$	$\boxed{81.53}$	79.41	81.50	80.35	80.51	$\boxed{81.91}$
	$n = 50$	$\boxed{82.61}$	80.01	82.45	81.00	81.20	$\boxed{82.99}$
	$n = 100$	$\boxed{83.75}$	80.75	83.45	81.77	82.00	$\boxed{84.11}$
	$n = 200$	$\boxed{84.55}$	81.27	84.22	82.38	82.61	$\boxed{84.91}$
Average dim. red	$n = 30$	10.5	11.0	$\boxed{9.2}$	9.7	$\boxed{9.4}$	100
	$n = 50$	10.5	11.1	$\boxed{9.3}$	9.7	$\boxed{9.6}$	100
	$n = 100$	10.7	11.3	$\boxed{9.6}$	10.0	$\boxed{9.9}$	100
	$n = 200$	10.9	11.5	$\boxed{9.7}$	10.1	$\boxed{9.9}$	100
Victories over Base	$n = 30$	37	39	$\boxed{49}$	$\boxed{43}$	42	-
	$n = 50$	42	34	$\boxed{56}$	44	$\boxed{46}$	-
	$n = 100$	$\boxed{49}$	32	$\boxed{57}$	41	47	-
	$n = 200$	48	29	$\boxed{59}$	42	$\boxed{49}$	-

methods. Moreover, RD is the most frequent winner with respect to the Base method (with all classifiers) keeping, in addition, a more stable general performance when compared with the other variable selection methods. In this sense, R-based methods seem both efficient and reliable. In agreement with the results in [7], the performance of the FC-based method is relatively poor. Finally, note that the Base option (which uses the entire curves) is never the winner, with the partial exception of the SVM classifier.

4.4. *Ranking the methods*

It is not easy to draw general conclusions, and clear recommendations for practitioners, from a large simulation study. A natural idea is to give some kind of quantitative assessment summarizing the relative merits of the different procedures. Many different ranking criteria might be considered. Following Berrendero *et al.* [4], we have considered here the following ones:

- **Relative ranking**: for each considered model and sample size the winner method (in terms of classification accuracy) gets 10 score points and the method with the worst performance gets 0 points. The score of any other method, with performance u, is defined by $10(u-w)/(W-w)$, where W and w denote, respectively, the performances of the best and the worst method.
- **Positional ranking**: The winner gets 10 points, the second best gets 9, etc.
- **F1 ranking**: the scores are assigned according to the current criteria in a Formula 1 Grand Prix: the winner gets 25 score points and the following ones get 18, 15, 10, 8, 6, and 4 points.

The summary results are shown in Tables 5 - 8 and a visual version of the complete (400 experiments) relative ranking outputs for the k-NN classifier is displayed in Figure 2 (analogous figures for the other classification methods can be found in the *Supplemental material* document). The conclusions are self-explanatory and quite robust with respect to the ranking criterion. The mRMR methods based on the distance correlation measure are the uniform global winners. The results confirm the relative stability of R, especially when compared with MI whose good performance is restricted to a few models.

Of course, the criteria for defining these rankings, as well as the idea of averaging over different models, are questionable (although one might think of a sort of Bayesian interpretation for these averages). Anyway, this is the only way we have found to provide an understandable summary for such a large empirical study. On the other hand, since we have made available the whole outputs of our experiments, other different criteria might be used by interested readers.

5. Real data examples

We have chosen three real-data examples on the basis of their popularity in the literature on Functional Data Analysis: we call them *Growth* (93 growth curves in boys and girls), *Tecator* (215, near-infrared absorbance spectra from finely chopped meat) and *Phoneme* (1717 log-periodograms corresponding to the pronounciation of the sounds 'aa' and 'ao'). The respective dimensions of the considered discretizations for these data are 31, 100 and 256. The second derivatives are used for the *Tecator* data. There are many references dealing with these data sets so we will omit here a detailed description of them. See, for example Ramsay and Silverman [27], Ferraty and Vieu [28] and Hastie *et al.* [24],

Table 5. Global scores of the considered methods under three different ranking criteria using NB. Each output is the average of 100 models

Ranking criterion (NB)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	2.43	5.10	$\boxed{8.67}$	7.08	$\boxed{8.10}$
	$n = 50$	3.04	4.31	$\boxed{9.16}$	6.97	$\boxed{7.86}$
	$n = 100$	3.38	3.92	$\boxed{9.28}$	6.84	$\boxed{7.82}$
	$n = 200$	3.84	3.57	$\boxed{9.20}$	6.56	$\boxed{7.59}$
Positional	$n = 30$	6.65	7.62	$\boxed{8.84}$	8.21	$\boxed{8.68}$
	$n = 50$	6.82	7.43	$\boxed{9.12}$	8.19	$\boxed{8.46}$
	$n = 100$	6.87	7.36	$\boxed{9.26}$	8.16	$\boxed{8.35}$
	$n = 200$	6.96	7.30	$\boxed{9.18}$	8.17	$\boxed{8.42}$
F1	$n = 30$	11.64	15.11	$\boxed{18.64}$	16.37	$\boxed{18.24}$
	$n = 50$	12.13	14.54	$\boxed{20.24}$	16.16	$\boxed{16.98}$
	$n = 100$	12.19	14.29	$\boxed{20.82}$	16.17	$\boxed{16.53}$
	$n = 200$	12.38	14.09	$\boxed{20.54}$	16.15	$\boxed{16.92}$

Table 6. Global scores of the considered methods under three different ranking criteria using k-NN. Each output is the average of 100 models

Ranking criterion (k-NN)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	4.01	3.50	$\boxed{9.38}$	6.63	$\boxed{6.64}$
	$n = 50$	4.66	3.09	$\boxed{9.07}$	6.19	$\boxed{6.34}$
	$n = 100$	5.64	2.74	$\boxed{8.96}$	$\boxed{5.94}$	5.78
	$n = 200$	$\boxed{6.58}$	2.34	$\boxed{8.70}$	5.89	5.81
Positional	$n = 30$	7.24	7.14	$\boxed{9.43}$	$\boxed{8.17}$	8.02
	$n = 50$	7.42	7.08	$\boxed{9.39}$	$\boxed{8.14}$	7.97
	$n = 100$	7.71	7.04	$\boxed{9.26}$	$\boxed{8.25}$	7.74
	$n = 200$	8.02	6.95	$\boxed{9.13}$	$\boxed{8.21}$	7.69
F1	$n = 30$	13.37	13.59	$\boxed{21.69}$	$\boxed{16.17}$	15.18
	$n = 50$	13.98	13.39	$\boxed{21.33}$	$\boxed{16.22}$	15.08
	$n = 100$	15.05	13.16	$\boxed{20.46}$	$\boxed{17.03}$	14.30
	$n = 200$	16.33	12.67	$\boxed{19.71}$	$\boxed{16.82}$	14.47

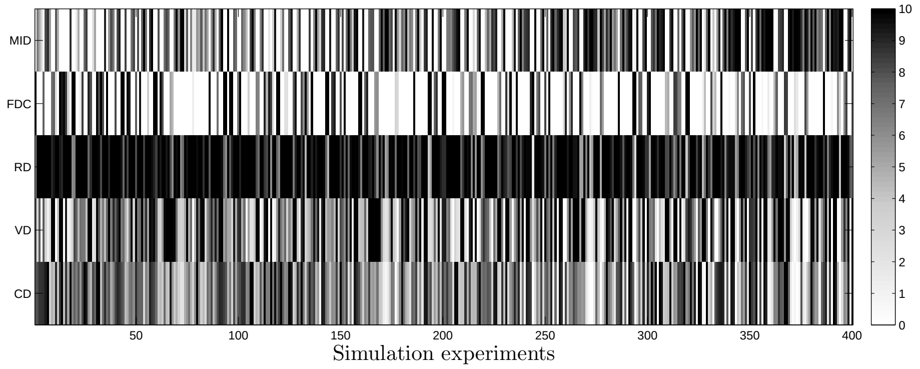

Figure 2. Cromatic version of the global relative ranking table taking into account the 400 considered experiments (columns) and the difference-based mRMR versions with the k-NN classifier: the darker de better.

Table 7. Global scores of the considered methods under three different ranking criteria using LDA. Each output is the average of 100 models

Ranking criterion (LDA)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	5.00	1.98	$\boxed{8.94}$	6.24	$\boxed{6.47}$
	$n = 50$	5.74	1.93	$\boxed{8.77}$	5.65	$\boxed{6.14}$
	$n = 100$	$\boxed{6.07}$	1.94	$\boxed{8.51}$	5.50	5.95
	$n = 200$	$\boxed{6.53}$	2.08	$\boxed{8.44}$	5.36	5.92
Positional	$n = 30$	7.57	6.68	$\boxed{9.31}$	8.17	$\boxed{8.27}$
	$n = 50$	7.78	6.78	$\boxed{9.28}$	8.00	$\boxed{8.16}$
	$n = 100$	7.85	6.90	$\boxed{9.14}$	8.02	$\boxed{8.09}$
	$n = 200$	7.99	6.86	$\boxed{9.11}$	8.01	$\boxed{8.03}$
F1	$n = 30$	14.69	11.81	$\boxed{20.86}$	$\boxed{16.51}$	16.13
	$n = 50$	15.56	12.13	$\boxed{20.60}$	15.72	$\boxed{15.99}$
	$n = 100$	15.81	12.39	$\boxed{19.86}$	$\boxed{16.07}$	15.87
	$n = 200$	$\boxed{16.29}$	12.25	$\boxed{20.11}$	15.79	15.56

Table 8. Global scores of the considered methods under three different ranking criteria using SVM. Each output is the average of 100 models

Ranking criterion (SVM)	Sample size	MID	FCD	RD	VD	CD
Relative	$n = 30$	$\boxed{6.32}$	2.99	$\boxed{8.10}$	5.34	5.57
	$n = 50$	$\boxed{6.63}$	3	$\boxed{8.28}$	5.07	5.70
	$n = 100$	$\boxed{6.82}$	2.87	$\boxed{8.13}$	4.97	5.59
	$n = 200$	$\boxed{7.19}$	2.45	$\boxed{8.24}$	5.06	5.28
Positional	$n = 30$	$\boxed{8.07}$	7.22	$\boxed{9.06}$	7.87	7.78
	$n = 50$	$\boxed{8.09}$	7.20	$\boxed{9.09}$	7.78	7.84
	$n = 100$	$\boxed{8.22}$	7.19	$\boxed{9.02}$	7.84	7.73
	$n = 200$	$\boxed{8.32}$	7.05	$\boxed{9.15}$	7.83	7.65
F1	$n = 30$	$\boxed{16.55}$	13.98	$\boxed{19.63}$	15.35	14.49
	$n = 50$	$\boxed{16.61}$	13.86	$\boxed{19.80}$	14.94	14.79
	$n = 100$	$\boxed{17.17}$	13.84	$\boxed{19.31}$	15.29	14.39
	$n = 200$	$\boxed{17.43}$	13.10	$\boxed{20.10}$	15.09	14.28

respectively, for additional details.

The methodology followed in the treatment of these data sets is similar to that followed in the simulation study, with a few technical differences. For *Tecator* and *Growth* data sets, a standard leave-one-out cross-validation is used. Such a procedure turns out to be too expensive (in computational terms) for the *Phoneme* data set. So in this case we have carried out 50-fold cross validation; see, for example, [24, Sec. 7.10] for related ideas.

A summary of the comparison outputs obtained for these data sets using the different mRMR criteria (as well as the benchmark 'Base' comparison, with no variable selection) is given in Table 9. Again, the letter D in MID, FCD, etc. indicates that the relevance and redundancy measures are combined by difference. The analogous outputs using the quotient (instead of the difference) can be found in the *Supplemental material* file.

The conclusions are perhaps less clear than those in the simulation study. The lack of a uniform winner is apparent. However, the R-based method is clearly competitive and might even be considered as the global winner, taking into account both, accuracy

Table 9. Performances of the different mRMR methods in three data sets. From top to bottom tables stand for Naive Bayes, k-NN, LDA and linear SVM outputs respectively.

NB outputs

Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	$\boxed{92.47}$	87.10	$\boxed{89.25}$	87.10	86.02	84.95
	Tecator	98.60	97.67	$\boxed{99.53}$	$\boxed{99.53}$	98.14	97.21
	Phoneme	79.03	$\boxed{80.27}$	$\boxed{80.49}$	79.39	80.14	74.08
Number of variables	Growth	2.0	$\boxed{1.1}$	2.2	$\boxed{1.0}$	1.3	31
	Tecator	2.0	5.9	$\boxed{1.0}$	$\boxed{1.0}$	3.3	100
	Phoneme	12.6	$\boxed{10.3}$	15.8	$\boxed{5.8}$	15.9	256

k-NN outputs

Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	$\boxed{95.70}$	83.87	94.62	91.40	84.95	$\boxed{96.77}$
	Tecator	99.07	99.07	$\boxed{99.53}$	$\boxed{99.53}$	99.07	98.60
	Phoneme	80.14	80.48	$\boxed{81.14}$	80.31	$\boxed{80.55}$	78.80
Number of variables	Growth	3.5	$\boxed{1.0}$	2.5	4.8	$\boxed{1.1}$	31
	Tecator	5.7	3.0	$\boxed{1.0}$	$\boxed{1.0}$	4.0	100
	Phoneme	15.4	$\boxed{13.3}$	17.7	16.5	$\boxed{10.7}$	256

LDA outputs

Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	$\boxed{94.62}$	91.40	$\boxed{94.62}$	$\boxed{94.62}$	89.25	-
	Tecator	$\boxed{95.81}$	93.95	94.88	$\boxed{95.81}$	94.88	-
	Phoneme	$\boxed{79.50}$	79.34	79.21	79.39	$\boxed{79.98}$	-
Number of variables	Growth	$\boxed{3.4}$	5.0	$\boxed{3.1}$	4.2	5.0	-
	Tecator	$\boxed{2.6}$	8.8	5.6	$\boxed{5.0}$	$\boxed{5.0}$	-
	Phoneme	19.1	$\boxed{8.8}$	14.6	17.1	$\boxed{12.0}$	-

SVM outputs

Output	Data	MID	FCD	RD	VD	CD	Base
Classification accuracy	Growth	94.62	87.10	94.62	$\boxed{95.70}$	86.02	$\boxed{95.70}$
	Tecator	98.14	99.07	$\boxed{99.53}$	$\boxed{99.53}$	98.60	99.07
	Phoneme	$\boxed{80.90}$	80.83	80.67	80.78	80.67	$\boxed{80.96}$
Number of variables	Growth	$\boxed{3.4}$	5.0	$\boxed{2.5}$	4.2	5.0	31
	Tecator	6.7	2.0	$\boxed{1.0}$	$\boxed{1.0}$	4.1	100
	Phoneme	18.5	$\boxed{8.6}$	16.2	16.7	$\boxed{16.0}$	256

and amount of dimension reduction. The *Tecator* outputs are particularly remarkable since RD and VD provide the best results (with three different classifiers) using just one variable. Again, variable selection methods beat here the 'Base' approach (except for the Growth example) in spite of the drastic dimension reduction provided by the mRMR methods.

6. Final conclusions and comments

The mRMR methodology has become an immensely popular tool in the machine learning and bioinformatics communities. For example, the papers by Ding and Peng [7] and Peng *et al.* [8] had 819 and 2430 citations, respectively on Google Scholar (by October 2, 2014). As we have mentioned, these authors explicitly pointed out the need of further research, in order to get improved versions of the mRMR method. The idea would be to keep the basic mRMR paradigm but using other association measures (besides the mutual information). This paper exactly follows such line of research, with a particular focus on the classification problems involving functional data.

We think that the results are quite convincing: our extensive simulation study (based on 1600 simulation experiments and real data) places the mRMR method based in the R association measure by Székely *et al.* [12] globally above the original versions of the mRMR paradigm. This is perhaps the main conclusion of our work. The good performance of the distance correlation in comparison with the other measures can be partially explained by the facts that this measure captures non-linear dependencies (unlike C and FC), has a simple smoothing-free empirical estimator (dissimilar to MI) and is normalized (different from V).

There are, however, some other more specific comments to be made.

(1) First of all, variable selection is worthwhile in functional data analysis. Accuracy can be kept (and often improved) using typically less than the 10% of the original variables, with the usual benefits of the dimension reduction. This phenomenon happens for all the considered classifiers.

(2) The average number of selected variables with the R- or V-based methods is also smaller than that of MI and FC (that is, the standard mRMR procedures). This entails an interpretability gain: the fewer selected variables, the stronger case for interpreting the meaning of such selection in the context of the considered problem.

(3) The advantage of the R-based methods over the remaining procedures is more remarkable for the case of small sample sizes. This looks as a promising conclusion since small samples are very common in real problems (e.g. in biomedical research).

(4) In those problems involving continuous variables there is a case for using nonparametric kernel density estimators in the empirical approximation of the mutual information criterion. However, these estimators are known to be highly sensitive to the selection of the smoothing parameter which can be seen as an additional unwelcome complication. On the other hand, the results reported so far (e.g. in [8]) do not suggest that kernel estimators will lead to a substantial improvement over the simplest, much more popular discretization estimators (see e.g. [21, 22]).

(5) Still in connection with the previous remark, it is worth noting the lack of smoothing parameters in the natural estimators of V and R.[12] This can be seen as an additional advantage of the R- or V-based mRMR method.

(6) The better performance of R when compared with V can be explained by the fact that R is normalized so that relevance (1) and redundancy (2) are always measured ‘in the same scale’. Otherwise, one of these two quantities could be overrated by the mRMR algorithm, specially when the difference criterion is used.

(7) The method FCD (sometimes suggested in the literature as a possible good choice) does not appear to be competitive. It is even defeated by the simple correlation-based method CD.

(8) In general, the difference-based methods are preferable to their quotient-based counterparts. The quotient-based procedures are only slightly preferable when combined

with methods (FC, V) where relevance and redundancy are expressed in different scales. The outputs for these quotient-based methods can be found in the complete list of results `www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx`, and a summary is available in *Supplemental material* document.

(9) We should emphasize again that the goal of this paper is to propose new versions of the mRMR method and to compare them with the standard ones. Therefore, a wider study involving comparisons with other dimension reduction methods, is beyond the scope of this work. The recent paper by Berrendero *et al.* [4] includes a study of this type (always in the functional setting) whose conclusions suggest that mRMR might be slightly outperformed by the Maxima-Hunting (MH) procedure proposed by these authors. It also has a very similar performance to that of Partial Least Squares (PLS), although PLS is harder to interpret. Moreover, the number of variables selected by MH is typically smaller than those required by mRMR.

(10) Finally, if we had to choose just one among the considered classification methods, we should probably take k-NN. The above commented advantages in terms of ease of implementation and interpretability do not entail any significant price in efficiency.

Acknowledgements

This research has been partially supported by Spanish grant MTM2010- 17366.

Supplemental material

The *Supplemental material* document contains: the complete list and description of all functional models, the summary Tables 1 - 9 with the quotient criterion instead of the difference one, figures analogous to Figure 2 with NB, LDA and SVM, and some new tables with a few simulation results. All outputs (with both difference and quotient criteria) of the 1600 simulation experiments and real data can be found at `www.uam.es/antonio.cuevas/exp/mRMR-outputs.xlsx`.

References

[1] Golub, T.R. and Slonim, D.K. and Tamayo, P. and Huard, C. and Gaasenbeek, M. and Mesirov, J.P. and Coller, H. and Loh, M.L. and Downing, J.R. and Caligiuri, M.A. and others. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science.*1999;286:531–537.

[2] Lindquist, M.A. and McKeague, I.W. Logistic regression with Brownian-like predictors. *Journal of the American Statistical Association.* 2009;104:1575-1585.

[3] Guyon, I. and Gunn, S. and Nikravesh, M. and Zadeh, L.A. Feature Extraction: Foundations and Applications. Springer. 2006.

[4] Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. Variable selection in functional data classification: a maxima hunting proposal. *Unpublished manuscript.* 2014.

[5] Baíllo, A., Cuevas, A. and Fraiman, R. Classification methods with functional data. In *Oxford Handbook of Functional Data Analysis*, 2011;pp-259–297. In: Ferraty F. and Romain Y., editors. Oxford University Press.

[6] Baíllo, A., Cuesta-Albertos, J. A. and Cuevas, A. Supervised classification for a family of Gaussian functional models. *Scand. J. Stat.* 2011;38:480–498.

[7] Ding, C. and Peng, H. Minimum redundancy feature selection from microarray gene expression data. J. Bioinform. Comput. Biol. 2005;3:185–205.

[8] Peng, H., Long, F. and Ding, C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 2005;27:1226–1238.

[9] Battiti, R. Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on.* 1994;5:537–550.

[10] Kwak, N. and Choi, C.H. Input feature selection by mutual information based on Parzen window. *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* 2002;24:1667–1671

[11] Yu, L. and Liu, H. Efficient feature selection via analysis of relevance and redundancy. *The Journal of Machine Learning Research.* 2004;5:1205–1224.

[12] Székely, G. J., Rizzo, M. L. and Bakirov, N. K. Measuring and testing dependence by correlation of distances. *Ann. Statist.* 2007;35:2769–2794.

[13] Székely, G. J. and Rizzo, M. L. Brownian Distance Covariance. *Ann. Appl. Stat.* 2009;3:1236–1265.

[14] Hall, P. and Miller, H. Determining and depicting relationships among components in high-dimensional variable selection. *J. Comput. Graph. Statist.* 2011;20:988–1006.

[15] Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R. McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M. and Sabeti, P. C. Detecting novel associations in large data sets. *Science* 2011;334:1518–1524.

[16] Székely, G. J. and Rizzo, M. L. On the uniqueness of distance covariance. *Statist. Probab. Lett.* 2012;82:2278–2282.

[17] Székely, G. J. and Rizzo, M. L. Energy statistics: a class of statistics based on distances. *J. Plann. Statist. Infer.* 2013;143:1249–1272.

[18] Wand, M.P. and Jones, M.C. *Kernel smoothing.* Chapman & Hill. 1995.

[19] Cao, R. and Cuevas, A. and Gonzalez-Manteiga, W. A comparative study of several smoothing methods in density estimation *Computational Statistics & Data Analysis.* 1994;17:153–176.

[20] Estévez, P.A. and Tesmer, M, and Perez, C.A. and Zurada, J.M. Normalized mutual information feature selection. *Neural Networks, IEEE Transactions on* 2009;20:189–201.

[21] Mandal, M. and n Mukhopadhyay, A. A novel PSO-based graph-theoretic approach for identifying most relevant and non-redundant gene markers from gene expression data. To appear in *International Journal of Parallel, Emergent and Distributed Systems.* 2014.

[22] Nguyen, X.V. and Chan, J. and Romano, S. and Bailey, J. Effective global approaches for mutual information based feature selection. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* 2014:512–521.

[23] Demler, D.V., Pencina, M.J. and D'Agostino, R.B. Impact of correlation on predictive ability of biomarkers. *Stat Med.* 2013;32:4196–4210.

[24] Hastie, T. and Tibshirani, R. and Friedman, J. and Franklin, J. *The elements of statistical learning: data mining, inference and prediction.* Springer. 2005.

[25] Hand, D. *Classifier technology and the illusion of progress. Statist. Sci.* 2006;21:1–34.

[26] Fan, R.-E. and Chang, K.-W. and Hsieh, C.-J, and Wang, X.-R. and Lin C.-J. *LIBLINEAR: A Library for Large Linear Classification. Journal of Machine Learning Research* 2008;9:1871–1874.

[27] Ramsay, J.O. and Silverman, B.W. *Functional data analysis.* Springer. 2005.

[28] Ferraty, F. and Vieu, P. *Nonparametric Functional Data Analysis: Theory and Practice.* Springer. 2006.