A COMPARISON OF MULTIVARIATE MUTUAL INFORMATION ESTIMATORS FOR FEATURE SELECTION

Gauthier Doquire and Michel Verleysen
Machine Learning Group - ICTEAM, Université catholique de Louvain,
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

Keywords: Mutual information estimation, Feature selection, Nearest neighbors.

Abstract: Mutual Information estimation is an important task for many data mining and machine learning applications. In particular, many feature selection algorithms make use of the mutual information criterion and could thus benefit greatly from a reliable way to estimate this criterion. More precisely, the multivariate mutual information (computed between multivariate random variables) can naturally be combined with very popular search procedure such as the *greedy forward* to build a subset of the most relevant features. Estimating the mutual information (especially through density functions estimations) between high-dimensional variables is however a hard task in practice, due to the limited number of available data points for real-world problems. This paper compares different popular mutual information estimators and shows how a nearest neighbors-based estimator largely outperforms its competitors when used with high-dimensional data.

1 INTRODUCTION

The ways to acquire and store data increase every day; machine learning practitioners often have to deal with databases of very large dimension (containing data described by a lot of features). When considering a prediction task, all the features are not equally relevant to predict the desired output while some can be redundant; irrelevant or redundant features can increase the variance of the prediction models without reducing their bias while most of distance-based methods are quite sensitive to useless features. More generally, learning with high-dimensional data is a hard task due to the problems related to the *curse of dimensionality* (Bellman, 1961).

Two main approaches exist to reduce the dimensionality of a data set. One solution is to project the data on a space of smaller dimension. Projections can be very effective but do not preserve the original features; this is a major drawback in many industrial or medical applications where interpretability is primordial. On the contrary, feature selection, by trying to find a subset of features with the largest prediction power, does allow such an interpretability.

Even if many ways of selecting features can be thought of, this paper focuses on filters. Filters are independent from the model used for regression or classification and thus do not require building any prediction model (including time-consuming learning and potential meta-parameters to tune by resampling methods). They are faster than wrappers which try to find the best subset of features for a specific model through extensive simulations. Filters are often based on an information-theoretic criterion measuring the quality of a feature subset and a search procedure to find the subset of features maximising this criterion; the mutual information (MI) criterion (Shannon, 1948) has proven to be very efficient for feature selection and has been used successfully for this task since many years (see e.g. (Rossi et al., 2007)).

As it is not possible in practice to evaluate the MI between all the 2^{f-1} (f being the initial number of features) possible feature subsets and the output vector when f is large, incremental greedy procedures are frequently used, whose most popular ones are forward, backward or forward/backward. Such procedures are said to be *multivariate*, in the sense that they require the evaluation of the MI (or of another chosen criterion) directly between a set of features and the output vector. These methods have the advantage over bivariate ones such as ranking that they are able to detect subsets of features which are jointly relevant or redundant. Consider the XOR problem as a simple example; it consists in two features and an output scalar which is zero if both features have the same value and one otherwise. Obviously, indi-

Doquire G. and Verleysen M. (2012).
A COMPARISON OF MULTIVARIATE MUTUAL INFORMATION ESTIMATORS FOR FEATURE SELECTION.
In *Proceedings of the 1st International Conference on Pattern Recognition Applications and Methods*, pages 176-185
DOI: 10.5220/0003726101760185

vidually each feature does not carry any information about the outptut; univariate procedures will never be able to detect them as relevant. However, when combined, the two features completely determine the output; when one is selected, a multivariate procedure will select the other one as relevant. A detailed introduction to the feature selection problem can be found in (Guyon and Elisseeff, 2003).

As will be seen, the MI generally cannot be computed analytically but has to be estimated from the data set. Even if this task has been widely studied, it remains very challenging for high-dimensional vectors. In this paper, it is shown how a MI estimator based on the principle of nearest neighbors (NN) outperforms traditional MI estimators with respect to three feature selection related criteria. This study is, to the best of our knowledge, the first one to compare MI estimators in such a context.

The rest of the paper is organized as follows. Section 2 briefly introduces the MI criterion and describes five of the most popular MI estimators. Section 3 presents the experiments carried out to compare these estimators and shows the results obtained on artificial and real-world data sets. Discussions and conclusions are given in Section 4.

2 MUTUAL INFORMATION

This section recalls basic notions about the MI and briefly presents the estimators used for comparison.

2.1 Definitions

Mutual information (Shannon, 1948) is a symmetric measure of the dependence between two (groups of) random variables X and Y, assumed to be continuous in this paper. Its interest for feature selection comes mainly from the fact that MI is able to detect nonlinear relationships between variables, whereas, as an example, it is not the case for the popular correlation coefficient which is limited to linear dependencies. Moreover, the MI can be naturally defined for groups of variables and is thus well-suited for multivariate search procedures. MI is formally defined as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \qquad (1)$$

where $H(X)$ is the entropy of X, defined for a continuous random variable as:

$$H(X) = -\int f_X(\zeta_X) \log f_X(\zeta_X)\, d\zeta_X. \qquad (2)$$

In this last equation, f_X is the probability density function (pdf) of X. The MI can then be rewritten as

$$I(X;Y) = \int\int f_{X,Y}(\zeta_X,\zeta_Y) \log \frac{f_{X,Y}(\zeta_X,\zeta_Y)}{f_X(\zeta_X) f_Y(\zeta_Y)}\, d\zeta_X\, d\zeta_Y. \qquad (3)$$

In practice, neither f_X, f_Y nor $f_{X,Y}$ are known for real-world problems; the MI has thus to be estimated.

2.2 Estimation

Plenty of methods have been proposed in the literature to estimate the MI. The great majority of them starts by estimating the unknown pdf before plugging these results into Equation (1) or an equivalent expression. However, the dimension of X increases at each step of a forward feature selection procedure (or is already very high at the beginning of a backward procedure) and most of these methods suffer dramatically from the curse of dimensionality (Bellman, 1961); indeed they require an exponentially growing number of samples as the dimension of X grows while the number of available samples is in practice often very limited. Such MI estimations do not thus seem well suited for feature selection ends. A NN-based MI estimator (Kraskov et al., 2004) avoiding the pdf estimation step has been used successfully in a feature selection context (Francois et al., 2007; Rossi et al., 2007). In the rest of this section, this estimator and four popular other ones are introduced.

2.2.1 The Basic Histogram

The histogram is one of the oldest and simplest ways to estimate a pdf. The basic idea is to divide the observation, prediction and joint spaces into non overlapping bins of fixed size and then to count the number of points falling in each of the bins. The entropy of X, Y and (X,Y) can be estimated using the discretized version of (2) and the estimation of the MI then naturally follows from (1). If histograms with bins of the same fixed size are considered, as it is the case in this paper, the size of the bins needs to be determined. Here, the approach by Sturges (Sturges, 1926) will be followed: the number k of bins will be $\lceil 1 + \log_2(N) \rceil$, where N is the number of samples in the data set; other approaches could also be thought of (Scott, 1979).

2.2.2 The Kernel Estimator

The basic histogram suffers from many drawbacks. Among others, it is sensitive to the choice of the origin and to the size of the bins. In order to avoid sharp steps between the bins (and hence discontinuities), one can use the kernel density estimator (KDE) given by:

$$\hat{f}_X(x) = \frac{1}{Nh}\sum_{i=1}^{N} K(\frac{x-x_i}{h}), \tag{4}$$

where N is the number of observations in X, h is the window width and K is the kernel function required to integrate to one, leading $\hat{f}$ to be a probability density (Parzen, 1962); x_i denotes the i^{th} observation of the data set X. One possible choice for K is the Gaussian kernel, leading to the following density estimator:

$$\hat{f}(x) = \frac{1}{Nh\sqrt{2\pi}}\sum_{i=1}^{N} \exp(\frac{-(x-x_i)^2}{2h^2}). \tag{5}$$

In practice, the choice of the bandwidth h is fundamental. In this paper, the approach by Silverman (Silverman, 1986) using a *rule of thumb* will be followed. It is often used as a good trade-off between performance and computational burden. The idea is to choose the width minimizing the asymptotic mean integrated square error (AMISE) between the estimation and the true density, assuming the underlying distribution is Gaussian. The resulting width is:

$$\hat{h}_{rot} \approx \sigma(\frac{4}{f+2})^{1/(f+4)} N^{-1/(f+4)} \tag{6}$$

where f is again the dimensionality of X. A large overview of different ways to select the kernel bandwidth is given in (Turlach, 1993).

2.2.3 The B-splines Estimator

Another generalisation of the simple binning approach is given by the use of B-splines functions (Daub et al., 2004). The idea is again to first discretize the X, Y and (X,Y) spaces. However, in this approach, the data points are allowed to be assigned to more than one bin a_i simultaneously in order to prevent the positions of the borders of the bins from affecting too much the estimation. The weights with which each point belongs to a bin are given by the B-spline functions $B_{i,k}$ (k being the spline order). Without getting too much into details, B-splines are recursively defined as:

$$B_{i,1}(x) := \begin{cases} 1 & \text{if } t_i \leq x \leq t_{i+1} \\ 0 & \text{otherwise} \end{cases}$$
$$B_{i,k}(x) := B_{i,k-1}(x)\frac{x-t_i}{t_{i+k-1}-t_i} + B_{i+1,k-1}(x)\frac{t_{i+k}-x}{t_{i+k}-t_{i+1}} \tag{7}$$

where t is a knot vector defined for a number of bins M and a spline order $k = 1...M-1$ as:

$$t_i := \begin{cases} 0 & \text{if } i < k \\ i-k+1 & \text{if } k \leq i \leq M-1 \\ M-1-k+2 & \text{if } i > M-1 \end{cases} \tag{8}$$

To estimate the density $\hat{f}_x$, M_X weights $B_{i,k}(x_u)$ are determined for each datapoint x_u (where M_X is the number of bins in the X space). As the sum of the weights corresponding to each data point is 1, the sum of the mean values of each bin is also 1. The weights can thus be seen as the probability of each bin ($p(a_i) = \frac{1}{N}\sum_{u=1}^{N} B_{i,k}(x_u)$) and the entropy of the distribution can be estimated. The process is repeated for the Y space and for the joint (X,Y) space to estimate the MI. The notion of B-splines can be extended to the multivariate case from univariate splines by the tensor produt construct. As an example, in two dimensions, the probability of a bin $a_{i,j}$ is given by $p(a_{i,j}) = \frac{1}{N}\sum_{u=1}^{N} B_{i,k}(x_u) \times B_{j,k}(y_u)$ where x denotes the first variable and y the second one.

2.2.4 The Adaptive Partition of the Observation Space

Darbellay and Vajda proved (Darbellay and Vajda, 1999) that the MI can be approximated arbitrarily closely in probability by calculating relative frequencies on appropriate partitions. More precisely, they use an adaptive partitioning of the observation scheme, different from the traditional product partitions, to take into account the fact that with such basic partitions, much of the bins are not used to estimate the MI and can be replaced by fewer bins; they proved the weak consistency of the proposed method. Mathematical details can be found in (Darbellay and Vajda, 1999). In the rest of this paper, this methodology will be denoted *adaptive histogram*.

2.2.5 The Nearest Neighbors-based or Kraskov Estimator

Since the hardest part when estimating the MI is the estimation of the underlying probability densities, another alternative is simply not to estimate densities and therefore directly estimating the MI by using NN statistics. The intuitive idea behind Kraskov's estimator (Kraskov et al., 2004) is that if the neighbors of a specific observation in the X space correspond to the same neighbors in the Y space, there must be a strong relationship between X and Y. More formally, the estimator is based on the Kozachenko-Leonenko estimator of entropy defined as:

$$\hat{H}(X) = -\psi(K) + \psi(n) + \log(c_d) + \frac{d}{N}\sum_{n=1}^{N} \log(\varepsilon_X(N,K)) \tag{9}$$

where ψ is the digamma function, N the number of samples in X, d the dimensionality of these samples, c_d the volume of a d-dimensional unitary ball and $\varepsilon_X(n,K)$, twice the distance (usually chisen as the Euclidean distance) from the n^{th} observation in X to its K^{th} NN. Two slightly different estimators are then derived whose most popular one is:

$$\hat{I}(X;Y) = \psi(N) + \psi(K) - \frac{1}{K} - \frac{1}{N}\sum_{i=1}^{N}(\psi(\tau_{x_i}) + \psi(\tau_{y_i})) \tag{10}$$

where τ_{x_i} is the number of points whose distance from x_i is not greater than $0.5 \times \varepsilon(n,K) = 0.5 \times \max(\varepsilon_X(n,K), \varepsilon_Y(n,K))$. By avoiding the evaluation of high-dimensional pdf, the hope is to reach better results than with the previously introduced estimators.

It is also important to note that other NN based density estimators have been proposed in the litterature, whose recent examples are (Wang et al., 2009; Li et al., 2011). However, as they are less popular than (Kraskov et al., 2004) for feature selection, they are not used in the present comparison.

3 EXPERIMENTS

Three sets of experiments are carried out in this section. The objective is to assess the interest of the different estimators for *incremental feature selection algorithms*. The criteria of comparison and the experimental setup are thus very different from the ones used in previous papers only focused on MI estimation (see e.g. (Walters-Williams and Li, 2009)). First, a suitable estimator should be accurate, i.e. it should reflect the true dependency between groups of features and increases (resp. decreases) when the dependance between groups of features increases (resp. decreases). Then it should also be able to detect uninformative features and return a value close to zero when two independent groups of features are given. Eventually, a good estimator should be quite independent from the value of its parameters or some fast heuristics to fix them should be available.

From a practical point of view, the implementation by Alexander Ihler has been used for KDE[1]. For the NN-based estimator, the parameter K is set to 6 unless stated otherwise. For the B-splines estimator, the degree of the splines is set to 3 and the number of bins to 3. These values correspond to those advised in the respective original papers (Kraskov et al., 2004; Daub et al., 2004).

3.1 Accuracy of the Estimators

The first set of experiments consists in comparing the precision of the MI estimators as the dimension of the data set increases. To this end, they will be used to estimate the MI between n correlated Gaussians $X_1 \ldots X_n$ with zero mean and unit variance. This way, the experimental results can be compared with exact analytical expressions as the MI for n such Gaussians is given by (Darbellay and Vajda, 1999):

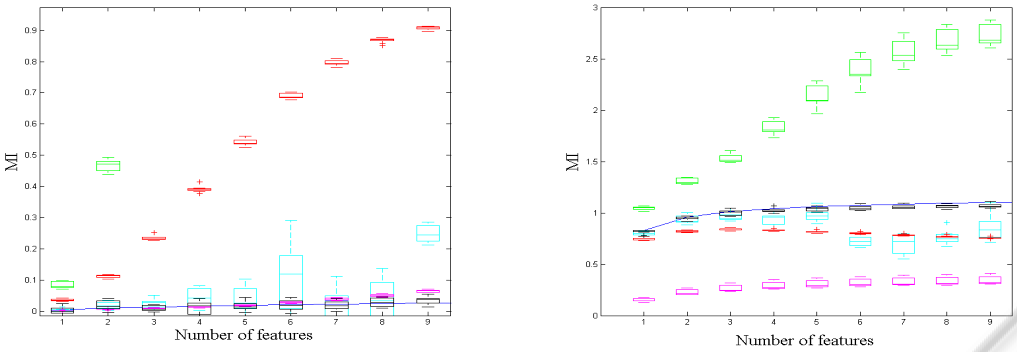
$$I(X_1 \ldots X_n) = -0.5 \times \log[det(\sigma)] \tag{11}$$

where σ is the covariance matrix.

All the correlation coefficients are set to the same value r which will be chosen to be 0.1 and 0.9. The estimation is repeated 100 times on randomly generated datasets of 1000 instances and the results are shown for $n = 1...9$. Even if this can be seen as a relatively small number of dimensions, there are practical limitations when using splines and histogram-based estimators in higher dimensions. Indeed the generalization of the B-splines-based estimator to handle vectors of dimension d involves the tensor product of d univariate B-splines, a vector of size M^d, where M is the number of bins. Histogram-based methods are also limited in the same way since they require the storage of the value of k^d bins, where k is the number of bins per dimension. Nearest neighbors-based methods are not affected by this kind of problems and have only a less restrictive limitation regarding the number n of data points since they require the calculation of $O(n^2)$ pairwise distances. As will be seen, the small number of dimensions used in the experiments is sufficient to underline the drawbacks and advantages of the compared estimators.

Figure 1 shows that, as far as the precision is concerned, Kraskov *et al.*'s estimator largely outperforms its competitors for the two values of r ($r = 0.1$ and $r = 0.9$). The estimated values are always very close to the true ones and show small variations along the 100 repetitions. The adaptive histogram provides on average accurate estimations up to dimension 8 and 6 for $r = 0.1$ and $r = 0.9$ respectively, with however very strong fluctuations observed accross the experiments. The B-spline estimator is also extremely accurate for the five first dimensions and $r = 0.1$. For $r = 0.9$ (and thus for higher values of MI), it severely underestimates the true values while the aspect of the true MI curve is preserved. This cannot be considered as a major drawback in a feature selection context where we are interested by the *comparison* of MI between groups of features. The results achieved by the kernel density estimator are very poor as soon as n exceeds 1, largely overestimating the true values for $r = 0.1$ while immediately decreasing for $r = 0.9$. Finally, as one could expect, the basic histogram produces the worst results; the estimated values are too

[1] http://www.ics.uci.edu/ ihler/code /



r=0.1

r=0.9

Figure 1: Boxplots of the approximation of the MI for correlated Gaussian vectors by several estimators: the basic histogram (green), a KDE (red), an adaptive histogram (cyan), a NN-based estimator (black) and a B-splines estimator (magenta). The solid line represents the true MI.

high to be reported on Figure 1 for $r = 0.1$ when the dimension of the the data exceeds two.

3.2 Mutual Information between Independent Variables

In a feature selection context, a suitable estimator should assign a value close to zero to the MI between independent (groups of) variables. More precisely, one has to make sure that a large (greatly above zero) value of MI is not the result of a weakness or a bias of the estimator but does correspond to a dependence between the variables. Moreover, as the MI is not bounded to a known interval (as $[-1,1]$ for the correlation coefficient), the relevance of each feature subset cannot be directly assessed based only on the value of the MI. A solution is to establish the relevance of a feature subset by looking if the MI between this subset and the outptut vector is significantly larger than the MI between this subset and a randomly permuted version of the output. It is thus important in practice to study how the MI between the actual data points and a randomly permuted objective vector is estimated. In theory, the MI estimated in this way should be 0 as no more relationship exists between the observations and the permuted outputs.

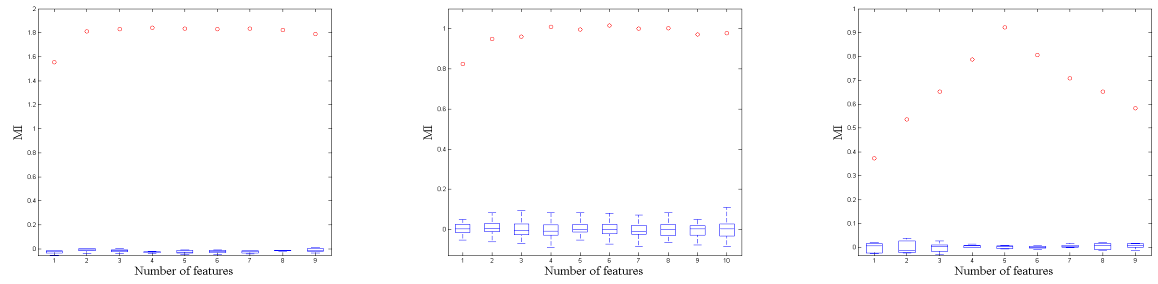
Experiments have been carried out on one artificial and two real-world data sets. The artificial problem is derived from Friedman (Friedman, 1991) and is often used as a benchmark for feature selection algorithms. It consists of 10 input variables X_i uniformly distributed over $[0,1]$ and an output variable Y given by $Y = 10\sin(X_1X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$ where ε is a Gaussian noise with zero mean and unit variance. The sample size is 1000 and 100 data sets are randomly generated. As can be deducted easily, only the five first features are relevant to predict Y.

The first real data set is the well known Delve census data set, available from the University of Toronto[2] for which the 2048 first entries of the training set are kept. The dimension of the data set is 104. The second real data set is the Nitrogen data set[3], containing only 141 spectra discretized at 1050 different wavelengths. The goal is to predict the Nitrogen content of a grass sample. As pre-processing, each spectrum is represented using its coordinates in a B-splines basis, in order to reduce the amount of features to a reasonable number of 105 ((Rossi et al., 2005)). For each data set, a forward feature selection procedure using the NN-based estimator is conducted (since it performed the best in the previous section) and is halted when nine features have been selected. The MI is then estimated as well as the MI with the permuted output for 100 random permutations of the output and for each of the nine subsets of features of increasing dimension. The performance of the estimators is thus compared on the same sets of relevant features.

In Figure 2, it can be seen that for the three problems, the NN-based estimator used with permuted output produces values very close to 0, even when working with few samples as for the Nitrogen data set (the variance is however larger in this case). This satisfactory observation is in good agreement with previous results found in (Kraskov et al., 2004) where the authors conjectured the fact that equation (10) is exact for independent variables, without proof of this result. Let us also notice two undesirable facts about the estimator. First it sometimes produces slightly negative values. Even if this has no theoretical justification (Cover and Thomas, 1991), this can easily

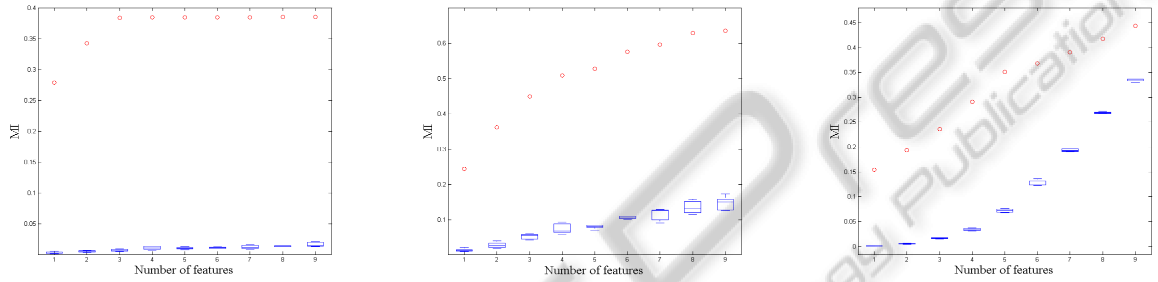
[2] http://www.idrc-chambersburg.org/index.html

[3] http://kerouac.pharm.uky.edu/asrg/cnirs/



(a) (b) (c)

Figure 2: Estimated MI between a group of features and the output (circles) and boxplots of the estimated MI between the same features and a permuted output for the NN-based estimator: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.



(a) (b) (c)

Figure 3: Estimated MI between a group of features and the output (circles) and boxplots of the estimated MI between the same features and a permuted output for the B-splines density estimator: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

be dealt with in practice, by setting negative values to 0. Secondly, it can be seen that the MI decreases after the addition of some variables. Once again, this phenomenon is not theoretically founded (Cover and Thomas, 1991) even if it has often been used as a stopping criterion in greedy feature selection algorithms.

The B-splines estimator (Figure 3) also performs well on the Delve data set. However, the results on the two other data sets contrast with this behaviour; as far as the artificial data set is concerned, the eight and nine first features have a higher MI with the permuted output than the first three with the actual output. This can also be understood as the eight and nine first permuted features having a higher MI with the output than the three first original features have. This is of course a very undesirable fact in the context of feature selection. Indeed, it is obvious that permuted features do not carry any information about Y while the first three original ones actually do.

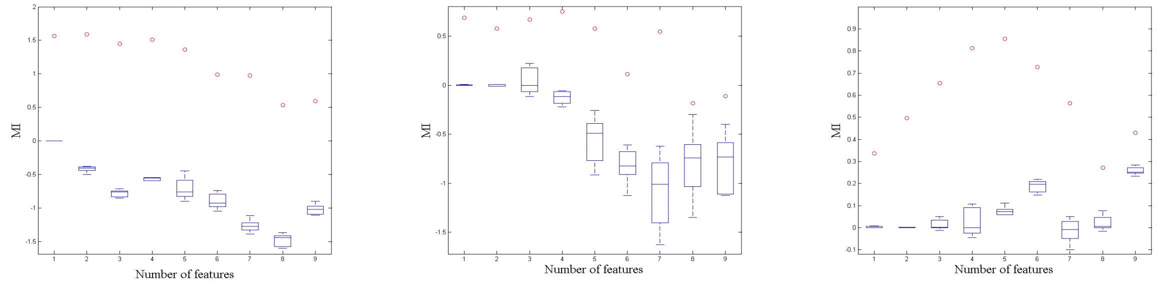
The adaptative histogram (Figure 4), produces highly negative values for the Delve and the Nitrogen data sets. Even if the same *trick* as the one used for the Kraskov estimator could also be applied here (setting the negative values to 0), things does not behave so well. First, the absolute values of the negative results are very large, traducing instabilities of the algorithm as the dimension increases. Next, for the Nitrogen data set, the first third and fourth features have a higher MI with the permuted output than the first eight and nine have with the actual output. For the artificial data set, the first nine features have a higher MI with the permuted output than the first six have with the true output.

The KDE (Figure 5) also returns values highly above 0 with the permuted output; on the artificial data set, the MI between the features and the actual or the permuted output becomes equal as the dimension increases. However, no confusion is possible for the two real-world data sets.

Eventually, the histogram (Figure 7) shows dramatically incorrect results, with almost equal values for the MI between any subset of features and the permuted or the actual output; things are however better for the Delve Census data set.

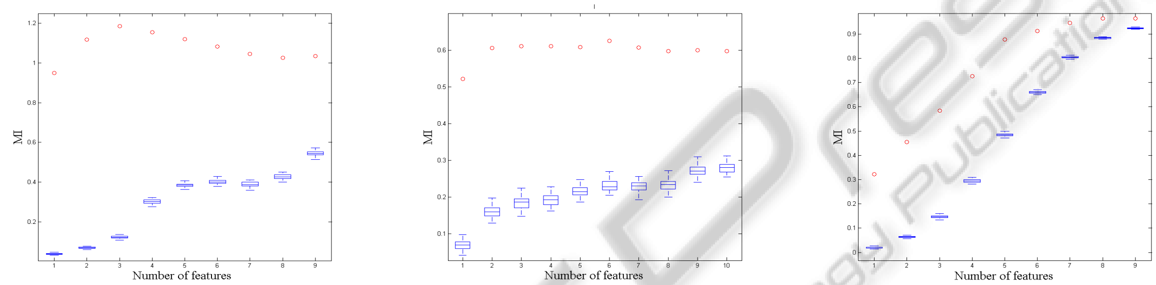
3.3 Choice of the Parameters

The last experiment is about the choice of the param-



(a) (b) (c)

Figure 4: Estimated MI between a group of features and the output (circles) and boxplots of the estimated MI between the same features and a permuted output for the adaptive histogram estimator: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.



(a) (b) (c)

Figure 5: Estimated MI between a group of features and the output (circles) and boxplots of the estimated MI between the same features and a permuted output for the kernel density estimator: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

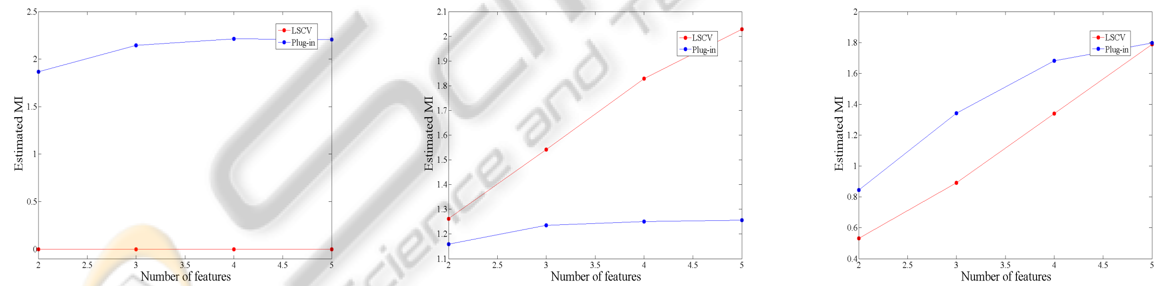


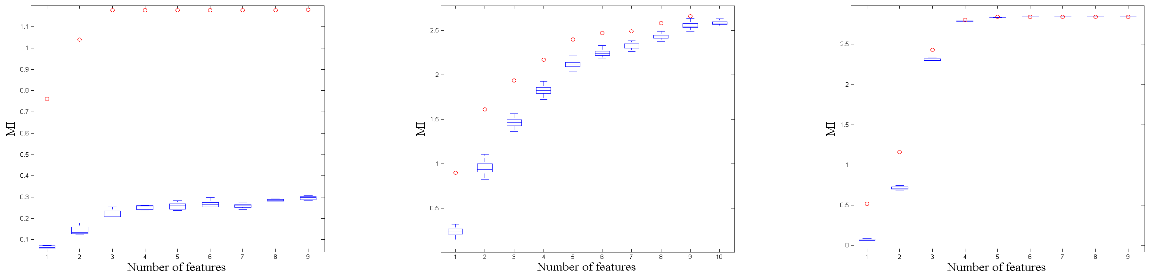
Figure 6: Estimated MI with the kernel density estimator for different values of the kernel width: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

eters in the estimators. As already mentioned, the basic histogram, the KDE, the B-splines approach and the NN-based estimator all have at least one parameter to fix, which can be fundamental for the quality of the estimations. Since the performances of the basic histogram in high-dimensional spaces are obviously dramatic, this estimator is not studied in more details.

To compare the different estimators, the same data sets are used as in the previous section and the MI estimations are shown for dimension 2 to 5. Once again this limitation is due to the time and space-consuming generalization of the B-splines approach in high-dimensional spaces. Moreover, the choice of the parameter is less related to feature selection.

3.3.1 The Kernel Density Estimator

For the KDE, the parameter to be fixed is the width of the kernel. As an alternative to the *rule of thumb* used so far (see Equation (6)), two other methods are considered. The first one is the very popular Least Squares Cross-Validation (LSCV) introduced



(a) (b) (c)

Figure 7: Estimated MI between a group of features and the output (circles) and boxplots of the estimated MI between the same features and a permuted output for the histogram based estimator: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

by Rudemo and Bowman (Bowman, 1984) (Rudemo, 1982) whose goal is to estimate the minimizer of the Integrated Square Error. The second one is the Plug-In method proposed by Hall, Sheater Jones and Marron (Hall et al., 1991). Figure 6 shows the extreme sensitivity of the KDE to the width of the kernel since the results obtained with both bandwidth determination strategies are totally different for the three data sets. Moreover, as illustrated in Figure 8 which shows the estimation of the MI for correlated Gaussians and $r = 0.9$, none of the methods used to set the kernel width clearly outperforms the other ones.

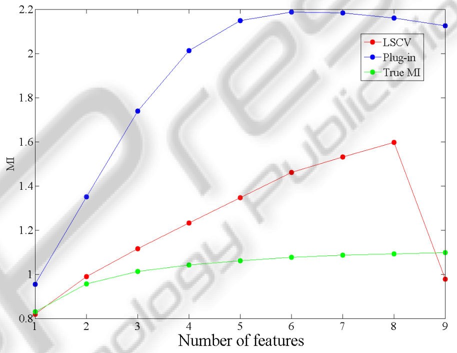


Figure 8: Estimated MI for correlated Gaussians with a kernel density estimator whose kernel's width has been determined by three different procedures.

3.3.2 The B-splines Estimator

Two parameters have to be determined in this approach: the degree of the splines and the number of bins. We fix the degree of the splines to three (as suggested in the original paper) and only focus on the number of bins per dimension as this parameter has been shown to influence much more the output (Daub et al., 2004); it will be taken between 2 and 5. Even if these values can seem surprisingly small, only three bins are used in (Daub et al., 2004).

The results presented in Figure 9 show that the estimated MI increases with the number of bins. These conclusions are consistent with those found in (Daub et al., 2004) for the one-dimensional case. However, even if the estimator is extremely sensitive to the number of bins, the relative values of the MI between the output and different groups of features is preserved, and so is the relative significance of the feature subsets. The sensitivity of the estimator is thus not a drawback for feature selection.

3.3.3 The Nearest Neighbors-based Algorithm

The only parameter to fix in the NN-based estimator is the number of neighbors K. Kraskov *et al.* suggest a value of 6, arguing it leads to a good trade-off between the variance and the bias of the estimator (Kraskov et al., 2004). Here, K is considered between 4 and 8.

Figure 10 shows very little sensitivity of the estimator in terms of absolute differences between estimations and thus a small sensitivity of the estimator to the number of neighbors used. However the results on the Delve data set indicate that even a small variation in the values of the estimated MI can lead to a different ranking of the features subsets in terms of relevance. As an example, in this data set, when using $K = 4$ or $K = 5$ neighbors, the subset of the five first features is less informative for the output than the subset of the four first features, while the opposite conclusion (which is in theory true) can be drawn when using 6, 7 or 8 neighbors. This is something that must be taken care of when performing feature selection because it could lead to the selection of irrelevant (or less relevant than other) features. One idea to overcome this issue is to average the estimations obtained within a reasonable range of values of K. In (Gomez-Verdejo et al., 2009), this principle is applied to feature selection using a version of the Kraskov estimator adapted for classification problems. Another

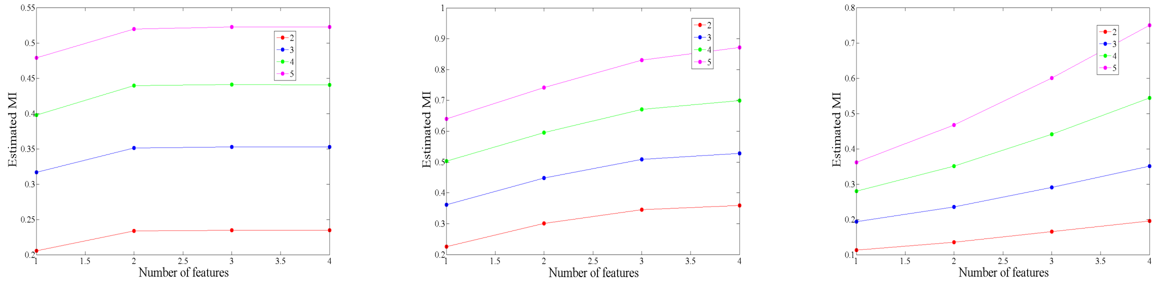


Figure 9: Estimated MI with the B-splines estimator for different values of bins per dimension: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

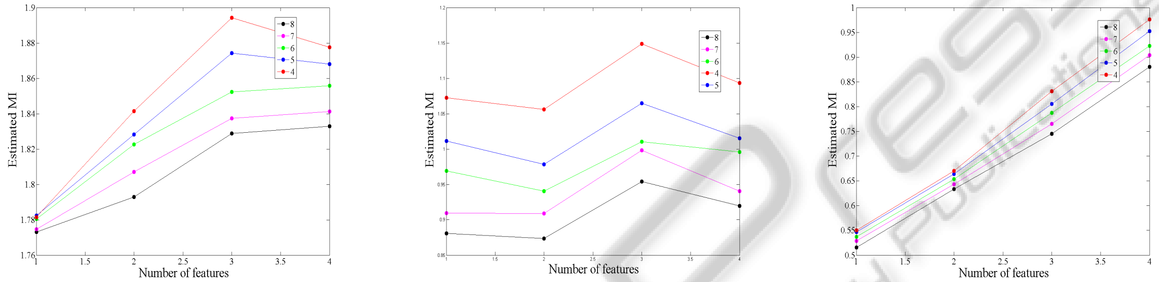


Figure 10: Estimated MI with the NN-based estimator for different values of the parameter k: (a) Delve dataset, (b) Nitrogen dataset , (c) Artificial dataset.

idea is to choose the value of K using the permutation test and resampling techniques (Francois et al., 2007).

4 CONCLUSIONS AND DISCUSSIONS

In this paper several popular approaches to the estimation of multi-dimensional MI are compared through three important criteria for feature selection: the accuracy, the consistency with an independence hypothesis and the sensitivity to the values of the parameter(s). The conclusion is the superiority of the NN-based algorithm which is by far the most accurate and the most consistent with an independent hypothesis (i.e. it returns values very close to 0 when estimating the MI between independent variables) on the three data sets used for comparison. The B-splines estimator presents interesting properties as well but can hardly be used when dimension becomes higher than 9 or 10, because of the exponential number of values to compute; the NN-based estimator is not affected by this major drawback, since it only requires the computation of the distances between each pair of points of the data set in the input, output and joint input-output spaces. By avoiding the hazardous evaluation of high-dimensional pdf, it is able to produce very robust results as the dimension of the data increases. It is also the less sensitive to the value of its single parameter, the number of neighbors K. However, as it has been seen, the choice of this parameter cannot be made at random since slight variations in the estimation of the MI can lead to a different ranking of the features subset relevance. Two approaches have been reported to deal with this issue, both producing satisfactory results. Being aware of all these facts, it thus appears to be a good choice to use the Kraskov estimator, or its counterpart for classification, to achieve MI-based multivariate filter feature selection.

ACKNOWLEDGEMENTS

G. Doquire is funded by a grant from the Belgian F.R.I.A.

REFERENCES

Bellman, R. E. (1961). *Adaptive control processes - A guided tour*. Princeton University Press, Princeton, New Jersey, U.S.A.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley-Interscience.

Darbellay, G. and Vajda, I. (1999). Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory*, 45(4):1315–1321.

Daub, C., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions - an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(1):118.

Francois, D., Rossi, F., Wertz, V., and Verleysen, M. (2007). Resampling methods for parameter-free and robust feature selection with mutual information. *Neurocomputing*, 70(7-9, Sp. Iss. SI):1276–1288.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67.

Gomez-Verdejo, V., Verleysen, M., and Fleury, J. (2009). Information-theoretic feature selection for functional data classification. *Neurocomputing*, 72:3580–3589.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Hall, P., Sheater, S. J., Jones, M. C., and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika*, 78(2):263–269.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Phys. Rev. E*, 69(6):066138.

Li, S., Mnatsakanov, R. M., and Andrew, M. E. (2011). k-nearest neighbor based consistent entropy estimation for hyperspherical distributions. *Entropy*, 13(3):650–667.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Rossi, F., Delannay, N., Conan-Guez, B., and Verleysen, M. (2005). Representation of functional data in neural networks. *Neurocomputing*, 64:183 – 210.

Rossi, F., Lendasse, A., François, D., Wertz, V., and Verleysen, M. (2007). Mutual information for the selection of relevant variables in spectrometric nonlinear modelling. *CoRR*.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66(3):605–610.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423,623–656.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman and Hall, London.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the American Statistical Association*, 21(153):65–66.

Turlach, B. A. (1993). Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*, pages 23–493.

Walters-Williams, J. and Li, Y. (2009). Estimation of mutual information: A survey. In *RSKT '09*, pages 389–396, Berlin, Heidelberg. Springer-Verlag.

Wang, Q., Kulkarni, S. R., and Verdu, S. (2009). Divergence Estimation for Multidimensional Densities Via k-Nearest Neighbor Distances. *Information Theory, IEEE Transactions on*, 55(5):2392–2405.