Applied Mathematical Sciences, Vol. 4, 2010, no. 29, 1417 - 1436

The Performance of Mutual Information for Mixture of Bivariate Normal Distributions Based on Robust Kernel Estimation

Kourosh Dadkhah[1] **and Habshah Midi**

Laboratory of Applied and Computational Statistics
Institute for Mathematical Research
University Putra Malaysia, Serdang, Malaysia

Olimjon Sh. Sharipov

Institute of Mathematics and Information Technologies
Uzbek Academy of Science, Tashkent, Uzbakestan

Abstract

Mutual Information (MI) measures the degree of association between variables in nonlinear model as well as linear models. It can also be used to measure the dependency between variables in mixture distribution. The MI is estimated based on the estimated values of the joint density function and the marginal density functions of X and Y. A variety of methods for the estimation of the density function have been recommended. In this paper, we only considered the kernel method to estimate the density function. However, the classical kernel density estimator is not reliable when dealing with mixture density functions which prone to create two distant groups in the data. In this situation a robust kernel density estimator is proposed to acquire a more efficient MI estimate in mixture distribution. The performance of the robust MI is investigated extensively by Monte Carlo simulations. The results of the study offer substantial improvement over the existing techniques.

Mathematics Subject Classification: 62G07, 62G35, 94A15, 94A15

Keywords: Mutual Information, Kernel Density, Minimum Volume Ellipsoid, Minimum Covariance Determinant, Outliers, Mixture Distribution, Robust Statistics

[1]Corresponding author: kdadkhah@yahoo.com

1 Introduction

The coefficient of determination, ρ^2 is a very popular measure of association between two variables because of tradition and ease of computation. However, ρ^2 only measures the degree of linear association and cannot be applied to nonlinear model. In real situation, many fitted models are not always linear and the use of ρ^2 may produce a misleading conclusion. An alternative approach is to use another measure of dependency between two variables which is called mutual information (MI). The attractive feature of the MI is that, it is capable of measuring the dependency of both linear and nonlinear relationships, making it more appropriate for use with complex nonlinear model. Most studies which use MI as a dependency measure only focused on a single distribution of random variables. In practice, we may encounter with mixture distributions, which incline to produce two distant groups in the data, particularly in some areas such as clustering analysis, data mining, machine learning etc. We usually consider the smaller of the two groups as outliers. Outliers are observations which are markedly different from the bulk of the data or from the pattern set by the majority of the observations. In this situation, by considering the estimates of MI on a single distribution instead of the mixture distributions may produce less efficient and misleading results.

This paper will focus only on mixture of two bivariate normal distributions. Several methods have been proposed for the MI estimation, such as kernel density estimators (KDE) [10], k-nearest neighbours (KNN) [7], Edgeworth approximation of differential entropy [5], MI carried by the rank sequences [17], and adaptive partitioning of the XY plane [1]. KDE is a well known method which is used extensively to estimate the MI. Accurate estimation of MI depends heavily on the precise estimate of density functions. Moon [10] stated that the ordinary KDE is not suitable in mixture distributions. This may be due to the composition of the observations in which their observations were usually separated into two groups with different locations and scales. In this situation, the computation of the covariance matrix of the probability function will be affected. It is now evident that the classical mean and classical standard deviation are easily affected by outliers. Outliers are the leading cause of bias estimation of parameters in the models [13]. In this respect, we have to incorporate robust statistical techniques in the computation of the KDE to estimate mutual information. Kim and Scott [6] utilized the iteratively reweighted least square (IRWLS) algorithm for M-estimation. They replaced the centroid with a robust M-estimate.

It is worth mentioning that the computation of the estimated KDE is based on the classical covariance matrix, S. As already been mentioned, the classical S is not robust and may have an unduly effect on the estimated KDE for a mixture distribution. Consequently, the estimated MI is not robust. In this

paper we propose to replace the classical S with the scale estimate acquired from robust estimates such as the Minimum Volume Ellipsoid (MVE) and the Minimum Covariance Determinant (MCD). This paper is organized as follows. Section 2 discusses the MI and mixture distribution. The development of the robust MI is exhibited in section 3. Section 4 reports the results of the simulation study. The conclusion is summarised in section 5.

2 Mutual Information and Mixture Distribution

Shannon [15] defined mutual information as a measure dependency of the bivariate random variables (X,Y), and has the form

$$I(X;Y) = \int\int f_{X,Y}(x,y) log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} dxdy. \quad (1)$$

where $f_{X,Y}(x,y)$ is the joint probability density function (pdf) of X and Y, and $f_X(x)$, $f_Y(y)$ are the marginal pdf of X and Y, respectively. The units of information of $I(X;Y)$ depend on the base of the logarithm, e.g. bits for the base of 2 and nats for the natural logarithm. MI is positive and symmetrical that is $I(X;Y) = I(Y;X)$. It is also invariant under one-to-one transformations that is, $I(X;Y) = I(U;V)$ where $u = f(x), v = f(y)$, and f is invertible. If X and Y are independent, the joint pdf is equal to the product of marginal pdfs leading to $I(X;Y) = 0$. If there is perfect dependence between X and Y, MI approaches infinity. It is always $I(X;Y) \leq H(X)$, where $H(X) = -\int f(x) log f(x) dx$.
MI between random variables X and Y can also be defined in terms of information entropies as [2]

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y).$$

where H(X) and H(Y) are called the marginal information entropies which measure the information content in X and Y, respectively, $H(Y|X)$ is the entropy of Y conditional on X which measures the information content remaining in Y if the information content in X is known completely, and H(X,Y) is the joint information entropy which measures the information content in a pair of random variables X and Y. The bivariate case is considered here for simplicity. The linear correlation coefficient between two variables X and Y denoted by $\rho(X,Y)$ is a measure of the strength of the linear dependence between the variables and varies from 0 to 1. The estimation of the most likely value and the corresponding uncertainties is relatively straightforward. However, the estimation of the mean and uncertainty bounds, for a MI-based

dependence measure that is normalized to scale between 0 to 1, is an area of ongoing research. If (X,Y) is bivariate normal, the MI and linear correlation coefficient are related as $I(X;Y) = -0.5log[1-\rho^2(X,Y)]$ [2].

Given a set of probability density functions $f_i(z)$ specified as $f_i(z;\theta_i)$ (called the mixture component), where θ_i; is the vector of unknown parameters in the postulated form for the ith component density in the mixture and weights $\alpha_1, \cdots, \alpha_l$ such that $\alpha_i \geq 0$ and $\sum \alpha_i = 1$, the sum (which is convex combination):

$$f(z;\psi) = \sum_{i=1}^{l} \alpha_i f_i(z;\theta_i).$$

is called the mixture density [9]. The vector ψ containing all the unknown parameters in the mixture model can be written as

$$\psi = (\alpha_1, \cdots, \alpha_{l-1}, \xi^T)^T \tag{2}$$

where ξ the vector containing all the parameters in $\theta_1, \cdots, \theta_l$ known a priori to be distinct. Since the mixing proportions α_i sum to unity, one of them is redundant. In defining ψ in Equation(2) , the lth mixing proportion α_l is omitted, arbitrarily. In practice, the components are often taken to belong to the normal family, leading to normal mixtures. In the case of multivariate normal components, the $f_i(z;\theta_i)$ is defined as

$$f_i(z;\theta_i) = \phi(z;\mu_i,\Sigma_i).$$

with mean (vector) μ_i and covariance matrix Σ_i $(i = 1, \cdots, l)$. This type of mixture, being a finite sum, is called a finite mixture, and in applications, an unqualified reference to a "mixture density" usually means a finite mixture. The mixture components are often not arbitrary probability density functions, but instead are members of a parametric family (such as normal distributions), with different values for a parameter or parameters [9].

To illustrate some property of MI in mixture models, we consider a mixture of two bivariate normal components with covariance matrixes Σ_1 and Σ_2 and means μ_1 and μ_2 with proportion $(1-\alpha)$ and α respectively, so that:

$$f_Z(z) = (1-\alpha)\phi(z;\mu_1,\Sigma_1) + \alpha\phi(z;\mu_2,\Sigma_2), \tag{3}$$

where $Z = (X,Y)$, $\mu_1 = (\mu_1^{(1)}, \mu_2^{(1)})$, $\mu_2 = (\mu_1^{(2)}, \mu_2^{(2)})$,

$$\Sigma_1 = \begin{pmatrix} \sigma_{11}^{(1)} & \sigma_{12}^{(1)} \\ \sigma_{21}^{(1)} & \sigma_{22}^{(1)} \end{pmatrix}, \Sigma_2 = \begin{pmatrix} \sigma_{11}^{(2)} & \sigma_{12}^{(2)} \\ \sigma_{21}^{(2)} & \sigma_{22}^{(2)} \end{pmatrix},$$

$$f_X(x) = (1-\alpha)\phi(x;\mu_1^{(1)},\sigma_{11}^{(1)}) + \alpha\phi(z;\mu_1^{(2)},\sigma_{11}^{(2)}), \tag{4}$$

$$f_Y(y) = (1-\alpha)\phi(x;\mu_2^{(1)},\sigma_{22}^{(1)}) + \alpha\phi(z;\mu_2^{(2)},\sigma_{22}^{(2)}). \tag{5}$$

Then the MI is defined as

$$I(X;Y) = \int\int f_{X,Y}(x,y) log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)} dxdy.$$

Since the closed form of the true MI is not easy to obtain theoretically, a Monte Carlo simulation will be conducted to estimate the true MI. The true MI can be approximated by

$$I(X;Y) \approx \frac{1}{n}\sum_{i=1}^{n} log \frac{f_{X,Y}(x_i,y_i)}{f_X(x_i)f_Y(y_i)}. \tag{6}$$

where the (x_i, y_i) are independent observations from a bivariate normal mixture. This is because the expected value of $E(\frac{1}{n}\sum_{i=1}^{n} log\frac{f_{X,Y}(x_i,y_i)}{f_X(x_i)f_Y(y_i)}) = I(X,Y)$. In this respect the precision of the estimated MI increases with the increase in n. Here n is the number of replications in the simulation study.

The MI in Equation(1) for any bivariate data set (X,Y) of size n can be estimated as

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} log \frac{\hat{f}_{X,Y}(x_i,y_i)}{\hat{f}_X(x_i)\hat{f}_Y(y_i)}. \tag{7}$$

where $\hat{f}_{X,Y}(x_i,y_i)$ is the estimated joint pdf and $\hat{f}_X(x_i)$ and $\hat{f}_Y(y_i)$ are the estimated marginal pdfs at (x_i, y_i).

For the bivariate data set $(z_1, \cdots, z_n)$, where each z is in d-dimensional space, the multivariate kernel density with kernel K is defined by

$$\hat{f}(z) = \frac{1}{n\lambda^d}\sum_{i=1}^{n} K(\frac{z-z_i}{\lambda}). \tag{8}$$

where λ is the smoothing parameter [16]. We choose the standard multivariate normal kernel's defined by

$$K(z) = \frac{1}{2\pi^{\frac{\lambda}{2}}} exp(-\frac{1}{2}z^T z). \tag{9}$$

Using Equations(8) and (9), the probability function is defined as

$$\hat{f}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}\lambda^d |S|^{\frac{1}{2}}} exp(-\frac{(z-z_i)^T S^{-1}(z-z_i)}{2\lambda^2}).$$

where S is the covariance matrix and $|S|$ is the determinant of S. For a normal kernel, Silverman [16] suggested an optimal smoothing parameter or Gaussian bandwidths in d-dimension given as

$$\lambda_{ref} = (\frac{4}{d+2})^{1/(d+4)} n^{(-1/(d+4))}.$$

where for bivariate (d=2), $\lambda_{ref} = n^{(-1/6)}$.Harrold et al. [4] recommended using smoothing parameter in estimating the MI. Their study showed that $\alpha = 1.5$ is a good scaling factor as $\lambda = \alpha\lambda_{ref}$ for sample of size 200 or more. For sample of size less than 200, a value of $\alpha = (1.8 - r(1))$ should be used, where r(1) is the sample estimate of the autocorrelation at lag 1.

3 Robust Estimation of Mutual Information

Robust estimation of multivariate location and covariance are the key to various statistical techniques. The simple idea is to substitute the location and the covariance with robust version of them. We employed this idea to develop a robust estimate of density in the MI function. Kernel function uses two parameters, namely the bandwidths and the covariance matrix. Rousseeuw [11, 12] introduced an affine equivariant estimator with maximal breakdown point, by putting:

T(Z) = center of the minimal volume ellipsoid covering (at least) h points of Z.

Let $Z = z_1, \cdots, z_n$ be a set of point in R^d for some constant d. Let $z_i = (z_{i1}, \cdots, z_{id})^T$. The (empirical) covariance matrix C(Z) of Z is the (d by d) matrix defined by

$$C(Z) = \frac{1}{h}\sum_{i=1}^{h} z_i z_i^T - T(Z)T^T(Z).$$

Where $T(Z) = \frac{1}{h}\sum_{i=1}^{h} z_i$ and h can be taken equal to [n/2]+1. This is called the minimum volume ellipsoid estimator (MVE). Another approach would be to generalize the least trimmed squares (LTS) estimator to multivariate location. This yields:

T(Z) = Mean of the h points of Z for which the determinant of the covariance matrix is minimal

They called this estimator as the minimum covariance determinant estimator (MCD). It corresponds to finding the h points for which the classical tolerance ellipsoid (for a given level) has minimum volume, and then taking its center. This estimator is also affine equivariant. The MCD has the same breakdown point as the MVE.

Rousseeuw and Van [14] improved the MCD algorithms to get a fast MCD. Both Davies [3] and Lopuha and Rousseeuw [8] independently found a way to increase the finite-sample breakdown point of the MVE to its best possible value. As in the case of the least median square (LMS) regression estimator, this is done by letting h to depend on the dimensionality. For the MVE and the MCD, the optimal variant is to consider "halves" containing

$$h = [(n + d + 1)/2].$$

observations. Then the resulting breakdown point is

$$\epsilon_n^* = \frac{[(n - d + 1)/2]}{n}.$$

It is important to point out that the classical MI is estimated by Equation(7) based on the classical KDN. As already been mentioned, the covariance matrix in the formulation of the probability function in the classical KDN is easily affected by the presence of outliers. A better approach is to consider a robust statistical approach. In order to develop a robust MI; we propose to replace the classical covariance matrix in the kernel density function with the robust covariance matrix which was acquired from the MVE and the MCD. The estimated robust MI is written as follows.

$$\hat{I}^R(X;Y) = \frac{1}{n}\sum_{i=1}^{n} log\frac{\hat{f^R}_{X,Y}(x_i, y_i)}{\hat{f^R}_X(x_i)\hat{f^R}_Y(y_i)}.$$

where

$$\hat{f^R}_Z(z) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{(2\pi)^{\frac{d}{2}}\lambda^d |S_{MCD}|^{\frac{1}{2}}}.exp(-\frac{(z - z_i)^T S_{MCD}^{-1}(z - z_i)}{2\lambda^2}). \tag{10}$$

4 Simulation Study

In this section, we report a Monte Carlo simulation study that is designed to illustrate the robustness of the estimators already discussed. Our main focus is to develop a reliable alternative robust approach to estimate the MI in mixture of two bivariate normal distributions in the situations where both distributions are distant apart. We assess the performance of each estimator by using the estimated errors. The error of each estimator is obtained by subtracting the estimated MI from the true MI. Since the theoretical MI is not possible, a simulation study is conducted to obtain the approximated true value. In this experiment, we consider a mixture of two bivariate normal distributions, each having different location and different correlation coefficient. The values of α

are varied from 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.9, 1.0. First, we create a mixture model with the joint and marginal probability density of the x and y as presented in (3), (4) and (5), respectively.

The first $100(1-\alpha)$ percent of observations are from the first distribution and the remaining 100α percent from the second distribution. The correlation coefficient of the first and the second bivariate normal distributions are fixed at $\rho_1 = 0$ and $\rho_2 = 0.99$, respectively. We consider 4 mixture models:

Model 1- Let $\mu_1 = (0,0),\ \mu_2 = (0,0),\ \Sigma_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$.
Model 2- Let $\mu_1 = (0,0),\ \mu_2 = (20,0),\ \Sigma_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$.
Model 3- Let $\mu_1 = (0,0),\ \mu_2 = (0,20),\ \Sigma_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$.
Model 4- Let $\mu_1 = (0,0),\ \mu_2 = (20,20),\ \Sigma_1 = \begin{pmatrix} 1 & \rho_1 \\ \rho_1 & 1 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & \rho_2 \\ \rho_2 & 1 \end{pmatrix}$.

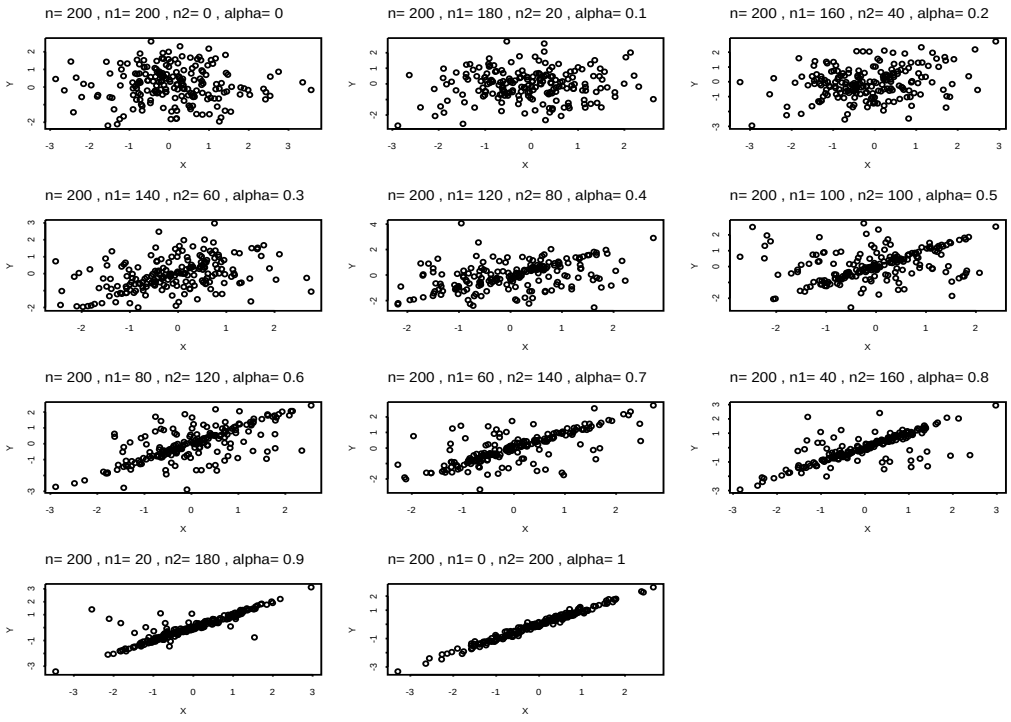

Figure 1: The scatter plot of y versus x at different α values, model 1.

For better understanding, the construction of the approximate MI true value for the first mixture model is illustrated by Monte Carlo simulation. Similarly, the estimated true MI for the other mixture models will follow the same procedure. Let us now consider the first mixture model and refer to the definition of the approximated true MI as shown in Equation(6).

It can be observed that in order to estimate the true MI, the joint and

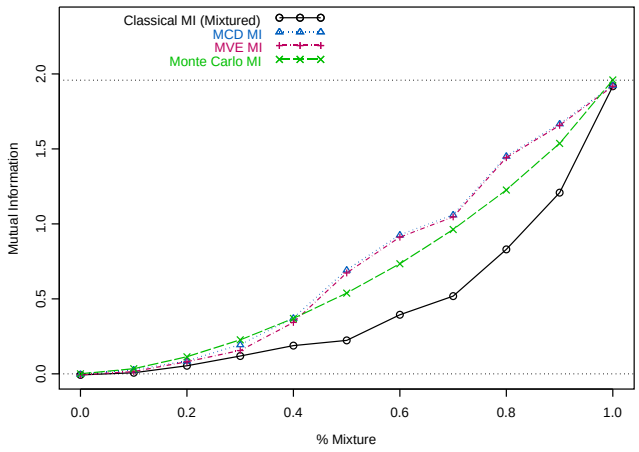

Figure 2: Graph of MI, MI(MVE), MI(MCD) at different values of α, model 1.

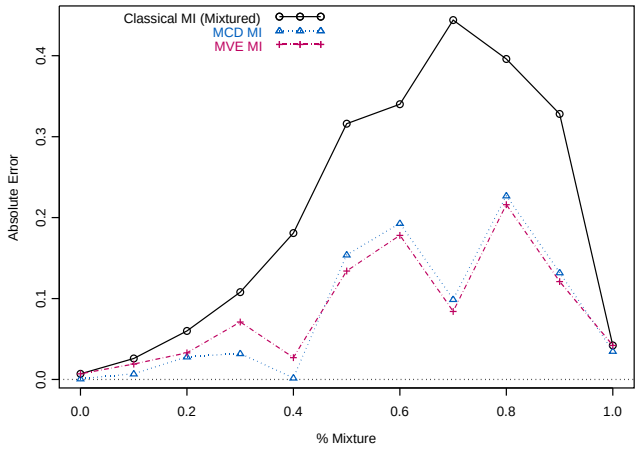

Figure 3: Plot of Absolute Error of MI, MI(MVE), MI(MCD) at different α, model 1.

the marginal probability density function of x and y need to be computed. For this purpose, we firstly generate the first $100(1-\alpha)$ percent of the bivariate normal observations x_1 and y_1, and the remaining 100α percent from the second bivariate normal observations x_2 and y_2 according to the parameters as specified in the first model with a sample of size 200. At α equals to 0.1, implies that the mixture model function in Equation(3) has 99 percents of the observations which come from the first and 1 percent of the observation from the second distribution. The value of MI is then computed using Equation(6) by substituting the true value of the joint and marginal probability density of the first and the second distribution as presented in Equation(3), (4) and (5), respectively. The above experiment is repeated for 10^6 times to obtain the estimated true value of MI. This implies that in each simulation run, there were 10^6 replications or $n = 10^6$. The estimated classical MI, the MI (MCD) and the MI (MVE) are applied to the generated data and their computations are based on the estimated KDN. When α equals to zero, indicates that all the observations comes from the first distribution while for α equals to one, all data comes from the second distribution. The scatter plots of y against x at different values of α, with fixed values of $\rho_1 = 0$ and $\rho_2 = 0.99$ are shown in Figures 1, 4, 7 and 10. The plots of the estimated MI, MI(MVE) and MI(MCD) at various percentage of mixtures are illustrated in Figures 2, 5, 8 and 11 while the estimated absolute errors of MI, MI(MVE) and MI(MCD) are displayed in Figures 3, 6, 9 and 12. The estimated values of MI, MI(MVE) and MI(MCD) with their respective errors for Mixture Model 1, 2, 3, 4 are exhibited in Tables 1, 2, 3, and 4, respectively.

Table 1: The Estimated MI, MI(MVE), MI(MCD) and their respective errors for Model 1

α	MI	$MI(MCD)$	$MI(MVE)$	$MI(Monte\ Carlo)$	$MI\ Error$	$MI(MCD)\ Error$	$MI(MVE)\ Error$
0	-0.007	-0.001	-0.007	0.000	-0.007	-0.001	-0.007
0.1	0.008	0.027	0.015	0.034	-0.026	-0.007	-0.020
0.2	0.054	0.086	0.081	0.114	-0.059	-0.028	-0.032
0.3	0.119	0.195	0.156	0.227	-0.107	-0.032	-0.070
0.4	0.189	0.372	0.343	0.370	-0.180	0.002	-0.027
0.5	0.223	0.693	0.673	0.539	-0.316	0.154	0.135
0.6	0.394	0.927	0.912	0.734	-0.340	0.193	0.178
0.7	0.519	1.062	1.047	0.963	-0.444	0.098	0.083
0.8	0.830	1.453	1.442	1.226	-0.396	0.227	0.216
0.9	1.209	1.669	1.658	1.537	-0.329	0.132	0.121
1	1.918	1.925	1.918	1.960	-0.042	-0.035	-0.042

Let us first focus to model 1 where there are two bivariate normal distributions, both having the same means. The scatter plots of the variable x versus y as shown in Figure 1 demonstrate that as the percentage of the second model increases (α) the observations x and y get more correlated. These results are as expected because the second bivariate normal distribution has higher value of $\rho_2 = 0.99$. In this situation, we also expect that the estimated MI will be in-

creased with the increased in α. It can be observed from Table 1 and Figure 2 that both MI(MVE) and MI(MCD) give closer estimates to the approximated true value of MI (true MI is approximated by Monte Carlo), compared to the classical MI. The classical MI reveals an underestimate while the MI(MCD) and the MI(MVE) show an overestimate values of the true MI. However, the absolute errors for both robust MIs are smaller than the classical MI as illustrated in Figure 3 and Table 1. For example, in the case where $\alpha = 0.9$ it suggests that the second distribution is mixed with 10% bivariate normal noise. In this situation, the estimated absolute error of the classical MI is equal to 0.329 which is larger than the absolute errors of the robust MI(MCD) and the MI(MVE) which equal to 0.132 and 0.121, respectively. The absolute errors of the MI(MCD) and MI(MVE) are constantly smaller than the absolute error of the classical MI. When $\alpha = 0$ or $\alpha = 1$, it suggests that the mixture distribution consists of only one distribution with no contamination in the data. It is very obvious that the dependency is almost zero when $\alpha = 0$. Nonetheless the results of Table 1 give misleading information because the MIs values are very small and negative.

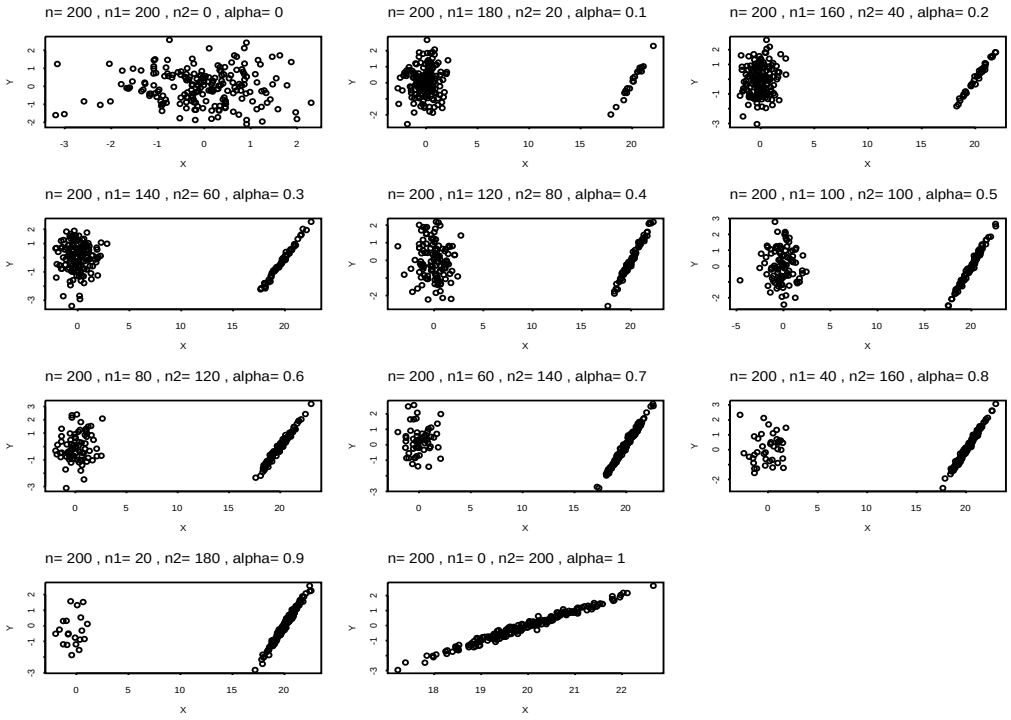

Figure 4: The scatter plot of y versus x at different α values, model 2.

Next, we consider Model 2 which is a mixture distribution of two bivariate normal distributions with their x-values are far from each other. The first

bivariate normal distribution having means $(0, 0)$ and the second distribution having means $(20, 0)$. From Figure 4, we can see that as the percentage of the second distribution increases the dependency also increases because the data were generated such that the association exists only on the second distribution. However, the results of Table 2 and Figure 5 show that the estimated MI gives misleading conclusions. As the α increases not only the values of classical MI get very small also the signs changed to negative which is not as expected. As mentioned earlier, it is important to mention here that the mixture of these two distributions produced two distant groups as illustrated in Figure 4. We may consider the smaller group as x-outliers for α values which are closer to 0.8.

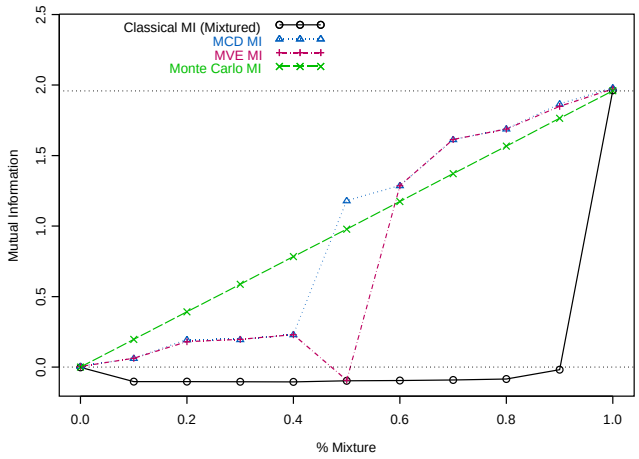

Figure 5: Graph of MI, MI(MVE), MI(MCD) at different values of α, model 2.

It can be observed from Table 2 that the MI produces misleading results in the situation where apparent outliers are present except when α equals 1.0 (strong dependency and no outlier). For example, when $\alpha = 0.9$ (apparent outliers), the MI value equals to -0.023 while the MI(MCD) and MI(MVE) equal to 1.869 and 1.858, respectively. From these results it seems that the classical MI is very sensitive to the presence of outliers and in their presence would give a misleading conclusion. The MI(MCD) and MI(MVE) seem to be only slightly affected by the outlier observations, as exemplified in Table 2 and Figure 6 where their absolute errors are smaller than the MI errors. In this situation, the performance of the MI(MCD) and MI(MVE) are equally good except when $\alpha = 0.5$ (the number of observations in the two distributions are equal). It is worth mentioning here that the value of the MI(MVE) immediately changed to small and negative value for $\alpha = 0.5$. This happened

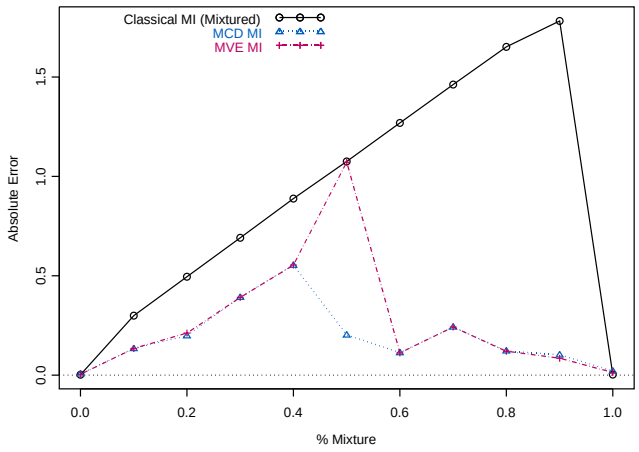

Figure 6: Plot of Absolute Error of MI, MI(MVE), MI(MCD) at different α, model 2.

as the consequence of the MVE algorithm whereby when α equals to 0.5, the algorithm tends to consider both groups. In the situation where the whole data were contaminated by small percentage of outliers, the MVE attempts to estimate the location and scale of the distribution for the majority of the observations (larger group) and pay less attention to the smaller group. Therefore when α less than 0.5, the MVE algorithm considers the first group that is less correlated and when α greater than 0.5, the algorithm considers the second group which is more correlated than the first group. In the situation where perfect association exist and no outlier in the data ($\alpha = 1$), all three methods are virtually indistinguishable with respect to the values of the errors.

Table 2: The Estimated MI, MI(MVE), MI(MCD) and their respective errors for Model 2

α	MI	$MI(MCD)$	$MI(MVE)$	$MI(Monte\ Carlo)$	$MI\ Error$	$MI(MCD)\ Error$	$MI(MVE)\ Error$
0	-0.006	0.001	-0.001	0.000	-0.006	0.001	-0.001
0.1	-0.102	0.058	0.048	0.196	-0.298	-0.138	-0.148
0.2	-0.104	0.119	0.120	0.392	-0.496	-0.273	-0.272
0.3	-0.092	0.251	0.251	0.587	-0.680	-0.337	-0.337
0.4	-0.106	0.228	0.229	0.784	-0.890	-0.556	-0.555
0.5	-0.102	1.112	-0.099	0.980	-1.082	0.132	-1.079
0.6	-0.093	1.357	1.357	1.176	-1.269	0.181	0.181
0.7	-0.083	1.698	1.696	1.371	-1.454	0.327	0.325
0.8	-0.075	1.845	1.846	1.566	-1.641	0.279	0.280
0.9	-0.023	1.869	1.858	1.759	-1.781	0.110	0.100
1	1.997	2.002	1.994	1.959	0.037	0.043	0.035

The scatter plots for Model 3 are illustrated in Figure 7 which shows that there are two bivariate normal distributions with their y-values are distant apart. Similar results can be drawn from Table 3, Figure 8 and Figure 9 as

is obtained for Model 2 where the observations are extreme in the x-values. The robust MIs repeat its excellence performances where it produces reliable approximations of the true MI when the two groups exist in the extreme y values. Similarly, the classical MI is not reliable as it overestimated the true MI with large errors.

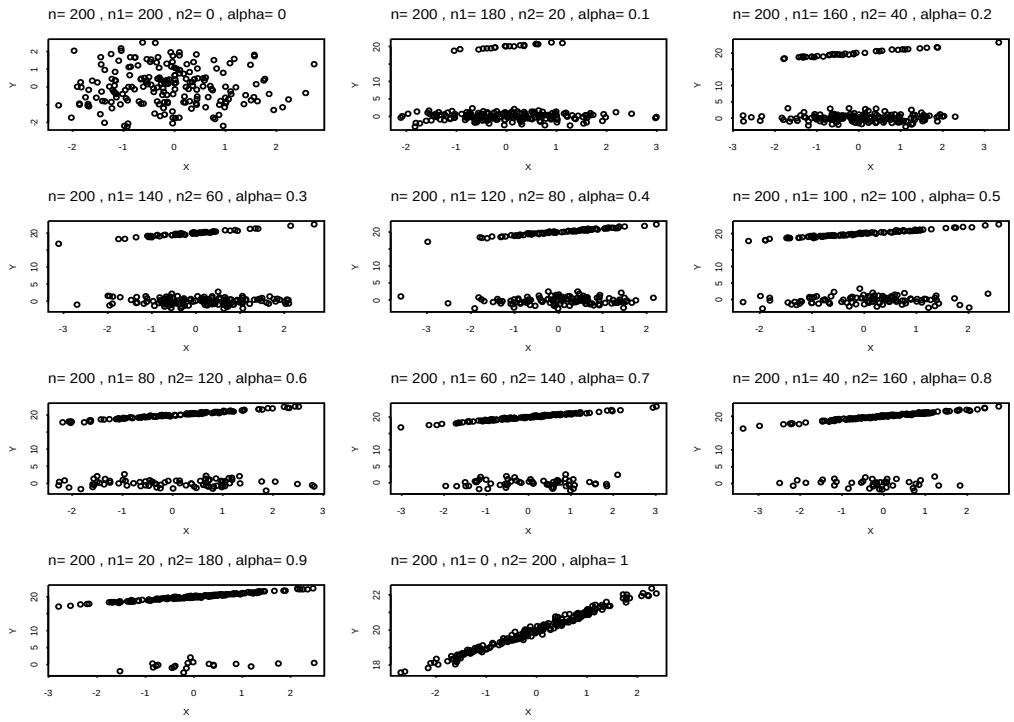

Figure 7: The scatter plot of y versus x at different α values, model 3.

According to model 4, two bivariate normal distributions are combined such that both their x-values and y-values are distanced apart. The pattern of this two mixture distribution can be seen in Figure 10. Just like the preceding models, as α increases, the number of observations in the uncorrelated group decreases and the number of observations in the correlated group increases. Therefore the dependency increases as well, as shown in Figure 10 and Table 4. This results can be confirmed from the values of MI, MI(MCD),MI(MVE), which were increasing as α increases. It is interesting to note about the values of MI of the three methods. Even though there exist two apparent groups, the classical MI is not much affected by their presence. These may happened perhaps due to the positions of the two groups which are extreme in both x and y directions. However, the robust MIs are slightly better than the classical MI as evidenced by their smaller errors as indicated in Table 4 and Figures 11 and 12.

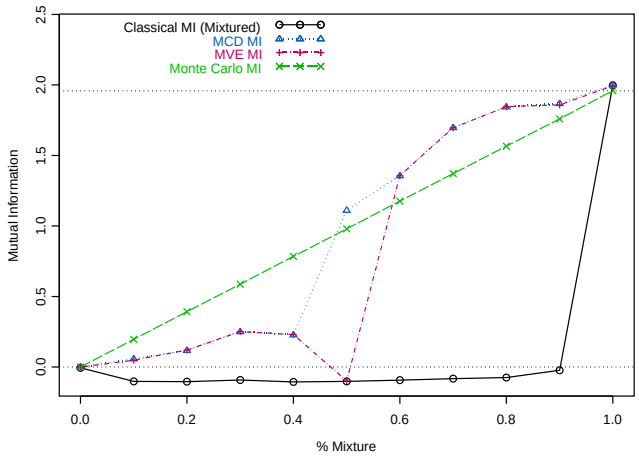

Figure 8: Graph of MI, MI(MVE), MI(MCD) at different values of α, model 3.

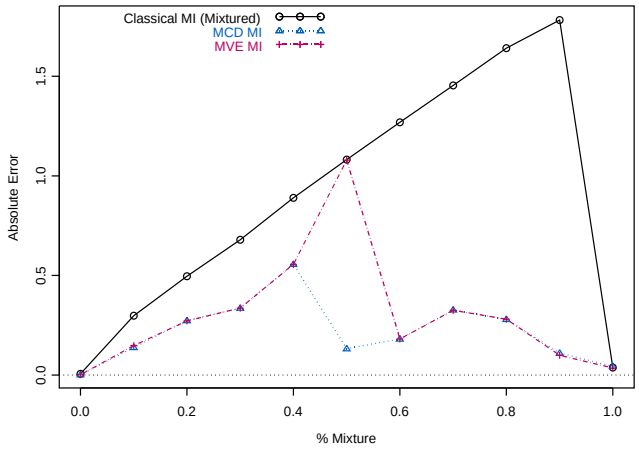

Figure 9: Plot of Absolute Error of MI, MI(MVE), MI(MCD) at different α, model 3.

Table 3: The Estimated MI, MI(MVE), MI(MCD) and their respective errors for model 3

α	MI	$MI(MCD)$	$MI(MVE)$	$MI(Monte\ Carlo)$	$MI\ Error$	$MI(MCD)\ Error$	$MI(MVE)\ Error$
0	-0.002	0.006	0.004	0.000	-0.002	0.006	0.004
0.1	-0.103	0.062	0.062	0.196	-0.299	-0.134	-0.134
0.2	-0.103	0.194	0.180	0.392	-0.495	-0.198	-0.212
0.3	-0.104	0.195	0.196	0.587	-0.691	-0.392	-0.391
0.4	-0.105	0.230	0.229	0.783	-0.889	-0.554	-0.554
0.5	-0.097	1.180	-0.093	0.978	-1.075	0.202	-1.071
0.6	-0.095	1.287	1.286	1.174	-1.269	0.113	0.111
0.7	-0.091	1.613	1.613	1.371	-1.462	0.242	0.243
0.8	-0.085	1.689	1.687	1.567	-1.652	0.122	0.119
0.9	-0.018	1.866	1.849	1.764	-1.782	0.102	0.085
1	1.962	1.980	1.972	1.959	0.002	0.021	0.012

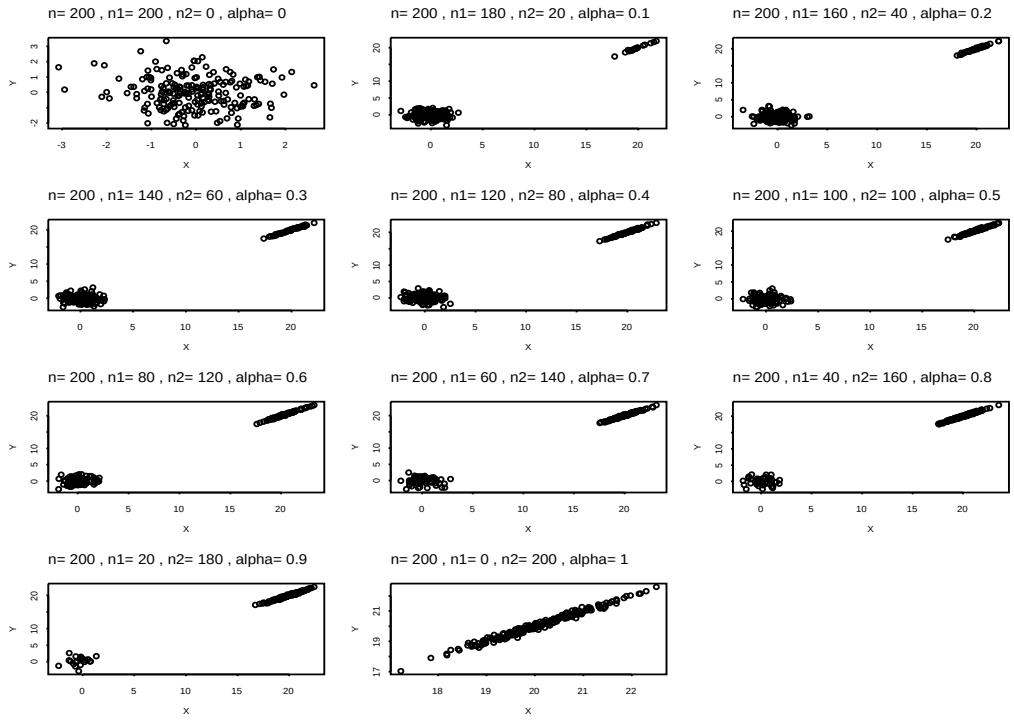

Figure 10: The scatter plot of y versus x at different α values, model 4.

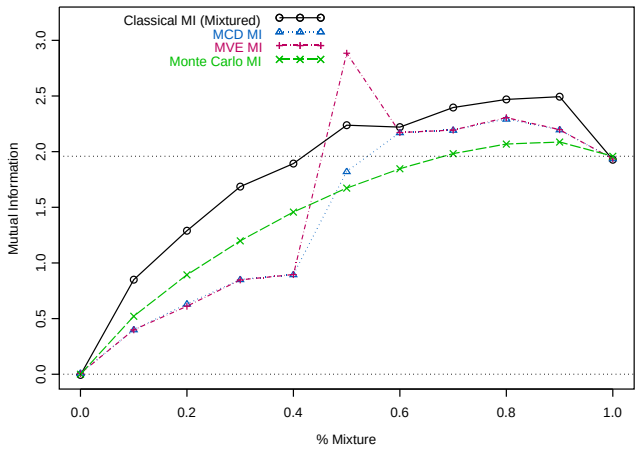

Figure 11: Graph of MI, MI(MVE), MI(MCD) at different values of α, model 4.

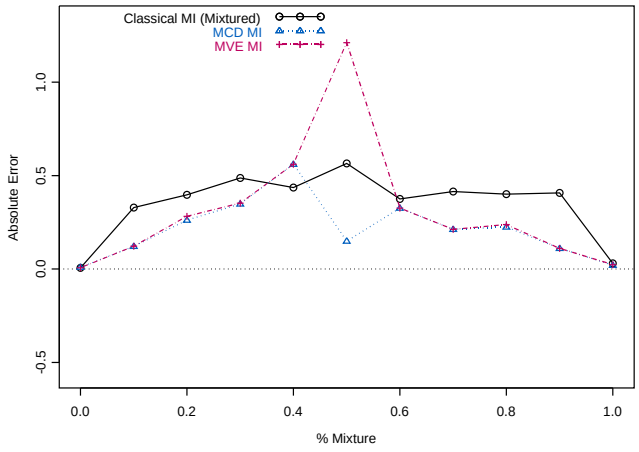

Figure 12: Plot of Absolute Error of MI, MI(MVE), MI(MCD) at different α, Model 4.

Table 4: The Estimated MI, MI(MVE), MI(MCD) and their respective errors for Model 4

α	MI	$MI(MCD)$	$MI(MVE)$	$MI(Monte\ Carlo)$	$MI\ Error$	$MI(MCD)\ Error$	$MI(MVE)\ Error$
0	-0.006	0.010	0.007	0.000	-0.006	0.010	0.007
0.1	0.850	0.400	0.399	0.521	0.328	-0.121	-0.123
0.2	1.290	0.632	0.611	0.893	0.397	-0.261	-0.282
0.3	1.686	0.851	0.847	1.199	0.487	-0.348	-0.352
0.4	1.893	0.896	0.896	1.457	0.436	-0.561	-0.561
0.5	2.238	1.822	2.884	1.673	0.565	0.149	1.211
0.6	2.221	2.172	2.172	1.846	0.375	0.326	0.326
0.7	2.397	2.194	2.194	1.982	0.416	0.212	0.212
0.8	2.469	2.293	2.306	2.068	0.401	0.225	0.238
0.9	2.494	2.197	2.197	2.087	0.407	0.111	0.110
1	1.928	1.938	1.935	1.958	-0.030	-0.020	-0.023

5 Conclusion

Outliers cause major interpretative problems in the classical MI, such as wrong sign problems, produce no dependency between variables, where in fact dependency exists. Thus it is very crucial to investigate the behaviour of the classical MI in the presence of outliers, to avoid making incorrect inferences. The main focus of this paper is to investigate the performance of the classical MI in four mixture distributions in the situations where the two bivariate normal distributions have different location parameters and different correlation coefficient. We also wanted to investigate, in which situations; these mixture distributions would produce outliers. In this paper we also attempt to develop robust MI(MCD) and MI(MVE) which are not easily affected by outliers and also distant groups. The simulation results signify that apparent outliers are created when the two distributions are extreme in either x or y observations. The classical MI performs poorly in the presence of outliers or when the two groups are distant apart. The MI(MCD) and the MI(MVE) estimators are indistinguishable and both of them give a reliable estimates of the true MI even when contaminations exist in the data. However, their performances are different when the two groups in the mixture distribution have the same number of observations ($\alpha = 0.5$) . This discrepancy may be due to the algorithm of the MVE whereby for equal number of observations in the two groups, the MVE considered observations in both groups to obtain the estimation of location and scatter. This inconsistency of the MI(MVE) is not only because of the equality in the number of observations in two groups, but also the positions of the two mixed groups. We have conducted many simulation experiments with different locations and scale parameters, and due to space limitations, we include only four models. The conclusions of other results are consistent and are not presented here due to space constraint.

References

[1] Cellucci, C. J., Albano, A. M., and Rapp, P. E. Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **200**, 6 (2005), 1.

[2] Cover, T. M., and Thomas, J. A. *Elements of Information Theory.* J. Wiley, New York, 1991.

[3] Davies, P. L. Asymptotic behavior of s-estimates of multivariate location parameters and dispersion matrices. *Ann.Statist.* **15** (1987), 1269.

[4] Harrold, T. I., Sharma, A., and Sheather, S. Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. *Stochastic Environmental Research and Risk Assessment* **15**, 4 (2001), 310.

[5] Hulle, M. M. V. Differential log likelihood for evaluating and learning gaussian mixtures. *Neural Computation* **18**, 2 (2005), 430.

[6] Kim, J., and Scott, C. Robust kernel density estimation. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (Las Vegas, NV, 2008), IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, p. 3381.

[7] Kraskov, A., Stgbauer, H., and Grassberger, P. Estimating mutual information. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* **69**, 62 (2004).

[8] Lopuhaa, H. P., and Rousseeuw, P. J. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann.Statist.* **19** (1991), 229.

[9] McLachlan, G. J., and Peel, D. *Finite mixture models.* Wiley series in probability and statistics. Applied probability and statistics section. Wiley, New York, 2000.

[10] Moon, Y. I., Rajagopalan, B., and Lall, U. Estimation of mutual information using kernel density estimators. *Physical Review E* **52**, 3 (1995), 2318.

[11] Rousseeuw, P. J. Least median of squares regression. *Journal of the American Statistical Association* **79** (1984), 871.

[12] Rousseeuw, P. J. Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications* **B** (1985), 283.

[13] Rousseeuw, P. J., and Leroy, A. M. *Robust regression and outlier detection*. Wiley series in probability and mathematical statistics. Applied probability and statistics. Wiley, New York, 1987.

[14] Rousseeuw, P. J., and Van Driessen, K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41**, 3 (1999), 212.

[15] Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal* **27** (1948), 379–423.

[16] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, 1986.

[17] Wang, Q., Shen, Y., and Zhang, J. Q. A nonlinear correlation measure for multivariable data set. *Physica D: Nonlinear Phenomena* **200**, 3-4 (2005), 287.

Received: October, 2009