Critical Values Of A Kernel Density-based Mutual Information Estimator

Robert J. May, *Student Member, IEEE*, Graeme C. Dandy, Holger R. Maier and T.M.K. Gayani Fernando

Abstract— **Recently, mutual information (MI) has become widely recognized as a statistical measure of dependence that is suitable for applications where data are non-Gaussian, or where the dependency between variables is non-linear. However, a significant disadvantage of this measure is the inability to define an analytical expression for the distribution of MI estimators, which are based upon a finite dataset. This paper deals specifically with a popular kernel density based estimator, for which the distribution is determined empirically using Monte Carlo simulation. The application of the critical values of MI derived from this distribution to a test for independence is demonstrated within the context of a benchmark input variable selection problem.**

I. Introduction

The identification of dependency within data is central to many algorithms used in a wide range of data analysis applications including function approximation, time-series analysis, and data mining. These applications often consider datasets where the variables are discrete, or where the structure of the dependency is potentially non-linear. In such cases, many conventional algorithms, in which correlation forms the basis for measuring dependence, fail to perform well because of a failure to identify, or accurately quantify, the dependency within the data. This is mostly due to the underlying assumptions that the data are Gaussian (i.e. continuous and normally distributed), and that the dependency between variables is linearly structured. Furthermore, correlation is not invariant under transformations of the data and these algorithms are therefore sensitive to any preprocessing that may alter the apparent dependency.

Recently, mutual information (MI) has gained recognition as a more suitable measure of dependence for applications that consider either discrete variables or potentially non-linear dependency. In contrast to correlation, MI is defined for both continuous and discrete distributions, and makes no assumptions regarding the structure of the dependency. Subsequently, a number of algorithms have been developed in which MI forms the underlying dependency measure. However, although MI can be considered an "ideal" measure from a theoretical perspective, a key practical shortcoming is that it is impossible to find an exact analytical expression for the distribution of a finite-sample estimate. The MI estimator is a stochastic measure of dependence and an understanding of the distribution is essential for formulating confidence bounds on sample-based estimates, which are needed to make a rigorous assessment of an observed degree of dependence in a particular case [1].

Several methods for approximating the distribution of MI have been defined in the case of discrete variables [2]–[5]. However, no equivalent expressions have been reported for continuous estimators. In the case of continuous estimators, practitioners have used Monte Carlo simulation (MCS) to estimate the distribution [6]–[8]. However, this can significantly increase the computational time of algorithms, particularly where multiple dependence tests are implemented. One proposed benchmarking methodology for kernel-density based estimators presents a possible alternative solution to this problem that could potentially improve the efficiency of algorithms, by providing a set of critical values [8] that could be used each time in place of MCS.

This paper describes the use of the benchmarking technique in [8] to obtain critical values for a popular MI estimator. The remainder of this paper is structured as follows: Section II provides some preliminary discussion of a popular kernel-density based estimator of mutual information, and highlights current methods for determining the distribution of MI estimators. Section III presents the methodology and results of a MCS undertaken in order to obtain critical values of the estimator described in Section II. In Section IV, the practical use of the critical values is demonstrated within the context of an input variable selection problem. Finally, concluding remarks are made in Section V.

II. Preliminaries

A. Estimation of Mutual Information

The MI between two random variables X and Y is defined as the net reduction in the uncertainty (or, *entropy*) surrounding the outcome of a given observation (x, y). Based on this formal definition, an indirect expression for MI is given in terms of the entropy, H, as [9]

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{1}$$

where I in (1) denotes the mutual information. Substitution of Shannon's formula for entropy [10] into (1) and rearrangement results in the following expression

$$I(X;Y) = \iint p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\, dxdy \tag{2}$$

where I is a direct function of the marginal probability density functions (pdfs) $p_X(x)$ and $p_Y(y)$; and of the joint pdf $p_{XY}(x,y)$. Both equations (1) and (2) will yield the "true" value of mutual information. However, for two practical reasons an estimator, $\hat{I}$, is usually constructed. First,

The authors are with the Centre for Applied Modelling in Water Engineering, School of Civil and Environmental Engineering, University of Adelaide, Adelaide, SA 5005, AUSTRALIA (phone: +618 8303 5451; fax: +618 8303 4539; email: robert.may@adelaide.edu.au).

in real-world circumstances the pdfs are often unknown and the "true" probabilities are replaced with density estimates. Second, the required integrations are numerically approximated as the summation of density estimates over a sample of observations. Consequently, a direct estimator for MI is given by

$$\hat{I}(X;Y) = \sum_{i=1}^{n} \log \frac{f_{XY}(x_i, y_i)}{f_X(x_i) f_Y(y_i)}. \tag{3}$$

where $f_X(x_i), f_Y(y_i)$ and $f_{XY}(x_i, y_i)$ denote the point estimates of the pdfs based on n sample observations. The base of the logarithm in (3) dictates the units of I. In this paper, the natural logarithm is assumed, and hence all values of mutual information reported have units of $nats$. Note that use of the binary logarithm is often reported elsewhere, which yields information in $bits$.

The MI estimator is therefore characterized by the density estimation technique. Non-parametric estimators (i.e. that use non-parametric density estimation) are preferred due to increased robustness, and one popular implementation is based on kernel density estimation (KDE). The histogram is also often used, in particular for discrete variables (e.g. in the case of classification datasets). However, for applications that consider continuous variables, KDE provides a more accurate density estimate [11], [12].

Based on the Gaussian kernel, a frequently adopted estimator for the pdf of a given sample of data is given by the expression

$$f(x) = \frac{1}{n\sqrt{2\pi h^d \left|\Sigma\right|}} \sum_{i=1}^{n} \exp - \frac{(x - x_i)^T \Sigma (x - x_i)}{2h^2} \tag{4}$$

where $f(x)$ is the estimate of the pdf at x based on the set of samples $\{x_1, \dots x_n\}$; d denotes the dimensions of the variable X; Σ is the sample covariance matrix; and h is the kernel bandwidth, or *smoothing parameter*.

Selection of an appropriate bandwidth is an important consideration for KDE, more so than the choice of kernel function itself [11]. Methods for selecting an optimal kernel bandwidth include cross-validation and plug-in bandwidth selections, which require some additional computation. However, for reasons of efficiency, [6], [13], [14] adopt the Gaussian reference bandwidth, which is given according to the Normal reference rule (or, *Scott's rule*) [12]

$$h = \left(\frac{1}{d+2}\right)^{1/(d+4)} \hat{\sigma} n^{-1/(d+4)} \tag{5}$$

where $\hat{\sigma}$ is the standard deviation of the sample data.

B. Distribution of Mutual Information

Applications dealing with dependence within data generally consider one of two cases: (1) whether one observed degree of dependence is greater than another, or (2) whether an observed degree of dependence is significantly greater than zero. In the case of stochastic measures of dependency, such as in the case of the MI estimator described, these cases must be considered within a statistical context [1]. Hence, knowledge of the distribution of the estimator is required in order to establish confidence bounds on estimates for a given sample, and to determine the critical values of the estimator. However, unlike the linear correlation coefficient, where the distribution of a sample-estimate follows a *t*-distribution, an equivalent analytical expression for $f(\hat{I})$ cannot be derived for the expression in (2) [5].

In [3], the distribution of discrete non-parametric MI estimators is considered to be the result of several contributing factors, which are:

1) sample variance;
2) sample bias;
3) discretization (quantization) bias; and
4) finite-histogram bias.

Following this, [3] describes a number of statistical methods for estimating the mean and variance of discrete (i.e. histogram-based) MI estimators. Alternatively, expressions for computing the mean, variance and conditional distribution $p(I|n)$ have been derived using the assumption of a prior distribution over the point density estimates [2], [4]. An expression for the distribution of MI has also been described based on a second-order Taylor series expansion of the discrete MI estimator [5].

In the case of KDE-based MI estimators, a similar assessment of the sources of bias and variance could be made considering the properties of the KDE approach (i.e. choice of bandwidth and finite-number kernels). However direct expressions for distribution parameters have not been derived. Hence, practitioners must resort to MCS in order to estimate $f(\hat{I})$, such as in [8] and [6]. MCS is a powerful method, since no assumptions regarding the distribution of the data are required. However, a large number of simulations (> 1000) are required to estimate the distribution accurately. This is of concern, since the kernel-density estimator is $O\{n^2\}$ and therefore analysis times can quickly become infeasible for large datasets. For example, only 100 Monte Carlo replicates were used in [6] to estimate the 95^{th} percentile MI, in order to maintain feasible run-times. Yet, with such a small number of replicates, this measure would be expected to exhibit significant variance.

Applications such as the one described in [6] are an example of an MI-based test for independence. In such cases, a possible means of avoiding the need to undertake MCS for each test is described in [8], where the distribution of a kernel-based estimator $\hat{s}_p$ was examined for a number of time-series models. In this study, the distributions of estimates of $\hat{s}_p$ obtained for lags of each model are benchmarked against the MI distribution for corresponding lags of a white-noise series, $y = \varepsilon_t$, from which critical values of $\hat{s}_p$ are inferred. Using this same approach, the distribution for the kernel density implementation of $\hat{I}$ could be derived and critical values obtained, thus allowing faster implementation of MI based tests for independence in algorithms such as that of [6]. In determining the critical values using this methodology, the effect of dependence on the distribution of MI is neglected. However, [7] observed that the variance in the MI estimator was reduced for increasingly correlated variables. Hence, the critical values obtained as part of this study would represent a worst-case set of confidence bounds on the MI estimated for any two variables.

III. Determining Critical Values of Mutual Information

A. Methodology

In this study, MCS was used to empirically determine the distribution for the MI estimator described in Section II for a bivariate dataset comprising i.i.d Gaussian white-noise data, with sample size n ranging from 50 to 5000 samples, in order to obtain a set of critical values that could be used for testing for independence based on MI.

For each sample size, a series $\varepsilon_y \sim N(0,1)$ was generated first and the marginal pdf f_{ε_y} estimated. A total of 100 000 independent replicates of series $\varepsilon_x \sim N(0,1)$ were generated, independent of ε_y. For each instance of ε_x the pdfs f_{ε_x} and $f_{\varepsilon_x \varepsilon_y}$ were estimated and $\hat{I}(\varepsilon_x, \varepsilon_y)$ subsequently evaluated. The critical values of the distribution of MI were then obtained from distribution formed by all computed values of $\hat{I}$. Code for this study was developed in C++ and compiled for Unix, with the Box-Muller transformation implemented for Gaussian pseudo-random number generation.

B. Approximate Distribution of Mutual Information

Although the intended purpose of generating the data was to extract critical values of the estimator, it is worth highlighting some features of the distribution, which are in good agreement with previous studies reported in the literature (see [4], for example) and thus provides verification of the simulation methodology and gives confidence in the critical values obtained. Figure 1 shows the pdf of $\hat{I}$ for the case $n = 500$, using the data generated by MCS. First, the empirical distribution is bounded by the condition $I > 0$, which agrees with the "true" mutual information. Second, the distribution about the mean is approximately Gaussian, which corresponds to the expected asymptotic behavior of the estimator for an infinite sample. Finally, it is observed that the distribution is skewed above the mean, indicating a tendency to over-estimate the MI.

C. Critical Values

Critical values of the MI estimator were obtained using the empirical distributions obtained for each sample size. These are summarized in Table I, which reports the mean, 90^{th}, 95^{th}, and 99^{th} percentile MI corresponding to the sample size, n. Figure 2, which graphically represents the tabulated data, more clearly indicates the finite-sample behavior of the MI estimator. Both the bias and variance of the estimator decrease monotonically with increasing sample size. This also provides some useful information as to the expected accuracy of an MI estimate for a given sample size.

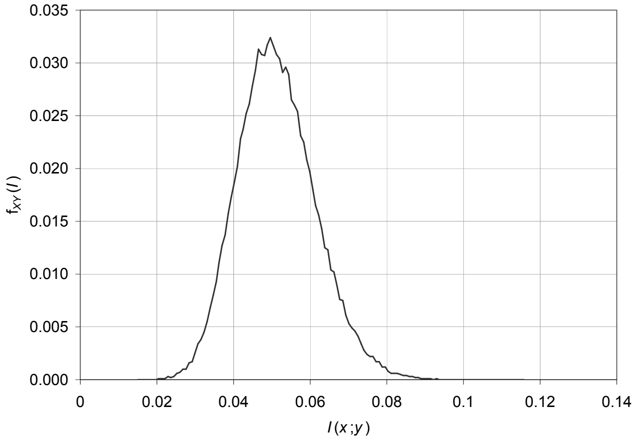


Fig. 1. Approximate pdf for the MI estimator, $n = 500$

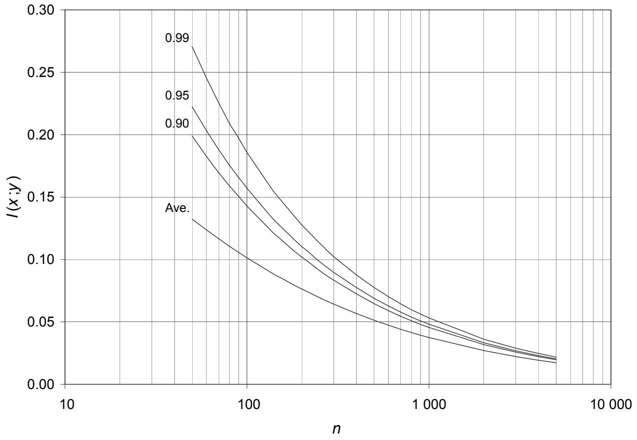


Fig. 2. Finite-sample behavior of the MI estimator

IV. Example Application

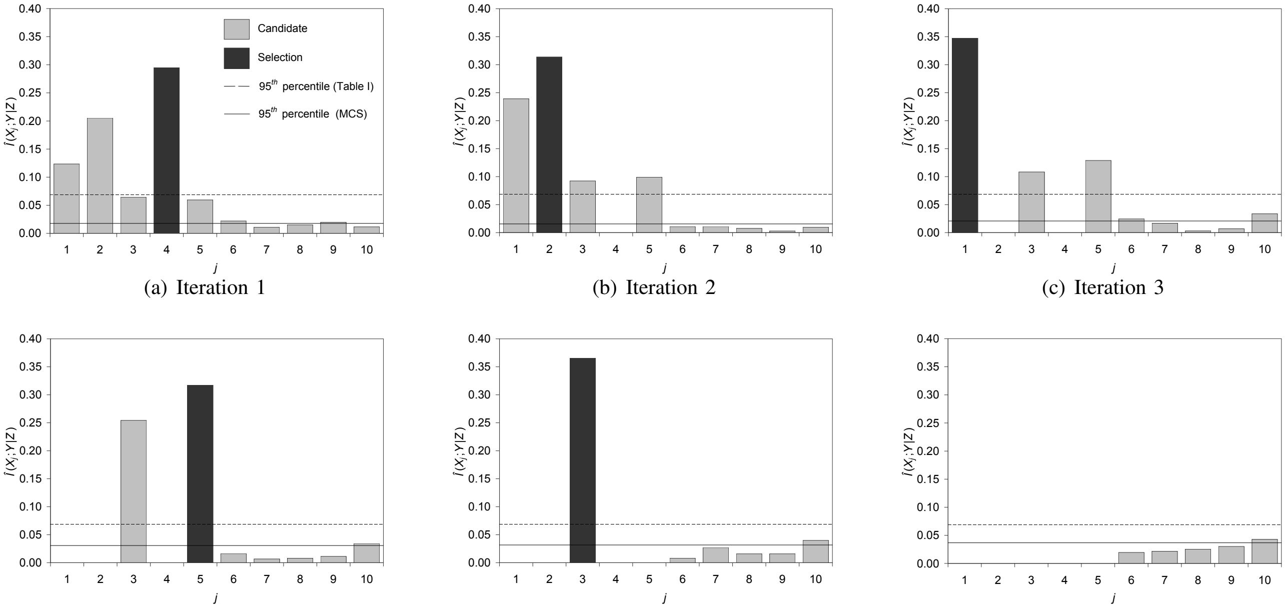
The problem of input variable selection provides an example of how the critical values in Table I can be used to improve the performance of algorithms, where a test for independence is required based on the MI estimator described in Section II. Considering a set $C = X_1, \ldots X_d$, which comprises all potential input variables, or *candidates*, for a model of some response variable Y, the input variable selection problem is defined as the task of appropriately selecting a subset of C that contains the minimum number variables required to achieve full mapping of Y [15]. This problem (also referred to as *feature selection* when constructing classifiers) has relevance to non-parametric regression and time-series model development, where popular techniques such as artificial neural networks (ANNs) are employed to map complex relationships based on a training dataset.

A. Selection Algorithm

The input selection algorithm used in this study was a forward selection filter originally proposed by Sharma [6] for determining the optimal inputs to models of hydrological

TABLE I
Critical values of the KDE-based mutual information estimator

n	$\bar{I}$	$I_{0.90}$	$I_{0.95}$	$I_{0.99}$	n	$\bar{I}$	$I_{0.90}$	$I_{0.95}$	$I_{0.99}$	n	$\bar{I}$	$I_{0.90}$	$I_{0.95}$	$I_{0.99}$
50	0.1323	0.1990	0.2224	0.2705	180	0.0798	0.1072	0.1166	0.1356	600	0.0473	0.0589	0.0627	0.0702
60	0.1236	0.1825	0.2031	0.2452	200	0.0763	0.1019	0.1103	0.1276	700	0.0441	0.0544	0.0578	0.0644
70	0.1166	0.1694	0.1879	0.2254	220	0.0735	0.0975	0.1055	0.1215	800	0.0415	0.0509	0.0539	0.0597
80	0.1106	0.1592	0.1756	0.2091	240	0.0707	0.0932	0.1005	0.1158	900	0.0393	0.0479	0.0507	0.0563
90	0.1057	0.1506	0.1657	0.1973	260	0.0682	0.0894	0.0965	0.1108	1000	0.0375	0.0455	0.0481	0.0531
100	0.1013	0.1429	0.1572	0.1858	280	0.0661	0.0862	0.0928	0.1062	2000	0.0270	0.0318	0.0333	0.0361
120	0.0943	0.1309	0.1434	0.1688	300	0.0642	0.0834	0.0896	0.1022	3000	0.0222	0.0257	0.0268	0.0289
140	0.0883	0.1211	0.1321	0.1546	400	0.0567	0.0724	0.0775	0.0876	4000	0.0192	0.0221	0.0230	0.0247
160	0.0839	0.1138	0.1237	0.1444	500	0.0513	0.0646	0.0689	0.0775	5000	0.0172	0.0196	0.0204	0.0218



(a) Iteration 1

(b) Iteration 2

(c) Iteration 3

(d) Iteration 4

(e) Iteration 5

(f) Iteration 6

Fig. 3. PMI-based selection of inputs for the non-linear ADD10 model based on a 500-sample dataset

time-series. The algorithm, which is based upon measuring the partial mutual information (PMI) of candidates, proceeds broadly as follows:

```
Initialize $Z \rightarrow \phi$
while $C \neq \phi$ do
  Find $s = \arg\max_j \hat{I}(X_j; Y|Z)$
  if $\hat{I}(X_s; Y|Z) > 0$ then
    $Z \rightarrow Z \cup X_s$
    $C \rightarrow C \setminus X_s$
  else
    break
  end if
end while
return $Z$
```

where Z denotes the subset of selected candidates; and $\hat{I}(X_j; Y|Z)$ denotes the PMI for candidate X_j (a more detailed description of the implementation of this algorithm can be found in [6] or [14]). An important feature of the algorithm is the test for independence, which forms the basis for the stopping criterion. The test adopts the 95^{th} percentile of the distribution of $I(X_s; Y|Z)$ as the critical value. In order to approximate the distribution, MCS is used in [6] during each iteration, where $I(X; Y|Z)$ is evaluated for 100 random permutations of the series X_s. In this study, the stopping criterion was modified by using the 95^{th} percentile from Table I as the critical value, thus removing the requirement for MCS at each iteration.

B. Dataset

A 500-sample dataset was generated by the ADD10 model, which is recommended as a model for benchmarking non-parametric regression techniques [16]. The ADD10 model is described by the function:

$$y = 5\left(2\sin(\pi x_1 x_2) + 4(x_3 - 0.5)^2 + 2x_4 + x_5\right) + \varepsilon \quad (6)$$

where $\varepsilon \sim N(0, 1)$ and $x_1, \ldots, x_5 \sim U[0, 1]$ are the uncorrelated input variables. The ADD10 model also includes an additional five uncorrelated noise variables $x_6, \ldots, x_{10} \sim U[0, 1]$, which represent irrelevant candidate inputs. This benchmarking methodology is therefore consistent with previous studies undertaken by [6]–[8], in which MI was used to determine the importance of variables for a number of non-linear data generating models.

C. Selection Results

Figure 3 shows the profile of the candidate set during the six iterations of the forward selection algorithm required to select the input set for the ADD10 model based on the 500-sample dataset. The PMI for each of the candidates X_j is indicated by the gray-shaded vertical bars, with the bar corresponding to the selected input shaded in dark gray. The dashed line indicates the 95^{th} percentile MI critical value of 0.069 obtained from Table I, which was used as the basis for testing independence. The solid line indicates the 95^{th} percentile obtained by MCS after each iteration.

The selected inputs (in order of selection) were X_4, X_2, X_1, X_5 and X_3, which indicates that the algorithm successfully specified the correct set of inputs for the ADD10 model. Figure 3(f) indicates clearly that, after the fifth iteration, none of the remaining candidates were significantly relevant and the selection was stopped. Hence, this result verifies that an accurate test for independence was achieved using the empirically derived critical values. In contrast, had Sharma's stopping criterion been used, the model would have been over-specified by at least one variable (and potentially more) since the critical value determined by MCS during the sixth iteration was too low, as also indicated in Figure 3(f).

A further comparison, between the accuracy of Sharma's MCS-based stopping criterion and the proposed stopping criterion, based on the critical values in Table I, showed an improvement in the accuracy of the forward selection algorithm. This was observed when the selection algorithm was applied to a number of independently generated datasets. The variability of the 95^{th} percentile MI stopping criterion, based on the permutation test performed after each iteration, led to inconsistent selections. The algorithm tended to over-specify the input set by including additional noise variables. In contrast, the stopping criterion based on the critical values selected the correct set of inputs each time.

Also, a significant reduction in the computational effort was achieved using the critical values from Table I. This is clearly indicated in Figure 4, which shows a comparison between cumulative run-time after each iteration of the forward selection algorithm, based on MI evaluations, with the test for independence implemented with and without MCS after each iteration. In the case of the small ADD10 dataset, MCS accounted for approximately 90% of the computation during each iteration. Given the difference in both accuracy and efficiency, there is a clear advantage to be gained by using critical values obtained by the benchmarking methodology presented in this paper.

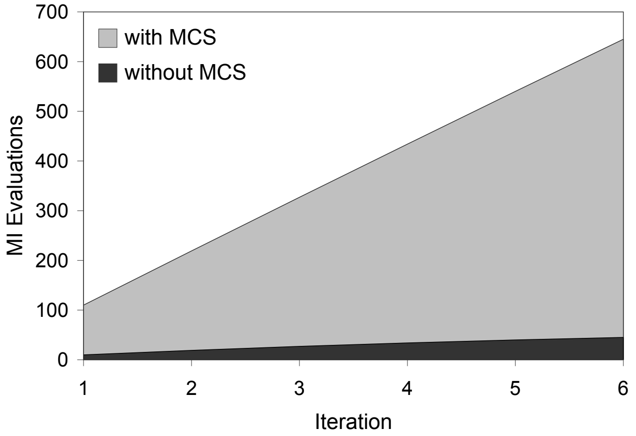


Fig. 4. Comparative run-time for PMI-based selection of inputs from the ADD10 dataset using critical values determined with and without MCS after each iteration.

V. Concluding Remarks

The motivation for this study was to formulate an efficient means of testing the independence of an estimate returned by a popular kernel-density based implementation of MI. In applications where this estimator is used as a dependency measure, such as in input variable selection algorithms, knowledge of the distribution of the finite-sample estimate of MI is paramount. Although analytical expressions to approximate the distribution of discrete MI estimators have been developed, to the authors' knowledge, a permutation test is the only method that has been reported for determining the distribution of the kernel density-based estimator.

This study has provided an alternative solution to the computationally intensive permutation test, by using a one-off MCS to determine a set of benchmark critical values of the MI estimator based on the analysis of white-noise data. The test for independence based on the critical values obtained has been verified within the context of a benchmark input selection problem, and was found to yield more accurate selections, but more importantly provide a significant reduction in the computational effort required.

Acknowledgment

This research was aided by the use of facilities provided by the South Australian Partnership for Advanced Computing (http://www.sapac.edu.au) and partially funded by United Water International Pty. Ltd. in conjunction with the Co-operative Research Centre for Water Quality and Treatment Project 2.5.0.1.

References

[1] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, pp. 1191–1253, 2003.

[2] D. H. Wolpert and D. R. Wolf, "Estimating functions of probability distributions from a finite set of samples," *Physical Review E*, vol. 52, no. 6, pp. 6841–6854, 1995.

[3] R. Moddemeijer, "A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations," *Signal Processing*, vol. 75, pp. 51–63, 1999.

[4] M. Hutter and M. Zaffalon, "Distribution of mutual information from complete and incomplete data," *Computational Statistics & Data Analysis*, vol. 48, pp. 633–657, 2005.

[5] B. Goebel, Z. Dawy, J. Hagenauer, and J. C. Mueller, "An approximation to the distribution of finite sample size mutual information estimates," in *IEEE International Conference on Communications (ICC-05)*, Seoul, South Korea, 2005.

[6] A. Sharma, "Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - a strategy for system predictor identification," *Journal of Hydrology*, vol. 239, pp. 232–239, 2000.

[7] A. Dionisio, R. Menezes, and D. A. Mendes, "Mutual information: A measure of dependency for nonlinear time series," *Physica A*, vol. 344, 2004.

[8] C. W. Granger, M. E., and J. Racine, "A dependence metric for possibly nonlinear processes," *Journal of Time Series Analysis*, vol. 24, no. 5, pp. 649–669, 2004.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York: John Wiley & Sons, Inc., 1991.

[10] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[11] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall, 1986.

[12] D. W. Scott, *Multivariate Density Estimation: Theory, Practice and Visualisation*. New York: John Wiley and Sons, 1992.

[13] T. W. S. Chow and D. Huang, "Estimating optimal feature subsets using efficient estimation of high-dimensional mutual information," *IEEE Transactions on Neural Networks*, vol. 16, no. 1, pp. 213–224, 2005.

[14] G. J. Bowden, G. D. Dandy, and H. R. Maier, "Input determination for neural network models in water resources applications. part 1 - background and methodology," *Journal of Hydrology*, vol. 301, no. 1-4, pp. 75–92, 2005.

[15] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537–550, 1994.

[16] J. Friedman, "Multivariate adaptive regression splines. technical report no. 102." Laboratory for Computational Statistics, Department of Statistics, Stanford University" Technical Report, 1988.