

ON THE ESTIMATION OF MUTUAL INFORMATION

Tomáš Marek, Petr Tichavský

Keywords: Mutual information, redundancy, adaptive histogram.

Abstract: The mutual information is useful measure of a random vector component dependence. It is important in many technical applications. The estimation methods are often based on the well known relation between the mutual information and the appropriate entropies. In 1999 Darbellay and Vajda [3] proposed a direct estimation methods. In this paper we compare some available estimation methods using different 2-D random distributions.

Abstrakt: Vzájemná informace je hojně užívanou mírou vzájemné závislosti jednotlivých složek vícerozměrných náhodných vektorů. Časté uplatnění nachází především v inženýrských aplikacích. Metody odhadu vzájemné informace většinou vychází ze známého vztahu mezi vzájemnou informací a entropiemi příslušných rozdělení, ale vzájemnou informaci je možné odhadnout také přímo. Na příkladech různých typů dvojrozměrných rozdělení srovnáme některé dostupné metody odhadu vzájemné informace.

1 Introduction

The mutual information is a very important tool in many engineering applications. Assuming 2-D random vector $(X, Y)^T$ with the joint density function $f_{X,Y}$ and marginal density functions f_X, f_Y, the mutual information $I(X, Y)$ is given as

$$I(X,Y) = \int_{\mathbf{R}^2} f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)}\, dx\, dy.$$

Traditional estimation methods are based on the well known relation between the mutual information and the appropriate entropies

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y), \\ &= H(X) - H(X|Y), \\ &= H(Y) - H(Y|X). \end{aligned}$$

In this paper we compare the available algorithms for direct computation of the mutual information with the estimate based on the maximum likelihood introduced by Miller [5]. In addition to the Miller-Madow's method we employ three methods described in the next section. All these methods of the direct computation of the mutual information are based on the 2-D histogram. The computer simulations bellow show the advantage of algorithms based on an adaptive histogram.

2 Methods and algorithms

2.1 Adaptive histogram methods

The adaptive histogram methods introduced by Darbellay [2] reach good efficiency. In this paper we use the algorithm published by Darbellay and Vajda [3]. The histogram generating process is based on the partitioning of the observation space into a finite number of nonoverlapping rectangular cells C_k, $1 \leq k \leq m$. The cells are generated by the recursive process described bellow.

Algorithm Mutin A

(i) The initial (the largest) cell is the smallest rectangular cell containing all data pairs $(X, Y)^T$.

(ii) Any cell containing less than two observations (data pairs) will not be partitioned.

(iii) Every cell containing at least two observations is tentatively partitioned by dividing each one of its edges into two equiprobable halves. It means four new cells are tentatively generated instead given 'mother' cell.

(iv) Assume that the partitioned 'mother' cell contains $N \leq 2$ observations. The new generated cells contain N_1, N_2, N_3 and N_4 observations. The partitioning is accepted if

$$T = \frac{4}{N} \sum_{i=1}^{4} \left(N_i - \frac{N}{4} \right)^2 > \chi^2_3(0.95) = 7.81, \tag{1}$$

where T is the goodness of fit statistic T intuitively testing the local independence of marginals at this 'mother' cell. If $T < 7.81$ then the tentative partition is refused and the 'mother' cell is admitted to the final computation.

(v) After stopping all (local) partitioning processes we denote generated cells C_k, $1 \leq k \leq m$ and the corresponding numbers of observations N_k. The estimate of the unknown mutual information $I(X, Y)$ we define as

$$\hat{I}_N(X, Y) = \sum_{k=1}^{m} \frac{N_k}{N} \log \frac{N_k/N}{(N_{x,k}/N)(N_{y,k}/N)}, \tag{2}$$

where $N_{x,k}$ is the number of observations that have the same x coordinate as observations in the cell C_k (analogically the $N_{y,k}$).

The computer simulation studies, e.g. Franěk [4], show that true expectation of the goodness of fit variable T defined in (1) is not constant but $\mathrm{E}\, T \leq 3$. Keeping notation of algorithm *Mutin A*, it approximately holds

$$\mathrm{E}\, T \approx 3 - \frac{N_{x,k}}{N_k} - \frac{N_{y,k}}{N_k}.$$

It follows that we should adjust the critical value of the goodness-of-fit test via (1). This adjustment decreases absolute value of the negative part of estimation bias, because the hypothesis of local independence is rejected in more cases and the partitioning process generates more cells with some positive contribution to the mutual information estimate (2). On the other hand the positive part of estimation bias slightly increases because of increasing the first type error probability in the test (1) which is not 0.05, but is lower in the algorithm *Mutin A*. The algorithm with adjusted critical values of the goodness-of-fit test will be called *Mutin B*.

2.2 Fixed histogram method

First we consider estimates using histograms based on a fixed partition of the observation space. There are two basic approaches to the partition making - the equidistant and the equiquantile partitioning of marginals, both followed by the 2-D product partition construction. The number of cells choice is the common problem of these methods. Generally, it is known that the optimal number of cells depends not only on the observation number but also on its 2-D distribution. This dependence is much stronger for estimators using equidistant cells.

In this paper, we denote *DirectD* $K \times K$ an algorithm based on the equidistantly generated partition containing K^2 rectangular cells. The cells are the Cartesian product of K equidistant intervals between maximal and minimal values on both marginals. The *DirectQ* $K \times K$ will denote an algorithm based on the partition also containing K^2 rectangular cells, but the cells are the Cartesian products of K intervals between maximal and minimal values on both marginals, which were chosen to contain approximately N/K observations. The estimates are calculated in the same way as in the case of the adaptive histogram methods, i.e. using the equation (2), where $m = K^2$.

2.3 Entropy method

The estimation of entropy is a well developed problem. There are many available methods in the literature. Consider independent identically distributed random vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N \in \mathbf{R}^p$ and a partition $B_1, \ldots, B_m$ of the space $\mathbf{R}^p$. Let us define a natural estimation of the entropy of $\mathbf{X}$'s distribution as

$$\widehat{H}_N(\mathbf{p}_N) = -\sum_{i=1}^{m} p_{N,i} \log p_{N,i}, \tag{3}$$

where $\mathbf{p}_N = (p_{N,i})_{i=1}^m$ are relative frequencies of the sets B_i. This estimate is also called maximum likelihood, plug-in (see Antos and Kontoyiannis [1]) or naive (see Strong [7]). The asymptotical properties of $\widehat{H}_N(\mathbf{p}_N)$ are summarized e.g. in Paninski [6]. Let the partition $B_1, \ldots, B_m$, be fixed, $H(p)$ be the entropy of discrete distribution $\mathbf{p} = (p_i)_{i=1}^m$ such that for any $1 \leq i \leq m$ holds

$p_{N,i} \to p_i$ if $N \to \infty$, and $p_i > 0$. The known results about the asymptotic bias and variance are

$$\mathrm{E}\left(\widehat{H}_N(\mathbf{p}_N) - H(p)\right) = -\frac{m-1}{2N} + O(N^{-1}), \tag{4}$$

$$\operatorname{Var}\widehat{H}_N(p_N) \leq \frac{(\log N)^2}{N}. \tag{5}$$

For the complete proof of (5) see Antos and Kontoyiannis [1]. The equation (4) is proved also by Miller [5].

At first we fix equidistant partition of marginals and the corresponding product 2-D partition of the observation space. Using the maximum likelihood entropy estimation with the Miller-Madow's bias correction

$$\widehat{H}_N(\mathbf{p}_N) = -\sum_{i=1}^{m} p_{N,i} \log p_{N,i} + \frac{\hat{m}-1}{2N}, \tag{6}$$

where $\hat{m}$ is number of nonempty cells in used partition, we calculate the entropy estimates $\widehat{H}_N(X)$, $\widehat{H}_N(Y)$ and $\widehat{H}_N(X,Y)$. The *Miller's* estimate of the mutual information can be defined as $\widehat{H}_N(X) + \widehat{H}_N(Y) - \widehat{H}_N(X,Y)$. The equidistant (*DirectD* 5×5) construction of partition is used for entropy estimation in the next section.

3 Simulation results

The computer simulations show the estimation results for 2-D data with various true mutual information using methods described in the previous section. The two types of data were generated. At first, the observations were linear dependent data pairs $(X,Y)^T$ with

$$Y = bX + Z,$$

where $X \sim U(0,1)$, b is linear dependence parameter and Z is random noise variable independent on X. The subplots in the right column of the figure 1 show comparable values of the standard deviations of all used methods without any strong dependence on the noise variable distribution. The left column subplots compare the mean values of the estimators. In the case of the Gaussian noise (a) the adaptive histogram methods are negligibly biased, the method *Mutin B* gives slightly higher values. The other methods have an observable negative bias. The dependence on the noise distribution is apparent in comparison of the Gaussian case (a), the uniform case (b) and the Cauchy case (c). The methods *Mutin A* and *Mutin B* have an observable negative bias in the uniform case, but a moderate positive bias in the Cauchy case. Both the other methods are much more biased. The Miller's estimator is inappropriate in the case of the Cauchy noise because of the partition problems during estimation of the needed entropies.

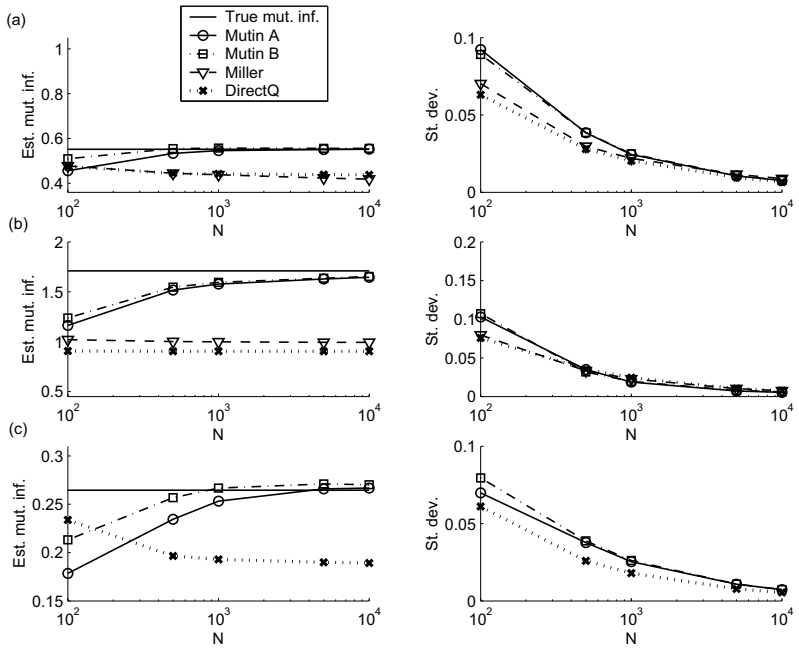

Figure 1: Results of 1000 simulated experiments with N random vectors $(X, Y)^T$ with uniformly distributed variable $X \sim U(0, 1)$ and $Y = 5X + Z$, where Z is random variable independent to X.
Figure (a): $Z \sim N(0, 1)$, $I(X, Y) = 0.5518$;
Figure (b): $Z \sim U(0, 1)$, $I(X, Y) = 1.7094$;
Figure (c): $Z \sim C(0, 1)$, $I(X, Y) = 0.2645$. The Miller's estimator is not displayed because of it's extreme bias in the Cauchy distribution case.

The figure 2 shows results obtained in the case of data that have zero linear correlation. The data are uniformly distributed on the annulus with the center in the origin and the various width. The observed standard deviations of all used methods are similar as in the previous case. The methods *Mutin A* and *Mutin B* have a lower bias than both other ones again, but it is seen that absolute bias decreases if the true mutual information increases. It means that the ratio MSE/$I(X, Y)$ of both *Mutin* estimates is notably lower in the case of higher true mutual information.

4 Conclusions

The computer simulation shows that the estimation methods based on the adaptive partition of observation space are more efficient than conservative methods based on a fixed partition. Especially the bias part of the MSE is

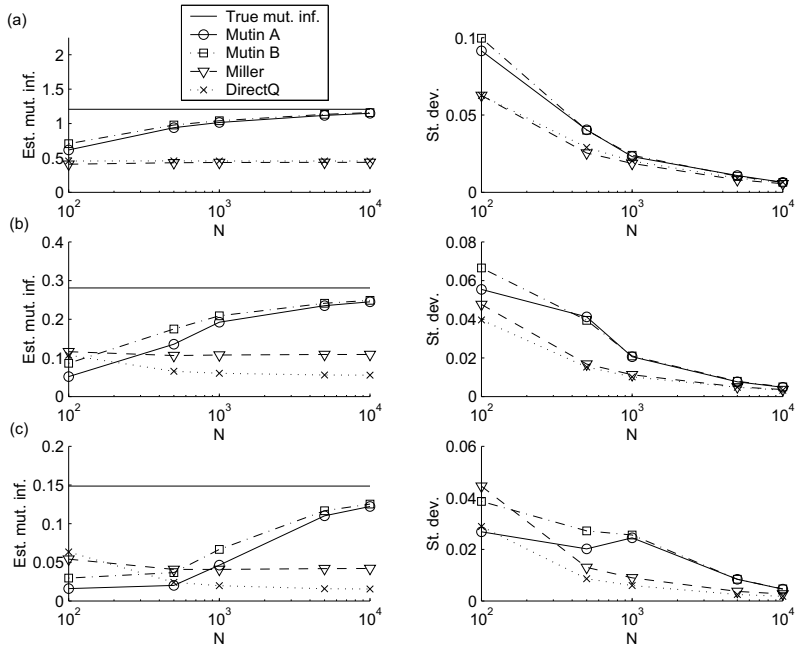

Figure 2: Results of 1000 simulated experiments with N random vectors $(X, Y)^T$ which are uniformly distributed on the annulus with center in the origin. The inner radius of this annulus is r_L and the outer radius is r_U.
Figure (a): $r_U = 1.1$ and $r_L = 0.9$, $I(X, Y) = 1.2081$;
Figure (b): $r_U = 1.5$ and $r_L = 0.5$, $I(X, Y) = 0.2809$;
Figure (c): $r_U = 1.9$ and $r_L = 0.1$, $I(X, Y) = 0.1485$.

seriously reduced. This reduction is stronger if the true mutual information is high and the methods work reasonably well also for data with heavy tailed distributions. The adaptive partitioning also removes the problem with the choice of the fixed partition. It is well known that the number of histogram cells has strong influence on the bias and the standard deviation of corresponding entropy estimators. Generally, the bias decreases and the standard deviation increases with increasing number of cells, see e.g. Paninski [6]. There are many methods optimizing the cells number with regard to the MSE in the literature. A different way to decrease MSE is to employ more accurate methods of density estimation as kernel estimates. We did not deal with this topic for a lack of space. Finally, it is seen that the methods employing the adaptive partitioning of the observation space are very user friendly regarding the implementation and the computation time.

References

[1] Antos A., Kontoyiannis I. (2001). *Convergence properties of functional estimates for discrete distributions*. Random Structures and Algorithms, **19**, 163 – 193.

[2] Darbellay G.A. (1999). *An estimator for the mutual information based on the criterion for independence*. Journal of the Computational Statistics and Data Analysis, **32**, 1 – 17.

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

[4] Franěk J. (2002). *A non parametric estimator of mutual information, redundancy and entropy for continuous random vectors (theory, implementation, applications)*. Ph.D. Thesis, Faculty of Nuclear Science and Physical Engineering Czech Technical University, Prague.

[5] Miller G. (1955). *Note on the bias of information estimates*. In H. Quastler (Ed.), Information Theory in Psychology II-B, Glencoe, IL: Free Press, 95 – 100.

[6] Paninski L. (2003). *Estimation of entropy and mutual information.* Neural Computation, **15**, 1191 – 1253.

[7] Strong S., Koberle R., de Ruyter van Steveninck R., Bialek W. (1998). *Entropy and information in neural spike trains.* Physical Review Letters, **80**, 197 – 202.

Acknowledgement: Tato práce byla podporována projektem 1M0572 MŠMT ČR.

Address: T. Marek, P. Tichavský, ÚTIA, AV ČR, Pod Vodárenskou věží 4, 182 08 Praha 8

E-mail: marek@utia.cas.cz