Brazilian Journal of Probability and Statistics (2004), 18, pp. 163–182.

Some data analyses using mutual information

David R. Brillinger
University of California

Abstract: This paper presents a number of data analyses making use of the concept of mutual information. Statistical uses of mutual information are seen to include: comparative studies, variable selection, estimation of pa rameters and assessment of model fit. The examples are taken from the fields of sports, neuroscience, and forest science. There is an Appendix providing proofs.

Key words: Data analysis, entropy, flooding, mutual information, football, time series, wildfires.

1 Introduction

> *"... . This shows that the notion of information, which is more closely related to the mutual information in communications theory than to the entropy, will play the most fundamental role in the future developments of statistical theories and techniques."* Akaike (1972)

This paper is a study of the usefulness of the coefficient of mutual information in statistical data analysis. The paper examines the utility in practice of estimates.

Mutual information (MI) is a measure of statistical dependence. The concept was introduced by Shannon (1948). Since then there has been substantial theoretical and practical development of the concept. For example MI has been proposed as a criterion on which to base a test of independence, Fernandes (2000) and as a quantity to maximize in order to estimate lag, Li (1990), Granger and Lin (1994) and in the spatial case to register images, Viola (1995). In particular in the bivariate case MI is the Kulback-Liebler distance between a joint distribution and the product of its marginals, see Joe (1989a,b), Cover and Thomas (1991) and the references therein. Dependence and association analysis are basic to statistics and science. In particular regression analysis and canonical correlation analysis may be mentioned. Some other questions to which MI would seem able to usefully contribute are: change? trend? serial correlation? dimension? model fit? variable selection?, model?, efficiency?, strength of association?

Re the last, the correlation coefficient is a long-standing measure of the strength of statistical dependence; however MI has advantages over it. These include that the variates involved do not have to be euclidian and that MI measures more than linear dependence.

There seems to have been substantial practical investigation of the related concept of entropy, including the introduction of some novel estimators. Papers concerned with the properties and estimation of entropy include: Miller (1955), Parzen (1983), Moddemeijer (1989, 1999, 2000), Hall and Morton (1993), Robinson (1991).

The paper begins with a brief discussion of the coefficient of determination, ρ^2, to contrast its properties with the coefficient of mutual information. Three empirical analyses are presented. There is discussion and then some formal development in an Appendix. This paper focuses on the case of independent identically distributed variates. It further concerns distributions described by a finite dimensional parameter.

2 Correlation analysis

Science studies relationships generally, while regression analysis studies the dependence of a variate Y with X. One can ask the question: what is the strength of a particular relationship? A common answer is the following: given the bivariate random variable (X, Y) employ the coefficient of determination,

$$\rho^2_{XY} \;=\; corr\{X, Y\}^2 \tag{2.1}$$

This measure is is symmetric and invariant and useful for studying: 1) implications of statistical independence, 2) explained variation, 3) strength of linear dependence, and 4) uncertainty of estimates.

For real-valued variates, X and Y, ρ^2_{XY} has long been estimated by

$$r^2 \;=\; [\sum (x_i - \bar{x})(y_i - \bar{y})]^2 \;/\; \sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2$$

There appears to be no such natural estimate for mutual information although several will be proposed.

3 Mutual information

3.1 Definition and properties

For the bivariate r.v. (X, Y) with pdf or pmf $p(x, y)$ the MI is defined as

$$I_{XY} \;=\; E\{log \frac{p(X, Y)}{p_X(X) p_Y(Y)}\} \tag{3.2}$$

The units of MI are sometimes referred to as *nats*.

In the case of a bivariate discrete distribution with pmf

$$Prob\{X_j \;=\; j, Y_k \;=\; k\} \;=\; p_{jk}, \quad j = 1, ..., J;\; k = 1, ..., K,$$

expression (3.2) becomes

$$I_{XY} \;=\; \sum_{j,k} p_{jk}\; log \frac{p_{jk}}{p_{j+}p_{+k}}$$

where $p_{j+} = Prob\{X = j\}$ and $p_{+k} = Prob\{Y = k\}$ and the sum is over $p_{jk} \neq 0$.

Consider a hybrid discrete-continuous variate with $p_j(y)$ given by

$$Prob\{X \;=\; j, Y \,\epsilon\, \Delta\} \;\approx\; p_j(y)|\Delta|$$

with Δ a small interval including y of length $|\Delta|$. Then the MI is

$$I_{XY} \;=\; \sum_j \int\; p_j(y)\; log \frac{p_j(y)}{p_{j+}p_Y(y)}\; dy, \qquad p_j(y) \;\neq\; 0 \tag{3.3}$$

p_{j+} and $p_Y(.)$ being the marginals.

Properties of I_{XY} include:

1. Non-negativity, $I_{XY} \;\geq\; 0$;
2. Invariance, $I_{XY} \;=\; I_{UV}$ if $u = u(x)$ and $v = v(y)$ are individually 1-1 measureable transformations;
3. Measuring strength of dependence in that,
   - i) $I_{XY} \;=\; 0$ iff X is independent of Y;
   - ii) For the continuous case, $I_{XY} \;=\; \infty$ if $Y \;=\; g(X)$;
   - iii) $I_{XZ} \;\leq\; I_{XY}$ if X independent of Z given Y;
   - iv) For the bivariate normal, $I_{XY} \;=\; -\,0.5 * log(1 \;-\; \rho^2_{XY})$;
   - v) There are ANOVA like decompositions.

A conditional form

$$I_{XY} \;=\; E\{log \frac{p_{Y|X}(Y)}{p_Y(Y)}\}$$

is sometimes employed.

A useful inequality is,

$$E\{Y \;-\; g(X)\}^2 \;\geq \frac{1}{2\pi e} \exp\{2(I_{YY} \;-\; I_{XY})\} \tag{3.4}$$

where g is measurable and I_{YY} is the entropy of Y, $E\{log\; p_Y(Y)\}$, see Cover and, Thomas (1991) supplementary problems. An implication of (3.4) is that the larger I_{XY} the smaller will be the lower bound for predicting Y via a function $g(X)$ of X. It is thus useful for investigating the efficiency of proposed estimates.

Joe (1989a,b) proposes the use of

$$1 \;-\; exp\{-2I_{XY}\} \tag{3.5}$$

as a ρ^2 or R^2 like measure.

3.2 Estimation

In a study of model identification Akaike (1972, 1974) has shown that there are important connections between the likelihood function and the Kullback-Liebler "distance", from the true model to any model. Taking the K-L distance from the model of independent marginals leads to the coefficient of mutual information, the K-L 'distance" of p to q being

$$E_U\{log\ p(U)/q(U)\}$$

where U is a random variable with density or pmf $p(u)$.

3.2.1 The parametric case

Consider a parametric model $p(x, y|\theta)$ where p is a pdf or a pmf or a hybrid depending on the circumstance. For the bivariate r.v. (X, Y) suppose realizations $(x_i, y_i),\ i = 1, ..., n$ are available. Suppose that one wishes to estimate the mutual information of X and Y,

$$I_{XY}(\theta)\ =\ E\left\{log\frac{p(X,Y|\theta)}{p_X(X|\theta)\ p_Y(Y|\theta)}\right\} \tag{3.6}$$

With $\hat{\theta}$ an estimate of θ, e.g. the mle, a natural estimate of the MI is

$$I_{XY}(\hat{\theta}) \tag{3.7}$$

This estimate has in mind that the expected value (6) can be well-evaluated numerically for any given θ.

3.2.2 Two particular examples

To begin consider two particular cases. The first example involves a bivariate discrete chance quantity (X, Y) with X taking on the values $1, ..., J$ and Y the values $1, ..., K$ and

$$Prob\{X\ =\ j,\ Y\ =\ k\}\ =\ p_{jk}$$

Write the marginals as $p_{j+},\ p_{+k}$. The MI here is

$$I_{XY}(\theta)\ =\ \sum_{j,k} p_{jk}\ log\frac{p_{jk}}{p_{j+}p_{+k}} \tag{3.8}$$

Represent the variate (X, Y) by $V\ =\ \{V_{jk}\}$ with $V_{jk} = 1$ if the result (j, k) occurs and $V_{jk} = 0$ otherwise. The probability mass function is

$$\frac{1}{\prod_{j,k} v_{jk}!}\prod_{j,k} {p_{jk}}^{v_{jk}},\qquad v_{jk}\ =\ 0\ or\ 1,\ \sum_{j,k} v_{jk}\ =\ 1$$

Suppose next that there are n independent realizations, $\{v_{jkl}, l = 1, ..., n\}$, of V. Suppose that θ, the unknown parameter, is $\{p_{jk}\}$. The maximum likelihood

estimates of the p_{jk} are the $\hat{p}_{jk} = \sum_l v_{jkl}/n$ and the plug-in estimate of the MI is

$$I_{XY}(\hat{\theta}) = \sum_{j,k} \hat{p}_{jk} \; log \frac{\hat{p}_{jk}}{\hat{p}_{j+}\hat{p}_{+k}} \tag{3.9}$$

Some statistical properties will be considered below.

Next consider now the likelihood ratio test statistic of the null hypothesis of the independence of X and Y, namely

$$G^2 = 2n \sum_{j,k} \hat{p}_{jk} \; log \frac{\hat{p}_{jk}}{\hat{p}_{j+}\hat{p}_{+k}} \tag{3.10}$$

see Christensen (1997). The quantity G^2 is seen to be proportional to the estimate (3.9). Further from classical statistical theory in the case that X and Y are independent the asymptotic null distribution of (3.10) is $\chi^2_{(J-1)(K-1)}$. One can conclude that the large sample distribution of the estimate (3.9) is $\chi^2_{(J-1)(K-1)}/2n$ in the null case of independence.

The non-null large sample distribution is more complicated. It is normal with mean (3.8) and variance

$$\frac{1}{n}\left(\sum_{j,k} p_{jk}[log\frac{p_{jk}}{p_{j+}p_{+k}}]^2 - [\sum_{jk} p_{jk} log\frac{p_{jk}}{p_{j+}p_{+k}}]^2\right) \tag{3.11}$$

according to Moddemeijer (1989). One notes that expression (3.11) is 0 when the variables are independent, consistent with the χ^2 expression above. The non-null distribution arises in power computations. There are a number of studies of power considering Pitman alternatives, see for example Mitra (1958).

As a second example consider the vector Gaussian case. Let Σ be the covariance matrix of the column variate $V = (X', Y')'$ with X r-vector-valued and Y s-vector-valued. The (differential) entropy is

$$E\{log\; p_V(V)\} = \frac{1}{2} log(|2\pi e \Sigma|) \tag{3.12}$$

with $|.|$ denoting the determinant, see Cover and Thomas (1991).

From (3.12) then the MI of X and Y is

$$I_{XY}(\theta) = - \frac{1}{2} log(|\Sigma|/|\Sigma_{XX}||\Sigma_{YY}|) \tag{3.13}$$

having partitioned Σ as

$$\begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}$$

One can write

$$|\Sigma|/|\Sigma_{XX}||\Sigma_{YY}| = \prod_i (1 - \rho_i^2)$$

with the ρ_i the canonical correlations and expression (3.13) becomes

$$-\frac{1}{2}\sum_i \ log(1-\rho_i^2) \tag{3.14}$$

The absence of much of the structure of Σ from (3.14) is to be noted. This follows from the invariance of $I_{XY}(\theta)$ under linear transformations of X and Y indicated in Section 3.1 above.

In what follows let the parameter θ be Σ. When the experiment is repeated n times the maximum likelihood estimate of Σ is

$$\hat{\Sigma} \ = \ \frac{1}{n}\sum_{i=1}^{n}(v_i-\bar{v})(v_i-\bar{v})'$$

and the plug-in estimate (3.13) becomes

$$\hat{I}_{XY}(\theta) \ = \ - \ \frac{1}{2}log(|\hat{\Sigma}|/|\hat{\Sigma}_{XX}||\hat{\Sigma}_{YY}|) \tag{3.15}$$

whose statistical properties will be considered below.

For this Gaussian case, consider the log-likelihood ratio criterion for testing the independence of X and Y. It is

$$\frac{n}{2}log\ |\hat{\Sigma}|/|\hat{\Sigma}_{XX}||\hat{\Sigma}_{YY}| \tag{3.16}$$

see Kendall and Stuart (1966), section 42.12. From classical likelihood ratio test theory the large sample null distribution of (3.15) is χ^2_{rs}. (It may be worth noting that some 'better" approximations have been proposed, ibid.) The statistic (3.16) is proportional to the plug-in estimate (3.15).

Turning to the large sample distribution in the non-null case, using (3.14) the statistic (3.15) may be written

$$-\frac{1}{2}\sum_i log(1-\hat{\rho}_i^2)$$

with the $\hat{\rho}_i$'s the sample canonical correlations. In the case that they are distinct and non-zero the ρ_i's are asymptically independent normal with means ρ_i and variances $\frac{1}{n}(1-\rho_i^2)^2$, see Hsu (1941). It follows that, in this case, the estimate (3.15) is asymptotically normal with mean $I_{XY}(\theta)$ and variance $\sum_i \rho_i^2/n$.

In summary, for these two circumstances the plug-in estimate of the MI is essentially the likelihood ratio statistic for testing independence. Distributional results that are available for the latter are directly applicable.

3.2.3 Approximate distributions

There are some general results.

Suppose that a sample of values $(x_i, y_i),\ i = 1, ..., n$ is available. Let θ_0 denote the true parameter. Let $\hat{\theta}$ denote the maximum likelihood estimate. Write I_0 for $I_{XY}(\theta_0)$ and $\partial I_0/\partial\theta'$ for $\partial I_{XY}(\theta)/\partial\theta'$ evaluated at θ_0. Write J_{XY} for the Fisher information of $(X', Y')'$ at θ_0 and $J_X,\ J_Y$ for that of X and Y respectively.

Both the cases of independent and dependent X and Y are considered in the theorem. Assumptions and derivations are provided in Appendix A.

Theorem 3.1. *Suppose that Assumption A.2 holds.*
a) In the case that X and Y are dependent and that $\partial I_0/\partial\theta$ is not 0, the variate $\sqrt{n}(I_{XY}(\hat{\theta}) - I_{XY}(\theta_0))$ is asymptotically normal with mean 0 and covariance matrix

$$\frac{\partial I_0}{\partial\theta}' J_{XY}^{-1} \frac{\partial I_0}{\partial\theta}$$

b) In the case of independence, $nI_{XY}(\hat{\theta})$ is distributed asymptotically as

$$\frac{1}{2} Z' J_{XY}^{-1/2} [J_{XY} - J_X - J_Y] J_{XY}^{-1/2} Z \tag{3.17}$$

where the entries of Z are independent standard normals.

The variate (3.17) will be $\frac{1}{2}\chi^2_\nu$ when $J_{XY}^{-1/2}[J_{XY} - J_X - J_Y]J_{XY}^{-1/2}$ is idempotent with trace ν.

In particular, the estimate, $I_{XY}(\hat{\theta})$, is consistent in both cases a) and b).

A second estimate of $I_{XY}(\theta)$ is provided by

$$\frac{1}{n}\sum_i log\left(p(x_i, y_i|\hat{\theta})\ /\ p_X(x_i|\hat{\theta})\ p_Y(y_i|\hat{\theta})\right) \tag{3.18}$$

with $\hat{\theta}$ again the overall maximum likelihood estimate. No integral needs to be evaluated in this case; however there are difficulties in developing its properties analagous to those arising in the estimation of entropy, see Robinson (1991), Granger and Li (1994), Hall and Morton (1993). Modified estimates of entropy are proposed in those papers.

As indicated by the discrete and multivariate normal examples above, another type of estimate of I_{XY} is sometimes available. Suppose that the parameter θ has the form $\theta = (\phi, \psi)$ and that the marginal distributions $p_x(.),\ p_y(.)$ only involve ϕ. Let $\hat{\phi}_*$ denote the mle of ϕ under the null hypothesis of independence. Consider the estimate

$$\frac{1}{n}\sum_i log\left(p(x_i, y_i|\hat{\theta})\ /\ p_X(x_i|\hat{\phi}_*)\ p_Y(y_i|\hat{\phi}_*)\right) \tag{3.19}$$

with $\hat{\theta}$ the full model mle. Expression (3.19) is the classic $log(likelihood\ ratio)/n$ test statistic for the hypothesis of independence.

Provided $\hat{\phi}_* \rightarrow \phi$ in probability generally, the statistic (3.19) will tend to $I_{XY}(\theta_0)$ in probability, i.e. (3.19) provides a consistent estimate of the MI. However the distinction is that the distribution of $\hat{\phi}_*$ is to be considered under the full distribution of (X, Y), not just the null.

An advantage when this situation obtains is that classical maximum likelihood theory indicates an asymptotic null distribution of

$$\chi^2_\nu \ / \ 2n, \quad \nu \ = \ dim(\psi) \tag{3.20}$$

for (3.19).

Theorem 3.2. *Suppose Assumption A.3 holds. Suppose that $\hat{\phi}_*$ converges in probability to ϕ. Then,*
a) the quantity (3.19) converges to $I_{XY}(\theta_0)$ in probability.
b) Suppose that X and Y are independent, then the large sample distribution of (3.19) is (3.20).

The statistic (3.19) has the advantage of being obtainable directly from the output of various mle programs.

3.2.4 The non-parametric case

Various inferential results have been developed for entropy. To mention one class of estimates studied, consider $\hat{p}(x, y)$ an estimate of $p(x, y)$, e.g. the histogram or a kernel-based one. Now one can consider plug-in estimates of mutual information, namely,

$$\hat{I}_{XY} \ = \ \sum_{j,k} \hat{p}(u_j, v_k) \ log \frac{\hat{p}(u_j, v_k)}{\hat{p}_X(u_j)\hat{p}_Y(v_k)} \ , \tag{3.21}$$

with (u_j, v_k) a grid of nodes, or

$$\hat{I}_{XY} \ = \ \int\int k(x, y)\hat{p}(x, y) log \frac{\hat{p}(x, y)}{\hat{p}_X(x)\hat{p}_Y(y)} dxdy \tag{3.22}$$

k being a kernel introduced to improve asymptotic properties. There are difficulties for $\hat{p}$ near 0.

A variety of authors have considered properties of this and related estimates. Antos and Kontoyiannis (2000) show that while plug-in estimates are uniformly consistent, under mild conditions, the rate of convergence can be arbitrarily slow, even in the discrete case. Beirlant et al (2001) provide a review of plug-in estimates of entropy of the type: integral, resubstitution, splitting data and cross-validation. Fernandes (2000) studies MI-like statistics for testing the assumption of independence between stochastic processes. The series are mixing. Robinson (1991) considered kernel-based estimates, as did Skaug and Tjostheim (1993). Joe (1989b) obtained consistency results for the estimates of type 1 and 2 above and obtained asymptotic mean-squared error results. Hall and Morton (1993) studied properties of Joe's estimates with emphasis on tail behavior, distribution smoothness and dimensionality. Hong and White (2000) develop asymptotic distributions

of estimates of Robinson (1991) and Granger and Li (1994). In a series of papers Moddemeijer (1989,1999,2000) studies various large sample properties of estimates of entropy.

3.3 Bias and statistical uncertainty

One needs statistical properties of estimates in order to make statistical inferences. As indicated above in certain cases the approximate null distribution of $\hat{I}_{XY}$ is chi-squared. In the case of (3.9) it is

$$\chi^2_\nu \;/\; 2n, \; where \;\; \nu \;=\; (J-1)*(K-1)$$

For example approximate p-values of the hypothesis of independence may be computed.

Both asymptotic developments and simulation experiments have shown that bias can be a problem in the estimation of entropy. This could have been anticipated because of the nonlinear character of mutual information as a function of its parameters. Miller (1955) proposed an elementary correction to (3.9). Woodfield (1982) studies estimate based on transforming marginals to uniforms and finds bias problems in a simulation study.

Because of the messiness of the expressions involved, nonparametric uncertainty procedures are often very helpful. These include the δ-method of propagation of error, the jackknife and the bootstrap. In particular the latter two can both reduce bias and provide estimates of uncertainty.

4 Examples

The histogram estimate (3.9) is used throughout when the data form a contingency table and the R/Splus function kde2d when X and Y are jointly continuous. It is assumed that the explanatory, X, is stochastic.

4.1 An example with two discrete variables

This is an example of the use of MI in a comparative study.

Soccer fans have often discussed the home team advantage and there are controversies. To study an interesting aspect of this, consider the specific question: in which country is the relationship strongest between the number of goals a team scores and and the circumstance that it is playing at home?

Lee (1997) used Poisson regression in a study of the English Premier Division specifically. He includes a home-away effect in the model. In contrast this paper presents a study of countries, not teams.

Data for the world's Premier Leagues of many countries are available at

sunsite.tut.fi/rec/riku/soccer2.html

The analysis that follows considers the 2001-2002 season and the countries: Argentina, Brazil, Canada, Chile, France, Germany, Italy, Portugal, Spain, Uruguay. These countries were studied because the example was developed for talks in Brazil.

The variates, X and Y, were defined as follows: $Y = 0, 1, 2, 3, 4+$ gives the number of goals a team scored in an away game while $X = 1, 0$ indicates whether the team was playing at home or away. The value 4+ represents 4 or more goals scored by a team.

The estimate (3.9) is employed and the formula for the independent identically distributed case has been used to obtain the upper 95% level. The results are given in Figure 1.

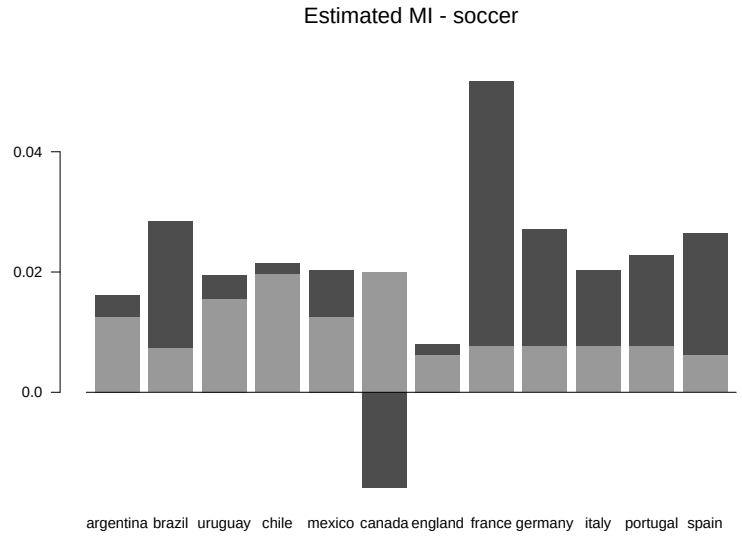

Figure 1 *Estimated MI between goals a team scored in a game and whether the team was playing at home or away. The heights of the grey bars provide the approximate 95% of the null points. The Canada value is below the line because it would have been hidden by the grey shading above.*

One sees France standing above all the other countries with a strong home effect and Canada showing none to speak of.. One colleague suggested that France stood out because its stadiums were such that the fans were particularly close to the field. In the case of Canada, its Premier Division is minor league.

The assumption of independence may be problematic because often a suite of games is played on a given day and, for example, weather conditions may be in common.

4.2 A real-valued time series example

This example involves checking a real-valued stationary time series for independence.

The data studied are based on a spike train of 951 firings of the neuron L10 of the sea hare, *Aplysia californica*, when it was firing spontaneously. Supposing the times of the spike train to be $\{\tau_k\}$ Let $\{Z_k = \tau_{k+1} - \tau_k\}$ denote the intervals between the firings.

When a neuron is firing spontaneously many of the proposed models imply that intervals are independent and identically distributed, i.e. the point process is renewal. An estimate of the MI was computed to address the question of whether the series of interspike intervals may be viewed as white noise.

Supposing $X_i = Z_i$ and $Y_i = Z_{i+h}$ the MI is estimated as a function of lag h. The results are shown in Figure 2.

The 99% critical level is estimated by repeating the MI estimation for random permutations of the intervals. It is the dashed line in the second panel.

The figures provide evidence against the assumption of a renewal process. Specifically there is a suggestion in both the top two panels of relationship at the very low lags.

What is different here from traditional studies is that serial association of the interval sequence has been examined over a broader range of possibilities.

4.3 A discrete-continuous example

This example involves selecting the variable most strongly associated with a given binary response variate and checking on the efficiency of some parametric models..

Estimates of the risks of wildfires are basic to governments' preparations for forest fires and their handling once detected. The problem is important because in many cases there are deaths and very large financial losses.

In dealing with the problem so called fire indices are often computed and promulgated. For example there are the Keetch-Byram Drought, the Fire Potential, the Spread Component and the Energy Release Component Indices, see Preisler et al. (2004).

One question of concern is whether a fire once started will become large. Mutual information will be employed to infer which of the four indices is most highly associated with a fire becoming large. Further the efficiencies of three parametric models of generalized linear model form will be studied.

The data employed are for the federal lands in the state of Orgeon for the years 1989 to 1996. They are discussed in Brillinger et al. (2003), Preisler et al. (2004). The state is divided into 1km by 1km pixels. The times and pixels in which fires occurred are recorded. Further the size of the fire is estimated.

For the mutual information analysis the response variable, Y, is defined to be 1 if a fire becomes large and 0 otherwise. The explanatory variable, X, is the value of the 4 indices in turn, i.e. four separate analyses are carried out.

The results are provided in Figure 3. The final panel is the nonparametric estimate while the previous three refer to the specific Bernoulli models employing

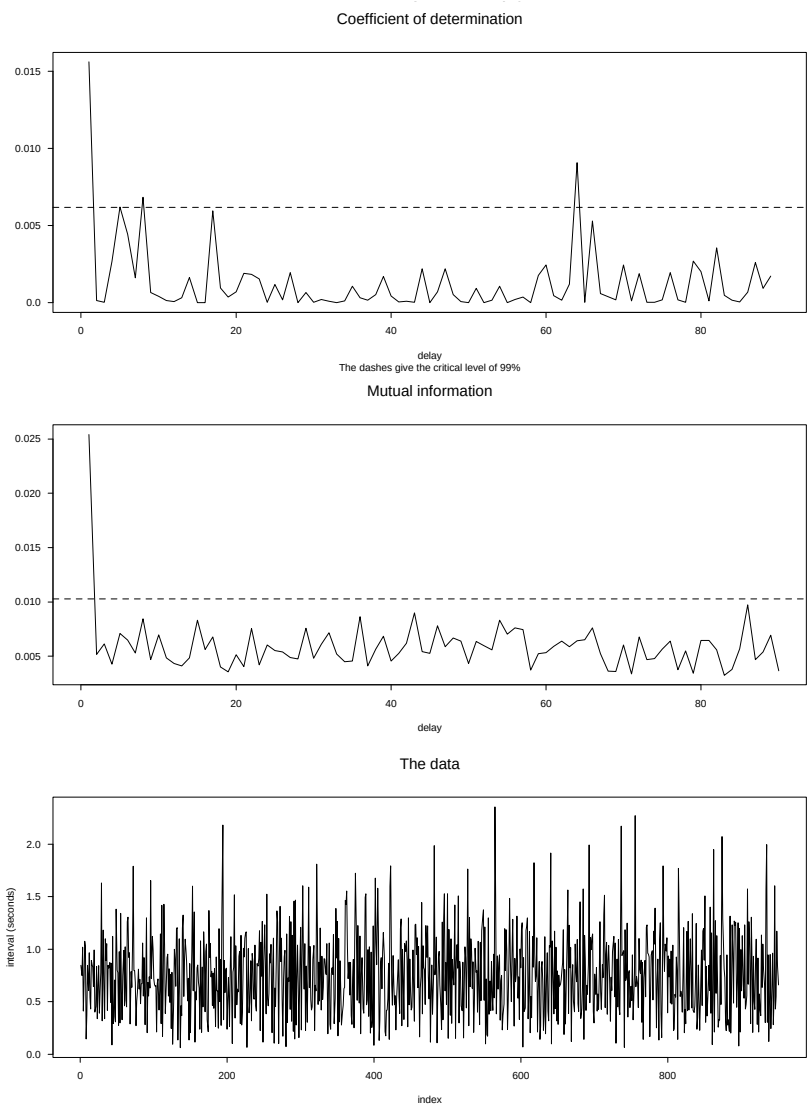

Figure 2 *From the bottom, the panels are respectively: a plot of the series, the estimated mutual information as a function of lag and the estimated coefficient of determination.*

the probit, logit and the complimentary loglog link respectively. The third, the so-called spread index lives up to its name and appears the most pertinent for inferring whether a fire becomes large. Turning to the question of the efficiency of the three parametric models, when their estimated MIs are compared with those of the nonparametric, they all appear to have performed reasonably. When focus is on the spread index, the complimentary loglog link looks the better. The dashed line in the final panel represents the approximate 95% point of the null distribution. The MIs for the parametric models are estimated via expression (3.19).

4.4 Discussion of the examples

A range of questions motivated the work carried out. The first example was a comparative study. The second involved model assessment. The third was concerned with both prediction and the efficiency of some specific parametric models.

Analyses might have been carried out using second-order moments; hypotheses of dependence have been examined against a much broader class of possibilities. Further the efficiencies of some parametric models have been examined.

5 Discussion and summary

Mutual information is a concept extending correlation, substituting for ρ^2 and R^2. It has a simple definition and a variety of uses.
Conclusions such as

"The hypothesis of independence is rejected."

become

"The estimated strength of dependence is $\hat{MI}$."

The mutual information provides another motivation for the use of ρ^2 in Gaussian case and for G^2 in the contingency table case. The efficiency of an estimate may be studied by considering parametric and nonparametric estimates as in Example 3.

There are some operational questions. Various estimates of MI have been proposed. Their practical properties need to be studied, in particular bias. Simulation studies can provide some guidance.

The mutual information is 'just" a non-negative number. In the examples it seemed that functional forms were to be preferred - MI as a function of country, or of lag, or of index, or of speed. Such thoughts can suggest new parameters for consideration.

The analysis is not complete for once a large value of the MI has been found in many cases one needs to look for an expression of the cause of the relationship, i.e. a model.

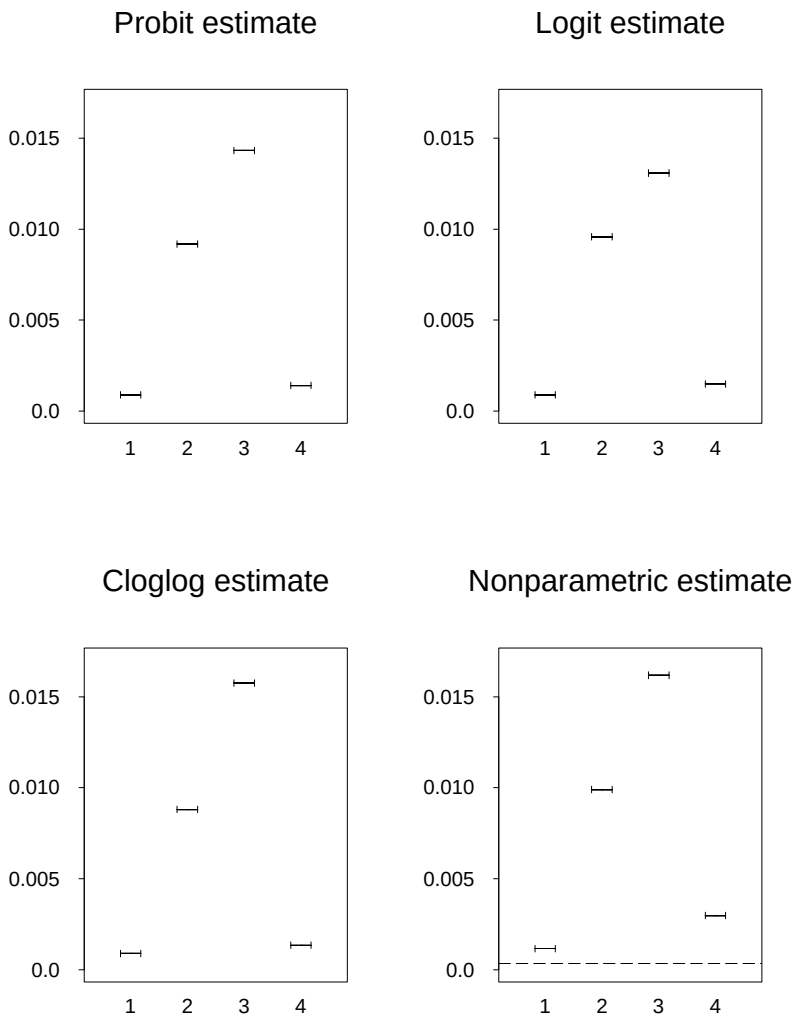

Figure 3 *The estimated MI as a function of four fire indices used in practice. The problem is that of infering which of these indices is most strongly associated with a fire becoming large.*

There are lots of problems to be worked upon. These include practical aspects of extensions to X *in* R^p and Y *in* R^q, higher-order analogs, robust/resistant variants for example based on M-estimates of θ.

There are other measures of independence and entropies, see Fernandes (2000), Hong and White (2000).

Joe's measure

$$1 \; - \; exp\{-2I_{XY}\}$$

has been mentioned. Nagelkirke (1991) proposes the use of an expression like this with $2I_{XY}$ replaced by the deviance. The discussion around Theorem 2 suggests that this may not be a reasonable quantity generally for the null estimate's distribution needs to be considered under the full distribution, not just the null.

Appendix A

A single variable problem is considered to begin.

Let V be a random variable with distribution depending on a finite dimensional parameter θ. Consider the problem of estimating

$$\Psi(\theta) \; = \; E\{g(V|\theta)\}$$

for some measureable function g. Assume that for given θ this expected value can be approximated numerically arbitrarily closely. (There is no problem in the finite discrete case.)

Assuming that the derivative involved exists, let J_V denote the Fisher information,

$$-E\{\frac{\partial^2 l(V|\theta)}{\partial\theta\partial\theta'}\}$$

evaluated at the point θ_0 where $l(v|\theta)$ denotes the log of the pdf (or the pmf) of the variate V.

Suppose that a sample of values, $\{v_1, ..., v_n\}$, is available and that $\hat{\theta}$ is the maximum likelihood estimate of θ. Consider as an estimate of $\Psi(\theta)$

$$\Psi(\hat{\theta}), \tag{A.1}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ. Large sample properties of (A.1) may be studied via the Taylor approximation

$$\Psi(\hat{\theta}) \; \approx \; \Psi_0 \; + \; \frac{\partial\Psi_0}{\partial\theta}'(\hat{\theta}-\theta_0) \; + \; \frac{1}{2}(\hat{\theta}-\theta_0)'\frac{\partial^2\Psi_0}{\partial\theta\partial\theta'}(\hat{\theta}-\theta_0) \tag{A.2}$$

with θ_0 the true parameter value, $\Psi_0 \; = \; \Psi(\theta_0)$, $\partial\Psi_0/\partial\theta$ is the first derivative evaluated at θ_0, and $\partial^2\Psi_0/\partial\theta\partial\theta'$ is the matrix of second derivatives evaluated at θ_0.

Assumption A.1. *The second derivatives of* Φ *exist and are continuous except in a set whose measure is 0. The matrix* J_V *is nonsingular. Further*

the large sample distribution of $\hat{\theta}$ is normal with mean θ_0 and covariance matrix ${J_V}^{-1}/n$.

Now one has as $n \to \infty$,

Theorem A.1. *Let the true parameter value be θ_0, and suppose that Assumption A.1 holds. Then*
a) In the case that the $\partial\Psi_0/\partial\theta$ is not 0, the variate $\sqrt{n}(\Psi(\hat{\theta}) - \Psi(\theta_0))$ is asymptotically normal with mean 0 and covariance matrix

$$\frac{\partial\Psi_0}{\partial\theta}' {J_V}^{-1} \frac{\partial\Psi_0}{\partial\theta}$$

b) In the case that $\partial\Psi_0/\partial\theta$ is 0, (as it is in the case of independence), the variate $n(\Psi(\hat{\theta}) - \Psi(\theta_0))$ has as large sample distribution that of

$$\frac{1}{2} Z' J_V^{-1/2} \frac{\partial^2\Psi_0}{\partial\theta\partial\theta'} J_V^{-1/2} Z, \tag{A.3}$$

Z being a vector of independent standard normals.

Corollary. *Under Assumption A.1, the estimate (A.1) is consistent.*

Consideration now turns to the mutual information case $V = (X, Y)$. Here

$$\Psi(\theta) = I_{XY}(\theta) = E\left\{ log \frac{p(X,Y|\theta)}{p_X(X|\theta)\ p_Y(Y|\theta)} \right\}. \tag{A.4}$$

Note that because $I_{XY}(\theta)$ is invariant under 1-1 transforms of X and Y, $I_{XY}(\theta)$ will sometimes not depend on all the coordinates of θ, i.e. $\partial I/\partial\theta$ will be of reduced rank.

Assumption A.2. *Derivatives up to order 2 exist. One can interchange the orders of integration and differentiation as necessary. The large sample distribution of the maximum likelihood estimate, $\hat{\theta}$, is normal with mean θ_0 and covariance matrix J_{XY}^{-1}/n.*

Then one has,

Lemma A.1. *Under Assumption A.2 and with Ψ given by (A.4) the gradient $\partial\Psi/\partial\theta$ vanishes in the case that X and Y are independent. Also in that case the Hessian matrix, $\partial^2\Psi/\partial\theta\partial\theta'$, is given by $J_{XY} - J_X - J_Y$, where the J are Fisher information matrices of the distributions $(X, Y), X, Y$ respectively.*

The quantity $J_{XY} - J_X - J_Y$ has an interpretation as the Fisher information re θ in (X, Y) minus that in X and further minus that in Y.

Proof of Lemma A.1. That the gradient vanishes is no surprise since the MI is minimized at independence. Still a proof is given. There is much changing of the order of differentiation and integration.

Consider the case that the random variable (X, Y) is continuous. The other cases follow similarly. Write, with abbreviated notation, $p_X(x)dx$ as p. The quantity in question, (A.4), may be written

$$\int\int p\ log\ p\ -\ \int p_X\ log\ p_X\ -\ \int p_Y\ log\ p_Y$$

with derivative

$$\int\int \frac{\partial p}{\partial \theta}[log\ p\ +\ 1] - \int \frac{\partial p_X}{\partial \theta}[log\ p_X\ +\ 1] - \int \frac{\partial p_Y}{\partial \theta}[log\ p_Y\ +\ 1]. \qquad (A.5)$$

Since $\int\int p,\ \int p_X,\ \int p_Y\ =\ 1$ one has

$$\int\int \frac{\partial p}{\partial \theta},\ \int \frac{\partial p_X}{\partial \theta},\ \int \frac{\partial p_Y}{\partial \theta}\ =\ 0$$

and the +1 terms drop out. Next from $\int p dy\ =\ p_X$

$$\int \frac{\partial p}{\partial \theta}\ =\ \frac{\partial p_X}{\partial \theta} \qquad (A.6)$$

and so

$$\int\int log\ p_X\ \frac{\partial p}{\partial \theta}\ =\ \int log\ p_X\ \frac{\partial p_X}{\partial \theta}. \qquad (A.7)$$

There is a similar result for p_Y. The gradient is thus

$$\int\int \frac{\partial p}{\partial \theta}[log\ p\ -\ log\ p_X\ -\ log\ p_Y] \qquad (A.8)$$

which is 0 at independence as $p\ =\ p_X p_Y$.

Turning to the Hessian, taking $\partial/\partial\theta'$ of (A.5) leads to

$$\int\int \frac{\partial^2 p}{\partial\theta\partial\theta'}[log\ p\ -\ log\ p_X\ -\ log\ p_Y]\ +\ \frac{\partial p}{\partial\theta}[\frac{1}{p}\frac{\partial p'}{\partial\theta}\ -\ \frac{1}{p_X}\frac{\partial {p_X}'}{\partial\theta}\ -\ \frac{1}{p_Y}\frac{\partial {p_Y}'}{\partial\theta}]$$

and from (A.6)

$$\int\int \frac{\partial p}{\partial\theta}\frac{1}{p_X}\frac{\partial {p_X}'}{\partial\theta}\ dxdy\ =\ \int \frac{1}{p_X}\frac{\partial p_X}{\partial\theta}\frac{\partial {p_X}'}{\partial\theta}dx$$

So when $p\ =\ p_X p_Y$

$$\frac{\partial^2\Psi}{\partial\theta\partial\theta}\ =\ \int\int[\frac{1}{p}\frac{\partial p}{\partial\theta}\frac{\partial p'}{\partial\theta}\ -\ \frac{1}{p_X}\frac{\partial p_X}{\partial\theta}\frac{\partial p'_X}{\partial\theta}\ -\ \frac{1}{p_Y}\frac{\partial p_Y}{\partial\theta}\frac{\partial p'_Y}{\partial\theta}]$$

i.e.

$$J_{XY} - J_X - J_Y$$

as claimed.

Proof of Theorem A.1. Both parts follow from the representation (A.3) and Corollary 3 of Mann and Wald (1943).

Proof of Theorem 3.1. Part a) follows directly from Theorem A.1 part a).

Consider next part b). In the case of independence, following Lemma A.1, the estimate $nI_{XY}(\hat{\theta})$ is asymptotically distributed as

$$\frac{1}{2}Z' J_{XY}^{-1/2}[J_{XY} - J_X - J_Y]J_{XY}^{-1/2}Z, \qquad (A.9)$$

where Z is a vector of independent standard normals. In the case that the inner matrix of (A.9) is idempotent the large sample distribution of $I_{XY}(\hat{\theta})$ is

$$\chi^2_\nu \ / \ 2n$$

with ν = the trace of $J_{XY}^{-1/2}[J_{XY} - J_X - J_Y]J_{XY}^{-1/2}$.

Assumption A.3. *Suppose that Assumption A.2 holds and that θ has been parametrized as (ϕ, ψ) and that the marginals of X and Y only depend on ϕ.*

Proof of Theorem 3.2.

In the case that X and Y are independent the asymptotic distribution of (3.19) is $\chi^2_\nu/2n$ with $\nu = dim(\psi)$. In the case that they are not

$$expression\ (9) \ \rightarrow \ E\{log\ p(X,Y|\theta)\} \ - \ E\{log\ p_X(X|\phi_*)p_Y(Y|\phi_*)\}$$

in probability where ϕ_* maximizes

$$E\{log\ p_X(X|\phi)p_Y(Y|\phi)\}$$

and one has the stated theorem.

Acknowledgements

The work was supported by the NSG Grant DMS-0203921. C.W.J. Granger, A. Guha, H. Joe, B. Jorgensen, H. Preisler, J.P. Segundo, A. Villa, H. White either made helpful comments or provided data or did both. I thank them all. Part of the material was presented in the 2002 Parzen Prize Lecture and part at the 2003 Regression School in Conservatoria.

(Received May, 2004. Accepted September, 2004.)

References

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Information Theory*, 267-281.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automatic Control*, **AC-19**, 716-723.

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

Beirlant, J., Dudewicz, E. J., Gyorfi, L. and van der Meulen, E. C. (1997). Nonparametric entropy estimation: an overview. *International J. Math. Sci.*, **6**, 17-39.

Brillinger, D. R., Preisler, H. K. and Benoit, J. W. (2003). Risk assessment: a forest fire example. *Science and Statistics.* Lecture Notes in Statistics, **40**, IMS 176-196.

Christensen, R. (1997). *Log-linear Models and Logistic Regression.* New York: Springer.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory.* New York: John Wiley.

Fernandes, M. (2000). Nonparametric entropy-based tests of independence between stochastic processes. Preprint. Rio de Janeiro: Fundação Getúlio Vargas.

Granger, C. W. J. and J-L. Lin, J-L. (1994). Using the mutual information coefficient to identify lags in nonlinear models. *J. Time Series Anal.*, **15**, 371-384.

Hall, P. and Morton, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.*, **45**, 69-88.

Hong, Y. and White, H. (2000). Asymptotic distribution theory for nonparametric entropy measures of serial dependence. Preprint. Cornell University: Department of Economics and San Diego: University of California.

Hsu, P. L. (1941). On the limiting distribution of the canonical correlation. *Biometrika*, **32**, 38-45.

Joe, H. (1989a). Relative entropy measures of multivariate dependence. *J. American Statistical Association*, **84**, 157-164.

Joe, H. (1989b). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.*, **41**, 683-697.

Kendall, M. G. and Stuart, A. (1966). *The Advanced Theory of Statistics.* London: Griffin.

Lee, A. J. (1997). Modelling scores in the Premier League: is Manchester United *really* the best? *Chance*, 15-19.

Li, W. (1990). Mutual information functions versus correlation functions. *J. Statistical Physics*, **60**, 823-837.

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. *Ann. Math. Statist.*, **14**, 217-226.

Miller, G. A. (1955). Note on the bias of information estimates. *Information Theory in Psychology*, 95–100. Illinois, Glencoe.

Mitra, S. K. (1958). On the limiting power of the frequency chi-square test. *Ann. Math. Statsit.*, **29**, 1221-1233.

Moddemeijer, R. (1989). On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, **16**, 233-248.

Moddemeijer, R. (1999). A statistic to estimate the variance of the histogram-based mutual information estimator based on dependent pairs of observations. *Signal Processing*, **75**, 51-63.

Moddemeijer, R. (2000). The distribution of entropy estimators based on maximum mean likelihood. *21st Symposium on Information Theory in Benelux* (Ed. J. Biemond).

Parzen, E. (1983). Time series model identification. *Studies in Econometrics, Time Series, and Multivariate Statistics* (Eds. S. Karlin, T. Amemiya and L. A. Goodman). New York: Academic, 279-298.

Preisler, H. K., Brillinger, D. R., Burgan, R. E. and Benoit, J. W. (2004). Probability based models for the estimation of wildfire risk. *Int. J. Wildfire Research*, **13**, 133-142.

Robinson, P. M. (1991). Consistent nonparametric entropy-based testing. *Rev. Econ. Studies*, **58**, 437-453.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.*, **27**, 379-423 and 623-656.

Skaug, H. J. and Tjostheim, D. (1993). Nonparametric tests of serial independence. *Athens Conference on Applied Probability and Time Series* II. Lecture Notes in Statistics, **115**, 363–378. New York: Springer.

Viola, P. (1995). *Alignment by Maximization of Mutual Information.* PhD thesis, Massachusetts Institute of Technology.

Woodfield, T. J. (1982). *Statistical Modelling of Bivariate Data.* Ph.D. Thesis, Institute of Statistics, Texas A & M University.

David R. Brillinger
Department of Statistics
University of California
Berkeley, California, USA, 94720-3860.
E-mail: brill@stat.berkeley.edu