Distribution of Mutual Information

Marcus Hutter
IDSIA, Galleria 2, CH-6928 Manno-Lugano, Switzerland
marcus@idsia.ch *http://www.idsia.ch/~marcus*

Abstract

The mutual information of two random variables $\imath$ and $\jmath$ with joint probabilities $\{\pi_{ij}\}$ is commonly used in learning Bayesian nets as well as in many other fields. The chances π_{ij} are usually estimated by the empirical sampling frequency n_{ij}/n leading to a point estimate $I(n_{ij}/n)$ for the mutual information. To answer questions like "is $I(n_{ij}/n)$ consistent with zero?" or "what is the probability that the true mutual information is much larger than the point estimate?" one has to go beyond the point estimate. In the Bayesian framework one can answer these questions by utilizing a (second order) prior distribution $p(\pi)$ comprising prior information about π. From the prior $p(\pi)$ one can compute the posterior $p(\pi|\mathbf{n})$, from which the distribution $p(I|\mathbf{n})$ of the mutual information can be calculated. We derive reliable and quickly computable approximations for $p(I|\mathbf{n})$. We concentrate on the mean, variance, skewness, and kurtosis, and non-informative priors. For the mean we also give an exact expression. Numerical issues and the range of validity are discussed.

1 Introduction

The mutual information I (also called cross entropy) is a widely used information theoretic measure for the stochastic dependency of random variables [CT91, Soo00]. It is used, for instance, in learning Bayesian nets [Bun96, Hec98], where stochastically dependent nodes shall be connected. The mutual information defined in (1) can be computed if the joint probabilities $\{\pi_{ij}\}$ of the two random variables $\imath$ and $\jmath$ are known. The standard procedure in the common case of unknown chances π_{ij} is to use the sample frequency estimates $\frac{n_{ij}}{n}$ instead, as if they were precisely known probabilities; but this is not always appropriate. Furthermore, the point estimate $I(\frac{n_{ij}}{n})$ gives no clue about the reliability of the value if the sample size n is finite. For instance, for independent $\imath$ and $\jmath$, $I(\pi)=0$ but $I(\frac{n_{ij}}{n})=O(n^{-1/2})$ due to noise in the data. The criterion for judging dependency is how many standard deviations $I(\frac{n_{ij}}{n})$ is away from zero. In [KJ96, Kle99] the probability that the true $I(\pi)$ is greater than a given threshold has been used to construct Bayesian nets. In the Bayesian framework one can answer these questions by utilizing a (second order)

prior distribution $p(\pi)$,which takes account of any impreciseness about π. From the prior $p(\pi)$ one can compute the posterior $p(\pi|\mathbf{n})$, from which the distribution $p(I|\mathbf{n})$ of the mutual information can be obtained.

The objective of this work is to derive reliable and quickly computable analytical expressions for $p(I|\mathbf{n})$. Section 2 introduces the mutual information distribution, Section 3 discusses some results in advance before delving into the derivation. Since the central limit theorem ensures that $p(I|\mathbf{n})$ converges to a Gaussian distribution a good starting point is to compute the mean and variance of $p(I|\mathbf{n})$. In section 4 we relate the mean and variance to the covariance structure of $p(\pi|\mathbf{n})$. Most non-informative priors lead to a Dirichlet posterior. An exact expression for the mean (Section 6) and approximate expressions for the variance (Sections 5) are given for the Dirichlet distribution. More accurate estimates of the variance and higher central moments are derived in Section 7, which lead to good approximations of $p(I|\mathbf{n})$ even for small sample sizes. We show that the expressions obtained in [KJ96, Kle99] by heuristic numerical methods are incorrect. Numerical issues and the range of validity are briefly discussed in section 8.

2 Mutual Information Distribution

We consider discrete random variables $\imath \in \{1,...,r\}$ and $\jmath \in \{1,...,s\}$ and an i.i.d. random process with samples $(i,j) \in \{1,...,r\} \times \{1,...,s\}$ drawn with joint probability π_{ij}. An important measure of the stochastic dependence of $\imath$ and $\jmath$ is the mutual information

$$I(\pi) \;=\; \sum_{i=1}^{r}\sum_{j=1}^{s} \pi_{ij} \log \frac{\pi_{ij}}{\pi_{i+}\pi_{+j}} \;=\; \sum_{ij} \pi_{ij} \log \pi_{ij} - \sum_{i} \pi_{i+} \log \pi_{i+} - \sum_{j} \pi_{+j} \log \pi_{+j}. \quad (1)$$

log denotes the natural logarithm and $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are marginal probabilities. Often one does not know the probabilities π_{ij} exactly, but one has a sample set with n_{ij} outcomes of pair (i,j). The frequency $\hat{\pi}_{ij} := \frac{n_{ij}}{n}$ may be used as a first estimate of the unknown probabilities. $n := \sum_{ij} n_{ij}$ is the total sample size. This leads to a point (frequency) estimate $I(\hat{\pi}) = \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}}$ for the mutual information (per sample).

Unfortunately the point estimation $I(\hat{\pi})$ gives no information about its accuracy. In the Bayesian approach to this problem one assumes a prior (second order) probability density $p(\pi)$ for the unknown probabilities π_{ij} on the probability simplex. From this one can compute the posterior distribution $p(\pi|\mathbf{n}) \propto p(\pi) \prod_{ij} \pi_{ij}^{n_{ij}}$ (the n_{ij} are multinomially distributed). This allows to compute the posterior probability density of the mutual information.[1]

$$p(I|\mathbf{n}) \;=\; \int \delta(I(\pi) - I)\, p(\pi|\mathbf{n})\, d^{rs}\pi \quad (2)$$

[2]The $\delta()$ distribution restricts the integral to π for which $I(\pi) = I$. For large sam-

[1] $I(\pi)$ denotes the mutual information for the specific chances π, whereas I in the context above is just some non-negative real number. I will also denote the mutual information *random variable* in the expectation $E[I]$ and variance $\mathrm{Var}[I]$. Expectaions are *always* w.r.t. to the posterior distribution $p(\pi|\mathbf{n})$.

[2]Since $0 \leq I(\pi) \leq I_{max}$ with sharp upper bound $I_{max} := \min\{\log r, \log s\}$, the integral may be restricted to $\int_0^{I_{max}}$, which shows that the domain of $p(I|\mathbf{n})$ is $[0, I_{max}]$.

ple size $n \to \infty$, $p(\pi|\mathbf{n})$ is strongly peaked around $\pi = \hat{\pi}$ and $p(I|\mathbf{n})$ gets strongly peaked around the frequency estimate $I = I(\hat{\pi})$. The mean $E[I] = \int_0^\infty I p(I|\mathbf{n})\, dI = \int I(\pi) p(\pi|\mathbf{n}) d^{rs}\pi$ and the variance $\mathrm{Var}[I] = E[(I - E[I])^2] = E[I^2] - E[I]^2$ are of central interest.

3 Results for I under the Dirichlet P(oste)rior

Most[3] non-informative priors for $p(\pi)$ lead to a Dirichlet posterior distribution $p(\pi|\mathbf{n}) \propto \prod_{ij} \pi_{ij}^{n_{ij}-1}$ with interpretation $n_{ij} = n'_{ij} + n''_{ij}$, where n'_{ij} are the number of samples (i,j), and n''_{ij} comprises prior information (1 for the uniform prior, $\frac{1}{2}$ for Jeffreys' prior, 0 for Haldane's prior, $\frac{1}{rs}$ for Perks' prior [GCSR95]). In principle this allows to compute the posterior density $p(I|\mathbf{n})$ of the mutual information. In sections 4 and 5 we expand the mean and variance in terms of n^{-1}:

$$\begin{aligned} E[I] &= \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}} + \frac{(r-1)(s-1)}{2n} + O(n^{-2}), \qquad (3) \\ \mathrm{Var}[I] &= \frac{1}{n} \sum_{ij} \frac{n_{ij}}{n} \left(\log \frac{n_{ij} n}{n_{i+} n_{+j}} \right)^2 - \frac{1}{n} \left(\sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}} \right)^2 + O(n^{-2}). \end{aligned}$$

The first term for the mean is just the point estimate $I(\hat{\pi})$. The second term is a small correction if $n \gg r \cdot s$. Kleiter [KJ96, Kle99] determined the correction by Monte Carlo studies as $\min\{\frac{r-1}{2n}, \frac{s-1}{2n}\}$. This is wrong unless s or r are 2. The expression $2E[I]/n$ they determined for the variance has a completely different structure than ours. Note that the mean is lower bounded by $\frac{const.}{n} + O(n^{-2})$, which is strictly positive for large, but finite sample sizes, even if $\imath$ and $\jmath$ are statistically independent and independence is perfectly represented in the data ($I(\hat{\pi}) = 0$). On the other hand, in this case, the standard deviation $\sigma = \sqrt{\mathrm{Var}(I)} \sim \frac{1}{n} \sim E[I]$ correctly indicates that the mean is still consistent with zero.

Our approximations (3) for the mean and variance are good if $\frac{r \cdot s}{n}$ is small. The central limit theorem ensures that $p(I|\mathbf{n})$ converges to a Gaussian distribution with mean $E[I]$ and variance $\mathrm{Var}[I]$. Since I is non-negative it is more appropriate to approximate $p(I|\pi)$ as a Gamma (= scaled χ^2) or log-normal distribution with mean $E[I]$ and variance $\mathrm{Var}[I]$, which is of course also asymptotically correct.

A systematic expansion in n^{-1} of the mean, variance, and higher moments is possible but gets arbitrarily cumbersome. The $O(n^{-2})$ terms for the variance and leading order terms for the skewness and kurtosis are given in Section 7. For the mean it is possible to give an exact expression

$$E[I] = \frac{1}{n} \sum_{ij} n_{ij} [\psi(n_{ij}+1) - \psi(n_{i+}+1) - \psi(n_{+j}+1) + \psi(n+1)] \qquad (4)$$

with $\psi(n+1) = -\gamma + \sum_{k=1}^{n} \frac{1}{k} = \log n + O(\frac{1}{n})$ for integer n. See Section 6 for details and more general expressions for ψ for non-integer arguments.

There may be other prior information available which cannot be comprised in a Dirichlet distribution. In this general case, the mean and variance of I can still be

[3] But not all priors which one can argue to be non-informative lead to Dirichlet posteriors. Brand [Bra99] (and others), for instance, advocate the entropic prior $p(\pi) \propto e^{-H(\pi)}$.

related to the covariance structure of $p(\pi|\mathbf{n})$, which will be done in the following Section.

4 Approximation of Expectation and Variance of I

In the following let $\hat{\pi}_{ij} := E[\pi_{ij}]$. Since $p(\pi|\mathbf{n})$ is strongly peaked around $\pi = \hat{\pi}$ for large n we may expand $I(\pi)$ around $\hat{\pi}$ in the integrals for the mean and the variance. With $\Delta_{ij} := \pi_{ij} - \hat{\pi}_{ij}$ and using $\sum_{ij} \pi_{ij} = 1 = \sum_{ij} \hat{\pi}_{ij}$ we get for the expansion of (1)

$$I(\pi) \;=\; I(\hat{\pi}) + \sum_{ij} \log\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}\right)\Delta_{ij} + \sum_{ij}\frac{\Delta_{ij}^2}{2\hat{\pi}_{ij}} - \sum_i \frac{\Delta_{i+}^2}{2\hat{\pi}_{i+}} - \sum_j \frac{\Delta_{+j}^2}{2\hat{\pi}_{+j}} + O(\Delta^3). \quad (5)$$

Taking the expectation, the linear term $E[\Delta_{ij}] = 0$ drops out. The quadratic terms $E[\Delta_{ij}\Delta_{kl}] = \mathrm{Cov}(\pi_{ij}, \pi_{kl})$ are the covariance of π under distribution $p(\pi|\mathbf{n})$ and are proportional to n^{-1}. It can be shown that $E[\Delta^3] \sim n^{-2}$ (see Section 7).

$$E[I] \;=\; I(\hat{\pi}) + \frac{1}{2}\sum_{ijkl}\left(\frac{\delta_{ik}\delta_{jl}}{\hat{\pi}_{ij}} - \frac{\delta_{ik}}{\hat{\pi}_{i+}} - \frac{\delta_{jl}}{\hat{\pi}_{+j}}\right)\mathrm{Cov}(\pi_{ij}, \pi_{kl}) + O(n^{-2}). \quad (6)$$

The Kronecker delta δ_{ij} is 1 for $i = j$ and 0 otherwise. The variance of I in leading order in n^{-1} is

$$\begin{aligned}\mathrm{Var}\, I(\pi) \;&=\; E[(I - E[I])^2] \;\stackrel{+}{=}\; E\left[\left(\sum_{ij}\log\left(\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}\right)\Delta_{ij}\right)^2\right] \;= \\ &=\; \sum_{ijkl}\log\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}\log\frac{\hat{\pi}_{kl}}{\hat{\pi}_{k+}\hat{\pi}_{+l}}\mathrm{Cov}(\pi_{ij}, \pi_{kl}), \end{aligned} \quad (7)$$

where $\stackrel{+}{=}$ means $=$ up to terms of order n^{-2}. So the leading order variance and the leading and next to leading order mean of the mutual information $I(\pi)$ can be expressed in terms of the covariance of π under the posterior distribution $p(\pi|\mathbf{n})$.

5 The Second Order Dirichlet Distribution

Noninformative priors for $p(\pi)$ are commonly used if no additional prior information is available. Many non-informative choices (uniform, Jeffreys', Haldane's, Perks', ... prior) lead to a Dirichlet posterior distribution:

$$\begin{aligned} p(\pi|\mathbf{n}) \;&=\; \frac{1}{N(\mathbf{n})}\prod_{ij}\pi_{ij}^{n_{ij}-1}\delta(\pi_{++} - 1) \quad \text{with normalization} \\ N(\mathbf{n}) \;&=\; \int\prod_{ij}\pi_{ij}^{n_{ij}-1}\delta(\pi_{++} - 1)d^{rs}\pi \;=\; \frac{\prod_{ij}\Gamma(n_{ij})}{\Gamma(n)}, \end{aligned} \quad (8)$$

where Γ is the Gamma function, and $n_{ij} = n'_{ij} + n''_{ij}$, where n'_{ij} are the number of samples (i,j), and n''_{ij} comprises prior information (1 for the uniform prior, $\frac{1}{2}$ for Jeffreys' prior, 0 for Haldane's prior, $\frac{1}{rs}$ for Perks' prior). Mean and covariance of $p(\pi|\mathbf{n})$ are

$$\hat{\pi}_{ij} := E[\pi_{ij}] = \frac{n_{ij}}{n}, \quad \mathrm{Cov}(\pi_{ij}, \pi_{kl}) = \frac{1}{n+1}(\hat{\pi}_{ij}\delta_{ik}\delta_{jl} - \hat{\pi}_{ij}\hat{\pi}_{kl}) \quad (9)$$

Inserting this into (6) and (7) we get after some algebra for the mean and variance of the mutual information $I(\pi)$ up to terms of order n^{-2}:

$$E[I] = J + \frac{(r-1)(s-1)}{2(n+1)} + O(n^{-2}), \tag{10}$$

$$\mathrm{Var}[I] = \frac{1}{n+1}(K - J^2) + O(n^{-2}), \tag{11}$$

$$J := \sum_{ij} \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}} = I(\hat{\pi}), \tag{12}$$

$$K := \sum_{ij} \frac{n_{ij}}{n} \left(\log \frac{n_{ij} n}{n_{i+} n_{+j}} \right)^2. \tag{13}$$

J and K (and L, M, P, Q defined later) depend on $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$ only, i.e. are $O(1)$ in $\mathbf{n}$. Strictly speaking we should expand $\frac{1}{n+1} = \frac{1}{n} + O(n^{-2})$, i.e. drop the $+1$, but the exact expression (9) for the covariance suggests to keep the $+1$. We compared both versions with the exact values (from Monte-Carlo simulations) for various parameters π. In most cases the expansion in $\frac{1}{n+1}$ was more accurate, so we suggest to use this variant.

6 Exact Value for $E[I]$

It is possible to get an exact expression for the mean mutual information $E[I]$ under the Dirichlet distribution. By noting that $x \log x = \frac{d}{d\beta} x^\beta |_{\beta=1}$, $(x = \{\pi_{ij}, \pi_{i+}, \pi_{+j}\})$, one can replace the logarithms in the last expression of (1) by powers. From (8) we see that $E[(\pi_{ij})^\beta] = \frac{\Gamma(n_{ij}+\beta)\Gamma(n)}{\Gamma(n_{ij})\Gamma(n+\beta)}$. Taking the derivative and setting $\beta = 1$ we get

$$E[\pi_{ij} \log \pi_{ij}] = \frac{d}{d\beta} E[(\pi_{ij})^\beta]_{\beta=1} = \frac{1}{n} \sum_{ij} n_{ij} [\psi(n_{ij}+1) - \psi(n+1)].$$

The ψ function has the following properties (see [AS74] for details)

$$\psi(z) = \frac{d \log \Gamma(z)}{dz} = \frac{\Gamma'(z)}{\Gamma(z)}, \quad \psi(z+1) = \log z + \frac{1}{2z} - \frac{1}{12z^2} + O(\frac{1}{z^4}),$$

$$\psi(n) = -\gamma + \sum_{k=1}^{n-1} \frac{1}{k}, \quad \psi(n+\tfrac{1}{2}) = -\gamma + 2\log 2 + 2 \sum_{k=1}^{n} \frac{1}{2k-1}. \tag{14}$$

The value of the Euler constant γ is irrelevant here, since it cancels out. Since the marginal distributions of π_{i+} and π_{+j} are also Dirichlet (with parameters n_{i+} and n_{+j}) we get similarly

$$E[\pi_{i+} \log \pi_{i+}] = \frac{1}{n} \sum_{i} n_{i+} [\psi(n_{i+}+1) - \psi(n+1)],$$

$$E[\pi_{+j} \log \pi_{+j}] = \frac{1}{n} \sum_{j} n_{+j} [\psi(n_{+j}+1) - \psi(n+1)].$$

Inserting this into (1) and rearranging terms we get the exact expression[4]

$$E[I] = \frac{1}{n} \sum_{ij} n_{ij} [\psi(n_{ij}+1) - \psi(n_{i+}+1) - \psi(n_{+j}+1) + \psi(n+1)] \tag{15}$$

[4] This expression has independently been derived in [WW93].

For large sample sizes, $\psi(z+1) \approx \log z$ and (15) approaches the frequency estimate $I(\hat{\pi})$ as it should be. Inserting the expansion $\psi(z+1) = \log z + \frac{1}{2z} + ...$ into (15) we also get the correction term $\frac{(r-1)(s-1)}{2n}$ of (3).

The presented method (with some refinements) may also be used to determine an exact expression for the variance of $I(\pi)$. All but one term can be expressed in terms of Gamma functions. The final result after differentiating w.r.t. β_1 and β_2 can be represented in terms of ψ and its derivative ψ'. The mixed term $E[(\pi_{i+})^{\beta_1}(\pi_{+j})^{\beta_2}]$ is more complicated and involves confluent hypergeometric functions, which limits its practical use [WW93].

7 Generalizations

A systematic expansion of all moments of $p(I|\mathbf{n})$ to arbitrary order in n^{-1} is possible, but gets soon quite cumbersome. For the mean we already gave an exact expression (15), so we concentrate here on the variance, skewness and the kurtosis of $p(I|\mathbf{n})$. The 3^{rd} and 4^{th} central moments of π under the Dirichlet distribution are

$$E[\Delta_a\Delta_b\Delta_c] = \frac{2}{(n+1)(n+2)}[2\hat{\pi}_a\hat{\pi}_b\hat{\pi}_c - \hat{\pi}_a\hat{\pi}_b\delta_{bc} - \hat{\pi}_b\hat{\pi}_c\delta_{ca} - \hat{\pi}_c\hat{\pi}_a\delta_{ab} + \hat{\pi}_a\delta_{ab}\delta_{bc}] \tag{16}$$

$$\begin{aligned} E[\Delta_a\Delta_b\Delta_c\Delta_d] = \frac{1}{n^2}[&3\hat{\pi}_a\hat{\pi}_b\hat{\pi}_c\hat{\pi}_d - \hat{\pi}_c\hat{\pi}_d\hat{\pi}_a\delta_{ab} - \hat{\pi}_b\hat{\pi}_d\hat{\pi}_a\delta_{ac} - \hat{\pi}_b\hat{\pi}_c\hat{\pi}_a\delta_{ad} \\ &- \hat{\pi}_a\hat{\pi}_d\hat{\pi}_b\delta_{bc} - \hat{\pi}_a\hat{\pi}_c\hat{\pi}_b\delta_{bd} - \hat{\pi}_a\hat{\pi}_b\hat{\pi}_c\delta_{cd} \\ &+ \hat{\pi}_a\hat{\pi}_c\delta_{ab}\delta_{cd} + \hat{\pi}_a\hat{\pi}_b\delta_{ac}\delta_{bd} + \hat{\pi}_a\hat{\pi}_b\delta_{ad}\delta_{bc}] + O(n^{-3}) \end{aligned} \tag{17}$$

with $a=ij$, $b=kl$,... $\in \{1,...,r\}\times\{1,...,s\}$ being double indices, $\delta_{ab} = \delta_{ik}\delta_{jl}$,... $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$. Expanding $\Delta^k = (\pi - \hat{\pi})^k$ in $E[\Delta_a\Delta_b...]$ leads to expressions containing $E[\pi_a\pi_b...]$, which can be computed by a case analysis of all combinations of equal/unequal indices a,b,c,... using (8). Many terms cancel leading to the above expressions. They allow to compute the order n^{-2} term of the variance of $I(\pi)$. Again, inspection of (16) suggests to expand in $[(n+1)(n+2)]^{-1}$, rather than in n^{-2}. The variance in leading and next to leading order is

$$\mathrm{Var}[I] = \frac{K-J^2}{n+1} + \frac{M + (r-1)(s-1)(\frac{1}{2} - J) - Q}{(n+1)(n+2)} + O(n^{-3}) \tag{18}$$

$$M := \sum_{ij}\left(\frac{1}{n_{ij}} - \frac{1}{n_{i+}} - \frac{1}{n_{+j}} + \frac{1}{n}\right) n_{ij}\log\frac{n_{ij}n}{n_{i+}n_{+j}}, \tag{19}$$

$$Q := 1 - \sum_{ij}\frac{n_{ij}^2}{n_{i+}n_{+j}}. \tag{20}$$

J and K are defined in (12) and (13). Note that the first term $\frac{K-J^2}{n+1}$ also contains second order terms when expanded in n^{-1}. The leading order terms for the 3^{rd} and 4^{th} central moments of $p(I|\mathbf{n})$ are

$$E[(I - E[I])^3] = \frac{2}{n^2}[2J^3 - 3KJ + L] + \frac{3}{n^2}[K + J^2 - P] + O(n^{-3}),$$

$$L := \sum_{ij}\frac{n_{ij}}{n}\left(\log\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^3, \quad P := \sum_i \frac{nJ_{i+}^2}{n_{i+}} + \sum_j \frac{nJ_{+j}^2}{n_{+j}},$$

$$J_{i+} := \sum_j \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}} \quad , \quad J_{+j} := \sum_i \frac{n_{ij}}{n} \log \frac{n_{ij} n}{n_{i+} n_{+j}},$$

$$E[(I - E[I])^4] = \frac{3}{n^2}[K - J^2]^2 + O(n^{-3}),$$

from which the skewness and kurtosis can be obtained by dividing by $\mathrm{Var}[I]^{3/2}$ and $\mathrm{Var}[I]^2$ respectively. One can see that the skewness is of order $n^{-1/2}$ and the kurtosis is $3+O(n^{-1})$. Significant deviation of the skewness from 0 or the kurtosis from 3 would indicate a non-Gaussian I. They can be used to get an improved approximation for $p(I|\mathbf{n})$ by making, for instance, an ansatz

$$p(I|\mathbf{n}) \propto (1 + \tilde{b} I + \tilde{c} I^2) \cdot p_0(I|\tilde{\mu}, \tilde{\sigma}^2)$$

and fitting the parameters $\tilde{b}$, $\tilde{c}$, $\tilde{\mu}$, and $\tilde{\sigma}^2$ to the mean, variance, skewness, and kurtosis expressions above. p_0 is the Normal or Gamma distribution (or any other distribution with Gaussian limit). From this, quantiles $p(I > I_*|\mathbf{n}) := \int_{I_*}^{\infty} p(I|\mathbf{n})\, dI$, needed in [KJ96, Kle99], can be computed. A systematic expansion of arbitrarily high moments to arbitrarily high order in n^{-1} leads, in principle, to arbitrarily accurate estimates.

8 Numerics

There are short and fast implementations of ψ. The code of the Gamma function in [PFTV92], for instance, can be modified to compute the ψ function. For integer and half-integer values one may create a lookup table from (14). The needed quantities J, K, L, M, and Q (depending on $\mathbf{n}$) involve a double sum, P only a single sum, and the $r+s$ quantities J_{i+} and J_{+j} also only a single sum. Hence, the computation time for the (central) moments is of the same order $O(r \cdot s)$ as for the point estimate (1). "Exact" values have been obtained for representative choices of π_{ij}, r, s, and n by Monte Carlo simulation. The $\pi_{ij} := x_{ij}/x_{++}$ are Dirichlet distributed, if each x_{ij} follows a Gamma distribution. See [PFTV92] how to sample from a Gamma distribution. The variance has been expanded in $\frac{r \cdot s}{n}$, so the relative error $\frac{\mathrm{Var}[I]_{approx} - \mathrm{Var}[I]_{exact}}{\mathrm{Var}[I]_{exact}}$ of the approximation (11) and (18) are of the order of $\frac{r \cdot s}{n}$ and $(\frac{r \cdot s}{n})^2$ respectively, *if* $\imath$ and $\jmath$ are dependent. If they are independent the leading term (11) drops itself down to order n^{-2} resulting in a reduced relative accuracy $O(\frac{r \cdot s}{n})$ of (18). Comparison with the Monte Carlo values confirmed an accurracy in the range $(\frac{r \cdot s}{n})^{1 \cdots 2}$. The mean (4) is exact. Together with the skewness and kurtosis we have a good description for the distribution of the mutual information $p(I|\mathbf{n})$ for not too small sample bin sizes n_{ij}. We want to conclude with some notes on *useful* accuracy. The hypothetical prior sample sizes $n''_{ij} = \{0, \frac{1}{rs}, \frac{1}{2}, 1\}$ can all be argued to be non-informative [GCSR95]. Since the central moments are expansions in n^{-1}, the next to leading order term can be freely adjusted by adjusting $n''_{ij} \in [0 \ldots 1]$. So one may argue that anything beyond leading order is free to will, and the leading order terms may be regarded as accurate as we can specify our prior knowledge. On the other hand, exact expressions have the advantage of being safe against cancellations. For instance, leading order of $E[I]$ and $E[I^2]$ does not suffice to compute the leading order of $\mathrm{Var}[I]$.

Acknowledgements

I want to thank Ivo Kwee for valuable discussions and Marco Zaffalon for encouraging me to investigate this topic. This work was supported by SNF grant 2000-61847.00 to Jürgen Schmidhuber.

References

[AS74] M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions.* Dover publications, inc., 1974.

[Bra99] M. Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 11(5):1155–1182, 1999.

[Bun96] W. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8:195–210, 1996.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. John Wiley & Sons, New York, NY, USA, 1991.

[GCSR95] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman, 1995.

[Hec98] D. Heckerman. A tutorial on learning with Bayesian networks. *Learnig in Graphical Models*, pages 301–354, 1998.

[KJ96] G. D. Kleiter and R. Jirousek. Learning Bayesian networks under the control of mutual information. *Proceedings of the 6th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-1996)*, pages 985–990, 1996.

[Kle99] G. D. Kleiter. The posterior probability of Bayes nets with strong dependences. *Soft Computing*, 3:162–173, 1999.

[PFTV92] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, Cambridge, second edition, 1992.

[Soo00] E. S. Soofi. Principal information theoretic approaches. *Journal of the American Statistical Association*, 95:1349–1353, 2000.

[WW93] D. R. Wolf and D. H. Wolpert. Estimating functions of distributions from A finite set of samples, part 2: Bayes estimators for mutual information, chi-squared, covariance and other statistics. Technical Report LANL-LA-UR-93-833, Los Alamos National Laboratory, 1993. Also Santa Fe Insitute report SFI-TR-93-07-047.