Distribution of Mutual Information from Complete and Incomplete Data*

Marcus Hutter and **Marco Zaffalon**

IDSIA, Galleria 2, CH-6928 Manno (Lugano), Switzerland
{marcus,zaffalon}@idsia.ch IDSIA-11-02

March 15, 2004

Abstract

Mutual information is widely used, in a descriptive way, to measure the stochastic dependence of categorical random variables. In order to address questions such as the reliability of the descriptive value, one must consider sample-to-population inferential approaches. This paper deals with the posterior distribution of mutual information, as obtained in a Bayesian framework by a second-order Dirichlet prior distribution. The exact analytical expression for the mean, and analytical approximations for the variance, skewness and kurtosis are derived. These approximations have a guaranteed accuracy level of the order $O(n^{-3})$, where n is the sample size. Leading order approximations for the mean and the variance are derived in the case of incomplete samples. The derived analytical expressions allow the distribution of mutual information to be approximated reliably and quickly. In fact, the derived expressions can be computed with the same order of complexity needed for descriptive mutual information. This makes the distribution of mutual information become a concrete alternative to descriptive mutual information in many applications which would benefit from moving to the inductive side. Some of these prospective applications are discussed, and one of them, namely *feature selection*, is shown to perform significantly better when inductive mutual information is used.

Keywords

*Preliminary results have been presented at the conferences NIPS-2001 [Hut02] and UAI-2002 [ZH02] and KI-2003 [HZ03]. This research was supported in parts by the NSF grants 2000-61847 and 2100-067961.

Mutual information, cross entropy, Dirichlet distribution, second order distribution, expectation and variance of mutual information, feature selection, filters, naive Bayes classifier, Bayesian statistics.

1 Introduction

Consider a data set of n observations (or units) jointly categorized according to the random variables $\imath$ and $\jmath$, in $\{1,...,r\}$ and $\{1,...,s\}$, respectively. The observed counts are $\boldsymbol{n} := (n_{11},...,n_{rs})$, with $n := \sum_{ij} n_{ij}$, and the observed relative frequencies are $\hat{\boldsymbol{\pi}} := (\hat{\pi}_{11},...,\hat{\pi}_{rs})$, with $\hat{\pi}_{ij} := n_{ij}/n$. The data $\boldsymbol{n}$ are considered as a sample from a larger population, characterized by the actual chances $\boldsymbol{\pi} := (\pi_{11},...,\pi_{rs})$, which are the population counterparts of $\hat{\boldsymbol{\pi}}$. Both $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$ belong to the rs-dimensional unit simplex.

We consider the statistical problem of analyzing the association between $\imath$ and $\jmath$, given only the data $\boldsymbol{n}$. This problem is often addressed by measuring indices of independence, such as the statistical coefficient ϕ^2 [KS67, pp. 556–561]. In this paper we focus on the index I called *mutual information* (also called *cross entropy* or *information gain*) [Kul68]. This index has gained a growing popularity, especially in the artificial intelligence community. It is used, for instance, in learning *Bayesian networks* [CL68, Pea88, Bun96, Hec98], to connect stochastically dependent nodes; it is used to infer classification trees [Qui93]. It is also used to select *features* for classification problems [DHS01], i.e. to select a subset of variables by which to predict the *class* variable. This is done in the context of a *filter approach* that discards irrelevant features on the basis of low values of mutual information with the class [Lew92, BL97, CHH$^+$02].

Mutual information is widely used in descriptive rather than inductive way. The qualifiers 'descriptive' and 'inductive' are used for models bearing on $\hat{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$, respectively. Accordingly, $\hat{\boldsymbol{\pi}}$ are called *relative frequencies*, and $\boldsymbol{\pi}$ are called *chances*. At descriptive level, variables $\imath$ and $\jmath$ are found to be either independent or dependent, according to the fact that the *empirical* mutual information $I(\hat{\boldsymbol{\pi}})$ is zero or is a positive number. At inductive level, $\imath$ and $\jmath$ are assessed to be either independent or dependent only with some probability, because $I(\boldsymbol{\pi})$ can only be known with some (second order) probability.

The problem with the descriptive approach is that it neglects the variability of the mutual information index with the sample, and this is a potential source of fragility of the induced models. In order to achieve robustness, one must move from the descriptive to the inductive side. This involves regarding the mutual information I as a random variable, with a certain distribution. The distribution allows one to make reliable, probabilistic statements about I.

In order to derive the expression for the distribution of I, we work in the framework of Bayesian statistics. In particular, we use a second order prior distribution $p(\boldsymbol{\pi})$ which takes into account our uncertainty about the chances $\boldsymbol{\pi}$. From the

prior $p(\boldsymbol{\pi})$ and the likelihood we obtain the posterior $p(\boldsymbol{\pi}|\boldsymbol{n})$, of which the posterior distribution $p(I|\boldsymbol{n})$ of the mutual information is a formal consequence.

Although the problem is formally solved, the task is not accomplished yet. In fact, closed-form expressions for the distribution of mutual information are unlikely to be available, and we are left with the concrete problem of using the distribution of mutual information in practice. We address this problem by providing fast analytical approximations to the distribution which have guaranteed levels of accuracy.

We start by computing the mean and variance of $p(I|\boldsymbol{n})$. This is motivated by the central limit theorem that ensures that $p(I|\boldsymbol{n})$ can be well approximated by a Gaussian distribution for large n. Section 2 establishes a general relationship, used throughout the paper, to relate the mean and variance to the covariance structure of $p(\boldsymbol{\pi}|\boldsymbol{n})$. By focusing on the specific covariance structure obtained when the prior over the chances is Dirichlet, we are then lead to $O(n^{-2})$ approximations for the mean and the variance of $p(I|\boldsymbol{n})$. Generalizing the former approach, in Section 3 we report $O(n^{-3})$ approximations for the variance, skewness and kurtosis of $p(I|\boldsymbol{n})$. We also provide an exact expression for the mean in Section 4, and improved tail approximations for extreme quantiles.

By an example, Section 5 shows that the approximated distributions, obtained by fitting some common distributions to the expressions above, compare well to the "exact" one obtained by Monte Carlo sampling also for small sample sizes. Section 5 also discusses the accuracy of the approximations and their computational complexity, which is of the same order of magnitude needed to compute the empirical mutual information. This is an important result for the real application of the distribution of mutual information.

In the same spirit of making the results useful for real applications, and considered that missing data are a pervasive problem of statistical practice, we generalize the framework to the case of incomplete samples in Section 6. We derive $O(n^{-1})$ expressions for the mean and the variance of $p(I|\boldsymbol{n})$, under the common assumption that data are *missing at random* [LR87]. These expressions are in closed form when observations from one variable, either $\imath$ or $\jmath$, are always present, and their complexity is the same of the complete-data case. When observations from both $\imath$ and $\jmath$ can be missing, there are no closed-form expressions in general but we show that the popular expectation-maximization (EM) algorithm [CF74] can be used to compute $O(n^{-1})$ expressions. This is possible as EM converges to the global optimum for the problem under consideration, as we show in Section 6.

We stress that the above results are a significant and novel step to the direction of robustness. To our knowledge, there are only two other works in literature that are close to the work presented here. Kleiter has provided approximations to the mean and the variance of mutual information by heuristic arguments [Kle99], but unfortunately, the approximations are shown to be crude in general (see Section 2). Wolpert and Wolf computed the exact mean of mutual information [WW95, Th.10] and reported the exact variance as an infinite sum; but the latter does not allow a straightforward systematic approximation to be obtained.

In Section 7 we move from the theoretical to the applied side, discussing the potential implications of the distribution of mutual information for real applications. For illustrative purposes, in the following Section 8, we apply the distribution of mutual information to feature selection. We define two new filters based on the distribution of mutual information that generalize the traditional filter based on empirical mutual information [Lew92]. Several experiments on real data sets show that one of the new filters is more effective than the traditional one in the case of sequential learning tasks. This is the case for complete data described in Section 9, as well as incomplete data in Section 10. Concluding remarks are reported in Section 11.

2 Expectation and Variance of Mutual Information

Setup. Consider discrete random variables $\imath \in \{1,...,r\}$ and $\jmath \in \{1,...,s\}$ and an i.i.d. random process with outcome $(i,j) \in \{1,...,r\} \times \{1,...,s\}$ having joint chance π_{ij}. The mutual information is defined by

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

where ln denotes the natural logarithm and $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ are marginal chances. Often the descriptive index $I(\hat{\boldsymbol{\pi}}) = \sum_{ij} \frac{n_{ij}}{n} \ln \frac{n_{ij} n}{n_{i+} n_{+j}}$ is used in the place of the actual mutual information. Unfortunately, the empirical index $I(\hat{\boldsymbol{\pi}})$ carries no information about its accuracy. Especially $I(\hat{\boldsymbol{\pi}}) \neq 0$ can have to origins; a true dependency of the random variables $\imath$ and $\jmath$ or just a fluctuation due to the finite sample size. In the Bayesian approach to this problem one assumes a prior (second order) probability density $p(\boldsymbol{\pi})$ for the unknown chances π_{ij} on the probability simplex. From this one can determine the posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{n}) \propto p(\boldsymbol{\pi}) \prod_{ij} \pi_{ij}^{n_{ij}}$ (the n_{ij} are multinomially distributed). This allows to determine the posterior probability density of the mutual information:[1]

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

The $\delta()$ distribution restricts the integral to $\boldsymbol{\pi}$ for which $I(\boldsymbol{\pi}) = I$. Since $0 \leq I(\boldsymbol{\pi}) \leq I_{max}$ with sharp upper bound $I_{max} := \min\{\ln r, \ln s\}$, the domain of $p(I|\boldsymbol{n})$ is $[0, I_{max}]$, hence integrals over I may be restricted to such interval of the real line.

[1] $I(\boldsymbol{\pi})$ denotes the mutual information for the specific chances $\boldsymbol{\pi}$, 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 $\mathrm{Var}[I]$. Expectations are *always* w.r.t. to the posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{n})$.

For large sample size, $p(\boldsymbol{\pi}|\boldsymbol{n})$ gets strongly peaked around $\boldsymbol{\pi}=\hat{\boldsymbol{\pi}}$ and $p(I|\boldsymbol{n})$ gets strongly peaked around the empirical index $I=I(\hat{\boldsymbol{\pi}})$. The mean $E[I]=\int_0^\infty I\cdot p(I|\boldsymbol{n})\,dI=\int I(\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{n})d^{rs}\boldsymbol{\pi}$ and the variance $\mathrm{Var}[I]=E[(I-E[I])^2]=E[I^2]-E[I]^2$ are of central interest.

General approximation of expectation and variance of I. In the following we (approximately) relate the mean and variance of I to the covariance structure of $p(\boldsymbol{\pi}|\boldsymbol{n})$. Let $\bar{\boldsymbol{\pi}}:=(\bar\pi_{11},...,\bar\pi_{rs})$, with $\bar\pi_{ij}:=E[\pi_{ij}]$. Since $p(\boldsymbol{\pi}|\boldsymbol{n})$ is strongly peaked around $\boldsymbol{\pi}=\hat{\boldsymbol{\pi}}\approx\bar{\boldsymbol{\pi}}$, for large n we may expand $I(\boldsymbol{\pi})$ around $\bar{\boldsymbol{\pi}}$ in the integrals for the mean and the variance. With $\Delta_{ij}:=\pi_{ij}-\bar\pi_{ij}\in[-1,1]$ and using $\sum_{ij}\pi_{ij}=1=\sum_{ij}\bar\pi_{ij}$ we get the following expansion of expression (1):

$$I(\boldsymbol{\pi}) \;=\; I(\bar{\boldsymbol{\pi}})+\sum_{ij}\ln\left(\frac{\bar\pi_{ij}}{\bar\pi_{i+}\bar\pi_{+j}}\right)\Delta_{ij}+\sum_{ij}\frac{\Delta_{ij}^2}{2\bar\pi_{ij}}-\sum_{i}\frac{\Delta_{i+}^2}{2\bar\pi_{i+}}-\sum_{j}\frac{\Delta_{+j}^2}{2\bar\pi_{+j}}+O(\Delta^3), \quad (3)$$

where $O(\Delta^3)$ is bounded by the absolute value of (and Δ^3 is equal to) some homogenous cubic polynomial in the $r\cdot s$ variables Δ_{ij}. Taking the expectation, the linear term $E[\Delta_{ij}]=0$ drops out. The quadratic terms $E[\Delta_{ij}\Delta_{kl}]=\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}]$ are the covariance of $\boldsymbol{\pi}$ under $p(\boldsymbol{\pi}|\boldsymbol{n})$ and they are proportional to n^{-1}. Equation (9) in Section 3 shows that $E[\Delta^3]=O(n^{-2})$, whence

$$E[I] \;=\; I(\bar{\boldsymbol{\pi}})+\frac{1}{2}\sum_{ijkl}\left(\frac{\delta_{ik}\delta_{jl}}{\bar\pi_{ij}}-\frac{\delta_{ik}}{\bar\pi_{i+}}-\frac{\delta_{jl}}{\bar\pi_{+j}}\right)\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}]+O(n^{-2}). \quad (4)$$

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

$$\begin{aligned}\mathrm{Var}[I] \;&=\; E[(I-E[I])^2] \;\simeq\; E\left[\left(\sum_{ij}\ln\left(\frac{\bar\pi_{ij}}{\bar\pi_{i+}\bar\pi_{+j}}\right)\Delta_{ij}\right)^2\right] \;= \\ &=\; \sum_{ijkl}\ln\frac{\bar\pi_{ij}}{\bar\pi_{i+}\bar\pi_{+j}}\ln\frac{\bar\pi_{kl}}{\bar\pi_{k+}\bar\pi_{+l}}\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}], \quad (5)\end{aligned}$$

where $\simeq$ denotes equality up to terms of order n^{-2}. So the leading order term for the variance of mutual information $I(\boldsymbol{\pi})$, and the leading and second leading order term for the mean can be expressed in terms of the covariance of $\boldsymbol{\pi}$ under the posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{n})$.

The (second order) Dirichlet distribution. Noninformative priors $p(\boldsymbol{\pi})$ are commonly used if no explicit prior information is available on $\boldsymbol{\pi}$. Most noninformative priors lead to a Dirichlet posterior distribution $p(\boldsymbol{\pi}|\boldsymbol{n})\propto\prod_{ij}\pi_{ij}^{n_{ij}-1}$ with interpretation[2] $n_{ij}=n'_{ij}+n''_{ij}$, where the n'_{ij} are the number of outcomes (i,j), and

[2]To avoid unnecessary complications we are abusing the notation: n_{ij} is now the sum of real *and* virtual counts, while it formerly denoted the real counts only. In case of Haldane's prior ($n''_{ij}=0$), this change is ineffective.

n''_{ij} comprises prior information Explicit prior knowledge may also be specified by using virtual units, i.e. by n''_{ij}, leading again to a Dirichlet posterior.

The Dirichlet distribution is defined as follows:

$$p(\boldsymbol{\pi}|\boldsymbol{n}) \;=\; \frac{1}{\mathcal{N}(\boldsymbol{n})}\prod_{ij}\pi_{ij}^{n_{ij}-1}\delta(\pi_{++}-1) \quad \text{with normalization}$$

$$\mathcal{N}(\boldsymbol{n}) \;=\; \int\prod_{ij}\pi_{ij}^{n_{ij}-1}\delta(\pi_{++}-1)d^{rs}\boldsymbol{\pi} \;=\; \frac{\prod_{ij}\Gamma(n_{ij})}{\Gamma(n)},$$

where Γ is the Gamma function. Mean and covariance of $p(\boldsymbol{\pi}|\boldsymbol{n})$ are

$$\bar{\pi}_{ij} := E[\pi_{ij}] = \frac{n_{ij}}{n} = \hat{\pi}_{ij}, \quad \mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] = \frac{1}{n+1}(\hat{\pi}_{ij}\delta_{ik}\delta_{jl} - \hat{\pi}_{ij}\hat{\pi}_{kl}). \tag{6}$$

Expectation and variance of I under Dirichlet priors. Inserting (6) into (4) and (5) we get, after some algebra, the mean and variance of the mutual information $I(\boldsymbol{\pi})$ up to terms of order n^{-2}:

$$E[I] \;\simeq\; J \;+\; \frac{(r-1)(s-1)}{2(n+1)}, \qquad J \;:=\; \sum_{ij}\frac{n_{ij}}{n}\ln\frac{n_{ij}n}{n_{i+}n_{+j}} \;=\; I(\hat{\boldsymbol{\pi}}), \tag{7}$$

$$\mathrm{Var}[I] \;\simeq\; \frac{1}{n+1}(K-J^2), \qquad K \;:=\; \sum_{ij}\frac{n_{ij}}{n}\left(\ln\frac{n_{ij}n}{n_{i+}n_{+j}}\right)^2. \tag{8}$$

J and K (and L, M, P, Q defined later) depend on $\hat{\pi}_{ij}$ only, i.e. are $O(1)$ in $\boldsymbol{n}$. Strictly speaking in (7) we should make the expansion $\frac{1}{n+1}=\frac{1}{n}+O(n^{-2})$, i.e. drop the $+1$, but the exact expression (6) for the covariance suggests to keep it. We compared both versions with the "exact" values (from Monte-Carlo simulations) for various parameters $\boldsymbol{\pi}$. In many cases the expansion in $\frac{1}{n+1}$ was more accurate, so we suggest to use this variant.

The first term for the mean is just the descriptive index $I(\hat{\boldsymbol{\pi}})$. The second term is a correction, small when n is much larger than $r\cdot s$. Kleiter [Kle99] determined the correction by Monte Carlo studies as $\min\{\frac{r-1}{2n},\frac{s-1}{2n}\}$. This is only correct if s or r is 2. The expression $2E[I]/n$ he 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 ($I(\hat{\boldsymbol{\pi}})=0$). 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 (where $f\sim g$ means that f and g have the same accuracy, i.e. $f=O(g)$ and $g=O(f)$).

Our approximations for the mean (7) and variance (8) are good if $\frac{r\cdot s}{n}$ is small. For dependent random variables, the central limit theorem ensures that $p(I|\boldsymbol{n})$ 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 $p(I|\boldsymbol{\pi})$ as a Gamma (= scaled χ^2) or a Beta distribution with mean $E[I]$ and variance $\mathrm{Var}[I]$, which are of course also asymptotically correct.

3 Higher Moments and Orders

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

$$E[\Delta_a\Delta_b\Delta_c] \;=\; \frac{2}{(n+1)(n+2)}[2\hat\pi_a\hat\pi_b\hat\pi_c - \hat\pi_a\hat\pi_b\delta_{bc} - \hat\pi_b\hat\pi_c\delta_{ca} - \hat\pi_c\hat\pi_a\delta_{ab} + \hat\pi_a\delta_{ab}\delta_{bc}] \quad (9)$$

$$\begin{aligned} E[\Delta_a\Delta_b\Delta_c\Delta_d] \;=\; \frac{1}{n^2}[&3\hat\pi_a\hat\pi_b\hat\pi_c\hat\pi_d - \hat\pi_c\hat\pi_d\hat\pi_a\delta_{ab} - \hat\pi_b\hat\pi_d\hat\pi_a\delta_{ac} - \hat\pi_b\hat\pi_c\hat\pi_a\delta_{ad} \\ &-\hat\pi_a\hat\pi_d\hat\pi_b\delta_{bc} - \hat\pi_a\hat\pi_c\hat\pi_b\delta_{bd} - \hat\pi_a\hat\pi_b\hat\pi_c\delta_{cd} \\ &+\hat\pi_a\hat\pi_c\delta_{ab}\delta_{cd} + \hat\pi_a\hat\pi_b\delta_{ac}\delta_{bd} + \hat\pi_a\hat\pi_b\delta_{ad}\delta_{bc}] + O(n^{-3}) \end{aligned} \quad (10)$$

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 (6). Many terms cancel out leading to the above expressions. They allow us to compute the order n^{-2} term of the variance of $I(\boldsymbol{\pi})$. Again, inspection of (9) suggests to expand in $[(n{+}1)(n{+}2)]^{-1}$, rather than in n^{-2}. The leading and second leading order terms of the variance are given below,

$$\mathrm{Var}[I] \;=\; \frac{K-J^2}{n+1} + \frac{M+(r-1)(s-1)(\frac{1}{2}-J)-Q}{(n+1)(n+2)} + O(n^{-3}) \quad (11)$$

$$M \;:=\; \sum_{ij}\left(\frac{1}{n_{ij}} - \frac{1}{n_{i+}} - \frac{1}{n_{+j}} + \frac{1}{n}\right) n_{ij}\ln\frac{n_{ij}n}{n_{i+}n_{+j}}, \quad (12)$$

$$Q \;:=\; 1 - \sum_{ij}\frac{n_{ij}^2}{n_{i+}n_{+j}}. \quad (13)$$

J and K are defined in (7) and (8). 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 $p(I|\boldsymbol{n})$ are

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

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

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

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. These expressions can be used to get an improved approximation for $p(I|\boldsymbol{n})$ by making, for instance, an ansatz

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

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 any distribution with Gaussian limit. ¿From this, quantiles $p(I>I_*|\boldsymbol{n}):=\int_{I_*}^{\infty}p(I|\boldsymbol{n})\,dI$, needed later (and in [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 (assuming convergence of the expansion).

4 Further Expressions

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 $x\ln x=\frac{d}{d\beta}x^{\beta}|_{\beta=1}$, $(x=\{\pi_{ij},\pi_{i+},\pi_{+j}\})$, one can replace the logarithms in the last expression of (1) by powers. From (6) 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

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

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

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

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

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

$$\begin{aligned}E[\pi_{i+}\ln\pi_{i+}] &= \frac{1}{n}\sum_i n_{i+}[\psi(n_{i+}+1)-\psi(n+1)],\\ E[\pi_{+j}\ln\pi_{+j}] &= \frac{1}{n}\sum_j n_{+j}[\psi(n_{+j}+1)-\psi(n+1)].\end{aligned}$$

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

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

(This expression has independently been derived in [WW95] in a different way.) For large sample sizes, $\psi(z+1)\approx\ln z$ and (15) approaches the descriptive index $I(\hat{\boldsymbol{\pi}})$ as it should. Inserting the expansion $\psi(z+1)=\ln z+\frac{1}{2z}+...$ into (15) we also get the correction term $\frac{(r-1)(s-1)}{2n}$ of (7).

The presented method (with some refinements) may also be used to determine an exact expression for the variance of $I(\boldsymbol{\pi})$. 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 [WW95].

Large and small I asymptotics. For extreme quantiles $I_*\approx 0$ or $I_*\approx I_{max}$, the accuracy of the derived approximations in the last sections can be poor and it is better to use tail approximations. In the following we briefly sketch how the scaling behavior of $p(I|\boldsymbol{n})$ can be determined.

We observe that $I(\boldsymbol{\pi})$ is small iff $\pi_{\imath\jmath}$ describes near independent random variables $\imath$ and $\jmath$. This suggests the reparametrization $\pi_{ij}=\tilde{\pi}_{i+}\tilde{\pi}_{+j}+\Delta_{ij}$ in the integral (2). In order to make this representation unique and consistent with $\pi_{++}=1$, we have to restrict the $r+s+rs$ degrees of freedom $(\tilde{\pi}_{i+},\tilde{\pi}_{+j},\Delta_{ij})$ to $rs-1$ degrees of freedom by imposing $r+s+1$ constraints, for instance $\sum_i\tilde{\pi}_{i+}=\sum_j\tilde{\pi}_{+j}=1$ and $\Delta_{i+}=\Delta_{+j}=0$ ($\Delta_{++}=0$ occurs twice). Only small $\boldsymbol{\Delta}$ can lead to small $I(\boldsymbol{\pi})$. Hence, for small I we may expand $I(\boldsymbol{\pi})$ in $\boldsymbol{\Delta}$ in expression (2). Inserting $\pi_{ij}=\tilde{\pi}_{i+}\tilde{\pi}_{+j}+\Delta_{ij}$ into (3), we get $I(\tilde{\pi}_{i+}\tilde{\pi}_{+j}+\Delta_{ij})=\boldsymbol{\Delta}^T\boldsymbol{H}(\tilde{\boldsymbol{\pi}})\boldsymbol{\Delta}+O(\Delta^3)$ with $H_{(ij)(kl)}=\frac{1}{2}[\delta_{ik}\delta_{jl}/\tilde{\pi}_{ij}-\delta_{ik}/\tilde{\pi}_{i+}-\delta_{jl}/\tilde{\pi}_{+j}]$ (cf. (4)) and $\boldsymbol{H}$ and $\boldsymbol{\Delta}$ interpreted as rs-dimensional matrix and vector. $\boldsymbol{\Delta}^T\boldsymbol{H}(\tilde{\boldsymbol{\pi}})\boldsymbol{\Delta}=I$ describes an rs-dimensional ellipsoid of linear extension $\propto\sqrt{I}$. Due to the $r+s-1$ constraints on $\boldsymbol{\Delta}$, the $\boldsymbol{\Delta}$-integration is actually only over, say, $\boldsymbol{\Delta}_\perp$ and $\boldsymbol{\Delta}_\perp^T\boldsymbol{H}_\perp(\tilde{\boldsymbol{\pi}})\boldsymbol{\Delta}_\perp=I$ describes the surface of a $\bar{d}:=(r-1)(s-1)$-dimensional ellipsoid only. Approximating $p(\boldsymbol{\pi}|\boldsymbol{n})$ by $p(\tilde{\boldsymbol{\pi}}|\boldsymbol{n})$ in (2), where $\tilde{\pi}_{ij}=\tilde{\pi}_{i+}\tilde{\pi}_{+j}$ we get

$$p(I|\boldsymbol{n}) \;=\; B(\boldsymbol{n})\cdot I^{\frac{\bar{d}}{2}-1}+o(I^{\frac{\bar{d}}{2}-1}) \quad\text{with}\quad B(\boldsymbol{n})=\int S_\perp(\tilde{\boldsymbol{\pi}})p(\tilde{\boldsymbol{\pi}}|\boldsymbol{n})d^{r+s-2}\tilde{\boldsymbol{\pi}}$$

where $S_\perp=\frac{\Pi^{\bar{d}/2}}{\Gamma(\bar{d}/2)\sqrt{\det\boldsymbol{H}_\perp}}$ is the ellipsoid's surface ($\Pi=3.14...$). Note that $d\tilde{\boldsymbol{\pi}}$ still contains a Jakobian from the non-linear coordinate transformation. So the small I asymptotics is $p(I|\boldsymbol{n})\propto I^{\frac{\bar{d}}{2}-1}$ (for any prior), but a closed form expression for the coefficient $B(\boldsymbol{n})$ has yet to be derived.

Similarly we may derive the scaling behavior of $p(I|\boldsymbol{n})$ for $I\approx I_{max}:=\min\{\ln r,\ln s\}$. $I(\boldsymbol{\pi})$ can be written as $H(\imath)-H(\imath|\jmath)$, where H is the entropy. Without loss of generality we may assume $r\leq s$. $H(\imath)\leq\ln r$ with equality iff $\pi_{i+}=\frac{1}{r}$ for all i. $H(\imath|\jmath)\geq 0$ with equality iff $\imath$ is a deterministic function of $\jmath$. Together, $I(\tilde{\boldsymbol{\pi}})=I_{max}$ iff $\tilde{\pi}_{ij}=\frac{1}{r}\delta_{i,m(j)}\cdot\sigma_j$, where $m:\{1...s\}\to\{1..r\}$ is any onto map and the $\sigma_j\geq 0$ respect the constraints $\sum_{j\in m^{-1}(i)}\sigma_j=1$. This suggests the reparametrization $\pi_{ij}=\frac{1}{r}\delta_{i,m(j)}\sigma_j+\Delta_{ij}$ in the integral (2) for each choice of $m()$ and suitable constraints on σ and Δ.

5 Numerics

In order to approximate the distribution of mutual information in practice, one needs consider implementation issues and the computational complexity of the overall method. This is what we set out to do in the following.

Computational complexity and accuracy. Regarding computational complexity, 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 $\boldsymbol{n}$) 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 $I(\hat{\boldsymbol{\pi}})$.

With respect to the quality of the approximation, let us briefly consider the case of the variance. The expression for the exact variance has been Taylor-expanded in $(\frac{rs}{n})$, so the relative error $\frac{\mathrm{Var}[I]_{approx}-\mathrm{Var}[I]_{exact}}{\mathrm{Var}[I]_{exact}}$ of the approximation is of the order $(\frac{rs}{n})^2$, *if* $\imath$ and $\jmath$ are dependent. In the opposite case, the $O(n^{-1})$ term in the sum drops itself down to order n^{-2} resulting in a reduced relative accuracy $O(\frac{rs}{n})$ of the approximated variance. These results were confirmed by numerical experiments that we realized by Monte Carlo simulation to obtain "exact" values of the variance for representative choices of π_{ij}, r, s, and n. The approximation for the variance, together with those for the skewness and kurtosis, and the exact expression for the mean, allow a good description of the distribution $p(I|\boldsymbol{n})$ to be obtained 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 second leading order term can be freely adjusted by adjusting $n''_{ij} \in [0...1]$. So one may argue that anything beyond the leading order term 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, the leading orders of $E[I]$ and $E[I^2]$ do not suffice to compute the leading order term of $\mathrm{Var}[I]$.

Approximating the distribution. Let us now consider approximating the overall distribution of mutual information based on the mean and the variance. Fitting a normal distribution is an obvious possible choice, as the central limit theorem ensures that $p(I|\boldsymbol{n})$ converges to a Gaussian distribution with mean $E[I]$ and variance $\mathrm{Var}[I]$. Since I is non-negative, it is also worth considering the approximation of $p(I|\boldsymbol{\pi})$ by a Gamma (i.e., a scaled χ^2). Another natural candidate is the Beta distribution, which is defined for variables in the [0,1] real interval. I can be made such a variable by a simple normalization. Of course the Gamma and the Beta are asymptotically correct, too.

We report a graphical comparison of the different approximations by focusing on the special case of binary random variables, and on three possible vectors of

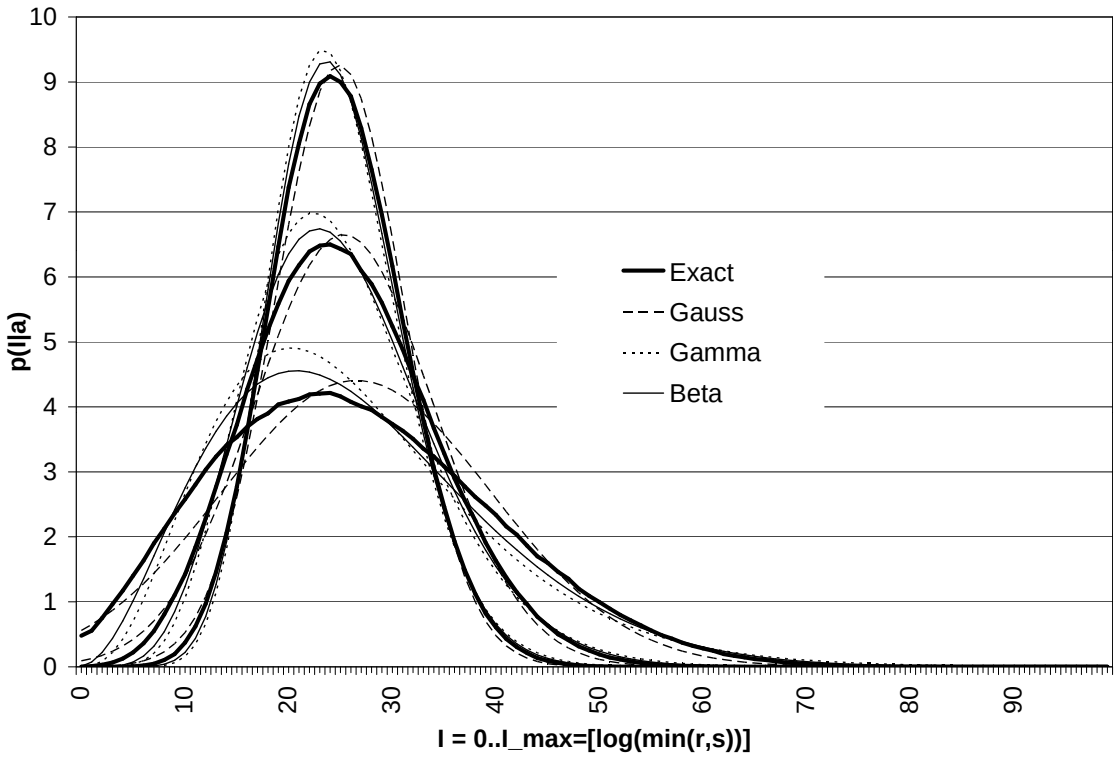

Figure 1: *Distribution of mutual information for two binary random variables (The labelling of the horizontal axis is the percentage of I_*max.*) There are three groups of curves, for different choices of counts* (n_{11},n_{12},n_{21},n_{22})*. The upper group is related to the vector* (40,10,20,80)*, the intermediate one to the vector* (20,5,10,40)*, and the lower group to* (8,2,4,16)*. Each group shows the "exact" distribution and three approximating curves, based on the Gaussian, Gamma and Beta distributions.*

counts. Figure 1 compares the "exact" distribution of mutual information, computed via Monte Carlo simulation, with the approximating curves. These curves have been fitted using the exact mean and the approximated variance of the preceding section. The figure clearly shows that all the approximations are rather good, with a slight preference for the Beta approximation. The curves tend to do worse for smaller sample sizes, as expected. Higher moments may be used to improve the accuracy (Section 3), or this can be improved using our considerations about tail approximations in Section 4.

6 Expressions for Missing Data

In the following we generalize the setup to include the case of missing data, which often occurs in practice. We extend the counts n_{ij} to include $n_{?j}$, which counts the number of instances in which only j is observed (i.e., the number of (?,j) instances),

and the counts $n_{i?}$ for the number of $(i,?)$ instances, where only $\imath$ is observed.

We make the common assumption that the missing data mechanism is ignorable (*missing at random* and *distinct*) [LR87]. The probability distribution of $\jmath$ given that $\imath$ is missing coincides with the marginal π_{+j}, and vice versa, as a consequence of this assumption.

Setup. The sample size n is now $n_c+n_{+?}+n_{?+}$, where n_c is the number of complete units. Let $\boldsymbol{n}=(n_{ij},n_{i?},n_{?j})$ denote as before the vector of counts, now including the counts $n_{i?}$ and $n_{?j}$, for all i and j. The probability of a specific data set $\boldsymbol{D}$, given $\boldsymbol{\pi}$, hence, is $p(\boldsymbol{D}|\boldsymbol{\pi},n_c,n_{+?},n_{?+})=\prod_{ij}\pi_{ij}^{n_{ij}}\prod_i\pi_{i+}^{n_{i?}}\prod_j\pi_{+j}^{n_{?j}}$. Assuming a uniform prior $p(\boldsymbol{\pi})\propto 1\cdot\delta(\pi_{++}-1)$, Bayes' rule leads to the posterior (which is also the likelihood in case of uniform prior)

$$p(\boldsymbol{\pi}|\boldsymbol{n}) = \frac{1}{\mathcal{N}(\boldsymbol{n})}\prod_{ij}\pi_{ij}^{n_{ij}}\prod_i\pi_{i+}^{n_{i?}}\prod_j\pi_{+j}^{n_{?j}}\delta(\pi_{++}-1)$$

where the normalization $\mathcal{N}$ is chosen such that $\int p(\boldsymbol{\pi}|\boldsymbol{n})d^{rs}\boldsymbol{\pi}=1$. With missing data there is, in general, no closed form expression for $\mathcal{N}$ any more (cf. (6)).

In the following, we restrict ourselves to a discussion of leading order (in n^{-1}) expressions. In leading order, any Dirichlet prior with $n''_{ij}=O(1)$ leads to the same results, hence we can simply assume a uniform prior. In leading order, the mean $E[\boldsymbol{\pi}]$ coincides with the mode of $p(\boldsymbol{\pi}|\boldsymbol{n})$, i.e. the maximum likelihood estimate of $\boldsymbol{\pi}$. The log-likelihood function $\ln p(\boldsymbol{\pi}|\boldsymbol{n})$ is

$$L(\boldsymbol{\pi}|\boldsymbol{n}) = \sum_{ij} n_{ij}\ln\pi_{ij} + \sum_i n_{i?}\ln\pi_{i+} + \sum_j n_{?j}\ln\pi_{+j} - \ln\mathcal{N}(\boldsymbol{n}) - \lambda(\pi_{++}-1),$$

where we have introduced the Lagrange multiplier λ to take into account the restriction $\pi_{++}=1$. The maximum is at $\frac{\partial L}{\partial\pi_{ij}}=\frac{n_{ij}}{\pi_{ij}}+\frac{n_{i?}}{\pi_{i+}}+\frac{n_{?j}}{\pi_{+j}}-\lambda=0$. Multiplying this by π_{ij} and summing over i and j we obtain $\lambda=n$. The maximum likelihood estimate $\hat{\boldsymbol{\pi}}$ is, hence, given by

$$\hat\pi_{ij} = \frac{1}{n}\left(n_{ij} + n_{i?}\frac{\hat\pi_{ij}}{\hat\pi_{i+}} + n_{?j}\frac{\hat\pi_{ij}}{\hat\pi_{+j}}\right). \tag{16}$$

This is a non-linear equation in $\hat\pi_{ij}$, which, in general, has no closed form solution. Nevertheless Eq. (16) can be used to approximate $\hat\pi_{ij}$. Eq. (16) coincides with the popular expectation-maximization (EM) algorithm [CF74] if one inserts a first estimate $\hat\pi^0_{ij}=\frac{n_{ij}}{n}$ into the r.h.s. of (16) and then uses the resulting l.h.s. $\hat\pi^1_{ij}$ as a new estimate, etc. This iteration (quickly) converges to the maximum likelihood solution (if missing instances are not too frequent). Using this we can compute the leading order term for the mean of the mutual information (and of any other function of π_{ij}): $E[I]=I(\hat{\boldsymbol{\pi}})+O(n^{-1})$. The leading order term for the covariance can be obtained from the second derivative of L.

Unimodality of $p(\boldsymbol{\pi}|\boldsymbol{n})$. The $rs\times rs$ Hessian matrix $\boldsymbol{H}\in\mathbb{R}^{rs\cdot rs}$ of $-L$ and the second derivative in the direction of the rs-dimensional column vector $\boldsymbol{v}\in\mathbb{R}^{rs}$ are

$$\boldsymbol{H}_{(ij)(kl)}[\boldsymbol{\pi}] \;:=\; -\frac{\partial L}{\partial\pi_{ij}\partial\pi_{kl}} = \frac{n_{ij}}{\pi_{ij}^2}\delta_{ik}\delta_{jl} + \frac{n_{i?}}{\pi_{i+}^2}\delta_{ik} + \frac{n_{?j}}{\pi_{+j}^2}\delta_{jl},$$

$$\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v} \;=\; \sum_{ijkl} v_{ij}\boldsymbol{H}_{(ij)(kl)}v_{kl} \;=\; \sum_{ij}\frac{n_{ij}}{\pi_{ij}^2}v_{ij}^2 + \sum_i\frac{n_{i?}}{\pi_{i+}^2}v_{i+}^2 + \sum_j\frac{n_{?j}}{\pi_{+j}^2}v_{+j}^2 \;\geq\; 0.$$

This shows that $-L$ is a convex function of $\boldsymbol{\pi}$, hence $p(\boldsymbol{\pi}|\boldsymbol{n})$ has a single (possibly degenerate) global maximum. L is strictly convex if $n_{ij}>0$ for all ij, since $\boldsymbol{v}^T\boldsymbol{H}\boldsymbol{v}>0$ $\forall\,\boldsymbol{v}\neq 0$ in this case. (Note that positivity of $n_{i?}$ for all i is not sufficient, since $v_{i+}=0$ for $\boldsymbol{v}\neq 0$ is possible. Actually $v_{++}=0$.) This implies a unique global maximum, which is attained in the interior of the probability simplex. Since EM is known to converge to a local maximum, this shows that in fact *EM always converges to the global maximum.*

Covariance of $\boldsymbol{\pi}$. With

$$\boldsymbol{A}_{(ij)(kl)} := \boldsymbol{H}_{(ij)(kl)}[\hat{\boldsymbol{\pi}}] = n\left[\frac{\delta_{ik}\delta_{jl}}{\rho_{ij}} + \frac{\delta_{ik}}{\rho_{i?}} + \frac{\delta_{jl}}{\rho_{?j}}\right],$$

$$\rho_{ij} := n\frac{\hat{\pi}_{ij}^2}{n_{ij}}, \quad \rho_{i?} := n\frac{\hat{\pi}_{i+}^2}{n_{i?}}, \quad \rho_{?j} := n\frac{\hat{\pi}_{+j}^2}{n_{?j}} \tag{17}$$

and $\boldsymbol{\Delta} := \boldsymbol{\pi} - \hat{\boldsymbol{\pi}}$, we can represent the posterior to leading order as an $(rs-1)$-dimensional Gaussian:

$$p(\boldsymbol{\pi}|\boldsymbol{n}) \sim e^{-\frac{1}{2}\boldsymbol{\Delta}^T\boldsymbol{A}\boldsymbol{\Delta}}\delta(\Delta_{++}). \tag{18}$$

The easiest way to compute the covariance (and other quantities) is to also represent the δ-function as a narrow Gaussian of width $\varepsilon\approx 0$. Inserting $\delta(\Delta_{++})\approx \frac{1}{\varepsilon\sqrt{2\pi}}\exp(-\frac{1}{2\varepsilon^2}\boldsymbol{\Delta}^T\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{\Delta})$ into (18), where $\boldsymbol{e}_{ij}=1$ for all ij (hence $\boldsymbol{e}^T\boldsymbol{\Delta}=\Delta_{++}$), leads to a full rs-dimensional Gaussian with kernel $\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{u}\boldsymbol{v}^T$, $\boldsymbol{u}=\boldsymbol{v}=\frac{1}{\varepsilon}\boldsymbol{e}$. The covariance of a Gaussian with kernel $\tilde{\boldsymbol{A}}$ is $\tilde{\boldsymbol{A}}^{-1}$. Using the Sherman-Morrison formula $\tilde{\boldsymbol{A}}^{-1}=\boldsymbol{A}^{-1}-\boldsymbol{A}^{-1}\frac{\boldsymbol{u}\boldsymbol{v}^T}{1+\boldsymbol{v}^T\boldsymbol{A}^{-1}\boldsymbol{u}}\boldsymbol{A}^{-1}$ [PFTV92, p. 73] and $\varepsilon\to 0$ we get

$$\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] := E[\Delta_{ij}\Delta_{kl}] \simeq [\tilde{\boldsymbol{A}}^{-1}]_{(ij)(kl)} = \left[\boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{e}\boldsymbol{e}^T\boldsymbol{A}^{-1}}{\boldsymbol{e}^T\boldsymbol{A}^{-1}\boldsymbol{e}}\right]_{(ij)(kl)}, \tag{19}$$

where $\simeq$ denotes equality up to terms of order n^{-2}. Singular matrices $\boldsymbol{A}$ are easily avoided by choosing a prior such that $n_{ij}>0$ for all i and j. $\boldsymbol{A}$ may be inverted exactly or iteratively, the latter by a trivial inversion of the diagonal part $\delta_{ik}\delta_{jl}/\rho_{ij}$ and by treating $\delta_{ik}/\rho_{i?}+\delta_{jl}/\rho_{?j}$ as a perturbation.

Missing observations for one variable only. In the case only one variable is missing, say $n_{?j}=0$, closed form expressions can be obtained. If we sum (16) over j

we get $\hat\pi_{i+}=\frac{n_{i+}+n_{i?}}{n}$. Inserting $\hat\pi_{i+}=\frac{n_{i+}+n_{i?}}{n}$ into the r.h.s. of (16) and solving w.r.t. $\hat\pi_{ij}$, we get the explicit expression

$$\hat\pi_{ij} = \frac{n_{i+}+n_{i?}}{n} \cdot \frac{n_{ij}}{n_{i+}}. \tag{20}$$

Furthermore, it can easily be verified (by multiplication) that $\boldsymbol{A}_{(ij)(kl)}=n[\delta_{ik}\delta_{jl}/\rho_{ij}+\delta_{ik}/\rho_{i?}]$ has inverse $[\boldsymbol{A}^{-1}]_{(ij)(kl)}=\frac{1}{n}[\rho_{ij}\delta_{ik}\delta_{jl}-\frac{\rho_{ij}\rho_{kl}}{\rho_{i+}+\rho_{i?}}\delta_{ik}]$. With the abbreviations

$$\tilde Q_{i?} := \frac{\rho_{i?}}{\rho_{i?}+\rho_{i+}} \quad \text{and} \quad \tilde Q := \sum_i \rho_{i+}\tilde Q_{i?}$$

we get $[\boldsymbol{A}^{-1}\boldsymbol{e}]_{ij}=\sum_{kl}[\boldsymbol{A}^{-1}]_{(ij)(kl)}=\frac{1}{n}\rho_{ij}\tilde Q_{i?}$ and $\boldsymbol{e}^T\boldsymbol{A}^{-1}\boldsymbol{e}=\tilde Q/n$. Inserting everything into (19) we get

$$\mathrm{Cov}_{(ij)(kl)}[\boldsymbol{\pi}] \;\simeq\; \frac{1}{n}\left[\rho_{ij}\delta_{ik}\delta_{jl} - \frac{\rho_{ij}\rho_{kl}}{\rho_{i+}+\rho_{i?}}\delta_{ik} - \frac{\rho_{ij}\tilde Q_{i?}\rho_{kl}\tilde Q_{k?}}{\tilde Q}\right].$$

Inserting this expression for the covariance into (5), using $\bar{\boldsymbol{\pi}} := E[\boldsymbol{\pi}] = \hat{\boldsymbol{\pi}}+O(n^{-1})$, we finally get the leading order term in $1/n$ for the variance of mutual information:

$$\mathrm{Var}[I] \;\simeq\; \frac{1}{n}[\tilde K - \tilde J^2/\tilde Q - \tilde P], \qquad \tilde K := \sum_{ij}\rho_{ij}\left(\ln\frac{\hat\pi_{ij}}{\hat\pi_{i+}\hat\pi_{+j}}\right)^2, \tag{21}$$

$$\tilde P := \sum_i \frac{\tilde J_{i+}^2 Q_{i?}}{\rho_{i?}}, \qquad \tilde J := \sum_i \tilde J_{i+}\tilde Q_{i?}, \qquad \tilde J_{i+} := \sum_j \rho_{ij}\ln\frac{\hat\pi_{ij}}{\hat\pi_{i+}\hat\pi_{+j}}.$$

A closed form expression for $\mathcal{N}(\boldsymbol{n})$ also exists. Symmetric expressions for the case when only $\imath$ is missing can be obtained. Note that for the complete data case $n_{i?}=0$, we have $\hat\pi_{ij}=\rho_{ij}=\frac{n_{ij}}{n}$, $\rho_{i?}=\infty$, $\tilde Q_{i?}=\tilde Q=1$, $\tilde J=J$, $\tilde K=K$, and $\tilde P=0$, consistent with (8).

There is at least one reason for minutely having inserted all expressions into each other and introducing quite a number definitions. In the presented form all expressions involve at most a double sum. Hence, the overall time for computing the mean and variance when only one variable is missing is $O(rs)$.

Expressions for the general case. In the general case when both variables are missing, each EM iteration (16) for $\hat\pi_{ij}$ needs $O(rs)$ operations. The naive inversion of $\boldsymbol{A}$ needs time $O((rs)^3)$, and using it to compute $\mathrm{Var}[I]$ time $O((rs)^2)$. Since the contribution from unlabelled-$\imath$ instances can be interpreted as a rank s modification of $\boldsymbol{A}$ in the case of when $\imath$ is not missing, one can use Woodbury's formula $[\boldsymbol{B}+\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T]^{-1}=\boldsymbol{B}^{-1}-\boldsymbol{B}^{-1}\boldsymbol{U}[\boldsymbol{D}^{-1}+\boldsymbol{V}^T\boldsymbol{B}^{-1}\boldsymbol{U}]^{-1}\boldsymbol{V}^T\boldsymbol{B}^{-1}$ [PFTV92, p. 75] with $\boldsymbol{B}_{(ij)(kl)}=\delta_{ik}\delta_{jl}/\rho_{ij}+\delta_{ik}/\rho_{i?}$, $\boldsymbol{D}_{jl}=\delta_{jl}/\rho_{?j}$, and $\boldsymbol{U}_{(ij)l}=\boldsymbol{V}_{(ij)l}=\delta_{jl}$, to reduce the inversion of the $rs\times rs$ matrix $\boldsymbol{A}$ to the inversion of a single s-dimensional matrix. The result can be written in the form

$$[\boldsymbol{A}^{-1}]_{(ij)(kl)} = \frac{1}{n}\left[F_{ijl}\delta_{ik} - \sum_{mn}F_{ijm}[\boldsymbol{G}^{-1}]_{mn}F_{kln}\right], \tag{22}$$

$$F_{ijl} := \rho_{ij}\delta_{jl} - \frac{\rho_{ij}\rho_{kl}}{\rho_{i?}+\rho_{i+}}, \qquad G_{mn} := \rho_{?n}\delta_{mn} + F_{+mn}.$$

The result for the covariance (19) can be inserted into (5) to obtain the leading order term for the variance:

$$\mathrm{Var}[I] \;\simeq\; \boldsymbol{l}^T\boldsymbol{A}^{-1}\boldsymbol{l} - (\boldsymbol{l}^T\boldsymbol{A}^{-1}\boldsymbol{e})^2/(\boldsymbol{e}^T\boldsymbol{A}^{-1}\boldsymbol{e}) \quad \text{where} \quad \boldsymbol{l}_{ij} := \ln\frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}\hat{\pi}_{+j}}. \tag{23}$$

Inserting (22) into (23) and rearranging terms appropriately, we can compute $\mathrm{Var}[I]$ in time $O(rs)$ plus the time $O(s^2r)$ to compute the $s\times s$ matrix $\boldsymbol{G}$ and time $O(s^3)$ to invert it, plus the time $O(\#\cdot rs)$ for determining $\hat{\pi}_{ij}$, where $\#$ is the number of iterations of EM. Of course, one can and should always choose $s \le r$. Note that these expressions converge to the exact values when n goes to infinity, irrespectively of the amount of missing data.

7 Applications

The results in the preceding sections provide fast and reliable methods to approximate the distribution of mutual information from either complete or incomplete data. The derived tools have been obtained in the theoretically sound framework of Bayesian statistics, which we regard as their basic justification. As these methods are available for the first time, it is natural to wonder what their possible uses can be on the application side or, stated differently, what can be gained in practice moving from descriptive to inductive methods. We believe that the impact on real applications can be significant, according to three main scenarios: *robust inference methods*, *inferring models that perform well*, and *fast learning from massive data sets*. In the following we use classification as a thread to illustrate the above scenarios. *Classification* is one of the most important techniques for knowledge discovery in databases [DHS01]. A classifier is an algorithm that allocates new objects to one out of a finite set of previously defined groups (or *classes*) on the basis of observations on several characteristics of the objects, called *attributes* or *features*. Classifiers are typically learned from data, making explicit the knowledge that is hidden in databases, and using this knowledge to make predictions about new data.

Robust inference methods. An obvious observation is that descriptive methods cannot compete, by definition, with inductive ones when robustness is concerned. Hence, the results presented in this paper lead naturally to a spin-off for reliable methods of inference.

Let us focus on classification problems, for the sake of explanation. Applying robust methods to classification means to produce classifications that are correct with a given probability. It is easy to imagine sensible (e.g., nuclear, medical) applications where reliability of classification is a critical issue. To achieve reliability, a necessary step consists in associating a posterior probability (i.e., a guarantee level) to classification models inferred from data, such as classification trees or Bayesian

nets. Let us consider the case of Bayesian networks. These are graphical models that represent structures of (in)dependence by directed acyclic graphs, where nodes in the graph are regarded as random variables [Pea88, Nea04]. Two nodes are connected by an arc when there is direct stochastic dependence between them. Inferring Bayesian nets from data is often done by connecting nodes with significant value of descriptive mutual information. Little work has been done on robustly inferring Bayesian nets, probably because of the difficulty to deal with the distribution of mutual information, with the notable exception of Kleiter's work [Kle99]. Joining Kleiter's work with ours might lead to inference of Bayesian network structures that are correct with a given probability. Some work has already been done to this direction [ZH03].

Feature selection might also benefit from robust methods. *Feature selection* is the problem of reducing the number of feature variables to deal with in classification problems. Features can reliably be discarded only when they are irrelevant to the class with high probability. This needs knowledge of the distribution of mutual information. In Section 8 we propose a filter based on the distribution of mutual information to address this problem.

Inferring models that perform well. It is well-known that model complexity must be in proper balance with available data in order to achieve good classification accuracy. In fact, unjustified complexity of inferred models leads classifiers almost inevitably to *overfitting*, i.e. to memorize the available sample rather than extracting regularities from it that are needed to make useful predictions on new data [DHS01]. Overfitting could be avoided by using the distribution of mutual information. With Bayesian nets, for example, this could be achieved by drawing arcs between nodes only if these are supported by data with high probability. This is a way to impose a bias towards simple structures. It has to be verified whether or not this approach can systematically lead to better accuracy.

Model complexity can also be reduced by discarding features. This can be achieved by including a feature only when its mutual information with the class is significant with high probability. This approach is taken in Section 8, where we show that it can effectively lead to better prediction accuracy of the resulting models.

Fast learning from massive data sets. Another very promising application of the distribution of mutual information is related to *massive data sets*. These are huge samples, which are becoming more and more available in real applications, and which constitute a serious challenge for machine learning and statistical applications. With massive data sets it is impractical to scan all the data, so classifiers must be reliably inferred by accessing only a small subset of the units. Recent work has highlighted [PM03] how inductive methods allow this to be realized. The intuition is the following: the inference phase stops reading data when the inferred model, say a Bayesian net, has reached a given posterior probability. By choosing such probability sufficiently high, one can be arbitrarily confident that the inferred model will not change much by reading the neglected data, making the remaining units

superfluous.

8 Feature Selection

Feature selection is a basic step in the process of building classifiers [BL97, DL97, LM98]. In fact, even if theoretically more features should provide one with better *prediction accuracy* (i.e., the relative number of correct predictions), in real cases it has been observed many times that this is not the case [KS96] and that it is important to discard irrelevant, or weakly relevant features.

The purpose of this section is to illustrate how the distribution of mutual information can be applied in this framework, according to some of the ideas in Section 7. Our goal is inferring simple models that avoid overfitting and have an equivalent or better accuracy with respect to models that consider all the original features.

Two major approaches to feature selection are commonly used in machine learning [JKP94]: *filter* and *wrapper* models. The filter approach is a preprocessing step of the classification task. The wrapper model is computationally heavier, as it implements a search in the feature space using the prediction accuracy as reward measure. In the following we focus our attention on the filter approach: we define two new filters and report experimental analysis about them, both with complete and incomplete data.

The proposed filters. We consider the well-known filter (F) that computes the empirical mutual information between features and the class, and discards low-valued features [Lew92]. This is an easy and effective approach that has gained popularity with time. Cheng reports that it is particularly well suited to jointly work with Bayesian network classifiers, an approach by which he won the *2001 international knowledge discovery competition* [CHH$^+$02]. The 'Weka' data mining package implements it as a standard system tool (see [WF99, p. 294]).

A problem with this filter is the variability of the empirical mutual information with the sample. This may cause wrong judgments of relevance, when those features are selected for which the mutual information exceeds a fixed threshold ε. In order for the selection to be robust, we must have some guarantee about the actual value of mutual information.

We define two new filters. The *backward filter* (BF) *discards* an attribute if $p(I < \varepsilon|\boldsymbol{n}) > \bar{p}$ where I denotes the mutual information between the feature and the class, ε is an arbitrary (low) positive threshold and $\bar{p}$ is an arbitrary (high) probability. The *forward filter* (FF) *includes* an attribute if $p(I > \varepsilon|\boldsymbol{n}) > \bar{p}$, with the same notations. BF is a conservative filter, along the lines discussed about robustness in Section 7, because it will only discard features after observing substantial evidence supporting their irrelevance. FF instead will tend to use fewer features (aiming at producing classifiers that perform better), i.e. only those for which there is substantial evidence about them being useful in predicting the class.

The next sections present experimental comparisons of the new filters and the original filter F.

Experimental methodology. For the following experiments we use the *naive Bayes classifier* [DH73]. This is a good classification model—despite its simplifying assumptions [DP97]—, which often competes successfully with much more complex classifiers from the machine learning field, such as C4.5 [Qui93]. The experiments focus on the incremental use of the naive Bayes classifier, a natural learning process when the data are available sequentially: the data set is read instance by instance; each time, the chosen filter selects a subset of attributes that the naive Bayes uses to classify the new instance; the naive Bayes then updates its knowledge by taking into consideration the new instance and its actual class. The incremental approach allows us to better highlight the different behaviors of the empirical filter (F) and those based on the distribution of mutual information (BF and FF). In fact, for increasing sizes of the learning set the filters converge to the same behavior.

For each filter, we are interested in experimentally evaluating two quantities: for each instance of the data set, the average number of correct predictions (namely, the prediction accuracy) of the naive Bayes classifier up to such instance; and the average number of attributes used. By these quantities we can compare the filters and judge their effectiveness.

The implementation details for the following experiments include: using the Beta approximation (Section 5) to the distribution of mutual information, with the exact mean (15) and the $O(n^{-3})$-approximation of the variance, given in (11); using the uniform prior for the naive Bayes classifier and all the filters; and setting the level $\bar{p}$ for the posterior probability to 0.95. As far as ε is concerned, we cannot set it to zero because the probability that two variables are independent ($I = 0$) is zero according to the inferential Bayesian approach. We can interpret the parameter ε as a degree of dependency strength below which attributes are deemed irrelevant. We set ε to 0.003, in the attempt of only discarding attributes with negligible impact on predictions. As we will see, such a low threshold can nevertheless bring to discard many attributes.

9 Experimental analysis with incomplete samples

Table 1 lists ten data sets used in the experiments for complete data. These are real data sets on a number of different domains. For example, Shuttle-small reports data on diagnosing failures of the space shuttle; Lymphography and Hypothyroid are medical data sets; Spam is a body of e-mails that can be spam or non-spam; etc.

The data sets presenting non-categorical features have been pre-discretized by MLC++ [KJL+94], default options, i.e. by the common entropy-based discretization [FI93]. This step may remove some attributes judging them as irrelevant, so the number of features in the table refers to the data sets after the possible discretization.

Name	#feat.	#inst.	mode freq.
Australian	36	690	0.555
Chess	36	3196	0.520
Crx	15	653	0.547
German-org	17	1000	0.700
Hypothyroid	23	2238	0.942
Led24	24	3200	0.105
Lymphography	18	148	0.547
Shuttle-small	8	5800	0.787
Spam	21611	1101	0.563
Vote	16	435	0.614

Table 1: *Complete data sets used in the experiments, together with their number of features, of instances and the relative frequency of the mode. All but the Spam data sets are available from the UCI repository of machine learning data sets [MA95]. The Spam data set is described in [AKC+00] and available from Androutsopoulos's web page.*

The instances with missing values have been discarded, and the third column in the table refers to the data sets without missing values. Finally, the instances have been randomly sorted before starting the experiments.

Results. In short, the results show that FF outperforms the commonly used filter F, which in turn, outperforms the filter BF. FF leads either to the same prediction accuracy as F or to a better one, using substantially fewer attributes most of the times. The same holds for F versus BF.

In particular, we used the *two-tails paired t test* at level 0.05 to compare the prediction accuracies of the naive Bayes with different filters, in the first k instances of the data set, for each k.

The results in Table 2 show that, despite the number of used attributes is often substantially different, both the differences between FF and F, and the differences between F and BF, were never statistically significant on eight data sets out of ten.

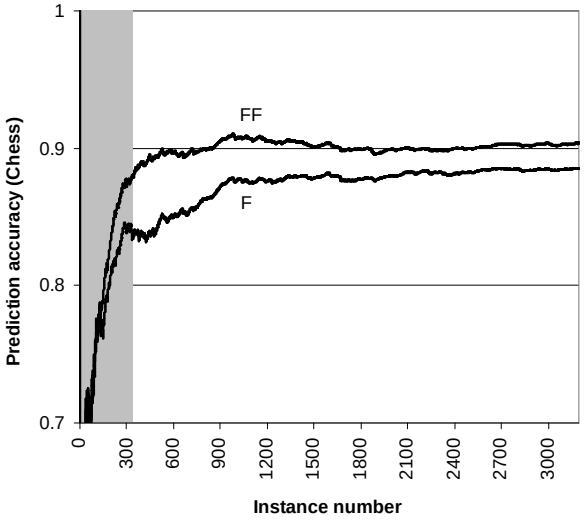

Figure 2: *Comparison of the prediction accuracies of the naive Bayes with filters F and FF on the Chess data set. The gray area denotes differences that are not statistically significant.*

The remaining cases are described by means of the following figures. Figure 2 shows that FF allowed the naive Bayes to significantly do better predictions than F

Data set	#feat.	FF	F	BF
Australian	36	32.6	34.3	35.9
Chess	36	12.6	18.1	26.1
Crx	15	11.9	13.2	15.0
German-org	17	5.1	8.8	15.2
Hypothyroid	23	4.8	8.4	17.1
Led24	24	13.6	14.0	24.0
Lymphography	18	18.0	18.0	18.0
Shuttle-small	8	7.1	7.7	8.0
Spam	21611	123.1	822.0	13127.4
Vote	16	14.0	15.2	16.0

Table 2: *Average number of attributes selected by the filters on the entire data set, reported in the last three columns. (Refer to the Section 'The proposed filters' for the definition of the filters.) The second column from left reports the original number of features. In all but one case, FF selected fewer features than F, sometimes much fewer; F usually selected much fewer features than BF, which was very conservative. Boldface names refer to data sets on which prediction accuracies where significantly different.*

for the greatest part of the Chess data set. The maximum difference in prediction accuracy is obtained at instance 422, where the accuracies are 0.889 and 0.832 for the cases FF and F, respectively. Figure 2 does not report the BF case, because there is no significant difference with the F curve. The good performance of FF was obtained using only about one third of the attributes (Table 2).

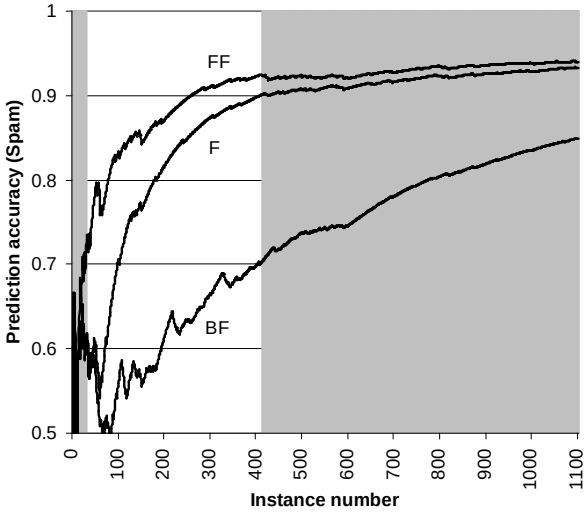

Figure 3: *Prediction accuracies of the naive Bayes with filters F, FF and BF on the Spam data set. The differences between F and FF are significant in the range of observations 32–413. The differences between F and BF are significant from observations 65 to the end (this significance is not displayed in the picture).*

Figure 3 compares the accuracies on the Spam data set. The difference between the cases FF and F is significant in the range of instances 32–413, with a maximum at instance 59 where accuracies are 0.797 and 0.559 for FF and F, respectively. BF is significantly worse than F from instance 65 to the end. This excellent performance

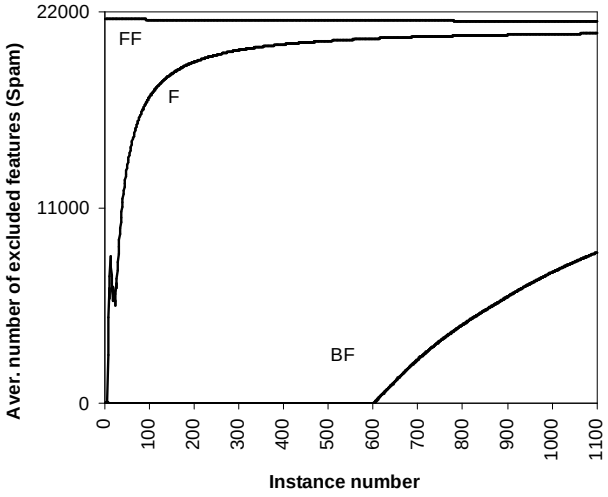

Figure 4: *Average number of attributes excluded by the different filters on the Spam data set.*

of FF is even more valuable considered the very low number of attributes selected for classification. In the Spam case, attributes are binary and correspond to the presence or absence of words in an e-mail and the goal is to decide whether or not the e-mail is spam. All the 21611 words found in the body of e-mails were initially considered. FF shows that only an average of about 123 relevant words is needed to make good predictions. Worse predictions are made using F and BF, which select, on average, about 822 and 13127 words, respectively. Figure 4 shows the average number of excluded features for the three filters on the Spam data set. FF suddenly discards most of the features, and keeps the number of selected features almost constant over all the process. The remaining filters tend to such a number, with different speeds, after initially including many more features than FF.

In summary, the experimental evidence supports the strategy of only using the features that are reliably judged to carry useful information to predict the class, provided that the judgment can be updated as soon as new observations are collected. FF almost always selects fewer features than F, leading to a prediction accuracy at least as good as the one F leads to. The comparison between F and BF is analogous, so FF appears to be the best filter and BF the worst. This is not surprising as BF was designed to be conservative and was used here just as a term of comparison. The natural use of BF is for robust classification when it is important not to discard features potentially relevant to predict the class.

10 Experimental analysis with incomplete samples

This section makes experimental analysis on incomplete data along the lines of the preceding experiments. The new data sets are listed in Table 3.

The filters F and FF are defined as before. However, now the mean and variance

Name	#feat.	#inst.	#m.d.	mode freq.
Audiology	69	226	317	0.212
Crx	15	690	67	0.555
Horse-colic	18	368	1281	0.630
Hypothyroidloss	23	3163	1980	0.952
Soybean-large	35	683	2337	0.135

Table 3: *Incomplete data sets used for the new experiments, together with their number of features, instances, missing values, and the relative frequency of the mode. The data sets are available from the UCI repository of machine learning data sets [MA95].*

of mutual information are obtained by using the results in Section 6, in particular the closed-form expressions for the case when only one variable is missing. In fact, in the present data sets the class is never missing, as it is quite common classification tasks. We remark that the mean is simply approximated now as $I(\hat{\boldsymbol{\pi}})$, where $\hat{\boldsymbol{\pi}}$ is given by (20), whereas the variance is reported in (21). Furthermore, note that also the traditional filter F, as well as the naive Bayes classifier, are now computed using the empirical probabilities (20). The remaining implementation details are as in the case of complete data.

Data set	#feat.	FF	F	BF
Audiology	69	64.3	68.0	68.7
Crx	15	9.7	12.6	13.8
Horse-colic	18	11.8	16.1	17.4
Hypothyroidloss	23	4.3	8.3	13.2
Soybean-large	35	34.2	35.0	35.0

Table 4: *Average number of attributes selected by the filters on the entire data set, reported in the last three columns. The second column from left reports the original number of featurs. FF always selected fewer features than F; F almost always selected fewer features than BF. Prediction accuracies where significantly different for the Hypothyroidloss data set.*

The results in Table 4 show that the filters behave very similarly to the case of complete data. The filter FF still selects the smallest number of features, and this number usually increases with F and even more with BF. The selection can be very pronounced, as with the Hypothyroidloss data set. This is also the only data set for which the prediction accuracies of F and FF are significantly different, in favor of FF. This is better highlighted by Figure 5.

Remark. The most prominent evidence from the experiments is the better performance of FF versus the traditional filter F. In this note we look at FF from another perspective to exemplify and explain its behavior.

FF includes an attribute if $p(I > \varepsilon | \boldsymbol{n}) > \bar{p}$, according to its definition. Let us

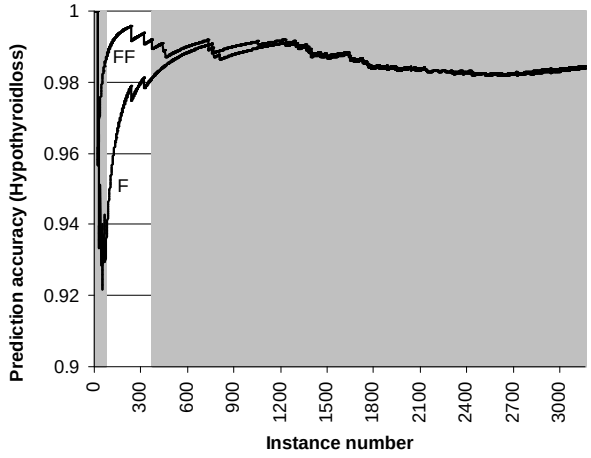

Figure 5: *Prediction accuracies of the naive Bayes with filters F and FF on the Hypothyroidloss data set. (BF is not reported because there is no significant difference with the F curve.) The differences between F and FF are significant in the range of observations 71–374. The maximum difference is achieved at observation 71, where the accuracies are 0.986 (FF) vs. 0.930 (F).*

assume that FF is realized by means of the Gaussian rather than the Beta approximation (as in the experiments above), and let us choose $\bar{p} \approx 0.977$. The condition $p(I > \varepsilon | \boldsymbol{n}) > \bar{p}$ becomes $\varepsilon < E[I] - 2 \cdot \sqrt{\mathrm{Var}[I]}$, or, in an approximate way, $I(\hat{\boldsymbol{\pi}}) > \varepsilon + 2 \cdot \sqrt{\mathrm{Var}[I]}$, given that $I(\hat{\boldsymbol{\pi}})$ is the first-order approximation of $E[I]$ (cf. (4)). We can regard $\varepsilon + 2 \cdot \sqrt{\mathrm{Var}[I]}$ as a new threshold ε'. Under this interpretation, we see that FF is approximately equal to using the filter F with the bigger threshold ε'. This interpretation makes it also clearer why FF can be better suited than F for sequential learning tasks. In sequential learning, $\mathrm{Var}[I]$ decreases as new units are read; this makes ε' a self-adapting threshold that adjusts the level of caution (in including features) as more units are read. In the limit, ε' is equal to ε. This characteristic of self-adaptation, which is absent in F, seems to be decisive to the success of FF.

11 Conclusions

This paper has provided fast and reliable analytical approximations for the variance, skewness and kurtosis of the posterior distribution of mutual information, with guaranteed accuracy from $O(n^{-1})$ to $O(n^{-3})$, as well as the exact expression of the mean. These results allow the posterior distribution of mutual information to be approximated both from complete and incomplete data. As an example, this paper has shown that good approximations can be obtained by fitting common curves with the mentioned mean and variance. To our knowledge, this is the first work that addresses the analytical approximation of the distribution of mutual information. Analytical approximations are important because their implementation is shown to lead to computations of the same order of complexity as needed for the empirical mutual information. This makes the inductive approach a serious competitor of the descriptive use of mutual information for many applications.

In fact, many applications are based on descriptive mutual information. We

have discussed how many of these could benefit from moving to the inductive side, and in particular we have shown how this can be done for feature selection. In this context, we have proposed the new filter FF, which is shown to be more effective for sequential learning tasks than the traditional filter based on empirical mutual information.

References

[AKC+00] I. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and D. Spyropoulos. An evaluation of naive Bayesian anti-spam filtering. In G. Potamias, V. Moustakis, and M. van Someren, editors, *Proceedings of the Workshop on Machine Learning in the New Information Age*, pages 9–17, 2000. 11th European Conference on Machine Learning.

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

[BL97] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1–2):245–271, 1997. Special issue on relevance.

[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.

[CF74] T. T. Chen and S. E. Fienberg. Two-dimensional contingency tables with both completely and partially cross-classified data. *Biometrics*, 32:133–144, 1974.

[CHH+02] J. Cheng, C. Hatzis, H. Hayashi, M. Krogel, S. Morishita, D. Page, and J. Sese. KDD cup 2001 report. *ACM SIGKDD Explorations*, 3(2), 2002.

[CL68] C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.

[DH73] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis*. Wiley, New York, 1973.

[DHS01] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern classification*. Wiley, 2001. 2nd edition.

[DL97] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis*, 1:131–156, 1997.

[DP97] P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29(2/3):103–130, 1997.

[FI93] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, pages 1022–1027, San Francisco, CA, 1993. Morgan Kaufmann.

[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. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 301–354. MIT press, 1998.

[Hut02] M. Hutter. Distribution of mutual information. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 399–406, Cambridge, MA, 2002. MIT Press.

[HZ03] M. Hutter and M. Zaffalon. Bayesian treatment of incomplete discrete data applied to mutual information and feature selection. In R. Kruse A. Günter and B. Neumann, editors, *Proceedings of the Twenty-sixth German Conference on Artificial Intelligence (KI-2003)*, volume 2821 of *Lecture Notes in Computer Science*, pages 396–406, Heidelberg, 2003. Springer.

[JKP94] G. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In W. W. Cohen and H. Hirsh, editors, *Proceedings of the Eleventh International Conference on Machine Learning*, pages 121–129, New York, 1994. Morgan Kaufmann.

[KJL+94] R. Kohavi, G. John, R. Long, D. Manley, and K. Pfleger. MLC++: a machine learning library in C++. In *Tools with Artificial Intelligence*, pages 740–743. IEEE Computer Society Press, 1994.

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

[KS67] M. G. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Griffin, London, 1967. 2nd edition.

[KS96] D. Koller and M. Sahami. Toward optimal feature selection. In *Proceedings of the Thirteenth International Conference on Machine Learning*, pages 284–292, 1996.

[Kul68] S. Kullback. *Information Theory and Statistics.* Dover, 1968.

[Lew92] D. D. Lewis. Feature selection and feature extraction for text categorization. In *Proceedings of Speech and Natural Language Workshop*, pages 212–217, San Francisco, 1992. Morgan Kaufmann.

[LM98] H. Liu and H. Motoda. *Feature Selection for Knowledge Discovery and Data Mining.* Kluwer, Norwell, MA, 1998.

[LR87] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* Wiley, New York, 1987.

[MA95] P. M. Murphy and D. W. Aha. UCI repository of machine learning databases, 1995.

[Nea04] Richard E. Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, 1988.

[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.

[PM03] D. Pelleg and A. Moore. Using Tarjan's red rule for fast dependency tree construction. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 825–832, Cambridge, MA, 2003. MIT Press.

[Qui93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 1993.

[WF99] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.

[WW95] D. H. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.

[ZH02] M. Zaffalon and M. Hutter. Robust feature selection by mutual information distributions. In A. Darwiche and N. Friedman, editors, *Proceedings of the 18th International Conference on Uncertainty in Artificial Intelligence (UAI-2002)*, pages 577–584, San Francisco, CA., 2002. Morgan Kaufmann.

[ZH03] M. Zaffalon and M. Hutter. Robust inference of trees. Technical Report IDSIA-11-03, IDSIA, 2003.