Estimating functions of probability distributions from a finite set of samples

David H. Wolpert[1,*] and David R. Wolf[2,†]
[1]*The Sante Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501*
[2]*DX-13, Image Analysis Section, MS P940, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*
(Received 15 July 1993; revised manuscript received 15 March 1995)

This paper addresses the problem of estimating a function of a probability distribution from a finite set of samples of that distribution. A Bayesian analysis of this problem is presented, the optimal properties of the Bayes estimators are discussed, and as an example of the formalism, closed form expressions for the Bayes estimators for the moments of the Shannon entropy function are derived. Then numerical results are presented that compare the Bayes estimator to the frequency-counts estimator for the Shannon entropy. We also present the closed form estimators, all derived elsewhere, for the mutual information, χ^2 covariance, and some other statistics.

PACS number(s): 02.70.Rw, 02.50.−r, 05.20.−y

I. INTRODUCTION

Consider a system with m possible states and an associated m vector of probabilities of those states $\boldsymbol{\rho}=(\rho_i)$, $1 \le i \le m$, $(\sum_{i=1}^{m}\rho_i=1)$. The system is repeatedly and independently sampled according to the distribution $\boldsymbol{\rho}$. Let the total number of samples be N and denote the associated vector of counts of states by $\mathbf{n}=(n_i)$, $1 \le i \le m$ $(\sum_{i=1}^{m}n_i=N)$. By definition, $\mathbf{n}$ is multinominally distributed.

In many cases what we are interested in is not $\boldsymbol{\rho}$ but some function of $\boldsymbol{\rho}$, $Q(\boldsymbol{\rho})$. In this paper we are concerned with the problem of estimating some such function $Q(\boldsymbol{\rho})$ from the data $\mathbf{n}$. This problem is ubiquitous in physics, arising, for example, in dimension estimation and in estimating correlations from data.

Example 1. In information dimension estimation [1], one imagines a discretization of a space containing an attractor. The attractor constitutes a probability density function across the space and therefore a probability distribution across the bins of the discretization. One is then interested in how the Renyi entropy of the distribution across the bins changes as the discretization becomes finer and finer. This behavior gives us the information dimension of the attractor, which is useful in nonlinear time-series analysis, especially in connection with estimating the embedding dimension [2].

We cannot directly measure these Renyi entropies, but must instead estimate them. These estimates are based on a limited amount of data consisting of a set of samples of the distribution. [Often it is assumed the samples are generated in an independent and identically distributed (IID) manner.] Therefore we are directly confronted with the problem of inferring a function of a distribution from a set of samples of the distribution.

It turns out that to accurately measure the information dimension we would like to make accurate estimates of the Renyi entropy for as wide a range of granularities of the discretization as possible. In particular, we would like to make accurate estimates when the discretization is quite fine. In such a regime, the number of counts per bin, i.e., the values n_i, will be quite small. Accordingly, we are unavoidably faced with the "small sample statistics problem" of how to meaningfully perform inference with small samples. This is precisely the regime in which Bayesian techniques, the subject matter of this paper, excel.

Example 2. In much of physics and engineering it is currently conventional to measure the degree to which a pair of random variables influence each other through correlations, covariances, and the like. More sophisticated information theoretic techniques such as mutual information [3–6] are also finding their way into physics [7], as well as into other fields such as computational biology [8]. (See also the discussion in Sec. VI.) All such measures can be cast as functionals $Q(\boldsymbol{\rho})$, where $\boldsymbol{\rho}$ is the underlying probability distribution over the variables in question.

Whatever the influence measure one prefers, to use it in practice, i.e., to estimate its value for a particular physical system using some experimental data, means using a data set $\mathbf{n}$ to estimate $Q(\boldsymbol{\rho})$. For example, in Ref. [8], the experimental data are a data base of HIV-1 V3 loop amino acid sequences. The goal is to find pairs of sites on the amino acid sequences that have high mutual information. (The idea is that such sites must vary together for the virus to be viable). So for any given pair of sites the data base provides a set of counts n_i of each of the $20\times20=400$ possible amino acid pairs (there are 20 possible amino acids per site). Our goal is to use such an $\mathbf{n}$ to estimate the mutual information of the $\boldsymbol{\rho}$ generating the set of HIV sequence values at that pair of sites.

In this paper we investigate the case where $Q(\,)$ is the Shannon entropy [9–11] $S(\boldsymbol{\rho})=-\sum_i\rho_i\ln(\rho_i)$. In Ref. [12] we analyze the cases where $Q(\,)$ is either the mutual information, χ^2, covariance, variance, or the average. (The conclusions of that analysis are presented in the

*Electronic address: dhw@santafe.edu
†Electronic address: wolf@lanl.gov

conclusion to this paper.) Some previous work on estimating $Q(\rho)$ from $\mathbf{n}$ (most closely related to the work of Ref. [12]) appears in Refs. [1,3,13–20].

We note in passing that the intuitive notion of the Shannon entropy of a distribution as the "amount of missing information" in that distribution is not usually considered meaningful if the information at hand consists of data $\mathbf{n}$ rather than the underlying distribution ρ, since Shannon entropy is a function of ρ rather than of $\mathbf{n}$. In the sense that the Bayes estimator discussed in this paper is optimal and produces a Shannon entropy value from information of the form of samples $\mathbf{n}$, the Bayes estimator can be viewed as a way of defining the "amount of missing information" when the information at hand consists of a finite data set $\mathbf{n}$ rather than a full distribution ρ.

In Sec. II of this paper we introduce the Bayes estimator for $Q(\rho)$ given $\mathbf{n}$. In Sec. III we discuss the optimal properties of Bayes estimators and discuss their relation to conventional statistical techniques. Section IV contains the central mathematical results needed to calculate Bayes estimators for $Q(\rho)$. We then apply these results to the case where $Q(\rho)$ is the Shannon entropy. Section V A contains a brief calculation showing that for small sample sizes there are significant differences between the Bayes and frequency-counts $[S(\mathbf{n})=-\sum_i (n_i/N)\ln(n_i/N)]$ estimators for the Shannon entropy. In Sec. V B we present graphs of the results of a numerical comparison of the Bayes and frequency-counts estimators for the Shannon entropy. These graphs illustrate the bias, variance, etc., associated with both the Bayes and frequency-counts estimators.

Fully formal justifications for some of the manipulations carried out in this paper (e.g., interchange of integration and differentiation) can be found as appendixes to Ref. [12]. Reference [12] also analyzes a number of priors not discussed in this paper.

II. BAYESIAN ESTIMATION OF $Q(\rho)$ FROM COUNTS

To estimate $Q(\rho)$ from the data $\mathbf{n}$, it is necessary to find the probability density function (PDF) $P(\rho|\mathbf{n})$. (See Ref. [21] for a formal discussion of our statistical notation.) First note that $P(\mathbf{n}|\rho)=N!\prod_{i=1}^m[\rho_i^{n_i}/n_i!]$. By Bayes's theorem the PDF $P(\rho|\mathbf{n})$ is given by

$$P(\rho|\mathbf{n})=P(\mathbf{n}|\rho)P(\rho)/P(\mathbf{n}) , \tag{1}$$

where $P(\mathbf{n})=\int d\rho\, P(\mathbf{n}|\rho)P(\rho)$ and $P(\rho)$ has support only on the simplex $R\equiv\{\rho\colon \rho_i\geq 0\ \forall i,\ \sum_i\rho_i=1\}$. $P(\rho|\mathbf{n})$ is called the "posterior PDF," $P(\mathbf{n}|\rho)$ is called the "likelihood," and $P(\rho)$ is called the "prior PDF." Unless otherwise stated, integrals over ρ are understood to be definite integrals over the region extending from 0 to ∞ in each ρ_i.

Note that because of cancellation, the constant $N!/(\prod_{i=1}^m n_i!)$ does not appear in $P(\rho|\mathbf{n})$. Accordingly we can simply write $P(\rho|\mathbf{n})\propto P(\rho)\prod_{i=1}^m\rho_i^{n_i}$ with the proportionality constant (dependent on $\mathbf{n}$ only) set by normalization. Finally, the PDF of $Q(\rho)$ given $\mathbf{n}$ is given in terms of $P(\rho|\mathbf{n})$ by

$$P(Q(\rho)=q|\mathbf{n})=\int d\rho\,\delta(Q(\rho)-q)P(\rho|\mathbf{n}) . \tag{2}$$

Consider the situation where what we know is $\mathbf{n}$ and what we wish to know is $Q(\rho)$. For this usual situation, it is the distribution $P(Q(\rho)=q|\mathbf{n})$ appearing in Eq. (2)—and this one alone—that tells us what we want. By way of contrast, distributions that do not depend on a prior (e.g., likelihood-based quantities) do *not* tell us what we wish to know. This, along with other arguments given below, is why we focus on $P(Q(\rho)=q|\mathbf{n})$ in this paper.

Rather than trying to find the density $P(Q(\rho)=q|\mathbf{n})$ directly, it is often simpler to find its moments. The kth moment of $Q(\rho)$ given $\mathbf{n}$ is given by $\int dq\, q^k P(Q(\rho)=q|\mathbf{n})=\int d\rho\, Q^k(\rho)P(\rho|\mathbf{n})$, i.e., the kth moment of $Q(\rho)$ given $\mathbf{n}$ is the posterior average of $Q^k(\rho)$ according to the posterior distribution $P(\rho|\mathbf{n})$. Define q_k by

$$q_k\equiv\int d\rho\, Q^k(\rho)P(\rho)\prod_{i=1}^m\rho_i^{n_i} . \tag{3}$$

(When we have a particular Q in mind we will change the letter 'q' appropriately, e.g., when Q is the entropy, we will refer to s_k rather than q_k.) Using Eq. (1) for $P(\rho|\mathbf{n})$ we see that the kth moment of $Q(\rho)$ given $\mathbf{n}$ is given by q_k/q_0. We refer to this ratio as the "Bayes estimator with prior $P(\rho)$ for $Q^k(\rho)$."

In particular, the Bayes estimators for $Q(\rho)$ and $Q^2(\rho)$ can be used to find the standard deviation of $Q(\rho)$. This in turn may be used with Chebyshev's inequality [20] to bound the probability of deviation of $Q(\rho)$ from the Bayes estimator's guess for $Q(\rho)$. This constitutes a "Bayesian error bar" associated with our estimating the value of $Q(\rho)$ as q_1/q_0. Note that this error bar does not rely on an assumption that the underlying distributions are Gaussian.

To proceed further it is necessary to make an assumption for the prior PDF $P(\rho)$; once this is done, $P(\rho|\mathbf{n})$ and q_k/q_0 are uniquely determined. In the calculations to follow, $P(\rho)$ will be assumed to be a uniform prior over the unit simplex, i.e., it will be assumed to have the form $P(\rho)\propto\Delta(\rho)\Theta(\rho)$, where $\Theta(\rho)\equiv\prod_i\theta(\rho_i)$, [$\theta$ is the Heaviside theta function $\theta(x)=1$ for $x\geq 0$ and 0 otherwise], $\Delta(\rho)\equiv\delta(\sum_i\rho_i-1)$, and the proportionality constant is set by the normalization condition $\int d\rho\, P(\rho)=1$. [The $\Theta(\ )$ enforces the non-negativity of the ρ_i and the $\Delta(\ )$ enforces the condition that the sum of the ρ_i equals 1.]

We emphasize that here we are using the uniform prior only for reasons of expository simplicity. In many problems the uniform prior is inappropriate and a different prior should be used. In Ref. [1] we consider the extension of our results to a broader class of priors than those considered in this paper. The general trick we use for such nonuniform priors is to express them as a linear combination of monomials $(\rho_i)^{k_i}$ and then note that the Bayes estimator with a prior (proportional to) $(\rho_i)^{k_i}\Delta(\rho)\Theta(\rho)$ is the same as the Bayes estimator with a prior $\Delta(\rho)\Theta(\rho)$, provided in the latter estimate each n_i is incremented by k_i. [See Eq. (3).]

As an example of such a nonuniform prior, the entro-

pic prior $P(\rho)=e^{\alpha S}$, where S is the Shannon entropy and α is some constant, is related to the popular technique of maximum entropy [22,23]. As another example, the Dirichlet prior, $P(\rho)\propto\sum_{i=1}^{m}\rho_i^a$ for some constant a, has also been considered in some contexts [24]. It is also sometimes appropriate to use a prior that does not allow the probability of certain states to differ from zero [25]. In Ref. [12], Bayes estimators for both entropic and Dirichlet priors are discussed.

For simplicity of presentation define

$$I[Q(\rho),\mathbf{n}]\equiv\int d\rho Q(\rho)\,\Delta(\rho)\Theta(\rho)\prod_{i=1}^{m}\rho_i^{n_i}\ . \tag{4}$$

Note that $I[\ ,\ ]$ is a functional of its first argument and a function of its second argument. With this notation the Bayes estimator with uniform prior for $Q^k(\rho)$, [i.e., q_k/q_0 with $P(\rho)$ uniform] is given by $I[Q^k(\rho),\mathbf{n}]/I[1,\mathbf{n}]$. [For nonuniform $P(\rho),q_k/q_0$ is given by a different ratio of integrals.]

III. BAYES ESTIMATORS MINIMIZE MEAN-SQUARED ERROR

Before evaluating the integrals $I[Q(\rho),\mathbf{n}]$ we briefly discuss an optimality property of Bayes estimators and relate these estimators to some classical estimation techniques. If the true probabilities are fixed to a particular ρ, then the mean-squared error when using an estimator $G(\mathbf{n})$ to estimate $Q(\rho)$ is given by

$$\sum_{\mathbf{n}}P(\mathbf{n}|\rho)[G(\mathbf{n})-Q(\rho)]^2\ . \tag{5}$$

For a fixed ρ, (5) is minimized by choosing $G(\mathbf{n})$ independent of the $\mathbf{n}$: $G(\mathbf{n})=Q(\rho)$.

More generally, when ρ is not fixed and is distributed according to $P(\rho)$, the mean-squared error is given by

$$\int d\rho\,P(\rho)\sum_{\mathbf{n}}P(\mathbf{n}|\rho)[G(\mathbf{n})-Q(\rho)]^2\ . \tag{6}$$

As conventional in the calculus of variations [26], to find the $G(\ )$ to minimize this expression write $G(\ )=G_0(\ )+\alpha\eta(\ )$, differentiate (6) with respect to α, and then evaluate the result at $\alpha=0$. Doing this yields

$$\sum_{\mathbf{n}}\eta(\mathbf{n})\int d\rho\,P(\mathbf{n}|\rho)P(\rho)[G_0(\mathbf{n})-Q(\rho)]=0\ . \tag{7}$$

Since this equality must hold for all $\eta(\ )$, for all $\mathbf{n}$

$$\int d\rho\,P(\mathbf{n}|\rho)P(\rho)[G_0(\mathbf{n})-Q(\rho)]=0\ . \tag{8}$$

Equation (8) is solved [assuming $\int d\rho\,P(\mathbf{n}|\rho)P(\rho)\neq 0$] by

$$G_0(\mathbf{n})=\int d\rho\,P(\mathbf{n}|\rho)P(\rho)Q(\rho)\Big/\int d\rho\,P(\mathbf{n}|\rho)P(\rho)$$
$$=q_1/q_0\ . \tag{9}$$

Note that Eq. (9) holds for any prior $P(\rho)$. Given the discussion in Sec. II, Eq. (9) shows that $G_0(\mathbf{n})$, the estimator having minimal mean-squared error from $Q(\rho)$, is identical to the Bayes estimator for $Q(\rho)$: $G_0(\mathbf{n})=\int d\rho\,P(\rho|\mathbf{n})Q(\rho)$. (One can derive this particular result more simply; the derivation here is pedagogical in that if one wants to optimize some other functionals, the relatively complicated approach presented here is needed. In particular, if one is interested in "least bias" estimators, this is the case. See Ref. [27].)

As an example consider the famous Laplace sample size correction estimator [24], in which the underlying ρ_i are estimated from counts $\mathbf{n}$ by $\hat{\rho}_i=(n_i+1)/(N+m)$. This estimator is precisely the Bayes estimator with uniform prior, for $Q(\rho)=\rho$ (see results in Ref. [12]). Note that for small n_i the Bayes estimator is notably different from the frequency count estimator $\hat{\rho}_i=n_i/N$.

Of course, none of this means that a Bayesian technique is optimal if the prior it uses is poorly chosen, i.e, if the prior the researcher uses does not match the one generating the data. This is a general feature of Bayesian approaches; they are "only as good as the prior." We admonish the reader to choose their prior with careful attention to the problem at hand when using the techniques we present here.

As an aside, note that when $Q(\ )$ is nonlinear and not injective [e.g., when $Q(\ )$ is the Shannon entropy], one cannot evaluate the Bayes estimate for $Q(\ )$ by calculating Q of the Bayes estimate for ρ, i.e., Q of an average is generally not the same as the average of Q. (Formally, for nonlinear $Q(\ )$, $Q[(n_i+1)/(N+m)]\neq q_1/q_0$, in general.) For these kinds of $Q(\ )$ one must take into account the probabilities of all ρ's to evaluate the Bayes estimator for $Q(\rho)$ and its associated error bars.

In general, one might not want to take the mean q evaluated according to the PDF $P(Q(\rho)=q|\mathbf{n})$ to form an estimate for $Q(\rho)$. For example, one might be interested in minimizing (the average of) $|G(\mathbf{n})-Q(\rho)|$ rather than $|G(\mathbf{n})-Q(\rho)|^2$, a goal that generically results in an estimate of the median of the PDF rather than its mean. As another example, it might be of interest to minimize something other than a functional of the error $G(\mathbf{n})-Q(\rho)$. An instance of this appears in Ref. [12], which discusses minimizing the mean-squared bias to find what might be called a "Bayes minimum-bias estimator." (See also Sec. V A, which discussed numerical calculations of biases and variances of the Bayes and frequency-counts estimators.)

As yet another example, in the non-Bayesian technique of maximum-likelihood estimation, for the case where $Q(\rho)=\rho$, one estimates $Q(\rho)$ as the $Q(\rho)$ that maximizes the likelihood $P(\mathbf{n}|Q(\rho))$. (See Ref. [28].) This corresponds to the Bayesian procedure of finding the mode of $P(Q(\rho)=q|\mathbf{n})$ [assuming the prior over $q=Q(\rho)$ is uniform]. When $Q(\rho)=\rho$ the result is the frequency-counts estimate $\rho_i=n_i/N$.

Note that techniques such as maximum likelihood have the advantage that (unlike Bayesian techniques) their predictions do not depend on (what is usually) an assumption for the prior. In this trivial sense, they do not degrade if one makes a poor assumption for the prior. On the other hand, such techniques usually cannot be cast as minimizing some functional of $Q(\rho)$ where ρ is not fixed. In this case, they cannot be cast as a technique that minimizes

some expected real world "cost" or "loss" [29]. Another advantage of Bayesian techniques is that they make all assumptions explicit, by putting them in the prior [30]. In addition, in that they are determined by the posterior $P(Q|\mathbf{n})$, as mentioned just below Eq. (2), Bayesian techniques concern themselves with "what we want." (See Ref. [31] for a more general discussion of the relative strengths and weaknesses of Bayesian and non-Bayesian techniques.)

IV. CALCULATION OF THE BAYES ESTIMATOR FOR SHANNON ENTROPY

As shown in Sec. II, finding the Bayes estimator with a uniform prior for $Q^k(\boldsymbol{\rho})$ reduces to evaluating integrals of the form $I[Q^k(\boldsymbol{\rho}),\mathbf{n}]$. This section presents the main techniques for calculating these integrals and uses them to calculate the Bayes estimator when Q is the Shannon entropy.

Readers interested only in the Shannon entropy results may skip directly to Sec. IV E. In Sec. IV A we derive an important result that allows integrals such as $I[\ ]$ to be recast as Laplace convolution products. In Sec. IV B we outline the general procedure, based on the results of Sec. IV A, for calculating the moments of $Q(\boldsymbol{\rho})$. In the remaining subsections we apply the procedure of Sec. IV B to the case in which $Q(\boldsymbol{\rho})$ is the Shannon entropy: Section IV C contains a calculation of $I[1,\mathbf{n}]$ and in Sec. IV D we present a calculation for those integrals that, along with $I[1,\mathbf{n}]$, give the Bayes estimator of the Shannon entropy when the prior is uniform.

A. Convolution form of the integrals

In this subsection two important results are given. First, in Theorem 1 it is shown that if a function $H(\boldsymbol{\rho})$ factors as $H(\boldsymbol{\rho})=\prod_{i=1}^m h_i(\rho_i)$, then the general form of the integral $\int d\boldsymbol{\rho}\, H(\boldsymbol{\rho})\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})$ is that of a convolution product of m terms (recall that m is the number of possible outcomes of the process under observation). Second, Laplace's convolution theorem is given.

Define the Laplace convolution operator $\otimes$ by $(f\otimes g)(\tau)\equiv\int_0^\tau dx\, f(x)g(\tau-x)$.

Theorem 1. If $H(\boldsymbol{\rho})=\prod_{i=1}^m h_i(\rho_i)$, then $\int d\boldsymbol{\rho}\,\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})H(\boldsymbol{\rho})=(\otimes_{i=1}^m h_i(\rho_i))(\tau)|_{\tau=1}$.

Proof. The ρ_i may not be independently integrated since the constraint $\sum_{i=1}^m\rho_i=1$ exists. This constraint is reflected in the explicit definition of the integral

$$\int d\boldsymbol{\rho}\,\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})H(\boldsymbol{\rho})=\int_0^\infty d\rho_1\cdots\int_0^\infty d\rho_m\{h_1(\rho_1)\times\cdots\times h_m(\rho_m)\}\delta\left[1-\sum_{i=1}^m\rho_i\right]$$

$$=\int_0^1 d\rho_1 h_1(\rho_1)\int_0^{1-\rho_1}d\rho_2 h_2(\rho_2)\cdots$$

$$\times\int_0^{1-(\rho_1+\cdots+\rho_{m-2})}d\rho_{m-1}h_{m-1}(\rho_{m-1})h_m(1-(\rho_1+\cdots+\rho_{m-1}))\ .$$

Define the m variables τ_k, $k=1,\ldots,m$, recursively by $\tau_1\equiv\sum_{i=1}^m\rho_i=1$ and $\tau_k\equiv\tau_{k-1}-p_{k-1}$. Since $\tau_k=\tau_1-\sum_{i=1}^{k-1}p_i$, our integral may be rewritten as

$$\int d\boldsymbol{\rho}\,\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})H(\boldsymbol{\rho})=\int_0^{\tau_1}d\rho_1h_1(\rho_1)\int_0^{\tau_2}d\rho_2h_2(\rho_2)\cdots\int_0^{\tau_{m-1}}d\rho_{m-1}h_{m-1}(\rho_{m-1})h_m(\tau_{m-1}-\rho_{m-1})\ .$$

Now, with the definition of the convolution, the integral can be rewritten as

$$\int d\boldsymbol{\rho}\,\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})H(\boldsymbol{\rho})=\int_0^{\tau_1}d\rho_1h_1(\rho_1)\cdots\int_0^{\tau_{m-2}}d\rho_{m-2}h_{m-2}(\rho_{m-2})(h_{m-1}\otimes h_{m-2})(\tau_{m-2}-\rho_{m-2})\ .$$

Since the convolution operator is both commutative and associative, we can repeat this procedure and write the integral above with obvious notation as

$$\int d\boldsymbol{\rho}\,\Delta(\boldsymbol{\rho})\Theta(\boldsymbol{\rho})H(\boldsymbol{\rho})=\left[\bigotimes_{i=1}^m h_i(\rho_i)\right](\tau)|_{\tau=1}\ .$$

Q.E.D.

Theorem 2 is the Laplace convolution theorem and is stated for completeness only. The proof may be found in Ref. [32]. Define the Laplace transform operator L by $L[h](s)=\int_0^\infty h(t)e^{-st}dt$.

Theorem 2. If $L[h_i(\rho_i)]$ exists for $i=1,\ldots,m$, then $L[\otimes_{i=1}^m h_i(\rho_i)]=\prod_{i=1}^m L[h_i(\rho_i)]$.

B. Outline of general procedure

Theorems 1 and 2 allow the calculation of integrals $I[Q^k(\boldsymbol{\rho}),\mathbf{n}]$ for functions of the form $Q(\boldsymbol{\rho})=\sum_{i=1}^k\prod_{j=1}^m Q_{ij}(\rho_j)$, which we call "factorable." Here we briefly summarize the procedure to be used.

(i) For each $Q_{ij}(\rho_j)$, calculate the Laplace transform of $h_{ij}(\rho_j)\equiv Q_{ij}(\rho_j)\rho_j^n$.

(ii) Calculate $\sum_{i=1}^k\prod_{j=1}^m L[h_{ij}(\rho_j)]$.

(iii) Take the inverse Laplace transform of the term calculated in (ii) and evaluate it for an argument of 1.

As an example, let $Q(\boldsymbol{\rho})=S(\boldsymbol{\rho})=-\sum_{i=1}^m\rho_i\ln(\rho_i)$. All powers of $S(\boldsymbol{\rho})$ are factorable terms. Therefore, the procedure outlined above may be used to find the Bayes esti-

mators with a uniform prior for any power of $S(\rho)$, as shown in detail in the remainder of this section.

C. Calculation of $I[1,n]$

In the next theorem the Laplace transform is used in concert with Theorems 1 and 2 to calculate the normalization constant $I[1,\mathbf{n}]$. With the Gamma function $\Gamma(z)$ given by $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$ for $\mathrm{Re}(z)>-1$ we have the following.

Theorem 3. If $\mathrm{Re}(n_i)>-1 \quad \forall i=1,\ldots,m$, then $I[1,\mathbf{n}]=\prod_{i=1}^m \Gamma(n_i+1)/\Gamma[N+m]$.

Proof. For the integral $I[1,\mathbf{n}]=\int d\rho\,\Delta(\rho)\Theta(\rho)\prod_{i=1}^m \rho_i^{n_i}$, the $h_i(\rho_i)$ of Theorem 1 are given by $h_i(\rho_i)=\rho_i^{n_i}$. Since

$$L[\rho^n](s)=\frac{\Gamma(n+1)}{s^{n+1}} \quad \text{for } n>-1\ ,$$

we have, by Theorems 1 and 2

$$\begin{aligned} I[1,\mathbf{n}] &= L^{-1}\left[\prod_{i=1}^m L[\rho^{n_i}](s)\right](\tau)\Big|_{\tau=1} \\ &= L^{-1}\left[\prod_{i=1}^m \Gamma(n_i+1)s^{-(n_i+1)}\right](\tau)\Big|_{\tau=1} \\ &= \frac{\prod_{i=1}^m \Gamma(n_i+1)}{\Gamma(N+m)}\ . \end{aligned}$$

Q.E.D.

The same result can also be derived using somewhat more laborious change-of-variable tricks. See the section on Dirichlet distributions in Ref. [33].

D. Calculation of $I[\rho_1^{q_1}\ln^{r_1}(\rho_1)\cdots\rho_m^{q_m}\ln^{r_m}(\rho_m),\mathbf{n}]$

As mentioned in Sec. IV B, since $S(\rho)=-\sum_i \rho_i\ln(\rho_i)$, powers of $S(\rho)$ are sums of terms, each of which have the form $\rho_1^{q_1}\ln^{r_1}(\rho_1)\cdots\rho_m^{q_m}\ln^{r_m}(\rho_m)$. Thus, to find the Bayes estimators for arbitrary powers of the Shannon entropy, expressions of the form

$$I[\rho_1^{q_1}\ln^{r_1}(\rho_1)\cdots\rho_m^{q_m}\ln^{r_m}(\rho_m),\mathbf{n}]$$

must be calculated. Using the fact that $\partial_n^r\rho^n=\rho^n\ln^r(\rho)$, we immediately have the following.

Theorem 4. For $\mathrm{Re}(n_i)>-1\ \forall i$,

$$I[\ln^{r_1}(\rho_1)\cdots\ln^{r_m}(\rho_m),\mathbf{n}]=\partial_{n_1}^{r_1}\cdots\partial_{n_m}^{r_m}I[1,\mathbf{n}]\ .$$

The justification for the needed interchange of derivative and integral is given in Appendix C of Ref. [1]. In using Theorem 4, note that since $N=\sum_{i=1}^m \mathbf{n}_i$, we have $\partial_{\mathbf{n}_i}N=1$.

For expository simplicity, before presenting Theorem 5 we introduce the definitions $\Phi^{(n)}(z)\equiv\Psi^{(n-1)}(z)$ and $\Delta\Phi^{(n)}(z_1,z_2)\equiv\Phi^{(n)}(z_1)-\Phi^{(n)}(z_2)$, where $\Psi^{(n)}(z)$ is the polygamma function $\Psi^{(n)}(z)=\partial_z^{n+1}\ln[\Gamma(z)]$ [34]. This definition of Φ is made to facilitate the clean presentation of results; $\Phi^{(n)}(z)=\partial_z^n\ln[\Gamma(z)]$.

Theorems 5 and 6 apply Theorem 4 to the calculation of the integral $I[\ln^{r_1}(p_1)\cdots\ln^{r_m}(p_m),\mathbf{n}]$ for some special cases.

Theorem 5. For $\mathrm{Re}(n_i)>-1\ \forall i$,

$$\begin{aligned} &I[\ln(\rho_u),\mathbf{n}] \\ &\quad=\Delta\Phi^{(1)}(n_u+1,N+m)\left\{\prod_{i=1}^m \Gamma(n_i+1)\right\}\Big/\Gamma(N+m)\ . \end{aligned}$$

Proof. $I[\ln(\rho_u),\mathbf{n}]=\partial_{n_u}I[1,\mathbf{n}]$ (by Theorem 4). Substituting the result from Theorem 3 for $I[1,\mathbf{n}]$ above we find

$$\begin{aligned} \partial_{n_u}\frac{\prod_{i=1}^m \Gamma(n_i+1)}{\Gamma(N+m)} &= \prod_{i\neq u}\Gamma(n_i+1)\partial_{n_u}\frac{\Gamma(n_u+1)}{\Gamma(N+m)} \\ &= \prod_{i\neq u}\Gamma(n_i+1)\frac{\Gamma(n_u+1)}{\Gamma(N+m)}\Delta\Phi^{(1)}(n_u+1,N+m) \quad \text{(by definition of } \Phi) \\ &= \frac{\prod_{i=1}^m \Gamma(n_i+1)}{\Gamma(N+m)}\Delta\Phi^{(1)}(n_u+1,N+m)\ . \end{aligned}$$

Q.E.D.

Theorem 6. For $\mathrm{Re}(n_i)>-1\ \forall i$,

$$\begin{aligned} I[\ln(\rho_u)\ln(\rho_v),\mathbf{n}] &= \left[\prod_{i=1}^m \Gamma(n_i+1)\right]\Big/\Gamma(N+m) \\ &\quad\times\{\Delta\Phi^{(1)}(n_u+1,N+m)\Delta\Phi^{(1)}(n_v+1,N+m)-\Phi^{(2)}(N+m)\}, \quad u\neq v \\ I[\ln^2(\rho_u),\mathbf{n}] &= \left[\prod_{i=1}^m \Gamma(n_i+1)\right]\Big/\Gamma(N+m)\{[\Delta\Phi^{(1)}(n_u+1,N+m)]^2+\Delta\Phi^{(2)}(n_u+1,N+m)\}\ . \end{aligned}$$

Proof. Similar to proof of Theorem 5.

E. Bayes estimators for moments of the Shannon entropy

In this subsection the results for the Bayes estimators with uniform prior for the first two powers of $S(\rho)$ are given, i.e., we give s_1/s_0 and s_2/s_0. Refer to Secs. IV A–IV D for the calculations used here and to Sec. IV D for the definitions of the functions $\Phi^{(n)}(z)$ and $\Delta\Phi^{(n)}(z_1,z_2)$.

Theorem 7. For $\mathrm{Re}(n_i) > -1\ \forall i$, $s_1/s_0 = -\sum_i [(n_i+1)/(N+m)]\Delta\Phi^{(1)}(n_i+2, N+m+1)$.

Proof. Define $\mathbf{e}_i$ as the m vector with a 1 in index i, 0's everywhere else. Then

$$
\begin{aligned}
s_1/s_0 &= I\left[\sum_i \rho_i \ln(\rho_i), \mathbf{n}\right] \Big/ I[1,\mathbf{n}] \\
&= -\sum_i I[\ln(\rho_i), \mathbf{n}+\mathbf{e}_i]/I[1,\mathbf{n}] \quad \text{(by definition of } I[,]) \\
&= -\sum_i \partial_{n_i} I[1,\mathbf{n}+\mathbf{e}_i]/I[1,\mathbf{n}] \quad \text{(by Theorem 4)} \\
&= -\sum_i \frac{n_i+1}{N+m}\Delta\Phi^{(1)}(n_i+2, N+m+1) \quad \text{(by Theorems 5 and 3)}.
\end{aligned}
$$

Q.E.D.

Theorem 8. For $\mathrm{Re}(n_i) > -1\ \forall i$,

$$
\begin{aligned}
s_2/s_0 = &\sum_{i\neq j} \frac{(n_i+1)(n_j+1)}{(N+m)(N+m+1)}\{\Delta\Phi^{(1)}(n_i+2, N+m+2)\Delta\Phi^{(1)}(n_j+1, N+m+2) - \Phi^{(2)}(N+m+2)\} \\
&+ \sum_i \frac{(n_i+1)(n_i+2)}{(N+m)(N+m+1)}\{[\Delta\Phi^{(1)}(n_i+3, N+m+2)]^2 + \Delta\Phi^{(2)}(n_i+3, N+m+2)\}.
\end{aligned}
$$

Proof. Similar to proof of Theorem 7.

Note that the computational evaluation of s_1/s_0 grows only linearly with m. Similarly, the calculation of s_2/s_0 grows quadratically. Moreover, the terms in the summands are no more difficult to evaluate than other typical transcendental functions; our entire procedure is extremely quick and easy, computationally speaking.

In a manner similar to the calculation of s_1 and s_2, all higher moments of $S(\rho)$ are calculable via differentiation [since $\partial_z \Phi^{(n)}(z) = \Phi^{(n+1)}(z)$]. Note that when no data have been observed, i.e., $\mathbf{n}=\mathbf{0}$, the estimator for s_1/s_0 is simply $-\Delta\Phi^{(1)}(2, m+1) = \sum_{i=2}^{m} i^{-1}$. (This may help the reader in deciding whether a uniform prior makes sense for the application they have in mind.) It should also be noted that as $N\to\infty$, the Bayes estimator $s_1/s_0 \to \sum_i (n_i/N)\ln(n_i/N)$ [34], i.e., it asymptotically becomes the frequency-counts estimator.

V. BAYES ESTIMATORS VS FREQUENCY-COUNTS ESTIMATORS

In this section we compare the Bayes estimator (see Theorem 7) and the frequency-counts estimation for the entropy in two ways. First, in Sec. V A an explicit calculation of the two estimators is made for two specific cases where a small number of counts are observed in two bins ($m=2$). This simple calculation points out that for small N there are significant differences in the values of the two estimators. Second, in Sec. V B the two estimators are graphically compared for a range of sample sizes and true underlying distributions.

A. Small N

It is always desirable to have a large amount of data. Often, however, this is not possible. One of the strengths of Bayesian analysis is its power for dealing with such small-data cases. In particular, not only are Bayesian estimators in many respects more "reasonable" than non-Bayesian estimators for small data, they also naturally provide error bars to govern one's use of their results.

For our case, for small N, the Bayes estimate s_1/s_0 can differ considerably from the estimate one would make using the frequency-counts estimator $S(\mathbf{n}) = -\sum_{i=1}^{m}(n_i/N)\ln(n_i/N)$. In addition, the Bayesian formalism automatically tells you when it is unsure of its estimate, through its error bars. Both points are illustrated by the following pair of examples:

Example 1. Assume two possible events ($m=2$). Let $n_1=0$ and $n_2=2$. $s_1/s_0 = 0.458$. The entropy estimate obtained using the frequency-counts estimator is 0. Note that the standard deviation of the Bayesian estimate [i.e., the square root of $s_1/s_0 - (s_1/s_0)^2$] is quite large. This indicates that we do not have strong confidence in the answer 0.458 and reflects the fact that the sample size is small.

Example 2. Again, $m=2$. Assume that $n_1=1$ and $n_2=4$. $s_1/s_0 = 0.533$. The entropy estimate obtained using the frequency-counts estimator is 0.5.

Note that there are "edge effects" in using s_1/s_0 as the estimate for the entropy. If the true ρ is uniform ($\rho_i = m^{-1}\ \forall i$), then $S(\rho)$ is maximal and always exceeds s_1/s_0, no matter what the observed $\mathbf{n}$ are. This is be-

cause the estimate s_1/s_0 takes into account all possible ρ that might have generated the observed $\mathbf{n}$, including all those with a smaller entropy than the true (maximal) entropy. In a similar fashion, if the true $S(\rho)$ is minimal, then it is always exceeded by s_1/s_0, regardless of the value of the observed $\mathbf{n}$.

B. Graphical results of numerical comparisons

The graphs appearing in Figs. 1–5 depict several comparisons of the Bayes and frequency-counts estimators for entropy. In all cases the solid line represents the Bayes estimator, the dash-dotted line represents the frequency-counts estimator, and the dotted line represents the true value of the entropy, where applicable. Figure 6 depicts the PDF of the Bayes estimator for a fixed ratio of counts as the number of counts increases. The graphs are the result of exact numerical computations of the various quantities represented.

Figure 1 explicitly demonstrates the result of Sec. III of this paper for the Shannon entropy with $m=2$. Recall that this section shows that the Bayes estimator is the minimal mean-squared error estimator. As is immediately seen in Fig. 1, for all N the Bayes estimator has a smaller mean-squared error than the frequency-counts estimator, where the mean-squared error for an estimator $S(\mathbf{n})$ is given by

$$\int d\rho\, P(\rho) \sum_{\mathbf{n}} P(\mathbf{n}|\rho)[S(\mathbf{n})-S(\rho)]^2 . \tag{10}$$

The curves were generated with $P(\rho)$ uniform. The Bayes estimator is that of Theorem 7, which assumes this uniform $P(\rho)$.

Figure 2 depicts the average over ρ of the sample variance, that is,

$$\int d\rho\, P(\rho) \sum_{n} P(\mathbf{n}|\rho) \left[S(\mathbf{n})- \sum_{\mathbf{n}'} P(\mathbf{n}'|\rho)S(\mathbf{n}')\right]^2 . \tag{11}$$

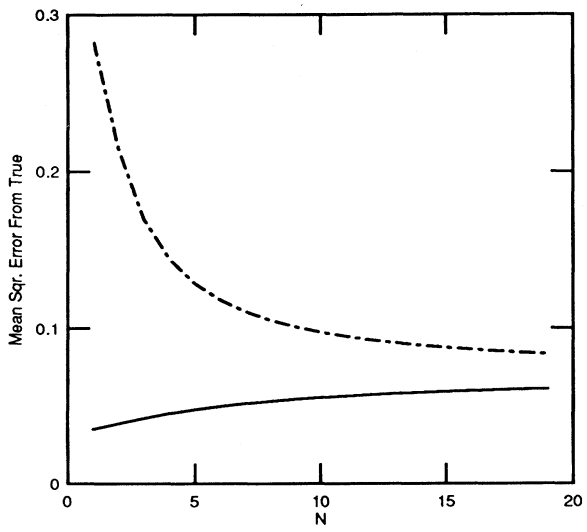

FIG. 1. Mean square error [Eq. (10)] for the Bayes (solid line) and frequency-counts (dash-dotted line) estimators of the entropy $S(\rho)=-\sum_{i=1}^{2}\rho \ln(\rho_i)$.

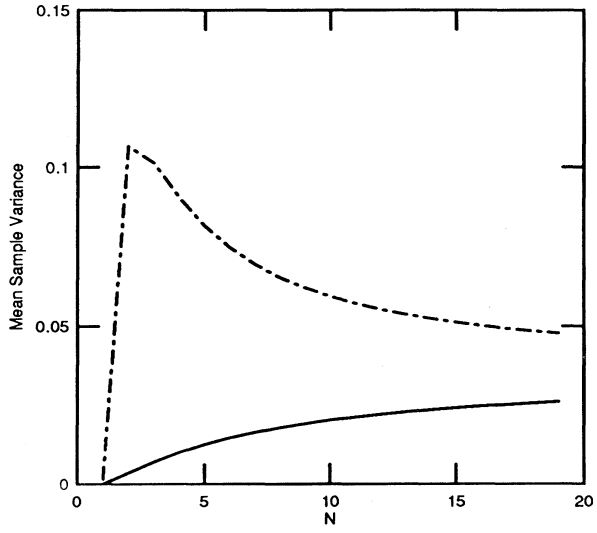

FIG. 2. Mean sample variance [Eq. (11)] of the Bayes (solid line) and frequency-counts (dash-dotted line) estimators of the entropy $S(\rho)=-\sum_{i=1}^{2}\rho \ln(\rho_i)$. Both variances are shown as functions of the sample size N.

[Again, $P(\rho)$ is uniform.] This figure shows how, for a particular sample size N, the estimators deviate from their sample averages. It is immediately seen that the Bayes estimator has a smaller sample variance. (This is in agreement with the conservative edge effects behavior of the Bayes estimator, which was mentioned in Sec. V A.) This result is useful for understanding Figs. 3 and 4.

Figure 3 shows the sample averages of the estimators as functions of the sample size N for various values of the true ρ, that is,

$$\sum_{\mathbf{n}} P(\mathbf{n}|\rho)S(\mathbf{n}) . \tag{12}$$

Figure 4 shows the same sample averages of the estimators, but now as functions of the true ρ for various values of the sample size N.

It is of interest to note that for a particular range of ρ values and sufficiently large N, the sample average of the frequency-counts estimator actually comes closer to the true entropy than does the sample average of the Bayes estimator [see Figs. 3(d)–3(f) and 4(d)–4(f)]. To see how this is possible in light of the fact that the Bayes estimator has lower mean-squared error, first note that

$$\begin{aligned}\int d\rho\, P(\rho) \sum_{\mathbf{n}} P(\mathbf{n}|\rho)[S(\mathbf{n})-S(\rho)]^2 \\ = \int d\rho\, P(\rho) \sum_{\mathbf{n}} P(\mathbf{n}|\rho) \left[S(\mathbf{n})- \sum_{\mathbf{n}'} P(\mathbf{n}'|\rho)S(\mathbf{n}')\right]^2 \\ + \int d\rho\, P(\rho) \left[\sum_{\mathbf{n}'} P(\mathbf{n}|\rho)S(\mathbf{n})-S(\rho)\right]^2 ,\end{aligned} \tag{13}$$

i.e., the mean-squared error is the sum of the mean sample variance and the mean-squared bias. The left-hand side of Eq. (13) is depicted in Fig. 1. The first integral on the right-hand side is depicted in Fig. 2. The integrand of the last integral on the right-hand side [excluding

$P(\rho)$] appears in Fig. 3 as the square of the difference between the curve for the estimator and the true value being estimated. This quantity favors the frequency-counts estimator for some values of ρ for sufficiently large N; however, the first integral on the right-hand side more than compensates to give a result favoring the Bayes estimator.

Figure 5 depicts the sample average of the estimator's deviations from true as a function of ρ for various sample sizes N,

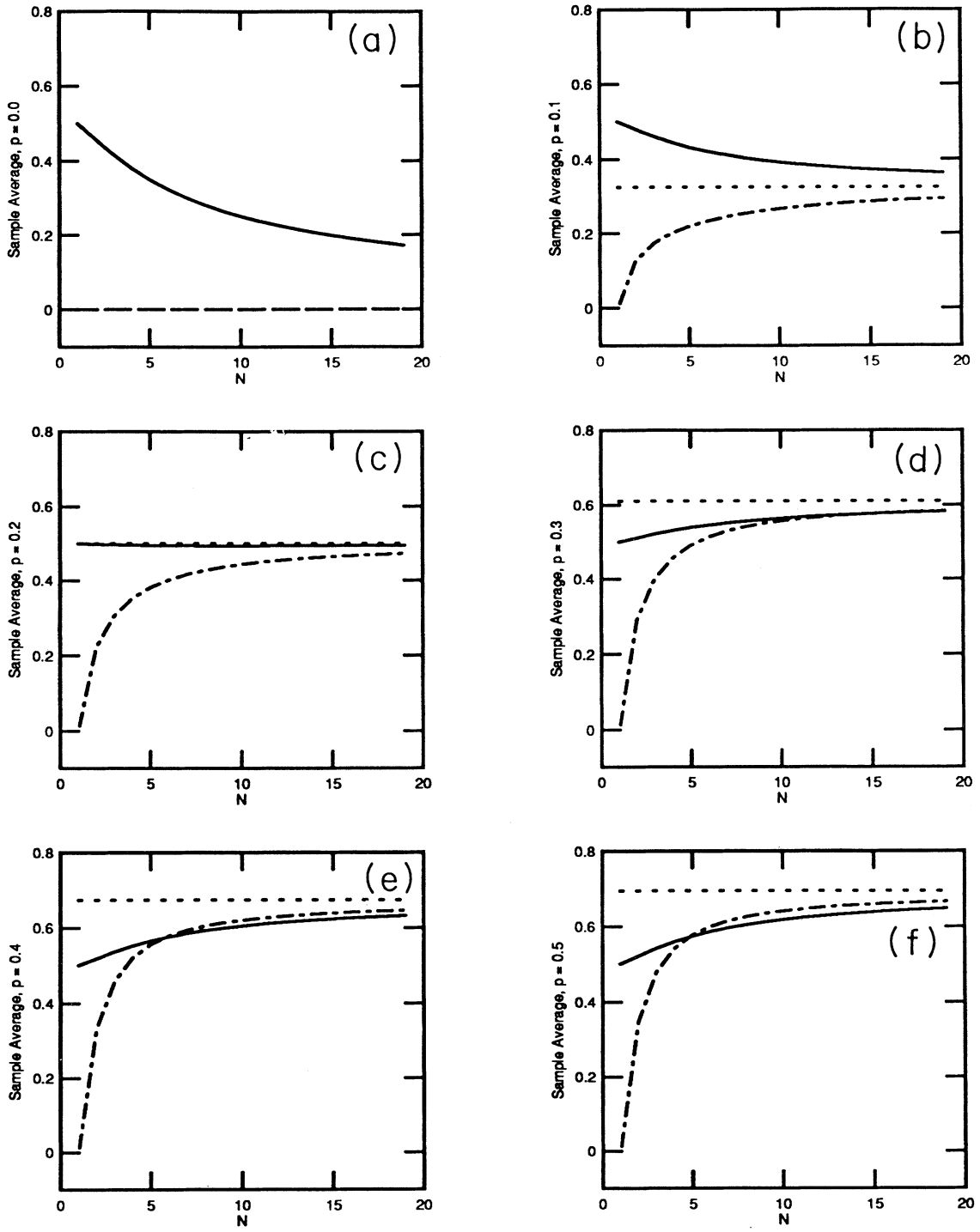

FIG. 3. Sample average [Eq. (12)] for the Bayes (solid line) and frequency-counts (dash-dotted line) estimators of the two-bin ($m=2$) entropy $S(\rho)$. Both are graphed as functions of N for various values of $\rho=(p,1-p)$. The true value of $S(\rho)$ is also graphed (dashed line).

$$\sum_{\mathbf{n}} P(\mathbf{n}|\rho)[S(\mathbf{n})-S(\rho)]^2 . \quad (14)$$

The integral of the expression in (14) multiplied by the density $P(\rho)$ (here uniform), depicted for various N, is shown in Fig. 1.

Finally, Fig. 6 shows the convergence of the PDF $P(s|\mathbf{n})$ given by

$$P(S(\rho)=s|\mathbf{n})=\int d\rho\, \delta(S(\rho)-s)P(\rho|\mathbf{n}) \quad (15)$$

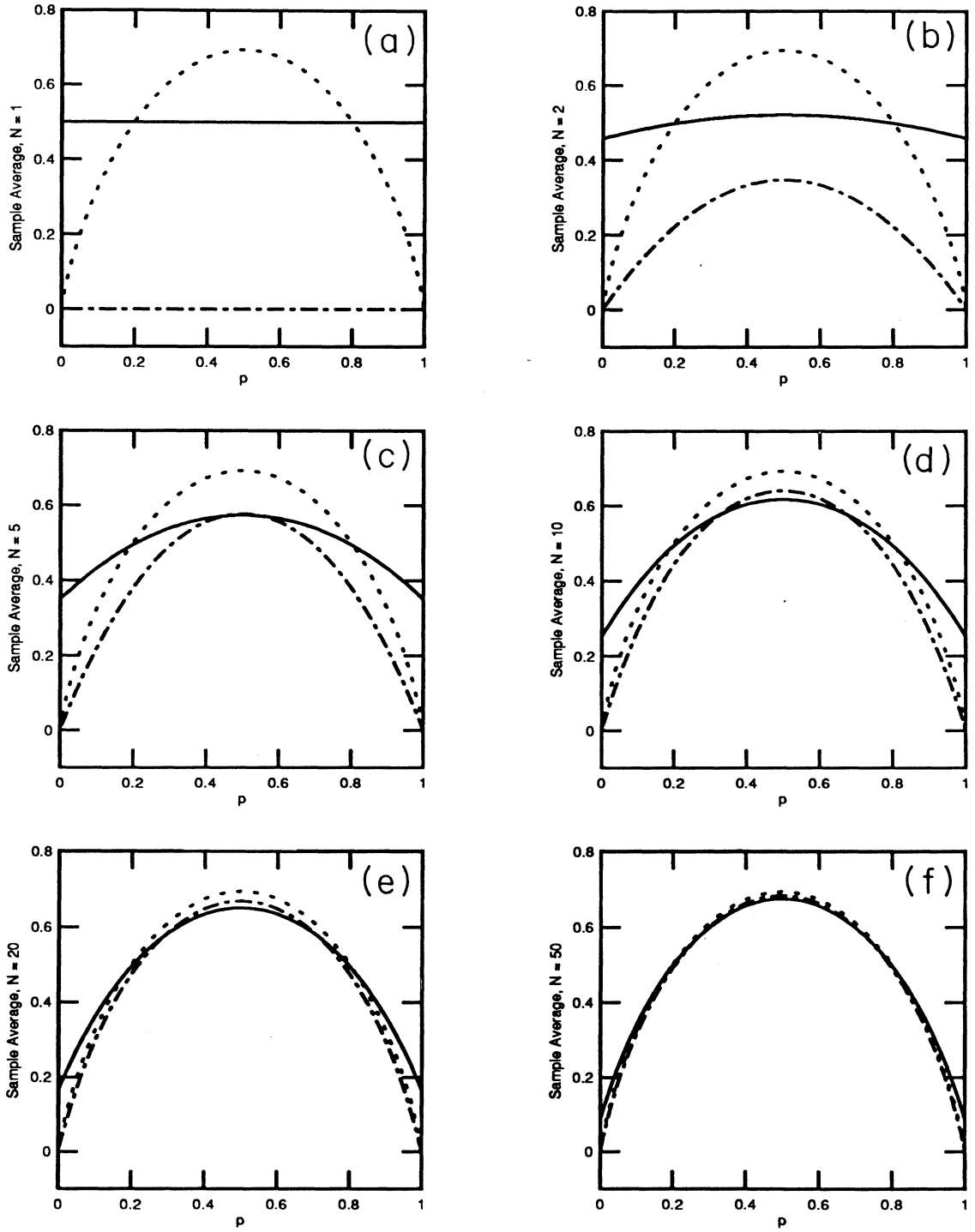

FIG. 4. Sample average [Eq. (12)] for the Bayes (solid line) and frequency-counts (dashed-dotted line) estimators of the two-bin ($m=2$) entropy $S(\rho)$. Both are graphed as functions of $\rho=(p,1-p)$ for various values of N. The true value of $S(\rho)$ is also graphed (dashed line).

for a fixed ratio (1:15) of observed counts $n_1:n_2$, as the overall number of counts $N=n_1+n_2$ increases. Note the increasing density placed upon the true entropy as the counts N increase. Note that the average of s according to this density $P(s|\mathbf{n})$, i.e., $\int ds\, s\, P(s|\mathbf{n})$, is the Bayes estimator for $S(\rho)$ given the observations $\mathbf{n}$. As mentioned previously, of all estimators, its squared error averaged over both ρ and $\mathbf{n}$ is minimal.

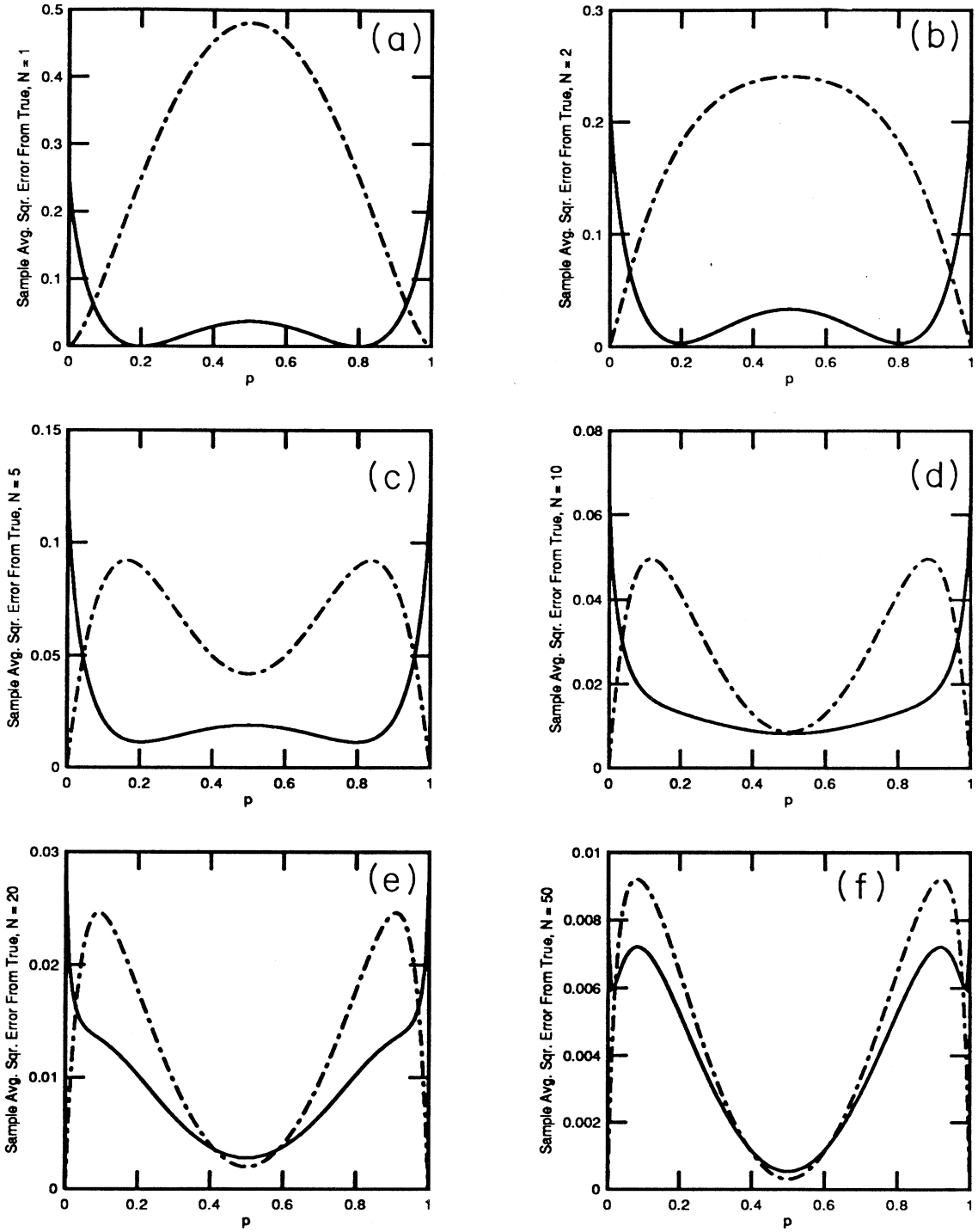

FIG. 5. Sample average deviation from true [Eq. (14)] for the Bayes (solid line) and frequency-counts (dash-dotted line) estimators of the two-bin ($m=2$) entropy. Both are graphed as functions of $\rho=(p,1-p)$ for various values of N. The integral over ρ with the density $P(\rho)$ appears as Fig. 1.

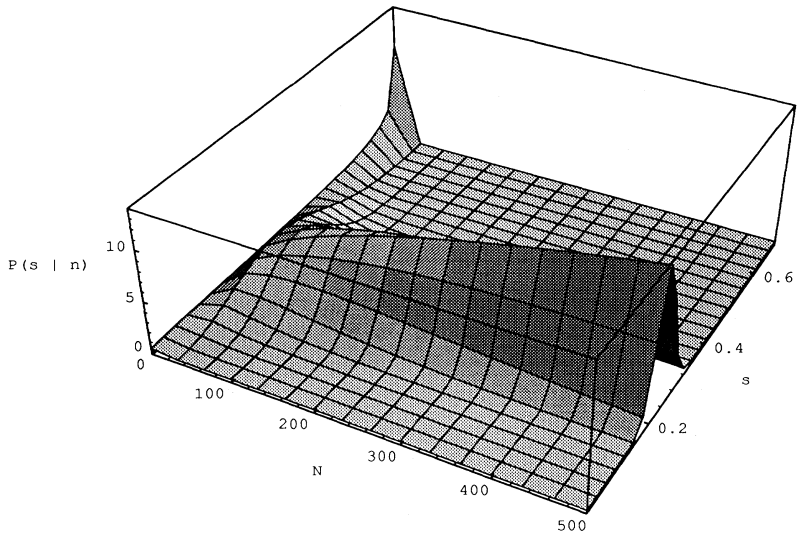

FIG. 6. Posterior PDF [Eq. (15)] of the entropy $S(\mathbf{p})$ for $m=2$ and fixed counts ratio $n_1{:}n_2=1{:}15$, but differing overall $N=n_1+n_2$. As N increases, the density converges to a δ function at the values $s=S(\frac{1}{16},\frac{15}{16})=0.2338$ of the entropy.

VI. RESULTS FOR OTHER $Q(\,)$

In Ref. [1] we performed calculations similar to those given above to derive the Bayes estimators with a uniform prior for some other quantities besides the Shannon entropy. This section defines those quantities and presents the results; the reader is referred to Ref. [1] for the rather involved derivation of these results. In the interest of brevity, not all the results are presented here; again, the interested reader is referred to Ref. [1].

In presenting these results sometimes $\boldsymbol{\rho}$ and $\mathbf{n}$ will be matrices. To avoid confusion with the previous discussion, sometimes the matrix $\boldsymbol{\rho}$ will be indicated by the symbol $\mathbf{p}$. It will also help clarify the discussion to define $\nu_{ij}\equiv n_{ij}+1$. Similarly, to denote row or column sums of n_{ij}'s we use $\nu_{i\cdot}\equiv\sum_j\nu_{ij}$ and $\nu_{\cdot j}\equiv\sum_i\nu_{ij}$ and define $\nu\equiv\sum_{ij}\nu_{ij}$.

The mutual information is defined in terms of a matrix $\mathbf{p}$ by

$$M(\mathbf{p})=S((p_{i\cdot}))+S((p_{\cdot j}))-S(\mathbf{p})\ . \tag{16}$$

Here $(p_{i\cdot})$ and $(p_{\cdot j})$ are the vectors of column and row sums of $\mathbf{p}=(p_{ij})$, respectively, i.e., $p_{i\cdot}\equiv\sum_j p_{ij}$ and similarly for $(p_{\cdot j})$. $S(\mathbf{p})$ is the usual Shannon entropy $S(\mathbf{p})=-\sum_{ij}p_{ij}\ln(p_{ij})$, while $S((p_{i\cdot}))=-\sum_i p_{i\cdot}\ln(p_{i\cdot})$ and similarly for $S((p_{\cdot j}))$. Mutual information is a measure of the amount of information shared between two symbol streams (symbolic dynamical systems) with joint probability p_{ij} [4]. It may also be seen as a measure of the correlation between two symbolic systems with joint probability p_{ij} [3]. The mutual information function has applications in areas such as communication theory [4] (e.g., the measurement of channel capacity), pattern recognition [5], and natural language analysis [6], to name but a few.

The χ^2 statistic for independence is also given by a function of a matrix $\mathbf{p}$,

$$\chi^2(\mathbf{p})\equiv\sum_{ij}[(p_{ij}-p_{i\cdot}p_{\cdot j})^2/(p_{i\cdot}p_{\cdot j})]\ . \tag{17}$$

χ^2 is commonly used in statistical tests of independence [35], where it appears in a form with the maximum-likelihood estimator of $\mathbf{p}$ substituted for $\mathbf{p}$ in Eq. (17). The form in Eq. (17) is proportional to the asymptotic (large data set) statistic used in these tests and it is easily shown to be a first-order approximation to the mutual information under certain conditions [2,36].

The covariance function of a matrix p is given by

$$\mathrm{cov}_{xy}(\mathbf{p})\equiv\sum_{ij}p_{ij}(X_i-\mu_x)(Y_j-\mu_y)\ , \tag{18}$$

where each of the m possible states is associated with some ordered pair (X_i,Y_j) of numbers [there are m index pairs (i,j) altogether]. The ijth state occurs with probability p_{ij}. The means are μ_x and μ_y; $\mu_x\equiv\sum_i p_{i\cdot}X_i$ and similarly for μ_y [35].

The variance function of a vector $\boldsymbol{\rho}$ is given by

$$\mathrm{var}(\boldsymbol{\rho})\equiv\sum_i\rho_i(X_i-\mu_x)^2\ , \tag{19}$$

where the ith state is associated with the number X_i and occurs with probability p_i. Finally, the average is a function of a vector $\mathbf{p}$ given by

$$\mathrm{avg}_x(\boldsymbol{\rho})\equiv\sum_i\rho_i X_i\ . \tag{20}$$

In calculating moments of these quantities, extensive use is made of the following result, which in Ref. [12] is established by using Mellin transforms and Theorem 3. This result allows us to calculate moments of real-valued powers of sums of the ρ_i.

Let σ_u mean a subset of the integers between 1 and m i.e., a delineation of certain ρ_i's. We define $\beta_u\equiv\sum_{i\in\sigma_u}(n_i+1)$, $\nu_i\equiv\mathbf{n}_i+1/\gamma_{\mathbf{n}}\equiv\prod_{i=1}^m\Gamma(\nu_i)/\rho$, and $\rho_u\equiv\sum_{i\in\sigma_u}\rho_i$.

Theorem 9. If the subsets σ_u, defined for $u=1,\ldots,k$, satisfy $\sigma_u\cap\sigma_v=\varnothing$ for all $u\neq v$, if $\mathrm{Re}(\beta_u+\eta_u)>0$ for all $u=1,\ldots,k$, and if $\mathrm{Re}(\nu_i)>0$ for all $i=1,\ldots,m$, then

$$I[\rho_1^{\eta_1}\cdots\rho_k^{\eta_k},\mathbf{n}]=\frac{\gamma_{\mathbf{n}}}{\Gamma(\beta+\eta)}\prod_{u=1}^k\frac{\Gamma(\beta_u+\eta_u)}{\Gamma(\beta_u)}\ ,$$

where $\beta \equiv \sum_i \nu_i$ and $\eta \equiv \sum_u \eta_u$.

(i) *Mutual information* $M(\mathbf{p}) = \sum_{ij} p_{ij} \ln(p_{ij}/p_{i\cdot}p_{\cdot j})$. In this case the observed counts form a matrix. Define $\bar{\nu}_{ij} \equiv \nu_{i\cdot} + \nu_{\cdot j} - \nu_{ij}$.

Theorem 10. If the ν_{ij} are non-negative integers for all ij, then (a) $E(M(\mathbf{p})|\mathbf{n}) = \mathcal{E}_{\overline{IJ}} - \mathcal{E}_{\overline{I}} - \mathcal{E}_{\overline{J}}$ where

$$\mathcal{E}_{\overline{IJ}} = E\left[\sum_{ij} p_{ij} \ln(p_{ij}) | \mathbf{n}\right] = \sum_{ij} \frac{\nu_{ij}}{\nu} \Delta\Phi^{(1)}(\nu_{ij}+1, \nu+1) ,$$

$$\mathcal{E}_{\overline{I}} = E\left[\sum_i p_{i\cdot} \ln(p_{i\cdot}) | \mathbf{n}\right] = \sum_i \frac{\nu_{i\cdot}}{\nu} \Delta\Phi^{(1)}(\nu_{i\cdot}+1, \nu+1) ,$$

$$\mathcal{E}_{\overline{J}} = E(p_{\cdot j} \ln(p_{\cdot j}) | \mathbf{n}) = \sum_j \frac{\nu_{\cdot j}}{\nu} \Delta\Phi^{(1)}(\nu_{\cdot j}+1, \nu+1) ,$$

and (b)

$$E(M^2(\mathbf{p})|\mathbf{n}) = \mathcal{E}_{\overline{IJMN}} + \mathcal{E}_{\overline{IM}} + \mathcal{E}_{\overline{JN}} - 2(\mathcal{E}_{\overline{IJM}} + \mathcal{E}_{\overline{IJN}} - \mathcal{E}_{\overline{IN}})$$

where

$$\mathcal{E}_{\overline{IJMN}} = E\left[\sum_{i,j,m,n} p_{ij} \ln(p_{ij}) p_{mn} \ln(p_{mn}) | \mathbf{n}\right]$$

$$= \sum_{i,j} \sum_{m,n \neq i,j} \frac{\nu_{ij}\nu_{mn}}{\nu(\nu+1)} \{\Delta\Phi^{(1)}(\nu_{ij}+1, \nu+2)\Delta\Phi^{(1)}(\nu_{mn}+1, \nu+2) - \Phi^{(2)}(\nu+2)\}$$

$$+ \sum_{ij} \frac{\nu_{ij}(\nu_{ij}+1)}{\nu(\nu+1)} \{[\Delta\Phi^{(1)}(\nu_{ij}+2, \nu+2)]^2 + \Delta\Phi^{(2)}(\nu_{ij}+2, \nu+2)\} ,$$

$$\mathcal{E}_{\overline{IM}} = E\left[\sum_{i,m} p_{i\cdot} \ln(p_{i\cdot}) p_{m\cdot} \ln(p_{m\cdot}|) \mathbf{n}\right]$$

$$= \sum_i \sum_{m \neq i} \frac{\nu_{i\cdot}\nu_{m\cdot}}{\nu(\nu+1)} \{\Delta\Phi^{(1)}(\nu_{i\cdot}+1, \nu+2)\Delta\Phi^{(1)}(\nu_{m\cdot}+1, \nu+2) - \Phi^{(2)}(\nu+2)\}$$

$$+ \sum_i \frac{\nu_{i\cdot}(\nu_{i\cdot}+1)}{\nu(\nu+1)} \{[\Delta\Phi^{(1)}(\nu_{i\cdot}+2, \nu+2)]^2 + \Delta\Phi^{(2)}(\nu_{i\cdot}+2, \nu+2)\}$$

(to find $\mathcal{E}_{\overline{JN}}$ substitute $\nu_{\cdot i}$ for $\nu_{i\cdot}$ and $\nu_{\cdot m}$ for $\nu_{m\cdot}$ in the expression for $\mathcal{E}_{\overline{IM}}$),

$$\mathcal{E}_{\overline{IJM}} = E\left[\sum_{ij} \sum_m p_{ij} \ln(p_{ij}) p_{m\cdot} \ln(p_{m\cdot}) | \mathbf{n}\right]$$

$$= \sum_{ij} \sum_{m \neq i} \frac{\nu_{ij}\nu_{m\cdot}}{\nu(\nu+1)} \{\Delta\Phi^{(1)}(\nu_{ij}+1, \nu+2)\Delta\Phi^{(1)}(\nu_{m\cdot}+1, \nu+2) - \Phi^{(2)}(\nu+2)\}$$

$$+ \sum_{i,j} \frac{\nu_{ij}(\nu_{i\cdot}+1)}{\nu(\nu+1)} \{[\Delta\Phi^{(1)}(\nu_{i\cdot}+2, \nu+2)]^2$$

$$+ \Delta\Phi^{(1)}(\nu_{ij}+1, \nu_{i\cdot}+1)\Delta\Phi^{(1)}(\nu_{i\cdot}+2, \nu+2) + \Delta\Phi^{(2)}(\nu_{i\cdot}+2, \nu+2)\}$$

(to find $\mathcal{E}_{\overline{IJN}}$ substitute $\nu_{\cdot m}$ for $\nu_{m\cdot}$ in the expression for $\mathcal{E}_{\overline{IJM}}$),

$$\mathcal{E}_{\overline{IN}} = E\left[\sum_i \sum_n p_{i\cdot} \ln(p_{i\cdot}) p_{\cdot n} \ln(p_{\cdot n}) | \mathbf{n}\right]$$

$$= \sum_{i,n} \frac{\bar{\nu}_{in}(\bar{\nu}_{in}+1)}{\nu(\nu+1)} \left\{ \{[\Delta\Phi^{(1)}(\bar{\nu}_{in}+2, \nu+2)]^2 + \Delta\Phi^{(2)}(\bar{\nu}_{in}+2, \nu+2)\} \right.$$

$$\times \left[1 - \frac{\nu_{i\cdot} + \nu_{\cdot n} - 2\nu_{\mathrm{in}}}{\nu} + \frac{(\nu_{i\cdot} - \nu_{in})(\nu_{\cdot n} - \nu_{in})}{\nu(\nu+1)}\right]$$

$$+ \Delta\Phi^{(1)}(\bar{\nu}_{in}+2, \nu+2) \sum_{r=0}^{\infty} \frac{Q_1(r,1)}{r!} \left[\frac{(\nu_{\cdot n} - \nu_{in})_r}{(\bar{\nu}_{in})_r} \left[1 + \frac{\nu_{i\cdot} - \nu_{in}}{\bar{\nu}_{in} + r}\right] + \frac{(\nu_{i\cdot} - \nu_{in})_r}{(\bar{\nu}_{in})_r} \left[1 + \frac{\nu_{\cdot n} - \nu_{in}}{\bar{\nu}_{in} + r}\right]\right]$$

$$\left. + \sum_{r=0}^{\infty} \sum_{s=0}^{\infty} \frac{(\nu_{i\cdot} - \nu_{in})_r (\nu_{\cdot n} - v_{in})_s}{(\bar{\nu}_{in})_{r+s}} \frac{Q_1(r,1)}{r!} \frac{Q_1(s,1)}{s!} \right\} ,$$

where Q_1 is given by

$$Q_1(j,\eta_1)\equiv[1-\theta(j-\eta_1-1)]\frac{(-1)^j}{(\eta_1-j)!}\sum_{r=0}^{j-1}\frac{1}{\eta_1-r}$$
$$+\theta(j-\eta_1-1)(-1)^{\eta_1+1}\Gamma(j-\eta_1)$$

and the notation a_b means $\Gamma(a+b)/\Gamma(a)$.

(ii) *Average* $A(\boldsymbol{\rho})=\sum_{i=1}^m \rho_i X_i$. [Note that for $X_i=\delta_{ij}$, part (a) of Theorem 11 gives the Laplace law of succession estimator for ρ_j.]

Theorem 11. If $\mathrm{Re}(\nu_i)>0\ \forall i$ then (a)

$$E(A(\boldsymbol{\rho})|\mathbf{n})=\sum_i \frac{\nu_i}{\nu}X_i$$

and (b)

$$E(A^2(\boldsymbol{\rho})|\mathbf{n})=\sum_{i\neq j}\frac{\nu_i\nu_j}{\nu(\nu+1)}X_iX_j+\sum_i\frac{\nu_i(\nu_i+1)}{\nu(\nu+1)}X_i^2\ .$$

(iii) *Variance* $V(\boldsymbol{\rho})=\sum_{i=1}^m \rho_i(X_i-\mu_x)^2$. Note that $E(V(\boldsymbol{\rho})|\mathbf{n})\neq E([A(\boldsymbol{\rho})-E(A(\mathbf{p})|\mathbf{n})]^2|\mathbf{n})$; μ_x is the true mean, not the expected mean, and $V(\boldsymbol{\rho})$ refers to the true variance, not the variance in the estimator $E(A(\boldsymbol{\rho})|\mathbf{n})$.

Theorem 12. If $\mathrm{Re}(\nu_i)>0\ \forall i$ then (a)

$$E(V(\mathbf{p})|\mathbf{n})=\sum_i\frac{\nu_i(\nu-\nu_i)}{\nu(\nu+1)}X_i^2-\sum_{i\neq j}\frac{\nu_i\nu_j}{\nu(\nu+1)}X_iX_j$$

and (b)

$$E(V^2(\mathbf{p})|\mathbf{n})=\sum_{i,j}E(\rho_i\rho_j|\mathbf{n})X_i^2X_j^2)$$
$$-2\sum_{i,j,k}E(\rho_i\rho_j\rho_k|\mathbf{n})X_i^2X_jX_k)$$
$$+\sum_{i,j,k,l}E(\rho_i\rho_j\rho_k\rho_l|\mathbf{n})X_iX_jX_kX_l)\ ,$$

where the expectations can be found by applying Theorem 9 or simply by augmenting the $\mathbf{n}$'s appropriately (see the proof of Theorem 7).

(iv) Covariance $C(\mathbf{p})=\sum_{ij}p_{ij}(X_i-\mu_x)(Y_i-\mu_y)$.

Theorem 13. If $\mathrm{Re}(\nu_{ij})>0\ \forall ij$, then

$$E(C(\mathbf{p})|\mathbf{n})=\sum_{i,j}X_iY_j\frac{1}{\nu(\nu+1)}[\nu\nu_{ij}-\nu_{i\cdot}\nu_{\cdot j}]\ ,$$

as can be verified by applying Theorem 14a of Ref. [12].

The second posterior moment of $C(\mathbf{p})$ is given in Theorem 26 of Ref. [12].

(v) $\chi^2(\mathbf{p})=\sum_{ij}(p_{ij}-p_{i\cdot}p_{\cdot j})^2/p_{i\cdot}p_{\cdot j}$.

Theorem 14. If $\mathrm{Re}(\nu_{ij})>0\ \forall ij$, $\mathrm{Re}(\nu_{i\cdot})>-1$, and $\mathrm{Re}(\nu_{\cdot j})>-1$, then

$$E(\chi^2(\mathbf{p})|\mathbf{n})$$
$$=-1+\sum_{i,j}\frac{(\nu-1)(\nu-2)}{(\nu_{i\cdot}+\nu_{\cdot j}-\nu_{ij}+1)(\nu_{i\cdot}+\nu_{\cdot j}-\nu_{ij})}$$
$$\times\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}\frac{(\nu_{i\cdot}-\nu_{ij})_m(\nu_{\cdot j}-\nu_{ij})_n}{(\nu_{i\cdot}+\nu_{\cdot j}-\nu_{ij}+2)_{m+n}}\ ,$$

as can be verified by applying Theorem 14a of Ref. [12].

The second posterior moment of $\chi^2(\mathbf{p})$ is given in Theorem 27 of Ref. [12].

VII. CONCLUSION

Many situations can be characterized as follows. There is some unknown probability distribution $\boldsymbol{\rho}$ across a set of m possible events. That distribution is IID sampled N times, to create a data set $\mathbf{n}$. From $\mathbf{n}$ we want to infer not $\boldsymbol{\rho}$ itself, but some functional of $\boldsymbol{\rho}$. The example considered in this paper is inferring the Shannon entropy of $\boldsymbol{\rho}$ from $\mathbf{n}$.

There are many different ways to go about performing such inference. The Bayesian approach is to directly estimate what we want, which is $P(Q(\rho)=q|\mathbf{n})$ as a function of q. This approach requires that we know (or assume) the distribution $P(\boldsymbol{\rho})$. This distribution should reflect one's prior knowledge concerning $\boldsymbol{\rho}$. In this paper we consider the case where $P(\boldsymbol{\rho})$ is uniform over all allowed $\boldsymbol{\rho}$. In Ref. [12] we consider the extension to arbitrary $P(\boldsymbol{\rho})$.

Rather than try to calculate $P(Q(\boldsymbol{\rho})=q|\mathbf{n})$ directly, in this paper we instead calculate its moments. The first moment gives the optimal guess (in the sense defined in Sec. III) if one wishes to minimize mean squared error. The second moment can then be used in conjunction with Chebyshev's inequality to bound the probability of deviation of $Q(\boldsymbol{\rho})$ from this optimal value.

These first two moments for the case where $Q(\ )$ is the Shannon entropy are presented in Sec. IV E. Numerically investigations of them are presented in Sec. V. In Ref. [12] we find the first two moments when $Q(\boldsymbol{\rho})$ is the mutual information of $\boldsymbol{\rho}$, the variance of $\boldsymbol{\rho}$, the covariance of $\boldsymbol{\rho}$, the χ^2 of $\boldsymbol{\rho}$, and the average of $\boldsymbol{\rho}$. These results are presented in Sec. VI.

ACKNOWLEDGMENTS

We would like to thank James Theiler for supplying the computer code XYPLOT used to produce the figures. We would also like to thank the Center for Nonlinear Studies and Theoretical Division, both of LANL, for their partial support during part of this work. This work was supported in part by the United States Department of energy under Contract No. W-7405-ENG-36. For part of this work D.H.W. was supported by the Santa Fe Institute and by National Library of Medicine Grant No. NLM F37 LM00011.

[1] P. Grassberger, Phys. Lett. A **128**, 369 (1988). A related calculation is carried out by H. Herzel, Syst. Anal. Model. Simul. **5**, 435 (1988).

[2] S. Eubank and D. Farmer, in *1989 Lectures in Complex Systems,* SFI Studies in the Sciences of Complexity Vol. II, edited by E. Jen (Addison-Wesley, Reading, MA, 1990).

[3] W. Li, J. Stat. Phys. **60**, 823 (1990). Some researchers have raised questions about the convergence accompanying the use of approximation (5.5) of this paper in its Eq. (5.4). It is also important to note that the assumption that $c_{\alpha\beta}/(c_\alpha c_\beta)$ is constant, made just below Eq. (5.8), is equivalent to assuming that the true $c_{\alpha\beta}$ has minimal mutual information. This observation explains Li's result that any estimated $c_{\alpha\beta}$ has higher mutual information than the true $c_{\alpha\beta}$.

[4] R. W. Hamming, *Coding and Information Theory,* 2nd ed. (Prentice-Hall, Englewood, Cliffs, NJ, 1986).

[5] S. Watanabe, *Pattern Recognition, Human and Mechanical* (Wiley, New York, 1985). Note that Watanabe's multidimensional generalization of the mutual information function, appearing in Chap. 6.5, may also be estimated using the techniques of this paper.

[6] W. Li, Santa Fe Institute Report No. TR 89-008, 1993 (unpublished).

[7] S. Lloyd, Phys. Rev. A **39**, 5378 (1989).

[8] B. T. M. Korber, R. M. Farber, D. H. Wolpert, and A. S. Lapedes, Proc. Natl. Acad. Sci. U.S.A. **90**, 7176 (1993).

[9] C. E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948).

[10] E. T. Jaynes, Phys. Rev. **106**, 620 (1957).

[11] Y Tikochinsky, N. Z. Tishby, and R. D. Levine, Phys. Rev. Lett. **52**, 1357 (1984); references therein for a preliminary exposition of the many uses of entropy in physics and engineering.

[12] D. R. Wolf and D. H. Wolpert, Los Alamos National Laboratory Report No. LA-UR-93-833, 1993 (unpublished). Send email to comp-gas@xyz.lanl.gov with subject "get 9403002" to get an encoded postscript version. "9403001" might also be helpful to the reader.

[13] B. Harris, Colloq. Math. Soc. Janos Bolyai **16**, 323 (1975).

[14] T. Leonard and J. Hsu, Ann. Stat. **20**, 1669 (1992), and references therein.

[15] G. P. Basharin, Theory Probab. Its Appl. (USSR) **4**, (3) 333 (1959).

[16] G. A. Miller, in *Information Theory in Psychology,* edited by H. Quastler (Free Press, Glencoe, IL, 1955), pp. 95–100.

[17] R. M. Fagen, J. Theor. Biol. **73**, 61 (1978).

[18] H. Herzel, A. O. Schnmitt, and W. Ebeling, Humboldt University Technical Report; 1993 (unpublished).

[19] A. O. Schmitt, H. Herzel, and W. Ebeling, Europhys. Lett. **23**, 303 (1993).

[20] V. S. Pugachev, *Probability Theory and Mathematical Statistics for Engineers* (Pergamon, New York, 1984).

[21] Formally, $\mathbf{n}$ and $\boldsymbol{\rho}$ are instances of some random variables N and T, respectively. We also have a random variable $Q=Q(T)$, with instances q. Let f indicate a probability density function. Then by $P(\boldsymbol{\rho}|\mathbf{n})$ we formally mean $f_{T|N}(\boldsymbol{\rho}|\mathbf{n})$. On the other hand, in expressions such as $P(\mathbf{n})$, the notation $P(\,)$ denotes a probability, not a probability density function. [The context will always make the meaning of $P(\,)$ clear.] In addition, below we will write expressions such as $P(Q(\boldsymbol{\rho})=q|\mathbf{n})$. Formally this means $f_{Q|N}(q|\mathbf{n})$. So, for example, the equation $\int dq\, q^k P(Q(\boldsymbol{\rho})=q|\mathbf{n})=\int d\boldsymbol{\rho}\, Q^k(\boldsymbol{\rho})P(\boldsymbol{\rho}|\mathbf{n})$ formally means $\int dq\, q^k f_{Q|N}(q|\mathbf{n})=\int d\boldsymbol{\rho}\, Q^k(\boldsymbol{\rho})f_{T|N}(\boldsymbol{\rho}|\mathbf{n})$.

[22] J. Skilling, in *Maximum Entropy and Bayesian Methods,* edited by J. Skilling (Kluwer Academic, Dordrecht, 1989), pp. 45–52.

[23] If one wishes to estimate $\boldsymbol{\rho}$ by finding the mode of $P(\boldsymbol{\rho}|\mathbf{n})$, then the entropic prior leads immediately to the technique of maximum entropy in the case where the data are not finite vectors of counts $\mathbf{n}$ but rather is some expectation value $b=\sum_i \rho_i B(\rho_i)$. This follows since $P(\sum_i \rho_i B(\rho_i)=b|\boldsymbol{\rho})$, considered as a function of ρ with b fixed, is everywhere either 1 or 0. As a result, by Bayes's theorem, for a prior of the form $e^{\alpha S(\rho)}$, finding the mode of $P(\boldsymbol{\rho}|\sum_i \rho_i B(\rho_i)=b)$ is equivalent to maximizing $S(\rho)$ subject to $\sum_i \rho_i B(\rho_i)=b$.

[24] E. T. Jaynes, University of Cambridge Report. No. 1189, 1984 (unpublished).

[25] D. Rubin, Ann. Stat. **9**, 130 (1981).

[26] H. Goldstein, *Classical Mechanics* (Addison-Wesley, Reading, MA, 1980).

[27] What we have shown here is that q_1/q_0 has the least mean-squared error from the true $Q(\boldsymbol{\rho})$, on average. This does not imply that it is the estimator of the entropy that is least *biased* on average. To find the "least average bias estimator," one instead searches for the $G(\,)$ minimizing $\int d\boldsymbol{\rho}\, P(\boldsymbol{\rho})[\sum_{\mathbf{n}}\{P(\mathbf{n}|\boldsymbol{\rho})G(\mathbf{n})-Q(\boldsymbol{\rho})\}]^2$. See Ref. [12] for a cursory exposition on finding this $G(\,)$.

[28] This issue of finding $P(\mathbf{n}|Q(\rho)=q)$ also arises in other classical sampling distribution problems such as hypothesis testing. (See, for example, Ref. [3].) This problem may sometimes be addressed using the techniques developed in this paper for calculating moments of the posterior $P(Q(\boldsymbol{\rho})=q|\mathbf{n})$. To do this, one estimates (rather than calculates) the distribution $P(Q(\boldsymbol{\rho})=q|\mathbf{n})$. For example, one could do this by evaluating its first few moments and then using the maximum entropy technique. After this one evaluates $P(\mathbf{n})=\int d\boldsymbol{\rho}\, P(\mathbf{n}|\boldsymbol{\rho})P(\boldsymbol{\rho})$ and then applies Bayes's theorem to deduce $P(\mathbf{n}|Q(\boldsymbol{\rho})=q)$ (up to overall normalization constants). Note that the answer will depend on the prior $P(\boldsymbol{\rho})$ in general.

[29] J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer-Verlag, New York, 1985).

[30] T. Loredo, in Maximum Entropy and Bayesian Methods, edited by P. Fougere (Kluwer, Dordrecht, 1990), pp. 81–142.

[31] D. Wolpert, in *Proceedings of the 1993 Maximum Entropy and Bayesian Methods Workshop,* edited by G. Heidbreder (Kluwer, Dordrecht, in press).

[32] D. L. Holl, C. G. Maple, and B. Vinograde, *Introduction to the Laplace Transform* (Appleton-Century-Crofts, New York, 1959), Chap. 2, p. 43.

[33] S. S. Wilks, *Mathematical Statistics* (Wiley, New York, 1962).

[34] *Handbook of Mathematical Functions,* edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1972), Eqs. (6.4.1) and (6.1.41).

[35] M. H. De Groot, *Probability and Statistics,* 2nd ed. (Addison Wesley, Reading, MA, 1986).

[36] E. T. Jaynes, *Papers on Probability, Statistics, and Statistical Physics* (Kluwer, Dordrecht, 1983).