Bayesian Estimation of the Entropy of the Multivariate Gaussian

Santosh Srivastava
Fred Hutchinson Cancer Research Center
Seattle, WA 98109, USA
Email: ssrivast@fhcrc.org

Maya R. Gupta
Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA
Email: gupta@ee.washington.edu

Abstract— **Estimating the entropy of a Gaussian distribution from samples drawn from the distribution is a difficult problem when the number of samples is smaller than the number of dimensions. A new Bayesian entropy estimator is proposed using an inverted Wishart distribution and a data-dependent prior that handles the small-sample case. Experiments for six different cases show that the proposed estimator provides good performance for the small-sample case compared to the standard nearest-neighbor entropy estimator. Additionally, it is shown that the Bayesian estimate formed by taking the expected entropy minimizes expected Bregman divergence.**

I. Introduction

Entropy is a useful description of the predictability of a distribution, and has use in many applications of coding, machine learning, signal processing, communications, and chemistry [1]–[5]. In practice, many continuous generating distributions are modeled as Gaussian distributions. One reason for this is the central limit theorem, another reason is because the Gaussian is the maximum entropy distribution given an empirical mean and covariance, and as such is a least assumptive model. For the multivariate Gaussian distribution, the entropy goes as the log determinant of the covariance; specifically, the differential entropy of a d-dimensional random vector X drawn from the Gaussian $\mathcal{N}(\mu, \Sigma)$ is

$$h(X) = \int \mathcal{N}(x) \ln \mathcal{N}(x) dx = \frac{d}{2} + \frac{d \ln(2\pi)}{2} + \frac{\ln|\Sigma|}{2}. \quad (1)$$

In this paper we consider the practical problem of estimating the entropy of a generating normal distribution $\mathcal{N}$ given samples $x_1, x_2, \ldots, x_n$ that are assumed to be realizations of random vectors $X_1, X_2, \ldots, X_n$ drawn iid from $\mathcal{N}(\mu, \Sigma)$. In particular, we focus on the difficult limited-data case that $n \leq d$, which occurs often in practice with high-dimensional data.

One approach to estimating $h(X)$ is to first estimate the Gaussian distribution, for example with maximum likelihood (ML) estimates of μ and Σ, and then plug-in the covariance estimate to (1) [6]. Such estimates are usually infeasible when $n \leq d$. The ML estimate is also negatively biased so that it underestimates the entropy on average [5]. We believe it is more effective and will create fewer numerical problems if one directly estimates a single value for the entropy, rather than first estimating the entire covariance matrix in order to produce the scalar entropy estimate. To this end, we propose a data-dependent Bayesian estimate using an inverted Wishart prior. The choice of prior in Bayesian estimation can dramatically change the results, but there is little practical guidance for choosing priors. The main contributions of this work are showing that the inverted Wishart prior enables estimates for $n \leq d$, and that using a rough data-dependent estimate as a parameter to the inverted Wishart prior yields a more robust estimator than ignoring the data when creating the prior.

II. Related Work

First we review related work in parametric entropy estimation, then in nonparametric estimation.

Ahmed and Gokhale investigated uniformly minimum variance unbiased (UMVU) entropy estimators for parametric distributions [6]. They showed that the UMVU entropy estimate for the Gaussian is,

$$\hat{h}_{\text{UMVU}} = \frac{d \ln \pi}{2} + \frac{\ln|S|}{2} - \frac{1}{2} \sum_{i=1}^{d} \psi\left(\frac{n+1-i}{2}\right), \quad (2)$$

where $S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$, and the digamma function is defined $\psi(z) = \frac{d}{dz} \ln \Gamma(z)$, where Γ is the standard gamma function.

Bayesian entropy estimation for the multivariate normal was first proposed in 2005 by Misra et al. [5]. They form an entropy estimate by substituting $\widehat{\ln|\Sigma|}$ for $\ln|\Sigma|$ in (1) where $\widehat{\ln|\Sigma|}$ minimizes the squared-error risk of the entropy estimate. That is, $\widehat{\ln|\Sigma|}$ solves,

$$\arg\min_{\delta \in \mathcal{R}} E_{\mu,\Sigma}\left[(\delta - \ln|\Sigma|)^2\right], \quad (3)$$

where here Σ denotes a random covariance matrix and μ a random vector, and the expectation is taken with respect to the posterior. We will denote by $\tilde{\mu}$ and $\tilde{\Sigma}$ respectively realizations of the random μ and Σ. Misra et al. consider different priors with support over the set of symmetric positive definite $\tilde{\Sigma}$. They show that given the prior $p(\tilde{\mu}, \tilde{\Sigma}) = \frac{1}{|\tilde{\Sigma}|^{\frac{d+1}{2}}}$, the solution to (3) yields the Bayesian entropy estimate

$$\hat{h}_{\text{Bayesian}} = \frac{d \ln \pi}{2} + \frac{\ln|S|}{2} - \frac{1}{2} \sum_{i=1}^{d} \psi\left(\frac{n-i}{2}\right). \quad (4)$$

For discrete random variables, a Bayesian approach that uses binning for estimating functionals such as entropy was proposed the same year as Misra et al.'s work [7].

It is interesting to note that the estimates $\hat{h}_{\text{UMVU}}$ and $\hat{h}_{\text{Bayesian}}$ were derived from different perspectives, but differ only slightly in the digamma argument. Like $\hat{h}_{\text{UMVU}}$, Misra et al. show that $\hat{h}_{\text{Bayesian}}$ is an unbiased entropy estimate. They also show that $\hat{h}_{\text{Bayesian}}$ is dominated by a Stein-type estimator,

$$\hat{h}_{\text{Stein}} = \ln |S + n\bar{x}\bar{x}^T| - c_1,$$

where c_1 is a function of d and n. Further, they also show that the estimate $\hat{h}_{\text{Bayesian}}$ is dominated by a Brewster-Zidek-type estimator $\hat{h}_{BZ}$,

$$\hat{h}_{\text{BZ}} = \ln |S + n\bar{x}\bar{x}^T| - c_2.$$

where c_2 is a function of $|S|$ and $\bar{x}\bar{x}^T$ that requires calculating the ratio of two definite integrals, stated in full in (4.3) of [5]. Misra et al. found that on simulated numerical experiments their Stein-type and Brewster-Zidek-type estimators achieved roughly only 6% improvement over the simpler Bayesian estimate $\hat{h}_{\text{Bayesian}}$, and thus they recommend using the computationally much simpler $\hat{h}_{\text{Bayesian}}$ in applications.

There are two practical problems with the previously proposed parametric estimators. First, the estimates given by (2), (4), and the other proposed Misra et al. estimators require calculating the determinant of S or $S + \bar{x}\bar{x}^T$, which can be numerically infeasible if there are few samples. Second, the Bayesian estimate $\hat{h}_{\text{Bayesian}}$ uses the digamma function of $n - d$ which requires $n > d$ samples so that the digamma has a non-negative argument, and similarly $\hat{h}_{\text{UMVU}}$ uses the digamma of $n - d + 1$, which requires $n \geq d$ samples. Thus, although the knowledge that one is estimating the entropy of a Gaussian should be of use, for the $n \leq d$ case one must turn to nonparametric entropy estimators.

A thorough review of work in nonparametric entropy estimation up to 1997 was written by Beirlant et al. [4], including density estimation approaches, sample-spacing approaches, and nearest-neighbor estimators. Recently, Nilsson and Kleijn show that high-rate quantization approximations of Zador and Gray can be used to estimate Renyi entropy, and that the limiting case of Shannon entropy produces a nearest-neighbor estimate that depends on the number of quantization cells [8]. The special case of their nearest-neighbor estimate that best validates the high-rate quantization assumptions is when the number of quantization cells is as large as possible. They show that this special case produces the nearest-neighbor differential entropy estimator originally proposed by Kozachenko and Leonenko in 1987 [9]:

$$\hat{h}_{\text{NN}} = \frac{d}{n}\sum_{i=1}^{n} \ln \|x_i - x_{i,1}\|_2 + \ln(n-1) + \gamma + \ln V_d \qquad (5)$$

where $x_{i,1}$ is x_i's nearest neighbor in the sample set, γ is the Euler-Mascheroni constant, and V_d is the volume of the d-dimensional hypersphere with radius 1: $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$. A problem with this approach is that in practice data samples may not be in general position; for example, image pixel data are usually quantized to 8 bits or 10 bits. Thus, it can happen in practice that two samples have the exact same measured value and thus $\|x_n - x_{n,1}\|$ is zero and thus the entropy estimate could be ill-defined. Though there are various fixes, such as pre-dithering the quantized data, it is not clear what effect these fixes could have on the estimated entropy.

A different approach is taken by Costa and Hero [2], [3], [10]; they use the Beardwood Halton Hammersley result that the function of the length of a minimum spanning graph converges to the Renyi entropy [11] to form an estimator based on the empirical length of a minimum spanning tree of data. Unfortunately, how to use this approach to estimate Shannon entropy remains an open question.

III. Bayesian Estimate with Inverted Wishart Prior

We propose to estimate the entropy as, $E_N[h(N)]$, where N is a random Gaussian, and the prior $p(N)$ is an inverted Wishart distribution with scale parameter q and parameter matrix B. We use a Fisher information metric to define a measure over the Riemannian manifold formed by the set of Gaussian distributions. These choices for prior and measure are very similar to the choices that we found worked very well for Bayesian quadratic discriminant analysis [12], and further details on this framework can be found in that work.

The resulting proposed inverted Wishart Bayesian entropy estimate is

$$\begin{aligned} \hat{h}_{iWBayesian} \quad &= \quad \frac{d \ln \pi}{2} + \frac{\ln |S + B|}{2} \\ &\quad - \frac{1}{2}\sum_{i=1}^{d} \psi\left(\frac{n+q+1-i}{2}\right). \qquad (6) \end{aligned}$$

Proof: To show that (6) is $E_N[h(N)]$, we will need to integrate

$$\begin{aligned} &\int_{\tilde{\Sigma}>0} \frac{\ln |\tilde{\Sigma}| \exp[-\text{tr}(\tilde{\Sigma}^{-1}\text{V})]}{|\tilde{\Sigma}|^{\frac{q}{2}}} d\tilde{\Sigma} \qquad (7) \\ \equiv \quad & E_\Sigma[\ln |\Sigma|] \frac{|V|^{\frac{q-d-1}{2}}}{\Gamma_d(\frac{q-d-1}{2})} \qquad (8) \end{aligned}$$

where Σ is a random covariance matrix drawn from an inverted Wishart distribution with scale parameter q and matrix parameter $2V$. Recall that for any matrix $2V$, $|\Sigma^{-1}|/|(2V)^{-1}| \sim \prod_{i=1}^{d} \chi^2_{q-d-i}$ [13, Corollary 7.3], where χ^2 denotes the chi-squared random variable. Take the natural log of both sides and use the fact that $|A^{-1}| = |A|^{-1}$ to show that $\ln |\Sigma| \sim \ln |2V| - \sum_{i=1}^{d} \ln \chi^2_{q-d-i}$. Then after taking $E_\Sigma[\ln |\Sigma|]$, (8) becomes

$$\begin{aligned} & \frac{\Gamma_d(\frac{q-d-1}{2})}{|V|^{\frac{q-d-1}{2}}} \left(\ln |2V| - \sum_{i=1}^{d} E\left[\ln \chi^2_{q-d-i}\right] \right) \\ = \quad & \frac{\Gamma_d(\frac{q-d-1}{2})}{|V|^{\frac{q-d-1}{2}}} \left(\ln |V| - \sum_{i=1}^{d} \psi\left(\frac{q-d-i}{2}\right) \right), \qquad (9) \end{aligned}$$

where the second line uses the property of the χ^2 distribution that $E[\ln \chi^2_q] = \ln 2 + \psi\left(\frac{q}{2}\right)$.

Now we will use the above integral identity to find $\hat{h}_{\text{iWBayesian}} = E_N[h(N)]$. Solving for $E_N[h(N)]$ only requires computing

$$E_N[\ln|\Sigma|] = \int_{\tilde{\Sigma}} \ln|\tilde{\Sigma}| \; p(\tilde{\Sigma}|x_1, x_2, \ldots, x_n) \frac{d\tilde{\Sigma}}{|\tilde{\Sigma}|^{\frac{d+2}{2}}},$$

where the term $1/|\tilde{\Sigma}|^{(d+2)/2}$ results from the Fisher information metric which converts the integral $E_N[h(N)]$ from an integral over the statistical manifold of Gaussians to an integral over covariance matrices, and the posterior $p(\tilde{\Sigma}|x_1, x_2, \ldots, x_n)$ is given in [12] such that $E_N[\ln|\Sigma|]$

$$\begin{align}
&= \left(\frac{|S+B|^{\frac{n+q}{2}}}{2^{\frac{(n+q)d}{2}} \Gamma_d(\frac{n+q}{2})} \right) \notag \\
&\quad \cdot \left(\int_{\tilde{\Sigma}>0} \frac{\ln|\tilde{\Sigma}| \; \exp[-\frac{1}{2}\mathrm{tr}(\tilde{\Sigma}^{-1}(\mathrm{S+B}))]}{|\tilde{\Sigma}|^{\frac{n+q+d+1}{2}}} d\tilde{\Sigma} \right) \notag \\
&= \left(\frac{|S+B|^{\frac{n+q}{2}}}{2^{\frac{(n+q)d}{2}} \Gamma_d(\frac{n+q}{2})} \frac{\Gamma_d(\frac{n+q}{2})}{\left|\frac{S+B}{2}\right|^{\frac{n+q}{2}}} \right) \notag \\
&\quad \cdot \left(\ln\left|\frac{S+B}{2}\right| - \sum_{i=1}^{d} \psi\left(\frac{n+q+1-i}{2}\right) \right) \tag{10} \\
&= \ln|S+B| - d\ln 2 - \sum_{i=1}^{d} \psi\left(\frac{n+q+1-i}{2}\right), \notag
\end{align}$$

where equation (10) follows by using the fact that (7) is given by (9) for $V = (S+B)/2$. Then replacing $\ln|\Sigma|$ in (1) with the computed $E_N[\ln|\Sigma|]$ produces the estimator (6).

A. Choice of Prior Parameters q and B

The inverted Wishart distribution is an unimodal prior that gives maximum a priori probability to Gaussians with $\Sigma = B/q$. Previous work using the inverted Wishart for Bayesian estimation of Gaussians has used the identity matrix for B [14], or a scaled identity where the scale factor was learned by cross-validation given labeled training data (for classification) [15]. Using $B = I$ sets the maximum of the prior at I/q, regardless of whether the data are measured in nanometers or trillions of dollars. To a rough approximation, the prior regularizes the likelihood towards the prior maximum at B/q, and thus the bias added by using the prior with $B/q = I$ can be ill-suited to the problem. Instead, setting B/q to be a rough estimate of the covariance can add bias that is more appropriate for the problem. For example, we have shown that using $B = q\mathrm{diag}(S)$ can work well when estimating Gaussians for classification by Bayesian quadratic discriminant analysis [12]. For entropy estimation, $B = q\mathrm{diag}(S)/n$ works excellently if the true covariance is diagonal, but can perform poorly if the true covariance is a full covariance because the determinant of $B = \mathrm{diag}(S)/n$ can be significantly higher than the determinant of S, biasing the entropy estimate to be too high. An optimal choice of B when there is no prior information about S remains an open question; we propose using

$$B = q\frac{\ln(\mathrm{diag}(S)+1)}{n}, \tag{11}$$

for $n \le d$, and for $n > d$ we let B be the $d \times d$ matrix of zeros.

The scalar prior parameter q changes the peaked-ness of the inverted Wishart prior, with higher q corresponding to a more peaked prior, and thus higher bias. For the entropy problem there is only one number to estimate, and thus we believe that as little bias as possible should be used. To achieve this, we set $q = \min(d-n, -1)$. Letting $q = -1$ when $n > d$ and using the zero matrix for B in this range makes $\hat{h}_{\text{iWBayesian}}$ equivalent to $\hat{h}_{\text{Bayesian}}$ when $n > d$.

B. Bregman Divergences and Bayesian Estimation

Misra et al. showed that $E_\Sigma[|\ln\Sigma|]$ minimizes the squared error loss as stated in (3). Here we show that $E_N[h(N)]$ minimizes any Bregman divergence loss. The Bregman divergences are a class of loss functions that include squared error and relative entropy [16], [17].

Lemma: The mean entropy with respect to uncertainty in the generating distribution $E_N[h(N)]$ solves

$$\arg\min_{\hat{h}\in\mathcal{R}} E_N\left[d_\phi(h(N), \hat{h})\right],$$

where $d_\phi(a,b) = \phi(a) - \phi(b) - \phi(b)'(a-b)$ for $a, b, \in \mathbb{R}$ is any Bregman divergence with strictly convex ϕ.

Proof: Set the first derivative to zero:

$$\begin{aligned}
0 &= \frac{d}{d\hat{h}} E_N\left[\phi(h(N)) - \phi(\hat{h}) - \frac{d}{d\hat{h}}\phi(\hat{h})(h(N) - \hat{h})\right] \\
&= -E_N\left[\left(\frac{d^2}{d\hat{h}^2}\phi(\hat{h})\right)\left(h(N) - \hat{h}\right)\right].
\end{aligned}$$

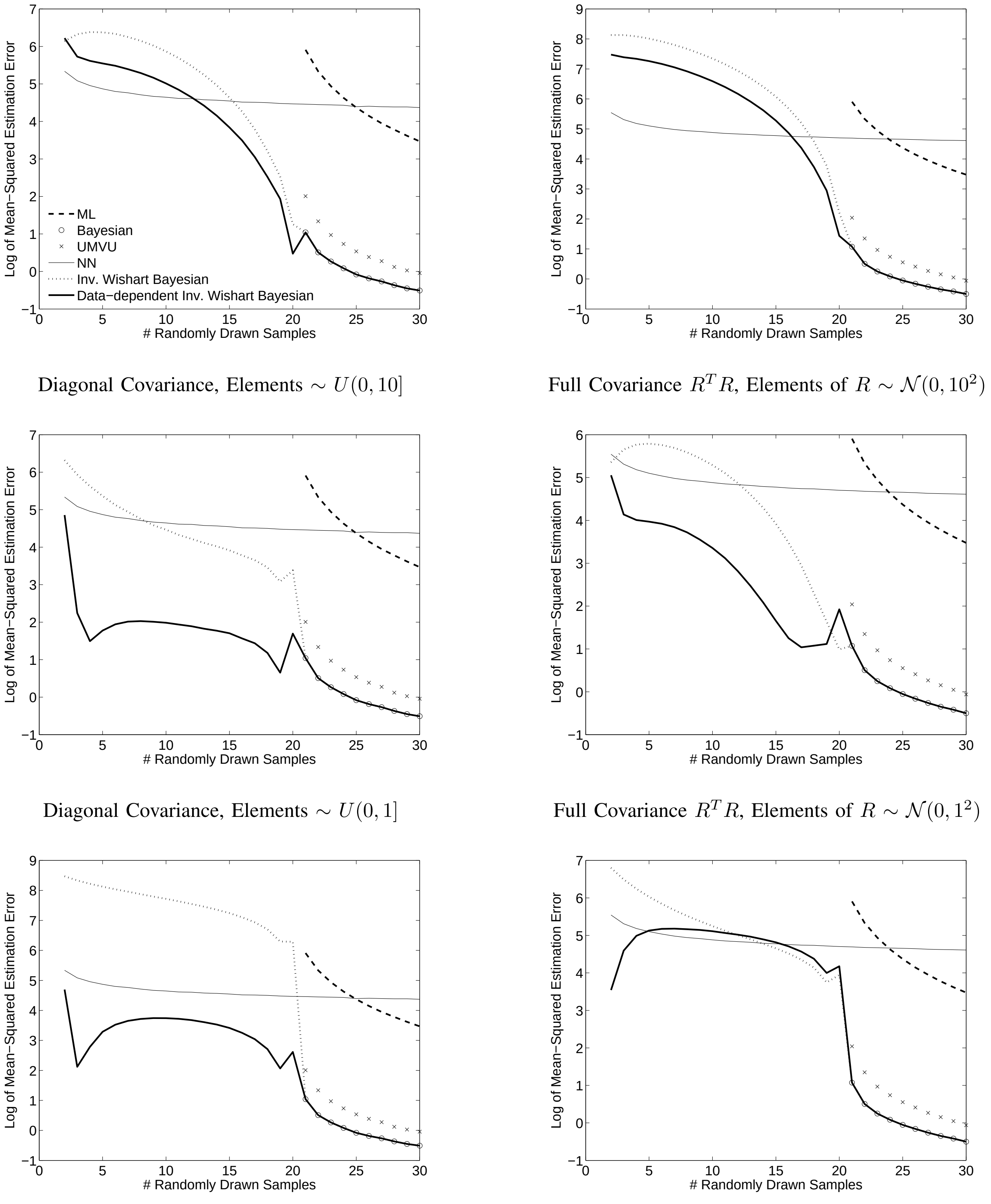
Because ϕ is strictly convex, $\frac{d^2}{d\hat{h}^2}\phi(\hat{h}) > 0$, and thus it must be that $E_N[h(N) - \hat{h}] = 0$, and thus by linearity that the minimizer is $\hat{h} = E_N[h(N)]$.

IV. Experiments

We compare the proposed inverted Wishart Bayesian estimator $\hat{h}_{\text{iWBayesian}}$ with the data-dependent given in (11) to $\hat{h}_{\text{iWBayesian}}$ with $B = I$, to the nearest-neighbor estimator given in (5), to the maximum likelihood estimator formed by replacing μ and Σ in (1) by maximum likelihood estimates of μ and Σ, to $\hat{h}_{\text{UMVU}}$ given in (2), and to $\hat{h}_{\text{Bayesian}}$ given in (4). All results were computed with Matlab 7.0. For the digamma function and Euler-Mascheroni constant we used the corresponding built-in Matlab commands.

Simulations were run for a fixed dimension of $d = 20$ with varying number of iid samples n drawn from random Gaussians. For $n \le d$ we it was not possible to calculate the digamma functions for $\hat{h}_{\text{Bayesian}}$ and $\hat{h}_{\text{UMVU}}$ or $|\Sigma|$, thus $\hat{h}_{\text{Bayesian}}$, $\hat{h}_{\text{UMVU}}$, and the maximum likelihood estimates are only reported for $n > d$. The nearest-neighbor estimate and the inverted Wishart Bayesian estimates are compared down to $n = 2$ samples.

In the first simulation, each generating Gaussian had a diagonal covariance matrix with elements drawn iid from the



Diagonal Covariance, Elements $\sim U(0,10]$

Full Covariance R^TR, Elements of $R \sim \mathcal{N}(0,10^2)$

Diagonal Covariance, Elements $\sim U(0,1]$

Full Covariance R^TR, Elements of $R \sim \mathcal{N}(0,1^2)$

Diagonal Covariance, Elements $\sim U(0,.1]$

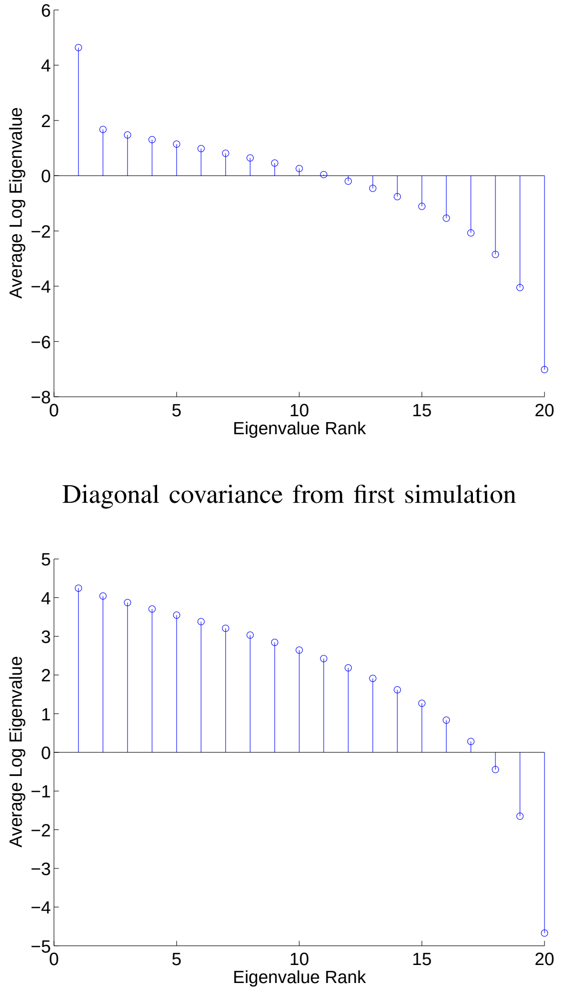
Full Covariance R^TR, Elements of $R \sim \mathcal{N}(0,.1^2)$

Fig. 1. Comparison of entropy estimators averaged over 10,000 iid runs of each simulation.

uniform distribution on $(0, \alpha]$. The results are shown in Fig. 1 (left) for $\alpha = 10$ (top), for $\alpha = 1$ (middle), and for $\alpha = .1$ (bottom). Fig. 2 shows the average eigenvalues.

In the second simulation, each generating Gaussian had a full covariance matrix $R^T R$, where each of the 20×20 elements of R was drawn iid from a normal distribution $\mathcal{N}(0, \alpha^2)$. The results are shown in Fig. 1 (right) for $\alpha = 10$ (top), for $\alpha = 1$ (middle), and for $\alpha = .1$ (bottom). Fig. 2 shows the average eigenvalues.

For each n and each of the six cases, the simulation was run 10,000 times and the results averaged to produce the plots.



Diagonal covariance from first simulation

Full covariance from second simulation

Fig. 2. Average log ranked eigenvalues for the first simulation (top) and second simulation (bottom).

For $n > d$ the three Bayesian estimates are equivalent and perform consistently better than $\hat{h}_{UMVU}$, the maximum likelihood estimate, or the nearest-neighbor estimate. The maximum likelihood estimator is always the worst parametric estimator. Throughout the simulations, the nearest-neighbor estimate makes the least use of additional samples, improving its estimate only slowly. This is reasonable because the nearest-neighbor estimator does not explicitly use the information that the true distribution is Gaussian.

Given very few samples the nearest-neighbor estimator is the best performer for two of the full covariance cases. This suggests that different prior parameter settings could be more effective when there are few samples, perhaps a prior that adds more bias.

As expected, in general the data-dependent $\hat{h}_{\text{iWBayesian}}$ achieves lower error than $\hat{h}_{\text{iWBayesian}}$ with $B = I$, sometimes significantly better, as in the case of true diagonal covariance with elements drawn from $U(0, .1]$, shown in Fig. 1 (bottom).

V. Conclusions

A data-dependent approach to Bayesian entropy estimation was given that minimizes expected Bregman divergence and performs consistently well compared to other possible estimators for the high-dimensional/limited data case that $n \leq d$.

References

[1] T. Cover and J. Thomas, *Elements of Information Theory*. United States of America: John Wiley and Sons, 1991.

[2] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Information Theory*, vol. 45, pp. 1921–1939, 1999.

[3] J. Costa and A. Hero, "Geodesic entropic graphs for dimension and entropy estimation in manifold learning," *IEEE Trans. Signal Processing*, vol. 52, no. 8, pp. 2210–2221, 2004.

[4] J. Beirlant, E. Dudewicz, L. Györfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal Mathematical and Statistical Sciences*, vol. 6, pp. 17–39, 1987.

[5] N. Misra, H. Singh, and E. Demchuk, "Estimation of the entropy of a multivariate normal distribution," *Journal of Multivariate Analysis*, vol. 92, pp. 324–342, 2005.

[6] N. A. Ahmed and D. V. Gokhale, "Entropy expressions and their estimators for multivariate distributions," *IEEE Trans. Information Theory*, pp. 688–692, 1989.

[7] D. Endres and P. Földiàk, "Bayesian bin distribution inference and mutual information," *IEEE Trans. Information Theory*, pp. 3766–3779, 2005.

[8] M. Nilsson and W. B. Kleijn, "On the estimation of differential entropy from data located on embedded manifolds," *IEEE Trans. on Information Theory*, vol. 53, no. 7, pp. 2330–2341, 2007.

[9] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of entropy of a random vector," *Problems in Information Transmission*, vol. 23, no. 1, pp. 95–101, 1987.

[10] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, pp. 85–95, 2002.

[11] J. Beardwood, J. H. Halton, and J. M. Hammersley, "The shortest path through many points," *Proc. Cambridge Philo. Soc.*, vol. 55, pp. 299–327, 1959.

[12] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1287–1314, 2007.

[13] M. Bilodeau and D. Brenner, *Theory of Multivariate Statistics*. New York: Springer Texts in Statistics, 1999.

[14] D. G. Keehn, "A note on learning for Gaussian properties," *IEEE Trans. on Information Theory*, vol. 11, pp. 126–132, 1965.

[15] P. J. Brown, T. Fearn, and M. S. Haque, "Discrimination with many variables," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1320–1329, 1999.

[16] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[17] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence and Bayesian estimation of distributions," *arXiv preprint cs.IT/0611123, 2006.*