*Entropy* **2010**, *12*, 1-x; doi:10.3390/——

OPEN ACCESS

*entropy*

**ISSN 1099-4300**

www.mdpi.com/journal/entropy

*Article*

# Parametric Bayesian Estimation of Differential Entropy and Relative Entropy

**Maya Gupta** [1,⋆] **and Santosh Srivastava** [2]

[1] Dept. of Electrical Engineering, University of Washington, Seattle WA 98195-2500, USA

[2] Computational Biology, Fred Hutchinson Cancer Research Center, Seattle WA 98109, USA

⋆ Author to whom correspondence should be addressed; E-Mail: gupta@ee.washington.edu.

*Received: 16 November 2009; in revised form: 28 March 2010 / Accepted: 2 April 2010 / Published: xx*

**Abstract:** Given iid samples drawn from a distribution with known parametric form, we propose the minimization of expected Bregman divergence to form Bayesian estimates of differential entropy and relative entropy, and derive such estimators for the uniform, Gaussian, Wishart, and inverse Wishart distributions. Additionally, formulas are given for a log gamma Bregman divergence and the differential entropy and relative entropy for the Wishart and inverse Wishart. The results, as always with Bayesian estimates, depend on the accuracy of the prior parameters, but example simulations show that the performance can be substantially improved compared to maximum likelihood or state-of-the-art nonparametric estimators.

**Keywords:** Kullback-Leibler; relative entropy; differential entropy; Pareto; Wishart

## 1. Introduction

Estimating differential entropy and relative entropy is useful in many applications of coding, machine learning, signal processing, communications, chemistry, and physics. For example, relative entropy between maximum likelihood-fit Gaussians has been used for biometric identification [1], differential entropy estimates have been used for analyzing sensor locations [2], and mutual information estimates have been used in the study of EEG signals [3].

In this paper we present Bayesian estimates for differential entropy and relative entropy that are optimal in the sense of minimizing expected Bregman divergence between the estimate and the uncertain true distribution. We illustrate techniques that may be used for a wide range of parametric distributions, specifically deriving estimates for the uniform (a non-exponential example), Gaussian (perhaps the most

popular distribution), and the Wishart and inverse Wishart (the most commonly used distributions for positive definite matrices).

Bayesian estimates for differential entropy and relative entropy have previously been derived for the Gaussian [4], but our estimates differ in that we take a distribution-based approach, and we use a prior that results in numerically stable estimates even when the number of samples is smaller than the dimension of the data. Performance of the presented estimates will depend on how well the user is able to choose the prior distribution's parameters, and we do not attempt a rigorous experimental study here. However, we do present simulated results for the uniform distribution (where no prior is needed), that show that our approach to forming these estimates can result in significant performance improvements over maximum likelihood estimates and over the state-of-the-art nearest-neighbor nonparametric estimates [5].

First we define notation that will be used throughout the paper. In Section II we review related work in estimating differential entropy and relative entropy. In Section III we show that the proposed Bayesian estimates are optimal in the sense of minimizing expected Bregman divergence loss. In the remaining sections, we present differential entropy and relative entropy estimates for the uniform, Gaussian, Wishart and inverse Wishart distributions given iid samples drawn from the underlying distributions.

All proofs and derivations are in the Appendix.

### *1.1. Notation and Background*

If $P$ and $Q$ were the known parametric distributions of two random variables with respective densities $p$ and $q$, then the differential entropy of $P$ is

$$h(P) = -\int_x p(x) \ln p(x) dx$$

and the relative entropy between $P$ and $Q$ is

$$\text{KL}(P||Q) = \int_x p(x) \ln \frac{p(x)}{q(x)} dx$$

For estimating differential entropy, we assume that one has drawn iid samples $\{x_1, x_2, \ldots, x_n\}$ from distribution $P$ where $x_i \in \mathbb{R}^d$ is a $d \times 1$ vector, and the samples have mean $\bar{x}$ and scaled sample covariance $S = \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})^T$. The notation $x_j[i]$ will be used to refer to the value of the $i$th component of vector $x_j$.

For estimating relative entropy, we assume that one has drawn iid $d$-dimensional samples from both distributions $P$ and $Q$, and we denote the samples drawn from $P$ as $\{x_{1,1}, x_{1,2}, \ldots, x_{1,n_1}\}$ and the samples drawn from $Q$ as $\{x_{2,1}, x_{2,2}, \ldots, x_{2,n_2}\}$. The empirical means are denoted by $\bar{x}_1$ and $\bar{x}_2$, and the scaled sample covariances are denoted by $S_1 = \sum_{j=1}^{n_1} (x_{1,j} - \bar{x}_1)(x_{1,j} - \bar{x}_1)^T$ and $S_2 = \sum_{j=1}^{n_2} (x_{2,j} - \bar{x}_2)(x_{2,j} - \bar{x}_2)^T$.

In some places, we treat variables such as the covariance $\Sigma$ as random, and we consistently denote realizations of random variables with a tilde, e.g., $\tilde{\Sigma}$. Expectations are always taken with respect to the posterior distribution unless otherwise noted. The digamma function is denoted by $\psi(z) \triangleq \frac{d}{dz} \ln \Gamma(z)$, where $\Gamma$ is the standard gamma function; and $\Gamma_d$ denotes the standard multi-dimensional gamma function.

Let $W$ be distributed according to a Wishart distribution with scalar degree of freedom parameter $q \geq d$ and positive definite matrix parameter $\Sigma \in \mathbb{R}^{d \times d}$ if

$$p(W = \tilde{W}) = \frac{|\tilde{W}|^{\frac{q-d-1}{2}} \exp\left(-\frac{1}{2}\text{tr}(\tilde{W}\Sigma^{-1})\right)}{2^{\frac{qd}{2}}\Gamma_d\left(\frac{q}{2}\right)|\Sigma|^{\frac{q}{2}}} \tag{1}$$

Let $V$ be distributed according to an inverse Wishart distribution with scalar degree of freedom parameter $q \geq d$ and positive definite matrix parameter $\Sigma \in \mathbb{R}^{d \times d}$ if

$$p(V = \tilde{V}) = \frac{|\Sigma|^{\frac{q}{2}} \exp\left(-\frac{1}{2}\text{tr}(\tilde{V}^{-1}\Sigma)\right)}{2^{\frac{qd}{2}}\Gamma_d(\frac{q}{2})|\tilde{V}|^{\frac{q+d+1}{2}}} \tag{2}$$

Note that $V^{-1}$ is then distributed as a Wishart random matrix with parameters $q$ and $\Sigma^{-1}$.

## 2. Related Work

First we review related work in parametric differential entropy estimation, then in nonparametric differential entropy estimation, and then in estimating relative entropy.

### *2.1. Prior Work on Parametric Differential Entropy Estimation*

A common approach to estimate differential entropy (and relative entropy) is to find the maximum likelihood estimate for the parameters and then substitute them into the differential entropy formula. For example, for the multivariate Gaussian distribution, the maximum likelihood differential entropy estimate of a $d$-dimensional random vector $X$ drawn from the Gaussian $\mathcal{N}(\mu, \Sigma)$ is

$$\hat{h}_{\text{ML}} = \frac{d}{2} + \frac{d\ln(2\pi)}{2} + \frac{\ln|\Sigma_{ML}|}{2}$$

Similarly, if samples $\{x_i\}$ are drawn iid from a one-dimensional uniform distribution, the maximum likelihood differential entropy estimate is $\hat{h}_{\text{ML}} = \ln(\max_i(\{x_i\}) - \min_i(\{x_i\}))$, which will always be an under-estimate of the true differential entropy.

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

$$\frac{d}{2} + \frac{d\ln\pi}{2} + \frac{\ln|S|}{2} - \frac{1}{2}\sum_{i=1}^{d}\psi\left(\frac{n+1-i}{2}\right) \tag{3}$$

However, they treated the random sample covariance of $n$ IID Gaussian samples as if it were drawn from a Wishart with $n$ degrees of freedom, when in fact it is drawn from a Wishart of $n-1$ degrees of freedom, and thus the UMVU estimator they derived should be stated:

$$\frac{d}{2} + \frac{d\ln\pi}{2} + \frac{\ln|S|}{2} - \frac{1}{2}\sum_{i=1}^{d}\psi\left(\frac{n-i}{2}\right) \tag{4}$$

Bayesian differential entropy estimation was first proposed for the multivariate normal in 2005 by Misra *et al.* [4]. They formed an estimate of the multivariate normal differential entropy by substituting

$\widehat{\ln|\Sigma|}$ for $\ln|\Sigma|$ in the differential entropy formula for the Gaussian, where their $\widehat{\ln|\Sigma|}$ minimizes the expected squared-difference of the differential entropy estimate:

$$\widehat{\ln|\Sigma|} = \arg\min_{\delta\in\mathbb{R}} E_{\mu,\Sigma}\left[\left(\delta - \ln|\Sigma|\right)^2\right] \tag{5}$$

They also considered different priors with support over the set of positive definite matrices. Using the prior $p(\tilde{\mu},\tilde{\Sigma}) = \frac{1}{|\tilde{\Sigma}|^{\frac{d+1}{2}}}$ to solve (5) results in the same estimate as the correct UMVU estimate [4], given in (4). Misra *et al.* show that (4) is dominated by a Stein-type estimator $\ln|S+n\bar{x}\bar{x}^T| - c_1$, where $c_1$ is a function of $d$ and $n$ [4]. In addition, they show that (4) is dominated by a Brewster-Zidek-type estimator $\ln|S+n\bar{x}\bar{x}^T| - c_2$, 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 [4]. Misra *et al.* found that on simulated numerical experiments their Stein-type and Brewster-Zidek-type estimators achieved roughly only $6\%$ improvement over (4), and thus they recommend using the computationally much simpler (4) for applications.

There are two practical problems with the previously proposed parametric differential entropy estimators. First, the estimates given by (3), (4), and the other estimators investigated by Misra *et al.* require calculating the determinant of $S$ or $S+\bar{x}\bar{x}^T$, which is problematic if $n < d$. Second, the estimate (4) uses the digamma function of $n-d$ which requires $n > d$ samples so that the digamma has a non-negative argument. Thus, although the knowledge that one is estimating the differential entropy of a Gaussian should be of use, for the $n \leq d$ case one must currently turn to nonparametric differential entropy estimators.

### *2.2. Prior Work on Nonparametric Differential Entropy Estimation*

Nonparametric differential entropy estimation up to 1997 has been thoroughly reviewed by Beirlant *et al.* [7], 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 that best validates the high-rate quantization assumptions is when the number of quantization cells is as large as possible, and they show that this special case produces the nearest-neighbor differential entropy estimator originally proposed by Kozachenko and Leonenko in 1987 [9]:

$$\hat{h}_{\text{NN}} = \frac{d}{n}\sum_{j=1}^{n}\ln\rho(j) + \ln(n-1) + \gamma + \ln V_d \qquad \text{for} \qquad \rho(j) = \min_{k=1,\ldots,n,k\neq j}\|x_j - x_k\|_2 \tag{6}$$

where $\gamma$ is the Euler-Mascheroni constant, and $V_d$ is the volume of the $d$-dimensional hypersphere with radius 1: $V_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$. Other variants of nearest-neighbor differential entropy estimators have also been proposed and analyzed [10,11]. A practical problem with the nearest-neighbor approach is that data samples are often quantized, for example, image pixel data are usually quantized to eight bits or ten bits. Thus, it can happen in practice that two samples $x_j$ and $x_k$ have the exact same measured value so that $\rho(j) = 0$ and the differential entropy estimate is ill-defined. Though there are various fixes, such as pre-dithering the quantized data, it is not clear what effect such fixes could have on the estimated differential entropy.

A different approach is taken by Hero *et al.* [12–14]. They relate a result of Beardwood-Halton-Hammersley on the limiting length of a minimum spanning graph to Renyi entropy, and form a Renyi entropy estimator based on the empirical length of a minimum spanning tree of data. Unfortunately, how to use this approach to estimate the special case of Shannon entropy remains an open question.

In other recent work on differential entropy estimation, Van Hulle took a semiparametric approach to nonparametric differential entropy estimation for a continuous density by using a 5th-order Edgeworth expansion about the maximum likelihood multivariate normal given the data samples drawn from a non-normal distribution [15].

### *2.3. Prior Work on Relative Entropy Estimation*

There is relatively little work on estimating relative entropy for continuous distributions. Wang *et al.* explored a number of data-dependent partitioning approaches for relative entropy between any two absolutely continuous distributions [16]. Nguyen *et al.* took a variational approach to relative entropy estimation [17], which was reported to work better for some cases than the data-partitioning estimators.

In more recent work [5,18], Wang *et al.* proposed a nearest-neighbor estimator based on nearest-neighbor density estimation:

$$\widehat{\mathrm{KL}}_{NN} = \ln \frac{n_2}{n_1 - 1} + \frac{d}{n_1} \sum_{j=1}^{n_1} \ln \frac{\nu(j)}{\rho(j)} \tag{7}$$

where

$$\nu(j) = \min_{k=1,\ldots,n_2} \|x_{1,j} - x_{2,k}\|_2 \qquad \text{and} \qquad \rho(j) = \min_{k=1,\ldots,n_1, k\neq j} \|x_{1,j} - x_{1,k}\|_2$$

They showed that (7) significantly outperforms their best data-partitioning estimators [5,18]. Peréz-Cruz has contributed additional convergence analysis for these estimators [19]. In practice, like the nearest-neighbor entropy estimate, $\widehat{\mathrm{KL}}_{NN}$ may be ill-defined if samples are quantized.

The nearest-neighbor relative entropy estimator can perform quite poorly for Gaussian distributed data given a reasonable number of finite samples, particularly in high-dimensions. For example, consider the case of two high-dimensional Gaussians each with identity covariance and a finite iid sample of points from the two distributions. Their true relative entropy is a function of $\|\mu_1 - \mu_2\|^2$, whereas the nearest neighbor estimated relative entropy is better approximated (though roughly so) as a function of $\ln \|\mu_1 - \mu_2\|^2$.

## 3. Functional Estimates that Minimize Expected Bregman Loss

Here we propose to form estimators of functionals (such as differential entropy and relative entropy) that are optimal in the sense that they minimize the expected Bregman loss, and that are always computable (assuming an appropriate prior is used).

Consider samples $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ drawn iid from some unknown distribution $A$, where we model $A$ as a random distribution drawn from a distribution over distributions $P_A$ that has density $p_A$. We use $\tilde{A}$ to denote a realization of the random distribution $A$.

The goal is to estimate some functional (such as differential entropy or relative entropy) $\xi$, where $\xi$ maps a distribution or set of distributions (e.g., relative entropy is a functional on pairs of distributions) to a real number $\xi : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A}$ is the Cartesian product of finite distributions $\mathcal{A} = A_1 \times A_2 \times \ldots \times A_M$, and we denote a realization of $\mathcal{A}$ as $\tilde{\mathcal{A}}$. For example, the functional relative entropy maps a pair of distributions $\mathcal{A} = A_1 \times A_2$ to a non-negative number.

We are interested in the Bayesian estimate of $\xi$ that minimizes an expected loss $L : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ [20]:

$$\begin{aligned} \xi^* &= \underset{\hat{\xi} \in \mathbb{R}}{\operatorname{argmin}} \int_{\tilde{\mathcal{A}}} L(\xi(\tilde{\mathcal{A}}), \hat{\xi}) dP_{\tilde{\mathcal{A}}} \\ &\equiv \underset{\hat{\xi} \in \mathbb{R}}{\operatorname{argmin}} E_{\mathcal{A}} \left[ L(\xi(\mathcal{A}), \hat{\xi}) \right] \end{aligned} \tag{8}$$

In this paper, we will focus on Bregman loss functions (Bregman divergences), which include squared error and relative entropy [21–24]. For any twice differentiable strictly convex function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, the corresponding Bregman divergence is $d_\phi(z, \hat{z}) = \phi(z) - \phi(\hat{z}) - \phi(\hat{z})'(z - \hat{z})$ for $z, \hat{z}, \in \mathbb{R}$.

The following proposition will aid in solving (8):

**Proposition 1.** *The expected functional $E_{\mathcal{A}}[\xi(\mathcal{A})]$ minimizes the expected Bregman loss such that*

$$E_{\mathcal{A}}[\xi(\mathcal{A})] = \arg\min_{z \in \mathbb{R}} E_{\mathcal{A}} \left[ d_\phi(\xi(\mathcal{A}), z) \right]$$

*if $E_{\mathcal{A}}[\xi(\mathcal{A})]$ exists and is finite.*

One can view this proposition as a special case of Theorem 1 of Banerjee *et al.* [22]; we provide a proof in the appendix for completeness.

In this paper we focus on estimating differential entropy and relative entropy, which by applying Proposition 1 we calculate respectively as:

$$\hat{h}_{\text{Bayesian}} = E_A[h(A)] \text{ and } \widehat{\text{KL}}_{\text{Bayesian}} = E_{A_1, A_2}[\text{KL}(A_1 \| A_2)]$$

assuming the expectations are finite.

## 4. Bayesian Differential Entropy Estimate of the Uniform Distribution

We present estimates of the differential entropy of an unknown uniform distribution over a hyperrectangular domain for two cases: first, that there is no prior knowledge about the uniform distribution; and second, that there is prior knowledge about the uniform given in the form of a Pareto prior.

### *4.1. No Prior Knowledge About the Uniform*

Given $n$ $d$-dimensional samples $\{x_1, x_2, \ldots, x_n\}$ drawn from a hyperrectangular $d$-dimensional uniform distribution, let $\Delta_i$ be the difference between the maximum and minimum sample in the $i$th dimension:

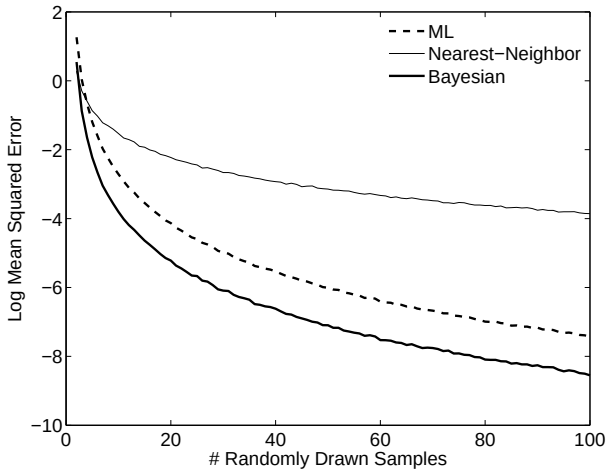
$$\Delta_i = \max_{j,k} x_j[i] - x_k[i]$$

Then because a hyperrectangular uniform is the product of independent marginal uniforms, its differential entropy is the sum of the marginal entropies. Given no prior knowledge about the uniform, we take the expectation with respect to the (normalized) likelihood, or equivalently using a non-informative flat prior. Then, the proposed differential entropy estimate is the sum over dimensions of the differential entropy estimate for each marginal uniform:

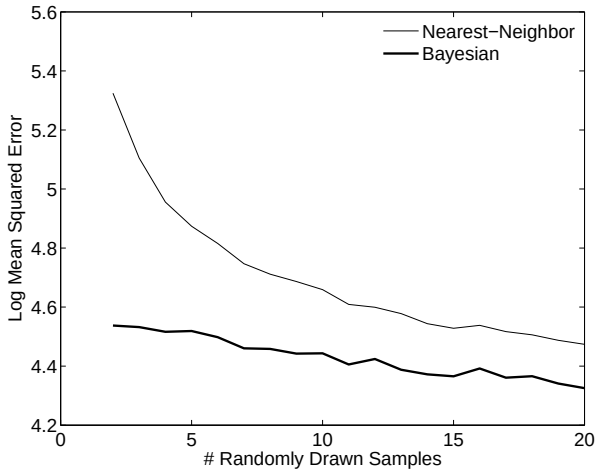
$$E_U[h(U)] = \sum_{i=1}^{d} \left( \ln \Delta_i + \frac{1}{n-1} + \frac{1}{n} \right) \tag{9}$$

To illustrate the effectiveness of the proposed Bayesian estimates, we show example results from two representative experiments in Figure 1.

**Figure 1**. Example comparison of differential entropy estimators. Left: For each of 10,000 runs of the simulation, $n$ samples were drawn iid from a uniform distribution on $[-5, 5]$. The proposed estimate (9) is compared to the maximum likelihood estimate, and to the nearest-neighbor estimate given in (6). Right: For each of 10,000 runs of the simulation, $n$ samples were drawn iid from a Gaussian distribution. For each of the 10,000 runs, a new Gaussian distribution with diagonal covariance was randomly generated by drawing each of the variances iid from a uniform on $[0, 1]$. The Bayesian estimator prior parameters were $q = d$ and $B = .5qI$. The proposed estimate (12) is compared to the only feasible estimator for this range of $n$, the nearest-neighbor estimate given in (6).




Uniform with Support $\delta = 10$




20-dimensional Gaussian

### *4.2. Pareto Prior Knowledge About the Uniform*

We consider the case that one has prior knowledge about the random uniform distribution $U$, where that prior knowledge is formulated as an independent Pareto prior for each dimension such that the prior

probability of the marginal $i$th-dimension uniform $\tilde{U}_\delta$ with support of length $\delta$ is:

$$p_i(\tilde{U}_\delta) = \begin{cases} \frac{\alpha_i \ell_i^{\alpha_i}}{\delta^{\alpha_i+1}} & \text{for } \delta \geq \ell_i \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $\alpha_i \in \mathbb{R}^+$ and $\ell_i \in \mathbb{R}^+$ are the Pareto distribution prior parameters for the $i$th dimension. The parameter $\ell_i$ defines the minimum length one believes the uniform's support could be in the $i$th dimension, and the parameter $\alpha_i$ specifies the confidence that $\ell_i$ is the right length; a larger $\alpha_i$ means one is more confident that $\ell_i$ is the correct length.

Then the differential entropy estimate for the $i$th dimension's one-dimensional uniform is:

$$E_U[h(U)]_i = \begin{cases} \ln \Delta_i + \frac{1}{n+\alpha_i} + \frac{1}{n+\alpha_i+1}, & \text{for } \Delta_i \geq \ell_i \\ \ln \ell_i + \frac{1}{(n+\alpha_i)+(n+\alpha_i)^2\left(\frac{\ell_i-\Delta_i}{\ell_i}\right)} + \frac{1}{n+\alpha_i+1} & \text{for } \Delta_i < \ell_i. \end{cases} \tag{11}$$

Note that the two cases given above do coincide for the boundary case that $\ell_i = \Delta_i$, so that this differential entropy estimate is a continuous function of $\Delta_i$. For the full $d$-dimensional uniform, the differential entropy estimate is the sum of the one-dimensional differential entropy estimates: $\sum_{i=1}^{d} E_U[h(U)]_i$.

## 5. Gaussian Distribution

The Gaussian is a popular model and often justified by central limit theorem arguments and because it is the maximum entropy distribution given fixed mean and covariance. In this section we assume $d$-dimensional samples have been drawn iid from an unknown Gaussian $N$, which we model as a random Gaussian and we take the prior to be an inverse Wishart distribution with scalar parameter $q \in \mathbb{R}$ and parameter matrix $B \in \mathbb{R}^{d \times d}$.

We use the Fisher information metric to define a measure over the Riemannian manifold formed by the set of Gaussian distributions [25–27]. We found these choices for prior and measure worked well for estimating Gaussian distributions for Bayesian quadratic discriminant analysis [27].

The performance of the proposed Gaussian entropy and relative entropy estimators will depend strongly on the choice of the prior. Generally, prior knowledge or subjective guesses about the data are used to set the parameters of the prior. Another choice to form a prior is to use a coarse estimate of the data, for example, in previous work we found that setting $B$ equal to the identity matrix times the trace of the sample covariance worked well as a data-adaptive prior in the context of classification [27]. Since the trace times the identity is the extremal case of maximum entropy Gaussian for a given trace, this specific approach is problematic as a coarse estimate for setting the prior for differential entropy estimation, but other coarse estimates based on a different statistic of the eigenvalue may work well.

### *5.1. Differential Entropy Estimate of the Gaussian Distribution*

Assume $n$ samples $\{x_1, x_2, \ldots, x_n\}$ have been drawn iid from an unknown $d$-dimensional normal distribution. Per Prop. 1, we estimate the differential entropy as: $E_N[h(N)]$, where the expectation is taken with respect to the posterior distribution over $N$ and the prior is taken to be inverse Wishart with

matrix parameter $B \in \mathbb{R}^{d\times d}$ and scale parameter $q \in \mathbb{R}$. See the appendix for full details and derivation. The resulting estimate is,

$$E_N[h(N)] = \frac{d\ln\pi}{2} + \frac{\ln|S+B|}{2} - \frac{1}{2}\sum_{i=1}^{d}\psi\left(\frac{n+q+i+1}{2}\right) \tag{12}$$

This estimate is well-defined for any number of samples $n$.

### *5.2. Relative Entropy Estimate between Gaussian Distributions*

Assume $n_1$ samples have been drawn iid from an unknown $d$-dimensional normal distribution $N_1$, and $n_2$ samples have been drawn iid from another $d$-dimensional distribution $N_2$, assumed independent from the first. Then following Prop. 1, we estimate the relative entropy as $E_{N_1,N_2}[\mathrm{KL}(N_1\|N_2)]$ where $N_1$ and $N_2$ are independent random Gaussians, the expectation is taken with respect to their posterior distributions, and the prior distributions are taken to be inverse Wisharts with scale parameters $q_1$ and $q_2$ and matrix parameters $B_1$ and $B_2$. See the appendix for full details and derivation. The resulting estimate is,

$$\begin{aligned}E_{N_1,N_2}[\mathrm{KL}(N_1\|N_2)] &= \frac{1}{2}\frac{n_2+q_2+d+1}{n_1+q_1}\mathrm{tr}((S_1+B_1)(S_2+B_2)^{-1}) - \frac{1}{2}\log\frac{|S_1+B_1|}{|S_2+B_2|}\\ &+ \frac{1}{2}\sum_{i=1}^{d}\left(\psi\left(\frac{n_2+q_2+1+i}{2}\right) - \psi\left(\frac{n_1+q_1+1+i}{2}\right)\right) - \frac{d}{2}\\ &+ \frac{1}{2}(n_2+q_2+d+1)\mathrm{tr}((\mathrm{S}_2+\mathrm{B}_2)^{-1}(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2)(\bar{\mathrm{x}}_1-\bar{\mathrm{x}}_2)^{\mathrm{T}}).\end{aligned} \tag{13}$$

This estimate is well-defined for any number of samples $n_1, n_2$. If the prior scalar parameters are taken to be the same, that is $q_1 = q_2$, then the digamma terms cancel.

## 6. Wishart and Inverse Wishart Distributions

The Wishart and inverse Wishart distributions are arguably the most popular distributions for modeling random positive definite matrices. Moreover, if a random variable has a Gaussian distribution, then its sample covariance is drawn from a Wishart distribution. The relative entropy between Wishart distributions may be a useful way to measure the dissimilarity between collections of covariance matrices or Gram (inner product) matrices.

We were unable to find expressions for differential entropy or relative entropy of the Wishart and inverse Wishart distributions, so we first present those, and then present Bayesian estimates of these quantities.

### *6.1. Wishart Differential Entropy and Relative Entropy*

The differential entropy of $W$ is

$$h(W) = \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{qd}{2} + \frac{d+1}{2}\ln|2\Sigma| - \frac{q-d-1}{2}\sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right) \tag{14}$$

The relative entropy between two Wishart distributions $p_1$ and $p_2$ with parameters $(q_1, \Sigma_1)$ and $(q_2, \Sigma_2)$ respectively is,

$$\mathrm{KL}(p_1||p_2) = \ln\left(\frac{\Gamma_d\left(\frac{q_2}{2}\right)}{\Gamma_d\left(\frac{q_1}{2}\right)}\right) + \frac{q_1}{2}\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1}\right) - \frac{q_1 d}{2} - \frac{q_2}{2}\ln|\Sigma_1\Sigma_2^{-1}| - \frac{q_2 - q_1}{2}\sum_{i=1}^{d}\psi\left(\frac{q_1 - d + i}{2}\right). \tag{15}$$

For the special case of $q_1 = q_2 = q$, we note that the relative entropy given in (15) is $q/2$ times Stein's loss function, which is itself a common Bregman divergence.

For the special case of $\Sigma_1 = \Sigma_2$, we find that the relative entropy between two Wisharts can also be written in the form a Bregman divergence [21] between $q_2$ and $q_1$. Consider the strictly convex function $\phi(q) = \ln\Gamma_d(q/2)$ for $q \in \mathbb{R}_+^d$, and let $\psi_d$ be the derivative of the $\Gamma_d$. Then (15) becomes,

$$\begin{aligned} &= \ln\Gamma_d\left(\frac{q_2}{2}\right) - \ln\Gamma_d\left(\frac{q_1}{2}\right) - \frac{q_2 - q_1}{2}\psi_d\left(\frac{q_2}{2}\right) \\ &= \phi(q_2) - \phi(q_1) - (q_2 - q_1)\phi'(q_1) \\ &= d_\phi(x, y). \end{aligned} \tag{16}$$

We term (16) the *log-gamma Bregman divergence*. We have not seen this divergence noted before, and hypothesize that this divergence may have physical or practical significance because of the widespread occurrence of the gamma function and its special properties [28].

### 6.2. *Inverse Wishart Differential Entropy and Relative Entropy*

Let $V$ be distributed according to an inverse Wishart distribution with scalar degree of freedom parameter $q \geq d$ and positive definite matrix parameter $\Sigma \in \mathbb{R}^{d\times d}$ as per (2).

Then $V$ has differential entropy

$$h(V) = \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{qd}{2} + \frac{d+1}{2}\ln\left|\frac{\Sigma}{2}\right| - \frac{q + d + 1}{2}\sum_{i=1}^{d}\psi\left(\frac{q - d + i}{2}\right) \tag{17}$$

The relative entropy between two inverse Wishart distributions with parameters $\Sigma_1, q_1$ and $\Sigma_2, q_2$ is

$$\ln\left(\frac{\Gamma_d\left(\frac{q_2}{2}\right)}{\Gamma_d\left(\frac{q_1}{2}\right)}\right) + \frac{q_1}{2}\mathrm{tr}(\Sigma_1^{-1}\Sigma_2) - \frac{q_1 d}{2} - \frac{q_2}{2}\ln|\Sigma_1^{-1}\Sigma_2| - \frac{q_2 - q_1}{2}\sum_{i=1}^{d}\psi\left(\frac{q_1 - d + i}{2}\right) \tag{18}$$

One sees that the relative entropy between two inverse Wishart distributions is the same as the relative entropy between two Wishart distributions with inverse matrix parameters $S_1^{-1}$ and $S_2^{-1}$ respectively. Like the Wishart distribution relative entropy, the inverse Wishart distribution relative entropy has special cases that are the Stein loss and the log-gamma Bregman divergence.

### 6.3. *Bayesian Estimation of Wishart Differential Entropy*

We present a Bayesian estimate of the differential entropy of a Wishart distribution $p$ where we make the simplifying assumption that the scalar parameter $q$ is known or estimated (for example, it is common to assume that $q = d$). We estimate the differential entropy $E_\Sigma[h(p)]$ where the estimation is with respect

to the uncertainty in the matrix parameter $\Sigma$. We assume the prior $p(\Sigma = \tilde{\Sigma})$ is inverse Wishart with scale parameter $r$ and parameter matrix $U$, which reduces to the non-informative prior when $r$ and $U$ are chosen to be zeros.

Then given sample $d \times d$ matrices $S_1, S_2, \ldots, S_n$ drawn iid from the Wishart $W$, the normalized posterior distribution $p(\tilde{\Sigma}|S_1, S_2, \ldots, S_n)$ is inverse Wishart with matrix parameter $\sum_{j=1}^{n} S_j + U$ and scalar parameter $nq + r$ (details in Appendix).

Then our differential entropy estimate $E_\Sigma[h(W)]$ where the expectation is with respect to the posterior $p(\tilde{\Sigma}|\{S_j\})$ is:

$$\ln \Gamma_d \left(\frac{q}{2}\right) + \frac{qd}{2} + \frac{d+1}{2} \ln \left| U + \sum_{j=1}^{n} S_j \right| - \frac{d+1}{2} \sum_{i=1}^{d} \psi \left( \frac{nq + r - d + i}{2} \right)$$
$$- \frac{q-d-1}{2} \sum_{i=1}^{d} \psi \left( \frac{q-d+i}{2} \right) \tag{19}$$

### *6.4. Bayesian Estimation of Relative Entropy between Two Wisharts*

We present a Bayesian estimate of the relative entropy between two Wishart distributions $p_1$ and $p_2$ where we make the simplifying assumption that the respective scalar parameters $q_1, q_2$ are known or estimated (for example, that $q_1 = q_2 = d$), and then we estimate the relative entropy $\mathrm{KL}(p_1||p_2)$ where the estimation is with respect to the uncertainty in the respective matrix parameters $\Sigma_1, \Sigma_2$. To this end, we treat the unknown Wishart parameters $\Sigma_1, \Sigma_2$ as random, and compute the estimate $E_{\Sigma_1,\Sigma_2}[\mathrm{KL}(p_1||p_2)]$. For $\Sigma_1$ and $\Sigma_2$ we use independent inverse Wishart conjugate priors with respective scalar parameters $r_1, r_2$ and parameter matrices $U_1, U_2$, which reduces to non-informative priors when $r_1, r_2$ and $U_1, U_2$ are chosen to be zeros.

Then given $n_1$ sample $d \times d$ matrices $\{S_j\}$ drawn iid from the Wishart $p_1$, and $n_2$ sample $d \times d$ matrices $\{S_k\}$ drawn iid from the Wishart $p_2$, the normalized posterior distribution $p(\tilde{\Sigma}_1|\{S_j\})$ is inverse Wishart with matrix parameter $\sum_{j=1}^{n_1} S_j + U_1$ and scalar parameter $n_1 q + r_1$, and the normalized posterior distribution $p(\tilde{\Sigma}_2|\{S_k\})$ is inverse Wishart with matrix parameter $\sum_{k=1}^{n_2} S_k + U_2$ and scalar parameter $n_2 q + r_2$.

Then our relative entropy estimate $E_{\Sigma_1,\Sigma_2}[\mathrm{KL}(p_1||p_2)]$ (where the expectation is with respect to the posterior distributions) is

$$\ln \left( \frac{\Gamma_d\left(\frac{q_2}{2}\right)}{\Gamma_d\left(\frac{q_1}{2}\right)} \right) - \frac{q_2 - q_1}{2} \sum_{i=1}^{d} \psi \left( \frac{q_1 - d + i}{2} \right) - \frac{q_1 d}{2}$$
$$+ \frac{q_1(r_1 + n_1 q_1)}{2(r_2 + n_2 q_2 - d - 1)} \mathrm{tr} \left( (U_1 + \sum_{j=1}^{n_1} S_j)(U_2 + \sum_{k=1}^{n_2} S_k)^{-1} \right) - \frac{q_2}{2} \ln |U_1 + \sum_{j=1}^{n_1} S_j|$$
$$+ \frac{q_2}{2} \ln |U_2 + \sum_{k=1}^{n_2} S_k| - \frac{q_2}{2} \sum_{i=1}^{d} \left( \psi \left( \frac{n_2 q_2 + r_2 - d + i}{2} \right) - \psi \left( \frac{n_1 q_1 + r_1 - d + i}{2} \right) \right)$$

### *6.5. Bayesian Estimation of Inverse Wishart Differential Entropy*

Let $S_i$ denote the $i$th of $n$ random iid draws from an inverse unknown Wishart distribution $p$ with parameters $(\Sigma, q)$. Taking the prior $p(\tilde{\Sigma})$ to be a Wishart distribution with parameter $r$ and $U$, our Bayesian estimate of the inverse Wishart differential entropy is

$$\ln \Gamma_d \left(\frac{q}{2}\right) + \frac{qd}{2} + \frac{d+1}{2} \ln |U^{-1} + \sum_j S_j^{-1}| + \frac{d+1}{2} \sum_{i=1}^{d} \psi \left(\frac{nq + r - d + i}{2}\right) - \frac{q+d+1}{2} \sum_{i=1}^{d} \psi \left(\frac{q-d+i}{2}\right) \tag{20}$$

### *6.6. Bayesian Estimation of Relative Entropy between Two Inverse Wisharts*

Given $q_1, q_2$, and assuming independent Wishart priors with respective scalar parameters $r_1, r_2$ and parameter matrices $U_1, U_2$, and given $n_1$ sample $d \times d$ matrices $\{S_j\}$ drawn iid from the inverse Wishart $p_1$, and $n_2$ sample $d \times d$ matrices $\{S_k\}$ drawn iid from the inverse Wishart $p_2$, our Bayesian estimate of the relative entropy is

$$\ln \left(\frac{\Gamma_d \left(\frac{q_2}{2}\right)}{\Gamma_d \left(\frac{q_1}{2}\right)}\right) + \frac{q_1}{2} \frac{n_2 q_2 + r_2}{n_1 q_1 + r_1 - d - 1} \text{tr} \left( \left( U_1^{-1} + \sum_{j=1}^{n_1} S_j^{-1} \right)^{-1} \left( U_2^{-1} + \sum_{k=1}^{n_2} S_k^{-1} \right) \right) - \frac{q_1 d}{2}$$
$$- \frac{q_2}{2} \ln \left( \frac{|U_2^{-1} + \sum_{k=1}^{n_2} S_k^{-1}|}{|U_1^{-1} + \sum_{j=1}^{n_1} S_j^{-1}|} \right) - \frac{q_2 - q_1}{2} \sum_{i=1}^{d} \psi \left( \frac{q_1 - d + i}{2} \right)$$
$$- \frac{q_2}{2} \sum_{i=1}^{d} \left( \psi \left( \frac{n_2 q_2 + r_2 - d + i}{2} \right) - \psi \left( \frac{n_1 q_1 + r_1 - d + i}{2} \right) \right). \tag{21}$$

## 7. Discussion

We have presented Bayesian differential entropy and relative entropy estimates for four standard distributions, and in doing so illustrated techniques that could be used to derive such estimates for other distributions. For the uniform case with no prior, the given estimators perform significantly better than previous estimators, and this experimental evidence validates our approach. However given a prior over distributions, the performance will depend heavily on the accuracy of the prior, and a thorough experimental study would be useful to practitioners but was outside the scope of this investigation.

In practice, there may not be a priori information available to determine a prior, and an open question is how to design appropriate data-dependent priors for differential entropy estimation. For example, for Bayesian quadratic discriminant analysis classification [27], we have shown that setting the prior matrix parameter for the Gaussian to be a coarse estimate of the data's covariance (the identity times the trace of the sample covariance) works well. However, for differential entropy estimation the trace forms an extreme estimate, and is thus not (by itself) suitable for forming a data-dependent prior for this problem.

Another open area is forming estimators for more complicated parametric models, for example estimating the differential entropy and relative entropy of Gaussian mixture models. Estimating the

differential entropy of Gaussian processes is also an important problem [29] that may be amenable to the present approach.

Lastly, the new estimators have been motivated by their expected Bregman loss optimality and by the practical consideration of producing estimates even when there are fewer samples than dimensions, but there are a number of theoretical questions about these estimators that are open, such as domination.

**Acknowledgments**

We would like to thank the United States Office of Naval Research for funding this research.

## A. Appendix

### *A.1. Proof of Proposition 1*

The proof is by contradiction. Let $\xi^* = E_{\mathcal{A}}[\xi(\mathcal{A})]$, and assume the true minimizer of $E_{\mathcal{A}}\left[d_\phi(\xi(\mathcal{A}), \hat{\xi})\right]$ occurs instead at some other value $s$. Then a contradiction occurs:

$$\begin{aligned}
&E_{\mathcal{A}}\left[d_\phi(\xi(\mathcal{A}), s)\right] - E_{\mathcal{A}}\left[d_\phi(\xi(\mathcal{A}), \xi^*)\right] \\
&\qquad \overset{(a)}{=} \phi(\xi^*) - \phi(s) - \frac{d\phi(s)}{ds}(E_{\mathcal{A}}[\xi(\mathcal{A})] - s) + \frac{d\phi(\xi^*)}{d\xi^*}(E_{\mathcal{A}}[\xi(\mathcal{A})] - \xi^*) \\
&\qquad \overset{(b)}{=} \phi(\xi^*) - \phi(s) - \frac{d\phi(s)}{ds}(E_{\mathcal{A}}[\xi(\mathcal{A})] - s) \\
&\qquad \overset{(c)}{=} d_\phi(\xi^*, s) \\
&\qquad \overset{(d)}{\geq} 0
\end{aligned}$$

where in $(a)$ we expanded $d_\phi$ and simplified, in $(b)$ we used the fact that $\xi^* = E_{\mathcal{A}}[\xi(\mathcal{A})]$, in $(c)$ we substituted $\xi^* = E_{\mathcal{A}}[\xi(\mathcal{A})]$ and used the definition of the Bregman divergence, and in $(d)$ we used the non-negativity of the Bregman divergence. Thus $\xi^* = E_{\mathcal{A}}[\xi(\mathcal{A})]$ must be the minimizer.

### *A.2. Derivation of Uniform Differential Entropy Estimate*

In this section we will repeatedly use the integral:

$$\int \frac{\ln u}{u^m} du = -\frac{\ln u}{(m-1)u^{m-1}} - \frac{1}{(m-1)^2 u^{m-1}} \tag{22}$$

To estimate the differential entropy of a multidimensional uniform distribution one only needs to consider the differential entropy for a one-dimensional uniform, because a multidimensional uniform can be written as a product of independent univariate distributions, and thus the differential entropy of the multidimensional uniform is the sum of the univariate entropies.

Thus we model the $n$ samples $\{x_1, x_2, \ldots, x_n\}$ as being drawn from a random one-dimensional uniform distribution $U$. Let $M$ be the two-dimensional statistical manifold composed of uniform distributions $\{\tilde{U}_{a,b}\}$, where $\tilde{U}_{a,b}$ has support on $[a, b]$ for $b > a$, $a, b, \in \mathbb{R}$. The measure should depend on the length $\delta = b - a$ of the uniform and be invariant to shifts in the support. To that end, we use the

Fisher information metric [25,26] based on the length,

$$dM = |I(\delta)|^{1/2} d\delta = \frac{d\delta}{\delta}$$

where $I$ is the Fisher information matrix,

$$I(\delta) = -E_X \left[ \frac{d^2 \log \frac{1}{\delta}}{d\delta^2} \right] = -\frac{1}{\delta^2}$$

Using $dM$ as a differential element and the normalized likelihood of the samples for $p(\tilde{U}_{a,b})$, the uniform differential entropy estimate is

$$\begin{aligned} E_U[h(U)] &= \frac{\int_M h(\tilde{U}_{a,b}) p(\tilde{U}_{a,b}) dM}{\int_M p(\tilde{U}_{a,b}) dM} \\ &= \frac{1}{\gamma} \int_{a=-\infty}^{x_{\min}} \int_{b=x_{\max}}^{\infty} \frac{\ln(b-a)}{(b-a)^n} \frac{da\, db}{(b-a)} \\ &= \frac{1}{\gamma n (n-1)(x_{\max} - x_{\min})^{n-1}} \left( \ln(x_{\max} - x_{\min}) + \frac{1}{n-1} + \frac{1}{n} \right), \end{aligned} \tag{23}$$

where the normalization factor $\gamma$ is

$$\begin{aligned} \gamma &= \int_{a=-\infty}^{x_{\min}} \int_{b=x_{\max}}^{\infty} \frac{1}{(b-a)^n} \frac{da\, db}{(b-a)} \\ &= \frac{1}{(n-1) n (x_{\max} - x_{\min})^{n-1}} \end{aligned}$$

Canceling terms in (23) due to the normalization factor $\gamma$ yields the one-dimensional uniform differential entropy $\ln(x_{\max} - x_{\min}) + \frac{1}{n-1} + \frac{1}{n}$. For the multidimensional uniform, one sums these marginal entropy terms over the dimensions, as given in (9).

### *A.3. Derivation of Uniform differential Entropy Given Pareto Prior*

As explained for the no-prior derivation, we need only consider a one-dimensional uniform. Although the Pareto distribution is a conjugate prior for the uniform with respect to its length, one must be careful because the data restrict $b > x_{\max}$ and $a < x_{\min}$, and these restrictions are not taken into account if one integrates with respect to the variable $\delta$. Throughout this section we use various flavors of $\gamma$ to denote normalization constants, and $\Delta = x_{\max} - x_{\min}$. We consider two cases separately.

Case I: $\ell \leq \Delta$:

$$\begin{aligned} p(\tilde{U}_{a,b} | \{x_i\}) &= \frac{1}{\gamma_1} p(\{x_i\} | \tilde{U}_{a,b}) p(\tilde{U}_{a,b}) \\ &= \begin{cases} \frac{1}{\gamma_1} \frac{\alpha \ell^\alpha}{(b-a)^{n+\alpha+1}} & \text{for } a \leq x_{\min}, b \geq x_{\max}, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{24}$$

where the normalizer is

$$\begin{aligned} \gamma_1 &= \int_{a=-\infty}^{x_{\min}} \int_{b=x_{\max}}^{\infty} p(\{x_i\} | \tilde{U}_{a,b}) p(\tilde{U}_{a,b}) \frac{da db}{b-a} = \int_{a=-\infty}^{x_{\min}} \int_{b=x_{\max}}^{\infty} \frac{\alpha \ell^\alpha}{(b-a)^{n+\alpha+1}} \frac{da db}{b-a} \\ &= \frac{\alpha \ell^\alpha}{(n+\alpha+1)(n+\alpha) \Delta^{n+\alpha}}. \end{aligned}$$

Then the posterior (24) becomes,

$$p(\tilde{U}_{a,b}|\{x_i\}) = \begin{cases} \frac{(n+\alpha)(n+\alpha+1)\Delta^{n+\alpha}}{(b-a)^{n+\alpha+1}} & \text{for } a \le x_{\min}, b \ge x_{\max}, \\ 0 & \text{otherwise.} \end{cases}$$

Using (22), it is straightforward to derive the differential entropy estimate given in the text as:

$$\begin{aligned} E_U[h(U)] &= (n+\alpha)(n+\alpha+1)\Delta^{n+\alpha}\int_{a=-\infty}^{x_{\min}}\int_{b=x_{\max}}^{\infty}\frac{\ln(b-a)}{(b-a)^{n+\alpha+1}}\frac{dadb}{b-a} \\ &= \ln\Delta + \frac{1}{n+\alpha} + \frac{1}{n+\alpha+1}. \end{aligned}$$

Case II: $\ell > \Delta$:

In this case, the posterior has an additional constraint compared to (24):

$$\begin{aligned} p(\tilde{U}_{a,b}|\{x_i\}) &= \frac{1}{\gamma_2}p(\{x_i\}|\tilde{U}_{a,b})p(\tilde{U}_{a,b}) \\ &= \begin{cases} \frac{1}{\gamma_2}\frac{\alpha\ell^\alpha}{(b-a)^{n+\alpha+1}} & \text{for } a \le x_{\min}, b \ge x_{\max}, \text{ and } b-a \ge \ell \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The normalization constant can be solved for as:

$$\begin{aligned} \gamma_2 &= \int_{a=-\infty}^{x_{\min}}\int_{b=x_{\max}}^{\infty} p(\{x_i\}|\tilde{U}_{a,b})p(\tilde{U}_{a,b})\frac{dadb}{b-a} \\ &= \int_{a=-\infty}^{x_{\max}-\ell}\int_{b=x_{\max}}^{\infty}\frac{\alpha\ell^\alpha}{(b-a)^{n+\alpha+1}}\frac{dadb}{b-a} + \int_{a=x_{\max}-\ell}^{x_{\min}}\int_{b=a+\ell}^{\infty}\frac{\alpha\ell^\alpha}{(b-a)^{n+\alpha+1}}\frac{dadb}{b-a} \\ &= \frac{\alpha\ell^\alpha}{(n+\alpha+1)(n+\alpha)\ell^{n+\alpha}} + \frac{\alpha\ell^\alpha(\ell-\Delta)}{(n+\alpha+1)\ell^{n+\alpha+1}} \\ &= \left(\frac{\alpha\ell^\alpha}{(n+\alpha+1)\ell^{n+\alpha}}\right)\left(\frac{1}{n+\alpha} + \frac{\ell-\Delta}{\ell}\right) \end{aligned} \tag{25}$$

Then the differential entropy estimate is

$$\begin{aligned} E_U[h(U)] &= \frac{\alpha\ell^\alpha}{\gamma_2}\left(\int_{a=-\infty}^{x_{\max}-\ell}\int_{b=x_{\max}}^{\infty}\frac{\ln(b-a)}{(b-a)^{n+\alpha+1}}\frac{dadb}{b-a} + \int_{a=x_{\max}-\ell}^{x_{\min}}\int_{b=a+\ell}^{\infty}\frac{\ln(b-a)}{(b-a)^{n+\alpha+1}}\frac{dadb}{b-a}\right) \\ &= \frac{\alpha\ell^\alpha}{\gamma_2}\left(\frac{\ln\ell}{(n+\alpha+1)(n+\alpha)\ell^{n+\alpha}} + \frac{1}{(n+\alpha+1)(n+\alpha)^2\ell^{n+\alpha}} + \frac{1}{(n+\alpha)(n+\alpha+1)^2\ell^{n+\alpha}}\right) \\ &+ \frac{\alpha\ell^\alpha}{\gamma_2}\left(\frac{(\ell-\Delta)\ln\ell}{(n+\alpha+1)\ell^{n+\alpha+1}} + \frac{\ell-\Delta}{(n+\alpha+1)^2\ell^{n+\alpha+1}}\right) \\ &= \left(\frac{\alpha\ell^\alpha}{\gamma_2(n+\alpha+1)\ell^{n+\alpha}}\right)\left(\frac{\ln\ell}{(n+\alpha)} + \frac{1}{(n+\alpha)^2} + \frac{1}{(n+\alpha)(n+\alpha+1)}\right. \\ &\left. + \frac{(\ell-\Delta)\ln\ell}{\ell} + \frac{\ell-\Delta}{(n+\alpha+1)\ell}\right) \\ &\overset{(a)}{=} \left(\frac{\ell(n+\alpha)}{\ell+(n+\alpha)(\ell-\Delta)}\right) \\ &\cdot\left(\frac{\ln\ell}{(n+\alpha)} + \frac{1}{(n+\alpha)^2} + \frac{1}{(n+\alpha)(n+\alpha+1)} + \frac{(\ell-\Delta)\ln\ell}{\ell} + \frac{\ell-\Delta}{(n+\alpha+1)\ell}\right) \end{aligned} \tag{26}$$

where in $(a)$ we substituted in (25). In the second factor of $(a)$ there are five terms. Combining the first and fourth term with the first factor results in the first term of the estimate given in (11). Combining the second term with the first factor results in the second term of (11). Lastly, combining the third and fifth term of $(a)$ with the first factor results in the third term of (11).

### *A.4. Propositions Used in Remaining Derivations*

The following identities and propositions will be used repeatedly in the derivations in the rest of the appendix.

**Identity 1.** This is a convenient re-statement of the fact that the normal distribution normalizes to one. For $x, \mu \in \mathbb{R}^d$ and positive definite $d \times d$ matrix $\Sigma$,

$$\int_{\mu} e^{-\frac{n}{2}\mathrm{tr}\left(\Sigma^{-1}(\mathrm{x}-\mu)(\mathrm{x}-\mu)^{\mathrm{T}}\right)} d\mu = \left(\frac{2\pi}{n}\right)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}$$

**Identity 2.** This is a convenient re-statement of the fact that the inverse Wishart distribution normalizes to one. For positive definite $\Sigma$:

$$\int_{\Sigma>0} \frac{e^{-\mathrm{tr}\left(\Sigma^{-1}\mathrm{B}\right)}}{|\Sigma|^{\frac{q}{2}}} d\Sigma = \frac{\Gamma_d\left(\frac{q-d-1}{2}\right)}{|B|^{\frac{q-d-1}{2}}}$$

**Proposition 2.**

For $W \sim$ Wishart$(S, q)$,

$$E[\ln|W|] = \ln|S| + d\ln 2 + \sum_{i=1}^{d} \psi\left(\frac{q-d+i}{2}\right) \equiv \ln|2S| + \sum_{i=1}^{d} \psi\left(\frac{q-d+i}{2}\right)$$

Proof: Recall that $|W|$ is distributed as $|S|\prod_{i=1}^{d}\chi^2_{q-d+i}$ ([30, Corollary 7.3]) where $\chi^2$ denotes the chi-squared random variable. Then the result is produced by taking the expected log and using the fact that $E[\ln\chi^2_q] = \ln 2 + \psi\left(\frac{q}{2}\right)$ [4]. Lastly the equivalence follows because $\ln|2S| = \ln 2^d|S| = d\ln 2 + \ln|S|$.

**Proposition 3.**

For $V \sim$ inverse Wishart$(S, q)$,

$$E[\ln|V|] = \ln|S| - d\ln 2 - \sum_{i=1}^{d} \psi\left(\frac{q-d+i}{2}\right) \equiv \ln\left|\frac{S}{2}\right| - \sum_{i=1}^{d} \psi\left(\frac{q-d+i}{2}\right)$$

Proof: Let $Z = V^{-1}$, then $Z \sim$ Wishart$(S^{-1}, q)$, and $E[\ln|V|] = E[\ln|Z|^{-1}] = -E[\ln|Z|] = -\ln|S^{-1}| - d\ln 2 - \sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right)$, by Prop. 2, and noting that $-\ln|S^{-1}| = \ln|S|$ produces the result. Lastly, the equivalence follows because $\ln|\frac{S}{2}| = \ln\frac{1}{2^d}|S| = -d\ln 2 + \ln|S|$.

**Proposition 4.**

For $W \sim \text{Wishart}(S, q)$ and any positive definite matrix $A \in \mathbb{R}^{d\times d}$,

$$E[\text{tr}(WA)] = q\text{tr}(SA)$$

Proof: $E[\text{tr}(WA)] = \text{tr}(E[W]A) = q\text{tr}(SA)$.

**Proposition 5.**

For $V \sim$ inverse Wishart $(S, q)$ and any positive definite matrix $A \in \mathbb{R}^{d\times d}$,

$$E[\text{tr}(VA)] = \frac{\text{tr}(AS)}{q-d-1}$$

Proof: $E[\text{tr}(VA)] = \text{tr}(E[V]A) = \text{tr}(AS)/(q-d-1)$.

**Proposition 6.**

For $V \sim$ inverse Wishart $(S, q)$ and any positive definite matrix $A \in \mathbb{R}^{d\times d}$,

$$E[\text{tr}(V^{-1}A)] = q\text{tr}(S^{-1}A)$$

Proof: By definition, $V^{-1} \sim$ Wishart $(S^{-1}, q)$, and so one can apply Prop. 4 to yield $E[\text{tr}(V^{-1}A)] = q\text{tr}(S^{-1}A)$.

### *A.5. Derivation of Bayesian Gaussian Differential Entropy Estimate*

We model the samples $\{x_i\}_{i=1}^n$ as being drawn from a random $d$-dimensional normal distribution $N$ and assume an inverse Wishart prior distribution for $N$ with parameters $(B, q)$. That is, the prior probability that the random normal $N$ is $\tilde{N}(\tilde{\mu}, \tilde{\Sigma})$ is

$$p(N = \tilde{N}(\tilde{\mu}, \tilde{\Sigma})) = \frac{|B|^{\frac{q}{2}} e^{-\frac{1}{2}\text{tr}\left(\tilde{\Sigma}^{-1}\text{B}\right)}}{2^{\frac{qd}{2}}\Gamma\left(\frac{q}{2}\right)|\tilde{\Sigma}|^{\frac{q+d+1}{2}}} \tag{27}$$

The likelihood can be written:

$$\begin{aligned} p(\{x_i\}_{i=1}^n | N = \tilde{N}(\tilde{\mu}, \tilde{\Sigma})) &= \frac{1}{(2\pi)^{\frac{nd}{2}}|\tilde{\Sigma}|^{\frac{n}{2}}} \prod_{i=1}^{n} e^{-\frac{1}{2}\text{tr}\left(\tilde{\Sigma}^{-1}(\text{x}_\text{i}-\tilde{\mu})(\text{x}_\text{i}-\tilde{\mu})^{\text{T}}\right)} \\ &= \frac{1}{(2\pi)^{\frac{nd}{2}}|\tilde{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}\text{tr}\left(\tilde{\Sigma}^{-1}(\text{x}_\text{i}-\bar{\text{x}}+\bar{\text{x}}-\tilde{\mu})(\text{x}_\text{i}-\bar{\text{x}}+\bar{\text{x}}-\tilde{\mu})^{\text{T}}\right)} \\ &= \frac{1}{(2\pi)^{\frac{nd}{2}}|\tilde{\Sigma}|^{\frac{n}{2}}} e^{-\frac{1}{2}\text{tr}\left(\tilde{\Sigma}^{-1}\text{S}\right)-\frac{\text{n}}{2}\text{tr}\left(\tilde{\Sigma}^{-1}(\bar{\text{x}}-\tilde{\mu})(\bar{\text{x}}-\tilde{\mu})^{\text{T}}\right)}. \end{aligned} \tag{28}$$

Then the posterior is the likelihood times the prior normalized, or sweeping all the constant terms into a normalization term $\alpha$ we can write the posterior as:

$$p(N=\tilde{N}(\tilde{\mu},\tilde{\Sigma})|\{x_i\}_{i=1}^n) = \frac{1}{\alpha}\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}\mathrm{S}\right)-\frac{\mathrm{n}}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\bar{\mathrm{x}}-\tilde{\mu})(\bar{\mathrm{x}}-\tilde{\mu})^{\mathrm{T}}\right)}}{|\tilde{\Sigma}|^{\frac{n}{2}}}\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}\mathrm{B}\right)}}{|\tilde{\Sigma}|^{\frac{q+d+1}{2}}} \tag{29}$$

Note that this is a density on the statistical manifold of Gaussians, so we integrate with respect to the Fisher information measure $1/|\tilde{\Sigma}|^{\frac{d+2}{2}}$ [27],[25] rather than the Lebesgue measure, such that

$$\begin{aligned}\alpha &= \int_{\tilde{\Sigma}}\int_{\tilde{\mu}} \frac{e^{-\frac{n}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\bar{\mathrm{x}}-\tilde{\mu})(\bar{\mathrm{x}}-\tilde{\mu})^{\mathrm{T}}\right)}}{|\tilde{\Sigma}|^{\frac{n+q+d+1}{2}}} e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)}\frac{d\tilde{\Sigma}d\tilde{\mu}}{|\tilde{\Sigma}|^{\frac{d+2}{2}}} \\ &= \left(\frac{2\pi}{n}\right)^{\frac{d}{2}}\frac{\Gamma_d\left(\frac{n+q+d+1}{2}\right)}{|S+B|^{\frac{n+q+d+1}{2}}}2^{\frac{d(n+q+d+1)}{2}},\end{aligned} \tag{30}$$

where the last line follows from Identities 1 and 2 stated in the previous subsection.

Then combining (30) and (29), the posterior is:

$$p(N=\tilde{N}(\tilde{\mu},\tilde{\Sigma})|\{x_i\}_{i=1}^n) = \left(\frac{n}{2\pi}\right)^{\frac{d}{2}}\frac{|S+B|^{\frac{n+q+d+1}{2}}}{\Gamma_d\left(\frac{n+q+d+1}{2}\right)}\frac{1}{2^{\frac{d(n+q+d+1)}{2}}}\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)-\frac{\mathrm{n}}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\bar{\mathrm{x}}-\tilde{\mu})(\bar{\mathrm{x}}-\tilde{\mu})^{\mathrm{T}}\right)}}{|\tilde{\Sigma}|^{\frac{n+q+d+1}{2}}} \tag{31}$$

Our differential entropy estimate is the integral $E_N[h(N)]$, which is an integral over the statistical manifold of Gaussians that we convert to an integral over covariance matrices by using the Fisher information metric $1/|\tilde{\Sigma}|^{(d+2)/2}$ [27],[25]. Then,

$$\begin{aligned}E_N[h(N)] &= \int_{\tilde{N}(\tilde{\mu},\tilde{\Sigma})}\left(\frac{d}{2}+\frac{d\ln(2\pi)}{2}+\frac{\ln|\tilde{\Sigma}|}{2}\right)p(\tilde{N}(\tilde{\mu},\tilde{\Sigma})|\{x_i\})d\tilde{N} \\ &= \frac{d}{2}+\frac{d\ln(2\pi)}{2} \\ &+ \left(\frac{n}{2\pi}\right)^{\frac{d}{2}}\frac{|S+B|^{\frac{n+q+d+1}{2}}}{2^{\frac{2+d(n+q+d+1)}{2}}\Gamma_d\left(\frac{n+q+d+1}{2}\right)}\int_{\tilde{\Sigma}>0}\int_{\tilde{\mu}}\ln|\tilde{\Sigma}|\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)-\frac{\mathrm{n}}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\bar{\mathrm{x}}-\tilde{\mu})(\bar{\mathrm{x}}-\tilde{\mu})^{\mathrm{T}}\right)}}{|\tilde{\Sigma}|^{\frac{n+q+d+1}{2}}}\frac{d\tilde{\Sigma}d\tilde{\mu}}{|\tilde{\Sigma}|^{\frac{d+2}{2}}}.\end{aligned} \tag{32}$$

We evaluate the third term of (32) as follows:

$$\left(\frac{n}{2\pi}\right)^{\frac{d}{2}}\frac{|S+B|^{\frac{n+q+d+1}{2}}}{2^{\frac{2+d(n+q+d+1)}{2}}\Gamma_d\left(\frac{n+q+d+1}{2}\right)}\int_{\tilde{\Sigma}>0}\ln|\tilde{\Sigma}|\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)}}{|\tilde{\Sigma}|^{\frac{n+q+2d+3}{2}}}\int_{\tilde{\mu}}e^{-\frac{n}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\bar{\mathrm{x}}-\tilde{\mu})(\bar{\mathrm{x}}-\tilde{\mu})^{\mathrm{T}}\right)}d\tilde{\mu}d\tilde{\Sigma}$$

$$= \left(\frac{n}{2\pi}\right)^{\frac{d}{2}}\frac{|S+B|^{\frac{n+q+d+1}{2}}}{2^{\frac{2+d(n+q+d+1)}{2}}\Gamma_d\left(\frac{n+q+d+1}{2}\right)}\int_{\tilde{\Sigma}>0}\ln|\tilde{\Sigma}|\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)}}{|\tilde{\Sigma}|^{\frac{n+q+2d+3}{2}}}\left(\frac{2\pi}{n}\right)^{\frac{d}{2}}|\tilde{\Sigma}|^{\frac{1}{2}}d\tilde{\Sigma} \tag{33}$$

$$= \frac{|S+B|^{\frac{n+q+d+1}{2}}}{2^{\frac{2+d(n+q+d+1)}{2}}\Gamma_d\left(\frac{n+q+d+1}{2}\right)}\int_{\tilde{\Sigma}>0}\ln|\tilde{\Sigma}|\frac{e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}^{-1}(\mathrm{S}+\mathrm{B})\right)}}{|\tilde{\Sigma}|^{\frac{n+q+2d+2}{2}}}d\tilde{\Sigma} \tag{34}$$

$$= \frac{1}{2}\ln|S+B| - \frac{d}{2}\ln 2 - \frac{1}{2}\sum_{i=1}^{d}\psi\left(\frac{n+q+1+i}{2}\right), \tag{35}$$

where (33) follows by Integral Identity 1; and (34) is half the expectation of $\ln|\Sigma|$ with respect to the inverse Wishart with parameters $(S+B, n+q+d+1)$, and thus (35) follows from (34) by Prop. 3.

Then (32) becomes

$$\frac{d}{2} + \frac{d \ln \pi}{2} + \frac{1}{2} \ln |S + B| - \frac{1}{2} \sum_{i=1}^{d} \psi \left( \frac{n + q + 1 + i}{2} \right) \tag{36}$$

### *A.6. Derivation of Bayesian Gaussian Relative Entropy Estimate*

Recall that the relative entropy between independent Gaussians $\mathcal{N}_1(x; \mu_1, \Sigma_1)$ and $\mathcal{N}_2(x; \mu_2, \Sigma_2)$ is

$$\text{KL}(\mathcal{N}_1 \| \mathcal{N}_2) = \frac{1}{2} \left( \text{tr}(\Sigma_1 \Sigma_2^{-1}) - \log |\Sigma_1 \Sigma_2^{-1}| - d + \text{tr} \left( \Sigma_2^{-1} (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \right) \right) \tag{37}$$

Here we derive $E_{N_1,N_2}[\text{KL}(N_1 \| N_2)]$. Analogous to the previous derivation of the Bayesian Gaussian entropy estimate, we form the posterior distributions using independent inverse Wishart priors with parameters $(B_1, q_1)$ and $(B_2, q_2)$. (Note that this is equivalent to having a non-informative prior on the mean parameter, and that a different prior on the mean would lead to a more regularized estimate). We consider the four terms of the expectation of (37), that is, of $E_{N_1,N_2}[\text{KL}(N_1 \| N_2)]$, in turn.

The first term is:

$$E_{N_1,N_2} \left[ \text{tr}(\Sigma_1 \Sigma_2^{-1}) \right] = \int \text{tr}(\Sigma_1 \Sigma_2^{-1}) p(\tilde{N}_1(\tilde{\mu}_1, \tilde{\Sigma}_1) | \{x_{1,i}\}_{i=1}^{n_1}) p(\tilde{N}_2(\tilde{\mu}_2, \tilde{\Sigma}_2) | \{x_{2,i}\}_{i=1}^{n_2}) d\tilde{N}_1 d\tilde{N}_2 \tag{38}$$

where the posteriors are given by (31). Using the Fisher information measure as above, (38) can be re-written as expectations of functions of independent random covariance matrices $\Sigma_1, \Sigma_2$ drawn from inverse Wishart distributions with respective parameters $(S_1 + B_1, n_1 + q_1 + d + 1)$ and $(S_2 + B_2, n_2 + q_2 + d + 1)$:

$$\begin{aligned} E_{\Sigma_1,\Sigma_2} \left[ \text{tr}(\Sigma_1 \Sigma_2^{-1}) \right] &\overset{(a)}{=} \frac{1}{n_1 + q_1} E_{\Sigma_2} \left[ \text{tr}((S_1 + B_1) \Sigma_2^{-1}) \right], \\ &\overset{(b)}{=} \frac{n_2 + q_2 + d + 1}{n_1 + q_1} \text{tr}((S_1 + B_1)(S_2 + B_2)^{-1}), \end{aligned}$$

where $(a)$ is by Proposion 5, and $(b)$ is by Proposition 6.

Similarly, we can write the second term as:

$$\begin{aligned} E_{\Sigma_1,\Sigma_2} \left[ \log |\Sigma_1 \Sigma_2^{-1}| \right] &= E_{\Sigma_1,\Sigma_2} \left[ \log |\Sigma_1| - \log |\Sigma_2| \right] = E_{\Sigma_1} \left[ \log |\Sigma_1| \right] - E_{\Sigma_2} \left[ \log |\Sigma_2| \right] \\ &= \log \frac{|S_1 + B_1|}{|S_2 + B_2|} + \sum_{i=1}^{d} \psi \left( \frac{n_2 + q_2 + 1 + i}{2} \right) - \psi \left( \frac{n_1 + q_1 + 1 + i}{2} \right), \end{aligned}$$

where the last line follows from Proposition 4.

The third term is simply $-d$. The fourth term simplifies by Proposition 6 to:

$$E_{\Sigma_2}[\text{tr}(\Sigma_2^{-1}(\bar{\text{x}}_1 - \bar{\text{x}}_2)(\bar{\text{x}}_1 - \bar{\text{x}}_2)^{\text{T}})] = (\text{n}_2 + \text{q}_2 + \text{d} + 1)\text{tr}((\text{S}_2 + \text{B}_2)^{-1}(\bar{\text{x}}_1 - \bar{\text{x}}_2)(\bar{\text{x}}_1 - \bar{\text{x}}_2)^{\text{T}}).$$

Combining the terms yields the relative entropy estimate given in (13).

### *A.7. Derivation of Wishart Differential Entropy:*

Using the Wishart density given in (1), the Wishart differential entropy $h(W)$ is $-E[\ln p(W)]$,

$$
\begin{aligned}
&= \frac{qd}{2}\ln 2 + \frac{q}{2}\ln|\Sigma| + \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{1}{2}E[\mathrm{tr}(W\Sigma^{-1})] - \frac{q-d-1}{2}E[\ln|W|],\\
&\overset{(a)}{=} \frac{qd}{2}\ln 2 + \frac{q}{2}\ln|\Sigma| + \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{qd}{2} - \frac{q-d-1}{2}E[\ln|W|],\\
&\overset{(b)}{=} \frac{qd}{2}\ln 2 + \frac{q}{2}\ln|\Sigma| + \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{qd}{2} - \frac{q-d-1}{2}\left(\ln|\Sigma| + d\ln 2 + \sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right)\right),\\
&\overset{(c)}{=} \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{qd}{2} + \frac{d+1}{2}\ln|2\Sigma| - \frac{q-d-1}{2}\sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right),
\end{aligned}
$$

where $(a)$ follows by applying Proposition 4 to show that $E[\mathrm{tr}(W\Sigma^{-1})] = q\mathrm{tr}(\Sigma\Sigma^{-1}) = qd$, and then in $(b)$ one applies Prop. 2 to $E[\ln|W|]$ and recalls that $\ln|2\Sigma| = \ln|\Sigma| + d\ln 2$.

### *A.8. Derivation of Wishart Relative Differential Entropy:*

The relative entropy $\mathrm{KL}(p_1, p_2)$ is

$$
\begin{aligned}
&\triangleq E_{p_1}[\ln\frac{p_1(W)}{p_2(W)}]\\
&= -h(p_1) - E_{p_1}[\ln p_2(W)]\\
&\overset{(a)}{=} -\ln\Gamma_d\left(\frac{q_1}{2}\right) - \frac{q_1 d}{2} - \frac{d+1}{2}\ln|2\Sigma_1| + \frac{q_1-d-1}{2}\sum_{i=1}^{d}\psi\left(\frac{q_1-d+i}{2}\right)\\
&- \frac{q_2-d-1}{2}E_{p_1}[\ln|W|] + \frac{1}{2}E_{p_1}[\mathrm{tr}(W\Sigma_2^{-1})] + \frac{q_2}{2}\ln|2\Sigma_2| + \ln\Gamma_d\left(\frac{q_2}{2}\right)\\
&\overset{(b)}{=} \ln\left(\frac{\Gamma_d\left(\frac{q_2}{2}\right)}{\Gamma_d\left(\frac{q_1}{2}\right)}\right) + \frac{q_1}{2}\mathrm{tr}\left(\Sigma_1\Sigma_2^{-1}\right) - \frac{q_1 d}{2} - \frac{q_2}{2}\ln|\Sigma_1\Sigma_2^{-1}| - \left(\frac{q_2-q_1}{2}\right)\sum_{i=1}^{d}\psi\left(\frac{q_1-d+i}{2}\right)
\end{aligned}
$$

where $(a)$ uses the formula for entropy given in (14), and $(b)$ follows by applying Proposition 2 and 4 and then simplifying.

### *A.9. Derivation of Inverse Wishart Differential Entropy:*

Using the inverse Wishart density given in (2), the inverse Wishart differential entropy is:

$$
\begin{aligned}
h(V) &= -\frac{q}{2}\ln|S| + \frac{E[\mathrm{tr}(V^{-1}S)]}{2} + \frac{qd}{2}\ln 2 + \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{q+d+1}{2}E[\ln|V|]\\
&\overset{(a)}{=} -\frac{q}{2}\ln|S| + \frac{q\mathrm{tr}(S^{-1}S)}{2} + \frac{qd}{2}\ln 2 + \ln\Gamma_d\left(\frac{q}{2}\right) + \frac{q+d+1}{2}\ln|S| - d\ln 2 - \sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right)\\
&\overset{(b)}{=} \frac{d+1}{2}\ln\left|\frac{S}{2}\right| + \frac{qd}{2} + \ln\Gamma_d\left(\frac{q}{2}\right) - \frac{q+d+1}{2}\sum_{i=1}^{d}\psi\left(\frac{q-d+i}{2}\right)
\end{aligned}
$$

where in $(a)$ we applied Proposition 6 and Proposition 3, and in $(b)$ used $\text{tr}(S^{-1}S) = \text{tr}(I) = d$ and simplified.

### *A.10. Derivation of Inverse Wishart Relative Entropy:*

Taking the expectation with respect to the first inverse Wishart $V_1$ of the log of the ratio of the two inverse Wishart distributions yields

$$\begin{aligned}
&\frac{(q_2-q_1)}{2}d\ln 2 + \frac{E[\text{tr}(\Sigma_2-\Sigma_1)V_1^{-1}]}{2} - \frac{q_2}{2}\ln|\Sigma_2| + \frac{q_1}{2}\ln|\Sigma_1| + \ln\Gamma_d\left(\frac{q_2}{2}\right) \\
&- \ln\Gamma_d\left(\frac{q_1}{2}\right) + \frac{q_2-q_1}{2}E[\ln|V_1|]. \\
&\overset{(a)}{=}\frac{(q_2-q_1)}{2}d\ln 2 + \frac{q_1\text{tr}(\Sigma_2\Sigma_1^{-1})}{2} - \frac{q_1 d}{2} - \frac{q_2}{2}\ln|\Sigma_2| + \frac{q_1}{2}\ln|\Sigma_1| + \ln\Gamma_d\left(\frac{q_2}{2}\right) \\
&- \ln\Gamma_d\left(\frac{q_1}{2}\right) + \frac{q_2-q_1}{2}E[\ln|V_1|], \\
&\overset{(b)}{=}\frac{q_2-q_1}{2}d\ln 2 + \frac{q_1\text{tr}(\Sigma_2\Sigma_1^{-1})}{2} - \frac{q_1 d}{2} - \frac{q_2}{2}\ln|\Sigma_2| + \frac{q_1}{2}\ln|\Sigma_1| + \ln\Gamma_d\left(\frac{q_2}{2}\right) - \ln\Gamma_d\left(\frac{q_1}{2}\right) \\
&+ \frac{q_2-q_1}{2}\left(\ln|\Sigma_1| - d\ln 2 - \sum_{i=1}^{d}\psi\left(\frac{q_1-d+i}{2}\right)\right), \\
&= \ln\frac{\Gamma_d\left(\frac{q_2}{2}\right)}{\Gamma_d\left(\frac{q_1}{2}\right)} + \frac{q_1}{2}\text{tr}(\Sigma_2\Sigma_1^{-1}) - \frac{q_1 d}{2} + \frac{q_2}{2}\ln|\Sigma_1\Sigma_2^{-1}| - \frac{q_2-q_1}{2}\sum_{i=1}^{d}\psi\left(\frac{q_1-d+i}{2}\right),
\end{aligned}$$

where $(a)$ results from distributing the trace and applying Proposition 6 to each term and recalling that $\text{tr}I = d$, $(b)$ applies Proposition 3, and the last line is simplifications.

### *A.11. Derivation of Bayesian Estimate of Wishart Differential Entropy:*

Given sample $d \times d$ matrices $S_1, S_2, \ldots, S_n$ drawn iid from the unknown Wishart $W$ with unknown parameters $\Sigma, q$, the normalized posterior distribution $p(\Sigma = \tilde{\Sigma}|S_1, S_2, \ldots, S_n)$ is the normalized product of the inverse Wishart prior $p(\tilde{\Sigma})$ and the product of $n$ Wishart likelihoods $\prod_j p(S_j|\tilde{\Sigma})$. To derive the posterior, we take the product of the prior and likelihood and sweep all terms that do not depend on $\tilde{\Sigma}$ into a normalization constant $\gamma$:

$$\begin{aligned}
p(\tilde{\Sigma}|\{S_j\}) &= \gamma\left(\prod_{j=1}^{n}\left(\frac{e^{-\frac{1}{2}\text{tr}(\tilde{\Sigma}^{-1}S_j)}}{|\tilde{\Sigma}|^{\frac{q}{2}}}\right)\right)\left(\frac{e^{-\frac{1}{2}\text{tr}(\tilde{\Sigma}^{-1}U)}}{|\tilde{\Sigma}|^{\frac{r+d+1}{2}}}\right) \\
&= \gamma\frac{e^{-\frac{1}{2}\text{tr}\left(\tilde{\Sigma}^{-1}\left(U+\sum_{j=1}^{n}S_j\right)\right)}}{|\tilde{\Sigma}|^{\frac{nq+r+d+1}{2}}} \\
&= \frac{|\left(U+\sum_{j=1}^{n}S_j\right)|^{\frac{nq+r}{2}}e^{-\frac{1}{2}\text{tr}\left(\tilde{\Sigma}^{-1}\left(U+\sum_{j=1}^{n}S_j\right)\right)}}{2^{\frac{(nq+r)d}{2}}\Gamma_d\left(\frac{nq+r}{2}\right)|\tilde{\Sigma}|^{\frac{nq+r+d+1}{2}}}
\end{aligned} \tag{39}$$

where in (39) we solved for the normalization constant $\gamma$. One sees from (39) that the posterior $p(\tilde{\Sigma}|\{S_j\})$ is inverse Wishart with parameters $U + \sum_j S_j$ and $nq + r$.

Then the differential entropy estimate $E[h(W)]$ can be computed by taking the expectation of $h(W)$ given in (14) where the $W$ is treated as random and the expectation is with respect to the posterior given in (39). Only one term requires the expectation:

$$\begin{aligned} E[\ln|2\Sigma|] &= E[\ln|\Sigma|] + d\ln 2 \\ &= \ln|\sum_{j=1}^{n} S_j + U| - \sum_{i=1}^{d} \psi\left(\frac{nq + r - d + i}{2}\right) \end{aligned}$$

where $(b)$ applies Proposition 3. Substituting this term into the differential entropy formula (14) produces the differential entropy estimate (19).

### *A.12. Derivation of Bayesian Estimate of Relative Entropy Between Wisharts:*

There are only two terms of (15) that require evaluating the expectation, taken with respect to the independent posteriors of the form given in (39).

The first term is evaluated by applying Proposition 4 and Proposition 5 sequentially:

$$E_{\Sigma_1,\Sigma_2}\left[\text{tr}\left(\Sigma_1\Sigma_2^{-1}\right)\right] = \frac{(r_1 + n_1 q_1)}{(r_2 + n_2 q_2 - d - 1)}\text{tr}\left(\left(U_1 + \sum_{j=1}^{n_1} S_j\right)\left(U_2 + \sum_{k=1}^{n_2} S_k\right)^{-1}\right)$$

The second term follows by applying Proposition 3 twice:

$$\begin{aligned} E_{\Sigma_1,\Sigma_2}\left[\ln|\Sigma_1\Sigma_2^{-1}|\right] &= E_{\Sigma_1,\Sigma_2}\left[\ln\left(|\Sigma_1||\Sigma_2^{-1}|\right)\right] \\ &= E_{\Sigma_1}\left[\ln|\Sigma_1|\right] - E_{\Sigma_2}\left[\ln|\Sigma_2|\right] \\ &= \ln|U_1 + \sum_{j=1}^{n_1} S_j| - \ln|U_2 + \sum_{k=1}^{n_2} S_k| \\ &+ \sum_{i=1}^{d}\left(\psi\left(\frac{n_2 q_2 + r_2 - d + i}{2}\right) - \psi\left(\frac{n_1 q_1 + r_1 - d + i}{2}\right)\right). \end{aligned}$$

### *A.13. Derivation of Bayesian Estimate of Inverse Wishart Differential Entropy:*

Given sample $d \times d$ matrices $S_1, S_2, \ldots, S_n$ drawn iid from the unknown inverse Wishart $V$ with unknown parameters $\Sigma, q$, the normalized posterior distribution $p(\Sigma = \tilde{\Sigma}|S_1, S_2, \ldots, S_n)$ is the normalized product of the Wishart prior $p(\tilde{\Sigma})$ and the product of $n$ inverse Wishart likelihoods $\prod_j p(S_j|\tilde{\Sigma})$.

To derive the posterior, we take the product of the prior and likelihood and sweep all terms that do not

depend on $\tilde{\Sigma}$ into a normalization constant $\gamma$:

$$\begin{aligned} p(\tilde{\Sigma}|\{S_j\}) &= \gamma \left( \prod_{j=1}^{n} |\tilde{\Sigma}|^{\frac{q}{2}} e^{-\frac{1}{2}\mathrm{tr}(\tilde{\Sigma} S_j^{-1})} \right) \left( |\tilde{\Sigma}|^{\frac{r-d-1}{2}} e^{-\frac{1}{2}\mathrm{tr}(\tilde{\Sigma} U^{-1})} \right) \\ &= \gamma |\tilde{\Sigma}|^{\frac{nq+r-d-1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}\left(U^{-1}+\sum_{j=1}^{n} S_j^{-1}\right)\right)} \\ &= \frac{|\tilde{\Sigma}|^{\frac{nq+r-d-1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left(\tilde{\Sigma}\left(U^{-1}+\sum_{j=1}^{n} S_j^{-1}\right)\right)}}{2^{\frac{(nq+r)d}{2}} \Gamma_d\left(\frac{nq+r}{2}\right) |U + \sum_{j=1}^{n} S_j^{-1}|^{\frac{nq+r}{2}}} \end{aligned} \tag{40}$$

where in (40) we solved for the normalization constant $\gamma$. One sees from (40) that the posterior $p(\tilde{\Sigma}|\{S_j\})$ is Wishart with parameters $U^{-1} + \sum_j S_j^{-1}$ and $nq + r$.

Then the differential entropy estimate $E[h(V)]$ can be computed by taking the expectation of $h(V)$ given in (17) where the $V$ is treated as random and the expectation is with respect to the posterior given in (40). Only one term requires the expectation:

$$\begin{aligned} E\left[\ln |\frac{\Sigma}{2}|\right] &\overset{(a)}{=} E[\ln |\Sigma| - d \ln 2 \\ &\overset{(b)}{=} \ln |U^{-1} + \sum_i S_i^{-1}| + \sum_{i=1}^{d} \psi\left(\frac{nq + r - d + i}{2}\right) \end{aligned}$$

where $(a)$ expands $\ln |\Sigma/2| = \ln |\Sigma| - d \ln 2$, and $(b)$ applies Proposition 2.
Substituting in this term to the differential entropy formula (17) produces the differential entropy estimate (20).

### *A.14. Derivation of Bayesian Estimate of Relative Entropy Between Inverse Wisharts:*

There are only two terms of (18) that require evaluating the expectation, taken with respect to the independent posteriors of the form given in (40).

The first term is:

$$\begin{aligned} E_{\Sigma_1,\Sigma_2}\left[\mathrm{tr}\left(\Sigma_1^{-1}\Sigma_2\right)\right] &\overset{(a)}{=} (n_2 q_2 + r_2) E_{\Sigma_1}\left[\mathrm{tr}\left(\Sigma_1^{-1}\left(U_2^{-1} + \sum_{j=1}^{n_2} S_{2j}^{-1}\right)\right)\right] \\ &\overset{(b)}{=} \frac{n_2 q_2 + r_2}{n_1 q_1 + r_1 - d - 1} \mathrm{tr}\left(\left(U_1^{-1} + \sum_{j=1}^{n_1} S_{1j}^{-1}\right)^{-1}\left(U_2^{-1} + \sum_{j=1}^{n_2} S_{2j}^{-1}\right)\right), \end{aligned}$$

where $(a)$ follows by Proposition 3, and $(b)$ follows because $\Sigma_1 \sim$ Wishart $(U_1^{-1} + \sum_{j=1}^{n_1} S_{1j}^{-1}, n_1 q_1 + r_1)$ and thus by definition $\Sigma_1^{-1} \sim$ inverse Wishart $((U_1^{-1} + \sum_{j=1}^{n_1} S_{1j}^{-1})^{-1}, n_1 q_1 + r_1)$, and thus $E[\mathrm{tr}(\Sigma_1^{-1} A)] = \mathrm{tr}((U_1^{-1} + \sum_{j=1}^{n_1} S_{1j}^{-1})^{-1} A)/(n_1 q_1 + r_1 - d - 1)$ by Proposition 5.

The second term is,

$$\begin{aligned} E_{\Sigma_1,\Sigma_2}\left[\ln|\Sigma_1^{-1}\Sigma_2|\right] &= E_{\Sigma_1,\Sigma_2}\left[\ln\left(|\Sigma_1|^{-1}|\Sigma_2|\right)\right] \\ &= E_{\Sigma_2}\left[\ln|\Sigma_2|\right] - E_{\Sigma_1}\left[\ln|\Sigma_1|\right] \\ &= \ln\left(\frac{|U_2^{-1}+\sum_{j=1}^{n_2} S_{2j}^{-1}|}{|U_1^{-1}+\sum_{j=1}^{n_1} S_{1j}^{-1}|}\right) \\ &+ \sum_{i=1}^{d}\left(\psi\left(\frac{n_2q_2+r_2-d+i}{2}\right)-\psi\left(\frac{n_1q_1+r_1-d+i}{2}\right)\right), \end{aligned}$$

where the last line follows by applying Proposition 2 twice.

Substituting these two terms into (18) produces the relative entropy estimate (21).

## References


1. El Saddik, A.; Orozco, M.; Asfaw, Y.; Shirmohammadi, S.; Adler, A. A novel biometric system for identification and verification of haptic users. *IEEE Trans. Instrumentation Measurement* **2007**, *56*, 895–906.
2. Choi, H. *Adaptive sampling and forecasting with mobile sensor networks*. MIT PhD Dissertation: Cambridge, MA, USA, 2009.
3. Moddemeijer, R. On estimation of entropy and mutual information of continuous distributions. *Signal Process.* **1989**, *16*, 233–246.
4. Misra, N.; Singh, H.; Demchuk, E. Estimation of the entropy of a multivariate normal distribution. *J. Multivariate Analysis* **2005**, *92*, 324–342.
5. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation for multi-dimensional densities via $k$ nearest-neighbor distances. *IEEE Trans. Inform. Theory* **2009**, *55*.
6. Ahmed, N.A.; Gokhale, D.V. Entropy expressions and their estimators for multivariate distributions. *IEEE Trans. Inform. Theory* **1989**, pp. 688–692.
7. Beirlant, J.; Dudewicz, E.; Györfi, L.; Meulen, E.V.D. Nonparametric entropy estimation: An overview. *Intl. J. Math. Stat. Sci.* **1997**, *6*, 17–39.
8. Nilsson, M.; Kleijn, W.B. On the estimation of differential entropy from data located on embedded manifolds. *IEEE Trans. Inform. Theory* **2007**, *53*, 2330–2341.
9. Kozachenko, L.F.; Leonenko, N.N. Sample estimate of entropy of a random vector. *Probl. Inform. Transm.* **1987**, *23*, 95–101.
10. Goria, M.N.; Leonenko, N.N.; Mergel, V.V.; Inverardi, P.L. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametric Stat.* **2005**, *17*, 277–297.
11. Mnatsakanov, R.M.;Misra, N.S.E. $k_n$-Nearest neighbor estimators of entropy. *Math. Methods of Stat.* **2008**, *17*, 261–277.
12. Hero, A.; Michel, O. Asymptotic theory of greedy approximations to minimal k-point random graphs. *IEEE Trans. Inform. Theory* **1999**, *45*, 1921–1939.
13. Hero, A.; Ma, B.; Michel, O.; Gorman, J. Applications of entropic spanning graphs. *IEEE Signal Process. Mag.* **2002**, pp. 85–95.

14. Costa, J.; Hero, A. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Trans. Signal Process.* **2004**, *52*, 2210–2221.
15. Hulle, M.M.V. Edgeworth approximation of multivariate differential entropy. *Neural Comput.* **2005**, pp. 1903–1910.
16. Wang, Q.; Kulkarni, S.R.; Verdú, S. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Trans. Inform. Theory* **2005**, *51*, 3064–3074.
17. Nguyen, X.; Wainwright, M.J.; Jordan, M.I. Estimating divergence functional and the likelihood ratio by penalized convex risk minimization. *Advances Neural Inform. Process. Syst.* **2007**.
18. Wang, Q.; Kulkarni, S.R.; Verdú, S. A nearest-neighbor approach to estimating divergence between continuous random vectors. *Proc. Intl. Symp. Inform. Theory* **2006**.
19. Pérez-Cruz, F. Estimation of information-theoretic measures for continuous random variables. *Adv. Neural Inform. Process. Syst. (NIPS)* **2008**.
20. Lehmann, E.L.; Casella, G. *Theory of Point Estimation*. Springer: New York, 1998, chapter 4.
21. Bregman, L. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **1967**, *7*, 200–217.
22. Banerjee, A.; Guo, X.; Wang, H. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inform. Theory* **2005**, *51*, 2664–2669.
23. Jones, L.K.; Byrne, C.L. General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis. *IEEE Trans. Inform. Theory* **1990**, *36*, 23–30.
24. Frigyik, B.A.; Srivastava, S.; Gupta, M.R. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Trans. Inform. Theory* **2008**, *54*, 5130–5139.
25. Amari, S.; Nagaoka, H. *Methods of Information Geometry*. Oxford University Press: New York, 2000.
26. Kass, R.E. The Geometry of Asymptotic Inference. *Statistical Science* **1989**, *4*, 188–234.
27. Srivastava, S.; Gupta, M.R.; Frigyik, B.A. Bayesian quadratic discriminant analysis. *J. Mach. Learn. Res.* **2007**, *8*, 1287–1314.
28. Havil, J. *Gamma*. Princeton University Press: Princeton, NJ, USA, 2003.
29. Bercher, J.; Vignat, C. Estimating the entropy of a signal with applications. *IEEE Trans. Signal Process.* **2000**, *48*, 1687–1694.
30. Bilodeau, M.; Brenner, D. *Theory of Multivariate Statistics*. Springer Texts in Statistics: New York, NY, USA, 1999.