Entropic Priors

Carlos C. Rodríguez
Department of Mathematics and Statistics
State University of New York at Albany
E-Mail: carlos@omega.albany.edu

October 4, 1991

Abstract: Entropic priors assign probabilities by combining in an inseparable way the information theoretic concept of entropy with the underlying Riemannian geometry of the hypothesis space. These priors form the cornerstone of a developing new and more objective Bayesian theory of inference.

Contents

1 Introduction

This assay is about a new method for building a prior model on the parameters of a regular family of distributions. Parts of this method have been know for some time but the full generality of entropic priors is a recent discovery. With this single procedure it is possible to reproduce useful and well-known results in diverse areas such as time series, image processing and empirical bayes. A deeper understanding of the basic ideas underlying the method could lead to better solutions for these applications, as well as to a new and more objective theory of inference.

The method is summarized with the following formula: "Given a regular parametric model containing probability measures P_θ the entropic priors on θ are given by

$$Pr(d\theta|\alpha,\theta_0) \propto \exp\{-\alpha I(\theta:\theta_0)\} g^{1/2} d\theta \tag{1}$$

where α is a positive scalar parameter, θ_0 is an initial guess for θ, $I(\theta:\theta_0)$ denotes the Kullback number between the probability measures P_θ and P_{θ_0} and g is the determinant of the Fisher information matrix at θ".

Equation (1) has a number of desirable properties: it is of general applicability, it is invariant under smooth changes in the coordinate systems of both the parameter and the data spaces, and it contains the methods of maximum entropy and Jeffreys' noninformative priors as two opposite poles ($\alpha \to \infty, \alpha \to 0$). Also, entropic priors are essentially implied by the assumptions of invariance and no apriori correlations (see Skilling ([15],[16]) and also Shore and Johnson ([13]), Rodríguez ([10],[11])).

There is a very simple rationale behind equation (1). The amount of probability assigned to the parameter θ decreases exponentially with the Kullback distance from the initial guess of θ_0. That is to say, the more difficult it is to discriminate the measure P_θ from P_{θ_0}, the more prior probability is given to it. The parameter α controls the sensitivity to changes in the distance. Thus, reliable $\theta_0' s$ should go with large α while unreliable guesses should be given small α. In the latter case the dependence of prior probabilities on θ is controlled not so much by the Kullback number (and therefore by the initial guess of θ_0) but by the surface area of the model given by the invariant measure $g^{1/2}d\theta$.

The entropic priors assign prior probabilities by combining in an intricate way the information theoretic concept of entropy with the underlying Riemannian structure of the parametric model. This combination of entropy and geometry is even more intricate than suggested by the notation of equation (1); for, the information metric, and therefore also its determinant g, arise from infinitesimal variations of the Kullback number (see equation (9) below).

2 Background: Entropy, Geometry, and Priors

Underlying the problem of inference we find the following fundamental question: How should the data and the prior information about a physical process be used to generate a manifold of distributions and a measure of prior uncertainty? The relevance of this question is clear, for once we fix the model and the prior, bayesian inference is reduced to the computational problem of posterior probabilities. Needless to say, unless the meaning of **prior information** is explicitly defined, there is no general solution to this problem. However, if a parametric model is available for the data and all the prior information (besides the knowledge of this parametric model) is contained in the initial value of the parameter, then the entropic priors, as defined by equation (1), provide invariant assignments of prior probabilities with desirable properties.

Entropic priors are, therefore, part of the answer to the fundamental question stated above. Although they solve only half of the problem, they solve the important half. Due to the lack of general methods for transforming prior information into assigments of prior probabilities, there is greater agreement among statisticians about models than about priors. The maximum entropy formalism (see Jaynes [12]) and the total ignorance priors obtained from invariance (see Jeffreys [7], and Jaynes [12]) are two isolated exceptions. It is remarkable that entropic priors include both methods as extreme special cases.

An overview of the main quatities that define formula (1) is presented below. To this aim it is necessary to introduce some standard definitions and results.

2.1 The Kullback Number

Let the data space be an arbitrary Haussdorf space $\mathcal{X}$ endowed with the Borel sigma field $\mathcal{B}$. In most applications, $\mathcal{X}$ is just a measurable piece of $\mathbb{R}^d$ but it is equally easy to deal with the more general case. The central quantity in the theory is the Kullback number $I(P:Q)$ between a probability measure P and a σ-finite measure Q both defined on the measurable data space $(\mathcal{X}, \mathcal{B})$. Define

$$I(P:Q) = \int P(dD) \log \frac{dP}{dQ}(D) \tag{2}$$

when P is absolutely continuous with respect to Q and the integral exists and define it as ∞ otherwise. When both P and Q are probability measures absolutely continuous with respect to each other, the above integral always exists even though it may be infinite (see Kullback [8] p. 5). Moreover, in this case, a simple application of Jensen's inequality shows that $I(P:Q)$ is non negative and that it is zero only when $P \equiv Q$. This last property makes it very similar to a metric, but of course is not symmetric and it does not satisfy the usual triangular inequality.

In statistical language, the Kullback number is the expected log-likelihood ratio of P to Q when P is the "true" model. In information theoretic terms, it is said to measure the mean information per observation for discrimination in favor of P and against Q when sampling from P. In stronger information theoretic language, it is the expected amount of information transmitted by the message: "The information source has been moved from Q to P". All these well stablished interpretations convey the same idea of separation of P from Q. But, the large deviation property reproduced below explicitly quantifies this separation:

Let $\mathcal{X} = \{1, 2, ..., k\}$ and let q be the probability measure that assigns probability q_j to the jth element of $\mathcal{X}$. We can imagine an urn with known proportions q_j of balls type j for $j = 1, 2, ..., k$. If we draw n balls from this urn and we denote by n_j the number of observed balls of type j then the probability of observing a frequency distribution $p_1, p_2, ..., p_k$ with $p_j = n_j/n$ is given by the multinomial

$$Pr(p|q) = Pr(n_1, ..., n_k|q) = \frac{n!}{n_1!n_2!...n_k!} {q_1}^{n_1} ... {q_k}^{n_k} \tag{3}$$

i.e. the chance of seeing a p–distribution when sampling n times from a q–distribution. Taking the logarithm on both sides of equation (3) and using Stirling's formula up to first order i.e.

$$\log m! = m \log m - m + o(m) \tag{4}$$

we obtain after a little simplification that,

$$\log Pr(p|q) = -n \sum_{j=1}^{k} p_j \log \frac{p_j}{q_j} + o(n) \tag{5}$$

Therefore when all the n_js are large, this last equation together with (2) imply

$$Pr(p|q) \propto e^{-nI(p:q)} \tag{6}$$

which has the exact form of (1). Notice that since n is large but finite, there are only a finite number of p-distributions and that is why the g in (6) is constant; there is no underlying continously parametrized manifold of distributions, only a finite number of them. Also, we obtain once more the parameter α associated to the amount of information about θ_0. In this case, however, the term **amount of information** takes the explicit form of the **number of observations** from the q-distribution.

2.2 Fisher Information Metric

The connection between α and amount of information in the sense of number of observations holds in general as it is shown below. Before we prove it we notice the classic relationship between entropy and geometry, i.e., the natural

Riemannian metric on a regular model appears from second variations of the entropy. More explicitly, the Riemannian length of a tangent vector v attached to the point θ is proportional to $I(P_{\theta+\epsilon v} : P_\theta)$ when ϵ is small. To see this, just consider (for fix θ and v) the function $u(\epsilon) = I(P_{\theta+\epsilon v} : P_\theta)$ that has a global minimun of zero at $\epsilon = 0$. Expanding up to second order terms in ϵ we have,

$$u(\epsilon) = u(0) + \epsilon u'(0) + \frac{1}{2}\epsilon^2 u''(0) + o(\epsilon^2) \tag{7}$$

by straight forward computation we find $u(0) = u'(0) = 0$ and

$$u''(0) = \sum_{i,j} v^i \left[\int \frac{1}{p_\theta}\frac{\partial p_\theta}{\partial \theta^j}\frac{1}{p_\theta}\frac{\partial p_\theta}{\partial \theta^i} P_\theta(dD) \right] v^j \tag{8}$$

where $p_\theta = p_\theta(D)$ denotes the density of P_θ with respect to an arbitrary, but fix, dominating measure. The last formula defines a quadratic form on tangent vectors v. Notice that the quadratic form must be positive definite because $u(\epsilon)$ has a global minimum at $\epsilon = 0$. The matrix of this quadratic form is known as the Fisher information matrix and the induced Riemannian metric is known as the information metric[1] (see [8],[1] or [10]). Thus,

$$I(P_{\theta+\epsilon v} : P_\theta) = \frac{\epsilon^2}{2}\sum_{i,j} v^i g_{ij}(\theta) v^j + o(\epsilon^2) \tag{9}$$

where $g_{ij}(\theta)$ are the elements of Fisher's matrix.

2.3 Prior Information *is* More Data

An interesting question suggested by (1) that turns out to have a surprisingly simple and useful answer is the following: What are the entropic priors when the data space consists of vectors $D = (x^1, ..., x^n)$ of independent and identically distributed components?

To answer the above question let $P_\theta^{(n)}$ and $g_{ij}^{(n)}(\theta)$ denote the corresponding quantities in the data space $\mathcal{X}^n$. The components are iid so,

$$P_\theta^{(n)}(dD) = \prod_{i=1}^{n} P_\theta(dx^i). \tag{10}$$

Therefore, from this, (2), and Fubini's theorem we have

$$I(P_\theta^{(n)} : P_{\theta_0}^{(n)}) = nI(P_\theta : P_{\theta_0}). \tag{11}$$

From equations (9) and (11) we can write

[1]Regular models are smooth Riemannian manifolds relative to the Haussdorf topology of the Hellinger distance.

$$g_{ij}^{(n)}(\theta) = n g_{ij}(\theta) \tag{12}$$

and taking the determinant on both sides of this matrix equation we obtain $g^{(n)} \propto g$ that together with (1) and (11) imply

$$Pr^{(n)}(d\theta|\alpha, \theta_0) = Pr(d\theta|n\alpha, \theta_0) \tag{13}$$

Equation (13) links the positive scalar parameter α to the number of observations. It suggests to regard α as a kind of continuous virtual number of independent observations supporting the choice of θ_0 as the initial guess. This interpretation squares nicely with our previous results. More over, from equation (13) we can write the posterior as

$$Pr^{(n)}(d\theta|D, \alpha, \theta_0) \propto \prod_{i=1}^{n} \left\{ P_\theta(dx^i) e^{-\alpha I(\theta:\theta_0)} \right\} g^{1/2}(\theta) d\theta \tag{14}$$

and therefore the posterior densities with respect to the invariant measure on the manifold satisfy

$$\pi^{(n)}(\theta|x^1, ..., x^n, \alpha, \theta_0) = \prod_{i=1}^{n} \pi(\theta|x^i, \alpha, \theta_0) \tag{15}$$

The last equation not only provides a recursive procedure that simplifies the computation of the posteriors when new data are available but it also reinforces the above interpretation for α. Given the classic connection between the Kullback number and the exponential family, a more detailed examination of the form of the posterior in the exponential family case (specially for θ close to θ_0) should produce interesting and informative results.

3 Applications

3.1 Empirical Bayes

The empirical Bayes approach to statistics is due to H. Robbins who has demonstrated the usefulness of this technique in both the parametric and nonparametric cases. However, in the parametric case a prior over the parameters is needed and, as it is shown here with an example, equation (1) supplies one that makes sense.

Efron and Morris were the first to provide an empirical Bayes context to Stein-type estimators, but their priors and unbiased method of estimation were chosen in an ad-hoc manner explicitly designed to obtain the result (see Robbins collection of selected papers [9] and list of references on pages 1–5).The simplest non-trivial application of entropic priors generates the celebrated Stein's shrinking phenomenon without the ad-hockeries and at the same time provides a rationale for it.

Let $x = (x^1, \ldots, x^k)$ be a single observation from a k-variate Gaussian distribution with unknown mean vector θ but known variance matrix $\sigma^2 I$ with $\sigma > 0$. i.e. $P_\theta \equiv N(\theta, \sigma^2 I)$. In this case the Kullback number is given by

$$I(\theta : \theta_0) = \frac{1}{2\sigma^2} E_\theta\{|x-\theta_0|^2 - |x-\theta|^2\} = \frac{1}{2\sigma^2}|\theta-\theta_0|^2 \tag{16}$$

i.e. proportional to the square of the euclidean distance between θ and θ_0. Thus, it follows directly from equation (9) that the information matrix is proportional to the identity and that the invariant measure is proportional to the Lebesgue measure on $\mathbb{R}^k$. Therefore, replacing in (1) we obtain the entropic prior

$$Pr(d\theta|\alpha,\theta_0) \propto \exp\left(-\frac{\alpha}{2\sigma^2}|\theta-\theta_0|^2\right) d\theta. \tag{17}$$

The prior is then $N\left(\theta_0, \frac{\sigma^2}{\alpha} I\right)$. Let $f(x|\alpha,\theta_0)$ denote the marginal density of the data vector with respect to the Lebesgue measure on $\mathbb{R}^k$. By conditioning on θ and using the above entropic prior we have

$$f(x|\alpha,\theta_0) \propto \int \exp\left\{\frac{-1}{2\sigma^2}\left[|x-\theta|^2 + \alpha|\theta-\theta_0|^2\right]\right\} d\theta \tag{18}$$

thus, after replacing the equation

$$|x-\theta|^2 = |x-\theta_0|^2 + |\theta_0-\theta|^2 - 2\langle x-\theta_0, \theta-\theta_0\rangle \tag{19}$$

and completing the square inside the exponential, it follows that

$$f(x|\alpha,\theta_0) \propto e^{\left\{\frac{-\alpha|x-\theta_0|^2}{2(\alpha+1)\sigma^2}\right\}} \int e^{\left\{\frac{-(\alpha+1)\left|\theta-\left(\theta_0-\frac{1}{\alpha+1}(\theta_0-x)\right)\right|^2}{2\sigma^2}\right\}} d\theta$$

from where we can immediately deduce that the marginal distribution of x is given by

$$f(x|\alpha,\theta_0) \equiv N\left(\theta_0, \frac{\alpha+1}{\alpha}\sigma^2 I\right) \tag{20}$$

and that the posterior distribution of θ after observing a single x is proportional to the expression inside the integral i.e.,

$$\pi(\theta|x,\alpha,\theta_0) \equiv N\left(x - \frac{\alpha}{\alpha+1}(x-\theta_0), \frac{\sigma^2}{\alpha+1} I\right). \tag{21}$$

Hence, for quadratic loss, the Bayes' estimate of θ is not the single observed vector x, but the mean of the posterior, that shows a shrinkage towards the guessed value θ_0. The amount of shrinkage can be estimated by extracting the information about α contained in the observed vector x. To see this, let $\lambda = \frac{\alpha}{\alpha+1}$ and notice that from (20) and (21) it follows that:

$$f(x|\lambda, \theta_0) \equiv \prod_{j=1}^{k} N\left(\theta_0^j, \frac{\sigma^2}{\lambda}\right) \tag{22}$$

and,

$$\mathbf{E}(\theta|x, \lambda, \theta_0) = x - \lambda(x - \theta_0). \tag{23}$$

The idea is to replace in (23) the most conservative estimator of the unknown λ that could be obtained from the inference problem defined by (22). Guided again by equation (1) it follows that in each independent component of x the total ignorance prior about the parameters of the j-th problem is given by:

$$\eta\left(d\theta_0^j, d\lambda\right) \propto \frac{d\lambda d\theta_0^j}{\lambda^2}. \tag{24}$$

Thus, the ignorance prior for all the parameters is:

$$\eta\left(d\theta_0, d\lambda\right) \propto \frac{d\lambda d\theta_0}{\lambda^2} \tag{25}$$

This prior makes λ ignorant not only of θ_0 but also about the dimensionality k. From the prior (25) and the likelihood (20) it follows that,

$$Pr(d\lambda|\theta_0, x) \propto \lambda^{\frac{k}{2}-2} \exp\left\{\frac{-\lambda|x-\theta_0|^2}{2\sigma^2}\right\} d\lambda$$

and this is a gamma distribution with mean

$$\mathbf{E}\left(\lambda|\theta_0, x\right) = \frac{(k-2)\sigma^2}{|x-\theta_0|^2}. \tag{26}$$

Replacing the unknown λ in (23) by the right hand side of (26) we obtain the estimator:

$$\theta^* = x - \frac{(k-2)\sigma^2}{|x-\theta_0|^2}(x - \theta_0) \tag{27}$$

which is the original Stein's estimator having uniformly better risk than x for all choices of θ_0 when $k \geq 3$. Thus, in dimensions grater than two, x is inadmissible (see for example [6, pages 25–27]). These are all classic results of the theory of estimation produced effortlessly from equation (1).

Stein's estimator exemplifies the general (Entropic) Empirical Bayes approach: first, reduce the problem of estimation of the (multi) parameter θ to the estimation of the positive scalar parameter α, as in (20). Second, replace in the posterior distribution of θ the unknown α by its estimator, as in (21) and (27).

3.2 Time Series

It is shown in this section that the general formalism of entropic priors could be applied to solve time series problems. Preliminirary results for the model of a parametric signal plus white noise are shown.

3.2.1 The Signal Manifold

Let us assume that the vector of observations $x = (x^1, \ldots, x^N)$ collected at times $t_1, t_2, \ldots, t_N$, not necessarily equally spaced, can be modeled by:

$$x^l = f(t_l, \theta) + e_l, \quad \text{for } l = 1, 2, \ldots, N \tag{28}$$

where $\theta \in \Theta \subset \mathbb{R}^k$, and the set

$$S = \left\{ r(\theta) = (f(t_1, \theta), \ldots, f(t_N, \theta)) \in \mathbb{R}^N : \theta \in \Theta \right\} \tag{29}$$

is a smooth k-dimensional surface in $\mathbb{R}^N$. The intrinsic geometry of S is characterized by its metric tensor:

$$\overline{g}_{ij}(\theta) = \langle \frac{\partial r}{\partial \theta_i}, \frac{\partial r}{\partial \theta_j} \rangle = \sum_{l=1}^{N} f_i(t_l, \theta) f_j(t_l, \theta) \tag{30}$$

where $f_i(t_l, \theta)$ denotes the partial derivative of $f(t_l, \theta)$ with respect to θ_i. The errors, e_l, are assumed[2] to be independent for different times and gaussian with mean 0 and variance σ^2 for all times. Hence, the joint likelihood function for θ and σ, denoted by $L(\theta, \sigma)$, is given by,

$$L(\theta, \sigma) \propto \sigma^{-N} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{l=1}^{N} \left(x^l - f(t_l, \theta) \right)^2 \right\} \tag{31}$$

that defines the hypothesis space of all the N-variate gaussian probability measures with mean vector on the surface S and variance matrix $\sigma^2 I$. The Riemannian geometry on this hypothesis space is intimately related to the geometry on S. The metric tensor (or Fisher information matrix) in this space can be obtained by taking the derivatives of the log-likelihood (see equation (8) or Amari ([1]), Rodríguez ([10])).

$$l(\theta, \sigma) = -N \log \sigma - \frac{1}{2\sigma^2} \sum_{l=1}^{N} \left(x^l - f(t_l, \theta) \right)^2 . \tag{32}$$

The derivative with respect to θ_j gives the tangent vector, ∂_j, in the j-th coordinate direction given by,

[2] This *assumption* follows from (1): make θ denote an arbitrary distribution for the errors with a given finite variance, flat initial, and large α.

$$\partial_j = \frac{1}{\sigma^2} \sum_{l=1}^{N} \left(x^l - f(t_l, \theta) \right) f_j(t_l, \theta) \quad \text{for } j = 1, \ldots, k. \tag{33}$$

and the tangent vector in the direction of σ is given by,

$$\partial_\sigma = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{l=1}^{N} \left(x^l - f(t_l, \theta) \right)^2 . \tag{34}$$

Thus, the components of the metric tensor in the θ-directions are:

$$\begin{aligned} g_{ij}(\theta, \sigma) &= E_{\theta,\sigma} \left(\partial_i \partial_j \right) \\ &= \frac{1}{\sigma^2} \sum_{l=1}^{N} f_i(t_l, \theta) f_j(t_l, \theta) \\ &= \frac{1}{\sigma^2} \overline{g}_{ij}(\theta) \end{aligned}$$

and it follows from (33) and (34) and the assumption of independence of the errors that,

$$\begin{aligned} g_{j\sigma}(\theta, \sigma) &= E_{\theta,\sigma} \left(\partial_j \partial_\sigma \right) \\ &= E \left\{ -\frac{N}{\sigma^3} \sum_{l} e_l f_j(t_l, \theta) + \frac{1}{\sigma^5} \sum_{l,m} e_l^2 e_m f_j(t_m, \theta) \right\} \\ &= 0. \end{aligned}$$

The last element of Fisher's matrix may be computed from the derivative of equation (34), ∂_σ^2. We have,

$$\begin{aligned} g_{\sigma\sigma}(\theta, \sigma) &= -E_{\theta,\sigma} \left(\partial_\sigma^2 \right) \\ &= -E \left\{ \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{l} e_l^2 \right\} \\ &= \frac{2N}{\sigma^2}. \end{aligned}$$

Hence, the matrices $[g(\theta, \sigma)]$ and $[\overline{g}(\theta)]$ satisfy,

$$[g(\theta, \sigma)] = \frac{1}{\sigma^2} \begin{pmatrix} [\overline{g}(\theta)] & 0 \\ 0 & 2N \end{pmatrix} .$$

The total ignorance prior, $\eta(d\theta, d\sigma)$, can be written in terms of the square root of the determinant of the previous matrix. We have,

$$\eta(d\theta, d\sigma) \propto \sigma^{-(k+1)} \left|\overline{g}(\theta)\right|^{\frac{1}{2}} d\theta d\sigma. \tag{35}$$

Notice that when σ is assumed to be known, the non-informative prior for θ coincides with the k-dimensional surface area of S and the Kullback number is similar to (16). By looking at equation (21) it is therefore reasonable to estimate θ by θ^* such that,

$$r(\theta^*) = x - \frac{\alpha}{\alpha+1} \left(x - r(\theta_0)\right). \tag{36}$$

It is natural to expect θ^* to show properties similar to Stein's estimator, but only when the surface S is sufficiently flat between θ^* and θ_0. The exact solution could be computed (or at least approximated in specific examples) and it would provide a new way for recovering a signal buried in white noise.

3.2.2 Separating Frequencies from Amplitudes

In the rest of this section we analize the time series estimation problem when a little more structure than just S is assumed for the signal. Computations are significantly simplified if linear (usually nuisance) parameters are separated from non-linear (frequencies like) parameters. Following Bretthorst ([2]) let us assume that the signal can be modeled by:

$$f(t, \theta) = \sum_{j=1}^{m} B_j G_j(t, \omega) \tag{37}$$

where $\theta = (B_1, \ldots, B_m, \omega_1, \ldots, \omega_{k-m}) \equiv (B, \omega)$. Using (30) we can write:

$$\overline{g}_{ij} = \sum_{l=1}^{N} G_i(t_l, \omega) G_j(t_l, \omega) \quad \text{for } i, j \leq m \tag{38}$$

which are independent of B. The other components are given by,

$$\overline{g}_{ia} = \sum_{j=1}^{m} B_j \left(\sum_{l=1}^{N} G_i(t_l, \omega) G_{ja}(t_l, \omega) \right) \quad \text{for } i \leq m, a > m \tag{39}$$

where G_{ja} denotes the partial derivative of G_j with respect to $\theta_a \equiv \omega_{a-m}$. These components are linear in B. Finally we can write:

$$\overline{g}_{ab} = \sum_{i,j} B_i B_j \left(\sum_{l=1}^{N} G_{ia}(t_l, \omega) G_{ib}(t_l, \omega) \right) \quad \text{for } a, b > m.$$

These components are quadratic in B. The whole matrix $[\overline{g}(B, \omega)]$ can be split into four blocks:

$$[\overline{g}(B,\omega)] = \begin{pmatrix} [\overline{g}_{B,B}] & [\overline{g}_{B,\omega}] \\ [\overline{g}_{B,\omega}] & [\overline{g}_{\omega,\omega}] \end{pmatrix} = \begin{pmatrix} [Indep.\ of\ B] & [Linear\ in\ B] \\ [Linear\ in\ B] & [Quadratic\ in\ B]\,. \end{pmatrix}$$

The determinant of this matrix is the sum of products of elements taken from each row and column. All having the form:

$$\pm\, u(\omega) B_1^{\alpha_1} B_2^{\alpha_2} \ldots B_m^{\alpha_m} \quad \text{with } \alpha_j \geq 0 \quad \text{and } \sum_{j=1}^{m} \alpha_j = 2(k-m). \tag{40}$$

Hence, the determinant shows the following homogeneity property:

$$|\overline{g}(\lambda B,\omega)| = \lambda^{2(k-m)} |\overline{g}(B,\omega)|\,. \tag{41}$$

More over, $[\overline{g}_{B,B}]$ is a symmetric positive definite matrix that appears also in the likelihood:

$$\begin{aligned} L(B,\omega,\sigma) &\propto \sigma^{-N} \exp\left\{ \frac{-1}{2\sigma^2} \sum_{j=1}^{m} \left(x^j - \sum_{l=1}^{m} B_j G_j(t_l,\omega) \right)^2 \right\} \\ &\equiv \sigma^{-N} \exp\left\{ -\frac{NQ}{2\sigma^2} \right\} \end{aligned}$$

where Q is defined by,

$$Q = \overline{d^2} - \frac{2}{N} \sum_j h_j B_j + \frac{1}{N} \sum_{i,j} B_i B_j \overline{g}_{ij} \tag{42}$$

also,

$$\overline{d^2} = \frac{1}{N} \sum_{l=1}^{N} \left(x^l\right)^2 \quad \text{and } h_j = \sum_{l=1}^{N} x^l G_j(t_l,\omega). \tag{43}$$

Thus, by diagonalizing $[\overline{g}_{B,B}]$ we simplify, not only the ignorance prior, but also the likelihood function. It is, therefore, straight forward to re-write the analysis in [2] using equation (1) as the prior. For $\alpha = 0$ (total ignorance case) we obtain the posterior distribution for the frequencies given by:

$$Pr(d\omega|x,\alpha=0) \propto \sqrt{|\overline{g}(h,\omega)|} \left(1 - \frac{m\overline{h^2}}{N\overline{d^2}} \right)^{-\frac{N+k-m}{2}} d\omega \tag{44}$$

which is different from the result obtained in [2]. However, when N is large, the generalized periodogram $\overline{h^2}$ peacks about the "true" ω but from equation (41) the term $\sqrt{\overline{g}(h,\omega)}$ increases only polynomially in h which is negligible compared to the other term that increases exponentially in h. Hence, for large N equation (44) coincides with the result in Bretthorst. We have computed both posteriors on simulated data containing one or two harmonics observing always that the two curves become indistinguishable very rapidly for N around 20.

Hence, non-informative entropic priors reproduce the results in the Bayesian spectral analysis of Bretthorst. More over, entropic priors make possible to incorporate definite prior knowledge (e.g. about the frequencies ω) inproving the estimate of the signal.

3.3 Image Reconstruction

All image reconstruction problems are inverse problems. The mathematics of image reconstruction can be reduced, in theory, to the inversion of a linear functional operator. The (approximate) linear functional relation between the input (object) and the output (data) is fixed by the physics of the particular situation, but the basic idea common to most methods is very simple (see [5]). The object f to be reconstucted, is hit with some kind of radiation, and a characteristic g, of the scattered output field is recorded. Typical examples include: radio frequency in NMR, x-rays in computed tomography, or sound waves in ultrasound imaging. Usually, g is linearly related to f through

$$g = Af \tag{45}$$

where A is a linear functional from the hypothesis space $\mathcal{F}$ (also known as image space or solution space) into the data space $\mathcal{G}$. The functional A, is typically a Fredholm integral operator that can be specified in terms of some kernel $K(x, y)$ in the form:

$$(Af)(y) = \int K(x,y)f(x)dx. \tag{46}$$

Hence, in theory the object f is obtained from g by,

$$f = A^{-1}g \tag{47}$$

but in practice, the value of g is only known **with noise** at N discrete points $y_1, y_2, \ldots, y_N$. What is observed are N noisy data,

$$\begin{aligned} D_j &= g(y_j) + \epsilon_j \qquad (48) \\ &= \int K(x,y_j)f(x)dx + \epsilon_j \quad \text{for } i = 1,\ldots,N. \end{aligned}$$

Therefore, in reality what needs to be solved is not the algebraic deterministic problem (48) but the inference problem: *Guess f from the N noisy data.* Notice that (47) is equivalent to (28) with a very large (possibly infinite) dimensional θ.

3.3.1 Digital Imaging

A discretization of x into pixels transforms problem (48) into a regression problem but with more parameters than data. The data alone is not enough for ranking all the possible pictures in $\mathcal{F}$, and a prior distribution over $\mathcal{F}$ is needed.

Prior knowledge about the set of possible images can be used to reduce the dimensionality of the hypothesis space $\mathcal{F}$, providing an encoding of the image with fewer number of parameters than the number of pixels. However, if nothing is known about the picture, the space $\mathcal{F}$ can be identified with the set of all the non-degenerate probability measures over the finite set of pixels. The Kullback number and invariant measure in this space of discrete distributions are easily computed from their respective definitions. Replacing in (1) it follows that the log of the entropic prior density is

$$\log\left\{\frac{Pr(df|\alpha, m)}{df}\right\} \propto -\alpha \sum_i f_i \log\frac{f_i}{m_i} + \frac{1}{2}\log\frac{1}{\prod_i f_i} \tag{49}$$

where $f_i = f(x_i)$ i.e. proportion of luminosity assigned to the i-th pixel and m denotes the initial guess for f.

The likelihood of f is fixed by fixing a distribution for the errors in (48) and hence, the most likely posterior choice for f is the solution of:

$$\max_{f\in\mathcal{F}}\left\{L(f) - \alpha\sum_i f_i \log\frac{f_i}{m_i} + \frac{1}{2}\log\frac{1}{\prod_i f_i}\right\}. \tag{50}$$

whrere $L(f)$ denotes the logarithm of the probability of the data given f. The MAP defined by the solution of (50) is what is computed by the celebrated MEMSYSs algorithm of Skilling, Gull and their co-workers (see [16]). It is interesting to notice that the last term in (50) i.e. the invariant measure on $\mathcal{F}$ was discovered experimentally by testing different functions of f to penalize values close to zero (Skilling personnal communication, see also [14, p. 99]).

3.3.2 An Example with Binary Images

To illustrate the use of the main formula (1) when specific prior information is available, consider the following simple problem: On a square of M pixels (e.g. the terminal screen) a few rectangles of random lengths and widths are allocated at random (e.g. by turning **on** the pixels of the rectangles). The image is then destroyed as follows: each of the M pixels is independently reversed with probability $0 < q < 0.5$. It is required to recover as close as possible the original image from these data.

Let's denote by $\theta = (\theta_1, \ldots, \theta_M)$ the unknown image, where

$$\theta_j = \begin{cases} 1 & \text{if } j\text{-th pixel was originally on} \\ 0 & \text{if } j\text{-th pixel was originally off} \end{cases}$$

similarly, let $z = (z_1, \ldots, z_M)$ be the observed pixels (i.e. the pixels of the destroyed image. Thus, the likelihood is given by:

$$Pr(z|\theta, q) = \prod_{j=1}^{M} \left\{ q^{|z_j - \theta_j|}(1-q)^{1 - |z_j - \theta_j|} \right\} \tag{51}$$

Let's denote by $P_{\theta,q}$ the probability measure in (51). The Kullback number for a given initial guess $P_{\mu,r}$ simplifies to,

$$I\left(P_{\theta,q} : P_{\mu,r}\right) = MI(q:r) - (1-2q)\log\left(\frac{r}{1-r}\right)\sum_{j=1}^{M} \mathbf{1}(\theta_j \neq \mu_j) \tag{52}$$

where $\mathbf{1}(A)$ is 1 if A is true and zero otherwise and,

$$I(q:r) = q\log\frac{q}{r} + (1-q)\log\frac{1-q}{1-r}.$$

The entropic prior density is then given by

$$\pi(\theta, q|\alpha, \mu, r) \propto \exp\left\{-\alpha I\left(P_{\theta,q} : P_{\mu,r}\right)\right\} \tag{53}$$

but relative to the invariant measure on the hypothesis space. The information given in the problem is not sufficiently precise to fix this measure from first principles. However, it is clear that it should decrease rapidly with the number of rectangles of the image θ to agree with the information that θ contains only a few rectangles. A useful choice is

$$\eta(d\theta, dq) \propto \exp\left\{-\kappa N_R(\theta)\right\} d\theta dq \tag{54}$$

where κ is a positive scalar parameter (like α) and $N_R(\theta)$ denotes the number or rectangles in the image θ. By using equations (51,52,53,54) it follows that the posterior mode is obtained by solving:

$$\max_{\theta, q} \quad \left\{ M\log(1-q) + \log\left(\frac{q}{1-q}\right)\sum_{j=1}^{M}\mathbf{1}(\theta_j \neq z_j) + \right. \\ \left. \alpha(1-2q)\log\frac{r}{1-r}\sum_{j=1}^{M}\mathbf{1}(\theta_j \neq \mu_j) - \alpha MI(q:r) - \kappa N_R(\theta) \right\}. \tag{55}$$

An algorithm to approximate the solution of problem (55) could be designed but only after deciding what to do with the parameters: α, κ, r, μ. A fast and simple smoothing of the data will fix r and μ and then (assumming enough computational muscle) solve (55) for different values of α and κ. Although, (55) is the problem to solve, when the proportion of pixels **on** in the original image

θ is small (i.e. few rectangles in θ) then q can be closely approximated from the data. In this case the optimization problem (55) simplifies to

$$\min_{\theta} \left\{ \sum_{j=1}^{M} \mathbf{1}(\theta_j \neq z_j) + \gamma \sum_{j=1}^{M} \mathbf{1}(\theta_j \neq \mu_j) + \beta N_R(\theta) \right\}. \tag{56}$$

where the positive parameters γ and β are defined by:

$$\gamma = \alpha(1-2q) \quad \text{and } \beta = \frac{-\kappa}{\log \frac{q}{q-1}}.$$

A further simplification is obtained by writing,

$$\sum_j [\mathbf{1}(\theta_j \neq z_j) + \gamma \mathbf{1}(\theta_j \neq \mu_j)] = (\gamma+1) \sum_{z_j = \mu_j} \mathbf{1}(\theta_j \neq z_j) + \sum_j \mathbf{1}(z_j \neq \mu_j)$$

Hence, replacing in (56), dividing through by γ (which is postive), and redefining β accordingly, the optimization problem to be solved reduces to:

$$\min_{\theta} \left\{ \sum_{z_j = \mu_j} \mathbf{1}(\theta_j \neq z_j) + \beta N_R(\theta) \right\}. \tag{57}$$

The exact solution of the above combinatorial optimization problem is out of the question with a regular desktop computer; the search space of θs has 2^M or more than $10^{3,000}$ elements for a square of 100×100 pixels. However, an stochastic optimization algorithm (like simulated annealing) rapidly produces high quality solutions[3] (see [3]).

It seems profitable to exploit the use of entropic priors in image processing problems of this kind.

4 Towards a Bayesian Theory of Information Geometry

The main advantage of having a geometric theory of inference is the enhancement on imagination that it produces. Geometry allows to **see** by relating familiar images from our every-day three dimensional space, to otherwise abstract mathematical objects living in obscure spaces. By thinking of the hypothesis space (i.e. parametric model) as a Riemannian manifold, we can picture the possible probability measures for our data living as points on a curved surface. With this imagery in mind, and the paraphernalia of modern geometry, the theory of inference is rapidly developing into a new science that Amari has baptized as **information geometry** [1, p. 7].

[3] I have implemented the algorithm in Quick-C and it takes a few minutes runnig on a 386/33Mhz to obtain solutions comparable to those in [3].

4.1 Robustness and the Lie Derivative

This section contributes to the ongoing development of information geometry by showing how the geometric concept of the Lie derivative is naturally related to the statistical concept of robustness.

Statistical robustness is a desirable property of an inferential procedure. It can be defined, qualitatively, as the stability of the procedure against small changes in the assumptions. The pictures that come to mind, almost simmultaneously with this definition, are those of solids able to relax back to their original form after feeling the effects of a small deformation or strain. It is therefore not surprising, that ideas developed in the physical theories of elasticity and the mechanics of continuous media, are found useful to quantify statistical robustness.

One aspect of the intrinsic robustness of a hypothesis space can be quantified by the **strain tensor** (see for example [4, pages 155 and 205–211]). Recall that the strain tensor is a covariant tensor of order two (i.e. a bilinear form on vectors) that quantifies how the metric of the space changes under deformations defined by a vector field. Evidently, this definition can only encode one aspect of robusteness: the robusteness of the metric. However, remember that the strain tensor is obtained as the Lie derivative of the metric with respect to a vector field, and that the Lie derivative is defined not only for metrics but for arbitrary tensors on the manifold. Hence, by taking Lie derivatives, it is possible to quantify, the rate of change with small deformations of the space, for many statistically meaningful quantities defined on the model. A non exhaustive list of quantities for wich it would be desirable to know the Lie derivative and their properties, includes:

1. Information matrix (metric tensor), $g_{ij}(\theta)$
2. Total ignorance prior, $\sqrt{|g(\theta)|}d\theta \equiv \sqrt{|g(\theta)|}d\theta^1 \wedge d\theta^2 \wedge \ldots \wedge d\theta^k$
3. Entropic prior, $\exp\{-\alpha I(\theta : \theta_0)\}\, g^{1/2} d\theta$
4. Posterior distribution, $P_\theta(dx) \exp\{-\alpha I(\theta : \theta_0)\}\, g^{1/2} d\theta$
5. A parameter (scalar function), $\nu = \nu(P_\theta)$
6. The gradient of a scalar function (Influence function)
7. The Riemann curvature tensor.

The total ignorance prior has been written with the wedge product of differential forms (or Clifford algebra) to emphasize the fact that measures on the (oriented) manifold are really totally antisymmetric tensors and therefore their Lie derivatives are well-defined.

There is no technical problem in computing the Lie derivatives of the above list. In fact, ready-made formulas for the items 1,2,5,6 and 7 can be found in any good book on modern geometry. The derivatives of item 4, and a naïve version of item 5, can be computed directly from the defining formula for the Lie derivative of a tensor. However, before rushing to write down the formulas, it would be convenient to subject the posterior to a more detailed examination. One is dealing here not only with the manifold of the model, but also with the manifold of the data space, and a deeper analysis should consider deformations of both of them. It would be very convenient to know how the posterior changes with small *kicks* applied to the model and the data space.

By looking at statistical robustness with this geometric eye, new avenues for the imagination open up, and previous approaches can be seen from a different perspective. Consider, for example, the approach based on the influence function i.e. the Gâteaux derivative of a parameter evaluated at a point mass distribution. The only deformations of the underlying space that can be encoded with this approach are those characterized by vector fields of central forces. Gâteaux derivatives apply to the model a very small class of local *kicks*. Only deformations of the space that send (locally) every probability measure closer to a fix point are considered.

From the more global perspective of information geometry, the Lie derivative, as a technical tool for quantifying robustness, is only one example of the many possibilities that Lie theory can bring to statistical inference. To illustrate the use of some Lie theory in statistics the gaussians and the discrete distributions are introduced below as examples of Lie groups.

4.2 Lie Theory and the Gaussians

The one dimensional gaussians, i.e., the space of all the probability measures with Lebesgue densities given by gaussian curves, not only has the manifold structure of the Lobachevskian plane, but also the algebraic structure of an affine group. The group operation is defined by

$$N\left(\mu,\sigma^2\right) \circ N\left(m,s^2\right) = N\left(\mu+\sigma m, \sigma^2 s^2\right) \quad \text{for } \mu, m \in \mathbb{R} \text{ and } \sigma, s > 0.$$

The identity and inverse elements are given by:

$$e = N(0,1) \qquad N\left(\mu,\sigma^2\right)^{-1} = N\left(-\frac{1}{\sigma}\mu, \frac{1}{\sigma^2}\right) \tag{58}$$

and it is not difficult to show that the group operation and the operation of taking the inverse are C^∞ maps.

It is interesting to note that, by decomposing the group of isometries into one parameter subgroups, the familiar statistical transformations of changing location and scale are automatically generated. To see this, recall that the group of direct isometries of the Lobachevskian plane, is the connected component of the identity of $SO(1,2)$, which is isomorphic to $SL(2,\mathbb{R})/\{\pm 1\}$ which is isomorphic to the group of linear-fractional transformations with coefficients given by the entries of the matrices in $SL(2,\mathbb{R})$. It is worth noting, by passing, that it follows from here that the group of symmetries of the hypothesis space of one dimensional gaussians are nothing but the orthochronous Lorentz transformations of $\mathbb{R}^3_1$ i.e. three dimensional space-time!.

The Lie albegra of the gaussians is then given by $sl(2,\mathbb{R})/\{\pm 1\}$ and the one parameter subgroups of isometries are obtained from the exponential function $\exp(tX)$ with $t \in \mathbb{R}$ and $X \in sl(2,\mathbb{R})/\{\pm 1\}$. The matrices X have zero trace and real entries i.e.

$$X = \begin{pmatrix} a & b \\ c & -a \end{pmatrix} \quad \text{and} \quad X^2 = \lambda I$$

with $a, b, c \in \mathbb{R}$ and $\lambda = (a^2 + bc)$. Therefore, it follows that,

$$e^{tX} = \begin{pmatrix} \cosh(\sqrt{\lambda}t) + \frac{a}{\sqrt{\lambda}}\sinh(\sqrt{\lambda}) & \frac{b}{\sqrt{\lambda}}\sinh(\sqrt{\lambda}) \\ \frac{c}{\sqrt{\lambda}}\sinh(\sqrt{\lambda}) & \cosh(\sqrt{\lambda}t) - \frac{a}{\sqrt{\lambda}}\sinh(\sqrt{\lambda}) \end{pmatrix}. \tag{59}$$

Replacing in (59) different values for a, b and c the following four one-parameter subgroups are obtained:

I. The Group of Location Transformations The generator is the matrix

$$X = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

obtained from (59) when $a = c = 0, b = 1$ and $\lambda = 0$. The group elements are:

$$\exp(tX) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

with corresponding linear-fractional transformation:

$$z' = \frac{1z + t}{0z + 1}$$

or, in terms of the components $z = (\mu, \sigma)$,

$$\begin{aligned} \mu' &= \mu + t \\ \sigma' &= \sigma \end{aligned}$$

with $t \in \mathbb{R}$.

II. The Group of Scale Transformations The generator is the matrix

$$X = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

obtained from (59) when $a = 1, b = c = 0$ and $\lambda = 1$. The group elements are:

$$\exp(tX) = \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$$

with associated linear-fractional transformation:

$$z' = \frac{e^t z + 0}{0z + e^{-t}} = e^{2t} z$$

or, in terms of the components $z = (\mu, \sigma)$,

$$\begin{aligned} \mu' &= \alpha\mu \\ \sigma' &= \alpha\sigma \end{aligned}$$

where $\alpha = e^{2t} > 0$.

III. The Group of Hyperbolic Rotations The generator is the matrix

$$X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

obtained from (59) when $a = 0, b = c = 1$ and $\lambda = 1$. The group elements are:

$$\exp(tX) = \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix}$$

with associated linear-fractional transformation:

$$z' = \frac{\cosh(t)z + \sinh(t)}{\sinh(t)z + \cosh(t)}$$

IV. The Group of Euclidean Rotations The generator is the matrix

$$X = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

obtained from (59) when $a = 0, b = 1, c = -1$ and $\lambda = -1$. The group elements are:

$$\exp(tX) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix}.$$

4.3 The Discrete Distributions as a Lie Group

Consider the hypothesis space of all the non-degenerate probability measures over a finite set. The elements of this set can be parametrized by vectors $p = (p_1, p_2, \ldots, p_k)$ with $p_j > 0$ and $\sum_j p_j = 1$. It is not difficult to see (e.g. using equation (9)) that this hypothesis space is topologically equivalent to a half sphere. A group structure can be defined by the operation:

$$(p_1, p_2, \ldots, p_k) \circ (q_1, q_2, \ldots, p_k) = \frac{1}{\sum_{j=1}^k p_j q_j} (p_1 q_1, p_2 q_2, \ldots, p_k q_k) \tag{60}$$

with identity:

$$e = (\frac{1}{k}, \frac{1}{k}, \ldots, \frac{1}{k}) \tag{61}$$

and inverse element,

$$p^{-1} = \left(\sum_j \frac{1}{p_j} \right)^{-1} (\frac{1}{p_1}, \frac{1}{p_2}, \ldots, \frac{1}{p_k})$$

and the operations of multiplication and inverse are C^∞[4]. Thus, the discrete model with the operation (60), is a Lie group.

This Lie group, let us call it G_k, is isomorphic to the abelian group of positive diagonal metrices with proportional metrices being identified, i.e.,

$$G_k \cong D_k^+ / \propto$$

where, D_k^+ is the group of k by k diagonal metrices with strictly positive entries, and $\propto$ is the equivalence relation:

$$A \propto B \quad \text{iff} \quad \exists c > 0 \quad \text{with } AB^{-1} = cI.$$

What is most remarkable about this, is that (60) is also Bayes theorem[5]; the normalized prior p, operated with the normalized likelihood q, produces the normalized posterior r given by the right hand side of equation (60).

It is not clear what all this implies but it seems worth it to find out.

References

[1] Shun-ichi Amari. *Differential-Geometrical Methods in Statistics*, volume 28 of *Lecture Notes in Statistics*. Springer-Verlag, 1985.

[4] Regular models are smooth Riemannian manifolds relative to the Haussdorf topology of the Hellinger distance.

[5] I recently learned about this from an unpublished manuscript by Michael Hardy a graduate student in statistics at the university of Minnesota

[2] G. Larry Bretthorst. *Bayesian Spectrum Analysis and Parameter Estimation*, volume 48 of *Lecture Notes in Statistics*. Springer-Verlag, 1988.

[3] P. Carnevalli, L. Colleti, and S. Patarnello. Image processing by simulated annealing. *IBM journal of Research and Devolopment*, 29:569–579, 1985.

[4] B.A. Dubrovin, A.T. Fomenko, and S.P. Novikov. *Modern Geometry–Methods and Applications, Part-I*, volume GTM 93 of *Graduate Texts in Mathematics*. Springer-Verlag, 1984.

[5] C.T. Herman. *Image reconstruction from projections. The fundamentals of Computarized Tomography.* Academic Press, 1980.

[6] I.A. Ibragimov and R.Z. Has'minskii. *Statistical Estimation*, volume 16 of *Applications of Mathematics*. Springer-Verlag, 1981.

[7] Harold Jeffreys. *Theory of Probability*. Oxford University Press, 1939.

[8] S. Kullback. *Information Theory and Statistics*. John Wiley, New York, 1959.

[9] T.L. Lai and D. Siegmund, editors. *Herbert Robbins Selected Papers*. Springer-Verlag, 1985.

[10] Carlos C. Rodríguez. The metrics induced by the kullback number. In John Skilling, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1989.

[11] Carlos C. Rodríguez. Objective bayesianism and geometry. In Paul F. Fougère, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1990.

[12] R.D. Rosenkrantz, editor. *E.T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, volume 158. Synthese Library, 1983.

[13] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Trans. on Information Theory*, IT-26:26–37, 1980.

[14] J. Skilling and S.F. Gull. Algorithms and applications. In C. Ray Smith and W. T. Grandy, Jr., editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*. D. Reidel Publishing Company, 1985.

[15] John Skilling. The axioms of maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1987.

[16] John Skilling. Classical Max Ent data analysis. In John Skilling, editor, *Maximum Entropy and Bayesian Methods*. Kluwer Academic Publishers, 1989.