Journal of Convex Analysis
Volume 29 (2022), No. 2, 623–647

Entropic Convex Duality in the Determination of Data-Constrained Kernel-Based Bayes-Jaynes Priors

Richard Le Blanc
FMSS, Université de Sherbrooke, Sherbrooke, Québec, Canada
richard.le.blanc@usherbrooke.ca

Received: January 1, 2021
Accepted: December 31, 2021

Dense datasets produced by high-throughput technologies allow for the objective determination of Bayesian priors when understood as solutions to Fredholm's integral equation. In order to select an appropriate methodology, we conducted a comprehensive review of alternative derivations for the Bayesian prior and posterior update rules comprising: classical iterative alternating algorithms, information theory-motivated variational approaches, and contemporary convex geometry methods. All of these share the key concepts of Kullback-Leibler (KL) relative entropy and of variational minimization. Since the KL relative entropy is a Bregman divergence allowing only linear exploration of tangent hyperplanes of the convex entropy functional, the resulting algorithms – reexpressed in terms of alternating Bregman projections – are necessarily iterative in nature. It has been recently argued that the prior can often only be understood in the context of the model likelihood/kernel, but iterative algorithms obscure the prior dependency on the kernel. In contrast, Jaynes maximal entropy principle yields data-constrained structural priors which make very explicit their kernel dependency, with the modelled density entropic convex dual – the convex analysis equivalent of coefficients from a Fourier decomposition – providing the model kernel weight function. Jaynesian-Bayesian modelling of dense datasets null hypothesis statistical testing (NHST) p-value distributions in terms of the Bayes factor $BF(p)$ allows for a natural extension of the NHST framework. We illustrate all concepts in terms of newly derived analytical expressions for the Fisher-Student-Snedecor's t and F noncentral hypersphere distributions which afford informative representations on compact 2D spaces.

Keywords: Fredholm equation, Kullback-Leibler relative entropy, Bregman divergence, Bregman projection, Mirror update, Bayes-Jaynes Prior.

2010 Mathematics Subject Classification: 45B05, 94A17, 49N15, 62C10, 60E05.

1. Introduction

Historical controversies have beset the elicitation, computation, and use of priors at the core of the Bayes-Laplace framework which is often set in opposition to the Frequentist framework. These controversies are the subject of the compelling narrative *"The theory that would not die"* by McGrayne [32]. In her opus, McGrayne reports on a "*glorious*" profusion of Bayesian theories in the 1960s (McGrayne [32, p. 129]):

> *Versions included subjective, personalist, objective, empirical Bayes (EB), semi-EB, epistemic, intuitionist, logical, fuzzy, hierarchical, pseudo, quasi, compound, parametric, nonparametric, hyperparametric, and non-hyperparametric Bayes.*

ISSN 0944-6532 / $ 2.50

Gelman et al. [18] discussed the relative merits of some related priors, and ultimately argued that the prior can often only be understood in terms of the likelihood function:

> *One might say that what makes a prior a prior, rather than simply a probability distribution, is that it is destined to be paired with a likelihood.*

Using the Fredholm theory kernel terminology, we shall abundantly illustrate this prior-likelihood/kernel pairing by exploiting recently derived analytical kernels for the t and F noncentral hypersphere distributions (Le Blanc [29]) which afford instructive 2D graphics on easily drawn compact domains.

Our starting point will be the classical Bayes-Laplace posterior update rule, recast by Csiszár and Tusnády as an iterative alternating prior-posterior specification algorithm which can objectively determine Bayesian priors associated with dense datasets (Chae et al. [12], Csiszár and Tusnády [13]). Being based on an alternating minimization procedure, their algorithm can be related to variational derivations of the Bayes-Laplace posterior put forward in the 1980s by Williams [42] and Zellner [43]. A key point of the latter developments was the use of the relative entropy measure introduced by Kullback and Leibler [27]. The Kullback-Leibler relative entropy being an example of a Bregman divergence, a bridge is drawn toward contemporary convex analysis. Bregman divergences have been used in inverse problems (Jones and Byrne [24], Le Besnerais [28]), in machine learning (Banerjee et al. [2], Taskar et al. [38], Kulis et al. [26], Vemuri et al. [40]), but their relevance in statistical inference and modelling have not been stressed as much (Grünwald et al. [20], Gneiting and Raftery [19]). The latter remark stands as a principal motivation for the present digression.

Ultimately, the 250 year-old Bayes-Laplace update rule will be argued to naturally arise from the convex geometry of a convex entropy function, negative of Shannon's famed entropy (Shannon [36]). Since the Kullback-Leibler relative entropy is a Bregman divergence which only allows linear exploration of a convex function tangent hyperplanes, classical iterative algorithms necessarily involve descent along iterated tangent hyperplanes toward the solution prior. As a consequence, these algorithms will determine the Bayesian priors, but will do so in an black-box manner with no explicit reference to the model kernels. In contrast, we shall argue that Jaynes maximal entropy (maxent) principle (Jaynes [23], Hestenes [22]) yields data-constrained structural priors which explicitly refer to their model kernels, with the modelled density entropic convex dual – the convex analysis equivalent of coefficients from a Fourier decomposition – providing the model kernel weight function.

Our exposition will begin with a brief review in Section 2 of the analytical t and F noncentral hypersphere distributions which define families of parameterized distributions for which both sample space $\mathcal{X}$ and model parameter space $\mathcal{M}$ are compact and one-dimensional. The resulting compact 2D joint density spaces have been exploited by Le Blanc [29] to provide informative graphics illustrating the differences between the Frequentist and Bayes-Laplace statistical frameworks. All concepts introduced herein will be provided with similar compact graphical representations. A brief exposition of Fredholm's theory of integral equations is given in Section 3.

We then proceed with a review of alternative derivations of the Bayes-Laplace prior and posterior update rules comprising: use of the classical definition of conditional probabilities in Section 4; the Laplacian variational approach à la Williams-Zellner together with its contemporary convex geometry generalization in Section 5; and the Csiszár and Tusnády iterative alternating algorithm in joint density (x, m) spaces in Section 6. Derivation of Jaynes maxent data-constrained structural priors will be carried out through minimization of the Gibbs-Jaynes potential function in Section 7.

A comprehensive Bayesian-Jaynesian graphical analysis of two genomic-scale microarray datasets will be given in Section 8, which includes a brief discussion concerning the definition of noninformative priors when likelihood/kernel-related geometric considerations are taken into account. A succinct review of Bregman divergences, the Kullback-Leibler relative entropy, Bregman projections, and convex duals is provided for convenience in the Appendix. We have tried throughout this manuscript to abide to a notation that aims to be both clear and precise, with our preference leaning toward clarity. Amongst others, we convene hereafter to designate the noncentral hypersphere distributions by the Greek letter υ (upsilon) as in $\upsilon\pi\varepsilon\rho\,\sigma\varphi\alpha\iota\rho\alpha$ (ypersfaíra), that is, hypersphere in English. We hope that the reader will find this compromise agreeable.

2. Noncentral hypersphere t- and F-distributions

The analytical Fisher-Student's noncentral hypersphere t-distribution for the t-statistic

$$t = \sqrt{\nu}\,\frac{\cos\theta}{\sin\theta}, \qquad 0 < \theta < \pi, \tag{1}$$

was shown by Le Blanc [29] to be given by the compact analytical kernel density

$$\upsilon_\nu^t(\theta|\theta_\delta) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})}\ \sin^{\nu-1}\theta \times T_\nu^t(\theta|\theta_\delta), \tag{2}$$

where the multiplicative translation factor $T_\nu^t(\theta|\theta_\delta)$ is given by

$$T_\nu^t(\theta|\theta_\delta) = \frac{1-\cos\theta\cos\theta_\delta}{[1-2\cos\theta\cos\theta_\delta+\cos^2\theta_\delta]^{\frac{\nu+1}{2}}}, \tag{3}$$

and where the trigonometric noncentrality parameter $\cos\theta_\delta$ is given by

$$\cos\theta_\delta = \frac{\delta/\sqrt{\nu}}{\sqrt{(\delta/\sqrt{\nu})^2+1}} = \frac{\delta}{\sqrt{\delta^2+\nu}}, \quad 0 < \theta_\delta < \pi, \tag{4}$$

in terms of the distribution's degree of freedom ν equal to dimension of the hypersphere surface, and the usual noncentrality parameter δ, $-\infty < \delta < \infty$. The distribution is thus characterized by a compact one-dimensional parameter space $0 < \theta_\delta < \pi$, and a compact one-dimensional sampling space $0 < \theta < \pi$. These properties allow for instructive illustrations on compact 2D joint density spaces. When $\delta = 0$, $\cos\theta_\delta = 0$, $T_\nu^t(\theta|\theta_\delta) = 1$, and the noncentral distribution simplifies to the Fisher-Student's central hypersphere t-distribution

$$\upsilon_\nu^t(\theta|\theta_{\delta=0}) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu}{2})}\sin^{\nu-1}\theta, \tag{5}$$

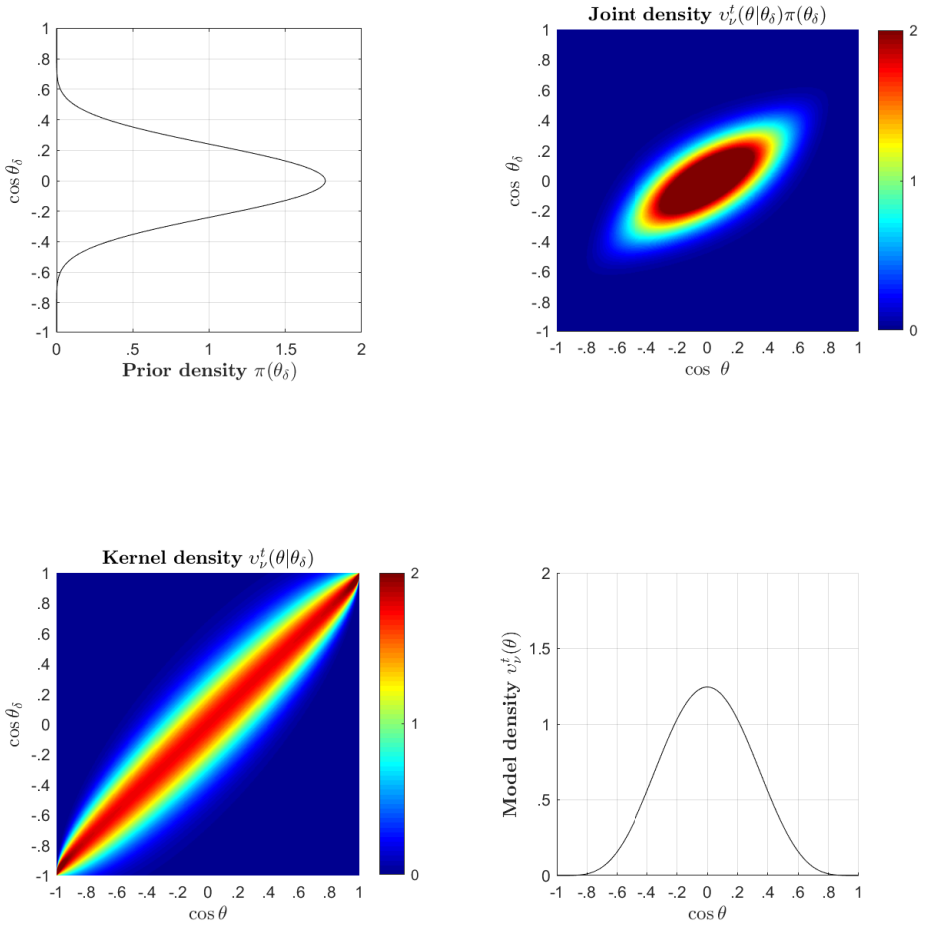

Figure 2.1: **Fisher-Student's noncentral hypersphere t-distribution** for $\nu = 20$. Lower left panel: kernel density $\upsilon_\nu^t(\theta|\theta_\delta)$. Upper left panel: maximal entropy prior density $\pi(\theta_\delta)$ given by Fisher-Students' central t-distribution, marginal density of the joint density to its right. Upper right panel: joint density $\upsilon(\theta, \theta_\delta) = \upsilon_\nu^t(\theta|\theta_\delta)\pi(\theta_\delta)$. Lower right panel: model density $\upsilon_\nu^t(\theta) = \int_{-\pi}^{\pi} \upsilon_\nu^t(\theta|\theta_\delta)\pi(\theta_\delta)\mathrm{d}\theta_\delta$, marginal density of the joint density above it.

which is the projection on the between-class variance unit vector of a uniform unconstrained maximal entropy distribution on the hypersphere $\mathcal{S}^\nu$ (Preda [34], Le Blanc [29]). When $\nu = 2$, one retrieves the well-known formula for the spherical surface element in terms of the polar angle θ. The central hypersphere t-distribution will be invoked as a maximal entropy prior distribution in the following. A graphical introduction to the Fisher-Student's noncentral hypersphere distribution is provided in Figure 2.1.

Similarly, the analytical Fisher-Snedecor's noncentral hypersphere F-distribution for the F-statistic

$$F = \frac{\nu_2}{\nu_1}\frac{\cos^2\theta}{\sin^2\theta}, \qquad 0 < \theta \leq \frac{\pi}{2}, \tag{6}$$

was shown by Le Blanc [29] to be given by the kernel density

$$\upsilon^F_{(\nu_1,\,\nu_2)}(\theta|\theta_\Lambda) = 2\frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\ \cos^{\nu_1-1}\theta\ \sin^{\nu_2-1}\theta \times T^F_{(\nu_1,\,\nu_2)}(\theta|\theta_\Lambda), \tag{7}$$

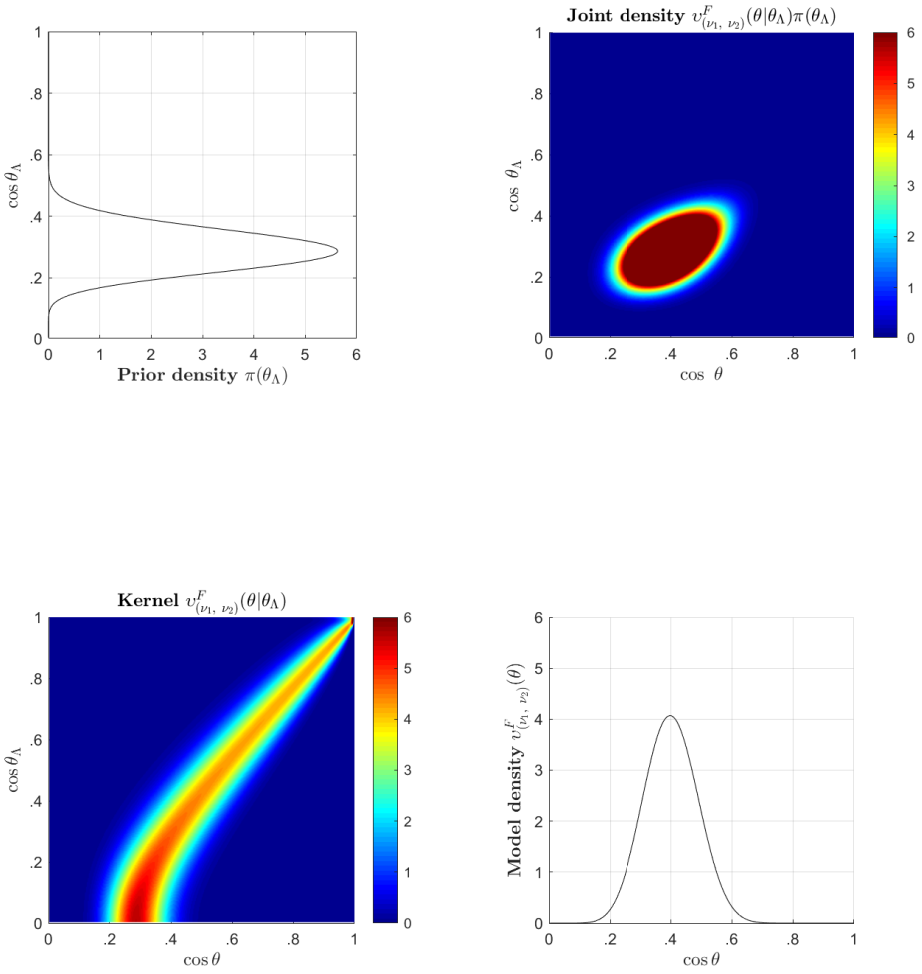

Figure 2.2: **Fisher-Snedecor's noncentral hypersphere F-distribution** for $\nu_1 = 9$, $\nu_2 = 90$. Upper left panel: maximal entropy prior density $\pi(\theta_\Lambda)$ given by the Fisher-Snedecor's central F-distribution, marginal density of the joint density to its right. Upper right panel: joint distribution $v^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)\pi(\theta_\Lambda)$. Lower left panel: kernel density $v^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)$. Lower right panel: model density $v^F_{(\nu_1,\nu_2)}(\theta) = \int_0^{\frac{\pi}{2}} v^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)\ \pi(\theta_\Lambda)\ \mathrm{d}\theta_\Lambda$, marginal density of the joint density above it.

where the multiplicative translation factor $T^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)$ is given by

$$T^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda) = \int_{\psi=0}^{\pi} T^F_{(\nu_1,\ \nu_2)}(\theta,\psi)\ \mathrm{d}\psi, \tag{8}$$

with

$$T^F_{(\nu_1,\ \nu_2)}(\theta,\psi) = \frac{\Gamma(\frac{\nu_1}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{\nu_1-1}{2})}\sin^{\nu_1-2}\psi \times \frac{(1-\cos\theta\cos\theta_\Lambda\cos\psi)}{[1-2\ \cos\theta\cos\theta_\Lambda\cos\psi+\cos^2\theta_\Lambda]^{\frac{\nu_1+\nu_2}{2}}},$$

and where the trigonometric noncentral parameter $\cos\theta_\Lambda$ is given by

$$\cos\theta_\Lambda = \sqrt{\Lambda/(\Lambda+\nu_2)}, \qquad 0 < \theta_\Lambda \leq \frac{\pi}{2}, \tag{9}$$

in terms of the usual noncentral parameter $0 < \Lambda < \infty$.

In an ANOVA, the distribution's parameter ν_1 and ν_2 stand for $K-1$ and $N-K$ respectively, where $K > 2$ is the number of classes, and N is the total number of samples. The distribution is thus characterized by a compact one-dimensional parameter space $0 < \theta_\Lambda \leq \pi/2$ and a compact one-dimensional sampling space $0 < \theta \leq \pi/2$. These properties again allow for instructive illustrations on compact 2D joint density spaces. When $\Lambda = 0$, $\cos\theta_\Lambda = 0$, $T^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda) = 1$, and the noncentral distribution simplifies to the central hypersphere F-distribution

$$\upsilon^F_{(\nu_1,\ \nu_2)}(\theta|\ \theta_{\Lambda=0}) = 2\ \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})}\ \cos^{\nu_1-1}\theta\ \sin^{\nu_2-1}\theta, \tag{10}$$

which is the projection on the between-class variance hyperplane of a uniform unconstrained maximal entropy distribution on the hypersphere. A graphical introduction to the Fisher-Snedecor's noncentral hypersphere distribution is provided in Figure 2.2.

In the null hypothesis statistical testing (NHST) framework, the effect sizes δ or Λ are assumed to be zero under the null hypothesis H_0, and the central hypersphere distributions (5) and (10) stand for the corresponding null distributions. Under the alternative hypotheses H_1 of non-vanishing effect sizes, the corresponding Bayes factors

$$BF = P(\mathcal{D}|H_1)/P(\mathcal{D}|H_0),$$

where $\mathcal{D}$ is the data observed, are given by

$$BF_t(\theta) \ = \ \int_{-\pi}^{\pi} T^t_\nu(\theta|\theta_\delta)\ \pi(\theta_\delta)\ d\theta_\delta, \tag{11}$$

$$BF_F(\theta) \ = \ \int_0^{\pi/2} T^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)\ \pi(\theta_\Lambda)\ d\theta_\Lambda, \tag{12}$$

respectively in terms of the translation factors (3) and (8), and of respective Bayesian priors $\pi(\theta_\delta)$ and $\pi(\theta_\Lambda)$ to be specified. Analytic expressions for the cumulative density functions – thus the null hypothesis statistical testing (NHST) p-values – for both central hypersphere distributions were provided by Le Blanc [29] who showed that the Bayes factors $BF(p)$ can accurately model empirical nonuniform $p-$value distributions. Recall that under the null hypothesis, the p-value distribution is the uninformative uniform distribution $U(0,1)$ on the range value $0 \leq p \leq 1$. In short, Jaynesian-Bayesian modelling of dense datasets NHST $p-$value distributions in terms of the Bayes factor $BF(p)$ allows for a natural extension of the NHST framework. Finally, a local false discovery rate (Efron et al. [16]) can be computed via the expression

$$\text{fdr}(p) = \frac{P(\mathcal{D}|H_0)}{P(\mathcal{D}|H_0) + P(\mathcal{D}|H_1)} = \frac{1}{1+BF(p)}. \tag{13}$$

3. Fredholm theory of integral equations

Given a kernel density $\rho(x|m)$ defined on a sample space $\mathcal{X}$ and parameterized according to coordinates m of a model parameter space $\mathcal{M}$, a generic model density can be written as a mixture or superposition density

$$\rho(x) = \int_{\mathcal{M}} \rho(x|m)\ \pi(m)\ \mathrm{d}m, \tag{14}$$

with $\pi(m)$ standing for a prior density. Given an evidence density $\rho(x)$, we shall seek herein to determine the prior density $\pi(m)$, an inverse problem amounting to solving a Fredholm equation of the 1st kind (Aster et al. [1]). According to Hadamard [21], such problems are usually ill-posed. Consider the linear equation $G\pi = \rho$, where π is to be found in terms of the evidence data ρ. The problem is *well-posed* if the following three conditions are satisfied: for every ρ, there exists a solution π; this solution is unique; and the inverse operator G^{-1} is continuous. Problems which do not satisfy all of these conditions are said to be *ill-posed.* For example, discretizing the hypersphere kernel density (2) yields a highly singular matrix which will, when inverted, amplify the noise inherent to any dataset. One way to resolve this ill-posedness is to use regularization methods (Tikhonov and Arsenin [39]). More recently, Bayes-Laplace methods have been advocated (Aster et al. [1], Calvetti and Somersalo [9], Stuart [37]). We shall follow suit with a Bayesian approach in the following. Recall that the Bayesian parametric framework is characterized by attribution of probability densities to both sample coordinates x and model parameters m, and that the celebrated Bayes-Laplace rule for the posterior conditional density $\rho(m|x)$ specification

$$\rho(m|x) = \frac{\rho(x|m)\ \pi(m)}{\rho(x)} \tag{15}$$

as a function of the prior $\pi(m)$ density is a simple and direct consequence of the definition of joint and conditional probabilities

$$e(x,m) := \rho(m|x)\rho(x) = \rho(x,m) = \rho(x|m)\pi(m) := h(x,m) \tag{16}$$

for two continuous variables. Note that, in Gelman et al. [18] terminology, the joint density $e(x,m) = \rho(m|x)\rho(x)$ refers to the posterior $\rho(m|x)$ predictive distribution, while the joint density $h(m,x) = \rho(x|m)\pi(m)$ refers to the prior $\pi(m)$ predictive distribution. As these authors argued, prior predictive utility functions assess the model hypothesis joint density $h(x,m)$ by how well it reproduces the data according to the marginal density $\rho(x) = \int h(x,m)\ \mathrm{d}m$, while posterior predictive utility functions inform the prior by considering the marginal density $\pi(m) = \int e(m,x)\ \mathrm{d}x$ of the evidence joint density $e(m,x)$.

4. Classical alternating prior/posterior specification procedure

In the parametric Bayesian framework, the inverse problem (14) has been historically solved via an iterative alternating algorithm. See Chae et al. [12] for a thorough contemporary digression on the subject. First, invoking definition of the Bayes-Laplace posterior, the $(k+1)^{\text{th}}$ update for the posterior distribution is given by

$$\rho^{(k+1)}(m|x) = \frac{\rho(x|m)\ \pi^{(k)}(m)}{\rho^{(k)}(x)} \tag{17}$$

where

$$\rho^{(k)}(x) = \int_{\mathcal{M}} \rho(x|m)\ \pi^{(k)}(m)\ \mathrm{d}m,$$

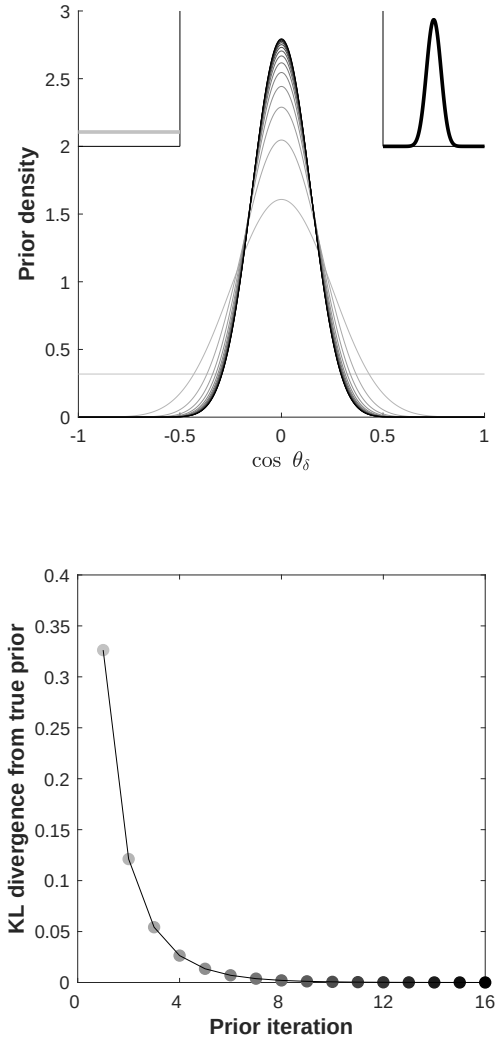

Figure 4.1: **Prior update iterative algorithm convergence in terms of the KL divergence** for the Fisher-Student's noncentral hypersphere t-distribution with $\nu = 50$ and a postulated maximal entropy prior. Upper panel: prior iteration. The flat seed prior $\pi^{(0)}(\theta_\delta) = U(0,\pi)/\pi$ is drawn to scale in the left medallion, while the target maximal entropy prior defined in terms of the Fisher-Student's central hypersphere distribution (5) is drawn to scale in the right medallion. Lower panel: Kullback-Leibler divergence $KL(\pi, \pi^{(k)})$ between the target prior π and the iterated prior $\pi^{(k)}$. The iterative algorithm necessitates about ten iterations before the estimate $\pi^{(k)}(\theta_\delta)$ finally converges unto the target maximal entropy prior in a KL-sense. See text for details. See also Figure 6.1.

and where it is assumed that the prior density $\pi^{(k)}(m)$ has been determined in the previous iteration. The prior $\pi^{(k)}(m)$ is thereafter updated by computing the marginal density of the joint density obtained by weighing the posterior distribution (17) with the evidence density $\rho(x)$:

$$\pi^{(k+1)}(m) = \int_{\mathcal{X}} \rho^{(k+1)}(m|x)\ \rho(x)\ \mathrm{d}x = \pi^{(k)}(m) \int_{\mathcal{X}} \rho(x|m) \frac{\rho(x)}{\rho^{(k)}(x)}\ \mathrm{d}x. \tag{18}$$

In the present considerations, this amounts to weighing the posterior distribution with a continuous dataset. In real word cases, as in Section 8, the continuous

data distribution will be approximated by dense granular data distributions. The iterative process starts with a normalized seed prior $\pi^{(0)}(m)$. Integrating both side of the latter equality with respect to the variable m, and inverting the order of integration over x and m, Fubini's theorem guarantees that $\pi^{(k+1)}(m)$ will be a normalized probability density. If

$$\lim_{k\to\infty} \rho^{(k)}(x) \to \rho(x),$$

the last integral will converge to one since the model distributions are assumed to be properly normalized with

$$\int \rho(x|m)\mathrm{d}x = 1$$

for all parameters m allowed by the model, and the iterated prior should become stationary at $\lim_{k\to\infty} \pi^{(k)}(m) = \pi^{(\infty)}(m)$. Convergence of the iterative alternating algorithm is discussed by Chae et al. [12]. Convergence of the algorithm is depicted in Figure 4.1 for models based on the Fisher-Student's noncentral hypersphere t-distribution (2).

Most introductory material to parametric Bayesian theory usually describe the Bayesian posterior density $\rho(m|x)$ as a prior-dependent conditional probability enabling one to attribute credible intervals to the parameters m after observing x. When dense datasets allow determination of empirical priors via the iterative procedure described above, it should be ascertained that the priors correspond to stationary solutions of the iterative procedure before using the posteriors to such ends. The latter observation begs answer to the question as to why such iterative procedures are necessary in classical parametric Bayesian theory, and whether they could be circumvented. It will be shown in the next sections that this need for iterations stems from the convex geometry underlying classical parametric Bayesian theory.

5. Bayesian updates as an optimization problem

The Bayes-Laplace posterior updating rule (15) was originally argued to stem from the definition of joint and conditional probabilities, and from the fact that both sample coordinates and parameters are provided with probability densities in the Bayes-Laplace framework. Williams [42] argued that the Bayes-Laplace rule was a special case of a *principle of minimum information* which he defined (using his notation) as follows:

> *Given the prior distribution P^o, the probability distribution P appropriate to a new state of information is one that minimises $I(P, P^o)$ subject to whatever constraints the new information imposes.*

In modern parlance, $I(P, P^o)$ should read as the Kullback-Leibler divergence (47) between P and P^o. Williams argued that this principle should not be considered a principle for setting up prior distributions, but rather as a general principle of *probability dynamics:*

> *It seeks to answer the question how to modify a probability distribution, in the light of new information, in the most conservative way. It is inapplicable in the absence of a prior distribution.*

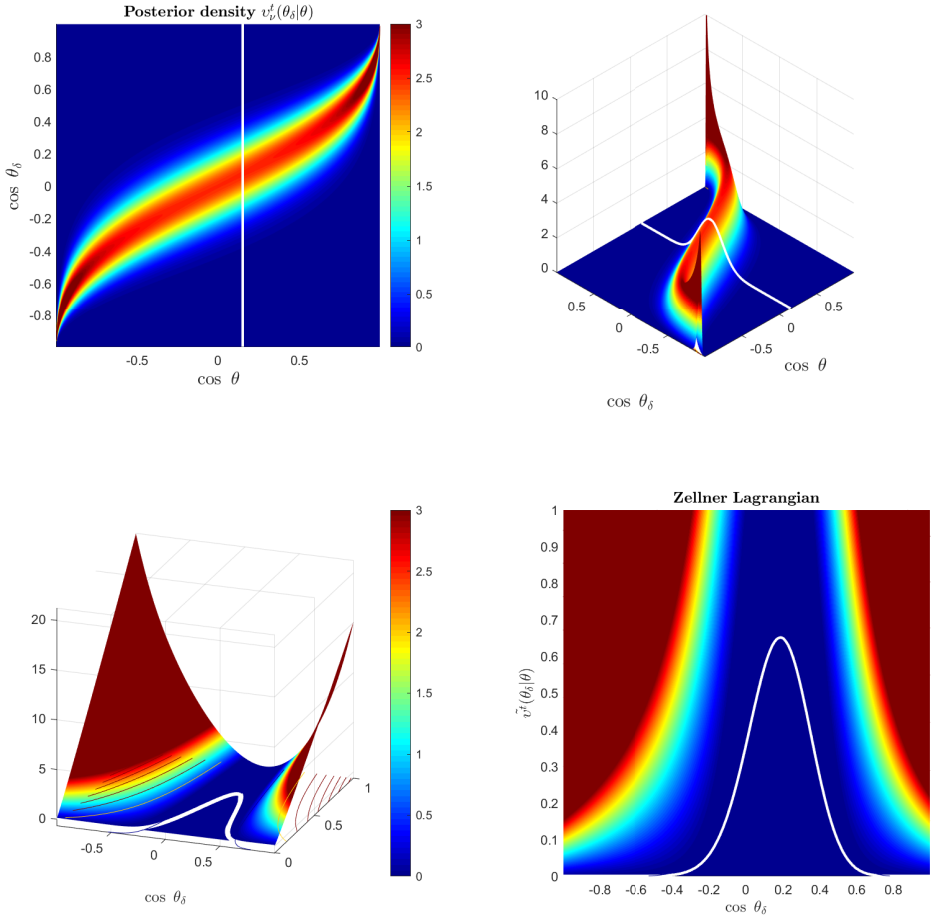

Figure 5.1: **Fisher-Student's noncentral hypersphere t-distribution posterior.** Upper left panel: posterior density $\upsilon_\nu^t(\theta_\delta|\theta)$ for the maximal entropy prior illustrated in Figure 2.1. Upper right panel: 3D view of the posterior density. Any posterior segment running the length of the parameter space θ_δ for fixed sample space θ as drawn in the two upper panels lies in a Zellner Lagrangian through. Lower left panel: 3D view of the Lagrangian well. Lower right panel: Lagrangian well viewed from above. In the latter panel, the posterior $\tilde{\upsilon}_\nu^t(\theta_\delta|\theta)$ has yet to be normalized since the Lagrangian drawn here does not include the normalization term multiplying the Lagrange parameter λ in (19). The normalization factor is a multiplicative scale factor which will preserve the shape of the posterior distribution.

In a similar vein, Zellner [43] invoked an *information conservation processing* rule to introduce (in our notation) the Lagrangian

$$L(\phi, \lambda) = KL\left(\phi, \pi\right) - \int_{\mathcal{M}} \log \rho(x|m)\ \phi(m)\ \mathrm{d}m + \lambda \left(\int_{\mathcal{M}} \phi(m)\ \mathrm{d}m - 1 \right), \qquad (19)$$

where the Lagrangian parameter λ term is introduced to insure normalization of the solution $\phi(m)$, and to solve for the posterior by solving the variational problem

$$\arg\min_{\phi,\ \lambda} L(\phi, \lambda). \qquad (20)$$

See also Bissiri et al. [5].

Performing functional derivatives with respect to ϕ and λ, one finds

$$\begin{aligned} \log \rho(m|x) &= \log[\rho(x|m)\pi(m)] - \lambda \\ &= \log[\rho(x|m)\pi(m)] - \log\left(\int_{\mathcal{M}} \rho(x|m)\pi(m)\ \mathrm{d}m\right), \end{aligned} \tag{21}$$

that is, the logarithm of the Bayes-Laplace posterior (15). Since the KL-divergence term alone would be minimized by setting $\phi = \pi$, while the log-likelihood term $\log \rho(x|m)$ alone would set ϕ as a maximum likelihood estimator, the Bayes-Laplace posterior (15) can be understood from a variational point of view as a trade-off between staying close to the prior and fitting of the data model as provided by the kernel $\rho(x|m)$. See Figure 5.1 for a graphical illustration of the optimization problem in terms of the Lagrangian (19) for models based on the noncentral hypersphere t-distribution (2). See also Reid et al. [35], Walker [41] and Zhu et al. [44]. Note that, as layed out, the variational problem (20) seeks a density $\phi(m)$ in model space $\mathcal{M}$ but yields a posterior $\rho(m|x)$ in joint space $\mathcal{X} \times \mathcal{M}$, and that the Lagrange parameter λ is not a scalar but a function $\lambda(x)$ on sample space $\mathcal{X}$. These notational and computational objective ambiguities will be addressed in the following.

While facilitating interpretation of the Bayes-Laplace posterior update rule from an information theoretic principled point of view, the present variational approach seems limited to reproduce the latter rule without providing an analytical framework which could suggest generalizations. It can be argued that, in terms of the convex geometry concepts concisely reviewed in the Appendix, the Bayes-Laplace posterior update rule (21) is a special case of the generic convex geometry optimization problem

$$\arg\min_{\mu} \left[D_\Phi(\mu, \pi) + \mathbb{E}_\mu\left(L(x|m)\right)\right], \text{ subject to } \mu \in \mathcal{C}, \tag{22}$$

where $D_\Phi(\mu, \pi)$ is the Bregman divergence (45) generated by a convex function Φ of interest (here the convex entropy function S defined in (46)), $L(x|m)$ is a convex loss function (here $-\log \rho(x|m)$), and $\mathcal{C}$ a constraint set (here the probability density normalization constraint). Solution to the optimization problem (22) is provided by the mirror update

$$\mu(m|x) = \nabla\Phi^*\left[\nabla\Phi(\pi) - L(x|m)\right], \text{ subject to } \mu \in \mathcal{C}, \tag{23}$$

which maps a point π from the primal function space to the linear dual tangent space via the gradient operator $\nabla\Phi$ (here $\nabla S = \log$), performs a gradient update in the dual space via the additive term L, and maps the new optimal point back in the primal function space via the dual operator $\nabla\Phi^*$ (here $\nabla S^* = \exp$). See Reed et al. [35]. The new point μ in the primal space might be outside of the constraint set $\mathcal{C}$, in which case it should be projected back into the constraint set via the Bregman projection $\mathbb{P}^{\mathcal{C}}_{\Phi}(\mu)$ defined in (48). Diagrammatically, we have

$$\begin{array}{ccc} \pi(m) & \xrightarrow{\nabla\Phi} & \nabla\Phi(\pi(m)) \\ & & \Big\downarrow {\scriptstyle -L(x|m)} \\ \mu(m|x) & \xleftarrow{\mathbb{P}^{\mathcal{C}}_{\Phi}\nabla\Phi^*} & \nabla\Phi(\pi(m)) - L(x|m) \end{array} \tag{24}$$

When Φ is given by the entropy $S(\pi)$ (46) and the loss function given by $L(x|m) = -\log \rho(x|m)$, we have from equations (23), (52) and (56) that

$$\rho(m|x) = \nabla S^*(\log \pi(m) + 1 + \log \rho(x|m)) = \frac{\rho(x|m)\ \pi(m)}{\int_{\mathcal{M}} \rho(x|m)\ \pi(m)\ \mathrm{d}m},$$

that is, we retrieve the Bayes-Laplace posterior update rule (15) via a mirror update. The 250 year-old Bayes-Laplace posterior update rule is thus rooted in the convex geometry of the entropy function $S(\pi)$.

While more enlightening, diagram (24) is still not entirely satisfactory. First, it does not form a closed loop: missing is an up-arrow on the left which would upgrade the posterior distribution $\mu(m|x)$ to an updated model space prior distribution $\pi(m)$. Second, it alternates between the model parameter space $\mathcal{M}$ and the joint model-sampling space $\mathcal{X} \times \mathcal{M}$, while our stated goal is a prior $\pi(m)$ in the model parameter space. Third, the loss term $L(x|m)$ is introduced somewhat arbitrarily without any weighing factor (learning rate). Fourth, since the Kullback-Leibler relative entropy is a Bregman divergence which only allow linear exploration of the convex entropy S tangent hyperplanes, prior specification procedures utilizing such divergence necessarily entail descent along iterated tangent hyperplanes toward the solution, and such iterative procedures are not providing us with a global solution pairing model kernels with solution priors.

6. Csiszár-Tusnády's update rules in joint density space (m, x)

The inner workings of the iterative alternating posterior/prior determination algorithm discussed in Section 4 can be better apprehended in terms of the joint densities $e(m, x)$ and $h(m, x)$ introduced at the end of Section 3. First, introduce on $\mathcal{X} \times \mathcal{M}$ the two sets of joint densities

$$\mathcal{E} = \{e(x, m) \mid \int e(x, m)\ \mathrm{d}m = \rho(x)\}, \tag{25}$$

$$\mathcal{H} = \{h(x, m) \mid h(x, m) = \rho(x|m)\tilde{\pi}(m)\}, \tag{26}$$

where $\tilde{\pi}$ is an arbitrary density on the model space $\mathcal{M}$. Note that the sets $\mathcal{E}$ and $\mathcal{H}$ define two constraint sets in the spirit of the Bregman projections (48). Specify thereafter the two joint densities

$$e^{(k)}(x, m) = \rho^{(k)}(m|x)\ \rho(x) \in \mathcal{E}, \tag{27}$$

$$h^{(k)}(x, m) = \rho(x|m)\ \pi^{(k)}(m) \in \mathcal{H}, \tag{28}$$

which refer to the two alternative ways of expressing the joint density $\rho(x, m)$ in terms of the conditional densities in (16). In Figure 6.1, these joint densities are drawn in the second and fourth column, respectively. Using the objective function defined by the joint densities' Kullback-Leibler relative entropy

$$KL(e, h) = \int_{\mathcal{X}} \int_{\mathcal{M}} e(x, m) \log \frac{e(x, m)}{h(x, m)}\ \mathrm{d}m\ \mathrm{d}x, \tag{29}$$

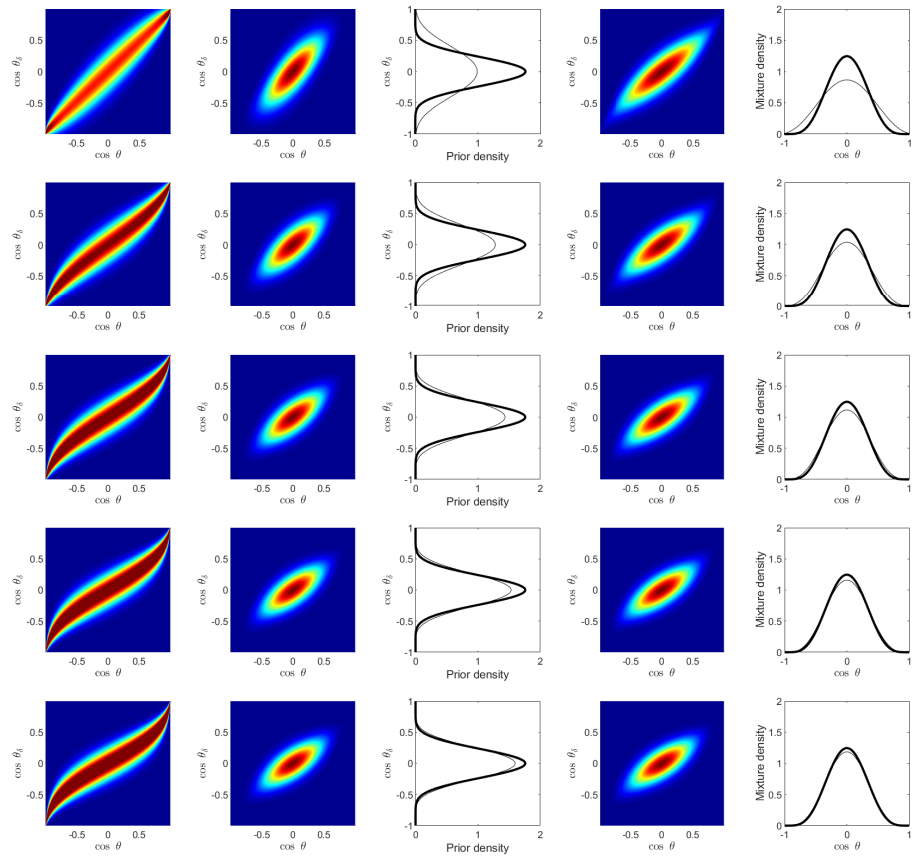

Figure 6.1: **Alternating prior/posterior specification procedure** for the Fisher-Student's noncentral hypersphere t-distribution. Here, the index k runs as $k = \{1, 2, ..., 5\}$ from top to bottom. The seed prior $\pi^{(0)}(\theta_\delta)$ is the flat prior $\pi^{(0)}(\theta_\delta) = U(0, \pi)/\pi$, and the algorithm starting point is the joint density $h^{(0)}(\theta, \theta_\delta) = \upsilon_\nu(\theta|\theta_\delta)\ \pi^{(0)}(\theta_\delta)$. First column: posterior density $\upsilon_\nu^{(k)}(\theta_\delta|\theta)$ according to Bayes-Laplace rule (15). Second column: joint density $e^{(k)}(\theta, \theta_\delta) = \upsilon_\nu^{(k)}(\theta_\delta|\theta)\upsilon_\nu(\theta)$, where $\upsilon_\nu(\theta)$ is the target mixture density. Third column: prior $\pi^{(k)}(\theta_\delta)$ computed as the vertical marginal (18) of the previous column joint density. The thick line is the target maximal entropy prior $\pi(\theta_\delta)$. Fourth column: joint density $h^{(k)}(\theta, \theta_\delta) = \upsilon_\nu(\theta|\theta_\delta)\pi^{(k)}(\theta_\delta)$. Fifth column: mixture density $\upsilon_\nu^{(k)}(\theta) = \int_{-\pi}^{\pi} \upsilon_\nu(\theta|\theta_\delta)\ \pi^{(k)}(\theta_\delta)\ \mathrm{d}\theta_\delta$, computed as the horizontal marginal density of the previous column joint density. The thick line is the target evidence mixture density $\upsilon_\nu(\theta) = \int_{-\pi}^{\pi} \upsilon_\nu(\theta|\theta_\delta)\ \pi(\theta_\delta)\ \mathrm{d}\theta_\delta$. See text for details.

Csiszár and Tusnády [13] argued that

$$e^{(k+1)}(x, m) = \arg\min_{e \in \mathcal{E}} KL(e, h^{(k)}) \doteq \mathbb{P}_S^{\mathcal{E}}(h^{(k)}), \tag{30}$$

$$h^{(k+1)}(x, m) = \arg\min_{h \in \mathcal{H}} KL(e^{(k+1)}, h) \doteq \mathbb{P}_S^{\mathcal{H}}(e^{(k+1)}), \tag{31}$$

iteratively. Note that these two optimization problems have been compactly and conveniently rewritten in terms of S entropy-generated Bregman projections, with implicit concurrent projections on the set of normalized joint distributions. The iteration starts with $h^{(0)}(x, m) = \rho(x|m)\ \pi^{(0)}(m)$, where $\pi^{(0)}(m)$ is the seed prior, and where $\rho(x|m)$ is the model kernel density under the Fredholm integral in (14).

The minimizing densities are easily computed: we find

$$e^{(k+1)}(x,m) = \mathbb{P}_S^{\mathcal{E}}(h^{(k)}) = \frac{\rho(x|m)\ \pi^{(k)}(m)}{\int_{\mathcal{M}} \rho(x|m)\ \pi^{(k)}(m)\ \mathrm{d}m}\ \rho(x) = \rho^{(k+1)}(m|x)\ \rho(x),$$

$$h^{(k+1)}(x,m) = \mathbb{P}_S^{\mathcal{H}}(e^{(k+1)}) = \rho(x|m) \int_{\mathcal{X}} \rho^{(k+1)}(m|x)\ \rho(x)\ \mathrm{d}x = \rho(x|m)\ \pi^{(k+1)}(m),$$

that is, we reproduce the iterative alternating posterior and prior specifications (17) and (18), but in terms of alternating Bregman projections (48) between joint densities. The iteration procedure is depicted in Figure 6.1 for models based on the noncentral hypersphere t-distribution. When considering the densities drawn in the latter figure, the iterated joint density in the second column is closest in a KL-sense to the iterated joint density in the fourth column of the preceding line while obeying its defining constraints, while the iterated joint density in the fourth column is closest in a KL-sense to the iterated joint density in the second column of the same line while obeying its defining constraints. Diagrammatically, the Csiszár-Tusnády iterative alternating algorithm can be simply reexpressed as the closed loop diagram

$$\mathbb{P}_S^{\mathcal{E}}(h) \rightleftarrows \mathbb{P}_S^{\mathcal{H}}(e) \tag{32}$$

in terms of alternating Bregman projections, with an implicit running index k incremented for each complete loop around. See also Dykstra [15], Censor and Reich [11] and Bauschke and Lewis [4]. The iteration can be stopped when the joint densities $e(x,m)$ and $h(x,m)$ become equal and stationary in a KL sense, as discussed by Chae et al. [12]. The alternating Bregman projection algorithm thus depicts a closed loop lying entirely in the joint space $\mathcal{X} \times \mathcal{M}$, and therefore addresses two of the four issues discussed at the end of Section 5. We shall see in the next section how the Gibbs-Jaynes maxent framework satisfactorily resolves all four issues while providing us with maximal entropy data-constrained structural priors.

7. Jaynes structural data-constrained maxent prior

Jaynes maximum entropy (maxent) principle (Jaynes [23], Hestenes [22], Mana [31]) states that parametric densities π of ill-posed problems should be attributed maximal entropy densities while being consistent with problem-specific functional constraints of the type

$$F_c(x) = \int_{\mathcal{M}} F_c(x,m)\ \pi(m)\ \mathrm{d}m, \quad c = 1, \ldots, C. \tag{33}$$

For the Bayesian prior determination problem, the parametric model kernel density function $\rho(x|m)$ provides the problem's unique functional constraint (14). Following Jaynes prescription, define the Lagrangian

$$L_\rho(\pi, \lambda) = -KL(\pi, \pi_o) + \int_{\mathcal{X}} \lambda(x) \left(\int_{\mathcal{M}} \rho(x|m)\pi(m)\ \mathrm{d}m - \rho(x) \right) \mathrm{d}x, \tag{34}$$

where the subscript ρ should be understood as a shorthand for the evidence density $\rho(x)$, where $\pi_o(m)$ is a reference prior distribution introduced to ascertain coordinate independence of the problem formulation, where $\lambda(x)$ is a Lagrange multiplier

introduced to impose the constraint provided by equation (14), and where $\pi(m)$ is implied to be a normalized prior distribution in model $\mathcal{M}$-space.

The primal problem is defined as the search for the saddle point of the concave function defined by the Lagrangian (34) which defines Jaynes maximal entropy

$$S_J(\rho) = \sup_{\pi,\ \lambda} L_\rho(\pi, \lambda), \tag{35}$$

a constrained optimization problem. The saddle point is obtained by nullifying the Lagrangian functional derivatives with respect to both π and λ : one obtains

$$0 = -\log \pi(m) + \log \pi_o(m) + \int_{\mathcal{X}} \lambda(x)\rho(x|m)\ \mathrm{d}x, \tag{36}$$

$$0 = \int_{\mathcal{M}} \rho(x|m)\pi(m)\ \mathrm{d}m - \rho(x). \tag{37}$$

Setting the reference prior π_o to a uniform prior, Jaynes prior and the mixture model can be, in terms of the partition function

$$Z(\lambda) = \int_{\mathcal{M}} \exp\left(\int_{\mathcal{X}} \lambda(x)\rho(x|m)\ \mathrm{d}x\right) \mathrm{d}m, \tag{38}$$

compactly reexpressed as

$$\pi(m) = \frac{1}{Z(\lambda)} \exp\left(\int_{\mathcal{X}} \lambda(x)\rho(x|m)\ \mathrm{d}x\right), \tag{39}$$

$$\begin{aligned} \rho(x) &= \frac{\delta}{\delta\lambda} \log Z(\lambda) \\ &= \frac{1}{Z(\lambda)} \int_{\mathcal{M}} \rho(x|m) \exp\left(\int_{\mathcal{X}} \lambda(x)\rho(x|m)\ \mathrm{d}x\right) \mathrm{d}m \\ &= \int_{\mathcal{M}} \rho(x|m)\pi(m)\mathrm{d}m. \end{aligned} \tag{40}$$

Equation (36) should be compared to equation (24): in the latter, the logarithm of the model kernel $\rho(x|m)$ is added as a loss term to the logarithm of the prior to be updated, yielding the Bayesian posterior (15), a conditional probability density residing in the joint $\mathcal{M} \times \mathcal{X}$ space; in the former, it is the integral $\int_{\mathcal{X}} \lambda(x)\rho(x|m)\ \mathrm{d}x$ that is added to the logarithm of the prior to be updated, yielding the structural Jaynes prior (39), a density residing in model space $\mathcal{M}$ as desired. The qualifier structural is meant here to stress the fact that the model kernel $\rho(x|m)$ is at the core of the prior expression (39).

We now proceed to the dual minimization problem by defining the Gibbs-Jaynes potential function

$$G_\rho(\lambda) = \log Z(\lambda) - \int_{\mathcal{X}} \lambda(x)\rho(x)\ \mathrm{d}x \tag{41}$$

which is a convex function of the Lagrange parameter λ. By convex duality, its minimum equates Jaynes maximal entropy (35), that is, we have

$$\inf_{\lambda} G_\rho(\lambda) = \inf_{\lambda} \left[\log Z(\lambda) - \int_{\mathcal{X}} \lambda(x)\rho(x)\ \mathrm{d}x\right] = \sup_{\pi,\ \lambda}\ L_\rho(\pi, \lambda) = S_J(\rho). \tag{42}$$

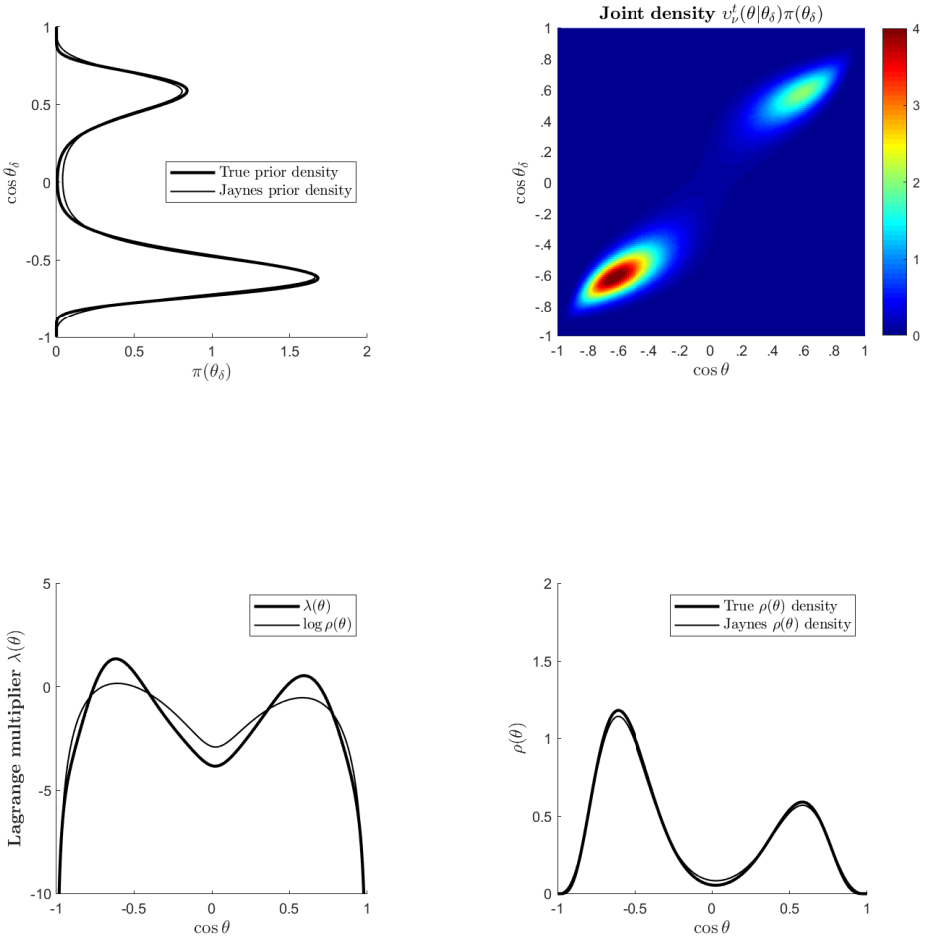

Figure 7.1: **Jaynes prior** for a mixture of noncentral hypersphere t-distributions, with a target bimodal prior composed of the normalized sum of two such distributions. Left upper panel: Jaynes prior compared to the target prior, marginal density of the joint density to its right. Right upper panel: joint density $\upsilon_\nu^t(\theta|\ \theta_\delta)\pi(\theta_\delta)$. Left lower panel: Gibbs-Jaynes model Lagrange multiplier $\lambda(\theta)$ compared to the seed Lagrange multiplier $\log\rho(\theta)$. Right lower panel: Gibbs-Jaynes model density compared to the target density, marginal density of the joint density above it. See text for details.

By definition, the left hand side of the latter equality is the negative Fenchel-Legendre transformation of $\log Z(\lambda)$, implying that $\rho(x)$ and $\lambda(x)$ are convex duals (Lucet [30]). In convex analysis, the Legendre-Fenchel transform stands for a natural counterpart to the Fourier transform. Paraphrasing Reid et al. [35], we shall refer to the convex duals $(\rho(x), \lambda(x))$ as *entropic convex duals.* Note that the dual minimization problem (42) is unconstrained and concerns only determination of the Lagrange multiplier $\lambda(x)$. By involution, $\log Z(\lambda)$ is the negative Legendre transformation of the negative Jaynes entropy, implying

$$\log Z(\lambda) = \inf_\rho \left[-S_J(\rho) - \int_{\mathcal{X}} \rho(x)\lambda(x) \right], \tag{43}$$

$$\lambda(x) = -\frac{\delta}{\delta\rho} S_J(\rho). \tag{44}$$

Through Jaynes maximal entropy formalism, specification of the structural prior (39) is thus formally given in model $\mathcal{M}$ space, the sampling space $\mathcal{X}$ dependency being absorbed by the integrals over $\mathcal{X}$ in the prior (39), the partition function (38), and the Gibbs-Jaynes potential (41), with the latter including the density $\rho(x)$ to be modelled. As such, the Gibbs-Jaynes formalism resolves all notational and computational ambiguities of the convex optimization mirror update procedure discussed at the end of Section 5, with no intermediate Bayesian posterior to compute nor Bregman divergence-induced iterative loop to close. Note that there is still an iterative procedure involved, but it is now implicit in the unconstrained dual minimization problem (42).

In practical terms, the unconstrained dual minimization problem (42) will be solved using one's numerical optimization algorithm of choice. Such algorithms usually require a seed Lagrange multiplier $\lambda(x)$. For the noncentral hypersphere t- and F-distributions (2) and (7) used here as model examples, it should be noted that their respective kernels behave like the diagonal Dirac delta functions $\delta(\theta - \theta_\delta)$ and $\delta(\theta - \theta_\Lambda)$, respectively, in the limit of large samples. In these limits, it is easy to deduce that the prior density π should equate the evidence density ρ, and that the Lagrange multiplier $\lambda(\theta)$ should be well approximated by $\lambda(\theta) \sim \log \rho(\theta)$. See Figure 7.1 for a nontrivial example in terms of the Fisher-Student's noncentral hypersphere t-distribution with a bimodal prior. All dual optimization problems considered herein were numerically solved using the MATLAB® subroutine *fminunc*, with $\log \rho(\theta)$ as seed for the Lagrange multiplier $\lambda(\theta)$. See also Lucet [30].

8. Genomic-scale dataset Bayesian-Jaynesian analysis

We first revisit in this section the cancerous vs normal head and neck tissue microarray dataset discussed by Le Blanc [29] in order to compute its empirical prior, and compare it to the central distribution maximal entropy prior postulated therein. The dataset comprises the differential expression for $11,302$ genes pertaining for $N = 44$ tissue samples ($\nu = 42$). We have plotted in Figure 8.1 Jaynes prior which is seen to be well approximated by the central distribution maximal entropy prior. It is also seen that Jaynes prior accurately models the dataset when paired with its kernel, and that the Bayes factor $BF_t(p)$ (11) fits very well the empirical nonuniform p-value distribution generated by a NHST analysis. The Bayesian local false discovery rate $\mathrm{fdr}(p)$ defined in (13) is $\sim 10\%$ at $p = .01$, and $\sim 1\%$ at $p = 10^{-6}$, which allows one to address in simple terms the multiple hypothesis testing problem (Dudoit et al. [14]). The alternating Bregman projection procedure expounded in Section 4 yields a nearly identical prior (data not shown), but necessitates about ten iterations to achieve convergence of its prior estimate.

We analyzed in a similar fashion a landmark multiclass microarray dataset pertaining to four small round blue cell tumor (SRBCT) types of childhood Khan et al. [25]. The dataset comprises $K = 4$ classes (Ewing's tumor, neuroblastoma, Burkitt's lymphoma, and rhabdomyosarcoma) of respective class cardinalities 29, 11, 18, and 25 totaling $N = 83$ samples ($\nu_1 = 3, \nu_2 = 79$), with 2308 genes interrogated across the four tumor types. A full Bayesian analysis in terms of the noncentral hypersphere F-distribution is carried out in Figure 8.2. Again, it is seen that Jaynes prior accurately models the dataset when paired with its kernel, while the Bayes factor

$BF_F(p)$ fits very well the empirical nonuniform p-value distribution generated by a NHST analysis. The Bayesian local false discovery rate $\text{fdr}(p)$ in this case is $\sim 10\%$ at $p = .01$, and $\sim 1\%$ at $p = 10^{-4}$, once again allowing one to address in simple terms the multiple hypothesis testing problem.

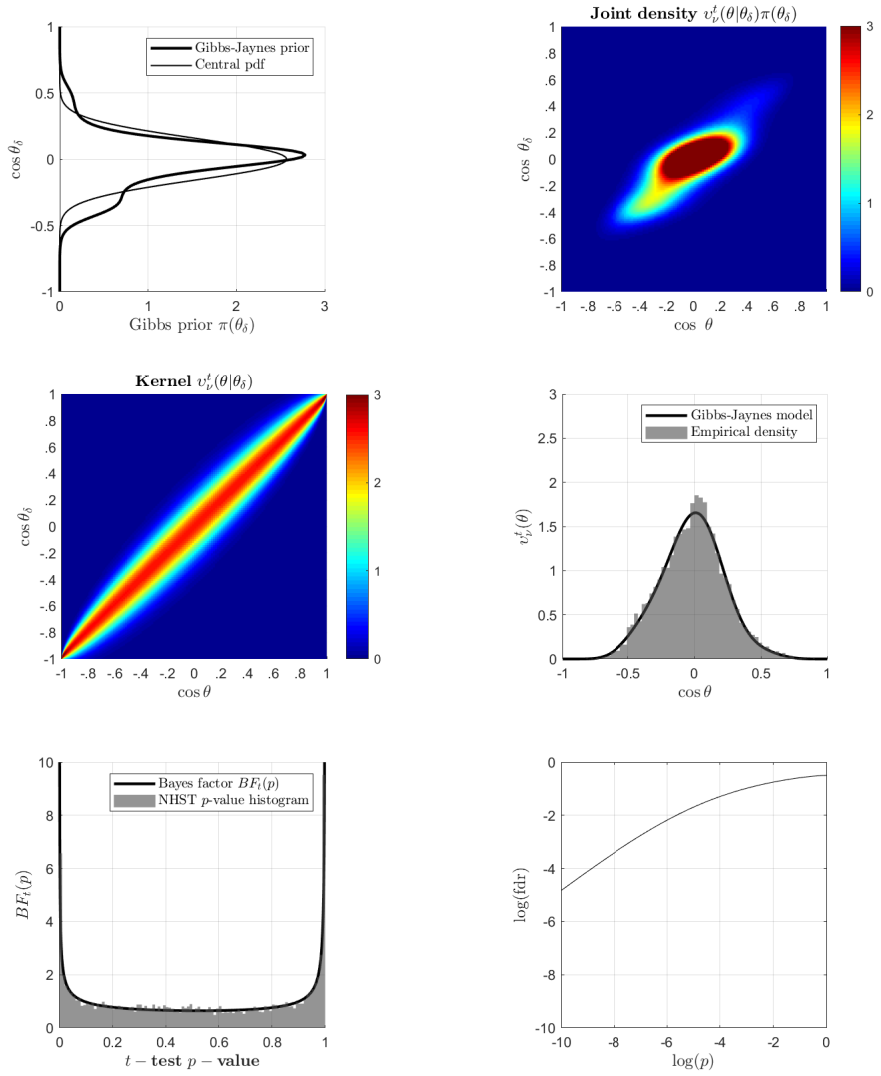

Figure 8.1: **Gibbs-Jaynes model** for a microarray dataset interrogating 11,302 genes while comparing cancerous versus normal head and neck tissue via a two-sample t-test for $N = 44$ ($\nu = 42$) tissue samples. Upper left panel: Jaynes prior density $\pi(\theta_\delta)$, marginal density of the joint density to its right, compared to the maximal entropy prior (5). Upper right panel: corresponding joint density $v_\nu^t(\theta|\theta_\delta)\ \pi(\theta_\delta)$. Middle left panel: noncentral hypersphere t-distribution kernel density (2). Middle right panel: dataset normalized histogram plotted against the Gibbs-Jaynes model density, marginal density of the joint density above it. Lower left panel: NHST t-test p-value normalized histogram modelled by the Bayes factor $BF_t(p)$. Lower right panel: log-log plot of the false discovery rate $\text{fdr}(p)$ against the NHST two-tail p-value. The fdr is $\sim 10\%$ at $p = .01$, and $\sim 1\%$ at $p = 10^{-6}$.

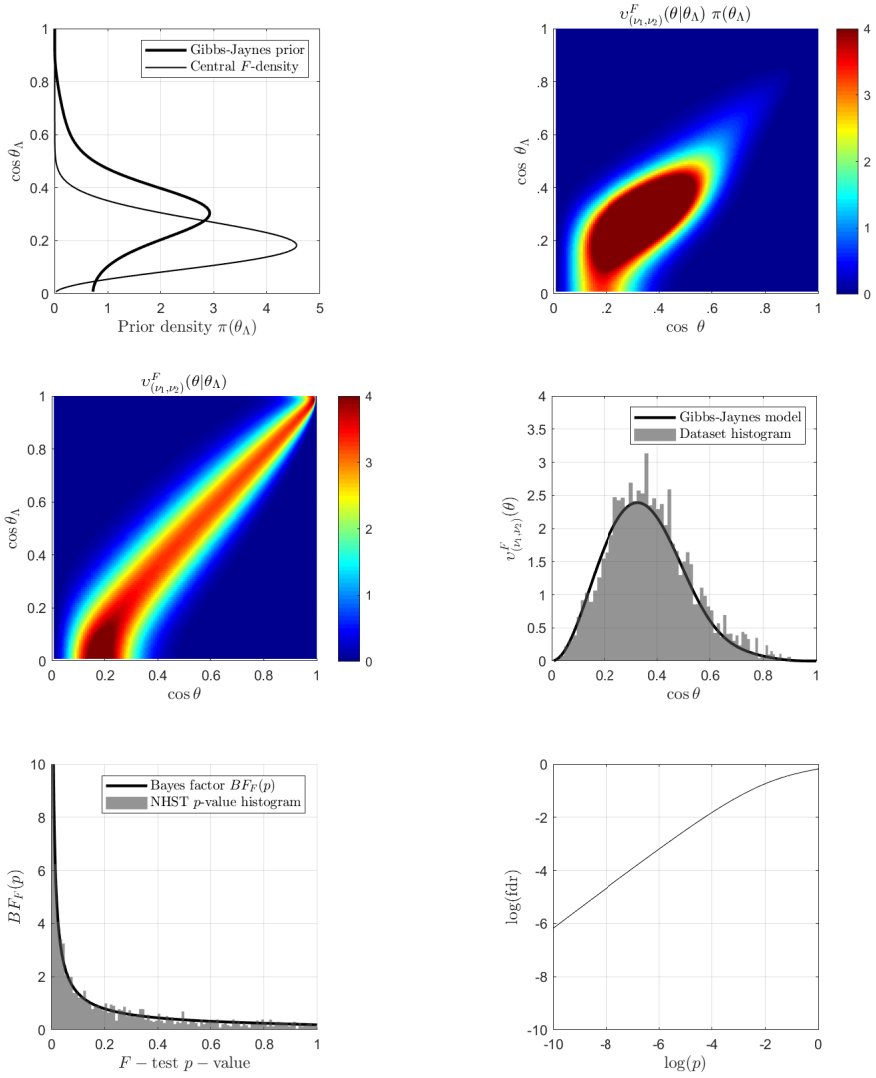

Figure 8.2: **Gibbs-Jaynes model** for a microarray dataset interrogating 2308 genes for $K = 4$ small round blue cell tumor types (SRBCT) of childhood, with respective class cardinalities 29, 11, 18, and 25 ($\nu_1 = 3, \nu_2 = 79$). Upper left panel: Jaynes prior density $\pi(\theta_\Lambda)$, marginal density of the joint density to its right, compared to the maximal entropy prior (10). Upper right panel: corresponding joint density $\upsilon^F_{(\nu_1,\ \nu_2)}(\theta|\theta_\Lambda)\ \pi(\theta_\Lambda)$. Middle left panel: noncentral hypersphere F-distribution kernel density (7). Middle right panel: dataset normalized histogram plotted against the Gibbs-Jaynes model density, marginal density of the joint density above it. Lower left panel: NHST F-test p-value normalized histogram modelled by the Bayes factor $BF_F(p)$. Lower right panel: log-log plot of the false discovery rate fdr(p) against the NHST p-value. The fdr is $\sim 10\%$ at $p = .01$, and $\sim 1\%$ at $p = 10^{-4}$.

We have graphically compared Jaynes priors to the central t- and F-distributions in both preceding examples. It is seen that the central distributions offer good approximations for the Jaynes priors. Would a biologist postulate such priors rather than compute the Jaynes data-constrained maxent priors, he would thereby assume that the biological system at hand can realize all possible size effects δ or Λ as they distribute themselves randomly but uniformly on the hypersphere. The latter

affirmation stems from the fact that the central t-distribution (5) and F-distribution (10) are projections of a uniform unconstrained maximal entropy distribution on the hypersphere unto specific $(K-1)$-dimensional between-class variance hyperplanes when discussed in terms of K-class ANOVA, with $K=2$ for the t-distribution and $K>2$ for the F-distribution (Le Blanc [29]). In short, these priors may not be flat, but they nevertheless stand for projections of an uninformative flat maximal entropy distribution on the hypersphere. Similar arguments have been invoked by Caticha and Preuss [10] in their construction of entropic priors:

> *Rather than seeking a totally non-informative prior, we translate information that we do in fact have: the knowledge of the likelihood function already constitutes valuable prior information.*

These results should help to further the long-lasting debate concerning definition of uninformative but geometrically meaningful priors for kernel-based parametric Bayesian models.

9. Conclusion

We sought in this manuscript to identify the key underpinnings of historical derivations of the Bayes prior and posterior update rules which successively emphasized: the classical probability theory approach; the more modern information-theory motivated variational approach and the contemporary convex geometry mirror update approach; and the Csiszár-Tusnády iterative alternating algorithm in joint density space. We concluded that the 250 year-old Bayes update rule naturally arises from the convex geometry of the convex entropy function S. With the exception of the original Bayes-Laplace derivation, all of these approaches shared two core notions: the concept of Kullback-Leibler relative entropy, and the principle of variational minimization. Since the Kullback-Leibler relative entropy is a Bregman divergence which only allows linear exploration of the convex entropy S tangent hyperplanes, all these algorithms necessarily involved descent along iterated tangent hyperplanes toward the solution, that is, they had to be iterative in nature. The iterative nature of these prior and posterior update rules unfortunately obscures the prior dependency on the model kernel.

In contrast, Jaynes maxent principle yields data-constrained structural priors which make very explicit their dependency on the model kernel, with the modelled density entropic convex dual providing the model kernel weight function. Through practical examples, we showed that Jaynes maximal entropy principle allows for specification of Bayesian priors which accurately model dense datasets produced by modern high-throughput technologies. We also showed that Jaynesian-Bayesian modelling of dense datasets NHST p-value distributions in terms of the Bayes factor $BF(p)$ allowed for a natural extension of the NHST framework. Finally, we empirically verified that, at least for the dense datasets considered herein, the data-constrained kernel-based Bayesian-Jaynesian priors for the noncentral t- and F-distributions are well approximated by the central t- and F-distributions which in turn are projections of a uniform unconstrained maximal entropy distribution on the hypersphere, an argument relevant to the definition of uninformative but geometrically meaningful priors for parametric Bayesian models.

Since dense datasets allow objective determination of the priors, we shall leave the concluding remark to McGrayne [32, p. 103]:

> *When they have little data, scientists disagree and are subjectivists; when they have piles of data, they agree and become objectivists.*

Appendix: Elements of convex analysis and information theory

Bregman divergences generalize the notion of distance between points (Bregman [8]). With an appropriate scalar product $\langle \cdot, \cdot \rangle$ and an appropriate gradient operator ∇, they can be defined for vectors, matrices, functions and probability distributions (Basseville [3]). Bregman divergences are formally defined in terms of convex functions. Let $\Phi : \Omega \to \mathbb{R}$ be a function that is strictly convex, continuously differentiable, and defined on a closed convex set Ω. The Bregman divergence for the convex generator function Φ is then defined as

$$D_\Phi(x, y) = \Phi(x) - (\Phi(y) + \langle \nabla\Phi(y), x - y \rangle), \ \forall x, y \in \Omega, \tag{45}$$

that is, as the difference between the value of Φ at x and the first-order Taylor expansion of Φ around y evaluated at point x. As such, the Bregman divergence allows for linear exploration of the tangent space at $\Phi(y)$. If, for example, Φ is the Euclidean distance $\Phi(x) = \frac{1}{2} \langle x, x \rangle$, then,

$$\begin{aligned} D_\Phi(x, y) &= \frac{1}{2} (\langle x, x \rangle - (\langle y, y \rangle + 2 \langle y, x - y \rangle)) \\ &= \frac{1}{2} \langle x - y, x - y \rangle . \end{aligned}$$

Note that, in this case, the Bregman distance equates the Euclidean distance since the latter is quadratic. This observation also applies to quadratic divergences of the form $x^T W x$ where W is a positive definite matrix, an example of which is the covariance matrix inverse which defines the Mahalanobis distance. Bregman divergences satisfy $D_\Phi(x, y) \geq 0$ for all x, y, and $D_\Phi(x, y) = 0$ if and only if $x = y$. Bregman divergences are not always metrics as they do not always satisfy the triangle inequality. Symmetry between their two arguments depends on the choice of Φ. The concept of Bregman divergence extends to that of functional Bregman divergence (Frigyik et al. [17]).

An entropy is a proper, convex, lower semi-continuous function $S : \triangle_\Xi \to \mathbb{R}$, where $\triangle_\Xi$ is the set of probability distributions over points in the probability simplex Ξ. The entropy S we shall consider herein is the negative of Shannon's entropy (Shannon [36]), that is, the convex function

$$S(\pi) = \int_{\mathcal{M}} \pi(m) \log \pi(m) \ \mathrm{d}m = \mathbb{E}_\pi[\log \pi]. \tag{46}$$

The Bregman divergence $D_S(\pi_1, \pi_2)$ associated with entropy S, where $\pi_1(m)$ and $\pi_2(m)$ assumed to be ℓ_1-normed, turns out to be equal to the Kullback-Leibler (KL)

divergence (Kullback and Leibler [27]), also known as the KL relative entropy:

$$\begin{aligned} D_S(\pi_1, \pi_2) &= \int_{\mathcal{M}} \left(\pi_1(m) \log \frac{\pi_1(m)}{\pi_2(m)} - (\pi_1(m) - \pi_2(m)) \right) \, \mathrm{d}m \\ &= \int_{\mathcal{M}} \pi_1(m) \log \frac{\pi_1(m)}{\pi_2(m)} \, \mathrm{d}m \\ &= KL(\pi_1, \pi_2) = \mathbb{E}_{\pi_1} \left[\log \frac{\pi_1}{\pi_2} \right]. \end{aligned} \tag{47}$$

In a Bregman divergence sense, the KL divergence is thus a linear measure on the tangent hyperplane at $S(\pi_2)$. As such, any Lagrangian-based variational computation in terms of the KL-divergence will necessarily entail iterations on successive such tangent hyperplanes till a stationary solution is met. In other words, a Bayes-Laplace prior cannot be computed in one single iteration with such linear tools.

Given a Bregman divergence induced by a convex function Φ, and given a constraint set $\mathcal{C}$, the Bregman projection of density ϕ unto the constraint set is defined as the infimum

$$\mathbb{P}_{\Phi}^{\mathcal{C}}(\phi) = \arg\min_{\pi \in \mathcal{C}} D_{\Phi}(\pi, \phi). \tag{48}$$

For example, when the convex function Φ is the entropy function S, and the constraint set $\mathcal{C}$ corresponds to the set of probability distributions $\triangle_{\Xi}$ over points in the probability simplex Ξ, the Bregman projection $\mathbb{P}_S^{\Xi}(\phi)$ simply amounts the a ℓ_1-normalization. The projection is achieved by minimizing the Lagrangian

$$L = \int_{\mathcal{M}} D_S(\pi, \phi) \, \mathrm{d}m + \lambda \left(\int_{\mathcal{M}} \pi(m) \, \mathrm{d}m - 1 \right), \tag{49}$$

with solution provided by

$$\mathbb{P}_S^{\Xi}(\phi) = \pi(m) = \frac{\phi(m)}{\lambda + 1} = \frac{\phi(m)}{\int_{\mathcal{M}} \phi(m) \, \mathrm{d}m} \in \triangle_{\Xi}. \tag{50}$$

Bregman projections underly the iterative alternating algorithm laid out in Section 6.

Convex duality is a powerful concept delineating a category of optimization problems for which powerful analytical, geometrical, and information theoretical tools are at hand (Borwein and Luke [7], Borwein [6]). Very succinctly, the concept uses the fact that analytical convex functions Φ have monotonic first derivatives and positive second derivatives. This implies that a convex function Φ can be fully described by either the doublet $(x, \Phi(x))$ or the doublet $(t, \Phi^*(t))$, where t stands for the first derivative (tangent) of the convex function Φ, and where the convex conjugate function $\Phi^*(t)$ is formally defined by

$$\Phi^*(t) := \sup_{x \in \mathcal{X}} \left(\langle t, x \rangle - \Phi(x) \right). \tag{51}$$

When Φ is given by the convex entropy S, we have:

$$\begin{aligned} S(x) &= x \log x, \\ t(x) &= \nabla_x S(x) = \log x + 1, \\ x(t) &= e^{(t-1)}, \qquad (52) \\ S^*(t) &= tx(t) - S(x(t)) \qquad (53) \\ &= te^{(t-1)} - e^{(t-1)} \log(e^{(t-1)}) = e^{(t-1)}, \\ \nabla_t S^*(t) &= e^{(t-1)} = x(t). \end{aligned}$$

Note that definition of entropy S varies in the literature, some authors using $S(x) = x \log x - x$ rather than $S(x) = x \log x$ in order to get rid of the cumbersome factor one in the equations above, and one must exercise some care when carrying over this factor one. The right hand side of equation (51) is a strictly concave maximization optimization problem with, for analytic functions, a unique maximum attained at $t = \nabla_x(\Phi(x))$ (Nielsen [33]). If Φ is strictly convex and twice differentiable, then so is Φ^*. If Φ is lower semi-continuous, Φ is the convex conjugate of Φ^*, implying a dual relationship between Φ and Φ^*. In this case, $\nabla_t \Phi^*(t)$ and $\nabla_x \Phi(x)$ are functional inverses, as demonstrated above for the entropy function S. For a density $\pi(m)$ defined in model space $\mathcal{M}$ and entropy $S(\pi)$, the dual function is similarly defined by

$$S^*(\varphi) := \sup_\pi \left[\int_{\mathcal{M}} \varphi(m)\ \pi(m)\ \mathrm{d}m - S(\pi) \right]. \qquad (54)$$

When including the Bregman projection (50) in the variational problem, we compute

$$S^*(\varphi) = \log \int_{\mathcal{M}} \exp \varphi(m)\ \mathrm{d}m, \qquad (55)$$

implying that

$$\nabla S^*(\varphi) = \frac{\exp(\varphi(m))}{\int_{\mathcal{M}} \exp(\varphi(m))\ \mathrm{d}m} = \frac{\pi(m)}{\int_{\mathcal{M}} \pi(m) \mathrm{d}m} \in \triangle_\Xi, \qquad (56)$$

for $\varphi = \nabla S(\pi) = \log \pi + 1$. The dual operator ∇S^* thus simply projects the functional entropy S derivative $\nabla S(\pi)$ onto a normalized distribution in the probability simplex Ξ, that is, $\nabla S^*(\nabla S(\pi)) \in \Delta_\Xi$. Reid et al. [35] refers to the entropy S convex dual S^* as its entropic dual.

References

[1] R. C. Aster, B. Borchers, C. H. Clifford: *Parameter Estimation and Inverse Problems*, Elsevier, Amsterdam (2018).

[2] A. Banerjee, X. Guo, H. Wang: *On the optimality of conditional expectation as a Bregman predictor*, IEEE Trans. Inform. Theory 51/7 (2005) 2664–2669.

[3] M. Basseville: *Divergence measures for statistical data processing – an annotated bibliography*, Signal Processing 93/4 (2013) 621–633.

[4] H. H. Bauschke, A. S. Lewis: *Dykstras algorithm with Bregman projections: a convergence proof*, Optimization 48/4 (2000) 409–427.

[5] P. G. Bissiri, C. C. Holmes, S. G. Walker: *A general framework for updating belief distributions*, J. Royal Stat. Society, Ser. B 78/5 (2016) 1103–1130.

[6] J. M. Borwein: *Maximum entropy and feasibility methods for convex and nonconvex inverse problems*, Optimization 61/1 (2012) 1–33.

[7] J. M. Borwein, D. R. Russell: *Duality and convex programming*, in: *Handbook of Mathematical Methods in Imaging*, O. Scherzer (ed.), Springer, New York (2011) 229–270.

[8] L. M. Bregman: *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Physics 7/3 (1967) 200–217.

[9] D. Calvetti, E. Somersalo: *Inverse problems: From regularization to Bayesian inference*, Wiley Interdisciplinary Reviews: Comp. Stat. 10/3 (2018) e1427.

[10] A. Caticha, R. Preuss: *Maximum entropy and Bayesian data analysis: entropic prior distributions*, Phys. Review E 70/4 (2004) 046127.

[11] Y. Censor, S. Reich: *The Dykstra algorithm with Bregman projections*, Comm. Appl. Analysis 2/3 (1998) 407–420.

[12] M. Chae, R. Martin, S. G. Walker: *On an algorithm for solving Fredholm integrals of the first kind*, Stat. Computing (2018) 1–10.

[13] I. Csiszár, G. Tusnády: *Information geometry and alternating minimization procedures*, Stat. Decisions 1 (1984) 205–237.

[14] S. Dudoit, J. P. Shaffer, J. C. Boldrick: *Multiple hypothesis testing in microarray experiments*, Stat. Science 18/1 (2003) 71–103.

[15] R. L. Dykstra: *An iterative procedure for obtaining I-projections onto the intersection of convex sets*, Annals of Probability (1985) 975–984.

[16] B. Efron: *Size, power and false discovery rates*, Annals of Statistics 35/4 (2007) 1351–1377.

[17] B. A. Frigyik, S. Srivastava, M. R. Gupta: *Functional Bregman divergence and Bayesian estimation of distributions*, IEEE Trans. Inform. Theory 54/11 (2008) 5130–5139.

[18] A. Gelman, D. Simpson, M. Betancourt: *The prior can often only be understood in the context of the likelihood*, Entropy 19/10 (2017) 555.

[19] T. Gneiting, A. E. Raftery: *Strictly proper scoring rules, prediction, and estimation*, J. Amer. Stat. Assoc. 102/477 (2007) 359–378.

[20] P. D. Grünwald, A. P. Dawid: *Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory*, Annals of Statistics 32/4 (2004) 1367–1433.

[21] J. Hadamard: *Sur les problèmes aux dérivées partielles et leur signification physique*, Princeton Univ. Bulletin (1902) 49–52.

[22] D. Hestenes: *Review of the book "E. T. Jaynes: Papers on Probability, Statistics, and Statistical Physics" by R. D. Rosenkrantz (ed.)*, Foundations of Physics 14/2 (1984) 187–191.

[23] E. T. Jaynes: *On the rationale of maximum-entropy methods*, Proc. IEEE 70/9 (1982) 939–952.

[24] L. K. Jones, C. L. Byrne: *General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis*, IEEE Trans. Inform. Theory 36/1 (1990) 23–30.

[25] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, P. S. Meltzer: *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks*, Nature Medicine 7/6 (2001) 673–679.

[26] B. Kulis, M. A. Sustik, I. S. Dhillon: *Low-rank kernel learning with Bregman matrix divergences*, J. Machine Learning Res. 10 (2009) 341–376.

[27] S. Kullback, R. A. Leibler: *On information and sufficiency*, Ann. Math. Statistics 22/1 (1951) 79–86.

[28] G. Le Besnerais, J.-F. Bercher, G. Demoment: *A new look at entropy for solving linear inverse problems*, IEEE Trans. Inform. Theory 45/5 (1999) 1565–1578.

[29] R. Le Blanc: *Bayesian analysis on a noncentral Fisher-Student's hypersphere*, Amer. Statistician 73/2 (2019) 126–140.

[30] Y. Lucet: *What shape is your conjugate? A survey of computational convex analysis and its applications*, SIAM Review 52/3 (2010) 505–542.

[31] P. G. L. Mana: *Geometry of maximum-entropy proofs: stationary points, convexity, Legendre transforms, exponential families*, arXiv: 1707.00624 (2017).

[32] S. B. McGrayne: *The Theory that Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*, Yale University Press, New Haven (2011).

[33] F. Nielsen: *Legendre transformation and information geometry*, https://www2.sonycsl.co.jp/person/nielsen/Note-LegendreTransformation.pdf (2010).

[34] V. C. Preda: *The Student distribution and the principle of maximum entropy*, Ann. Inst. Statist. Math. 34/1 (1982) 335–338.

[35] M. D. Reid, R. M. Frongillo, R. C. Williamson, N. Mehta: *Generalized mixability via entropic duality*, arXiv: 1406.6130 (2014).

[36] C. E. Shannon: *A mathematical theory of communication*, Bell Sys. Tech. J. 27/3 (1948) 379–423.

[37] A. M. Stuart: *Inverse problems: a Bayesian perspective*, Acta Numerica 19 (2010) 451–559.

[38] B. Taskar, S. Lacoste-Julien, M. I. Jordan: *Structured prediction, dual extragradient and Bregman projections*, J. Machine Learning Res. 7 (2006) 1627–1653.

[39] A. N. Tikhonov, Y. Y. Arsenin: *Solutions for Solving Ill-Posed Problems*, John Wiley and Sons, New York (1977).

[40] B. C. Vemuri, M. Liu, S.-I. Amari, F. Nielsen: *Total Bregman divergence and its applications to DTI analysis*, IEEE Trans. Medical Imaging 30/2 (2011) 475–483.

[41] S. G. Walker: *Bayesian inference via a minimization rule*, Sankhyā: Indian J. Statistics (2006) 542–553.

[42] P. M. Williams: *Bayesian conditionalisation and the principle of minimum information*, British J. Phil. Science 31/2 (1980) 131–144.

[43] A. Zellner: *Optimal information processing and Bayes's theorem*, Amer. Statistician 42/4 (1988) 278–280.

[44] J. Zhu, N. Chen, E. P. Xing: *Bayesian inference with posterior regularization and applications to infinite latent SVMs*, J. Machine Learning Res. 15/1 (2014) 1799–1847.