Entropic priors : corrections, new observations and applications

Tai Kaufmann
University at Albany, State University of New York, 7beggars@gmail.com

The University at Albany community has made this article openly available. Please share how this access benefits you.

Follow this and additional works at: https://scholarsarchive.library.albany.edu/legacy-etd

Part of the Statistics and Probability Commons

Recommended Citation
https://scholarsarchive.library.albany.edu/legacy-etd/658

This Dissertation is brought to you for free and open access by the The Graduate School at Scholars Archive. It has been accepted for inclusion in Legacy Theses & Dissertations (2009 - 2024) by an authorized administrator of Scholars Archive. Please see Terms of Use. For more information, please contact scholarsarchive@albany.edu.
ENTROPIC PRIORS:
CORRECTIONS, NEW OBSERVATIONS
AND APPLICATIONS

by

Tai Kaufmann

A Dissertation
Submitted to the University at Albany, State University of New York
in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Philosophy

College of Arts & Sciences
Department of Mathematics & Statistics

2012
Abstract. Carlos C. Rodríguez [6] [7] has published a family of priors, so named (by John Skilling [4]) the Entropic Priors, that decay as a function of Kullback-Leibler divergence in order to encode a blend of proximity near a prior estimate, with uniformity over the hypothesis space of probability distributions. The asymmetry of the KL divergence gives rise to a continuum of Entropic Priors. Here we examine only the extreme cases, the 0- and 1-entropic priors, the former coinciding with the standard and convenient conjugate priors, but the latter being the unique optimizer of a simple notion of ignorance. The original contributions of this paper are:

(1) We have developed quantitative criteria for comparing the informativeness of the Entropic Priors, that is, how much they affect Bayesian posteriors. We have examined many common regular models and found that the 1-prior is always strictly informative; we have found no so-called antidata effect [10] of the prior canceling information in the likelihood. Further, the 1-prior is often more informative than the 0-prior. By examining the definitions of the Entropic Priors we explain why these facts are so. We correct some formulas of the 1-prior published in [10]. We produce graphics that help visualise differences in the 0- and 1-priors.

(2) We make new interpretations of naive flat priors for Bitnets by interpreting them as presumed observations, showing how they distort the true volume-uniform elements. We observe that in particular, the posterior variance of bottom-most nodes, if the observations are (almost) all zeros or ones, benefits from vast reduction in variance by using the correct volume-uniform element versus a naive diffuse prior.

(3) As an application of the above, we demonstrate that Entropic Priors can be used with ease as a hierarchical priors in random effects models, providing flexibility in modeling that is differential-geometrically correct, which is not so with ad-hoc “standard diffuse priors.” We show that this leads to an improvement in a ranking problem where a league table based on observations from Binomial distributions is to be built.
ACKNOWLEDGMENTS

Thanks to God and to my parents.

Thanks to my wife for supporting and enduring my work.

Thanks to my advisor Carlos Rodríguez, whose innovative and rigorous work has opened the way for this dissertation, for showing me new frontiers in mathematics and computation, and for guiding and supporting me through the dissertation process.

Thanks to my committee members. Ariel Caticha has provided a third-party view of Carlos Rodríguez’s work which has helped guide me. With Kevin Knuth I have had the pleasure to take a Bayesian Inference course, which has left me with perspectives and examples that have been essential. And with Malcolm Sherman I have had the pleasure to take a regression course which has also left me with important lessons and examples.

Thanks to Philip Goyal for inviting me to present the results that are in Section 1 at the MaxEnt 2011 conference, and to John Skilling for his helpful review and acceptance of my paper to the MaxEnt Proceedings.

Thanks to the R, LaTeX and WinBUGS developers and communities.
## Contents

Acknowledgments .......................................................... iii
Introduction ............................................................... 1

1. Entropic 1-Prior is Strictly Informative (i.e. No Antidata); Comparisons of 0- and 1-Priors
   1.1. Definition of the Entropic 0- and 1-Priors ...................... 5
   1.2. Definition and Measure of Antidata; Findings ................. 7
   1.3. Inference on the Bernoulli Model ......................... 8
   1.4. Inference on the Exponential Model ......................... 9
   1.5. Inference on the Gaussian with Unknown Variance .......... 10
   1.6. Inference on the Gaussian with Both Parameters Unknown ... 12

2. Observations and Conjectures Regarding Entropic Priors ......... 17
   2.1. Regarding Strict Informativeness of the 1-Prior ............ 17
   2.2. Regarding Comparison of the 0- and 1-priors ............... 17
   2.3. Theory for Explaining the Above Observations Regarding
        Informativeness of the 1-Prior, and a Proof Sketch of the
        Conjectures .................................................. 20

3. Interpreting Naive Volume Elements in Bitnets .................. 23
   3.1. Bernoulli .................................................. 24
   3.2. Bitnet of Two Nodes ...................................... 26
   3.3. Higher Bitnets .......................................... 26

4. Applications .......................................................... 27

References ............................................................... 37
INTRODUCTION

If we have some estimate of a parameter that we would like to do inference on by Bayes’ Theorem, the prior should be denser near the estimate, and otherwise if we are ignorant the prior should be uniform in the hypothesis space. To achieve uniformity over the hypothesis space for regular statistical models, the Fisher information metric will yield the invariant volume element, aka the Jeffreys prior. But what if the volume element is not normalizable, as is the case with parameters such as Gaussian mean, Gaussian variance, and Exponential rate? Improper priors present hazards of nonsense or paradoxical results (e.g. the envelope problem, [5] p.9). Normalizing the volume element is one motivation for using Entropic Priors, which provide a formula for systematically, objectively, and flexibly peaking the prior around an estimate and decaying at points that are distant from the estimate in the hypothesis space. What “points,” and “distance” and “hypothesis space” are we referring to?

A regular statistical model should be understood as a differentiable manifold ([1]; [8] p. 3). Let \( P = \{ p(x|\theta) : \theta \in \Theta \} \) be a regular statistical model (e.g. Gaussian with unknown mean and/or variance, or Bernoulli). Imbed \( P \) in the Hilbert space \( L^2(p) \)

\[
L_2 = \{ \sqrt{p} f | p \in P, f \in L^2(p) \},
\]

where \( L^2(p) \) are the square integrable functions wrt \( p \),

\[
L^2(p) = \left\{ f | \int f^2(x) p(x) \, dx < \infty \right\},
\]

and \( L_2 \) has dot product

\[
< \sqrt{p} f, \sqrt{q} g > = \int (\sqrt{p} f)(\sqrt{q} g) \, dx
\]

and hence the metric

\[
\| \sqrt{p} f \|^2 \equiv < \sqrt{p} f, \sqrt{p} f > = \int p f^2 \, dx;
\]

\[\text{More generally, we can imbed in Banach spaces } L_\delta; \text{ see } [8].\]
imbed the probability density function (pdf) \( p \) in \( L_2 \) via

\[
p \mapsto 2\sqrt{p}.
\]

Thus we represent the set of densities \( P \) as a surface of vectors \( c(\theta) = 2\sqrt{p(x|\theta)} \) for which we can compute volume, curvature, etc. In particular, the volume element coincides with the Jeffreys prior, the square root of the determinant of the Fisher information matrix, because the volume element is

\[
\sqrt{\det(c_{\theta_i}, c_{\theta_j})_{i,j}} d\theta = \sqrt{\det\left( \int c_{\theta_i} c_{\theta_j} dx \right)_{i,j}} d\theta
\]

\[
= \sqrt{\det\left( \int \frac{p_{\theta_i}}{\sqrt{p}} \frac{p_{\theta_j}}{\sqrt{p}} dx \right)_{i,j}} d\theta
\]

\[
= \sqrt{\det\left( \int \frac{p_{\theta_i} p_{\theta_j}}{p^2} dx \right)_{i,j}} d\theta
\]

\[
= \sqrt{\det(Fisher)} d\theta,
\]

where the subscript \( \theta_i \) denotes partial derivative wrt \( \theta_i \).

In Section 1 we review the definitions of Kullback-Leibler Divergence, Entropic Priors, and Fisher Information. As a new contribution, we introduce quantitative definitions for measuring the informativeness of priors, and we document that contrary to the claim of [10] the 1-prior is strictly informative in all common regular statistical models that we have examined. Furthermore, we report that the 1-prior is more informative than the 0-prior in all but one of the common models we have examined, the exception being the Bernoulli model and the Bitnet models in general, specifically when the parameters are not extreme (that is, when the node probabilities \( p_i \) are near zero or one, the 1-prior (which approaches degeneracy as \( p_i \to 0 \) or 1) becomes more informative than the 0-prior, and otherwise the 0-prior is more informative than the 1-prior). We show graphics that are the first published visualizations comparing the 0-prior and 1-priors. We make corrections to the formulas for the Gaussian 1-prior and 1-posterior (with \( \mu \) and \( \nu \) both unknown) that were published in [10]. Section 1 is also being published in the MaxEnt 2011 Proceedings as “A Note on Antidata.” Since
the submission of that monograph, more models have been searched for antidata according to the aforementioned definitions, with no findings of antidata, and the results are summarized in Section 2.

In Section 2, our observations from Section 1 are summarized, and we make conjectures as to their generality. We remark that the 1-prior, being non-conjugate in general, can lead to a posterior that is bimodal, which is atypical for the 0-prior, the conjugate prior; and we produce an example.

In Section 3 we analyze the volume elements of some simple Bitnets and show how much of a distortion can be introduced by a naive “flat” prior that seems to be uniform, and how much inferential power can be thereby wasted. Formulas for the true volume elements have been published by C. Rodríguez in [9], but here we interpret the naive flat priors in terms of virtual data to show how much distortion is introduced by such priors; as far as we know, these interpretations are new contributions. We apply the correct volume priors to the examples in Section 4. Further, we observe that in particular, the posterior variance of bottom-most nodes, if the observations are almost all zeros or ones, benefits from vast reduction in variance by using the correct volume-uniform element versus a naive diffuse prior.

In Section 4 we apply the correct volume prior and Entropic Priors to a simple example in order to demonstrate their application and advantages of use. We revisit an example given in the WinBUGS software documentation, namely a problem of ranking hidden Bernoulli parameters \( p_i, i = 1..N \) based observations \( y_i \) that are \( Bin(n_i, p_i) \). We use the volume prior in contrast to the Beta(1,1) prior used in the WinBUGS documentation when considering the fixed effects model, and we use various 0-priors when considering the random effects model. In the latter case, the flexibility furnished by the 0-prior is made evident, and in particular we show that by using a 0-prior with a higher weight and with a higher estimated population variance

\[ \alpha = 0.5 \]

2Although multimodality probably cannot occur in the natural parameterization, other parameterizations may present multimodality. An example is Beta with parameters less than 1, for example the 0-prior \( Beta(0.75, .75) \) which consists of \( \alpha = 0.5 \) virtual observations having statistic \( \bar{x} = 0.5 \).
we can produce a conclusive ranking, in contrast to the inconclusive posteriors resulting from the method in the WinBUGS documentation that uses an ad-hoc “standard non-informative prior.” In the process we discover how informative this prior really is, and we generalize, defining a way to quantify how much a prior informs the results of an inference. We also discover that in hierarchical models, conditional on the intermediate parameters, the volume element and the Entropic Prior only depend on the likelihood defined on the topmost hyperparameter.
1. Entropic 1-Prior is Strictly Informative (i.e. No Antidata); Comparisons of 0- and 1-Priors

Note: An earlier version of this section has been accepted for publication as “A Note on Antidata” in the *AIP Conference Proceedings* (“MaxEnt 2011”), due in April 2012. Some changes and corrections have been made here, and the more substantial ones will have footnotes to indicate the change from the MaxEnt 2011 version. In particular, a discussion has been included here regarding why the 1-prior merits interest; this was omitted from the earlier version due to space constraint.

1.1. Definition of the Entropic 0- and 1-Priors. For brevity, let \( \theta_0 \) and \( \theta \) be the names of Exponential Family pdfs \( p(x|\theta_0) \) and \( p(x|\theta) \).\(^3\) A measure of the dissimilarity of the pdfs is the Kullback number, which is the mean divergence of \( \theta \) from \( \theta_0 \) when averaged according to \( \theta \):

\[
I(\theta_0 : \theta) = E_{p(x|\theta_0)}[\ln \frac{p(x|\theta_0)}{p(x|\theta)}] = \int_{\Omega} p(x|\theta_0) \ln \frac{p(x|\theta_0)}{p(x|\theta)} dV. \tag{1.1}
\]

This is a parameterization-independent way to quantify the closeness of two pdfs. It is not symmetric, however, for we may choose to use \( I(\theta : \theta_0) \) instead. These are the limiting cases (respectively \( \delta = 0 \) or \( 1 \)) of a continuum of what are essentially the only ways to quantify the closeness of two pdfs, the delta-information-deviations ([1] and [8] p. 4)\(^4\)

\[
I_\delta(\theta_0 : \theta) \equiv \frac{1}{\delta(1-\delta)} \int_{\Omega} \delta p(x|\theta_0) + (1 - \delta)p(x|\theta_0) - p^\delta(x|\theta)p^{1-\delta}(x|\theta_0) dV, \tag{1.2}
\]

\[
0 \leq \delta \leq 1. \tag{1.3}
\]

The *0-entropic family of prior pdfs* for the parameter \( \theta \) is parametrized by \( \theta_0 \) (an estimated \( \theta \) value) and a weight \( \alpha > 0 \) (a measure of belief assigned to the estimate)

\(^3\)Actually these names are the coordinates of the vectors in \( L_2 \) mapped by the embedding (0.5).

\(^4\)The MaxEnt 2011 version has this partially correct statement: “This function, though not symmetric, is parameterization-independent and is regarded as essentially the only way to compare the closeness of two pdfs.”
and is defined by an exponential decay as $\theta_0$ diverges from $\theta$:

\begin{equation}
p_0(\theta|\alpha, \theta_0) \propto \exp(-\alpha I(\theta_0 : \theta)).
\end{equation}

The **1-priors** however are defined by exponential decay as $\theta$ diverges from $\theta_0$:

\begin{equation}
p_1(\theta|\alpha, \theta_0) \propto \exp(-\alpha I(\theta : \theta_0)).
\end{equation}

The 0-priors coincide with the natural conjugate priors, which are convenient. A salient difference\(^5\) is that if $p(x|\theta_0) \to 0$ anywhere then $p_1 \to 0$ for all $\theta \neq \theta_0$ so the 1-prior can become degenerate; but $p_0$ supports all $\theta$ in any case. So why does the 1-prior merit interest?\(^6\) As discussed and proved in [10] p.18, the 1-prior has the distinction that it is the distribution that is the unique solution $\pi^*$ to the following constrained variational problem: Let $\pi$ be the distribution that minimizes the Kullback divergence of the joint pdf of $(x^\alpha, \theta)$, i.e.

\begin{equation}
p(x_1|\theta)p(x_2|\theta) \ldots p(x_\alpha|\theta)\pi(\theta) \equiv p^\alpha \pi,
\end{equation}

away from the factorized model

\begin{equation}
p(x_1|\theta_0)p(x_2|\theta_0) \ldots p(x_\alpha|\theta_0)\omega(\theta) \equiv p^0 \omega
\end{equation}

where $\omega$ is the Uniform prior with respect to the information volume in $\Theta$. (This is related to what Kullback called the *Minimum Discrimination Information* (MDI) Principle). That is, the 1-prior alone is the $\pi^*$ that satisfies the simple notion of ignorance

\begin{equation}
\pi^* = \arg \min \pi I(p^\alpha \pi : p^0 \omega),
\end{equation}

so that $\pi$ maximizes the stochastic independence of the parameter $\theta$ from the data $x$. This can be also be viewed with the help of a straightforward computation of the

---

\(^5\)This was pointed out to me by John Skilling.

\(^6\)The following discussion of the actions for ignorance was not included in the MaxEnt proceedings, due to space constraints.
above entropy expression:

\[ I(p^\alpha \pi : p_0^\alpha \omega) = \int p^\alpha \pi \ln \frac{p(x^\alpha | \theta) \pi(\theta)}{p(x^\alpha | \theta_0) \omega(\theta)} dx^\alpha d\theta \]  
(1.9)

\[ = \alpha \int \pi(\theta) I(\theta : \theta_0) d\theta + I(\pi : \omega) \]  
(1.10)

\[ = \alpha E_\pi[I(\theta : \theta_0)] + I(\pi : \omega). \]  
(1.11)

That is, the 1-prior is this compromise between mass concentration of \( \theta \) around the estimate \( \theta_0 \), and uniform spread of \( \theta \) over the information volume, when sampling relative to \( \pi \). Likewise the 0-prior minimizes

\[ \alpha E_\pi[I(\theta_0 : \theta)] + I(\pi : \omega). \]  
(1.12)

These are two ways to balance a uniform spread over \( \omega \) along with a mass concentration near \( p(x|\theta_0) \). The quantity (1.11) that the 1-prior minimizes has a symmetry lacking in that of the 0-prior (1.12) — notice that \( \theta \) and \( \pi \) both occur in the first positions in \( I \) with the 1-prior, whereas with the 0-prior \( \theta \) is second while \( \pi \) is first.

1.2. Definition and Measure of Antidata; Findings. Antidata is the occurrence of encoding a parameter estimate in the 1-prior and having it result in a 1-posterior that is more ignorant (more similar to a uniform prior) than if we had not encoded this estimate; it is as if the estimate weight in the prior annihilates part of the information in the likelihood, rather than compounding it. Since Kullback divergence is invariant under reparameterization, it promises objective ways to quantify the informativeness of a prior. One way to define antidata is to compare the information gain of the 1-posterior, versus the information gain of the posterior that follows from the Uniform prior. That is, to take the difference of their respective Kullback divergences relative to the Uniform prior:

**Definition 1.13. Definition of Antidata in Terms of Information Gain.** If the 1-posterior has less information gain than the posterior that follows from the Uniform prior, this is antidata. We restrict the definition of antidata to the condition that the sample statistic coincides with the parameter estimate. The reason
for this restriction is that if the sample statistic and the parameter are different, the 1-posterior can indeed reduce the information gain, however the same is true for the 0-posterior.

If a Uniform prior does not exist (i.e. when the hypothesis space has infinite volume), we may approximate it with a MIPPr, defined as follows:

Definition 1.14. **Minimally informative proper prior (MIPPr).** A 1-prior (or other δ-prior) with infinitesimal estimate weight α.

In all cases we have examined, the 1-posterior is more informative than the posterior that follows from the Uniform prior. With Bernoulli, the 1-posterior becomes more informative than the 0-posterior as the sample proportion \( \theta_0 \) approaches 0 or 1 — see Fig. 1. But with the Exponential and Gaussian models the information gain surplus beyond the information gain of the posterior that follows from the MIPPr is constant (or nearly constant) regardless of the sample statistic, and the 1-posterior is always more informative than the 0-posterior. These observations are the original contributions of this paper. It remains to prove these observations in general (or disprove them with an instance of antidata in some model). Details follow.

Another way to quantify antidata, proposed by Carlos C. Rodríguez, is as follows:

Definition 1.15. **The \( \beta^* \) criterion for Defining Antidata.** If the argument \( \beta^* \) that minimizes the divergence of a 1-posterior with prior weight \( \alpha > 0 \), from a 0-posterior with weight \( \beta \) is such that \( \beta^* < 0 \), this is antidata.

This too was measured and no negative \( \beta^* \) were found for models mentioned in this paper.

1.3. Inference on the Bernoulli Model. Let \( \theta \) be the probability that \( x = 1 \), and \( 1 - \theta \) the probability that \( x = 0 \). Write \( p(x|\theta) = (\theta)^x(1-\theta)^{1-x} \). Fisher information is \( g = -E_x[\partial^2 \ln p] = \ln \frac{1}{\theta (1-\theta)} \). So the hypothesis space volume element as well as the ignorant **Uniform prior** is \( m(\theta) d\theta = 1 dV \propto \sqrt{g} d\theta = \frac{d\theta}{\sqrt{\theta(1-\theta)}} \) which is \( \text{Beta} \left( \frac{1}{2}, \frac{1}{2} \right) \).
Multiplying the likelihood yields the posterior that follows from the Uniform prior, denoted \( p_U \), which is \( \text{Beta} \left( \sum x_i + \frac{1}{2}, n - \sum x_i + \frac{1}{2} \right) \). Suppose we wish to encode some prior knowledge of \( \theta \) by using an estimate of \( \theta_0 = .01 \) with weight \( \alpha \). This might be the case for example if a merchant of some items claims 99% are free of defect, and we wish to give his estimate a weight of \( \alpha \) against the \( n \) observations that we would normally take in order to get a posterior confidence that the defect percentage is below some limit. For general \( \theta_0 \) and \( \alpha \) we have \( I(\theta_0 : \theta) = \sum x p(x|\theta_0) \ln \frac{p(x|\theta_0)}{p(x|\theta)} = \theta_0 \ln \frac{\theta_0}{\theta} + (1 - \theta_0) \ln \frac{1 - \theta_0}{1 - \theta} \). So the 0-prior is \( p_0(\theta|\theta_0, \alpha) dV \propto \exp(-\alpha I(\theta_0 : \theta)) dV \) which is \( \text{Beta}(\alpha \theta_0 + \frac{1}{2}, \alpha (1 - \theta_0) + \frac{1}{2}) \). (Note that for \( \alpha = 1 \), \( \theta_0 = \frac{1}{2} \) uniquely this yields \( 1 d\theta \).) By swapping \( \theta \) and \( \theta_0 \) in \( I \) the 1-prior is

\[
I(\theta : \theta_0) = \frac{\exp(-\alpha (\theta_0 - \theta))}{\sqrt{\theta (1 - \theta)}} d\theta \propto \left( \frac{\theta_0}{\theta} \right)^{\alpha \theta} \left( \frac{1 - \theta_0}{1 - \theta} \right)^{\alpha (1 - \theta)} \frac{d\theta}{\sqrt{\theta (1 - \theta)}}.
\]

Note that the 1-prior is degenerate if \( \theta_0 \) is 0 or 1. The 0- and 1-posteriors, assuming the sample coincides with the estimate, are respectively

\[
p_0(\theta|\theta_0, \alpha, x^n) dV \propto \theta^{\sum x_i + \alpha \theta_0} (1 - \theta)^{n - \sum x_i + \alpha (1 - \theta_0)} \frac{d\theta}{\sqrt{\theta (1 - \theta)}}
\]

\[
p_1(\theta|\theta_0, \alpha, x^n) dV \propto \left( \frac{\theta_0}{\theta} \right)^{\alpha \theta} \left( \frac{1 - \theta_0}{1 - \theta} \right)^{\alpha (1 - \theta)} \theta^{\sum x_i} (1 - \theta)^{n - \sum x_i} d\theta \frac{d\theta}{\sqrt{\theta (1 - \theta)}}.
\]

Note that \( p_0 \) encodes more decay for \( \theta \) near 0 or 1. The difference of information gains \( I(p_1 : m) - I(p_U : m) \) remained positive as we varied \( \theta_0 \) from .001 to .999, \( n \) from 2 to 20, and \( \alpha \) from 1 to 1e-5, so there was no antidata. The plot of \( I(p_0 : m) - I(p_U : m) \), though having mostly opposite concavity, also remained positive, as expected. Figure 1 shows the information gain differences for \( n = 15 \) and \( \alpha = 1 \) as \( \theta_0 \) varies; the values \( n = 15 \) and \( \alpha = 1 \) were of particular interest to me because if \( \sum x_i/n = \theta_0 = .01 \) then approximately 16 samples are needed to get a posterior \( P(\theta < .10) \geq .90 \), which was the goal in a quality control problem that inspired my consideration of the Bernoullis.

1.4. Inference on the Exponential Model. Write \( x \sim \text{Exp}(\lambda) \) for \( \lambda, x > 0 \) if \( p(x|\lambda) = \lambda \exp(-\lambda x) \). Fisher information is \( g = E_{x|\lambda} \left[ -\frac{\partial^2}{\partial \lambda^2} \ln p(x|\lambda) \right] = \lambda^{-2} \). Volume element and (improper) Uniform Prior is \( 1 dV \propto \sqrt{g} d\lambda = d\lambda/\lambda \). The posterior
Bernoulli trials: Divergences of Posteriors from Uniform Prior

\[ I(p_0 : m) - I(p_U : m) \]

\[ I(p_1 : m) - I(p_U : m) \]

\[ \theta_0 \]

**Figure 1.** Difference of Posterior Divergences for Bernoulli. Assuming \( n = 15 \) samples, \( \sum x_i = n\theta_0 \), and estimate weight \( \alpha = 1 \). Vary \( \theta_0 \). The difference remains positive. No Antidata.

that follows from the Uniform prior is \( p_U(\lambda|\lambda_0, \alpha, x^n) \propto \lambda^n \exp(-n\bar{x}_n\lambda) \) which is Gamma\((n, \frac{1}{n\bar{x}_n})\). Divergence of \( \lambda_0 \) from \( \lambda \) is \( I(\lambda_0 : \lambda) = \ln \lambda_0 - \ln \lambda - 1 + \lambda/\lambda_0 \). The 0-prior is \( p_0(\lambda|\lambda_0, \alpha) \propto \exp(-\alpha I(\lambda_0 : \lambda)) \propto \lambda^\alpha \exp(-\frac{\alpha\lambda}{\lambda_0}) \) (which is Gamma\((\alpha, \frac{\lambda_0}{\alpha})\) with mean \( \lambda_0 \), as we would expect). The 1-prior is \( p_1(\lambda|\lambda_0, \alpha) \propto \exp(-\alpha I(\lambda : \lambda_0)) \propto \lambda^{-\alpha} \exp(-\frac{\alpha\lambda}{\lambda_0}) \) (which is InverseGamma\((\alpha, \lambda_0^\alpha)\) with mean \( \frac{\alpha\lambda_0}{\alpha-1} \) for \( \alpha > 1 \) (else the mean diverges), and mode \( \frac{\alpha\lambda_0}{\alpha+1} \)). The respective posteriors after \( n \) samples are

\begin{align*}
    p_0(\lambda|\lambda_0, \alpha, x^n) &\propto \lambda^{n+\alpha} \exp(-\lambda(\frac{\alpha}{\lambda_0} + n\bar{x}_n)) \quad \text{Note: more decay for } \lambda \text{ large} \\
p_1(\lambda|\lambda_0, \alpha, x^n) &\propto \lambda^{n-\alpha} \exp(-\lambda_0(\frac{\alpha}{\lambda} - \lambda n\bar{x}_n)) \quad \text{Note: more decay for } \lambda \text{ small}
\end{align*}

Comparing the information gains of the 1-posterior and the posterior that follows from the Uniform prior, relative to the MIPPr, we found no antidata. (Method: Kullback divergence was compared as we varied \( \alpha \) from 1e-6 to .1, \( n \) from 2 to 6, and \( \bar{x}_n \) from 1e-3 to 1e+3, and used a 1-prior with weight \( \alpha = 1e-10 \) as the MIPPr. See Fig. 2 for a plot of the pdfs.

Furthermore, the 1-posterior gains more information than the 0-posterior. Using \( n = 2 \) and \( \alpha = .1 \), the 1-posterior gained .113, whereas the 0-posterior gained .042. The gains were constant as we varied the sample means.

1.5. **Inference on the Gaussian with Unknown Variance.** Given a Gaussian variable \( x \) with density \( \frac{1}{\sqrt{2\pi\nu}} \exp\left(-\frac{(x-\mu)^2}{2\nu}\right) \) where \( \mu \) is known and fixed but we are
Figure 2. Plot of pdfs in Inference of Exponential Parameter, \( n = 2, \alpha = .2 \).

ignorant of \( \nu \), then by straightforward computation, the Fisher information is \( g = \nu^{-2} \) so the Uniform prior for \( \nu \) would be \( m(\nu) \, d\nu = 1 \, d\nu \propto \sqrt{g} \, d\nu = \frac{d\nu}{\nu} \). But being unnormalizable this is an improper distribution. The divergence of two Gaussians having respective parameters \( \theta = (\mu, \nu) \) and \( \theta_0 = (\mu_0, \nu_0) \) is

\[
I(\theta : \theta_0) = \frac{(\mu - \mu_0)^2}{2\nu_0} + \frac{\nu}{2\nu_0} - \frac{1}{2} - \frac{1}{2} \ln \frac{\nu}{\nu_0}.
\]

(1.21)

The 1-prior is \( \propto \exp(-\alpha I(\theta : \theta_0)) \, dV \propto \nu^{\frac{\alpha}{2}} \exp\left(\frac{-\alpha\nu}{2\nu_0}\right) \frac{d\nu}{\nu} \) and by swapping \( \theta \) and \( \theta_0 \) in \( I \), the 0-prior is \( \propto \exp(-\alpha I(\theta_0 : \theta)) \, dV \propto \nu^{-\frac{\alpha}{2}} \exp\left(\frac{-\alpha\nu_0}{2\nu}\right) \frac{d\nu}{\nu} \). By multiplying the likelihood of a sample of size \( n \) we get the respective posteriors:

\[
p_1(\nu) \, dV \propto \nu^{-\frac{n+\alpha}{2}} \exp\left(-\frac{n\hat{\sigma}_n^2}{2\nu} - \frac{\alpha\nu}{2\nu_0}\right) \frac{d\nu}{\nu} \quad \text{Note: more decay for \( \nu \) large}
\]

(1.22)

\[
p_0(\nu) \, dV \propto \nu^{-\frac{n+\alpha}{2}} \exp\left(-\frac{(n\hat{\sigma}_n^2 + \alpha\nu_0)}{2\nu}\right) \frac{d\nu}{\nu} \quad \text{Note: more decay for \( \nu \) small}
\]

(1.23)

In [10] the 1-posterior was claimed to show antidata via the following argument: The 0-posterior (1.23) is a Scaled Inverse Chi Squared pdf\(^7\), \( \chi^{-2}(n + \alpha, \hat{\sigma}_{n+\alpha}^2) \), where \( \hat{\sigma}_{n+\alpha}^2 \) stands for the pooled variance \( \frac{n\hat{\sigma}_n^2 + \alpha\nu_0}{\nu + \alpha} \). The 1-posterior is a Generalized Inverse Gaussian (GIG). Change the variable to \( 1/\nu \) and take the best second-order Gamma approximation (assuming \( \alpha << n \)), then change the variable back to \( \nu \). This was called the Gamma approximation to the 1-posterior, \( p_G(\nu) \propto \chi^{-2} \left(n - \alpha, \hat{\sigma}_{n-\alpha}^2\right) \),

\(^7\text{defined by } \nu \sim \chi^{-2}(\alpha = \text{df}, \nu_0 = \text{rate}) \text{ iff } \frac{\nu_0}{\nu} \sim \chi_\alpha^2.\)
where $\hat{\sigma}_{n-\alpha}^2$ is short for $\frac{n}{n-\alpha} \hat{\sigma}_n^2$. The approximation essentially drops the exponential decay factor $\exp\left(-\frac{\alpha \nu}{2 \nu_0}\right)$ from (1.22) on the basis that $\alpha$ is small. It was rightly claimed in [10] that $p_G$ is more ignorant than the $p_0$ in the sense that it is like the $p_0$ but with $n - \alpha$ degrees of freedom compared to the $n + \alpha$ df of $p_0$. And at first glance the $p_1$ and the $p_G$, plotted in terms of $1/\nu$, appear close. Their tails are close. See [10] p. 16 Fig. 3, where they are plotted ($x = 1/\nu$ there). But look at the low $x$ range and see that the curves diverge radically. By omitting the exponential decay factor, the tail of $p_G$ for high $\nu$ is much thicker. In fact, as indicated above, the $p_1$ has less spread in terms of $\nu$ than the $p_0$.

We want to compare the information gains of the 1-posterior and the posterior that follows from the Uniform prior, relative to the Uniform prior $m$. Since the volume element $m$ is not normalizable, in order to be cautious of misleading results we used a MIPPr and the corresponding posterior and we computed the difference $I(p_1 : m_1) - I(p_{1,\beta} : m_1)$ where $m_1$ is the 1-prior with infinitesimal estimate weight $\beta$ and $p_{1,\beta}$ is the resulting posterior. Figure 3 shows the relevant pdfs and the results for a typical iteration. (The pdf $m_0$, a 0-prior with weight $\beta$ is also shown, and is indistinguishable from $m_1$ for $\beta$ small. We were able to use as small as $\beta = 1e - 10$ in computations). The difference remained positive as we varied $\alpha$ from 1e-10 to 1, $n$ from 2 to 10, and $\nu_0$ from 1e-3 to 1e+6, so there was no show of antidata. Furthermore, for $n = 2$ and $\alpha = 1$, the 1- and 0-posteriors gained respectively .165 and .027 more than $p_U$ did, regardless of the sample variance value.

1.6. Inference on the Gaussian with Both Parameters Unknown. Suppose now the entire Gaussian parameter $\theta = (\mu, \nu)$ is unknown. For the two-dimensional space the Fisher information matrix entries are$^8$:

$$g_{ij} = \mathbb{E}_{x|\theta}[\partial_{\theta_i}(\ln p) \partial_{\theta_j}(\ln p)] = 4 \int_\Omega \partial_i \sqrt{p(x|\theta)} \partial_j \sqrt{p(x|\theta)} \, dx. \tag{1.24}$$

$^8$Note that [10] p.4 had an extraneous factor of $p$ in this integral.
The volume element (and improper Uniform prior) is $dV \propto \sqrt{|g|} \propto \nu^{-1.5} \, d\mu \, d\nu$. From (1.21) compute the 1-prior:

\begin{equation}
\frac{1}{\nu} \exp\left(-\frac{\alpha}{2\nu_0}((\mu - \mu_0)^2 + \nu)\right) \, d\mu \, d\nu
\end{equation}

\begin{equation}
\nu^{\frac{1}{2}} \exp\left(-\frac{\nu}{2\nu_0}\right) \, \nu \, d\mu \, d\nu \propto N(\mu_0, \frac{\nu_0}{\alpha}) \, d\mu \cdot \chi^2(\alpha - 1, \frac{\nu_0(\alpha - 1)}{\alpha}) d\nu,
\end{equation}

where the Scaled Chi Squared is given by $\zeta \sim \chi^2(\alpha = df, \nu_0 = \text{scale})$ iff $\frac{\alpha \zeta}{\nu_0} \sim \chi^2_\alpha$ iff $p(\zeta) d\zeta \propto \zeta^{\frac{\alpha}{2}} \exp\left(-\frac{\alpha \zeta}{2\nu_0}\right) \frac{d\zeta}{\zeta}$. The 1-prior parameters are independent, unlike the 0-prior case (see Observation 2.10 and commentary thereafter). The correct 1-posterior is:

\begin{equation}
\frac{1}{\nu} \exp\left(-\frac{\alpha}{2\nu_0}((\mu - \mu_0)^2 + \nu)\right) \, d\mu \, d\nu
\end{equation}

\begin{equation}
\nu^{\frac{1}{2}} \exp\left(-\frac{\nu}{2\nu_0}\right) \, \nu \, d\mu \, d\nu \propto GIG\left(\frac{\alpha - n - 1}{2}, \frac{n(\sigma_n^2 + (\bar{x}_n - \mu)^2)}{2\nu_0}, \frac{\alpha}{2\nu_0}\right) d\nu \cdot N(\mu_0, \frac{\nu_0}{\alpha}) d\mu.
\end{equation}

\textbf{Figure 3.} Comparison of $I(p_1 : m_1)$ versus $I(p_{1,\beta} : m_1)$: an iteration.
where $\tilde{\sigma}_n^2 = \frac{\sum(x_i - \bar{x}_n)^2}{n}$ and $GIG$ is Generalized Inverse Gaussian. For comparison, the 0-posterior is $p_0(\mu, \nu | x^n) dV \propto \nu^{-\frac{n+1}{2}} \exp \left( -n(\mu - \bar{x}_n)^2 + \tilde{\sigma}_n^2 - \alpha (\mu - \mu_0)^2 + \nu_0 \right) \frac{d\mu d\nu}{\nu^\nu \tau^\tau}$. Note that again the 1-posterior has more decay for large $\nu$ and the 0-posterior has more decay for small $\nu$. To get the marginal $\mu$ distribution, integrate over $\nu$.

What was done in [10] p. 17 was that the GIG was given a Gamma approximation. After such approximation, the marginal $\mu$ density will be $(n(\tilde{\sigma}_n^2 + (\bar{x}_n - \mu)^2)/2)^{-\frac{n+1}{2}}$. 

$N(\mu_0, \frac{\nu_0}{\alpha}) d\mu$ which is the product of a scaled $t$ pdf centered at $\bar{x}_n$ and a Normal pdf centered at $\mu_0$: $\mu \sim M \cdot N(\mu_0, \frac{\nu_0}{\alpha}) d\mu$ where $M$ is given by $\frac{(\mu - \bar{x}_n)\sqrt{n-\alpha}}{\sigma_n} \sim t_{n-\alpha}$. If we would ignore the Normal pdf on account of its typically high variance $\frac{\nu_0}{\alpha}$ then $\mu$ has roughly $n - \alpha$ degrees of freedom — this might make us think we see antidata because the marginal $\mu$ pdf of the posterior that follows from the Uniform prior is precisely $M$ but with $n$ df: $\frac{(\mu - \bar{x}_n)\sqrt{n}}{\sigma_n} \sim t_n$ and the 0-posterior’s has $n + \alpha$ df; essentially this was called antidata in [10] p. 18. But as we saw in the previous section, the impact of dropping the exponential $\nu$ decay factor when we get the Gamma approximation is not negligible, and to thereby estimate the 1-posterior’s df as $n - \alpha$ is inaccurate.

The array in Fig. 4 shows a snapshot of the priors and posteriors with parameters $n = 2$, $\alpha = .1$, and $\tilde{\sigma}_n^2 = 1$. It also shows the sample likelihood, and the marginal $\mu$ and $\nu$ posteriors. All posteriors are normalized. The plots were made in natural ($\mu, \tau$) coordinates, where $\tau = \nu^{-5}$, because then $dV \propto 1 d\mu d\tau$ and the pdfs are better

---

\[ p_\alpha(\mu, \nu | x^n) dV \propto \nu^{-\frac{n+1}{2} - 1} \exp(-n(\mu - \bar{x}_n)^2 + \tilde{\sigma}_n^2) \exp(-\alpha(\mu - \mu_0)^2) \frac{d\mu d\nu}{\nu_0^{\nu_0} \tau^{\tau}}. \]

Using the constant of integration for $\chi^{-2}$, we get a marginal $\mu \sim (n(\tilde{\sigma}_n^2 + (\bar{x}_n - \mu)^2)/2)^{-\frac{n+1}{2}} \cdot N(\mu_0, \frac{\nu_0}{\alpha}) d\mu$. This is a correction to [10] p. 18 which had $\frac{(\mu - \mu_n)\sqrt{n+\alpha}}{\sigma_n} \sim t_{n-\alpha}$ where $\mu_n$ is the pooled mean $(n\bar{x}_n + \alpha \mu_0)/(n + \alpha)$ of the $n$ real samples and $\alpha$ virtual samples.
behaved. Noteworthy is that the 1-posterior is more informative for $\mu$ and has thicker tail for low $\nu$ and thinner tail for high $\nu$.

A comparison of the information gains of the 1-posterior and of the posterior following from the MIPPr, relative to a MIPPr, as in the last section, revealed no antidata. The divergences were compared as we used $\bar{x}_n = 0$ and we varied $n$ from 2 to 8, $\hat{\sigma}^2_n$ from $\frac{1}{8}$ to 8 and $\alpha$ from .1 to 1e-4, and the MIPPr $p_{1,\beta}$ was a 1-prior with a negligible weight of $\beta=1e-8$. Furthermore, for $n = 2$ and $\alpha = .1$, the 1- and 0-posteriors gained resp. .190 and .042 more than the posterior that follows from the MIPPr, nearly regardless of $\hat{\sigma}^2_n$. 
Figure 4. Gaussian with parameters $(\mu, \nu)$ unknown: Priors, Likelihood, Posteriors and Marginals for $n = 2$, $\alpha = .1$, $\tilde{\sigma}_n^2 = 10$, and $\bar{x}_n = 0$. The vertical axes are $\tau = 1/\sqrt{\nu}$. 
2. Observations and Conjectures Regarding Entropic Priors

The observations that follow in Sections 2.1 and 2.2 are based on numerical computations and searches that we performed using R or Maple. See Sections 1.3-1.6 for descriptions of how the parameters were varied for the searches. Section 2.3 contains theory to explain these observations.

2.1. Regarding Strict Informativeness of the 1-Prior.

Observation 2.1. Cases where the 1-prior is strictly informative (that is, no antidata), in the sense of contributing positive information gain (see Section 1.13): In the cases of the Bernoulli, Exponential, Gaussian (with only $\nu$ unknown or with both params unknown), Bitnets $K_2$, $K_3$ and $L_3$ (see [9]), and Pareto models, the 1-posterior has greater information gain — that is, it has more Kullback divergence away from the Uniform Prior — than the posterior that follows from the Uniform prior has.

Conjecture 2.2. Strict informativeness of the 1-prior in terms of information gain: The above is true in general.

Observation 2.3. Cases where the 1-prior is strictly informative (that is, there is no antidata), in terms of the $\beta^*$ criterion (see Definition 1.15): If the 1-posterior with given prior weight $\alpha > 0$, is closest to the 0-posterior with prior weight $\beta^*$ as $\beta^*$ is allowed to vary, then $\beta^*$ is nonnegative, in the case of Bernoulli and Gaussian ($\nu$ unknown) models, Bitnets $K_2$, $K_3$ and $L_3$, and Pareto. For Gaussian (with only $\nu$ unknown) this was also verified by comparing numerical derivatives at $\beta^* = 0$ and at $\alpha = 0$.

Conjecture 2.4. Strict informativeness of the 1-prior, in terms of the $\beta^*$: The above is true in general.

2.2. Regarding Comparison of the 0- and 1-priors.
Observation 2.5. The 1-prior is usually — but not always — more informative than 0-prior: See Section 1 where most cases showed the 1-prior resulting in more posterior information gain than the 0-prior, and where the counterexample Obs. 2.6 below is also shown.

Observation 2.6. The 1-prior is more informative than the 0-prior in Bernoulli or Bitnet models the when a node parameter is extreme, and vice-versa when it is moderate: See Figure 1 in Section 1 which plots the divergences of the posterior from the uniform prior for the Bernoulli. The generalization to Bitnets follows from the same degeneracy of the 1-prior for Bitnets as occurs for Bernoulli when a parameter $p_i$ approaches zero or one.

Observation 2.7. For Bernoulli, a simple analytical illustration that the 1-prior can be more informative or less informative than the 0-prior, depending on whether the parameter is at the extremes: For any $\alpha > 0$, the 1-prior becomes degenerate (a spike) as the parameter $p$ approaches 0 or 1. The 0-prior does not do this. This shows that for extreme $p$ the 1-prior is more informative. Conversely, observe that $\alpha = 1$ and $p_0 = .5$ uniquely make the 0-prior into the naive flat prior $1 \, dp$ (that is, the estimate is enough to pull the curve $dV(p) = (\pi^2 p (1-p))^{-5}$ up in the middle, until it is “flat.” But the 1-prior for the same $\alpha$ and $\theta_0$ values maintains asymptotes at 0 and 1 since it will still have negative powers of $p$ and $1-p$ (and thus the estimate pulls the curve up in the middle (i.e. at the estimate $p_0 = .5$) to a lesser extent). This shows the 0-prior is more informative. But it is interesting to note that the 1-prior for these parameter values, though it flares up at the extremes, is still flat at $p = .5$ (i.e. it has a second derivative of zero there).

Observation 2.8. A 1-prior that is almost identical to the 0-prior: For Bernoulli with $n = 15$, $\alpha = 1$, $p_0 = .5$, the 1-prior with $\alpha = 1$ is numerically almost identical to (i.e. has nearly zero KL divergence from) the 0-prior with $\beta^* = .89$ and $p_0 = .5$ also). We computed the divergence as $3.28e-06$ with absolute error less than $5.4e-07$. 

18
Figure 5. A bimodal 1-posterior. The axis $\tau$ is $1/\sqrt{\nu}$. Gaussian with unknown parameters; prior estimate $(\mu_0, \tau_0) = (0, 2)$ with weight $\alpha = .25$, and sample statistics $(\bar{x}_n, 1/\sigma_n) = (4, 8)$ with $n = 2$.

Observation 2.9. The 1-prior offers possibilities of more spread, or even multiple modes, in the posterior, in certain cases when the estimate and sample are “distant”: The 1-prior is generally non-conjugate, so if the estimate differs from the sample, it results in a posterior that can be more diffuse than the 0-posterior, or even multimodal (see Figure 5), in contrast to the conjugate prior which is typically unimodal.\footnote{see footnote 2} As Carlos Rodríguez has noted (May 2011, at the author’s Oral Exam), the fact that the 1-posterior can be multimodal is an advantage (and, I add, a sign of more informativeness) in that it thereby indicates the estimate is “wrong,” that is, distant from the sample.

In relation to this see Jaynes [2] p. 311: “...if the predictions prove to be wrong, then induction has served its real purpose; we have learned that our hypotheses are wrong
or incomplete, and from the nature of the error we have a clue as to how they might be improved... As Harold Jeffreys explained long ago, induction is most valuable to a scientist just when it turns out to be wrong; only then do we get new fundamental knowledge;” and p. 326, “Put more strongly, it is only when our inductive inferences are wrong that we learn new things about the real world.”

Observation 2.10. For Gaussian with both parameters unknown, the 1-prior has \( \mu \) and \( \nu \) independent: See formula 1.26.

This contrasts with the 0-prior, where \( \mu \) is dependent upon \( \nu \). And this gives some justification for priors that do have \( \mu \) and \( \nu \) independent, such as the naive “standard non-informative prior” in (4.7).

2.3. Theory for Explaining the Above Observations Regarding Information of the 1-Prior, and a Proof Sketch of the Conjectures. Let \( x \) be an Exponential Family random variable with unknown parameter \( \theta \), and suppose the distribution has more spread as \( \theta \) increases (for, if \( \theta \) is a location parameter then the 0- and 1-priors coincide; see [10] p.11). We simplify the circumstances for clarity. Namely, let \( x \) have likelihood function

\[
p(x|\theta) \propto \exp(C_1(\theta)T_1(x) - T(x) - \ln Z(\theta))
\]

and suppose that \( T_1(x) \) is a sufficient statistic for \( \theta \) such that

\[
\tau_1(\theta) \equiv E[T_1(x)|\theta] = \theta.
\]

We will give the entropic prior a weight of \( \alpha \) virtual observations, with an estimated \( \theta \) value of \( \theta_0 \), and will take \( n \) actual samples to get the posteriors. Suppose further that the estimate is a perfect guess, so that the sample statistics coincide with \( \theta_0 \):

\[
T_1(x) = \theta_0.
\]

We will need the computation

\[
I(\theta_0 : \theta) \propto (C_1(\theta_0) - C_1(\theta))\tau_1(\theta_0) - \ln \left( \frac{Z(\theta_0)}{Z(\theta)} \right).
\]
Then the 0-prior is

\[(2.15) \quad p_0(\theta | \theta_0, \alpha) \propto \exp(-\alpha I(\theta_0 : \theta))\]

\[(2.16) \quad \propto \exp(\alpha C_1(\theta) \tau_1(\theta_0)) / Z^\alpha(\theta)\]

\[(2.17) \quad = \exp(\alpha C_1(\theta) \theta_0) / Z^\alpha(\theta)\]

and the 1-prior is

\[(2.18) \quad p_1(\theta | \theta_0, \alpha) \propto \exp(-\alpha I(\theta : \theta_0))\]

\[(2.19) \quad \propto \exp(\alpha(C_1(\theta_0) - C_1(\theta)) \tau_1(\theta)) \cdot Z^\alpha(\theta)\]

\[(2.20) \quad = \exp(\alpha(C_1(\theta_0) - C_1(\theta)) \theta) \cdot Z^\alpha(\theta)\]

The key to the theory lies in these three observations:

#1. The function $C_1(\theta)$ is a negative-valued function diverging monotonically to $-\infty$ as $\theta \to 0$, and converging to 0 as $\theta \to \text{Large}$.

#2. Where does the prior have exponential decay — on the high $\theta$ end or the low $\theta$ end? Answer: high for 1-prior, low for 0-prior.

#3. The volume element is typically inversely proportional to $\text{Polynomial}(\theta)$.

To illustrate these points, consider the following examples. For a Gaussian with fixed mean and unknown variance $\theta$, we have $C_1(\theta) = -\frac{1}{2\theta}$. For a Bernoulli with likelihood $p(x|p) = \theta^x(1-\theta)^{1-x}$, we have $C_1(p) = \ln(p/q)$. This illustrates #1. Now look at $C_1$ in (2.17) and (2.20) and observe #2.

The consequence is that the 1-posterior has an exponential decay as $\theta$ gets large; and this exponential decay is absent on the low $\theta$ end. Conversely, the 0-prior has an exponential decay as $\theta$ becomes small; and this exponential decay is absent on the high $\theta$ end. Under our assumptions, the likelihood of the $n$ observations is identical to the 0-prior but with weight $n$ instead of $\alpha$. Now, of the two posteriors below, which should we expect to be most informative? —

\[(2.21) \quad p_0(\theta | \theta_0, \alpha, x^n) \propto p_0^n \cdot p_0^\alpha\]
or

\[ p_1(\theta|\theta_0, \alpha, x^n) \propto p_0^n \cdot p_1^\alpha ? \]

We should expect the 1-posterior (2.22) to be more informative. This is simply because the likelihood (like the 0-prior) already has an exponential decay in the low \( \theta \) end. Multiplying it by another \( \alpha \) factors of \( p_0 \) which have no exponential decay in the high \( \theta \) will not nearly as much heap the peak upwards as will multiplying by a complementary \( p_1 \) factor which does decay in the high \( \theta \)'s. In other words, the 1-posterior \( p_0^n \cdot p_1^\alpha \) applies an exponential decay on both sides of the curve instead of only on one side, the latter being what the 0-posterior does. This exponential decay, even though its argument has only a small coefficient \( \alpha \), has a profound effect on the scale of the distribution. Recall that exponentiality is what makes the difference between a t-distribution with no higher moments and a Gaussian distribution with finite moments of every order.

But these two observations, #1 and #2 are not conclusive enough without observation #3, that the volume element typically decays as \( 1/\text{Poly}(\theta) \); for example \( dV \propto d\theta/\theta \) for the Gaussian with \( \nu \) unknown, and \( dV \propto \frac{1}{\sqrt{pq}} \) for the Bernoulli, which means that heaping the pdf up more towards the low \( \theta \) end (which is what the \( p_1^\alpha \) factor does) rather than more towards the high end (which is what \( p_0^n \) would do) makes more of an information gain impact. This completes the proof sketch and explains why the 1-prior is either usually more informative than the 0-prior, or at least strictly informative.
3. INTERPRETING NAIVE VOLUME ELEMENTS IN BITNETS

Suppose we are given a Bernoulli random variable, and our state of knowledge of the parameter \( p \), the probability that the node is “On,” is that we are completely ignorant of it. Suppose now we are given that there has been one observation of On. What is the probability that the next observation will be On? If taken as a single-number summary of our knowledge, the method of maximum likelihood clearly overfits here, for the modal likelihood is \( p = 1 \). This is an example of that which is known, but worth repeating, namely that when data is scant, maximum likelihood (i.e. mode of the likelihood) overfits, and Bayesian methods are more useful.

But in fact, if we do the Bayesian analysis and use the Jeffreys prior here (i.e., volume element proportional to square root of Fisher information), the MAP (maximum a posteriori) \( p \) will also be one. So, rather, let us consider the mean of the posterior (the minimizer of the quadratic loss function) as our single-number summary, which should provide more information about our state of knowledge. The mean of the likelihood is \( \frac{2}{3} \). What is the mean of the posterior?

Statisticians are becoming more aware that in Bayesian analysis we should not assume that a naive “flat” prior \( 1 \, d\theta \) (where \( \theta \) is the chosen parameter) represents true ignorance. Such a prior is generally not invariant under coordinate transformations. For example, in the unknown Bernoulli mentioned above, \( 1 \, dp \) which naively appears flat actually transforms to \( \frac{e^{-\psi}}{(1-e^{-\psi})^2} \, d\psi \), where \( \psi \) is the log odds \( \log \left( \frac{p}{1-p} \right) \), which no longer appears flat in \( \psi \).

The true uniform prior (if normalizable) is the normalized information volume element, the normalized square root of the Fisher metric tensor, which is a geometric fact of the Riemannian manifold, the latter being what a regular parametric model (e.g. Bernoulli (a curve), Gaussian(\( \mu, \nu \)) (a surface, the hyperbolic half-plane), etc.) truly is. The volume prior is invariant under the arbitrary choice of parameterization; so for example for the change of variable to \( \psi \) we have

\[
1 \, dV(p) = \frac{1}{\pi \sqrt{p(1-p)}} \, dp = (2\pi \cosh(\psi/2))^{-1} \, d\psi = 1 \, dV(\psi).
\]
The Bernoulli distribution (= solo Bitnet) and Bitnets (see [9]) illustrate very simply and saliently the impact of a naive flat prior, versus the volume prior. Let us compare:

3.1. **Bernoulli.** Recall that for a Bernoulli likelihood \( p(x|\theta) = \theta^x(1 - \theta)^{1-x} \), \( x \in \{0, 1\} \), \( 0 \leq \theta \leq 1 \), the volume element of the hypothesis space \( \{\theta|0 \leq \theta \leq 1\} \) is \( 1\ dV \propto \sqrt{g} \ d\theta = \frac{d\theta}{\sqrt{\theta(1-\theta)}} \) where \( g \) is Fisher information; this is the true uniform prior. Therefore putting a naive flat prior of \( 1\ d\theta \) can be interpreted as actually presuming one unwarranted observation of \( \bar{x}_{n-1} = 0.5 \). In other words,

\[
(3.2) \quad 1\ d\theta = \sqrt{p(0|\theta)p(1|\theta)} \ dV.
\]

And it corresponds to an entropic 0-prior of \( p_0(\theta|\alpha = 1, \theta_0 = .5) \ dV \). If we have only a small sample size of for instance between one and four then this prior is a very significant assumption. Even with more observations, much inferential power can be wasted by this presumption. For example, if we continue to get “On” observations, using the naive flat prior causes us to take twice as long (or twice as many lives, resources, money, etc.) to reach the same estimates that the true uniform prior will generate; see Table 1. For instance, as the table shows, with the volume uniform prior it takes only four Ons to get a posterior probability \( p(On) = .9 \), whereas with the naive flat prior it will take eight Ons for this. In terms of the 1-prior the naive flat prior is even worse, for it corresponds to an even greater number of assumed observations, namely 2.575, for \( p_1(\theta|\alpha = 2.575, \theta_0 = .5) \ dV \) is the 1-prior that is closest to \( 1\ d\theta \), at a divergence of 0.03264. (See Figure 1 in *A Note on Antidata* for an depiction of how the informativeness of these two priors varies as \( \theta_0 \) varies).

Formulae and computations indicate that higher bitnets follow this pattern, that the 1-prior is less informative than the 0-prior for a moderate central node estimate, but then becoming more informative as the estimate becomes extreme. But despite this, no antidata has been found in any of the Bitnets \( K_1, K_2, K_3, \) and \( L_3 \). In particular, the computation of posterior variances as shown in the table tells us that extreme
### Table 1. Priors and Posteriors for Bernoulli

<table>
<thead>
<tr>
<th>Prior</th>
<th>Naive flat</th>
<th>True volume-uniform</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>$1 , d\theta$</td>
<td>$\frac{1 , d\theta}{\pi \sqrt{\theta(1-\theta)}}$</td>
</tr>
<tr>
<td>Posterior with</td>
<td>$\propto \text{Prior} \cdot \text{Likelihood}$</td>
<td>$\propto \text{Prior} \cdot \text{Likelihood}$</td>
</tr>
<tr>
<td>$a$ Ons and $b$ Offs</td>
<td>$\text{Beta}(a + 1, b + 1) , d\theta$</td>
<td>$\text{Beta}(a + \frac{1}{2}, b + \frac{1}{2}) , d\theta$</td>
</tr>
<tr>
<td></td>
<td>Its mean is $\frac{a+1}{a+b+2}$.</td>
<td>Its mean is $\frac{a+2}{a+b+1}$.</td>
</tr>
<tr>
<td></td>
<td>Its variance is $\frac{(a+1)(b+1)}{(a+b+2)^2(a+b+3)}$.</td>
<td>Its variance is $\frac{(a+\frac{1}{2})(b+\frac{1}{2})}{(a+b+1)^2(a+b+2)}$.</td>
</tr>
</tbody>
</table>

| $p(x=1|(a,b))$        | $\int \theta \cdot \text{Beta}(a + 1, b + 1) \, d\theta$ | $\int \theta \cdot \text{Beta}(a + 1/2, b + 1/2) \, d\theta$ |
|                       | $= \frac{\Gamma(a+b+2)}{\Gamma(a+1)\Gamma(b+1)} \cdot \frac{\Gamma(a+2)\Gamma(b+1)}{\Gamma(a+b+3)}$ | $= \frac{\Gamma(a+b+1)}{\Gamma(a+1/2)\Gamma(b+1/2)} \cdot \frac{\Gamma(a+3/2)\Gamma(b+1/2)}{\Gamma(a+b+2)}$ |
|                       | $= \frac{a+1}{a+b+2}$ | $= \frac{a+1/2}{a+b+1}$ |

| $p(x=1|(0,0))$        | $1/2$ | $1/2$ |
| $p(x=1|(1,0))$        | $2/3$ | $3/4$ |
| $p(x=1|(2,0))$        | $3/4$ | $5/6$ |
| $p(x=1|(3,0))$        | $4/5$ | $7/8$ |
| $p(x=1|(4,0))$        | $5/6$ | $9/10$ |
| $p(x=1|(5,0))$        | $6/7$ | $11/12$ |
| $p(x=1|(6,0))$        | $7/8$ | $13/14$ |
| $p(x=1|(7,0))$        | $8/9$ | $14/15$ |
| $p(x=1|(8,0))$        | $9/10$ | $15/16$ |
| ...                   | ... | ... |
| $p(x=1|(49,0))$       | $50/51$ | $99/100$ |
| $p(x=1|(98,0))$       | $99/100$ | $197/198$ |
| $p(x=1|(a,0))$        | $\frac{a+1}{a+2}$ | $\frac{2a+1}{2a+2}$ |

Data samples (i.e. those with (nearly) all Ons or Offs) enjoy a great reduction in posterior variance by using the correct volume element versus a naive flat one:

**Observation 3.3.** Posterior variance reduction enjoyed by extreme samples when using the correct volume prior. The posterior variances computed in Table 1 show that for $a$ or $b$ (nearly) zero, the posterior variance of a Bernoulli
parameter $p$ by using the correct volume-element prior is nearly half of that given by using a naive flat prior. Even more variance reduction is thereby enjoyed in the lower (bottom-most) node(s) of higher Bitnets, because naive flat priors especially distort the volume-uniform element with regard to the lower nodes — see below.

This is likely the first explicit statement of this simple observation.

3.2. Bitnet of Two Nodes. Here the hypothesis space is $\{(\theta_1, \theta_2, \theta_3)|0 \leq \theta_i \leq 1\}$ where $\theta_1$ is the probability that node 1 is On, and $\theta_2$ and $\theta_3$ are the conditional probabilities that node 2 is On given that node 1 is Off or On. The volume element is $dV \propto \theta_1(1-\theta_1) d\theta$. Therefore a naive flat $d\theta$ prior is equivalent to assuming two hypothetical observations — one observation of $(x_2|x_1 = 0) = .5$, and one of $(x_2|x_1 = 1) = .5$. In other words,

$$1d\theta = \sqrt{p(x_2 = 0|x_1 = 0)p(x_2 = 1|x_1 = 0)p(x_2 = 0|x_1 = 1)p(x_2 = 1|x_1 = 1)} dV.$$ (3.4)

3.3. Higher Bitnets. The complete bitnet $K_3$ has

$$1dV \propto \frac{\theta_1(1-\theta_1) d\theta}{\sqrt{\theta_4(1-\theta_4)\theta_5(1-\theta_5)\theta_6(1-\theta_6)\theta_7(1-\theta_7)}}.$$ (3.5)

So using the naive flat prior is equivalent to having four hypothetical observations of $x = \frac{1}{2}$ at the final (leaf) node, along with the annihilation of the two observations 0 and 1 at node 1. In other words,

$$1d\theta = \sqrt{\prod_{x_3,x_1,x_2 \in \{0,1\}} p(x_3|x_1,x_2) p(x_1 = 0|\theta_1) p(x_1 = 1|\theta_1)} dV.$$ (3.6)

The directed line $L_3$ has

$$1dV \propto \sqrt{\frac{[\theta_2 + \theta_1(\theta_3 - \theta_2)](1 - [\theta_2 + \theta_1(\theta_3 - \theta_2)])}{\theta_4(1-\theta_4)\theta_5(1-\theta_5)\theta_6(1-\theta_6)\theta_7(1-\theta_7)}} d\theta$$

(3.7)

so the naive flat prior is also like four hypothetical conditional observations at the leaf node, but combined with the annihilation of one unconditional observation of the central node. For, the numerator $[\theta_2 + \theta_1(\theta_3 - \theta_2)](1 - [\theta_2 + \theta_1(\theta_3 - \theta_2)])$ under the radical is the probability that the central node is On, times the probability that it is Off.
4. Applications

We revisit a simple example from the WinBUGS documentation, showing that the fixed effects model analysis can be improved by using the correct volume prior, and that the random effects model analysis can be performed more flexibly and more satisfactorily by using the correct volume element and an entropic 0-prior. In the process we make a new observation regarding the ease of implementing an entropic prior for the hyperparameters thanks to the independence of the observations and the hyperparameters, conditional on the intermediate parameters. We also discover how in general to quantify the informativeness of a prior in a given inference problem.

Analysis of the WinBUGS Surgical Example

WinBUGS has an example that analyses data consisting of mortality frequencies after a certain surgical procedure (cardiac surgery in infants) performed at twelve hospitals, labeled A through L. Namely, there were

\begin{equation}
    y = (0, 18, 8, 46, 8, 13, 9, 31, 14, 8, 29, 24)
\end{equation}

\begin{equation}
    n = (47, 148, 119, 810, 211, 196, 148, 215, 207, 97, 256, 360)
\end{equation}

deaths in procedures.

Fixed Effects Model. The example states ([3]):

The number of deaths \( y_i \) for hospital \( i \) are modelled as a binary response variable with “true” failure probability \( p_i \):

\begin{equation}
    y_i \sim \text{Binomial}(p_i, n_i)
\end{equation}

We first assume that the true failure probabilities are independent (i.e. fixed effects) for each hospital. This is equivalent to assuming a standard non-informative prior distribution for the \( p_i \)'s, namely:

\begin{equation}
    p_i \sim \text{Beta}(1.0, 1.0)
\end{equation}
As we have stated in Section 3.1, in order to be a truly non-informative prior this ought be $p_i \sim \text{Beta}(0.5, 0.5)$. And as an application of what we have computed in Section 3, for hospital $A$ with the true volume prior we have posterior $p_i$ having mean $\frac{0.5}{48} = .0104$ and variance $\frac{47.5/2}{48^2 \cdot 49} \approx 2.101e-4$, versus with the naive flat prior a mean $\frac{1}{49} = .0204$ and variance $\frac{48}{49^2 \cdot 50} \approx 4.001e-4$ — thus the posterior variance has been cut roughly in half for this extreme sample by using the correct volume prior, as stated in Obs. 3.3. However the variance shrinkage is less for hospitals with more moderate data. The other samples have been reworked as well and they are compared with WinBUGS’s analysis in Fig. 6, which shows central 95% posterior credible intervals (CrI’s):

As the boxplots in Fig. 7: show, since the sample proportions are close to 0, the Volume Prior results in means that are somewhat lower.
Figure 7. WinBUGS Surgical Example, Independent Binomials: Posterior Means. The boxplots show the quartiles of the twelve posterior means for each of the two methods, with the volume prior resulting in slightly lower means.

More interestingly, we see substantial shrinkage in the 95% CrI widths by the volume prior versus the naive prior. See Fig. 8.

Random Effects Model. The WinBUGS Documentation points out that a random effects model is more realistic:

A more realistic model for the surgical data is to assume that the failure rates across hospitals are similar in some way. This is equivalent to specifying a random effects model for the true failure probabilities as follows:

\[
\text{logit}(p_i) = b_i
\]

(4.5)

\[
b_i \sim \text{Normal}(\mu, \tau[\equiv 1/\nu])
\]

(4.6)
Figure 8. WinBUGS Surgical Example, Independent Binomials: Posterior 95% CrI Widths

Standard non-informative priors are then specified for the population mean (logit) probability of failure, $\mu$, and precision, $\tau$.

The WinBUGS code goes on to specify the “non-informative priors” — ad-hoc — as

$$\mu \sim N(0, \tau = 1 \times 10^{-6});$$

(4.7)  

$$\tau \equiv 1/\nu \sim Gamma(1 \times 10^{-3}, 1 \times 10^{-3}),$$

which accomplishes a convenient normalization of the prior space that, being very disperse, would naively seem to be not far from volume element (however, we will see that this prior is significantly informative). For, when we consider the likelihood of $(\mu, \tau)$ which is the density of the complete data,

(4.8)  

$$p(y_i, b_i | \mu, \tau) = p(y_i | b_i) p(b_i | \mu, \tau)$$

(where the true proportion $p_i$ is censored to us) then conditionally on $b_i$ the volume element (by Fisher information) and entropic prior for the hyperparameters $\mu$ and $\tau$. 

30
depend only on the latter term. Namely, the volume element is the improper

\[ d\mu d\tau/\sqrt{\tau}, \]  

the same as in the inference of a Gaussian with unknown parameters.

In fact, what we have just mentioned is also the case in general for hierarchical models, so we state it as an observation that is, to our knowledge, new:

**Observation 4.10. Independence, conditionally on immediate parameter, of observations and hyperparameter.** Given a likelihood \( p(y|\theta) \), suppose \( \theta \) is dependent on hyperparameter \( \phi \). Then the likelihood of the complete data factors as

\[ p(y, \theta|\phi) = p(y|\theta) p(\theta|\phi), \]  

and therefore, conditionally on \( \theta \), the volume element of, and entropic prior for, \( \phi \) depend only on the latter factor, since the first factor is constant wrt \( \phi \).

This makes the entropic prior easy to apply. And even if we have many levels of hyperparameters, then conditionally on the intermediate parameters, the volume elements and entropic priors for the topmost hyperparameters depend only on the likelihoods specified at the topmost node(s).

Now let us execute this model more objectively than with the ad-hoc prior above, as an exercise and in order to see what differences will arise.

**Random Effects with the 0-Entropic Prior.** The volume element (4.9) for the hyperparameter \( \theta = (\mu, \nu) \) is improper. We will use a 0-prior for \( \theta \). Thus for each \( i = 1 \ldots 12 \), we have

\[ p(b_i|y_i, \mu, \nu) \propto p(y_i|b_i) p(b_i|\mu, \nu); \]  

\[ (\mu, \nu) \sim 0\text{-prior}(\mu_0, \nu_0, \alpha), \text{ namely} \]  

\[ \mu \sim N(\mu_0, \nu/\alpha) \]  

\[ \nu \sim \chi^{-2}(\alpha, \nu_0). \]
From the data above, we chose prior estimates $\theta_0$ given by

\begin{align}
\hat{p}_i &= \frac{y_i + 0.5}{n_i + 1} \\
(4.17) \\
\hat{b}_i &= \text{logit}(\hat{p}_i) \\
(4.18) \\
\mu_0 &= \hat{b}_i \\
(4.19) \\
\nu_0 &= \left(\hat{b}_i - \mu_0\right)^2 \\
(4.20)
\end{align}

$\approx -2.72$ 

We can use $\alpha = 0.1$ as a somewhat informative value, borrowing strength from the samples. Then the 0-prior is

\begin{align}
p_0(\mu, \nu|y_i, n_i, 0.5) \propto N(-2.72, \nu/1) \, d\mu \cdot \chi^{-2}(0.1, 0.514) \, d\nu. \\
(4.21)
\end{align}

This is quite different from WinBUGS’s prior (4.7),

\begin{align}
\text{WinBUGS’s ad-hoc} \propto N(0, 1 \times 10^6) \, d\mu \cdot \chi^{-2}(0.002, 1) \, d\nu, \\
(4.22)
\end{align}

since the latter has $\mu$ and $\nu$ independent (as the 1-prior (1.26) does), whereas the 0-prior has $\mu$ dependent on $\nu$. Also, the 0-prior has $\nu$ mode $\frac{0.1-0.514}{2+0.1} = 0.245$, compared to WinBUGS’s 1e-4; the means of $\nu$ are both divergent in either case. (Recall $\chi^{-2}(\nu, \sigma^2)$ has divergent mean for $\nu \leq 2$.) Our mode 0.245 seems much more reasonable, for, the observed variance is 0.514.

The results are compared with WinBUGS’s analysis in Fig. 9. Of the $\alpha$ weights used, namely, $1e - 5, .001, 1, 1$ and 10, the posteriors with $\alpha = 1$ look most similar to those of the WinBUGS documentation; with smaller $\alpha$ the posteriors were more shrunk towards the mean, and with higher $\alpha$ they were more spread apart. Using a lower $\nu_0$ estimate, for example, one-fourth of the sample variance, caused the posteriors to be more shrunk towards the mean with increasing $\alpha$, whereas using a higher $\nu_0$ cause more spread with increasing $\alpha$. See Fig. 10, where $\nu_0$ equal to four times the sample variance was used, which is not a very unreasonable adjustment, since the number of hospitals, twelve, is small, and more importantly, the population of $b_i$’s may be more heavy-tailed than Gaussian. Actually the very modelling of the $b_i$ as
Gaussian-generated is a major contribution of prior information; the uniform prior on $b_i$’s would be inverse-cosh, which is thicker-tailed.

We claim success by stating that with $\alpha = 10$ and this higher $\nu_0$, we can clearly distinguish four leagues, according to the upper 97.5 posterior percentiles (that is, using an upper bound on the fatality rate as the criterion). Namely, rounding to the nearest five percent, hospitals A, D and E have no more than five percent fatalities; hospitals C, F, G, I and L have no more than ten percent; hospitals B, J and K no more than fifteen percent; and hospital H has no more than twenty percent. This is a vast improvement in spread compared to the method of the WinBUGS documentation, where the posteriors were so homogenous that no conclusive league partitioning was derived. Thus we have overcome some of what the documentation called ”the considerable uncertainty associated with ‘league tables’”. Further, note that all the other the settings of $\alpha$ and $\nu_0$ mentioned, besides the WinBUGS documentation’s
method, result in the upper bound of hospital A — which alone, at 0 of 47, has perfect record — to be higher than those of D and E, which we think is undesirable. Hence we consider our set of rankings by this last method to be most desirable, and we expect such a strategy to be applicable to other ranking problems. To summarize, when we wish to build a ranking or league table by a random effects model when we do not have reason to believe that the tails are not heavy, we should consider using a 0-entropic prior with a moderately high $\alpha$ weight on the hyperparameters and an estimated variance that is higher than the sample variance, especially if the sample is small.

Performing the inference with the 1-prior showed similar results, but with posteriors more shrunk towards the mean than with the 0-prior, as we expect due to the complementarity of the 1-prior with the likelihood mentioned in Section 2.3. See Figure 11. In particular, Hospital A is especially shrunk towards the other posteriors by
Figure 11. Posterior 95% CrI’s, 0-prior vs. 1-prior: \( \nu_0 = 4 \cdot var(\hat{b_i}) \)
and \( \alpha = 10 \).

the 1-prior, and, building the league table by rounding the 97.5th posterior percentile to the nearest .05 as previously, the only change is that Hospital A is demoted to the second highest league. Preferring to not allow the perfect-record hospital to be so demoted, the author favors using the 0-prior here.

Our choice of the \( \alpha, \nu_0 \), and the \( \delta \) above is somewhat ad-hoc and motivated by the aim of producing spread for the sake of ranking; but it is no less arbitrary than the choice of modelling the \( b_i \)'s as being Gaussian-generated versus something heavier-tailed, closer to the uniform inverse-cosh. Given a lack of prior information we may be more faithful to our state of ignorance by proceeding to assign priors to these hyperparameters and to the tail-thickness of the model.

This basic example has served to show that the fixed effects model analysis can be improved by using the correct hypothesis space volume element, and that the random effects model analysis can be performed more flexibly and more satisfactorily.
by using the correct volume element and an entropic 0-prior. And in the process we have made a new observation regarding the ease of implementing an entropic prior for the hyperparameters which is thanks to the independence of the observations and the hyperparameters, conditional on the intermediate parameters. In larger models we expect these benefits to be more pronounced, due to the greater distortion of the true hypothesis space volume element caused by so-called standard diffuse priors in larger models, as shown in Section 3, and the sometimes unexpected geometry of hypothesis spaces (see the surprising shape of the hypothesis space for simple logistic regression in [11] pp. 6-7 and the superior analysis achieved there by using the correct volume element).

We make one more observation. As we mentioned above, Figure 9 shows that the results from the “standard noninformative prior” used in the WinBUGS documentation are practically indistinguishable from the 0-prior with weight $\alpha = 1$. This gives a way to quantify the actual contextual “informativeness” of such a prior — namely, it has informativeness of order approximately 1 (with $\theta_0$ equal to the estimated sample statistics). Were it truly uninformative, it would yield results closer to those shown for $1e^{-5}$. We do not object to it being informative per se — we ultimately chose to borrow much strength and use $\alpha = 10$. But we ought to know how much a prior is actually informing the results, and this demonstrates how to do so, given that the results of some 0-prior are close to those results. To generalize —

**Observation 4.23. Quantifying the informativeness of prior.** If an entropic 0-prior is available for an inference problem, and the results of an inference by using some other prior are (nearly) indistinguishable from those resulting from a 0-prior with parameters $\alpha$ and $\theta_0$, then these $\alpha$ and $\theta_0$ quantify how much the prior informs the results of that inference.

In our case we were fortunate to have $\alpha$ and $\theta_0$ values that mimic the naive prior’s results, but in general there may be no $\alpha$ and $\theta_0$ that so closely mimic them with one
0-prior. In such cases we might generalize by quantifying in terms of a mimicry by a mixture of 0-priors, or by $\delta$-prior(s) that may fit better.

**References**


