Wrong Priors

Carlos C. Rodríguez

http://omega.albany.edu:8008/
Department of Mathematics and Statistics
The University at Albany, SUNY
Albany, NY 12222
USA

Abstract. All priors are not created equal. There are right and there are wrong priors. That is the main conclusion of this contribution. I use, a cooked-up example designed to create drama, and a typical textbook example to show the pervasiveness of wrong priors in standard statistical practice.

Keywords: Information Geometry, Volume Prior, Bayesian Inference, Bayesian Information Geometry, Ignorance Priors
PACS: 02.40.-k,02.50.Tt

INTRODUCTION

The information geometry available in regular statistical models can be used to build objectively meaningful prior distributions. When the information volume of the model is finite, the uniform distribution over the model manifold coincides with Jeffreys invariant rule. A simple example in two dimensions shows that the popular naive diffuse prior over the parameters of this model is in fact a *wrong* prior, requiring more than ten thousand observations to match Jeffreys rule with only 100 samples. Bayesian inference is suffering from an epidemic of wrong priors and to prove it I consider standard simple logistic regression with naive diffuse priors and with uniform priors over the manifold. The results are obviously less dramatic but similar to the previous cooked-up example. When the information volume of the model is infinite, the uniform distribution over the model does not exist. However, the available geometry can still be exploited and it provides a semiparametric family of invariant objectively ignorant priors.

A SIMPLE EXAMPLE

Consider bivariate normals with unit covariance matrix and mean vector restricted to a region of the euclidean plane. Specifically, for given values a and b the experiment consists of choosing (x,y) at random on the euclidean plane with,

$$\begin{aligned} x &= \exp\left(-50(a^2+b^2)\right)+\varepsilon_1 \\ y &= \exp\left(-50((a-c)^2+(b-c)^2)\right)+\varepsilon_2. \end{aligned}$$

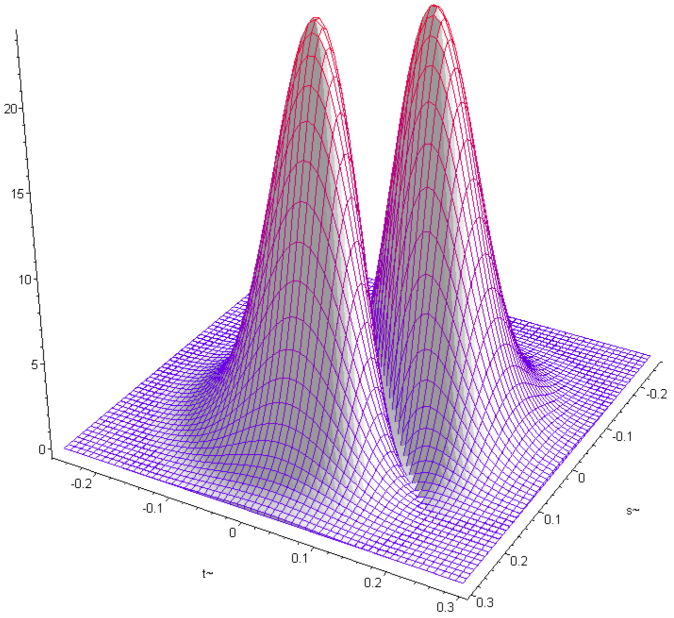

FIGURE 1. The prior π_1 is the true uniform over the model. Notice that the picture is not drawn at scale. The actual peaks should be more than 40 times taller than the ones displayed.

The unknown parameters are $a, b \in R$ but $c = 0.1$ is assumed known, and ε_1 and ε_2 are independent standard normals. The problem consists of learning the parameters $\theta = (a, b)$ from n independent observations $(x_1, y_1), \ldots, (x_n, y_n)$. We want to compare the performance of two priors on (a, b). The naive "‘ignorant"’ prior π_0 that takes a and b independently from $N(0, 100)$ and the uniform prior over the manifold model, π_1 given by,

$$\pi_1(a,b) = \frac{|a-b|}{Z} \exp\left(-100(a^2+b^2) + 10(a+b)\right) \tag{1}$$

where Z is a finite normalization constant. Equation (1) is just the normalized volume form of the model computed trivially as $\sqrt{\det g}/Z$ with g as the information matrix (minus the expected values of the second derivatives of the log likelihood). This prior puts positive mass on the entire (a, b) plane (except on the line $a = b$ but that region has measure 0) but it is very far from uniform as it is shown in figure 1. Notice also that there are two peaks because the likelihood is invariant under the exchange of a with b. The volume prior respects this symmetry.

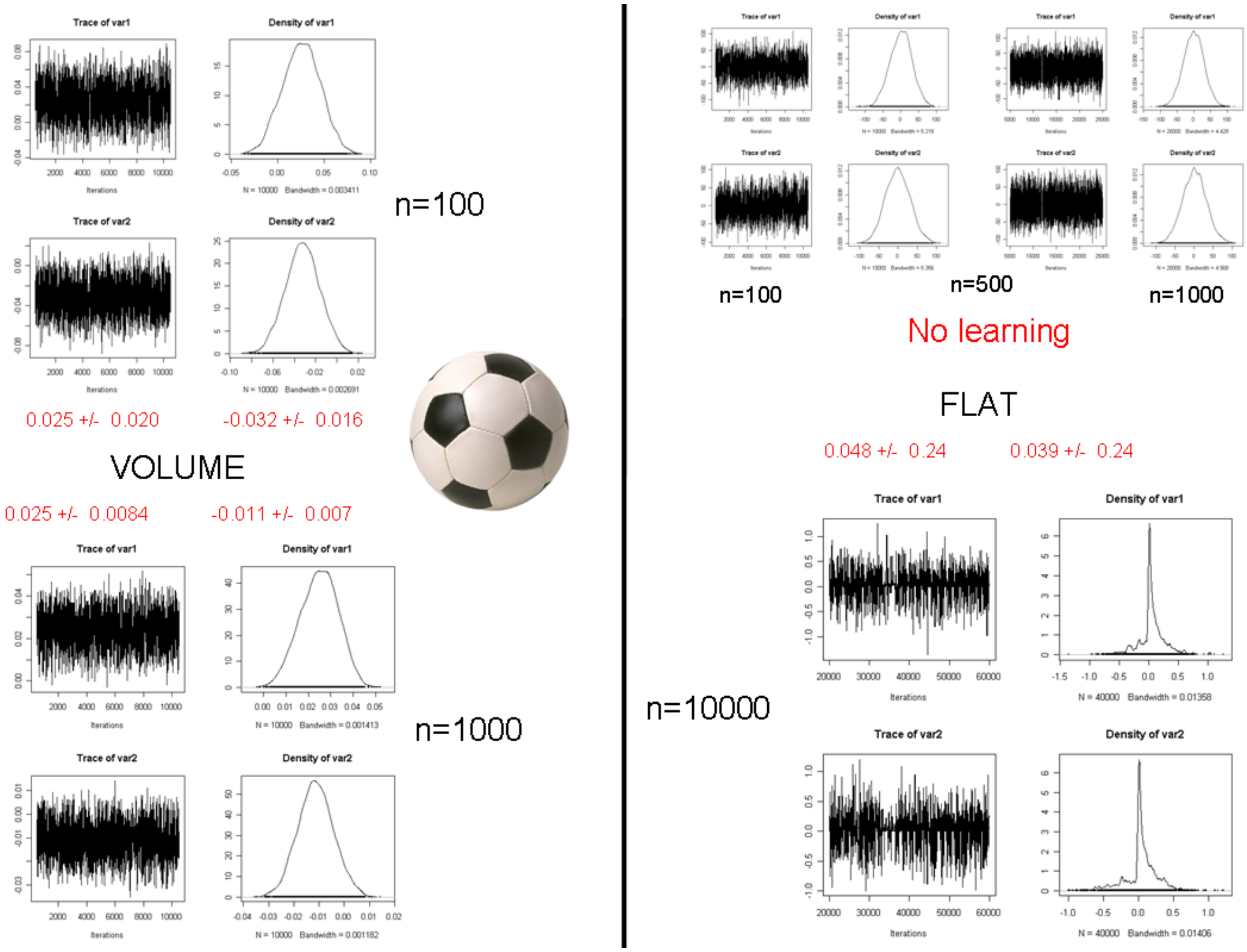

FIGURE 2. The right side shows the posteriors for the naive flat prior. The left side shows the posterior computed with the true uniform over the model.

Posterior Inference

With the help of the free MCMC package [1, 2] it only takes a few lines of code to realize the inadequacy of the naive prior for this example. The results of the MCMC simulations are summarized in figure 2. The true parameters where fixed at $a = 0.025$ and $b = -0.01$ and independent samples were chosen from the distribution with those parameters. With the naive flat prior the posteriors after observing $100, 500$ and 1000 samples were essentially identical to the priors $N(0, 100)$, i.e. nothing was learned from the data. With 10000 observations the program was able to learn the values $(0.048 \pm 0.24, 0.039 \pm 0.24)$ for the true parameters. In contrast, just after 100 observations the posterior with the true uniform prior estimates the parameters very precisely as $(0.025 \pm 0.020, -0.032 \pm 0.016)$, still one order of magnitude of extra accuracy over the posterior with the flat prior with two orders of magnitude of extra data!

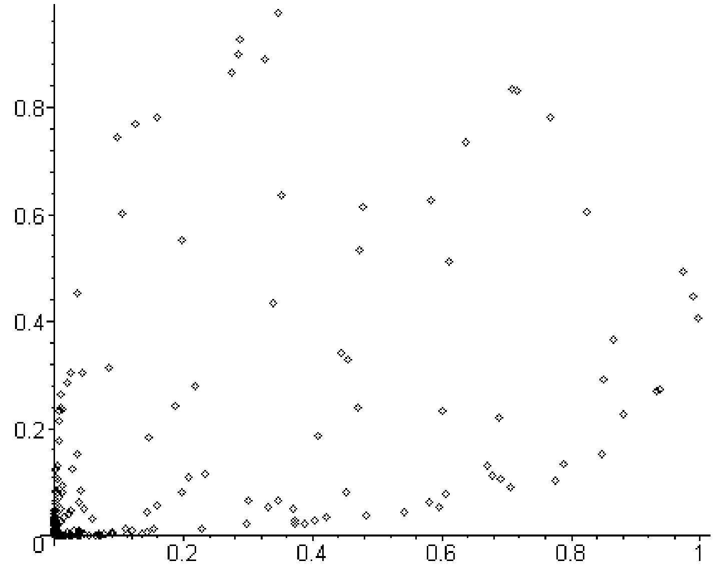

FIGURE 3. Ten thousand (u,v) points from (a,b) points chosen uniformly inside a ball of radius 3 centered at 0. Notice that more than 9000 points disappear into the origin.

Why is the volume prior so good?

To understand why the naive flat prior is so bad and the volume prior so good let's identify the transformed region of means (u,v) given by,

$$\begin{aligned} u &= \exp\left(-50(a^2+b^2)\right) \\ v &= \exp\left(-50((a-c)^2+(b-c)^2)\right) \end{aligned}$$

as (a,b) range over the entire plane. An easy way to find the shape of this region is to pick points (a,b) at random on the plane and plot the corresponding (u,v) points. Figure 3 shows $10000(u,v)$ points obtained from $10000(a,b)$ points uniformly distributed inside a circle centered at the origin of radius 3. Notice that lots of points disappear into the origin!. Now take another $10000(a,b)$ points but now distributed according to π_1 with density given in (1). Figure 4 shows these (a,b) points. Notice that they are all highly concentrated about two points close to the origin. The corresponding (u,v) points are shown in figure 5. Got it?

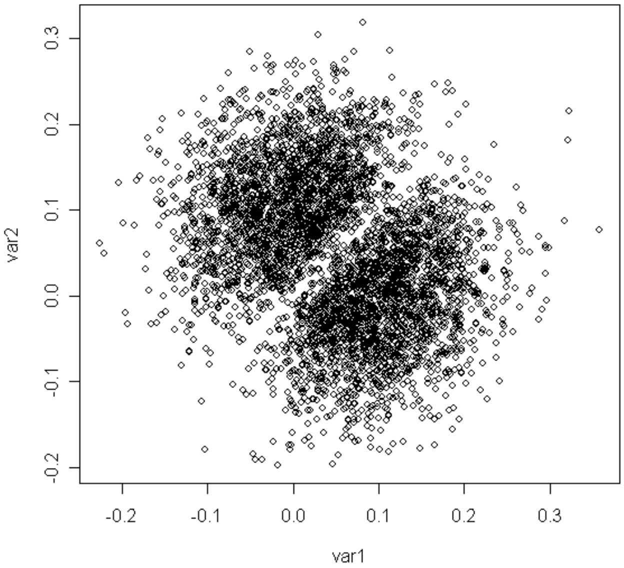

FIGURE 4. Ten thousand (a,b) points from the true uniform over the model

TABLE 1. Bioassay data from [3, p.88]

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.863	5	0
-0.296	5	1
-0.053	5	3
0.727	5	5

The equation of the boundary of the leaf of (u,v) points

The computation of the exact equation of the leaf boundary in figure 5 is a nice exercise in simple optimization: Find max and min of v subject to the constraint that $u=t$. The max is given by the Red (dark) $R(t)$ curve in figure 6, with

$$R(t)=\exp(-(\sqrt{-\log t}-1)^2). \tag{2}$$

The min is given by the Green (light) curve,

$$G(t)=\exp(-(\sqrt{-\log t}+1)^2) \tag{3}$$

with $0<t<1$ in both cases.

Notice that there is a non-removable corner singularity at $t=0$ but it is a piece of euclidean space so the curvature is zero at every point.

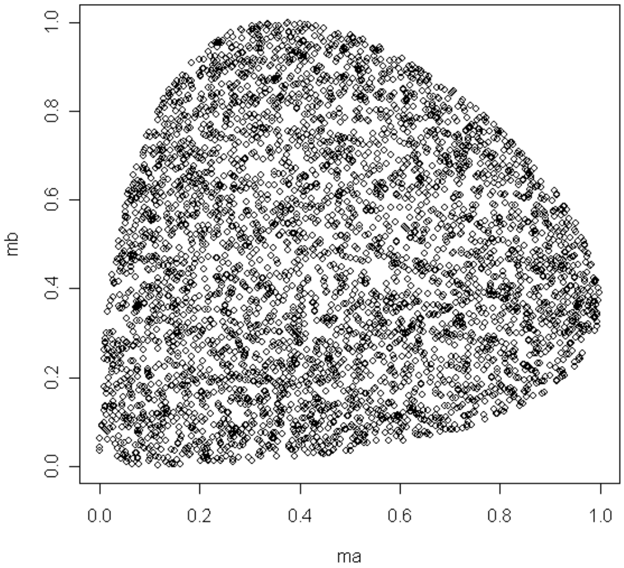

FIGURE 5. Ten thousand (u,v) points from (a,b) chosen from the density (1).

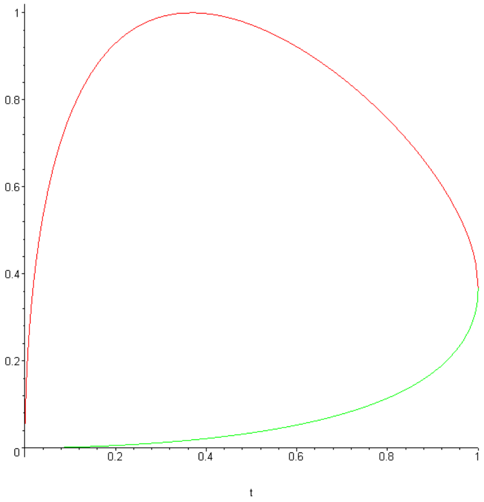

FIGURE 6. The exact boundary of the (u,v) points. Upper dark (red) part is $R(t)$. Lower light (green) part is $G(t)$

TEXTBOOK EXAMPLE

Perhaps the first non-trivial example of a multiparameter bayesian model is simple logistic regression (see [3, p.88]). Twenty animals were tested, five at each of four dose levels (see Table 1). The standard model for this kind of data is,

$$(x_i, n_i, y_i);\ i = 1, \ldots, k,$$

assumed independent with,

$$y_i | \theta_i \sim \mathrm{Bin}(n_i, \theta_i),$$

where θ_i is the probability of death for animals given dose x_i. The standard logistic dose-response relation is:

$$\log \frac{\theta_i}{1 - \theta_i} = a + bx_i \tag{4}$$

The joint distribution of $(y_1, \ldots, y_k)$ is a function of the unknown parameters (a, b) and straight (but tedious) calculations give the volume element $dV = \sqrt{\det g}\ dadb$ in the (a, b) parameterization as

$$\begin{aligned} dV &= T\ \sigma\ dadb \\ T &= \sum_j w_j = \sum_j n_j \theta_j (1 - \theta_j) \\ \sigma &= \text{ stand. dev. of } X \text{ defined as,} \\ & \quad P\left\{X = x_j \,\middle|\, \theta_j\right\} \propto w_j \\ \theta_j &= \frac{1}{1 + \exp(-a - bx_j)} \end{aligned}$$

This is a strange looking density (see figure 7). In particular this prior is proper and it assigns correlation of about 0.5 between a and b. This correlation is known a priori from the underlying geometry. In fact, the volume prior provides a better fit to the data than the standard diffuse naive prior that models a and b as independent variables with large variances. Figure 8 shows the results of the posterior simulations with both priors. Left panel with naive prior, right panel with volume prior. The red (dark) middle curves represent the logistic curves associated to the mean posterior values for (a, b) (100 thousand of them). The pictures also show 500 logistic curves obtained by sampling 500 (a, b) pairs from the available posterior samples. There is clearly more spread of logistic curves on the right than on the left panel. This is compatible with the fact that the volume prior samples uniformly over the manifold. Just like in the cooked-up example the over-spread (a, b) points cover only a small region of the manifold.

BEYOND FINITE VOLUMES

When the information volume $V(M)$ of the model M,

$$V(M) = \int_M dV = \int_\Theta \sqrt{\det g}\ d\theta$$

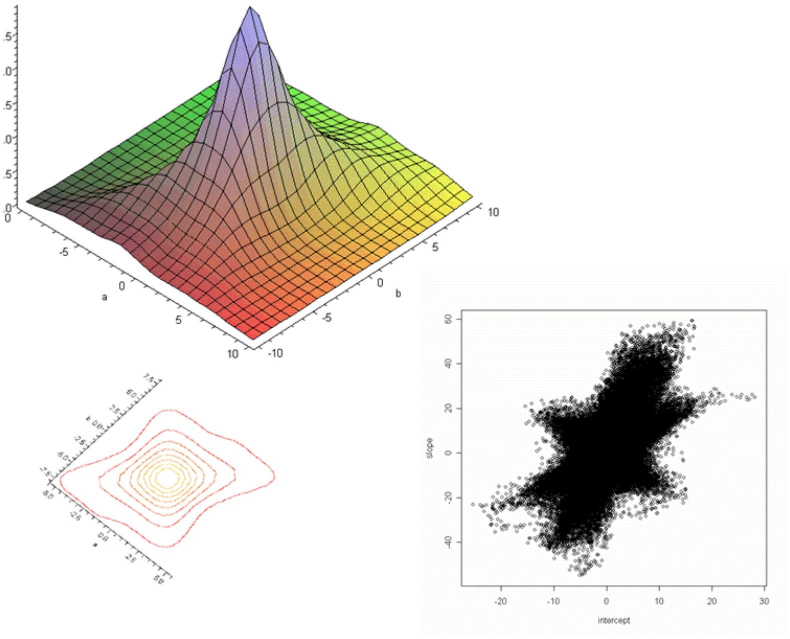

FIGURE 7. Uniform prior for logistic regression in the (a,b) parameterization. Bottom left: Contours. Bottom right: 2.5×10^5 samples from this prior.

is infinite; there is no uniform distribution over M. However, the underlying information geometry provides the following class of priors given as scalar density fields defined invariantly on M by,

$$\pi(p|t,\nu,\delta,\alpha) = \frac{1}{Z}[1+\alpha\nu I_\delta(p:t)]^{-\frac{1}{\nu}} \tag{5}$$

where $p \in M$, t is a probability distribution guessing the actual distribution of the data, δ, ν are scalar parameters in $[0,1]$, $\alpha > 0$ large enough so that $Z < \infty$ and $I_\delta(p:t)$ is the δ-information deviation between (unnormalized) distributions p and t given by,

$$I_\delta(p:t) = \frac{1}{\delta(1-\delta)} \int [\delta p + (1-\delta)t - p^\delta t^{1-\delta}]$$

where the integral is over the whole data space manifold. This family of priors exists for any regular model and it has many remarkable properties. In particular this family maximizes a simple and objective notion of ignorance. For details see my *A geometric theory of ignorance* . The hyper parameters can be estimated with priors of the same kind or with a nonparametric prior of the Dirichlet Process type (which could itself be

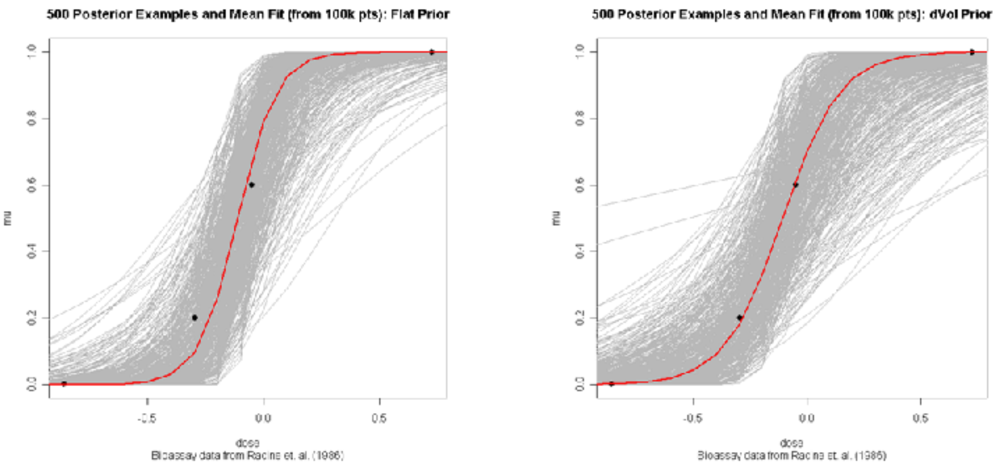

FIGURE 8. Posterior logistic curves. Left panel with naive prior. Right panel with volume prior.

seen as part of this family if we allow M to be infinite dimensional). There are still many open problems but the road ahead seems clear: More geometry.

ACKNOWLEDGMENTS

I am in debt to Phil Dawid whose invitation to talk at UCL prompted the finding of the example in the paper and to Ariel Caticha, Kevin Knuth, and John Skilling for many interesting discusions.

REFERENCES

1. A. D. Martin, , and K. M. Quinn, *MCMCpack: Markov chain Monte Carlo (MCMC) Package* (2007), URL `http://mcmcpack.wustl.edu`, r package version 0.8-2.
2. R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria (2007), URL `http://www.R-project.org`, ISBN 3-900051-07-0.
3. A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis, second edition*, London: CRC press, 2004.