Bayesian General Linear Models with T-Priors

W. Penny
Wellcome Trust Centre for Neuroimaging,
University College London,
12 Queen Square,
London WC1N 3BG.

April 30, 2013

Introduction

This report has three components. Firstly, we describe Bayesian inference for GLMs where we assume T-priors over regression coefficients. This naturally leads to T-posteriors. Previously in neuroimaging, Gaussian posteriors have been widely used. But in the limit of small samples, eg. typically $N = 12$ for population inference, the Gaussian approximation may not be sufficiently accurate. Secondly, we describe the use of loss functions, showing that the Bayes estimate of an effect is one that minimises the posterior expected loss. Thirdly, we compare Bayesian inferences based on Bayes factors to classical inferences based on p-values.

Theory

We consider the General Linear Model (GLM)

$$y = Xw + e \tag{1}$$

where y is an $N \times 1$ vector of data points, X is an $N \times K$ design matrix, w is a $K \times 1$ vector of regression coefficients and e is an $N \times 1$ vector of zero mean Gaussian errors with precision λ. This gives rise to the following likelihood

$$p(y|w, \lambda) = \mathsf{N}(y; Xw, \lambda I_N) \tag{2}$$

Priors

We assume a Normal-Gamma prior over regression coefficients and noise precision

$$\begin{aligned} p(w, \lambda) &= p(w|\lambda)p(\lambda) \\ p(\lambda) &= \mathsf{Ga}(\lambda; b_0, c_0) \\ p(w|\lambda) &= \mathsf{N}(w; w_0, B_0\lambda) \end{aligned} \tag{3}$$

The prior mean and variance over the precision is

$$\begin{aligned} \lambda_0 \equiv E[\lambda] &= b_0 c_0 \\ V[\lambda] &= b_0^2 c_0 \end{aligned} \tag{4}$$

This implicitly defines a multivariate non-central t-distribution over regression coefficients [1]

$$\begin{aligned} p(w) &= \int p(w|\lambda) p(\lambda) d\lambda \\ &= \mathsf{St}_k(w; w_0, B_0\lambda_0, 2c_0) \end{aligned} \tag{5}$$

The prior mean and variance over regression coefficients is therefore

$$\begin{aligned} E[w] &= w_0 \\ V[w] &= \frac{c_0}{\lambda_0(c_0 - 1)} B_0^{-1} \end{aligned} \tag{6}$$

Posteriors

For the precisions we have

$$\begin{aligned} p(\lambda|y) &= \mathsf{Ga}(\lambda; b_N, c_N) \\ \frac{1}{b_N} &= \frac{1}{b_0} + \frac{1}{2}(y - Xw_N)^T(y - Xw_N) \\ &\quad + \frac{1}{2}(w_N - w_0)^T B_0 (w_N - w_0) \\ c_N &= c_0 + \frac{N}{2} \\ \lambda_N &= b_N c_N \end{aligned} \tag{7}$$

For the regression coefficients we have

$$\begin{aligned} p(w|y) &= \mathsf{St}_k(w; w_N, B_N\lambda_N, 2c_N) \\ B_N &= B_0 + X^T X \\ w_N &= B_N^{-1}(B_0 w_0 + X^T y) \end{aligned} \tag{8}$$

The posterior mean and variance over regression coefficients is therefore

$$\begin{aligned} E[w] &= w_N \\ V[w] &= \frac{c_N}{\lambda_N(c_N - 1)} B_N^{-1} \end{aligned} \tag{9}$$

Contrasts

For a contrast

$$u = m^T w \tag{10}$$

with contrast vector m, we have

$$\begin{aligned} p(u) &= \mathsf{St}_1(u; \hat{u}, \hat{\lambda}, 2c_N) \\ &= \mathsf{St}(z; 2c_N) \end{aligned} \tag{11}$$

where

$$\begin{aligned} z &= \hat{\lambda}^{1/2}(u - \hat{u}) \\ \hat{u} &= m^T w_N \\ \hat{\lambda} &= \frac{\lambda_N}{m^T B_N^{-1} m} \end{aligned} \tag{12}$$

Evidence

The evidence is given by (see eg. Equation 3.34 in [**?**])

$$p(y) = \frac{\Gamma(c_N)\left(\frac{2}{b_0}\right)^{c_0}}{\Gamma(c_0)\pi^{N/2}} \left(\frac{|B_0|}{|B_N|}\right)^{1/2} \left(\frac{2}{b_N}\right)^{-c_N} \tag{13}$$

If the evidence for model m_i is $p(y|m_i)$ then the Bayes factor is defined as

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)} \tag{14}$$

Loss functions

If we define a Loss Function, $L(w, a)$, which is the 'cost' of estimating a parameter to be w when the true value is a, then the Posterior Expected Loss (PEL) is given by (see eg. page 113 in [2])

$$PEL(a) = \int L(w, a) p(w|Y) dw \tag{15}$$

A 'Bayes estimate' is then one that minimises this loss

$$w_B = \underset{a}{\text{argmin}} PEL(a) \tag{16}$$

For the quadratic loss function $L(a, w) = (w - a)^2$ the Bayes estimate is the posterior mean. For $L(a, w) = |w - a|$ its the posterior median.

For neuroimaging we might expect an asymmetric cost to be more appropriate. In this paper we use

$$L(w, a) = \begin{cases} -k_0(a - w) & \text{if } w \geq a \\ k_1(a - w) & \text{if } w < a \end{cases} \tag{17}$$

A Bayes estimate is then given by the $k_0/(k_0 + k_1)$ quantile. For example, if $k_0 = 1$ and $k_1 = 19$ then its 19 times worse to overestimate than underestimate the effect. The Bayes estimate is given by the 5th percentile.

Results

One sample t-test

A one-sample t-test can be implemented by setting $X = 1_N$. Notice we have degrees of freedom $DF = N + 2c_0$, whereas for classical inference we have DF=$N-1$. Figure 1 shows the posterior distribution for $p(w|y)$ for two different sized data sets $N = 4$ and $N = 12$ with $w = 1$ and $\lambda = 1$. We used $b_0 = 10, c_0 = 0.1, B = 1$.

The t-posteriors are compared to a Normal approximation to the posterior. This shows that, as is well known, the posterior is better approximated by a Normal as N increases.

We also computed the cost function given in equation 17 with $k_0 = 1$ and $k_1 = 19$ for different point estimates of the effect size. These were (a) the posterior mean, (b) the Bayes estimate (here, given by the 5th percentile) and (c) the null $w = 0$. Averaged over 1000 data sets these gave losses of (a) $L = 1.63$, (b) $L = 0.60$ and (c) $L = 1$[1].

Two sample t-test

This section compares Bayesian and classical inference for two-sample t-tests. This is implemented in a GLM by specifying $X_{2T} = I_2 \otimes 1_N$ where $\otimes$ denotes the Kronecker product and 1_N denotes a column vector of ones.

We generate 100 data sets using $N = 12$, $w = [2, 1]^T$ and $\lambda = 1$. For the Bayesian analysis we used $b_0 = 10, c_0 = 0.1$ and fitted two models to the data. Model 1 used a design matrix $X = 1_{2N}$ and prior precision $B = 1$. Model 2 used $X = X_{2T}$ with $B = I_2$. The evidence was computed for each model using equation 13. A Bayes Factor BF_{21} greater than 3 provides (weak) evidence in favour of an effect. For the classical inference we used the F-test described in [3].

Figure 2 plots log Bayes factor versus log p-value showing a good correlation between classical and Bayesian inference. Classical inferences with smaller p-values (ie. higher significance) correspond to larger Bayes factors. The figure also plots decision boundaries corresponding to p-values of 0.05 and Bayes factors of 3. Points in the lower left quadrant of Figure 2 are data sets where Bayesian and classical inference disagree. Agreement would be better with a cut-off of eg. $p < 0.01$

Figures 3 and 4 show how this correlation changes as a function of sample size and effect variability. To produce Figure 4 (a) we used $\lambda = 10$. This is a strong signal. Here all p-values were less than 0.05 and most Bayes Factors were greater than 3.

[1] This does not show that the Bayes estimate minimises the posterior expected loss. Here we are averaging over multiple realisations of data.

Appendix

The multivariate Normal density is given by

$$\mathsf{N}(\mu, \Lambda) = (2\pi)^{-d/2}|\Lambda|^{1/2} \exp\left(-\frac{1}{2}(x-\mu)^T\Lambda(x-\mu)\right) \tag{18}$$

The Gamma density is defined as

$$Ga(b, c) = \frac{1}{\Gamma(c)}\frac{x^{c-1}}{b^c}\exp\left(\frac{-x}{b}\right) \tag{19}$$

The multivariate non-central T-distribution, $\mathsf{St}_k(x; \mu, B, \alpha)$, has mean and variance

$$\begin{aligned} E[x] &= \mu \\ V[x] &= \frac{\alpha}{\alpha - 2}B^{-1} \end{aligned} \tag{20}$$

Of course the expression for the variance only holds if $\alpha \geq 2$.

For $k = 1$ we have a univariate non-central t-distribution. If

$$p(x) = \mathsf{St}_1(x; \mu, b, \alpha) \tag{21}$$

and

$$y = b^{1/2}(x - \mu) \tag{22}$$

then y has a t-distribution with degrees of freedom α. This standard, univariate, central form is what is usually referred to as a t-distribution in neuroimaging. We denote this as

$$p(y) = \mathsf{St}(y; \alpha) \tag{23}$$

Unit Information Prior

Rouder et al [4] recommend the use of a 'unit information prior' to define a Bayesian one-sample t-test. This first defines an 'effect-size', δ, as the ratio of the mean to the standard deviation. In the above notation, the mean will be the regression coefficent, w, (assuming the design matrix is a single column of ones), and the standard deviation of the data is given by the standard deviation of the 'observation noise' , $\lambda^{1/2}$. Thus

$$\delta = \lambda^{1/2}w \tag{24}$$

A unit information prior is then specified by setting the prior variance of δ to unity. We have

$$\begin{aligned} Var[\delta] &= \lambda^{1/2}V[w] \\ &= \frac{\lambda^{1/2}c_0}{\lambda_0(c_0 - 1)}B_0^{-1} \end{aligned} \tag{25}$$

First of all we set λ equal to the observed data precision from a GLM fit. If we then set $c_0 = 2$, $\lambda_0 = \lambda$, $b_0 = \lambda/2$, and $B_0 = \sqrt{2}/\sqrt{b_0}$ then $Var[\delta] = 1$. Thus, a Bayesian GLM with a T-prior can be reduced to the unit information prior. Note that the empirical results in this report do not use this approach. See Wetzels et al [5] for a comparison of these Bayesian t-tests to those from classical inference.

References

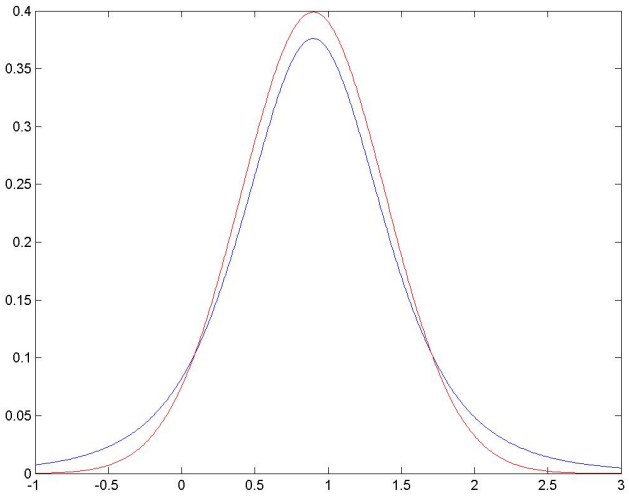
[1] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory.* Wiley, Chichester, 2000.

[2] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, Boca Raton, 1995.

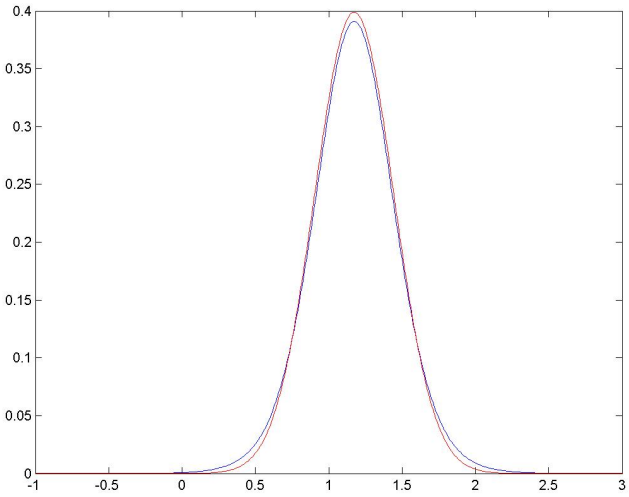
[3] D.G. Kleinbaum, L.L. Kupper, and K.E. Muller. *Applied Regression Analysis and Other Multivariable Methods.* PWS-Kent, Boston, 1988.

[4] J N. Rouder, P L. Speckman, D Sun, R. Morey, and G Iverson. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev*, 16(2):225–237, Apr 2009.

[5] R Wetzels, D. Metzke, M. Lee, J. Rouder, G. Iverson, and E. Wagenmakers. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298, 2011.



(a)



(b)

Figure 1. Posterior distribution over effect size w for t-posteriors (blue) and Normal posteriors (red) for (a) N=4 and (b) N=12 samples.

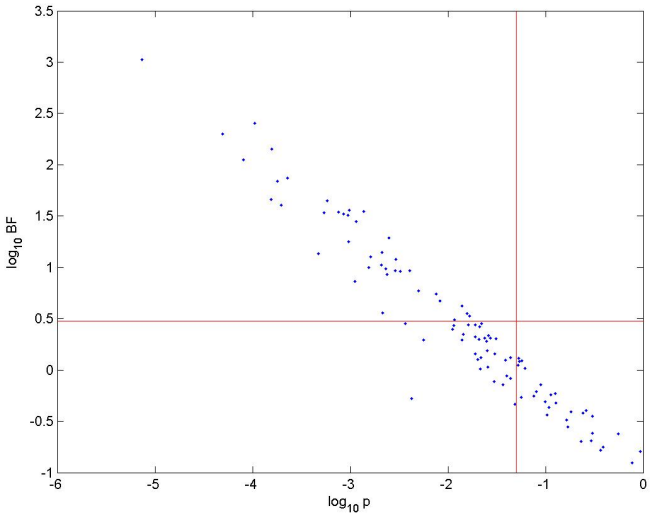
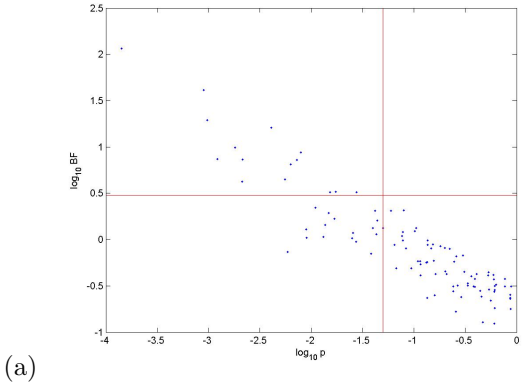
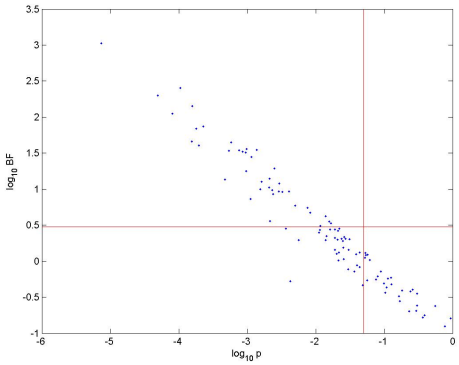


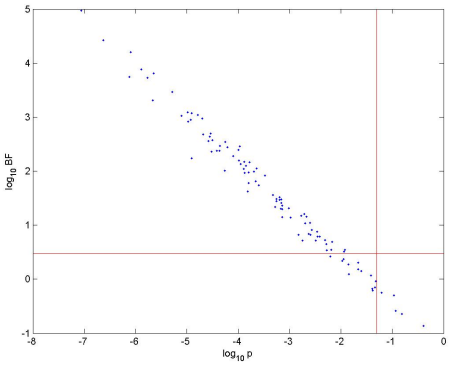
Figure 2. Plot of log Bayes factor versus log p-value for inference with a two-sample t-test with common noise precision $\lambda = 1$. Each dot is the value for a single data set. The horizontal lines correspond to a Bayes factor of 3 and the vertical lines to a p-value of 0.05.



(a)

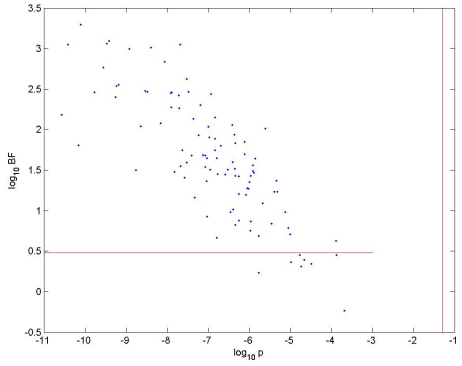


(b)

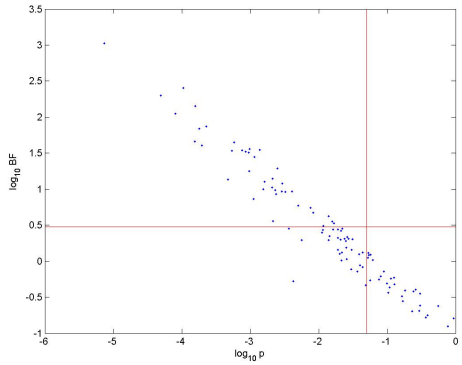


(c)

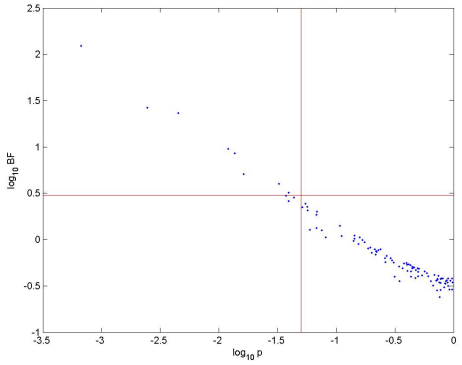
Figure 3. Plots of log Bayes factor versus log p-value for inference with a two-sample t-test with (a) $N = 6$, (b) $N = 12$ and (c) $N = 24$ subjects per group. Each dot is the value for a single data set. The horizontal lines correspond to a Bayes factor of 3 and the vertical lines to a p-value of 0.05.



(a)



(b)



(c)

Figure 4. Plots of log Bayes factor versus log p-value for inference with a two-sample t-test with common noise precision (a) $\lambda = 10$, (b) $\lambda = 1$ and (c) $\lambda = 0.1$. Each dot is the value for a single data set. The horizontal lines correspond to a Bayes factor of 3 and the vertical lines to a p-value of 0.05.