Bayesian Inference for the Multivariate Normal

Will Penny

Wellcome Trust Centre for Neuroimaging,
University College, London WC1N 3BG, UK.

November 28, 2014

Abstract

Bayesian inference for the multivariate Normal is most simply instantiated using a Normal-Wishart prior over the mean and covariance. Predictive densities then correspond to multivariate $\mathcal{T}$ distributions, and the moments from the marginal densities are provided analytically or via Monte-Carlo sampling. We show how this textbook approach is applied to a simple two-dimensional example.

1 Introduction

This report addresses the following question. Given samples of multidimensional vectors drawn from a multivariate Normal density with mean m and precision Λ, what are the likely values of m and Λ ? Here, the precision is simply the inverse of the covariance i.e. $\Lambda = C^{-1}$. Whilst it is simple to compute the sample mean and covariance, for inference we need the probability distributions over these quantities. In a Bayesian conception of this problem we place prior distributions over all quantities of interest and use Bayes rule to compute the posterior. We follow the formulation in Bernardo and Smith [1] (tabularised on page 441).

2 Preliminaries

2.1 Multivariate $\mathcal{T}$ distribution

The multivariate $\mathcal{T}$ distribution over a d-dimensional random variable x is

$$p(x) = \mathcal{T}(x; \mu, \Lambda, v) \tag{1}$$

with parameters μ, Λ and v. The mean and covariance are given by

$$\begin{aligned} E(x) &= \mu \\ Var(x) &= \frac{v}{v-2}\Lambda^{-1} \end{aligned} \tag{2}$$

The multivariate $\mathcal{T}$ approaches a multivariate Normal for large degrees of freedom, v, as shown in Figure 1. For $v = 1$, $\mathcal{T}$ is a multivariate Cauchy distribution.

For $d = 1$, we have a univariate t-distribution

$$p(x) = t(x; \mu, \lambda, v) \tag{3}$$

A side issue here relates to the marginal one-dimensional distributions produced e.g. by applying a contrast matrix to a multivariate vector x. Generally, such linear contrasts are not univariate t-distributions. However, the multivariate T's that are defined in [4] (in terms of correlation rather than precision matrices) do have this property. Nevertheless, this issue is not vitally important as quantities based on the marginals (moments, exceedances) can be computed using sampling, as we show in the Results section.

2.2 Wishart Distribution

The Wishart distribution, as defined in Bernardo and Smith (p. 435), over a $[d \times d]$ matrix Λ is

$$\begin{aligned} p(\Lambda) &= \mathcal{W}(\Lambda; a, B) \\ E(\Lambda) &= aB^{-1} \end{aligned} \tag{4}$$

where B is a symmetric, nonsingular matrix and $2a > d - 1$. For $d = 1, B = 1$ it reduces to a χ^2 distribution with a degrees of freedom. In Bayesian statistics the Wishart is the conjugate prior of the precision matrix.

For $d = 1$, the Wishart reduces to a Gamma distribution [2](p. 693)

$$\begin{aligned} p(\lambda) &= Ga(\lambda; a, b) \\ E(\lambda) &= \frac{a}{b} \\ Var(\lambda) &= \frac{a}{b^2} \end{aligned} \tag{5}$$

For $a = 1$ we have the exponential distribution (Bishop p. 688) However, the marginals $p(\lambda_i)$ of $p(\Lambda)$ are not Gamma densities [4]. Again, this is not terribly important as we can use sampling to compute quantities based on the marginals (moments, exceedances). Samples can be drawn from the Wishart density using the wishrnd.m function in the matlab statistics toolbox. They can also be generated directly, as described in Gelman [3] (page 481), by generating v Gaussian random vectors, x_n with zero mean and precision $2B$ and letting $\Lambda = \sum_{n=1}^{2a} x_n x_n^T$.

3 Multivariate Normal Model

3.1 Priors

Following Bernardo and Smith (p. 441) the prior is chosen to have the following form

$$\begin{aligned} p(w_n|m, \Lambda) &= \mathcal{N}(w_n; m, \Lambda) \\ p(m|\Lambda) &= \mathcal{N}(m; m_0, \beta_0 \Lambda) \\ p(\Lambda) &= \mathcal{W}(\Lambda; a_0, B_0) \end{aligned} \tag{6}$$

with mean m and precision matrix Λ. The samples are w_n with $n = 1..N$ and each is d-dimensional. We have also specified a Normal-Wishart prior over $\{m, \Lambda\}$ which is specified by the parameters m_0, β_0, a_0, B_0.

In the above formulation w_n varies around m with precision Λ, whereas m varies about m_0 with precision $\beta_0\Lambda$. Thus the value of β_0 specifies the prior precision of the mean relative to that of the parameters. The overall generative model is shown in Figure 2.

3.1.1 Marginal Prior Precision

As the precision is at the top of the hierarchy in the generative model, its marginal prior distribution is exactly as written above

$$p(\Lambda) = \mathcal{W}(\Lambda; a_0, B_0) \tag{7}$$

The mean prior precision matrix is the mean of a Wishart density

$$\begin{aligned} \bar{\Lambda} &= a_0 B_0^{-1} \\ \bar{C} &= \frac{1}{a_0} B_0 \end{aligned} \tag{8}$$

We have also written the equivalent mean prior covariance matrix of $\bar{C} = \bar{\Lambda}^{-1}$. The parameter matrix B_0 is set to reflect our prior beliefs. For example, if the data points are a priori believed to be independent, B_0 can be set to an appropriate diagonal matrix.

3.1.2 Marginal Prior Mean

The marginal distribution over m corresponds to a multivariate T-distribution (Bernardo and Smith, p435)

$$\begin{aligned} p(m) &= \mathcal{T}(m; \mu_{m0}, \Lambda_{m0}, v_{m0}) \\ \mu_{m0} &= m_0 \\ \Lambda_{m0} &= \beta_0 a_0 B_0^{-1} \\ v_{m0} &= 2a_0 - d + 1 \end{aligned} \tag{9}$$

where the covariance $C_0 = \Lambda_0^{-1}$. We have

$$\begin{aligned} Var(m) &= \left[\frac{v_{m0}}{v_{m0} - 2}\right] \Lambda_{m0}^{-1} \\ &= \left[\frac{1}{\beta_0(a_0 - 1)}\right] B_0 \end{aligned} \tag{10}$$

3.1.3 Prior Predictive Density

The prior predictive density, or sample density, is also given by a multivariate T-distribution

$$\begin{aligned} p(w) &= \mathcal{T}(w; \mu_{w0}, \Lambda_{w0}, v_{w0}) \\ \mu_{w0} &= m_0 \\ \Lambda_{w0} &= \frac{\beta_0[a_0 - 0.5(d-1)]}{\beta_0 + 1} B_0^{-1} \\ v_{w0} &= 2a_0 - d + 1 \end{aligned} \tag{11}$$

We have

$$\begin{aligned} Var(w) &= \left[\frac{v_{w0}}{v_{w0}-2}\right]\Lambda_{w0}^{-1} \\ &= \left[\frac{2(1+\beta_0)}{\beta_0(v_{w0}-2)}\right]B_0 \end{aligned} \tag{12}$$

If we sum up the sample density over the N observations we get the marginal likelihood, or model evidence.

3.2 Posteriors

Given w, the posterior distribution over m, Λ is a Normal-Wishart density

$$\begin{aligned} p(m|\Lambda, w) &= \mathcal{N}(m; m_N, \beta_N\Lambda) \\ p(\Lambda|w) &= \mathcal{W}(\Lambda; a_N, B_N) \end{aligned} \tag{13}$$

where

$$\begin{aligned} m_N &= (\beta_0 m_0 + N\bar{w})/\beta_N \\ \beta_N &= \beta_0 + N \\ a_N &= a_0 + \frac{N}{2} \\ B_N &= B_0 + \frac{N}{2}\left[\bar{\Sigma} + \frac{\beta_0}{\beta_N}(\bar{w}-m_0)(\bar{w}-m_0)^T\right] \end{aligned} \tag{14}$$

and

$$\begin{aligned} \bar{w} &= \frac{1}{N}\sum_{n=1}^{N} w_n \\ \bar{\Sigma} &= \frac{1}{N}\sum_{n=1}^{N}(w_n-\bar{w})(w_n-\bar{w})^T \end{aligned} \tag{15}$$

In the equation for B_N we see that β_0 determines the relative weight given to the term describing the covariation about the prior mean versus the term relating to the sample covariance, $\bar{\Sigma}$. In the equation for a_N we see that $2a_0$ can be interpreted as the number of pseudo data points that have been used to construct B_0. The least informative prior is given by $a_0 = d$.

3.2.1 Marginal Posterior Precision

The marginal posterior precision matrix is distributed as $p(\Lambda|w)$ as shown above. The mean posterior precision and covariance matrices are

$$\begin{aligned} \bar{\Lambda}_N &= a_N B_N^{-1} \\ \bar{C}_N &= \frac{1}{a_N}B_N \end{aligned} \tag{16}$$

3.2.2 Marginal Posterior Mean

The marginal posterior distribution over m corresponds to a multivariate $\mathcal{T}$(Bernardo and Smith, p441)

$$\begin{aligned} p(m|w) &= \mathcal{T}(m; \mu_{mN}, \Lambda_{mN}, v_{mN}) \\ \mu_{mN} &= m_N \\ \Lambda_{mN} &= \beta_N a_N B_N^{-1} \\ v_{mN} &= 2a_N - d + 1 \\ &= v_{m0} + N \end{aligned} \tag{17}$$

We have

$$\begin{aligned} Var(m|w) &= \left[\frac{v_{mN}}{v_{mN} - 2}\right] \Lambda_{mN}^{-1} \\ &= \left[\frac{1}{\beta_N(a_N - 1)}\right] B_N \end{aligned} \tag{18}$$

The mean posterior covariance used in drawing m is then $\bar{C}_N/\beta_N$. The square roots of the diagonals of this latter quantity are analagous to the Standard Error of the Mean (SEM). As $\beta_N = \beta_0 + N$, this shows that the uncertainty regarding m decreases in proportion to the number of data points. Notice that the posterior variance has the same mathematical form as the prior variance (equation 10), as we expect from the use of conjugate priors.

3.2.3 Posterior Predictive Density

The posterior predictive density for a new sample $\tilde{w}$ is a multivariate $\mathcal{T}$ distribution

$$\begin{aligned} p(\tilde{w}|w) &= \mathcal{T}(\tilde{w}; \mu_{wN}, \Lambda_{wN}, v_{wN}) \\ \mu_{wN} &= m_N \\ \Lambda_{wN} &= \frac{\beta_N[a_N - 0.5(d-1)]}{\beta_N + 1} B_N^{-1} \\ v_{wN} &= 2a_N - d + 1 \\ &= v_{w0} + N \end{aligned} \tag{19}$$

We have

$$\begin{aligned} Var(\tilde{w}|w) &= \left[\frac{v_{wN}}{v_{wN} - 2}\right] \Lambda_{wN}^{-1} \\ &= \left[\frac{2(1 + \beta_N)}{\beta_N(v_{wN} - 2)}\right] B_N \end{aligned} \tag{20}$$

Notice that the posterior variance has the same mathematical form as the prior variance (equation 12), as we expect from the use of conjugate priors.

4 Two-Dimensional Example

We generate samples w_n, $n = 1..32$, from a multivariate Normal density with mean $\mu = [10, 7]^T$ and covariance $C = [4, -0.7; -0.7, 0.25]$ (underlying correlation $r = -0.7$).

We then fitted the multivariate Normal model to the data using values for the priors of: $m_0 = [0, 0]^T$, $\beta_0 = 0.01$ and $a_0 = 2$. The sufficient statistics of the posterior distributions were computed as in equations 13 to 15. Figure 3 (left panel) shows (log of) the posterior predictive density computed using parameters from equation 19. For comparison, we plot the (log of) the Normal density computed using maximum likelihood values of the mean and covariance (using the usual formulae).

Marginal distributions over quantities of interest are readily computed using a sampling approach as follows. Figure 4 plots samples from the posterior distribution over $p(\sigma_1, \sigma_2|w)$. These were computed by drawing 1000 samples from the posterior precision $p(\Lambda|w)$ shown in Equation 13. For each sample, we inverted to get the covariance matrix and recorded the σ_1 and σ_2 values. For each sample, we also computed the correlation r_{12}. These samples were then used to construct the posterior correlation $p(r_{12}|w)$ shown in Figure 5.

References

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

[2] C.M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

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

[4] S. Kotz. *Multivariate T-Distributions and Their Applications.* Cambridge University Press, Cambridge University Press, 2004.

A Densities

The gamma density

$$Ga(x; a, b) = \frac{1}{\Gamma(a)} b^a x^{a-1} \exp(-bx) \tag{21}$$

is implemented in the function `p=spm_Gpdf(x,a,b)`.

The multivariate $\mathcal{T}$

$$\begin{aligned} \mathcal{T}(x; \mu, \Lambda, v) &= \frac{1}{Z} \left(1 + \frac{1}{v}(x-\mu)^T \Lambda (x-\mu)\right)^{-(v+d)/2} \\ Z &= \frac{\Gamma((v+d)/2)}{\Gamma(v/2)(v\pi)^{d/2}} |\Lambda|^{1/2} \end{aligned} \tag{22}$$

is implemented in the function `p=spm_mvtpdf(x,\mu,\Lambda,v)`.

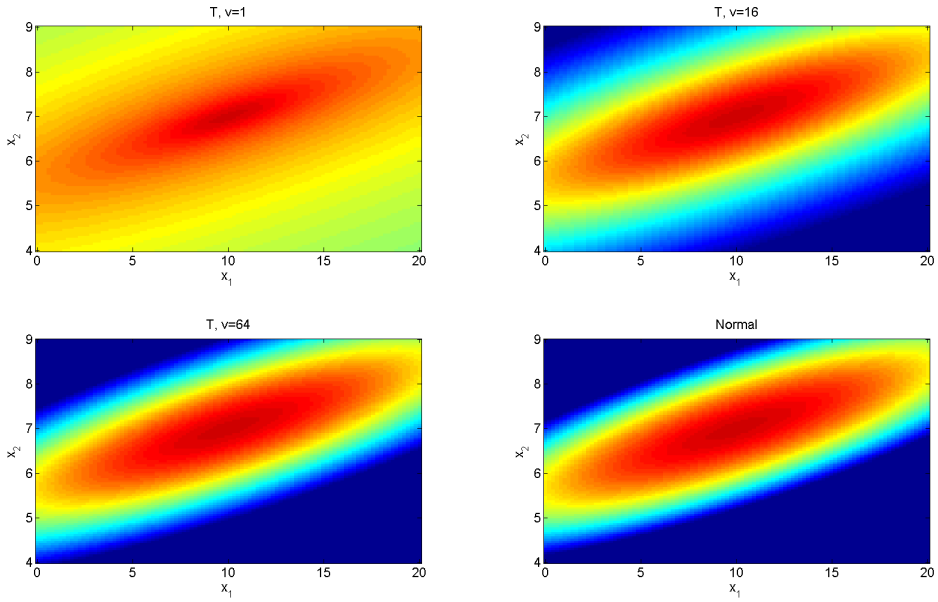

Figure 1: The multivariate $\mathcal{T}$ approaches a multivariate Normal for large degrees of freedom, v. The above plots show the log probability density of a bivariate $\mathcal{T}$ with mean $\mu = [10, 7]^T$ and covariance $C = [9, 1.2; 1.2, 0.25]$ (underlying correlation $r = 0.8$) for degrees of freedom $v = 1, 16, 64$. The bottom right panel plots the log probability of a multivariate Normal with the same mean and covariance. The scales are the same on all plots. For $v = 1$ (top left) $\mathcal{T}$ is a multivariate Cauchy distribution.

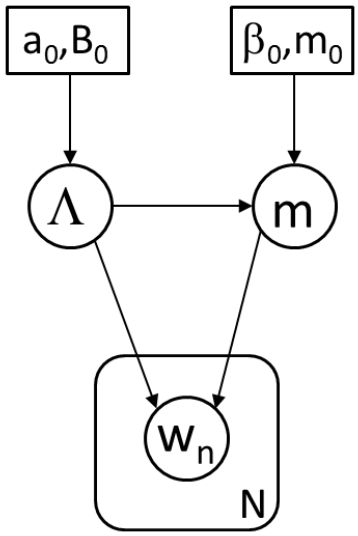

Figure 2: The graphical model shows that the overall joint density can be written $p(w, m, \Lambda) = p(\Lambda)p(m|\Lambda)\prod_{n=1}^{N} p(w_n|m, \Lambda)$. The quantity $2a_0$ can be interpreted as the number of pseudo data points that have been used to construct B_0. The least informative prior is given by $a_0 = d/2$. Here m_0 is the prior mean and β_0 determines the strength of the prior covariation about the prior mean.

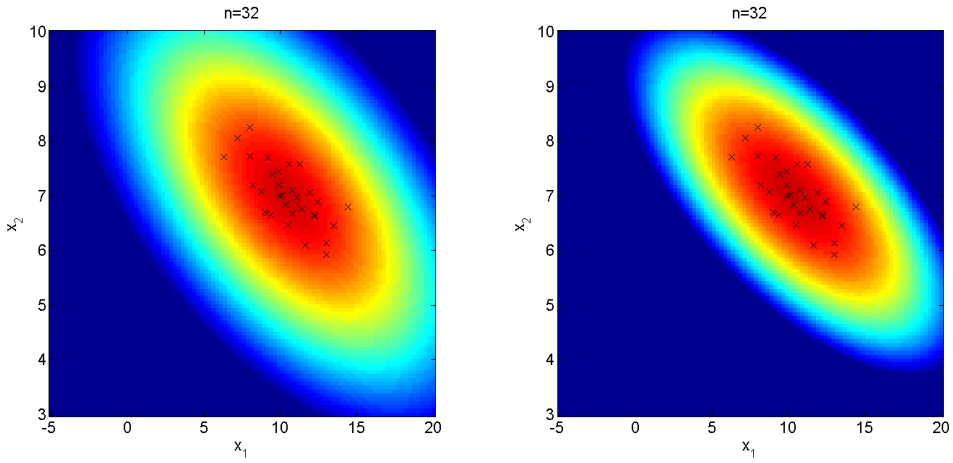

Figure 3: Data points ($N = 32$) and log predictive density of Normal model (left panel). This is a $\mathcal{T}$ distribution specified by equation 19. We also show a Normal distribution with maximum likelihood estimates of the mean and covariance (right panel).

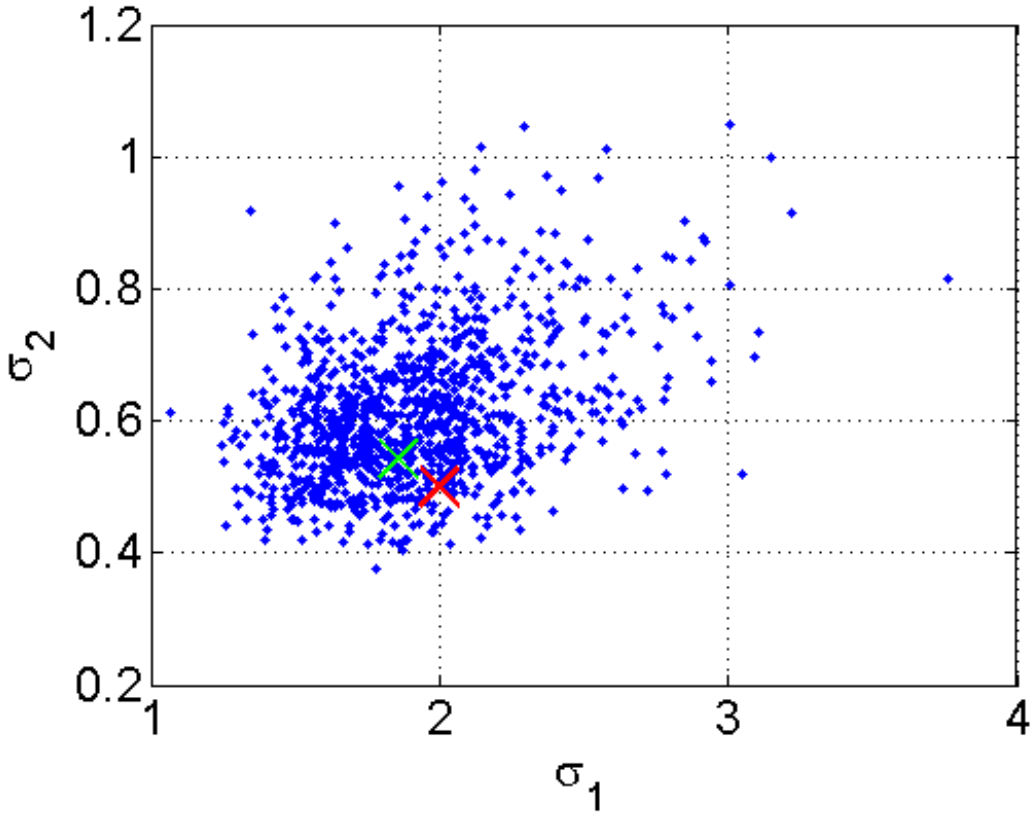

Figure 4: Samples from the posterior distribution $p(\sigma_1, \sigma_2|w)$ given $N = 32$ data points, true values (red cross), maximum likelihood estimates (green cross).

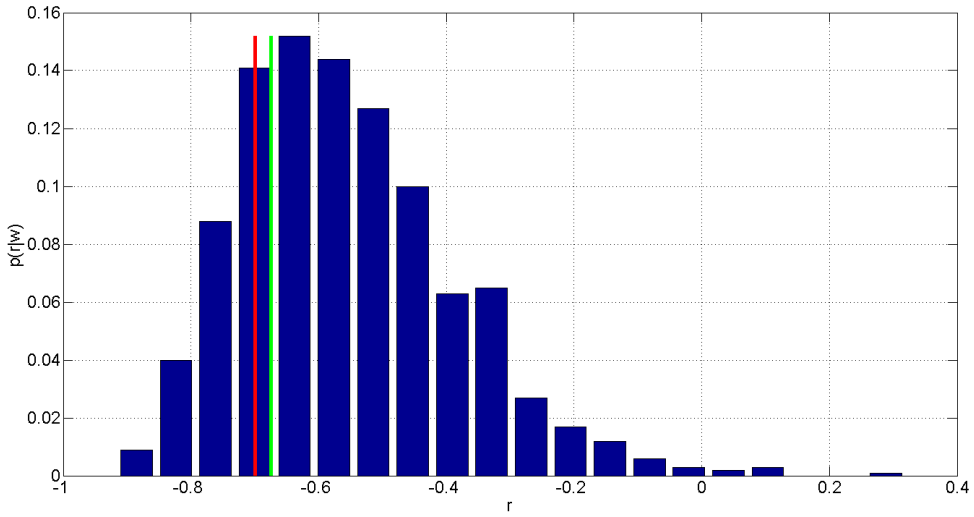

Figure 5: Posterior distribution over correlation $p(r_{12}|w)$ given $N = 32$ data points, true value (red line), maximum likelihood estimate (green line).