Bayesian data analysis in practice:
Three simple examples

Martin P. Tingley

Introduction

These notes cover three examples I presented at Climatea on 25 October 2011. Matlab code is available by request to demonstrate the ideas in Sections 1 and 2, and to fit the model in Section 3 and perform various manipulations on the draws from the posterior.

Most of what follows is based on Bayes' rule:

$$P(A|B) \propto P(B|A) \cdot P(A), \tag{1}$$

where the term to the left is called the *posterior*, and the terms to the right the likelihood and the *prior*, respectively.

Another helpful identity is,

$$P(A,B) = P(A|B) \cdot P(B). \tag{2}$$

In what follows, the short-hand notation $[A]$ denotes the distribution of the random variable A, while $[A|B]$ denotes the distribution of A conditional on B, $[A|\cdot]$ the distribution of A conditional on all other variables in the model, and $[A, B]$ the joint distribution of A and B.

For additional information, see Gelman et al. [2003], Banerjee et al. [2004], Tingley and Huybers [2010a,b]

1 Bivariate normal with known covariance

Say we have N independent and identically distributed (IID) draws from a bivariate normal distribution, and we seek inference on the mean vector. For the sake of simplicity, assume that the covariance matrix is known (put differently, all results in this section are conditional on the true value of the covariance matrix).

We'll use a normal prior, as it is conjugate (see below).

Data:

$$\mathbf{Y}_i|\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1 \dots N. \tag{3}$$

Prior:

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\eta}, \boldsymbol{\Gamma}) \tag{4}$$

Apply Bayes' rule to solve for the posterior distribution of $\boldsymbol{\mu}$, using the notation $\mathbf{Y}_{all}$ to denote the collection of N observations $\mathbf{Y}_i$:

$$P(\boldsymbol{\mu}|\mathbf{Y}_{all}) \propto P(\mathbf{Y}_{all}|\boldsymbol{\mu}) \cdot P(\boldsymbol{\mu}) \tag{5}$$

Now substitute in the likelihood and the prior, which are both normal, discarding all leading terms that do not contain $\boldsymbol{\mu}$:

$$P(\boldsymbol{\mu}|\mathbf{Y}_{all}) \propto \prod_{1}^{N} \exp\left\{-\frac{1}{2}(\mathbf{Y}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\mu})\right\} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\eta})^T \boldsymbol{\Gamma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\eta})\right\}. \tag{6}$$

Turn the product into a sum inside the exponential, use the notation $\bar{\mathbf{Y}}$ to denote the mean of the $\mathbf{Y}_i$, expand all terms, and discard any that do not contain $\boldsymbol{\mu}$:

$$P(\boldsymbol{\mu}|\mathbf{Y}_{all}) \propto \exp\left\{-\frac{1}{2}\left(-N\bar{\mathbf{Y}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - N\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\bar{\mathbf{Y}} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\eta} \;- \boldsymbol{\eta}\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}^T + \boldsymbol{\mu}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\mu}\right)\right\}. \tag{7}$$

The exponent is quadratic in $\boldsymbol{\mu}$, so immediately we recognize this as a normal distribution. All that remains is to complete the square to determine the mean and covariance. **If you've never done this calculation, you should!** Here's the answer:

$$\boldsymbol{\mu}|\mathbf{Y}_{all} \sim \mathcal{N}\left((N \cdot \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}(N \cdot \boldsymbol{\Sigma}^{-1}\bar{\mathbf{Y}} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}), (N \cdot \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Gamma}^{-1})^{-1}\right). \tag{8}$$

Recall that the prior for $\boldsymbol{\mu}$ was also normal. The property of the prior and posterior having the same distributional form is known as *conjugacy*, and the use of conjugate or conditionally conjugate priors generally results in simpler calculations and computations. Priors are often chosen to be conjugate or conditionally conjugate for these practical (rather than scientific) reasons.

The posterior mean is an inverse variance weighted average of the mean of the data and the prior mean. To facilitate interpretation, make the simplifying assumption that $\boldsymbol{\Gamma} = \frac{1}{c}\Sigma$, where c is some constant. In other words, the prior covariance matrix is proportional to the covariance of the observations. In this case, the posterior simplifies to,

$$\boldsymbol{\mu}|\mathbf{Y}_{all} \sim \mathcal{N}\left(\frac{N\bar{\mathbf{Y}} + c\boldsymbol{\eta}}{N + c}, \frac{1}{N + c}\Sigma\right). \tag{9}$$

The prior is thus equivalent to taking c additional observations, with a mean of $\boldsymbol{\eta}$. If c is small relative to N, the prior adds little, while if c is large relative to N, the prior can dominate the posterior.

For general $\boldsymbol{\Gamma}$, the prior can be interpreted as adding the same information as an additional observation, with value $\boldsymbol{\eta}$ and covariance matrix $\boldsymbol{\Gamma}$.

Matlab Demo: Relative influence of prior and likelihood for bivariate normal.

2 Gibbs sampler for the bivariate normal

The end goal of (most) Bayesian data analysis is to infer the posterior distribution of a set of parameters. In the case presented in Section 1, the joint posterior was bivariate normal – a well-known distributional form. In many cases (see Section 3 for an example), the joint distribution of the unknowns does not follow a standard distribution, and it becomes necessary to turn to various computational tools to draw samples from the posterior. The simplest such tool is the Gibbs sampler.

Say we have some nasty expression for the joint posterior of three random variables, Y_1, Y_2, and Y_3. The posterior does not follow a known distribution, so we can't just tell Matlab to produce samples for us. However, it may be the case that the distributions $[Y_1|Y_2,Y_3]$, $[Y_2|Y_1,Y_3]$, and $[Y_3|Y_1,Y_3]$ (the full conditional posteriors; often abbreviated as $[Y_1|\cdot]$ and so on) are all well-known forms. In this case, the Gibbs sampler provides us with a recipe for producing (correlated) samples from the joint posterior of $\mathbf{Y}$:

1. Produce/guess some initial value for each element of $\mathbf{Y}$, call them $\mathbf{Y}^0$.

2. For each sample $k = 1 \ldots N$, draw Y_1^k from $[Y_1^k|Y_2^{k-1},Y_3^{k-1}]$, draw Y_2^k from $[Y_2^k|Y_1^k,Y_3^{k-1}]$, and draw Y_3^k from $[Y_3^k|Y_1^k,Y_2^k]$.

In other words, draw from each conditional posterior in turn, using the most recent draws of the other variables.

The first time I built a Gibbs sampler I thought it was magic. Let's build one to sample from a bivariate normal. (Matlab can, of course, draw directly from the joint distribution in this case). Assume that,

$$\mathbf{Y} \sim \mathcal{N}\left(\left(\begin{array}{c}\mu_1\\ \mu_2\end{array}\right),\left(\begin{array}{cc}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{array}\right)\right) \tag{10}$$

The distribution of Y_1 conditional on a particular value y_2 of the random variable Y_2 is univariate normal, and likewise for $Y_2|y_1$:

$$Y_2|y_1 \sim \mathcal{N}\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1-\mu_1), \sigma_2^2(1-\rho^2)\right) \tag{11}$$

$$Y_1|y_2 \sim \mathcal{N}\left(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2), \sigma_1^2(1-\rho^2)\right). \tag{12}$$

The Gibbs sampler then proceeds by first initializing with some value y_1^0 (or y_2^0), and then iteratively drawing from the conditional distributions, always conditioning on the most recent draw of the other element of Y.

If some/all of the conditional posteriors do not follow well-known distributions, then the Metropolis-Hastings algorithm, or a variety of other tools, can be used to produce samples. See references for more details.

Matlab Demo: Gibbs sampler for the bivariate normal.

3 A simple hierarchical spatial model

The goal in this section is to infer a complete spatial field from noisy, incomplete, point referenced data. The Matlab example will use data from a number of weather stations to infer monthly average temperatures on a fine grid for the state of Colorado.

Let the vector $\mathbf{Y}$ refer to the spatial field at a set of locations $\mathbf{s}_i, i = 1 \ldots N$, which includes both the set of points where there are observations and the set of points where inference will be made.

At the **process level**, assume that $\mathbf{Y}$ is Gaussian, with constant mean and exponential covariance:

$$\mathbf{Y}|\mu, \phi, \sigma^2 \sim \mathcal{N}(\mu \mathbf{1}_N, \Sigma) \tag{13}$$

$$\Sigma_{ij} = \sigma^2 \exp\left(-\phi ||\mathbf{s}_i - \mathbf{s}_j||\right), \tag{14}$$

where $||\mathbf{s}_i - \mathbf{s}_j||$ is the distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$, and $\mathbf{1}_N$ is a vector of ones. Note that μ can be thought of as a regression coefficient for the covariate $\mathbf{1}$. Conceptually, we could replace $\mu\mathbf{1}$ with $\mu\mathbf{1} + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \ldots$, where the $\mathbf{X}_i$ are covariates such as elevation, latitude, or whatever else may be relevant. The machinery developed below can easily be generalized to this situation.

For the purposes of this example, we'll assume ϕ is known and fixed. Inference on ϕ is certainly possible, and the references provide a number of examples. However, inference requires use of the Metropolis-Hastings algorithm, which has not been covered in these notes. Think of the Metropolis-Hastings algorithm as a tool to learn and use when one or more of the full conditional posteriors does not follow a known form.

At the **data level**, assume that the observations $Z_j, j = 1 \ldots M$ can be expressed as the corresponding value of the field, plus IID white noise:

$$Z_i | Y_i, \tau^2 \sim \mathcal{N}(Y_i, \tau^2). \tag{15}$$

By defining an M by N selection matrix H of zeros and ones, we can write this in vector form:

$$\mathbf{Z}|\mathbf{Y}, \tau^2 \sim \mathcal{N}(\mathsf{H}\mathbf{Y}, \tau^2 \mathsf{I}_M), \tag{16}$$

where I_M is the M by M identity matrix.

At the **prior level**, assume conditionally conjugate priors for μ, σ^2 and τ^2:

$$\mu \sim \mathcal{N}(\eta, \delta^2) \tag{17}$$

$$\sigma^2 \sim IG(a_\sigma, b_\sigma) \tag{18}$$

$$\tau^2 \sim IG(a_\tau, b_\tau). \tag{19}$$

IG refers to the Inverse Gamma distribution, which is the conjugate prior for the variance parameters; see the Appendix for more information.

The joint posterior then follows from Bayes' rule:

$$[\mathbf{Y}, \mu, \sigma^2, \tau^2 | \mathbf{Z}] \propto [\mathbf{Z} | \mathbf{Y}, \mu, \sigma^2, \tau^2][\mathbf{Y} | \mu, \sigma^2, \tau^2][\mu][\sigma^2][\tau^2]. \tag{20}$$

In order to sample from the joint posterior, we make use of a Gibbs sampler. To derive the conditional posteriors, plug in the distributional forms, and then, for each unknown, consider the resulting expression as a function of only that parameter. See Section 2 for how to do so in the case of a Normal mean. Denote the correlation matrix by R, i.e. $\mathsf{R} = \sigma^{-2}\mathsf{\Sigma}$. The resulting conditional posteriors are,

$$\mathbf{Y}|\cdot \sim \mathcal{N}\left(\left(\mathsf{\Sigma}^{-1} + \tau^{-2}\mathsf{H}^T\mathsf{I}_M\mathsf{H}\right)^{-1} \cdot \left(\mu\mathsf{\Sigma}^{-1}\mathbf{1}_N + \tau^{-2}\mathsf{H}'\mathbf{Z}\right), \left(\mathsf{\Sigma}^{-1} + \tau^{-2}\mathsf{H}^T\mathsf{I}_M\mathsf{H}\right)^{-1}\right) \tag{21}$$

$$\mu|\cdot \sim \mathcal{N}\left(\left(\mathbf{1}_N^T\mathsf{\Sigma}^{-1}\mathbf{1}_N + \delta^{-2}\right)^{-1} \cdot \left(\mathbf{1}_N^T\mathsf{\Sigma}^{-1}\mathbf{Y} + \eta/\delta^2\right), \left(\mathbf{1}_N^T\mathsf{\Sigma}^{-1}\mathbf{1}_N + \delta^{-2}\right)^{-1}\right) \tag{22}$$

$$\sigma^2|\cdot \sim IG\left(N/2 + a_1, \frac{1}{2}\left(\mathbf{Y} - \mu\mathbf{1}_N\right)^T \mathsf{R}^{-1}\left(\mathbf{Y} - \mu\mathbf{1}_N\right) + b_1\right) \tag{23}$$

$$\tau^2|\cdot \sim IG\left(M/2 + a_2, \frac{1}{2}\left(\mathbf{Z} - \mathsf{H}\mathbf{Y}\right)^T\left(\mathbf{Z} - \mathsf{H}\mathbf{Y}\right) + b_2\right). \tag{24}$$

Suggestion: go through these calculations, and make sure I haven't made a mistake!

To sample from the joint posterior, we start with some initial guess of the parameters μ, σ^2 and τ^2, and then successively draw from each of the conditional posteriors. After discarding an initial series of draws to allow the chain to 'burn in', we can then use the resulting ensemble of draws to perform just about any analysis we can think of.

Matlab Demo: Implement this model for monthly Colorado temperature data. Note that the priors used in the demo for the parameters are strongly informative due to the paucity of data. It is worthwhile to change the priors (and also the value of ϕ) and see how the results are affected.

This analysis can clearly be improved in a number of ways. We've left out some pretty important geographic co-variates, and have thrown away the temporal dimension of a space-time data set.

Assignment for the ambitious: Add in the relevant co-variates, include a temporal model with a seasonal cycle, and perform inference on ϕ.

Appendix: The Inverse-Gamma distribution

The pdf of the Inverse Gamma is of the form,

$$P(\sigma^2|a_1, b_1) \propto \left(\sigma^2\right)^{-a_1-1}\exp\left(-b_1/\sigma^2\right). \tag{25}$$

Note that this functional form is equivalent to treating the Normal pdf as a function of the variance parameter.

Consider estimating the variance from M independent draws from a $\mathcal{N}(\mu, \sigma^2)$ distribution, where we assume that μ is known, and specify an $IG(a, b)$ prior for σ^2. Think of this as one step in a Gibbs sampler. Applying Bayes' rule, we have:

$$P(\sigma^2|,\mu,y_i,\ldots,y_M) \propto \prod_{i=1}^{M} P(y_i|\mu,\sigma^2)\cdot P(\sigma^2|a_1,b_1). \tag{26}$$

Substituting in the distributional forms, we have,

$$P(\sigma^2|,\mu,y_i,\ldots,y_M) \propto (\sigma^2)^{-M/2}\exp\left(-\frac{1}{\sigma^2}\frac{\sum_{i=1}^{M}(y_i-\mu)^2}{2}\right)\cdot\left(\sigma^2\right)^{-a_1-1}\exp\left(-b_1/\sigma^2\right). \tag{27}$$

Collecting terms yields,

$$P(\sigma^2|,\mu,y_i,\ldots,y_m) \propto (\sigma^2)^{-M/2-a_1-1}\exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{M}(y_i-\mu)^2}{2}+b_1\right)\right). \tag{28}$$

The posterior is thus Inverse Gamma:

$$\sigma^2|,\mu,y_i,\ldots,y_M \sim IG(M/2+a_1,\frac{1}{2}\sum_{i=1}^{M}(y_i-\mu)^2+b_1). \tag{29}$$

The Inverse Gamma prior for σ^2 can be interpreted as $2a_1$ prior observations with an average squared deviation of b_1/a_1.

References

S. Banerjee, B.P. Carlin, and A.E. Gelfand. *Hierarchical modeling and analysis for spatial data.* Chapman & Hall, 2004.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis.* Chapman & Hall/CRC, Boca Raton, 2 edition, 2003.

M.P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 1: Development and applications to paleoclimate reconstruction problems. *Journal of Climate*, 23(10):2759–2781, 2010a.

M.P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part 2: Comparison with the Regularized Expectation-Maximization Algorithm. *Journal of Climate*, 23(10), 2010b.