of fit. We might choose a different distribution for the likelihood. We might alter our deterministic model by adjusting the number of parameters or changing its functional form. We would not proceed to make inferences until we can be confident that the distribution of data arising from our model does not differ from the distribution of the original data more than we would expect from chance alone.

8.2 Marginal posterior distributions

Many researchers use software that implements the MCMC algorithm (e.g., JAGS, Plummer, 2003 or WinBUGS, Lunn et al., 2000). These packages allow users to easily produce crisp density plots of posterior distributions of parameters and lovely tables of statistics summarizing those distributions – means, medians, standard deviations, quantiles, and so on. It is a good idea to know where these plots and tables come from. Moreover, if you program your own MCMC algorithm, then it will be useful to know how to summarize the converged chains you have worked so hard to produce.

Assume we have MCMC output from a model that has been checked (as in Section 8.1). We want to use it to make inferences. An especially useful property of MCMC is marginalization. Recall that the purpose of MCMC is to help us gain insight about the joint distribution of the unobserved quantities (parameters and latent states) conditional on the observed ones (the data). We usually want to make statements about these quantities one at a time rather than jointly. We discussed in Section 3.4 that you can learn about a single random variable that is part of a joint distribution by marginalizing over all of the random variables except the one we seek to understand.

The marginalization property of MCMC allows us to do this for each parameter and latent state in MCMC output. This can be seen in the illustrative MCMC output in Box 8.1. The table's rows hold values for different parameters and columns hold their values at each iteration in the MCMC chain. The marginalization property of MCMC allows us to examine the posterior distributions of each parameter and latent state by treating its "row" as a vector without paying any attention to the other rows. The frequency of values in that vector defines the posterior distribution of the parameter (Box 8.1). The posterior distribution of a single parameter or latent state can be shown in scientific papers and proposals by simply plotting a normalized histogram of its values from the converged chain. Alternatively, you can use a kernel density estimator to fit a smooth curve.[8] This

[8]The advantage of a smooth curve is that you can overlay the posterior on the prior, which is good practice. This overlay can be hard to see if you use a histogram for the MCMC output, particularly when the prior is diffuse.

is what is going on under the hood when you output smooth curves of the posterior distribution when using popular MCMC software.

How do we summarize the marginal posterior distribution statistically? The marginalization property means that we can calculate summary statistics directly from the chain for each parameter (i.e., from its "row" in Box 8.1) to estimate the moments of the distribution. So, for example we approximate the mean of θ, that is, its expected value $\mathrm{E}(\theta|\boldsymbol{y})$ using

$$\mathrm{E}(\theta|\boldsymbol{y}) \approx \frac{1}{K}\sum_{k=1}^{K}\theta^{(k)}, \tag{8.2.1}$$

where K is the total number of iterations in the converged chain and $\theta^{(k)}$ is the sample of θ at the k^{th} iteration. This is called Monte Carlo integration (Box 8.1). Recall from Section 3.4.1 that equation 8.2.1 is an approximation of the integral $\mathrm{E}(\theta|\mathbf{y}) = \int \theta\,[\theta|\mathbf{y}]\,d\theta$. The posterior variance is similarly approximated by

$$\mathrm{var}(\theta|\mathbf{y}) \approx \frac{\sum_{k=1}^{K}\left(\theta^{(k)} - \mathrm{E}(\theta|\mathbf{y})\right)^2}{K}. \tag{8.2.2}$$

Again, we see this is an approximation of the second central moment,

$$\mathrm{E}\left((\theta - \mathrm{E}(\theta|\mathbf{y}))^2\right) = \int \left(\theta - \mathrm{E}(\theta|\mathbf{y})^2\right)[\theta|\mathbf{y}]\,d\theta$$

(Section 3.4.1). We can also calculate functions of the mean and variance, e.g., the standard deviation of θ conditional on the data is

$$\mathrm{sd}(\theta|\mathbf{y}) \approx \sqrt{\mathrm{var}(\theta|\mathbf{y})}. \tag{8.2.3}$$

and the coefficient of variation,

$$\mathrm{cv}(\theta|\mathbf{y}) \approx \frac{\sqrt{\mathrm{var}(\theta|\mathbf{y})}}{\mathrm{E}(\theta|\mathbf{y})}. \tag{8.2.4}$$

Don't allow these formulas to make life complicated. What this means in practice is that you calculate the mean, variance, standard deviation, or coefficient of variation over all of the θ in the MCMC output (i.e., the row for any parameter in Box 8.1).

Functions for kernel density estimators are widely available in statistical software – for example see the `density( )` function in R.

How do we show uncertainty associated with the point estimate of a parameter? We can define a $1-\alpha$ Bayesian credible interval on θ as the interval between L and U such that

$$\Pr\left(L < \theta < U\right) = \int_L^U [\theta|\mathbf{y}]\, d\theta = 1 - \alpha\,, \tag{8.2.5}$$

which is to say that the probability that the true value of the random variable θ falls between L and U is $1-\alpha$. Credible intervals can be calculated in different ways. We cover the two most widely used ones here.

An equal tailed interval gives the points L and U such that $\Pr\left(\theta < L\right) = \frac{\alpha}{2}$ and $\Pr\left(\theta > U\right) = \frac{\alpha}{2}$ (Figure 8.2.1). This means that

$$\int_{-\infty}^{L} [\theta|\mathbf{y}]\, d\theta = \int_{U}^{+\infty} [\theta|\mathbf{y}]\, d\theta = \frac{\alpha}{2}. \tag{8.2.6}$$

Given a mathematical expression for $[\theta|\mathbf{y}]$, we find L and U analytically from its cumulative distribution function (Section 3.4.1.4). We approximate U and L empirically from the quantiles of the converged MCMC samples: L is the $\frac{\alpha}{2}$ quantile and U is the $1-\frac{\alpha}{2}$ quantile. They are not necessarily symmetric, which obviates the use of a conventional interval expression like $\mu \pm \frac{w}{2}$, where w is the width of the interval.

Equal tailed intervals are easily understood and widely used in the ecological literature. Some Bayesians prefer a different approach that does not require equal areas in tails. A highest posterior density interval (HPDI) for the parameter θ is defined as a subset H of all of the possible values of θ such that

$$H = \{\theta : \Pr\left(\theta|\mathbf{y}\right) \geq c\} \tag{8.2.7}$$

where c is the largest number such that

$$\int_{\theta:[\theta|\mathbf{y}]\geq c} [\theta|\mathbf{y}] = 1 - \alpha. \tag{8.2.8}$$

This integral can be understood this way. The quantity c represents a horizontal line intersecting the posterior distribution (Figure 8.2.1). The points where the line intersects with the distribution curve are the lower and upper limits of the HDPI, L and U. We want to find the *longest* line

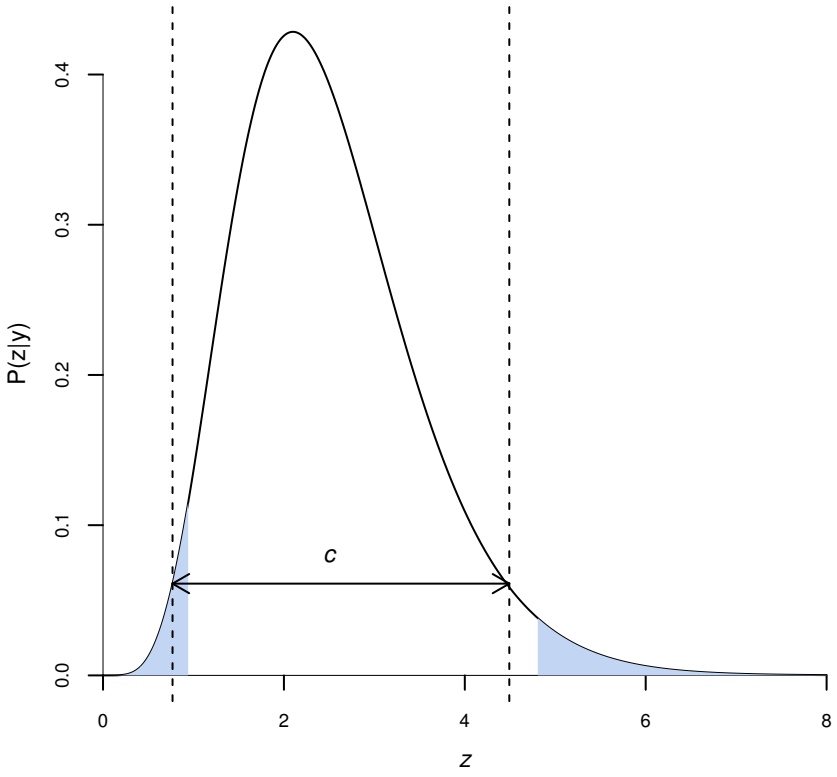

Figure 8.2.1: Illustration of equal-tailed intervals and highest posterior density intervals (HPDI) for $\alpha = .05$. The distribution is gamma with mean = 2.5 and variance = 1. The equal-tailed intervals give the upper (.975) and lower (.025) quantiles of the distribution. The shaded areas define the values of z outside of the equal-tailed interval. The HPDI is defined by the c arrow. This is the longest horizontal line that can be placed within the distribution such that the area between the vertical dashed lines and beneath the distribution curve = $1 - \alpha$. Note that the HPDI (.74, 4.48) is substantially narrower than the equal tailed interval (.94,4.81) because the distribution is skewed. For symmetric distributions, they would be the same.

between L and U for which the area under the curve and between intersections of the line with the curve $= 1-\alpha$. Starting at the peak of the distribution we move the line downward until area defined by its intersection with the curve $= 1 - \alpha$ (Figure 8.2.1). Although this sounds like a formidable task, there are functions available[9] to estimate HPDI's from vectors extracted from MCMC chains, at least for univariate distributions.

There are three situations when HPDIs are preferable to equal tailed intervals. Highest posterior density intervals are particularly useful for posterior distributions that are asymmetric. When this

[9]For example, `HPDinterval()` in the `coda` package (Plummer et al., 2010) or the `HPDregion()` in the `emdbook` package Bolker (2013) in R (R Core Team, 2013).

is the case, an HPDI will be narrower than an equal tailed interval (Figure 8.2.1). HPDI's can also include values in the interval that would be excluded by equal tailed intervals. For example, it may make sense for an interval to contain 0 for a random variable for distributions with non-negative support or support on $(0, 1)$. In these cases Bayesian equal-tailed credible intervals will never include 0 but HPDI's can. Finally, HPDI's are called for when the posterior distribution is multi-modal, that is, has more than one peak, although in this case the HPDI can be challenging to calculate.

8.3 Derived quantities

A second, exceedingly useful property of MCMC is *equivariance*. The idea of equivariance is that any quantity that is calculated from a random variable becomes a random variable with its own probability distribution. We have learned how Bayesian analysis supports inference on parameters and latent states. However, it is often the case that we seek inference on quantities *derived* from parameters and latent states. For example, we might estimate survivals and fertilities in a stage-structured, matrix model and seek inference on the dominant eigenvalue of the projection matrix. We might estimate the proportional contribution of species in plots and seek inference using Shannon's diversity index. We might estimate the mean body mass of a species and want to make inference on its metabolic rate using a scaling equation. The equivariance property of MCMC allows us to easily obtain posterior distributions of these derived quantities, a result that is difficult if not impossible using classical statistics.

Continuing our example from above (equations 8.1.3, 8.1.4), assume we calculate a derived quantity, $\varphi^{(k)}$ at each of the K iterations of an MCMC algorithm using a deterministic function, $\varphi^{(k)} = h\left(\boldsymbol{\theta}^{(k)}, z^{(k)}\right)$. We can summarize the distribution of the $\varphi^{(k)}$ from MCMC output in the same way we summarize the distributions of each θ (Section 8.2).

- A particularly useful application of derived quantities is to examine the size of effects of treatments, locations, groups, and so on, an application analogous to the use of contrasts in an analysis of variance. Imagine that we have a model that estimates θ_j for $j = 1, ..., J$ levels of a treatment. We want to know the magnitude of the difference between level 1 and level 2, that is the posterior distribution of the difference $\delta_{1,2}$ attributable to the treatment. We can

do this by simply calculating a derived quantity at each MCMC iteration:

$$\delta_{1,2}^{(k)} = \theta_1^{(k)} - \theta_2^{(k)} \tag{8.3.1}$$

and using the converged chain for $\delta_{1,2}$ to make an inference. For example, we use the chain as a basis for statements like "The probability that treatment 1 exceeds treatment 2 is p", where p is the proportion of MCMC samples where $\delta_{1,2}$ is positive. This, of course is our estimate of the area of the posterior distribution of the $\delta_{1,2}$ that is greater than 0. It is also a Bayesian p-value.

Though one of the most commonly desired quantities involves a sum or difference of parameters, it is important to note that we can use any mathematical function $h(\boldsymbol{\theta})$ as a derived quantity. For example, recall the matrix population model we used to illustrate the use of multiple sources of data (equation 6.2.57). Often these models are used to learn about the population growth rate (λ) and stable stage structure $\boldsymbol{\omega}$, which for a linear projection matrix, can be derived as the dominant eigenvalue and associated normalized eigenvector of the matrix. In the example above, we could compute the dominant eigenvalue and eigenvector of the matrix $\mathbf{A}$ at each MCMC iteration and use the vector of these accumulated computations across all of the iterations as a basis for inference about growth rate and stable age structure. If we did this for more than one population, we could compute the difference in λ for each iteration and in so doing, make inferences about the probability of differences in growth rates among populations.

In all of the examples above, the posterior distribution of derived quantities can be calculated directly from the MCMC output. All that is required is that we calculate the derived quantity at each iteration based on the values of relevant parameters at that iteration. Alternatively, we can use the MCMC output for the relevant parameters as input for a function for a derived quantity of interest *after* the model has been fit. This allows us to use complex numeric functions that would be difficult to embed in MCMC software, for example, the eigenvalue of a large matrix.

The derived function could involve model parameters, latent processes, and/or data (i.e., $h(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$). However, when obtaining the desired inference about this derived quantity using MCMC samples, it is important to follow the general rules of Monte Carlo integration. That is, when we approximate integrals using sums based on random samples from the distribution of interest, we need to make

sure that the samples for each quantity being combined in the function are "aligned" such that they occurred at the same step in the MCMC algorithm as we illustrated in Box 8.1. This point is subtle, but absolutely critical because the samples arise from an MCMC algorithm in such a way that they are correlated with each other and it is this correlation that provides the proper dependence in the joint posterior distribution.

For example, if we are interested in the posterior mean of the derived quantity $\theta_1^{y_i}/\theta_2$, where y_i is a certain observation in our data set, then we could approximate the necessary integral using

$$\begin{aligned} \mathrm{E}(\theta_1^{y_i}/\theta_2|\mathbf{y}) &= \int\int (\theta_1^{y_i}/\theta_2)\,[\theta_1,\theta_2|\mathbf{y}]d\theta_1 d\theta_2 && (8.3.2) \\ &\approx \frac{\sum_{k=1}^{K}(\theta_1^{(k)})^{y_i}/\theta_2^{(k)}}{K}\,, && (8.3.3) \end{aligned}$$

if $\theta_1^{(k)}$ and $\theta_2^{(k)}$ are samples arising from the same iteration of the MCMC algorithm, for $k = 1, \ldots, K$.

8.4 Predictions of unobserved quantities

A distinguishing feature of Bayesian inference is that unobserved quantities (e.g., parameters, latent state variables, and "future data") are treated as random variables in a Bayesian statistical model. Consequently, inference pertaining to them could be more accurately described as "prediction" rather than "estimation." This interpretation differs from that in non-Bayesian inference, where "unknown observables" (i.e., future data) are treated as random variables but "unknown unobservables" (i.e., parameters) are often treated as fixed quantities. Latent state variables, if stochastic, are treated as random quantities in both Bayesian and non-Bayesian approaches. Therefore, inference pertaining to them could be best summarized as prediction, especially when inference is desired at different times or locations than the data were collected.

We usually build models to make predictions. Predictions provide the foundation for evaluating the ability of our model to adequately represent the distribution of the data (Section 8.1). We will use predictions to evaluate competing models in the next chapter. But predictions are also one of the most useful applications of a model beyond their value in model checking and evaluation. We often want to predict what we would *expect to observe* conditional on what we *have observed*. We covered the principles allowing us to make these predictions in the Section on model checking (8.1) when we described posterior predictive distributions. Here we elaborate further on the utility of

these distributions.

Model predictions are most often used by ecologists in two cases. In the first, most simple case, we have a deterministic model $g(\boldsymbol{\theta}, \boldsymbol{x})$ that we fit to data $\boldsymbol{y}$, $\left[y_i | g(\boldsymbol{\theta}, x_i), \sigma^2\right]$. We want to know the posterior predictive distribution of an unobserved response variable $\tilde{y}_i$ based on a newly observed or postulated value of the predictor variable, $\tilde{x}_i$, that is, a prediction of a new observation, $\tilde{y}_i$ conditional on the data in hand, $\mathbf{y}$.

A critical choice at this point is whether we seek the posterior predictive distribution of the *mean* of the new observation ($[\mathrm{E}(\tilde{y}_i)|\mathbf{y}]$) or if we want the posterior distribution of an *individual observation* ($[\tilde{y}_i|\mathbf{y}]$). This concept might be familiar from training in classical linear regression teaching the difference between the confidence intervals on a single observation of y_i at a given x_i versus confidence intervals on the mean of y_i at a given x_i (Figure 8.4.1). To approximate the posterior predictive distribution of the mean of $\tilde{y}_i$ $\left(\mu_i = \int [\mathrm{E}(\tilde{y}_i)|\boldsymbol{\theta}]\,[\boldsymbol{\theta}|\mathbf{y}]\,d\boldsymbol{\theta}\right)$ and its quantiles (Figure 8.4.1, dashed lines) we calculate a fixed quantity $\mu_i = g\left(\boldsymbol{\theta}^{(k)}, \tilde{x}_i\right)$ at each MCMC iteration. This makes sense because our model predicts the mean (or perhaps some other central tendency) of the independent variable at each value of the dependent variable. We make inference on the mean of the $\tilde{y}_i$ by summarizing the samples of μ_i following the procedures outlined in Section 8.2. If you are using MCMC software, this means that you simply include the output of the deterministic model, $\mu_i = g(\boldsymbol{\theta}, x_i)$ as a line in your code.

To obtain the posterior distribution of an individual observation $\left(\int_{\boldsymbol{\theta}} [\tilde{y}_i|\mathbf{y}, \boldsymbol{\theta}]\,[\boldsymbol{\theta}|\mathbf{y}]\,d\boldsymbol{\theta}\right)$, we calculate $g\left(\boldsymbol{\theta}^{(k)}, \tilde{x}_i\right)$ at each MCMC iteration and then make a draw from the likelihood,

$$\left[\tilde{y}_i^{(k)} | g\left(\boldsymbol{\theta}^{(k)}, \tilde{x}_i\right), \sigma^{2(k)}\right],$$

which you will recognize as the same series of steps we used to simulate new data for posterior predictive checks (equation 8.1.20) except that we are using a newly observed or postulated $\tilde{x}_i$ rather than an x_i from the original dataset. The iterative composition sampling scheme effectively computes the necessary integral and Monte Carlo predictive inference can be obtained using $\tilde{y}_i^{(k)}$ for the $k = 1, \ldots, K$ converged MCMC samples. For example, a posterior predictive mean for $\tilde{y}_i$