MCMC Checklist

Basic Bayesian Model Checking

John W. Seaman, Jr. David J. Kahle James D. Stamey

Baylor University
Department of Statistical Science
1 September 2022
Version 2.0

Contents

The Bayesian paradigm affords a remarkably flexible framework for statistical inference. While posterior inference can make use of prior information and avoids the need for large sample approximations required by much of frequentist modeling, there are other practical burdens to bear. Prior distributions must be constructed and justified, requiring acquisition of historical data, elicitation from subject-matter experts, or both. Posterior sensitivity to prior choices must be examined carefully. Iterative simulation methods such as Markov chain Monte Carlo (MCMC) algorithms allow the practitioner to obtain posterior distributions and examine their features with relative ease, but raise questions of convergence. In what follows we provide concise guidelines for checking convergence and questioning prior choices with suggestions for further reading. These items correspond to a subset of what should be reported in a Bayesian analysis. For an extensive discussion about such issues see Gabry *et al.* (2019) and Kruschke (2021).

In Section 1 we begin with a "quick kill" list of items allowing a rapid assessment of convergence and a list of "due diligence" checks on model features. Each item in these lists refers to a more elaborate discussion in subsequent sections. In Section 2 we consider convergence issues in more detail. Posterior accuracy and suitability are discussed in Section 3. The justification of prior distributions and prior-to-posterior sensitivity are covered in Section 4. We conclude with a brief look at model performance assessment in Section 5.

1 "Quick Kill" and "Due Diligence" Checks

We have found the following items to be essential in assessing convergence. The list can be used by the modeler, or in evaluating a model constructed by others. Often, a lack of convergence can be spotted very quickly. The "quick kill list" can be used to rapidly detect a lack of convergence for the MCMC algorithm. If any of them do not hold, then failure to converge to a stable distribution, model identifiability, or other problems should be investigated. The "due diligence" list details key model features that should always be examined carefully. We consider each in more detail in subsequent sections, as indicated.

Quick Kill Checklist:

1. Kernel density plots of the posterior marginals should be relatively "smooth". See 3, Section 2.

2. History or trace plots should exhibit good "mixing". See 2, Section 2.

3. Autocorrelation plots should dampen quickly. See 6, Section 2.

4. Brooks-Gelman-Rubin diagnostic plots should approach 1. See 8, Section 2. This assumes the modeler employed at least two chains. (Multiple chains are recommended, starting from "overdispersed" initial values.)

5. For fully identified models, prior-to-posterior comparisons (as in 3, Section 2 and 4, Section 4) should reveal change from the prior to the posterior. That is, the prior should be "updated" by the data. Models that are not fully identified, such as purposefully over-parameterized measurement error models, are exceptions.

6. Interval estimates for parameters of interest should be sufficiently narrow to be of practical use. See 3, Section 3.

7. For priors with interval support, the corresponding posterior density plot should not exhibit mass "piling up" near either endpoint without reason. See 4d of Section 4.

Due Diligence Checklist:

1. Prior choices should be justified, whether based on expert opinion or historical data. Commensurability of the current data with historical data sets used in constructing the prior should be examined. Independence assumptions in prior construction should be justified. See Section 4.

2. If informative priors are used, the prior equivalent sample size and prior predictive probability of success should be reported. If either is "too large", then the prior might be too influential. See 5 in Section 4.

3. Prior-to-posterior sensitivity should be examined. Small changes to the prior should yield only small posterior changes. See 4b and 4c in Section 4.

4. Where "diffuse" or "non-informative" priors are used on parameters, induced priors on functions of those parameters should be examined, as they may be unintentionally informative. See 7 in Section 4.

5. Model characteristics such as the use of fixed vs. random effects, constant error variances, exchangeability assumptions, and identifiability should be noted and justified. See 4a of Section 4.

6. The model should be tested using simulated data representing various truth scenarios. The prior predictive distribution is particularly useful for this purpose. See Section 5.

2 Basic Convergence

Before MCMC samples can be used for inference, the chains from which they are drawn must be checked for convergence. By "convergence" we mean that the chain has progressed sufficiently for subsequent draws to be considered approximately distributed as the posterior. To assess convergence of the the chains for a model, there are several things to consider.[1]

1. The number of chains, their initial values, burn-in length, and subsequent chain length should be noted. How the initial values were chosen should be detailed. See Lunn *et al.* (2013), p. 70ff, or Carlin & Louis (2009), pp. 158-159.[2] Attempts to dampen autocorrelation such as thinning should be noted, as discussed in 6 below.

2. Trace (history) plots including all chains should be provided for each parameter. These plots can indicate problems with convergence.[3] Chain lengths needed for convergence

[1] For a more detailed development, see Gelman & Shirley (2011).

[2] For an alternate view on the use of multiple chains see, for example, Geyer (2011), who recomends the use of single, very long chains.

[3] There are R packages that provide a variety of posterior plots and diagnostic tools. For example, bayesplot returns ggplot objects and includes familiar tools like trace plots, autocorrelation function plots, and kernel density estimates. MCMCvis and Shinystan are other options. These can be used with output from most iterative simulation software, such as BUGS, JAGS, Nimble, Stan, and the SAS MCMC procedure.

may vary across parameters and starting values. Examples for good and bad single chains are shown in Figures 1 (left) and 2. Figure 3 exhibits trace plots for multiple chains. Trace plots suggest a sufficient number of iterations for convergence if they "mix". That is, the plots of the individual chains merge after convergence.[4]

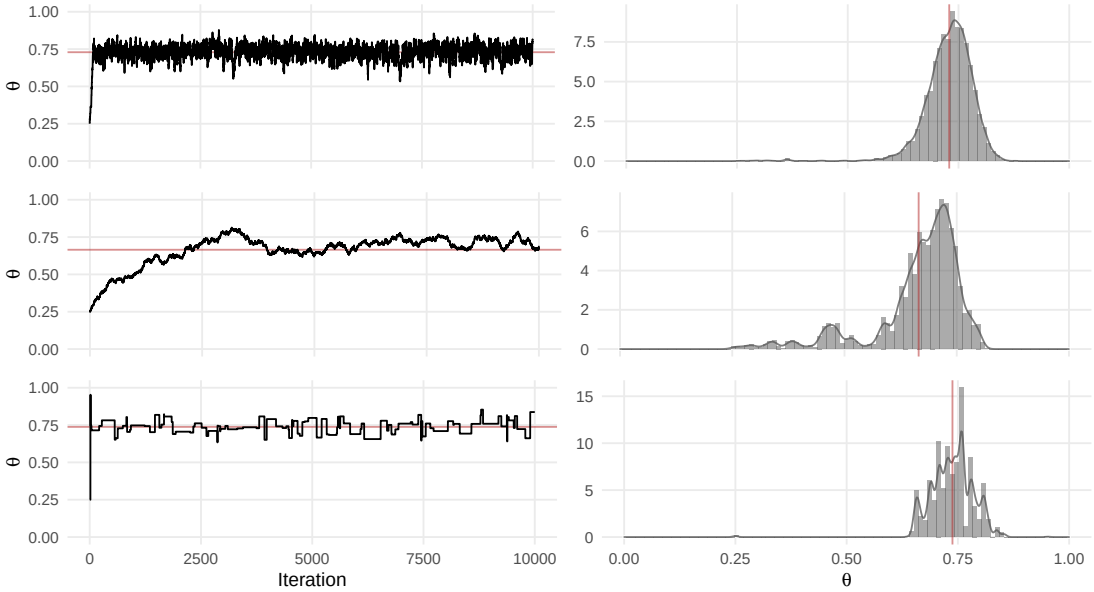

Figure 1: Trace plots along with histograms and KDEs for three different chains. At the top, a well-behaved chain yields a relatively smooth KDE. The the poor convergence depicted in the middle and bottom trace plots results in correspondingly rough KDEs.

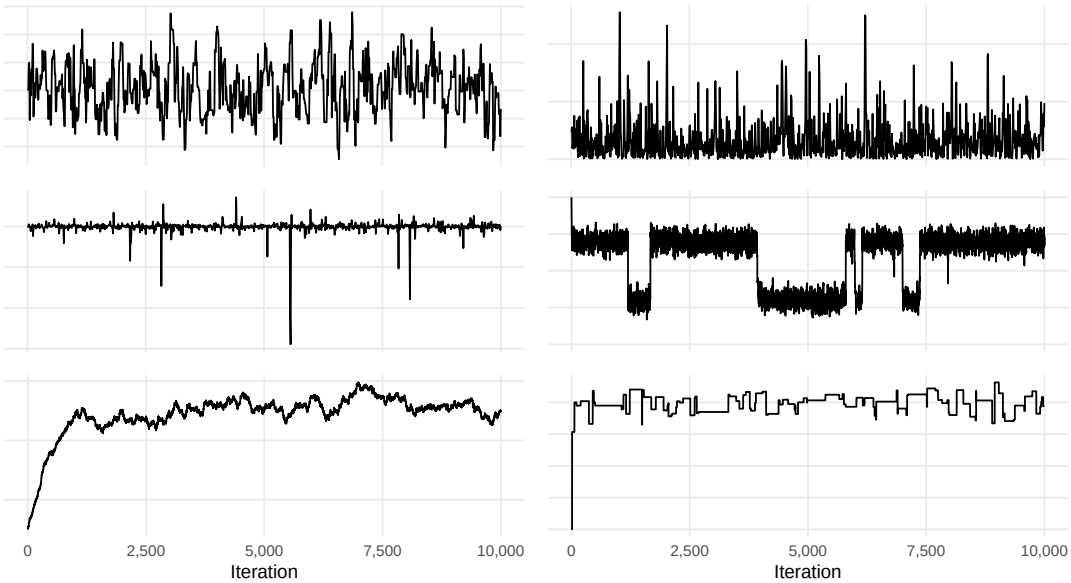

Figure 2: History (trace) plots for a hypothetical parameter. Top left: The chain appears to have converged to a symmetric stable distribution; Top right: This chain appears to have converged to a positively skewed distribution; Middle left: The chain has converged to a heavy-tailed distribution; Middle right: This chain exhibits poor mode switching from a bimodal target distribution; Bottom left: Here, burn-in values (the "tail" on the left) should be discarded and high autocorrelation is slowing convergence. Bottom right: This trace suggests an inefficient sampler due to a proposal distribution with an overly large variance. This is a frequent problem when using standard Metropolis-Hastings algorithms to sample from a joint posterior with probability mass concentrated in a relatively low-dimensional subset of a high-dimensional space.

[4]See Rosner *et al.* (2021), Section 4.4.5 for a thorough discussion of this and other aspects of convergence diagnostics. See also Gelman *et al.* (2013), Section 11.4, and Carlin & Louis (2008), p. 156ff.

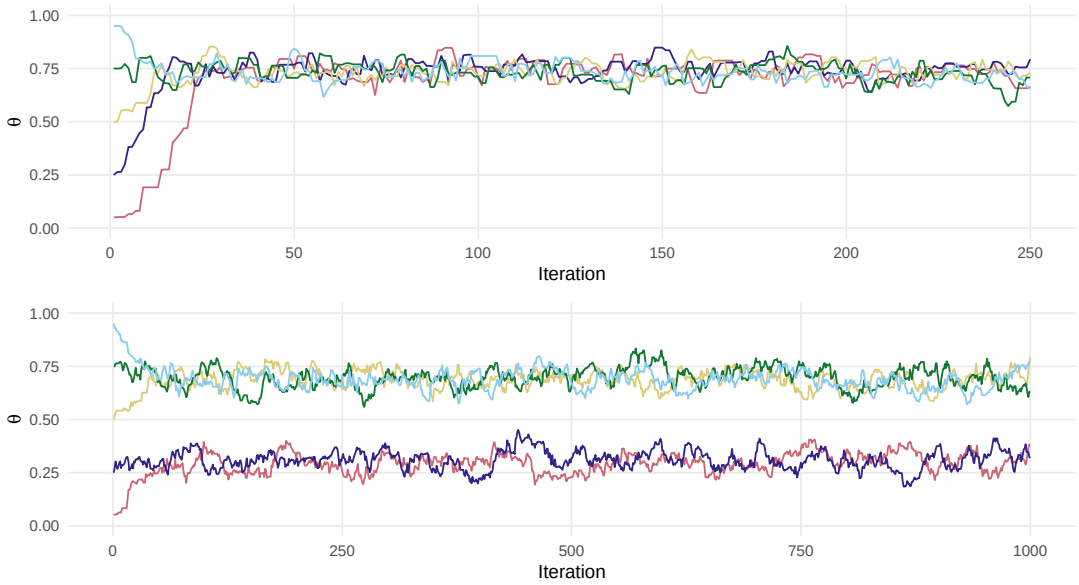

Figure 3: History (trace) plots for multiple chains from the posterior of a parameter. Top: These five chains exhibit good mixing after a burn-in of about 50 iterations. Bottom: Poor mixing of the plots is evident for these five chains, even after burn-in iterations.

3. Histograms and kernel density estimates of all marginal priors and corresponding marginal posteriors should be examined.[5] Functions of model parameters (e.g. hazard ratios or ED50's) will require computation of induced priors via separate simulation. The "smoothness" of a kernel-density estimate (KDE) can suggest the extent of convergence, as illustrated in Figure 1. This is a subjective assessment and will depend on graphics options, such as features of the KDE. Histograms should be examined as well because kernel density smoothers can yield misleading results, especially where there are large probability concentrations in a relatively small subset of the support.

4. To gauge the prior's influence on the posterior, it helps to make comparisons to a posterior obtained with a relatively diffuse prior on the same parameter. Alternatively, the maximum likelihood estimate of the parameter can be indicated on the posterior plot. This is approximately the posterior mode under a very diffuse prior for that parameter and suggests what the data "says" without prior influence.

5. Estimates of posterior means and quantiles should stabilize with increasing chain length. To check this, plot the posterior mean and, say, two quantiles of interest, as depicted in Figure 4. See Rosner *et al.* (2021), p. 104ff and Section 3.

6. MCMC algorithms produce correlated iterates, by construction. Thus, autocorrelation plots for each parameter should be examined. As autocorrelation in the chain increases, requisite chain-lengths required for convergence and subsequent inference also increase. That is, the effective sample size from the posterior is reduced as autocorrelation in the chain increases. It is often helpful to center continuous covariates. Autocorrelation can be alleviated by thinning, that is, retaining only every kth iterate, where k is typically similar to the lag required to dampen autocorrelation evident in plots such as these. However, thinning necessarily decreases the amount of information in the MCMC sample. Such thinning should be reported. See Figure 5 for an illustration.

[5]Such plots may be problematic if the prior is extremely diffuse compared to the posterior; that is, the relative maxima of prior and posterior densities may make plotting difficult.

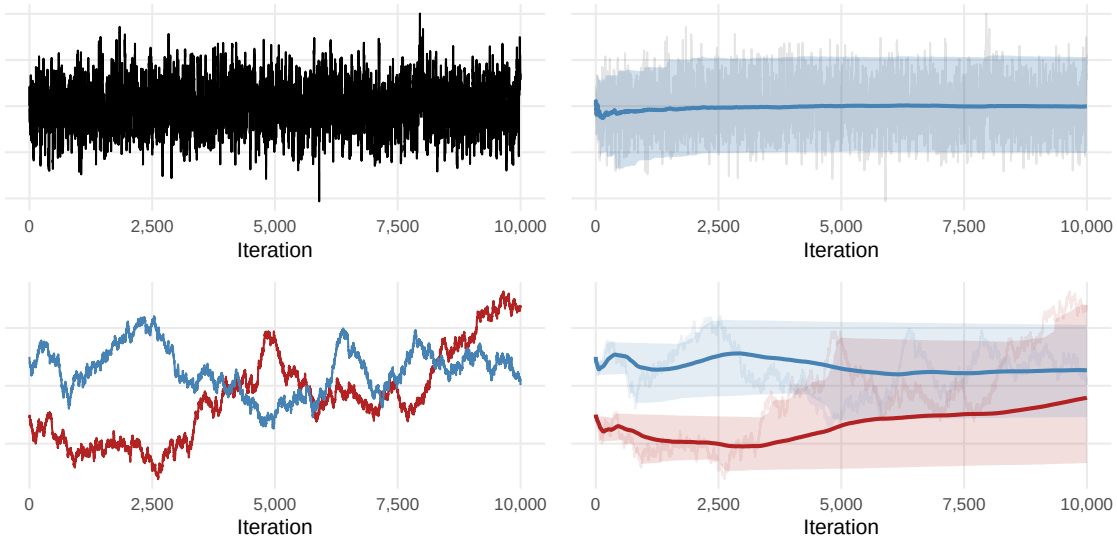

Figure 4: History (trace) plots of two chains that start at different values along with their mean and quantile plots. Top left: The trace of two well-behaved chains. Top right: The corresponding history plot of posterior means and quantiles, all well-behaved. Bottom left: Two chains exhibiting poor mixing and high autocorrelation. Bottom right: The corresponding badly behaved mean and quantile plots.

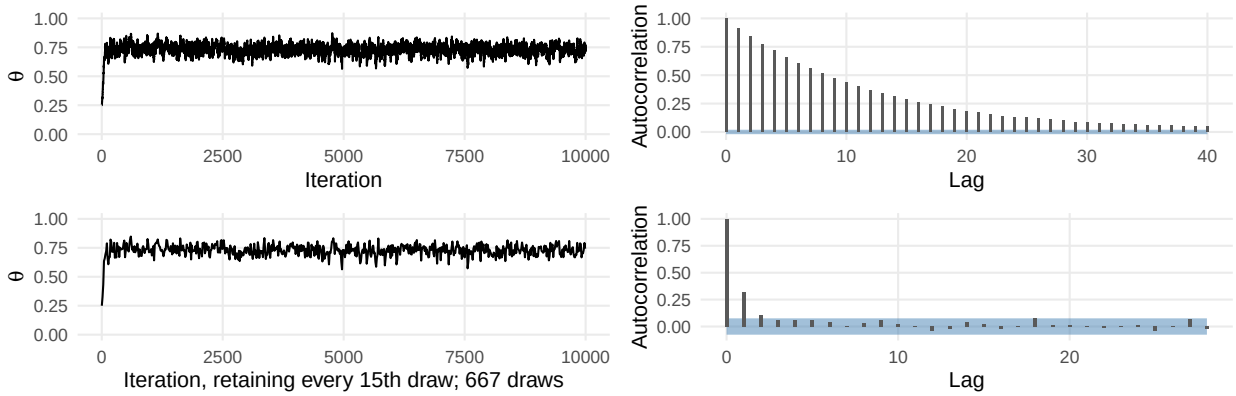

Figure 5: The trace displayed in the upper left plot exhibits possibly problematic correlation, as revealed in the autocorrelation function plot at top right. Thinning at $k = 15$ or so will render the autocorrelation negligible, as seen in the autocorrelation function plot depicted bottom right. However, such thinning reduces the effective sample size to 667 draws for inference, down from 10,000.

7. Parameters can be dependent *a posteriori* even when corresponding priors are constructed to be independent. If parameters are highly dependent, using posterior marginals for inference can be problematic. See, for example, Carlin & Louis (2008), p. 159. Thus, it is prudent to examine bivariate plots for each pair of the parameters. This can be done, for example, by exporting the chains to R and use the `pairs()` function in the `GGally` package, which employs `ggplot2` graphics. See Figure 6.

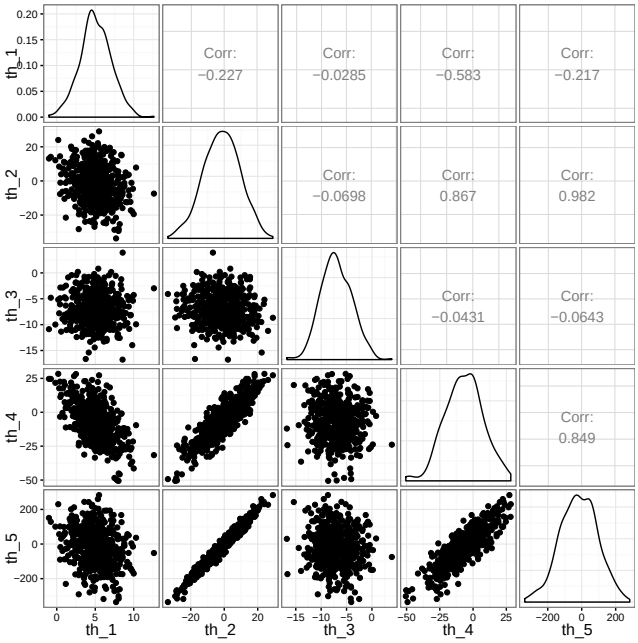

Figure 6: Scatter plots of posterior samples from pairs of five parameters. Parameter pairs $(\theta_1, \theta_4), (\theta_2, \theta_4), (\theta_2, \theta_5)$, and (θ_4, θ_5) are evidently dependent.

8. In general, multiple chains should be run starting at over-dispersed initial values. After convergence, the within-chain variability should resemble the between chain variability. The Brooks-Gelman-Rubin (BGR) diagnostic assesses between- and within-chain variability and should be plotted. The R package `coda` computes this diagnostic using the `gelman.diag` function and plots can be constructed with `gelman.plot`. Figure 7 provides an illustration. See, for example, Rosner *et al.* (2021), p. 108ff.

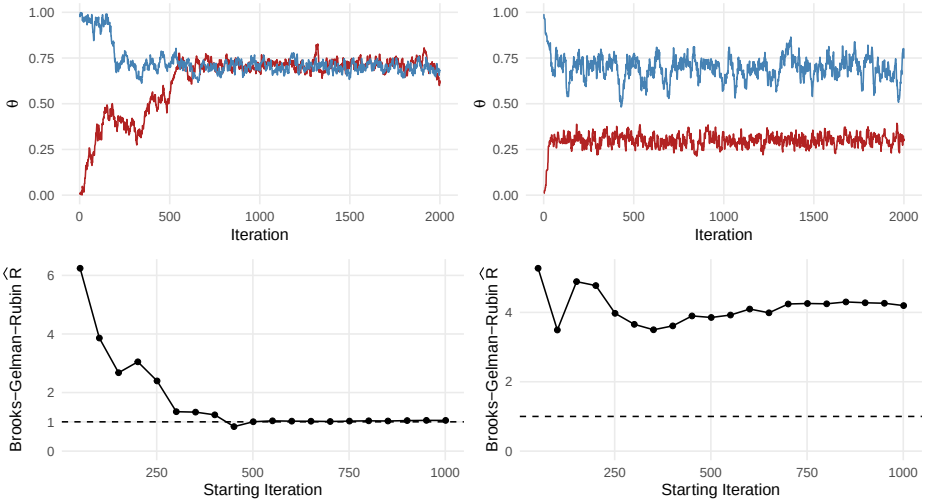

Figure 7: Brooks-Gelman-Rubin plots for two chains. The left column exhibits two chains with good mixing as indicated by the BGR plot, which converges to 1 at about the 500th iteration. The right column also shows two chains, however mixing does not occur within the depicted chain length. Note that the BGR plot does not apporach 1.

3 Posterior Accuracy and Suitability

Markov chain Monte Carlo algorithms produce a Markov chain from which posterior samples can be drawn. However, the chain length must be sufficient to adequately represent its stationary distribution. How we recognize adequate chain length is as much of an art as a science. In fact, we must manage a compound approximation here. The chain must be long enough that new iterates may be considered as sampled from the posterior (the stationary distribution) and the number of subsequent iterates must be large enough to adequately approximate the posterior feature of interest. That is, once "convergence" obtains, we must run the chain long enough to obtain a good approximation of the posterior feature of interest.

1. To assess such approximation goals, the Monte Carlo standard error (MCSE) and the posterior standard deviation should be included for each parameter. A rule of thumb is that the chain should extend beyond convergence long enough to render the MCSE no more than 5% of the posterior standard deviation for the parameter. See Cowles (2013), pp. 137-138, and Lunn *et al.* (2013), p. 77ff.[6]

2. For a visual assessment, plot posterior summary values (mean, standard deviation, percentiles) as a function of chain-length. Once all such summaries have stabilized within prescribed boundaries, chain length adequate for accuracy can be said to have been attained. This is illustrated in Figure 8. See, for example, Lunn *et al.* (2009), p. 79. The chain lengths you utilize should be at least that required by the most "demanding" parameter.

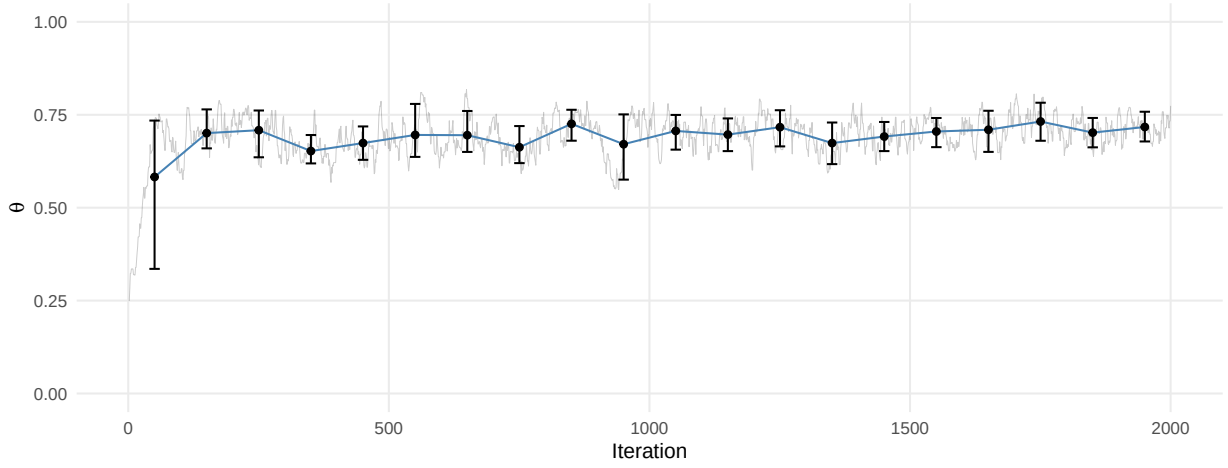

Figure 8: Posterior means plotted as a function of iterations.

3. Interval estimates for parameters of interest should be sufficiently narrow for practical purposes. For a fixed sample size, this need must be carefully balanced against the justifiability of informative priors. Priors can usually be made sufficiently informative as to render credible sets arbitrarily narrow. However, doing so may result in unreasonably high prior probability of study success, or unrealistically large prior effective sample size. For more on justification of priors, see Section 4.

[6] A more extensive discussion is provided in Flegal *et al.* (2008).

4 Prior Justification and Prior-to-Posterior Sensitivity Analysis

The posterior distribution clearly depends on both the choice of data model and the joint prior distribution. The former is an issue for any approach to modeling. The latter is unique to Bayesian methods. Priors should be carefully justified, even if using relatively non-informative priors. Furthermore, posterior results should be reasonably robust with respect to prior choices. That is, "small" changes in the joint prior structure should not produce "large" changes in posterior results. Thus, checking the sensitivity of focal posterior results to changes in the prior structure is a necessary task in Bayesian model building. The list below provides guidance for prior-to-posterior sensitivity analyses.[7]

1. The choice of each prior distribution should be justified.
   (a) Justification is needed even for relatively non-informative (diffuse) priors on parameters, especially if functions of those parameters are of interest, as their induced priors may in fact be informative. See 7 below.
   (b) Furthermore, priors intended to be relatively non-informative can be chosen to be unnecessarily diffuse, sometimes resulting in convergence problems. For example, suppose a relatively non-informative prior is desired for the mean shelf life (in years) of a drug product. A normal prior with a standard deviation of 100 years is surely overly diffuse. For more on this, see 4d and 4e below.
2. Parameters are often given independent priors for convenience, especially if relatively non-informative priors are desired. Assuming independent priors for a normal mean and variance may be quite plausible in many situations, for example, as might modeling variance components as independent. Of course, *a priori* independence need not yield *a posteriori* independence. Examining posterior plots of pairs of parameters can be revealing, even when independent priors are used, as we saw in 7 of Section 2. Strong posterior bivariate relationships suggest that modeling dependency in the joint prior might have afforded more efficient use of the data.
3. Assuming independence in the presence of evident dependencies should be avoided if possible. For example, it would be unreasonable to assume parameters related to product safety and efficacy are independent.
4. Gauging the influence of the prior on the posterior is clearly a critical aspect of Bayesian inference.
   (a) It may be that the posterior is little changed from the prior for some parameters. Indeed, in the absence of MCMC convergence problems, identifiability issues manifest themselves in the failure of a prior to be updated *a posteriori*.
      i. That is, the posterior for the unidentified parameter looks pretty much like the prior, with the same location and dispersion. In effect, as Robert (2001, p. 24) notes, "An important aspect of the Bayesian paradigm in non-identifiable settings is ... that the prior distribution can be used as a tool to *identify* the

[7] More formal approaches have been considered in the literature. Such methods are beyond the scope of this document. For a good brief overview of formal methods see Roos *et al.* (2015). In addition, they offer a relatively practical approach to formal sensitivity analysis that has the virtue of reproducibility.

parts of the parameter that are not covered by the likelihood, even though the choice of prior may have a bearing on the identifiable part."

ii. An important consequence of a lack of identifiability when implementing iterative simulation methods like the Gibbs sampler is the possibility of "drifting". If the prior on a parameter is not informative, then trajectories of the Markov chain for components of that parameter may drift to extreme values. The diffuse prior structure offers no centering effect. This can compromise convergence assessment and yield inaccurate estimates.

iii. There is nothing about the mechanics of Bayesian inference that requires indentifiability in models. In fact, no statistical methodology is immune from its consequences. In Bayesian inference, nonidentifiability may cause convergence issues with MCMC methods, as noted above. One area in which Bayesian methods provide a considerable advantage in the presence of non-identified parameters is models with mismeasured variables. Gustafson (2004, p. 55) notes that

> "An interesting facet of Bayesian inference is that in one sense it 'works' whether or not one has parameter identifiability. That is, the mechanics of forming a posterior distribution and obtaining parameter estimates from this distribution can be carried out equally well in nonidentifiable situations, modulo some technical concerns about MCMC... Of course in another sense no form of inference can work in a nonidentifiable model, as estimators which tend to true parameter values do not result. One intuitive way of thinking about Bayesian inference in the absence of parameter identifiability is that the prior distributions play more of a role than usual in determining the posterior belief about the parameters having seen the data."

(b) Small changes in prior location or scale should result in small corresponding posterior changes. In particular, slightly altering the prior should not change decisions based on the posterior.

(c) The effect of substantial changes in the prior should also be examined. It can be useful to compare posterior results under vague ("relatively non-informative" or "diffuse"), skeptical, enthusiastic, clinical (i.e., expert-elicited), and just-significant (i.e., barely meeting success criteria). For examples of such "archetypal" priors see Lesaffre & Lawson (2012) and Lunn *et al.* (2011). The latter illustrates use of skeptical priors in two clinical trials, and an enthusiastic prior for another. Speigelhalter *et al.* (2004) discuss the use of skeptical priors at length.

(d) Assumptions made about variance components can greatly affect posterior results. Choosing priors on such parameters is highly problem-dependent. Gamma priors for precision and inverse-gamma priors for variances, both with very small shape parameters (e.g., 0.001), were once widely used. However, they are now known to be problematic (Gelman, 2006; Gelman *et al.*, 2014). Alternatives, such as half-normal, half-t, or uniform distributions on standard deviations are often preferable. Placing a uniform prior with support $(0, B], B > 0$, on a standard deviation requires specification of the upper bound, B. This choice should be large enough to prevent probability mass from "piling up" at B. See Figure 9 for an illustration.

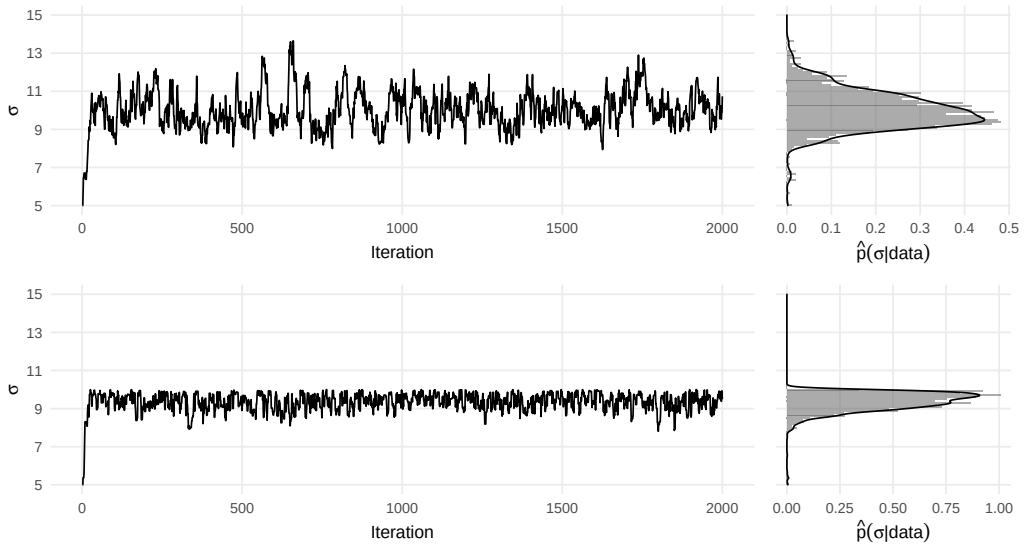

Figure 9: Posterior plots for a standard deviation with a $U(0, B)$ prior. The KDEs are approximate marginal posteriors for the standard deviation. In the lower plot, $B = 10$, resulting in a "pile-up" of probability mass at that bound. In the upper plot, $B > 14$, resulting in a reasonable posterior plot.

(e) Justification of priors on variance components should be done in terms of standard deviations, arguing from operational scales.

i. As an example, suppose you have a model including the mean systolic blood pressure, μ, in mm/HG, for subjects in some population of interest. If we use the prior $\mu \sim \mathrm{N}(120, \sigma_0^2)$, then selecting, say, $\sigma = 100$, is clearly silly.

ii. Spiegelhalter *et al.* (2004, p. 170) provide an excellent illustration of this problem in the context of a meta-analysis involving log-odds ratios.

- Suppose we model log-odds ratios as $\theta_i \sim \mathrm{N}(\mu_\theta, \sigma_\theta^2)$. That means we possess prior information implying 95% of the θ_i's would fall within $\mu_\theta \pm 1.96\sigma_\theta$. That is, for the ith study,

$$\theta_i \sim \mathrm{N}(\mu_\theta, \sigma_\theta^2) \implies \Pr\underbrace{(\mu_\theta - 1.96\sigma_\theta < \theta_i < \mu_\theta + 1.96\sigma_\theta)}_{\text{width } 3.92\sigma_\theta} = 0.95$$
$$\implies \exp(3.92\sigma_\theta) \text{ approximate "range" of the odds ratios.}$$

- If, for instance, you think the odds-ratios are unlikely to vary by more than an order of magnitude, then you might take

$$\exp(3.92\sigma_\theta) = 10 \implies \sigma_\theta = \log(10)/3.92 \approx 0.59$$

as a "high" value for the standard deviation. So, choosing $\sigma_\theta = 0.6$ would be conservative, given your beliefs. A prior-to-posterior sensitivity analysis should then be conducted, bracketing $\sigma_\theta = 0.6$ with a grid of plausible values.

iii. As another illustration, suppose the posterior feature of interest is an 80% credible interval on some parameter for which a $\mathrm{U}(0, B)$ must be specified. It is straightforward, if computationally intensive, to perform a sensitivity analysis *a posteriori* to verify that the choice of B has not impacted posterior

inference. We illustrate this in Figure 10. To aid in choosing B, a similar examination might be performed *a priori* using samples drawn from the prior predictive distribution. A similar sensitivity analysis can be performed for any prior parameter.

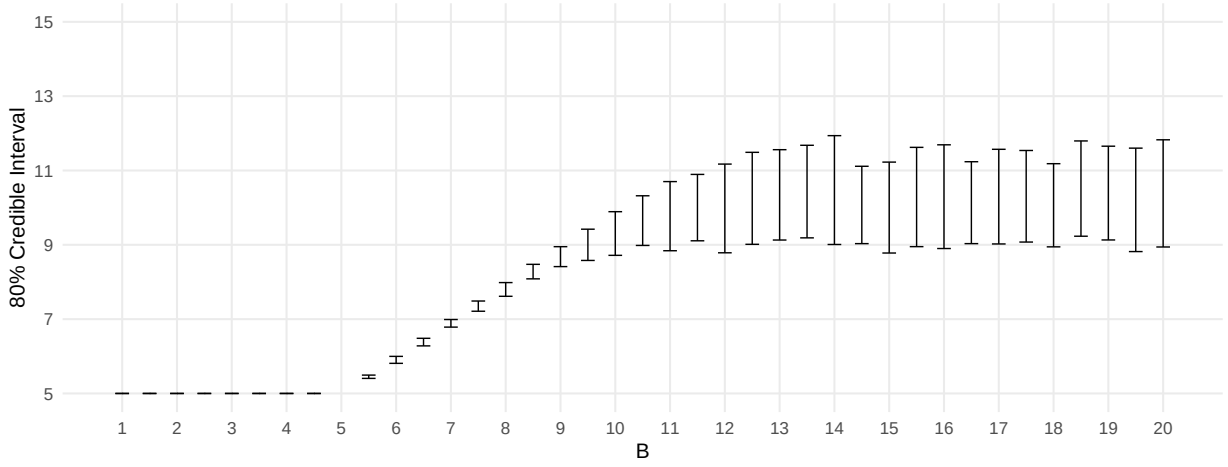

Figure 10: Prior-to-posterior sensitivity analysis for a parameter with a $\mathrm{U}(0, B)$ prior. The model is refit for each value of B, so this is a potentially computationally intensive task. The widths appear to stabilize beyond $B \approx 13$.

5. The prior equivalent sample size (ESS) is an important and intuitive measure of prior informativeness. See Morita *et al.* (2008, 2010, 2012) and the R code referred to therein. This ESS approximates the sample size that would be necessary to achieve similar posterior results if a vague prior structure had been used. Following the FDA (2010) guidance on the use of Bayesian methods in device trials, the prior probability of study success should be carefully considered.

6. The use of large historical studies in prior construction can yield inappropriately small prior variances with correspondingly high posterior influence. Methods such as power priors can be used to attenuate this influence (see, for example, Ibrahim *et al.*, 2015 and Neuenschwander *et al.*, 2009). Power priors effectively attenuate the variance of the resulting prior. For example, in the normal sampling case, attenuation is by a factor of $1/a_0$, $0 < a_0 < 1$. Clearly, a_0, known as the power parameter, must be very carefully chosen. Placing a prior on a_0 is not recommended. Instead, one can plot the prior probability of study success against values of a_0 to assess the power parameter's influence. Computing the prior ESS as a function of a_0 is another approach. See Figure 11 for illustrations. Again see the papers by Morita *et al.* See also FDA (2010), pp. 26, 39. Thompson *et al.* (2021) introduce a new measure of similarity between the current and historical data that limits similarity by a pre-specified margin.

7. Induced priors should be examined carefully. Suppose you construct a joint prior on a parameter vector $\boldsymbol{\theta} \in \Theta$. Let $\varphi(\boldsymbol{\theta})$ be a function of interest. If φ is a linear function, and the elements of $\boldsymbol{\theta}$ are given, say, independent diffuse normal distributions, then $\varphi(\boldsymbol{\theta})$ can be expected to be diffuse. However, if $\varphi(\boldsymbol{\theta})$ is nonlinear, the induced prior may be very, and unintentionally, informative. See Seaman *et al.* (2012).

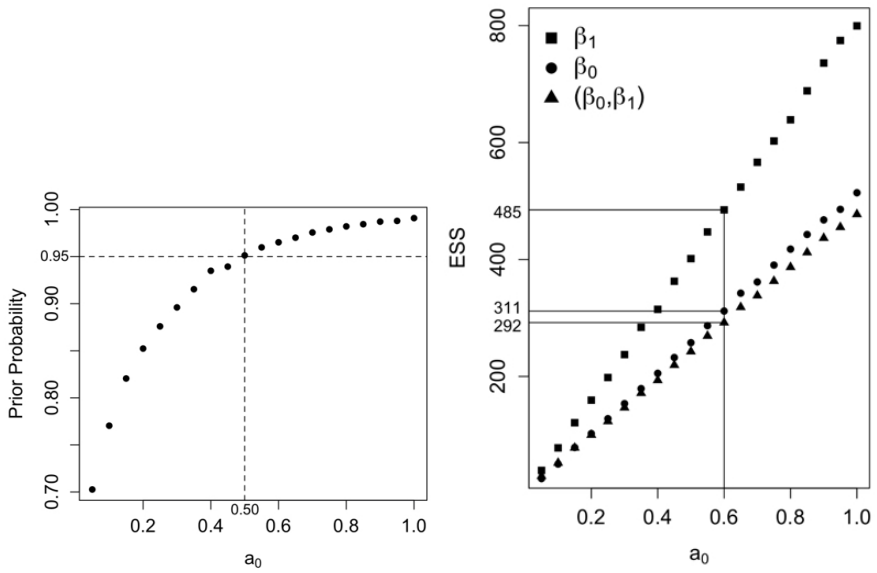

Figure 11: Two methods for justifying the choice of a fixed power parameter, a_0, when using power priors. In the plot at left, the prior probability of success for some model is plotted against values of the power parameter, a_0. The value of a_0 can be chosen so as to achieve a prior probability of success no larger than some specified value. In this case, all values of a_0 yield relatively large values for the prior probability of study success. In the plot at right, we have illustrated selection of a_0 with respect to prior ESS. Here, prior ESS is computed for each parameter, as well as the joint parameter, for a hypothetical two-parameter model. Selection of $a_0 = 0.6$ corresponds to a prior ESS of approximately 292 for the vector parameter.

8. If several models are under consideration, sensitivity analyses should precede model selection. Model choice involves deciding which among several models is most appropriate.[8] Candidates may differ in both their data models (likelihood functions) and corresponding prior structures. The candidate models should each have been subjected to a sensitivity analysis *before* the selection process.

5 Model Performance

A thorough treatment of model performance analysis is beyond the scope of this guide, but here are some brief comments.

1. A Bayesian model requires specification of a likelihood function, $l(\boldsymbol{\theta}|\mathbf{x})$, for parameter vector $\boldsymbol{\theta} \equiv (\theta_1, \ldots, \theta_p)$ and data $\mathbf{x}$, as well as a joint prior structure, $p(\boldsymbol{\theta})$. The joint probability distribution, $f(\mathbf{x}|\boldsymbol{\theta})$, associated with the likelihood function can be used to simulate data sets, with parameter values generated using a joint "design prior", $p_D(\boldsymbol{\theta})$, analogous to Bayesian sample size determination.[9] Location of the design prior is chosen to reflect θ_i values of interest. Dispersion parameters for $p_D(\boldsymbol{\theta})$ are chosen to reflect uncertainty about which values of the θ_i's will obtain in an experiment. For each generated data set, the model is fit using the joint prior, $p(\boldsymbol{\theta})$. More than one design prior can be chosen to represent different scenarios, e.g., optimistic or skeptical.

[8] Or the use of more sophisticated methods such as Bayesian model averaging or mixture models.
[9] See, for example, Brutti *et al.* (2008) for more detail.

2. Testing a model under various truth scenarios (values of $\boldsymbol{\theta}$) in the fashion described above can result in hundreds or even thousands of MCMC runs. Clearly not all can be checked for convergence, but we recommend doing so for a small subset of such runs, chosen to represent the range of parameters generated by the design prior.

3. Summarizing the results of testing a model across multiple model characteristics (diffuse priors vs. informative priors, for example) and/or truth scenarios can be done compactly with graphical tools. A particularly useful graphical summary can be constructed like that illustrated in Figure 12.

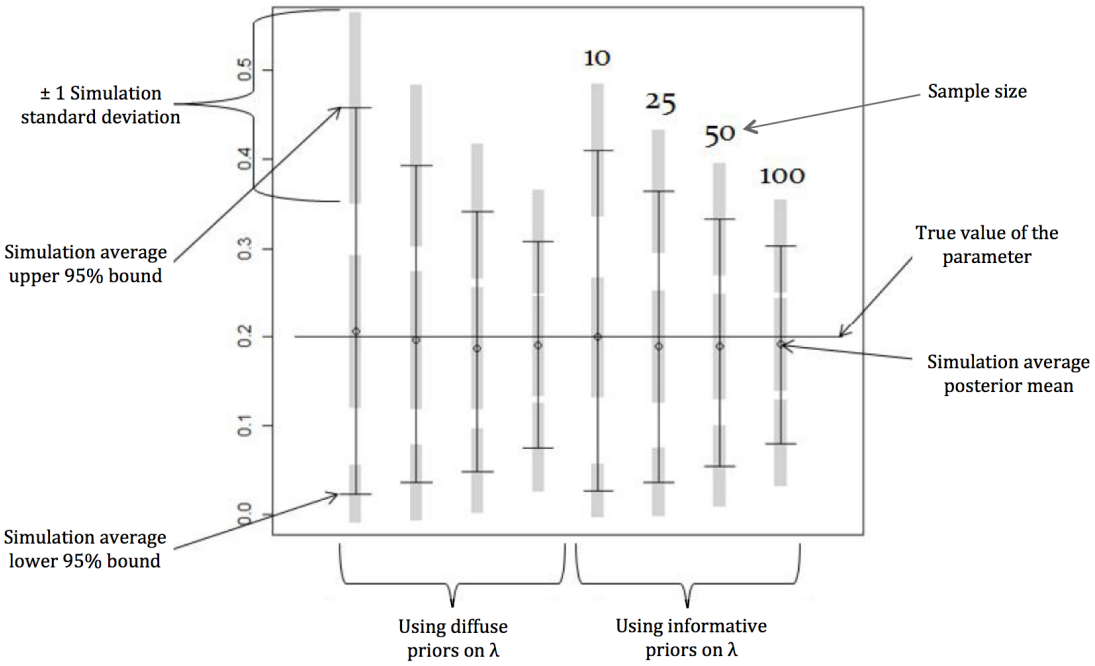

Figure 12: Simulation summary plot for 500 samples simulated to study model performance under a single truth scenario, indicated by the horizontal line at 0.2. In this case, the plot compares model performance for a diffuse prior structure to an informative prior structure for a hypothetical parameter, λ, at various simulated sample sizes. Each vertical bar summarizes the results of multiple simulations. Included are the average of 500 posterior means, 500 upper and lower 95% credible set bounds, and associated ± 1 simulation standard deviation bars (in gray). In some cases, simulation and/or posterior medians or other measures may be preferred to simulation averages.

6 References

Brutti P, De Santis F, Gubbiotti S, (2008), Robust Bayesian sample size determination in clinical trials, *Statistics in Medicine*, **27**, 2290-2306.

Carlin, B. and Louis, T. (2008) *Bayesian Methods for Data Analysis*, 3rd. Ed., CRC Press: Boca Raton.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. (2011) *Bayesian Ideas and Data Analysis*, CRC Press: Boca Raton.

Cowles M. (2013) *Applied Bayesian Statistics*, Springer: New York.

Food and Drug Administration (2010) "Guidance for the use of bayesian statistics in medical device clinical trials," at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials

Flegal, J., Haran, M. and Jones, G. (2008) "Markov chain Monte Carlo: can we trust the third significant figure?," *Statistical Science*, **23**(2), 250-260.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019) "Visualization in Bayesian workflow," *J. Royal Statistical Society A*, **182** Part 2, 389-402.

Gelman, A. (2006) "Prior distributions for variance parameters in hierarchical models (Comment on Article by Browne and Draper)," *Bayesian Analysis*, **1**(3), 515-534.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2014) *Bayesian Data Analysis*, 3rd Ed., CRC Press: Boca Raton.

Gelman, A. and Shirley, K. (2011) "Inference from simulations and monitoring convergence," in *Handbook of Markov Chain Monte Carlo*, edited by Brooks, S., Gelman, A, Jones, G, and Meng, X., CRC Press: Boca Raton, pp. 163-174.

Geyer, C. (2011) "Introduction to Markov chain Monte Carlo," in *Handbook of Markov Chain Monte Carlo*, edited by Brooks, S., Gelman, A, Jones, G, and Meng, X., CRC Press: Boca Raton, pp. 3-48.

Gustafson, P. (2004) *Measurement Error and Misclasification in Statistics and Epidemiology*, Chapman and Hall.

Gustafson, P. (2015) *Bayesian Inference for Partially Identified Models*, CRC Press: Boca Raton.

Ibrahim, G., Chen, M., Gwon, Y., and Chen, F. (2015) "The power prior: theory and applications," *Statistics in Medicine*, **34**(28), 3724-3749.

Kruschke, J. (2021) "Bayesian analysis reporting guidelines," *Nature Human Behaviour*, **5**, 1282-1291.

Lesaffre, E. and Lawson, A. (2012) *Bayesian Biostatistics*, Wiley: New York.

Lunn, D., Jackson, C., Best, N., Thomas, A., and Spiegelhalter, D. (2013) *The BUGS Book*, CRC Press: Boca Raton.

Morita, S., Thall, P., and Müller, P. (2008) "Determining the effective sample size of a parametric prior," *Biometrics*, **64**(2), 595-602.

Morita, S., Thall, P., and Müller, P. (2010) "Evaluating the impact of prior assumptions in Bayesian biostatistics," *Statistics in Biosciences*, **2**, 1-17.

Morita, S., Thall, P., and Müller, P. (2012) "Prior effective sample size in conditionally independent hierarchical models," *Bayesian Analysis*, **6**(3), 591-614.

Murphy, J., Hofer, J. (2002) "Establishing shelf life, expiry limits, and release limits," *Drug Information Journal*, **36**(4):769–781.

Ntzoufras, I. (2009) *Bayesian Modeling Using WinBUGS*, Wiley: New York.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. (2009) "A note on the power prior," *Statistics in Medicine*, **28**, 3562-3566.

Robert, C. (2001) *The Bayesian Choice*, 2nd. Ed., Springer-Verlag: New York.

Roos, M., Martins, T., Held, L., and Rue, H. (2015) "Sensitivity analysis for Bayesian hierarchical models," *Bayesian Analysis*, **10**(2). 321-349.

Rosner, G., Laud, P. Johnson, W. (2021) *Bayesian Thinking in Biostatistics*, CRC Press: Boca Raton.

Seaman, J. III, Seaman, J. Jr., and Stamey, J. (2012) "Hidden dangers of specifying noninformative priors," *The American Statistician*, **66**(2), 77-84.

Spiegelhalter, D., Abrams, K., and Myles, J. (2004) *Bayesian Approaches to Clinical Trials and Health-Care Evaluation*, Wiley: Chichester, England.

Thompson, L., Chu, J., Xu, J., Li, X., Nair, R., and Tiwari, R. (2021) "Dynamic borrowing from a single prior data source using the conditional power prior," *J. of Biopharmaceutical Research*, **31**(4), 403-424.