Markov Chain Monte Carlo

Galin L. Jones

School of Statistics

University of Minnesota

Draft: March 22, 2024

Contents

1 Introduction

Later.

2 Markov Chains

Let $(\mathsf{X}, \mathcal{B})$ be a measurable space. A sequence of X-valued random variables $\{X_1, X_2, X_3, \ldots\}$ is a Markov chain if for all g

$$E\left[g(X_{n+1}, X_{n+2}, \ldots) \mid X_n, X_{n-1}, \ldots, X_1\right] = E\left[g(X_{n+1}, X_{n+2}, \ldots) \mid X_n\right].$$

Then P is a *Markov kernel* if $P : \mathsf{X} \times \mathcal{B} \to \mathbb{R}$ satisfying (i) for each fixed $x \in \mathsf{X}$, $P(x, \cdot)$ is a probability measure and (ii) for each fixed $B \in \mathcal{B}$, $P(\cdot, B)$ is a measurable function.

When X is a discrete set a Markov kernel can be represented as a square matrix whose entries are nonnegative and whose rows sum to 1.

Example 2.1. Suppose

$$P = \begin{pmatrix} 1/2 & 1/2 \\ 1/3 & 2/3 \end{pmatrix}$$

Then P is a Markov matrix on two states $\{0, 1\}$, say. The first row for example, is interpreted as the probability of moving in one step from state 0 to state 0 is $1/2$ which is the same as the probability of moving in one step from state 0 to state 1.

Example 2.2. Let $\mathsf{X} = \mathbb{Z}$ and let $0 < \theta < 1$. If $x \geq 1$, then a Markov kernel is defined by the matrix P with elements

$$P(x, x+1) = P(-x, -x-1) = \theta, \qquad P(x, 0) = P(-x, 0) = 1 - \theta,$$

and $P(0, 1) = P(1, 0) = 1/2$.

Often X will be uncountable and $\mathcal{B}$ will be countably generated. If X is topological, then $\mathcal{B}$ will be the Borel σ-algebra generated by X.

Example 2.3. Let $\mathsf{X} = (0, 1)$ and consider the Markov chain that evolves as follows. Draw $U \sim \text{Uniform}(0, 1)$. If $u \leq 0.5$, $X_{n+1} \sim \text{Uniform}(0, X_n)$, but if $u > 0.5$, $X_{n+1} \sim \text{Uniform}(X_n, 1)$. Then if $X_n = x$ and $B \in \mathcal{B}$

$$P(x, B) = \int_B \left[\frac{1}{2} \frac{1}{x} I_y((0, x)) + \frac{1}{2} \frac{1}{1-x} I_y(x, 1) \right] \mathrm{d}y.$$

In Example 2.3, the integrand in the Markov kernel is a conditional density on X. This is a setting that will be encountered repeatedly throughout. If there is a conditional density $k(y \mid x)$, with respect to a measure λ, such that the Markov kernel satisfies for $B \in \mathcal{B}$

$$P(x, B) = \int_B k(y \mid x) \lambda(\mathrm{d}y),$$

then say k is a *Markov transition density.*

Example 2.4. Suppose $f(x, y)$ is a joint density with support $\mathbb{R}^2$ and conditional densities $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. Then

$$k(x', y' \mid x, y) = f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x')$$

is a Markov transition density. The Markov chain evolves from $(X_k = x, Y_k = y)$ to (X_{k+1}, Y_{k+1}) by drawing $X_{k+1} \sim F_{X|Y}(\cdot \mid y)$ followed by $Y_{k+1} \sim F_{Y|X}(\cdot \mid X_{k+1})$. This is a special case of the so-called two-variable Gibbs sampler.

Suppose λ is a positive measure on $(\mathsf{X}, \mathcal{B})$, define

$$\lambda P(B) = \int_{\mathsf{X}} \lambda(\mathrm{d}x) P(x, B). \tag{1}$$

When λ is a probability measure, the encouraged interpretation is that $X_{n+1} \mid X_n \sim P(X_n, \cdot)$ and $X_n \sim \lambda$, the product $\lambda(\mathrm{d}x)P(x, \cdot)$ is the joint distribution of (X_n, X_{n+1}) and λP is the marginal distribution of X_{n+1}.

Since Markov kernels act to the left on measures (1),

$$P^2(x, B) = \int_{\mathsf{X}} P(x, \mathrm{d}x_k) P(x_k, B).$$

Continuing in this fashion obtain for every $n \geq 2$

$$P^n(x, B) \int_{\mathsf{X}} P(x, \mathrm{d}x_k) P(x_k, \mathrm{d}x_{k+1}) \cdots P(x_{k+n-2}, B).$$

More generally, the so-called Chapman-Kolmogorov equations hold for $n \geq m \geq 0$

$$P^n(x, B) \int_{\mathsf{X}} P^m(x, \mathrm{d}y) P^{n-m}(y, B).$$

If $\lambda = \lambda P$, then λ is *invariant* for P. Notice that if λ is invariant for P and $X_n \sim \lambda$, then $X_{n+1} \sim \lambda$. That is, the marginal distribution does not depend upon n in which case the Markov chain is *stationary*.

Come back to these examples.

Example 2.5.

Example 2.6. Recall the Markov chain defined in Example (2.4)

One common way of establishing invariance of MCMC Markov chains is to verify a *detailed balance condition*; see Exercise 3.1. Detailed balance holds if

$$\lambda(\mathrm{d}x) P(x, \mathrm{d}y) = \lambda(\mathrm{d}y) P(y, \mathrm{d}x). \tag{2}$$

When λ is a probability measure, one interpretation is that the joint distribution of (X_k, X_{k+1}) is the same as the distribution of (X_{k+1}, X_k) so that this is also often called the *reversibility condition*. Another name often encountered is that P is λ-*symmetric*.

2.1 Stability

MCMC applications typically are constructed so that a specific probability distribution F is invariant. However, in applications where MCMC is required it is typically difficult to simulate from the invariant distribution. The most that can be hoped for is that the simulation will eventually produce a representative sample from F. This long-run behavior is in not guaranteed without additional assumptions. The following simple examples illustrate that the problems can arise due to the either the way the kernel is specified or the properties of the state space X.

Example 2.7. Suppose F lives on $\{1,2\}$ with $F(1) = 1 - F(2) = 1/4$ and

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

Since the Markov chain moves deterministically between the two states, it will over represent state 1 and underrepresent state 2 no matter how many iterations there are.

Example 2.8. Suppose F lives on $\{1,2,3\}$ with $F(1) = F(2) = F(3) = 1/3$ and

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Thus starting at state $\{3\}$ the chain remains there forever while starting from $\{1,2\}$ the chain will never visit $\{3\}$. Thus the chain cannot represent F in the long run.

Example 2.9. Suppose for $i = 1, 2$, f_i is a pf on $\mathsf{X}_i \subseteq \mathbb{R}$ and g_i is a pf on $\mathsf{Y}_i \subseteq \mathbb{R}$. Set

$$f(x,y) = \frac{1}{2} f_1(x) g_1(y) + \frac{1}{2} f_2(x) g_2(y).$$

Then

$$f_{X|Y}(x \mid y) = \frac{f_1(x) g_1(y) + f_2(x) g_2(y)}{g_1(y) + g_2(y)}$$

and

$$f_{Y|X}(y \mid x) = \frac{f_1(x) g_1(y) + f_2(x) g_2(y)}{f_1(y) + f_2(y)}$$

and the Gibbs sampler MTD is

$$k(x', y' \mid x, y) = f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x').$$

When $\mathsf{X}_i = \mathsf{Y}_i = \mathbb{R}$ this Gibbs sampler will produce a representative sample eventually. However, complications may arise if the spaces are constrained.

Suppose $\mathsf{X}_1 = \mathsf{X}_2 = (0,1)$ and that $f_1 = f_2$ is the Uniform density. Let $Y_1 = (0,1)$ and $Y_2 = (2,3)$ and g_1 and g_2 be Uniform densities. Easy calculation yields that

$$f_{X|Y}(x \mid y) = I(0 < x < 1) \left[I(0 < y < 1) + I(2 < y < 3) \right]$$

and

$$f_{Y|X}(y \mid x) = \frac{1}{2} I(0 < x < 1) \left[I(0 < y < 1) + I(2 < y < 3) \right]$$

and that, no matter which square, $\mathsf{X}_1 \times \mathsf{Y}_1$ or $\mathsf{X}_2 \times \mathsf{Y}_2$, the current state is in there is a positive probability of the next state being in either square. This Gibbs sampler will eventually produce a representative sample.

Now consider the setting with $\mathsf{X}_1 = \mathsf{Y}_1 = (0, 1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2, 3)$ so that

$$f_{X|Y}(x \mid y) = \frac{I(0 < x < 1) I(0 < y < 1) + I(2 < x < 3) I(2 < y < 3)}{I(0 < y < 1) + I(2 < y < 3)}$$

and

$$f_{Y|X}(y \mid x) = \frac{I(0 < x < 1) I(0 < y < 1) + I(2 < x < 3) I(2 < y < 3)}{I(0 < x < 1) + I(2 < x < 3)}.$$

If $y \in (0, 1)$, then $f_{X|Y}(x \mid y) = I(0 < x < 1)$ and, similarly, if $X \in (0, 1)$, then $f_{Y|X}(y \mid x) = I(0 < y < 1)$. Thus if the current state is in the square $\mathsf{X}_1 \times \mathsf{Y}_1$, then the next step will be in $\mathsf{X}_1 \times \mathsf{Y}_1$. That is, there is no chance for the chain to visit $\mathsf{X}_2 \times \mathsf{Y}_2$. This Gibbs sampler will not produce a representative sample from the target distribution.

The above examples demonstrate that one way problems arise is when the Markov chain cannot access all of the space eventually and hence properties which avoid this are required.

Let ϕ be a non-trivial positive measure on $\mathcal{B}$. Then $A \in \mathcal{B}$ is *ϕ-communicating* if for all $B \subseteq A$ such that $\phi(B) > 0$ and for all $x \in A$, there exists n such that $P^n(x, B) > 0$. This is a weak property that does not alone ensure desirable long run properties. Consider the Gibbs sampler from Example 2.9 with $\mathsf{X}_1 = \mathsf{Y}_1 = (0, 1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2, 3)$. If ϕ denotes Lebesgue measure, then this Markov chain is ϕ-communicating but does not have desirable long run properties.

The Markov kernel, P, is *ϕ-irreducible* if for all $x \in \mathsf{X}$ and for all $A \in \mathcal{B}$ such that $\phi(A) > 0$ there exists n such that $P^n(x, A) > 0$. This is a key property in many MCMC settings, but is not enough to ensure desirable long run properties; consider the Markov chain in Example 2.7 which is irreducible.

There is some arbitrariness in the definition of ϕ-irreducibility, but if for some ϕ, P is ϕ-irreducible, then there exists a maximal irreducibility measure ψ (meaning that $\psi(A) = 0$ implies

$\phi(A) = 0$ for all irreducibility measures ϕ).

Proposition 2.1. *If λ is an invariant measure for the Markov kernel P and, for some ϕ, P is ϕ-irreducible, then P is λ-irreducible.*

Corollary 2.1. *If P is λ-symmetric and ϕ-irreducible, then P is λ-irreducible.*

Often, inspection of the kernel P and the underlying space X is enough to establish ϕ-irreducibility. For example, many kernels obviously satisfy a positivity condition so that $P(x, A) > 0$ for all $x \in \mathsf{X}$ and $A \in \mathcal{B}$ with $\phi(A) > 0$. The following is a special case of a more general result [1, Theorem 3], but is often useful.

Proposition 2.2. *Suppose X is a connected, separable metric space. If every nonempty, open set A satisfies $\phi(A) > 0$ and every point has a ϕ-communicating neighborhood, the Markov kernel is ϕ-irreducible.*

3 Constructing MCMC Algorithms

3.1 Metropolis-Hastings

Give credit to R?

The fundamental MCMC algorithm is Metropolis-Hastings [3, 5]. It serves as a building block for many, if not most, applications of MCMC. The following will be generalized later, but, for now, suppose $\mathsf{X} \subseteq \mathbb{R}^d$ and that $q(x, y)$ is a proposal conditional density that is easy to sample. Define the *Hastings ratio*

$$r(x, y) = \frac{f(y)q(y, x)}{f(x)q(x, y)}.$$

Algorithm 1 Metropolis-Hastings

1: *Input:* Current value $X_n = x$.

2: Draw $Y \sim Q(x, \cdot)$

3: Draw $U \sim \text{Uniform}(0, 1)$

4: If $u \leq r(x, y) \wedge 1$, accept y and set $X_{n+1} = y$, else set $X_{n+1} = x$.

Algorithm 2 is the formal definition, but is not at all how the algorithm should be implemented in practice. The following implementation will help avoid overflow issues.

Algorithm 2 Metropolis-Hastings Implementation

1: *Input:* Current value $X_n = x$.

2: Draw $Y \sim Q(x, \cdot)$

3: Draw $U \sim \text{Uniform}(0, 1)$

4: If $\log r(x, y) \geq 0$ set $X_{n+1} = y$, else if $u \leq r(x, y)$, set $X_{n+1} = y$, else set $X_{n+1} = x$.

Metropolis-Hastings defines a Markov kernel. Specifically, if $\alpha(x, y) = 1 \wedge r(x, y)$, then

$$P(x, \mathrm{d}y) = Q(x, \mathrm{d}y) + \delta_x(\mathrm{d}y) \int [1 - \alpha(x, u)]\, Q(x, \mathrm{d}u). \tag{3}$$

Proposition 3.1. *The Metropolis-Hastings kernel* (3) *is F-symmetric.*

Proof. It suffices to consider $x \neq y$. Then

$$\begin{aligned} F(\mathrm{d}x)P(x, \mathrm{d}y) &= f(x)q(x, y)\,[1 \wedge r(x, y)]\,\mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= [f(x)q(x, y) \wedge f(y)q(y, x)]\,\mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= f(y)q(y, x)\,[1 \wedge r(y, x)]\,\mu(\mathrm{d}x)\mu(\mathrm{d}y) \\ &= F(\mathrm{d}y)P(y, \mathrm{d}x). \end{aligned}$$

□

The proof of the following result is easy and is left as an exercise.

Proposition 3.2. *If $q(x, y) > 0$ for all $x, y \in \mathsf{X}$, then the Metropolis-Hastings kernel P is F-irreducible.*

Example 3.1. Suppose continuous density f has support $\mathsf{X} = \mathbb{R}$ and consider three settings. Firstly, if the proposal distribution is the Student's t distribution with $d \geq 1$ degrees of freedom, then by Proposition 3.2 the resulting Metropolis-Hastings kernel is F-irreducible. Secondly, if the proposal distribution is Uniform$(0, 1)$, then the resulting Metropolis-Hastings kernel is obviously reducible since no value outside of $(0, 1)$ will be proposed. Thirdly, if $X_n = x$ and the proposal is Uniform$(x - 1, x + 1)$, then Proposition 2.2 can be used to establish F-irreducibility. The first two conditions of the proposition are easy consequences of the properties of the real line. Let $0 < \delta < 1$ so that $N_{x,\delta} = (x - \delta, x + \delta)$ is a neighborhood of $\{x\}$ and

$$P(x, N_{x,\delta}) = \int_{x-\delta}^{x+\delta} \frac{1}{2} \left(1 \wedge e^{-0.5(x^2+y^2)}\right) \mathrm{d}y > 0$$

from which it is easy to see that $N_{x,\delta}$ is an F-communicating neighborhood.

3.1.1 Classical Strategies for Choosing a Proposal

An *Independence Metropolis-Hastings* results if the proposal is independent of the previous state. That is, $Y \sim Q$. For example, suppose the proposal is a d-dimensional normal distribution with fixed mean m and covariance C, so that $X'_t \sim N_d(m, C)$.

Independence MH is one of the simplest and best understood MH Markov chains. However, it is unlikely to be effective in many settings, as will be shown later.

Another well-studied version is the *Symmetric Metropolis-Hastings*, which results when the proposal distribution which is symmetric about the current state X_{t-1}, so $q(x, y) = q(y, x)$. For example, suppose the proposal is a d-dimensional normal distribution centered at the previous iteration with covariance matrix hI_d so that $X'_t \mid X_{t-1} \sim N_d(X_{t-1}, hI_d)$. The scaling parameter $h > 0$ is user-specified.

If the proposal of a symmetric Metropolis-Hastings also satisfies $q(x, y) = q(\|x - y\|)$, then a *Random-Walk Metropolis-Hastings* sampler results. Note that the proposal, $X'_t \mid X_{t-1} \sim N_d(X_{t-1}, hI_d)$,

in the previous example results in random walk Metropolis-Hastings. Another example would be a Uniform$(x-h, x+h)$, $h>0$, proposal.

Suppose the proposal distribution is a d-dimensional normal distribution where the mean takes a gradient descent step and covariance hI_d so that

$$X_t' \mid X_{t-1} \sim N_d(X_{t-1} + (h/2)\nabla \log(f(X_{t-1})), hI_d).$$

The gradient does not require the normalization constant of f. The motivation for this construction comes from discretized Langevin dynamics [7] and, consequently, is known as Metropolis-Adjusted Langevin Algorithm (MALA).

These examples barely scratch the surface of Metropolis-Hastings variations. For example, much recent research has gone into Hamiltonian Monte Carlo which uses a proposal based on discretized Hamiltonian dynamics [6]. Extensions to Riemanian manifold MALA and Hamiltonian variants also exist [2]. Some of these will be encountered later.

3.2 Combining Markov Kernels

Markov kernels may be combined in order to create more effective algorithms. There are two basic ways of doing so that will be exploited below. For now, suppose $P_1, \ldots, P_d$ are Markov kernels such that $FP_i = F$ for $i = 1, \ldots, d$. The *composition* kernel is defined by

$$P_C(x, \cdot) = (P_1 \cdots P_d)(x, \cdot)$$

and corresponds to cycling through the kernels in a specified order. If each $r_i > 0$ such that $r_1 + \cdots + r_d = 1$, then the *mixing* kernel is defined by

$$P_m(x\cdot) = r_1 P_1(x, \cdot) + \cdots + r_d P_d(x, \cdot)$$

and corresponds to updating via the kernel selected by the mixing probabilities r_i. It is an easy exercise to verify that $FP_c = F$ and $FP_m = F$.

3.3 Component-wise Updates

It is rare that Metropolis-Hastings can be used without modification in practically relevant settings. The distribution F is often too complicated or too high-dimensional for a block Metropolis-Hastings update to be effective. It is natural to seek to work with smaller problems.

Since this is introductory, the focus will be on the setting where $f(x, y)$ is a density function on $\mathsf{X} \times \mathsf{Y}$, that is, the so-called *two-variable* setting. The extension to the setting with more than two variables is straightforward through the usual properties of joint probability functions.

3.3.1 Linchpin Variables

Let $f_{X|Y}$ be the pf of the conditional distribution of X given Y. Let f_Y be the pf of the marginal distribution of Y. If sampling from $f_{X|Y}$ is straightforward, recall that Y is a linchpin variable since

$$f(x, y) = f_{X|Y}(x|y)\, f_Y(y). \tag{4}$$

Recall that exact samples can be obtained by first simulating $Y \sim f_Y$ followed by $X \sim f_{X|Y}$. However, F_Y may be complicated or difficult to sample directly so it is natural to turn to MCMC. Let $P_Y(y, \cdot)$ be a Markov kernel such that F_Y is invariant. Then the linchpin variable sampler is specified in Algorithm 3.

Algorithm 3 Linchpin variable sampler

1: *Input:* Current value (X_j, Y_j)

2: Draw $Y_{j+1} \sim P_Y(Y_j, \cdot)$.

3: Draw $X_{j+1} \sim F_{X|Y}(\cdot \mid Y_{j+1})$.

4: Set $j = j + 1$

The resulting Markov kernel is given by

$$P((x, y), A) = \int_A F_{X|Y}(dx' \mid y') P_Y(y, dy') \tag{5}$$

That F is invariant is left as an exercise.

Linchpin variable samplers have been employed in a variety of scenarios. Their success is typically due to either (1) superior mixing in the lower-dimensional space, (2) de-correlation of components via the linchpin variables, or (3) lower post-processing costs.

3.3.2 Conditional Samplers

2.4

Exercises

Exercise 3.1. Prove that if Equation 2 holds, then λ is invariant for P.

Exercise 3.2. Consider the Gibbs samplers in Example 2.9. establish that the Gibbs sampler on $\mathsf{X}_1 = \mathsf{Y}_1 = (0,1)$ and $\mathsf{X}_2 = \mathsf{Y}_2 = (2,3)$ is not irreducible while the one on $\mathsf{X}_1 = \mathsf{Y}_1 = \mathsf{X}_2 = (0,1)$ and $\mathsf{Y}_2 = (2,3)$ is irreducible.

Exercise 3.3. Establish Proposition 2.1.

Exercise 3.4. Establish Proposition 2.2.

Exercise 3.5. Prove that Metropolis-Hastings has Markov kernel given in (3). Hint: Note that if $A \in \mathcal{B}$ and $j \geq 1$ then

$$\Pr(X_{j+1} \in A \mid X_j) = \Pr(X_{j+1} \in A, U \leq \alpha(x,y) \mid X_j) + \Pr(X_{j+1} \in A, U > \alpha(x,y) \mid X_j).$$

Exercise 3.6. Establish that F is invariant for the linchpin variable kernel specified in (5).

Exercise 3.7. Let $Y \in \mathbb{R}^n$, $\beta \in \mathbb{R}^p$, $u \in \mathbb{R}^k$, X be a known $n \times p$ full column rank design matrix, and Z a known $n \times k$ full column rank matrix. Also assume that $\max\{p,k\} < n$. Then a Bayesian linear model is given by the following hierarchy

$$\begin{aligned} Y|\beta,u,\lambda_E,\lambda_R &\sim \mathrm{N}_n\left(X\beta + Zu, \lambda_E^{-1} I_n\right) \\ g(\beta) &\propto 1 \\ u|\lambda_E,\lambda_R &\sim \mathrm{N}_k\left(0, \lambda_R^{-1} I_k\right) \\ \lambda_E &\sim \mathrm{Gamma}\left(e_1,e_2\right) \\ \lambda_R &\sim \mathrm{Gamma}\left(r_1,r_2\right) . \end{aligned} \tag{6}$$

Assume that $e_1, e_2, r_1, r_2 > 0$ are known hyper-parameters. This hierarchy results in a proper posterior [8] . Let $\xi = (\beta^T, u^T)^T$, $\lambda = (\lambda_E, \lambda_R)^T$ and let y denote all of the data. Then the posterior density satisfies

$$f(\beta, u, \lambda_E, \lambda_R \mid y) = f(\xi, \lambda \mid y) = f_{\xi|\lambda}(\xi \mid \lambda, y)\, f_\lambda(\lambda \mid y)\,. \tag{7}$$

Appendix

References

[1] Geyer, C. J. (2014). Stat 8501 lecture notes.

[2] Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 73:123–214.

[3] Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1).

[4] Huber, M. L. (2016). *Perfect Simulation*, volume 148. CRC Press.

[5] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21.

[6] Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In Brooks, S., Gelman, A., Meng, X.-L., and Jones, G. L., editors, *Handbook of Markov chain Monte Carlo*. Chapman & Hall, Boca Raton.

[7] Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. E., editors, *Markov chain Monte Carlo in practice*, pages 45–57. Chapman & Hall, Boca Raton.

[8] Sun, D., Tsutakawa, R. K., and He, Z. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica*, pages 77–95.