DOI: 10.1177/00222437241255307



JOURNAL of MARKETING RESEARCH

# Revisiting Scalable Targeted Marketing with Distributed Markov Chain Monte Carlo

| | |
|---|---|
| Journal: | *Journal of Marketing Research* |
| Manuscript ID | JMR-22-0154.R5 |
| Manuscript Type: | Revised Submission |
| Topics and Methods: | Bayesian estimation < Theoretical Foundation, Measurement and inference < Theoretical Foundation, Target marketing, Big data |
| | |

# Revisiting Scalable Targeted Marketing with Distributed Markov Chain Monte Carlo

Michael Braun
Cox School of Business
Southern Methodist University
braunm@smu.edu

April 29, 2024

**Abstract**

Bumbaca, Misra, and Rossi (2020) propose a parallelizable algorithm for estimating a large number of customer-level parameters in a Bayesian hierarchical model. However, the algorithm follows from a mathematical error in the derivation of the target posterior density, which calls into question the theoretical support for the algorithm sampling from the specified model. Adapting the algorithm to be consistent with the corrected math nullifies the claimed benefits in scalability and efficiency. Notwithstanding that error, unbiasedness requires the number of customers to be asymptotic per computational node, which is more restrictive than being asymptotic in the size of the dataset as a whole. The more the algorithm is parallelized, the greater the bias. Potential adopters should be aware that the algorithm does not sample from the exact posterior distribution, and that its ability to take advantage of distributed computing infrastructure is limited.

**Keywords:** parallel Bayesian estimation, Bayesian hierarchical models, target marketing, big data,hierarchical models

Tailoring marketing strategies to specific consumers can often improve their effectiveness (e.g., Ascarza 2018; Danaher 2023). Doing so requires estimating each customer's unobserved tendencies and propensities to respond to marketing interventions, which can be methodologically difficult. One approach is to conduct inference on customer-level parameters in a hierarchical Bayesian model (Rossi et al. 1993). But standard Bayesian estimation algorithms like Markov chain Monte Carlo (MCMC) become computationally intractable when the number of customers in a database is very large.

Some researchers have proposed scalable MCMC variants that employ parallel computing, where the model is estimated from separate "shards" (partitions) of the dataset on distributed computational nodes. MCMC draws are then post-processed to allow for approximate posterior inference (e.g., Neiswanger et al. 2014; Scott et al. 2016; Vyner et al. 2023). The post-processing stage is necessary because combining samples generated from separate nodes is not equivalent to sampling from a posterior that conditions on the aggregate database. Potential adopters of these algorithms must consider the tradeoffs among computational efficiency, the scope of applications for which the various methods are valid, and the magnitude of the biases that these methods may introduce in practice.

Bumbaca, Misra, and Rossi (2020, henceforth BMR) introduce an algorithm in that same class of "divide and conquer" parallel MCMC methods, with a specific focus on estimating customer-level parameters. However, their paper contains a material mathematical error where they treat a proportional relationship as an equality, and make an impermissible substitution in their derivation of the target posterior density. This calls into question BMR's claim of asymptotic exactness. This paper identifies and corrects the error, and shows that the published version of the algorithm is

inconsistent with the corrected math. Adapting the algorithm to be consistent with the corrected math reduces its scalability and efficiency.

Furthermore, the absence of a post-processing step means that the distributed nature of the BMR algorithm induces "parallelization bias." In practice, this is a problem because the algorithm's asymptotic unbiasedness assumes a sufficiently large number of customers *per shard*, which is a stronger (and less typical) assumption than a large number of customers in the aggregate dataset. Because real-world databases contain data for a finite number of customers, the more a practitioner tries to parallelize the algorithm (which is the primary motivation of BMR's paper), the more biased the results will be. While the BMR algorithm could produce a sample that may be close to the exact posterior, potential adopters of the BMR algorithm should be aware of these theoretical inconsistencies and practical limitations.



## THE BMR ALGORITHM AND THE MATHEMATICAL ERROR

The objective of BMR's algorithm is to estimate customer-level parameters $\beta_i$ in a Bayesian hierarchical model when $N$, the number of customers in the dataset, is large. It does this by partitioning the dataset Y into $S$ shards of size $N_s = N/S$ (denoted as $\mathrm{Y}_{1:S} = \{\mathrm{Y}_1, \ldots, \mathrm{Y}_S\}$) and parceling computation across $S$ computational nodes. In Stage 1 of the algorithm, $R$ samples of a population-level parameter $\theta$ are generated from each of $p(\theta \mid \mathrm{Y}_1, \tau), \ldots, p(\theta \mid \mathrm{Y}_S, \tau)$ in parallel. These samples are gathered from the various nodes and combined into $\{\theta_s^r\}$, a collection of $R \times S$ draws. Identical copies of $\{\theta_s^r\}$ are redistributed to the nodes in preparation for Stage 2, which involves iteratively sampling $\beta_i$ from $p(\beta_i \mid \theta_s^r) \propto p(y_i \mid \beta_i) p(\beta_i \mid \theta_s^r)$. BMR claim that samples from $p(\beta_i \mid \theta_s^r)$ constitute an asymptotically unbiased estimate of the target posterior

distribution $p(\beta_i \mid \mathrm{Y}, \tau)$. Web Appendix A summarizes the BMR algorithm and the hierarchical model it is designed to estimate.

BMR's error is in their derivation of $p(\beta_i \mid \mathrm{Y}, \tau)$. To begin, consider their Eqs. 5 and 7:

(BMR-5) $$p(\beta_i \mid y_i, \theta) \propto p(\beta_i \mid \theta) p(y_i \mid \beta_i)$$

(BMR-7) $$p(\beta_i \mid \mathrm{Y}, \tau) = \int p(\beta_i \mid y_i, \theta) p(\theta \mid \mathrm{Y}, \tau) \mathrm{d}\theta.$$

BMR substitute the $p(\beta_i \mid y_i, \theta)$ term in the integrand of BMR-7 with BMR-5, resulting in

(BMR-8) $$p(\beta_i \mid \mathrm{Y}, \tau) \propto \int p(\beta_i \mid \theta) p(y_i \mid \beta_i) p(\theta \mid \mathrm{Y}, \tau) \mathrm{d}\theta$$

(BMR-13) $$= \frac{p_{\theta \mid \mathrm{Y}, \tau}(\beta_i) p(y_i \mid \beta_i)}{p(y_i)},$$

where $p_{\theta \mid \mathrm{Y}, \tau}(\beta_i) = \int p(\beta_i \mid \theta) p(\theta \mid \mathrm{Y}, \tau) \mathrm{d}\theta$. But this step is incorrect because BMR-5 is a proportional relationship ($\propto$), not an equality ($=$). Instead, by Bayes' Theorem,

(1) $$p(\beta_i \mid y_i, \theta) = \frac{p(y_i \mid \beta_i, \theta) p(\beta_i \mid \theta)}{p(y_i \mid \theta)}.$$

The $p(y_i \mid \theta)$ term in the denominator of Eq. 1 is consequential because it depends on $\theta$, the variable of integration in BMR-7, and cannot be factored out of the integrand. Web Appendix B shows that substituting Eq. 1 into BMR-7 corrects the target posterior density in BMR-13.

(2) $$p(\beta_i \mid \mathrm{Y}, \tau) = \frac{p_{\theta \mid \mathrm{Y}_{-i}, \tau}(\beta_i) p(y_i \mid \beta_i)}{p(y_i \mid \mathrm{Y}_{-i})}.$$

In Eq. 2, $\mathrm{Y}_{-i}$ is data for all customers excluding $i$, and $p_{\theta \mid \mathrm{Y}_{-i}, \tau}(\beta_i) = \int p(\beta_i \mid \theta) p(\theta \mid \mathrm{Y}_{-i}, \tau) \mathrm{d}\theta$.

As published, the BMR algorithm is inconsistent with the corrected math, and appears to be scalable only because it samples from the incorrect BMR-13. Using all $S$ nodes to generate a single $\{\theta_s^r\}$ in Stage 1 was justified only because of how BMR derived $p_{\theta|\mathrm{Y},\tau}(\beta_i)$: integrating $p(\beta_i \mid \theta)$ over the same $p(\theta \mid \mathrm{Y}, \tau)$ for all customers. BMR's reports of computational efficiency depend on all customers' draws from $p(\beta_i \mid \theta_s^r)$ being conditional on elements of a common $\{\theta_s^r\}$.

But the $p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$ term in the corrected Eq. 2 is an integral over $p(\theta \mid \mathrm{Y}_{-i}, \tau)$, which is different for each customer. Fixing the algorithm to be consistent with the corrected math would involve conditioning Stage 2 samples on a distinct $\{\theta_s^r\}_i$ for each customer. The efficiency of running only $S$ parallel MCMC instances in Stage 1 would be lost, especially since $N/S$ must be sufficiently large for the algorithm to be asymptotically unbiased (see below). Also, Stage 2 would incur additional communication overhead because each customer's $\theta_s^r$ draws would come from a different $\{\theta_s^r\}_i$. Gains in computational efficiency would be much more modest. See Web Appendix C.

Web Appendix D explains how BMR algorithm systematically biases the estimate of $p(\beta_i \mid \mathrm{Y}, \tau)$, relative to the target posterior density in the model. The distinction between BMR-13 and Eq. 2 is in whether the prior on $\beta_i$ conditions on Y, which includes on $y_i$, or on $\mathrm{Y}_{-i}$, which does not. The practical effect of this error may be small, but remains to be proven. What is true is that even after conditioning on $\theta_s^r$, the BMR algorithm does not sample from the model that is specified in their paper. The corrected derivation calls into question the theoretical support for BMR's claims that their algorithm samples from the exact posterior.

Further, BMR's salient theoretical contribution is that their approximations to $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ and

$p(\beta_i \mid \mathrm{Y}, \tau)$ converge in distribution to their true densities (Theorems 5 and 6). The error means that many of their theorems need to be either restated or reproven. See Web Appendix E.

## PARALLELIZATION BIAS

The key to BMR's faster computational speed comes from dividing the data into shards so that computation can be performed in parallel. BMR's claimed exactness also depends on each shard including a very large number of customers (approaching infinity). One concern for potential adopters is that the algorithm induces additional "parallelization bias" into estimates of $p(\beta_i \mid \mathrm{Y}, \theta)$. This is likely of more concern to potential adopters than the mathematical error.

The problem comes from how in Stage 1, $\{\theta_s^r\}$ is not actually sampled from $p(\theta \mid \mathrm{Y}, \tau)$, but rather from a mixture of posterior distributions that are conditional on each shard:

$$p^*(\theta \mid \mathrm{Y}_{1:S}, \tau) = \frac{1}{S}\sum_{s=1}^{S} p(\theta \mid \mathrm{Y}_s, \tau) \tag{3}$$

The algorithm is asymptotically unbiased as $p^*(\theta \mid \mathrm{Y}_{1:S}, \tau)$ converges to $p(\theta \mid \mathrm{Y}, \tau)$, which requires the number of customers *per shard* ($N/S$) to be sufficiently large (see BMR's Theorem 4.1.2). This is more restrictive than the typical asymptotic assumption on $N$ ($N/S \to \infty$ implies $N \to \infty$ trivially), and $N/S$ gets smaller the more the practitioner parallelizes the algorithm (increasing $S$). Intuitively, BMR achieve asymptotic unbiasedness because when each shard has an infinite number of customers, each summation term in Eq. 3 conditions on the same infinitely-sized dataset. Web Appendix F shows that for *finite* $N$ and $S > 1$, $p^*(\theta \mid \mathrm{Y}_{1:S}, \tau)$ has higher variance than $p(\theta \mid \mathrm{Y}, \tau)$, which attenuates Bayesian shrinkage. In BMR-13, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ acts as an infor-

mative prior on $\beta_i$ (as does $p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$ in Eq. 2). Parallelization bias weakens that information, allowing $y_i$ to pull the estimate of $p(\beta_i \mid \mathrm{Y}, \tau)$ away from the target posterior in the model. That is, the estimates of $\beta_i \mid \mathrm{Y}$ potentially overfit customer $i$'s data when the shards are small.

Because the size of each shard decreases as the number of computing nodes increases, this bias will increase as the practitioner takes greater advantage of parallel computing. BMR address this problem with a heuristic that computes $S_{\max}$, which is an upper bound on $S$ that is a function of a *practitioner-specified* bias tolerance, $\varepsilon^2_{\max}$. This approach has some limitations. First, by constraining the number of shards that can be used, using the BMR algorithm as specified fails to fully take advantage of distributed computing infrastructure. BMR's heuristic asks practitioners to process larger shards on fewer nodes than are available, and to leave the remaining nodes idle. And if $S_{\max}$ is very small, the time to run Stage 1 may still be prohibitively large.

Second, it is not clear how a practitioner should select or interpret the bias tolerance $\varepsilon^2_{\max}$, which is a maximum allowable squared difference between densities. BMR's Theorem 9 requires that $N'/S'$ in the *pilot run* be large enough for asymptotics to hold, which is even more restrictive than $N/S$ being large. It is not clear how confident researchers can be that $S_{\max}$ bounds the bias to be within $\varepsilon^2_{\max}$. A tolerance based on moments would be more useful.

Finally, researchers should consider whether their data are a good fit for the BMR algorithm. In BMR's simulation study, customers are generated with either 5, 15 or 45 observations. These are situations in which the customer's data will overwhelm $p^*_{\theta|\mathrm{Y},\tau}(\beta_i)$ anyway, so attenuated shrinkage from parallelization bias is less likely to come into play. The algorithm should predict well for those customers. In BMR's empirical application, the algorithm's predictive ability is tested only among customers with at least 16 observations. Yet, BMR's Fig. 4 shows many

instances where the BMR posterior sample is quite different from the Gibbs sample ("a source of truth"). If the database includes customers with only a few observations, for whom the posterior depends more on the prior, or if both the data and prior on individual level parameters are highly informative, parallelization bias is likely to have a more detrimental effect on the results.

## DISCUSSION

In their "What is Novel" section, BMR write: "The shard-splitting idea in each of the two stages is not new and common to both Scott et al. (2016) and Neiswanger et al. (2014). The novelties of the first stage are (1) constructing the posterior predictive density and (2) drawing from the posterior predictive density in parallel to reduce communication overhead between stages" (Bumbaca et al. 2020, p.1004). *These are the two aspects of their paper that are directly affected by their error*. Even if a practitioner were to look past that error, estimating population-level parameters conditional on shards of data induces bias in finite samples that practitioners cannot ignore, particularly when making maximal use of available parallel computing resources. While BMR do not promise unbiased distributions of population-level parameters as a "deliverable," there is no avoiding the fact that Stage 1 generates biased population-level samples, and nothing in Stage 2 adjusts for that.

Despite the concerns about the BMR algorithm that are described in this paper, some practitioners may still find the trade-offs between real-time efficiency gains and biased estimation to be favorable. Potential adopters should consider whether these biases are likely to be small enough for their own practical applications.

## REFERENCES

Ascarza, Eva (2018). "Retention Futility: Targeting High-Risk Customers Might Be Ineffective." *Journal of Marketing Research*, 60(1):80–98.

Bumbaca, Federico, Sanjog Misra, and Peter E. Rossi (2020). "Scalable Target Marketing: Distributed Markov Chain Monte Carlo for Bayesian Hierarchical Models." *Journal of Marketing Research*, 57(6):999–1018.

Danaher, Peter J. (2023). "Optimal Microtargeting of Advertising." *Journal of Marketing Research*, 60(3):564–584.

Neiswanger, Willie, Chong Wang, and Eric Xing (2014). "Asymptotically Exact, Embarrassingly Parallel MCMC." *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence*. Quebec City. ARXIV:1311.4780v2.

Rossi, Peter E. and Greg M. Allenby (1993). "A Bayesian Approach to Estimating Household Parameters." *Journal of Marketing Research*, 30(2):171–182.

Scott, Steven L., Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch (2016). "Bayes and Big Data: The Consensus Monte Carlo Algorithm." *International Journal of Management Science and Engineering Management*, 11(2):78–88.

Vyner, Callum, Christopher Nemeth, and Chris Sherlock (2023). "SwISS: A Scalable Markov chain Monte Carlo Divide-and-Conquer Strategy." *Stat*, 12(1):1–11.

# Web Appendices

## Revisiting Scalable Targeted Marketing with Distributed Markov Chain Monte Carlo

Michael Braun
Cox School of Business
Southern Methodist University
braunm@smu.edu



**Contents**

## WEB APPENDIX A: CONTEXT, BACKGROUND, AND OBJECTIVES

To set the scene, Fig. A1 illustrates a hierarchical model with conditional independence across $N$ customers (BMR's Eqs. 1 to 3). $\mathrm{Y} = \{y_1, \ldots, y_N\}$ represents the entire dataset, where each $y_i$ is a (possibly multivariate) observed outcome for customer $i$. The customer-level latent parameters $\beta = \{\beta_1, \ldots, \beta_N\}$ are associated only through their common dependence on a population parameter $\theta$, which affects customer data $y_i$ only through its corresponding $\beta_i$. $p(y_i \mid \beta_i)$ is the likelihood of each customer's data, $p(\beta_i \mid \theta)$ is the customer-level prior, and $p(\theta \mid \tau)$ is the population-level hyperprior.

BMR's managerial task involves sampling from the marginal posterior density $p(\beta_i \mid \mathrm{Y})$ for individual customers. BMR's proposed alternative is to sample customer-level parameters from an approximation to

(BMR-13) $$p(\beta_i \mid \mathrm{Y}) = \frac{p_{\theta|\mathrm{Y},\tau}(\beta_i)p(y_i \mid \beta_i)}{p(y_i)},$$

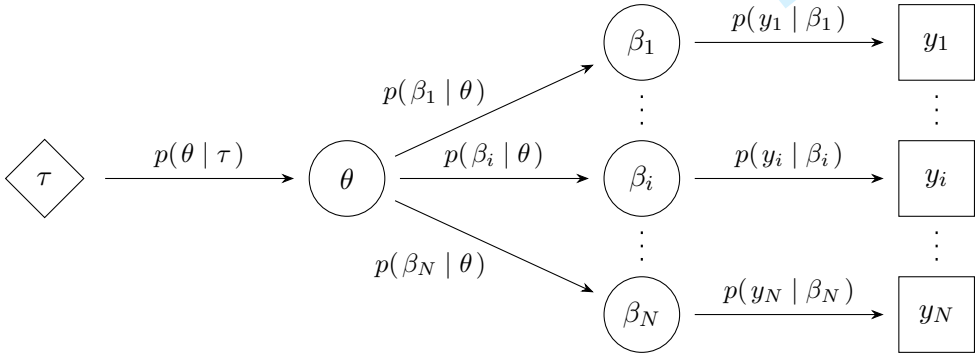
where

(A1) $$p_{\theta|\mathrm{Y},\tau}(\beta_i) = \int p(\beta_i \mid \theta)p(\theta \mid \mathrm{Y}, \tau)\mathrm{d}\theta$$

Here, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ acts as an informative prior distribution on $\beta_i$ that conditions on the entire dataset by marginalizing $p(\beta_i \mid \theta)$ over $p(\theta \mid \mathrm{Y}, \tau)$.

To provide some intuition, $p(\beta_i \mid \theta)$ is not only a prior on a single customer's $\beta_i$, but it also describes the heterogeneous mixture of $\beta_i$ across the population. By averaging $p(\beta_i \mid \theta)$ over $p(\theta \mid \mathrm{Y}, \tau)$, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ takes two sources of uncertainty or variation into account: variation in $\beta_i$ for a given value of $\theta$, and the

Figure A1: A Bayesian Hierarchical Model with Conditional Independence




Note: Observed data: $\mathrm{Y} = \{y_1, \ldots, y_N\}$. Heterogeneous parameters: $\beta = \{\beta_1, \ldots, \beta_N\}$. Population parameter: $\theta$. Hyperprior parameter: $\tau$.

informed uncertainty in the value of $\theta$ itself.

In Stage 1 of BMR's algorithm, customers are randomly assigned into $S$ cohorts, with $N_S = N/S$ customers in each cohort, and with each cohort assigned to a computational node. $\mathrm{Y}_s$ is the "shard" of data for customers in cohort $s$ in the partition $\mathrm{Y} = \{\mathrm{Y}_1, \ldots, \mathrm{Y}_S\}$. Using a standard MCMC algorithm, $R$ samples are generated from each of $p(\theta \mid \mathrm{Y}_1, \tau), \ldots, p(\theta \mid \mathrm{Y}_S, \tau)$ in parallel, across $S$ separate nodes, conditioning on each node's cohort's respective shard of data. These samples are gathered from the various nodes and combined into a pool of $R \times S$ draws, denoted as $\{\theta_s^r\}$. Identical copies of $\{\theta_s^r\}$ are redistributed to the nodes in preparation for Stage 2.

Stage 2 flows from how Eq. A1 is a posterior expected value of $p(\beta_i \mid \theta)$, and therefore can be approximated as a sample mean across $\{\theta_s^r\}$.

(BMR-20) $$p_{\theta|\mathrm{Y},\tau}(\beta_i) \approx \ddot{p}_{\theta|\mathrm{Y},\tau}(\beta_i) = \frac{1}{RS}\sum_{r,s} p(\beta_i \mid \theta_s^r).$$

The goal of Stage 2 is to sample each $\beta_i$ from an approximation to $p(\beta_i \mid \mathrm{Y})$, denoted as $\ddot{p}(\beta_i \mid \mathrm{Y})$. BMR do this by sampling uniformly from $\{\theta_s^r\}$, and then sampling $\beta_i \mid \theta_s^r$ from

(A2) $$p(\beta_i \mid \theta_s^r) \propto p(y_i \mid \beta_i) p(\beta_i \mid \theta_s^r)$$

Repeating these steps samples from the marginal distribution $\ddot{p}(\beta_i \mid \mathrm{Y})$ by numerically integrating

(BMR-21) $$\ddot{p}(\beta_i \mid \{\theta_s^r\}, \mathrm{Y}) \propto p(y_i \mid \beta_i) \ddot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$$

over the empirical posterior distribution of $\{\theta_s^r\}$, which is conditional on Y. In this stage the only data needed to sample $\beta_i \mid \mathrm{Y}$ for a single customer are $y_i$ and $\{\theta_s^r\}$. Because $\{\theta_s^r\}$ is the same for all customers, $\beta_i$ can be sampled from Eq. A2 in parallel.

## WEB APPENDIX B: DERIVING THE CORRECT POSTERIOR

### *Definitions and Axioms*

Define $\mathrm{Y} = \{y_1, \ldots, y_N\}$ and $\beta = \{\beta_1, \ldots, \beta_N\}$. Further, define $\mathrm{Y}_{-i}$ as all elements of Y excluding $y_i$ and $\beta_{-i}$ as all elements of $\beta$ excluding $\beta_i$. Hence, $\mathrm{Y} = \{y_i, \mathrm{Y}_{-i}\}$ and $\beta = \{\beta_i, \beta_{-i}\}$. A consequence of these definitions is that

$$p(\mathrm{Y} \mid \beta, \cdot) = f(y_i, \mathrm{Y}_{-i} \mid \beta_i, \beta_{-i}, \cdot) \tag{B1}$$

$$p(\beta \mid \cdot) = p(\beta_i, \beta_{-i} \mid \cdot) \tag{B2}$$

The following expressions formalize BMR's definition of conditional independence.

$$p(y_i \mid \beta_{-i}, \theta) = p(y_i \mid \theta) \tag{B3}$$

$$p(\mathrm{Y} \mid \beta, \theta) = p(\mathrm{Y} \mid \beta) \tag{B4}$$

$$p(\mathrm{Y} \mid \beta) = \prod_{i=1}^{N} p(y_i \mid \beta_i) \tag{B5}$$

$$p(\beta \mid \theta) = \prod_{i=1}^{N} p(\beta_i \mid \theta) \tag{B6}$$

It follows from Eqs. B3 to B6 that

$$p(\mathrm{Y} \mid \beta, \theta) = p(y_i \mid \beta_i, \theta) p(\mathrm{Y}_{-i} \mid \beta_{-i}, \theta) \tag{B7}$$

$$p(\mathrm{Y} \mid \beta) = p(y_i \mid \beta_i) p(\mathrm{Y}_{-i} \mid \beta_{-i}) \tag{B8}$$

$$p(\beta \mid \theta) = p(\beta_i \mid \theta) p(\beta_{-i} \mid \theta) \tag{B9}$$

$$p(y_i \mid \mathrm{Y}_{-i}, \beta_i, \theta) = p(y_i \mid \beta_i, \theta) \tag{B10}$$

### *Full Posterior*

Bayes Theorem gives the following posterior distributions.

$$p(\beta \mid \mathrm{Y}, \theta) = \frac{p(\mathrm{Y} \mid \beta, \theta) p(\beta \mid \theta)}{p(\mathrm{Y} \mid \theta)} \tag{B11}$$

$$p(\theta \mid \mathrm{Y}, \tau) = \frac{p(\mathrm{Y} \mid \theta) p(\theta \mid \tau)}{p(\mathrm{Y})} \tag{B12}$$

The full posterior distribution is

$$p(\beta, \theta \mid Y) = p(\beta \mid Y, \theta) p(\theta \mid Y, \tau) \tag{B13}$$

$$= \frac{p(Y \mid \beta) p(\beta \mid \theta) p(\theta \mid \tau)}{p(Y)} \tag{B14}$$

***Correcting the BMR Error***

Because conditional independence assumptions are defined at the customer level, we need establish independence of $y_i$ and $Y_{-i}$ after marginalizing over $\beta_{-i}$.

**Lemma 1.** $p(Y \mid \theta) = p(y_i \mid \theta) p(Y_{-i} \mid \theta)$.

*Proof.* By the Law of Total Probability and Eq. B4,

$$p(Y \mid \theta) = \int p(Y \mid \beta, \theta) p(\beta \mid \theta) d\beta \tag{B15}$$

$$= \int p(Y \mid \beta) p(\beta \mid \theta) d\beta \tag{B16}$$

Substitute Eqs. B8 and B9 into Eq. B16.

$$p(Y \mid \theta) = \iint p(y_i \mid \beta_i) p(Y_{-i} \mid \beta_{-i}) p(\beta_i \mid \theta) p(\beta_{-i} \mid \theta) d\beta_i d\beta_{-i} \tag{B17}$$

Rearrange terms, factor the integral into two parts, and integrate over $\beta_i$ and $\beta_{-i}$ separately to get the result.

$$p(Y \mid \theta) = \int p(y_i \mid \beta_i) p(\beta_i \mid \theta) d\beta_i \cdot \int p(Y_{-i} \mid \beta_{-i}) p(\beta_{-i} \mid \theta) d\beta_{-i} \tag{B18}$$

$$= p(y_i \mid \theta) p(Y_{-i} \mid \theta) \tag{B19}$$

■

Next, we show that conditional on $\theta$, the only data that affects $\beta_i$ is $y_i$.

**Lemma 2.** $p(\beta_i \mid Y, \theta) = p(\beta_i \mid y_i, \theta)$

*Proof.* By Bayes' Theorem,

$$p(\beta_i \mid Y, \theta) = \frac{p(Y \mid \beta_i, \theta)}{p(Y \mid \theta)} p(\beta_i \mid \theta) \tag{B20}$$

Factor the numerator in Eq. B20.

$$p(\mathrm{Y} \mid \beta_i, \theta) = f(y_i, \mathrm{Y}_{-i} \mid \beta_i, \theta) \tag{B21}$$

$$= p(y_i \mid \mathrm{Y}_{-i}, \beta_i, \theta) p(\mathrm{Y}_{-i} \mid \beta_i, \theta) \tag{B22}$$

A corollary to Eq. B3 is that $p(\mathrm{Y}_{-i} \mid \beta_i, \theta) = p(\mathrm{Y}_{-i} \mid \theta)$. Substitute that and Eq. B10 into Eq. B22.

$$p(\mathrm{Y} \mid \beta_i, \theta) = p(y_i \mid \beta_i, \theta) p(\mathrm{Y}_{-i} \mid \theta) \tag{B23}$$

Back to Eq. B20, substitute Eq. B23 in the numerator, and the result of Lemma 1 in the denominator.

$$p(\beta_i \mid \mathrm{Y}, \theta) = \frac{p(y_i \mid \beta_i, \theta) p(\mathrm{Y}_{-i} \mid \theta)}{p(y_i \mid \theta) p(\mathrm{Y}_{-i} \mid \theta)} p(\beta_i \mid \theta) \tag{B24}$$

After cancelling terms in Eq. B24, applying Bayes Theorem to the RHS gives the result. ■

The next lemma states the recursive property of Bayesian updating.

**Lemma 3.**

$$p(\theta \mid \mathrm{Y}, \tau) = \frac{p(y_i \mid \theta)}{p(y_i \mid \mathrm{Y}_{-i})} p(\theta \mid \mathrm{Y}_{-i}, \tau) \tag{B25}$$

*Proof.* By Bayes' Theorem,

$$p(\theta \mid \mathrm{Y}, \tau) = \frac{p(\mathrm{Y} \mid \theta) p(\theta \mid \tau)}{p(\mathrm{Y})} \tag{B26}$$

$$p(\theta \mid \mathrm{Y}_{-i}, \tau) = \frac{p(\mathrm{Y}_{-i} \mid \theta) p(\theta \mid \tau)}{p(\mathrm{Y}_{-i})} \tag{B27}$$

In Eq. B26, factor $p(\mathrm{Y} \mid \theta)$ using Lemma 1 and factor $p(\mathrm{Y})$ using the definition of joint probability.

$$p(\theta \mid \mathrm{Y}, \tau) = \frac{p(y_i \mid \theta)}{p(y_i \mid \mathrm{Y}_{-i})} \frac{p(\mathrm{Y}_{-i} \mid \theta) p(\theta \mid \tau)}{p(\mathrm{Y}_{-i})} \tag{B28}$$

Replace the second fraction in the RHS of Eq. B28 with the LHS of Eq. B27 to get the result. ■

We can now derive the correct $p(\beta_i \mid \mathrm{Y}, \tau)$ to replace BMR-13.

**Proposition B1.**

(B29) $$p(\beta_i \mid \mathrm{Y}, \tau) = \frac{p(y_i \mid \beta_i) p_{\theta \mid \mathrm{Y}_{-i}, \tau}(\beta_i)}{p(y_i \mid \mathrm{Y}_{-i})}$$

*Proof.* We restate BMR-7 here. Note that $p(\beta_i \mid \mathrm{Y}, \theta) = p(\beta_i \mid y_i, \theta)$ is established in Lemma 2.

(BMR-7) $$p(\beta_i \mid \mathrm{Y}, \tau) = \int p(\beta_i \mid y_i, \theta) p(\theta \mid \mathrm{Y}, \tau) \mathrm{d}\theta,$$

Replace $p(\beta_i \mid y_i, \theta)$ with Eq. 1, and replace $p(\theta \mid \mathrm{Y}, \tau)$ with the result from Lemma 3.

(B30) $$p(\beta_i \mid \mathrm{Y}, \tau) = \int \frac{p(y_i \mid \beta_i) p(\beta_i \mid \theta)}{p(y_i \mid \theta)} \frac{p(y_i \mid \theta)}{p(y_i \mid \mathrm{Y}_{-i})} p(\theta \mid \mathrm{Y}_{-i}, \tau) \mathrm{d}\theta$$

Cancel the $p(y_i \mid \theta)$ terms and factor out terms that do not depend on $\theta$.

(B31) $$p(\beta_i \mid \mathrm{Y}, \tau) = \frac{p(y_i \mid \beta_i)}{p(y_i \mid \mathrm{Y}_{-i})} \int p(\beta_i \mid \theta) p(\theta \mid \mathrm{Y}_{-i}, \tau) \mathrm{d}\theta$$

The integrand in Eq. B31 is defined in Eq. D2, giving the result.

(B32) $$p(\beta_i \mid \mathrm{Y}, \tau) = \frac{p(y_i \mid \beta_i)}{p(y_i \mid \mathrm{Y}_{-i})} p_{\theta \mid \mathrm{Y}_{-i}, \tau}(\beta_i)$$

■

## WEB APPENDIX C: PARALLELIZING A CORRECTED ALGORITHM

A practitioner may wonder if there is any way to amend the BMR algorithm to be consistent with the corrected math. To maintain the same spirit of BMR, a modified algorithm should retain two parallelizable stages with minimal communication overhead between nodes. Otherwise, it would be a different algorithm altogether.

As a constructive measure for BMR and potential adopters, we describe one possible approach, with no claims of optimality. Stage 1 would have to generate a separate $\{\theta_s^r\}_i$ for each customer, where $y_i$ is held out for each run. First, generate "full shard" samples from $p(\theta \mid \mathrm{Y}_s, \tau)$ on each node. Then, sample from $p(\theta \mid \mathrm{Y}_{s-i}, \tau)$ on each customer $i$'s cohort's node, where $\mathrm{Y}_{s-i}$ is the shard for cohort $s$ after removing $y_i$. This would happen $N_S$ times on each node, once for each customer in the cohort. The pool for customer $i$, $\{\theta_s^r\}_i$, would combine the full shard samples generated on all the other nodes with samples from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$. This approach would take $N_S + 1$ times as long as BMR's algorithm.

## WEB APPENDIX D: THE BMR ALGORITHM OVERFITS CUSTOMER DATA

In this section we offer a Bayesian theoretical perspective on why BMR's derivation of the target posterior density could not have been correct. BMR express the posterior density of $\beta_i$ as

(BMR-13) $$p(\beta_i \mid \mathrm{Y}) = \frac{p_{\theta|\mathrm{Y},\tau}(\beta_i)p(y_i \mid \beta_i)}{p(y_i)},$$

where

(D1) $$p_{\theta|\mathrm{Y},\tau}(\beta_i) = \int p(\beta_i \mid \theta)p(\theta \mid \mathrm{Y}, \tau)\mathrm{d}\theta$$

In BMR-13, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ acts as an informative prior on $\beta_i$. But BMR describe $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ as both "the posterior predictive density of $\beta_i$: the density of $\beta_i$ before observing $y_i$, given $p(\theta \mid \mathrm{Y}, \tau)$" and "a highly informative prior distribution for $\beta_i$, before observing $y_i$" (p. 1003, col. 1). These statements are difficult to reconcile because $y_i$ is already included in Y. BMR-13 is double-counting the focal customer's data, once as part of the information contained in $p(\theta \mid \mathrm{Y}, \tau)$, and then again in the data likelihood $p(y_i \mid \beta_i)$. Therefore, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ cannot be a valid prior on $\beta_i$.

A more suitable prior on $\beta_i$ would contain information about $\mathrm{Y}_{-i}$: all customer data *except* $y_i$.

(D2) $$p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i) = \int p(\beta_i \mid \theta)p(\theta \mid \mathrm{Y}_{-i}, \tau)\mathrm{d}\theta$$

Because $p(\theta \mid \mathrm{Y}_{-i}, \tau)$ does not condition on data from customer $i$, $p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$ truly captures prior information about $\beta_i$ before $y_i$ is observed. Then, once $y_i$ is observed,

(D3) $$p(\beta_i \mid \mathrm{Y}, \tau) \propto p(y_i \mid \beta_i)p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$$

reflects the properly updated posterior distribution of $\beta_i$. This is precisely what we get when we correct BMR's error.

The BMR algorithm reflects their incorrect derivation, so it creates *systematic* bias in $p(\beta_i \mid \mathrm{Y}, \tau)$. Note that $y_i$ is not a *random* observation, but the data that corresponds to the specific $\beta_i$ being inferred. Unlike $p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$, $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ is influenced by $y_i$, so BMR-estimated posterior distributions of $\beta_i \mid \mathrm{Y}$ are biased

toward the observed data. If $y_i$ were, say, an outlier data point, and the algorithm were deployed on a large number of nodes (large $S$), the magnitude of the bias could be substantial. Potential adopters of the algorithm should be aware that the error and associated bias exist.

## WEB APPENDIX E: NOTES ON BMR'S THEORETICAL PROOFS

In this section we explain how the mathematical error affects BMR's theorems in their Web Appendix. We do not discuss theorems whose subject matter falls outside the scope of this paper. We include this section to assist BMR and the reader, and we do not claim rigorous results.

The following expressions are defined in BMR and used in this section, but were not used in our main text.

(BMR-10) $$p(\beta_i \mid \mathrm{Y}) = \mathrm{E}_{\theta|\mathrm{Y},\tau}(p(\beta_i \mid \theta))p(y_i \mid \beta_i)$$

(BMR-14) $$\dot{p}_{\theta|\mathrm{Y},\tau}(\beta_i) = \frac{1}{R}\sum_r p(\beta_i \mid \theta^r)$$

(BMR-15) $$\dot{p}(\beta_i \mid \{\theta^r\}, \mathrm{Y}) \propto \dot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)p(y_i \mid \beta_i)$$

Because $\mathrm{E}_{\theta|\mathrm{Y},\tau}(p(\beta_i \mid \theta)) = p_{\theta|\mathrm{Y},\tau}(\beta_i)$, BMR-10 is equivalent to BMR-13.

***Theorem 1***

Using our notation, Gelman et al. (2004, Eq. 1.4) defines a posterior predictive density (PPD) as

(E1) $$p(y_i \mid \mathrm{Y}_{-i}) = \int p(y_i \mid \beta_i, \theta)g(\beta_i, \theta \mid \mathrm{Y}_{-i})\mathrm{d}\beta_i\,\mathrm{d}\theta$$

That is, it is a predictive distribution for new data, conditional on previously observed data. BMR call $p(\beta_i \mid \mathrm{Y}_{-i})$ a PPD, but it is better described as a "marginal posterior distribution." Note that Theorem 1 is not referenced in the BMR paper at all. In the main text, BMR refer to $p_{\theta|\mathrm{Y},\tau}(\beta_i)$ (conditional on Y) as a PPD (p. 1003, col. 1).

***Theorem 2***

Restate this theorem as

(E2) $$\mathrm{E}\big(\dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)\big) = \mathrm{E}_{\theta|\mathrm{Y}_{-i}}(p(\beta_i)) = p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$$

Let $\theta^r$ be a sample from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$. Replacing BMR-14 with

(E3) $$\dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i) = \frac{1}{R}\sum_r p(\beta_i \mid \theta^r)$$

makes the restated theorem correct.

***Theorem 3***

This theorem does not need to be restated. It would be correct after making the following changes.

1. Replace BMR-10 with $p(\beta_i \mid \mathrm{Y}) = \mathrm{E}_{\theta|\mathrm{Y}_{-i},\tau}(p(\beta_i \mid \theta))p(y_i \mid \beta_i)$.
2. Define $\{\theta^r\}$ as samples from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$ instead of $p(\theta \mid \mathrm{Y}, \tau)$.
3. Replace BMR-15 with $\dot{p}(\beta_i \mid \{\theta^r\}, \mathrm{Y}, \tau) \propto \dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)p(y_i \mid \beta_i)$
4. Apply the restated Theorem 2.

***Theorem 4***

BMR acknowledge taking a frequentist perspective to their asymptotic analysis, where "Y is a random sample from a distribution for some fixed, nonrandom, unknown parameter." In that sense, $\sqrt{N}\left(\theta_N^r - \theta\right) \xrightarrow{P} N\left(0, I_\theta^{-1}\right)$ represents convergence in probability to the fixed parameter $\theta$. But in the rest of the paper, the treatment of $\theta$ is Bayesian, as a random variable. In the convergence proofs, the fixed $\theta$ should be noted as something else, like $\theta^*$.

We think Theorem 4.1.1 should be correct if $\dot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$ were replaced with $\dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$ and $\{\theta^r\}$ are samples from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$ instead of $p(\theta \mid \mathrm{Y}, \tau)$. However, partitioning $\mathrm{Y}_{-i}$ into shards, with $\mathrm{Y}_{-i} = \{\mathrm{Y}_{1:S-1}, \mathrm{Y}_{S-i}\}$ would leave one shard with $N_S - 1$ observations. It is possible that the theorem is still valid in the limit, but that remains to be proven.

***Theorem 5***

Restate the theorem as $\lim_{N\to\infty}\left[\ddot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)\right] = \dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$. *If* we accept that Theorem 4 holds, then the theorem can be corrected with the following changes.

1. Denote $\theta_N^r$ as a draw from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$ instead of $p(\theta \mid \mathrm{Y}, \tau)$.
2. Replace all instances of $\dot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$ with $\dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$.

***Theorem 6***

This theorem does not need to be restated. It would be correct after making the following changes.

1. In BMR-21, replace $\ddot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$ with $\ddot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$.

2. Define $\{\theta^r\}$ as samples from $p(\theta \mid \mathrm{Y}_{-i}, \tau)$ instead of $p(\theta \mid \mathrm{Y}, \tau)$.

3. Apply the restated Theorem 5 and Theorem 2.

***Theorem 8***

Because one observation needs to be held out of $\mathrm{Y}_S$, it is not clear that the assumptions of Theorem 4 apply.

***Theorem 11***

By Eq. B29, BMR-13 should be

$$p(\beta_i \mid \mathrm{Y}) = \frac{p(y_i \mid \beta_i) p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)}{p(y_i \mid \mathrm{Y}_{-i})} \tag{E4}$$

To correct this theorem, replace $p_{\theta|\mathrm{Y},\tau}(\beta_i)$, $\dot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$, and $\ddot{p}_{\theta|\mathrm{Y},\tau}(\beta_i)$ with $p_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$, $\dot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$, and $\ddot{p}_{\theta|\mathrm{Y}_{-i},\tau}(\beta_i)$, respectively. Also, replace the normalizing constant $p(y_i)$ with $p(y_i \mid \mathrm{Y}_{-i})$.

However, it is not immediately clear if the conclusion of this theorem can be supported by Theorem 4.

## WEB APPENDIX F: PARALLELIZATION BIAS

**Proposition F1.** *For $S > 1$, the expected variance of $p^*(\theta \mid Y_{1:S}, \tau)$ is strictly greater than the variance of $p(\theta \mid Y, \tau)$.*

*Proof.* Stage 1 of the published BMR algorithm samples $\theta$ from

$$p^*(\theta \mid Y_{1:S}, \tau) = \frac{1}{S}\sum_{s=1}^{S} p(\theta \mid Y_s, \tau) \tag{F1}$$

The first two moments of this distribution are

$$\mathrm{E}^*(\theta \mid Y_{1:S}, \tau) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau) \tag{F2}$$

$$\mathrm{E}^*(\theta^2 \mid Y_{1:S}, \tau) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta^2 \mid Y_s, \tau) \tag{F3}$$

So the variance is

$$\mathrm{var}^*(\theta \mid Y_{1:S}, \tau) = \mathrm{E}^*(\theta^2 \mid Y_{1:S}, \tau) - \mathrm{E}^*(\theta \mid Y_{1:S}, \tau)^2 \tag{F4}$$

$$= \frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta^2 \mid Y_s, \tau) - \left[\frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau)\right]^2 \tag{F5}$$

By the Law of Total Variance,

$$\mathrm{E}(\theta^2 \mid Y_s, \tau) = \mathrm{var}(\theta \mid Y_s, \tau) + \mathrm{E}(\theta \mid Y_s, \tau)^2 \tag{F6}$$

Substitute Eq. F6 into the first summand in Eq. F5.

$$\mathrm{var}^*(\theta \mid Y_{1:S}, \tau) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{var}(\theta \mid Y_s, \tau) + \frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau)^2 - \left[\frac{1}{S}\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau)\right]^2 \tag{F7}$$

By Jensen's inequality, $\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau)^2 \geq \left[\sum_{s=1}^{S} \mathrm{E}(\theta \mid Y_s, \tau)\right]^2$. Therefore,

$$\mathrm{var}^*(\theta \mid Y_{1:S}, \tau) \geq \frac{1}{S}\sum_{s=1}^{S} \mathrm{var}(\theta \mid Y_s, \tau) \tag{F8}$$

So we know that the variance of $p^*(\theta \mid Y_{1:S}, \tau)$ will be at least as great as the variance of $p(\theta \mid Y_{1:S}, \tau)$.

If $S > 1$, then $N_S < N$ and $\mathrm{E}_Y(\mathrm{var}(\theta \mid Y_s, \tau)) > \mathrm{var}(\theta \mid Y, \tau)$ (where $\mathrm{E}_Y(\cdot)$ is a expectation across

random partitions in a frequentist sense). Averaging across shards,

$$\frac{1}{S}\sum_{s=1}^{S} \mathrm{E}_{\mathrm{Y}}(\mathrm{var}(\,\theta \mid \mathrm{Y}_s, \tau\,)) > \mathrm{var}(\,\theta \mid \mathrm{Y}, \tau\,) \tag{F9}$$

Taking the expectation across all terms in Eq. F8, and combining with Eq. F9, gives us

$$\mathrm{E}_{\mathrm{Y}}(\mathrm{var}(\,\theta \mid \mathrm{Y}_{1:S}, \tau\,)) \geq \frac{1}{S}\sum_{s=1}^{S} \mathrm{E}_{\mathrm{Y}}(\mathrm{var}(\,\theta \mid \mathrm{Y}_s, \tau\,)) > \mathrm{var}(\,\theta \mid \mathrm{Y}, \tau\,) \tag{F10}$$

■

In Eq. F7, we see that $\mathrm{var}(\,\theta \mid \mathrm{Y}_{1:S}, \tau\,)$ represents two sources of variation in $\theta$. The first term on the RHS is the average of the shard-level posterior variances. The second and third terms on the RHS define the variation in the shard-level means across shards.

**Proposition F2.** *For $S > 1$, the expected variance of $p^*(\,\beta_i \mid \mathrm{Y}_{1:S}, \tau\,)$ is strictly greater than the variance of $p(\,\beta_i \mid \mathrm{Y}, \tau\,)$.*

*Proof.* By the Law of Total Variance,

$$\mathrm{var}(\,\beta_i \mid \mathrm{Y}, \tau\,) = \mathrm{E}(\mathrm{var}(\,\beta_i \mid \theta, \tau\,)) + \mathrm{var}(\mathrm{E}(\,\beta_i \mid \theta, \tau\,)) \tag{F11}$$

$$\mathrm{var}^*(\,\beta_i \mid \mathrm{Y}, \tau\,) = \mathrm{E}^*(\mathrm{var}(\,\beta_i \mid \theta, \tau\,)) + \mathrm{var}^*(\mathrm{E}(\,\beta_i \mid \theta, \tau\,)) \tag{F12}$$

If $\mathrm{E}(\,\theta \mid \mathrm{Y}, \tau\,) = \mathrm{E}^*(\,\theta \mid \mathrm{Y}_{1:S}, \tau\,)$, then $\mathrm{E}(\mathrm{var}(\,\beta_i \mid \theta, \tau\,)) = \mathrm{E}^*(\mathrm{var}(\,\beta_i \mid \theta, \tau\,))$. Therefore, if $\mathrm{var}^*(\mathrm{E}(\,\beta_i \mid \theta, \tau\,)) > \mathrm{var}(\mathrm{E}(\,\beta_i \mid \theta, \tau\,))$, then $\mathrm{var}^*(\,\beta_i \mid \mathrm{Y}, \tau\,) > \mathrm{var}(\,\beta_i \mid \mathrm{Y}, \tau\,)$. Intuitively, this occurs because an increase in the posterior variance of $\theta \mid \mathrm{Y}$ results in greater variance in the means of $\beta_i \mid \theta$, so $\mathrm{var}(\,\beta_i \mid \mathrm{Y}, \tau\,)$ increases as well. ■

## WEB APPENDIX REFERENCES

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin (2004). *Bayesian Data Analysis*. 2nd ed. Boca Raton, Fla.: Chapman and Hall.