Supplement to "Scalable Empirical Bayes Inference and Bayesian Sensitivity Analysis"

Hani Doss
Department of Statistics
University of Florida
doss@stat.ufl.edu

Antonio Linero
Department of Statistics and Data Science
University of Texas at Austin
antonio.linero@austin.utexas.edu

Abstract

This document provides supporting material for "Scalable Empirical Bayes Inference and Bayesian Sensitivity Analysis" by Hani Doss and Antonio Linero, specifically an additional illustration, proofs of Theorems 1–3, and a proof that the importance weighted marginal density estimator is an unbiased estimate of the true marginal density.

Throughout this document, sections, equations, figures, and tables are labelled with the prefix "S". We do this in order to avoid confusion with the equations, figures, etc. of the main paper.

S-1 A Model Selection Problem in Binary Regression

Here we illustrate our methodology on a data set analyzed through the robit model for binary regression, and before we consider the data set, we review the robit model and discuss its salient features. Suppose that (x_j, Y_j), $j = 1, \ldots, m$ are independent observations, where x_j is a predictor variable of length p and Y_j is a binary response, taking the values 0 and 1. The two most common regression models for binary data are the probit and logistic models. The probit model relates Y_j to x_j through the equation $P(Y_j = 1) = \Phi(x_j^\top \beta)$, where $\beta_{p\times 1}$ is a vector of regression coefficients and Φ is the cumulative distribution function (cdf) of the standard normal distribution. The logistic model relates Y_j to x_j through the equation $P(Y_j = 1) = \exp(x_j^\top \beta)/(1 + \exp(x_j^\top \beta))$. It is well known that the maximum likelihood estimates of the regression parameters for the probit and logistic models are not robust to outliers (more precisely, points with outlying design points), and Albert and Chib (1993) (see also Mudholkar and George 1978) proposed replacing the normal cdf Φ with the cdf of the t distribution with d degrees of freedom, which we denote by T_d. The logistic cdf given by $L(u) = \exp(u)/(1 + \exp(u))$ is closely approximated by T_d for d equal to about seven, and Φ is well approximated by T_d when d is large. Gelman and Hill (2007, chapter 6) showed that the maximum likelihood estimator for robit models with small d is robust against discordant data points. Thus, the robit family of models effectively includes the probit and logistic models, but also includes models which are more suitable when the data set has influential outlying observations.

An important modeling issue is the determination of the parameter d: small values of d give rise to robust estimators, but at the cost of inefficiency if the true model governing the data is the logistic or probit

model. The problem of choosing d can be cast in the empirical Bayes framework discussed in Section 1 of this paper. There is one distinction, however. In the empirical Bayes setup described earlier, the likelihood is fixed and the prior varies; here the prior is fixed and the likelihood varies. Nevertheless, the principle that the parameter d should be selected by maximizing the marginal likelihood still applies. This was noted by Roy (2014), who used umbrella sampling to estimate the marginal likelihood function $m(\cdot)$ (up to a constant) and hence its argmax. Our methodology applies with only notational changes. Let $\ell_{d,y}(\beta)$ be the likelihood of β under the robit model with parameter d, and let π_β and π_D denote the prior densities on β and d, respectively. We then have $\pi_{D\,|\,y}(d)/\pi_D(d) \propto m(d)$, and the situation is as before.

Let X be the design matrix. For the prior on β, we will follow the recommendation of Gelman et al. (2008) to use $t(0, \Sigma_0, d_0)$ (the multivariate t-distribution with degrees of freedom parameter equal to d_0, center equal to 0, and scatter matrix equal to Σ_0), with $\Sigma_0 = a(X^\top X)^{-1}$; Gelman et al. (2008) advocate taking d_0 to be moderately small and a to be some large number, and we will take $d_0 = 3$ and $a = 10^4$. (It is possible to treat d_0 and a as hyperparameters to be estimated, i.e. consider $h = (d, d_0, a)$, but we did not follow that route in order to keep the focus on d.) We will take $\pi_D = \mathrm{gam}(2, 1/10)$, which was noted by Juárez and Steel (2010) to be an accurate approximation to Jeffrey's prior for the degrees of freedom parameter in a skew-t response model.

In our illustrations, we use the No-U-Turn Sampler (NUTS, Hoffman and Gelman 2014) as implemented in the `STAN` software package (Carpenter et al., 2017) to sample from the posterior distribution $\pi_{(\beta,D)\,|\,y}$. NUTS implements a form of HMC in which the integration time is selected adaptively on a per-iteration basis. Recent work of Livingstone et al. (2016) identified conditions for geometric ergodicity of HMC; however, to our knowledge, establishing conditions under which NUTS produces a geometrically ergodic chain remains an open problem.

To estimate $\pi_{D\,|\,y}$, we view the robit model via a standard data augmentation scheme as follows:

$$\begin{aligned} &\epsilon_j \overset{\text{indep}}{\sim} \mathcal{N}(0, \lambda_j^{-1}), \quad Z_j = x_j^\top \beta + \epsilon_j, \qquad j = 1, \ldots, m, \\ &\lambda_j \overset{\text{iid}}{\sim} \mathrm{gam}(d/2, d/2), \qquad j = 1, \ldots, m, \\ &\beta \sim t(0, \Sigma_0, d_0). \end{aligned} \tag{S-1.1}$$

Taking $Y_j = I(Z_j > 0)$, (S-1.1) induces the robit model with parameter d. In (S-1.1), as usual, each distributional statement is understood to hold conditionally on the parameters defined in the lines below it. We now note that the conditional distribution of d given (Z, λ, β) and the data is given by

$$\begin{aligned} f(d\,|\,Z, \lambda, \beta, y) &\propto \mathrm{gam}(d; 2, .1) \prod_{j=1}^m \Big[\mathrm{gam}\Big(\lambda_j; \frac{d}{2}, \frac{d}{2}\Big) \cdot \mathcal{N}(Z_j; x_j^\top \beta, \lambda_j^{-1})\Big] \cdot t(\beta; 0, \Sigma_0, d_0) \\ &\propto \mathrm{gam}(d; 2, .1) \prod_{j=1}^m \mathrm{gam}\Big(\lambda_j; \frac{d}{2}, \frac{d}{2}\Big). \end{aligned} \tag{S-1.2}$$

If $\alpha = d/2$, then the marginal distribution of α is $\mathrm{gam}(2, 0.2)$. Consider the very simple model where $\alpha \sim \mathrm{gam}(2, 0.2)$, and given α, $\lambda_i \overset{\text{iid}}{\sim} \mathrm{gam}(\alpha, \alpha)$. In this model, the posterior distribution of α is the expression given in the last line of (S-1.2), and we are in precisely the situation considered by Miller (2019) and discussed in Section 3.2. Therefore, we may use the approximation of Miller (2019) to implement the

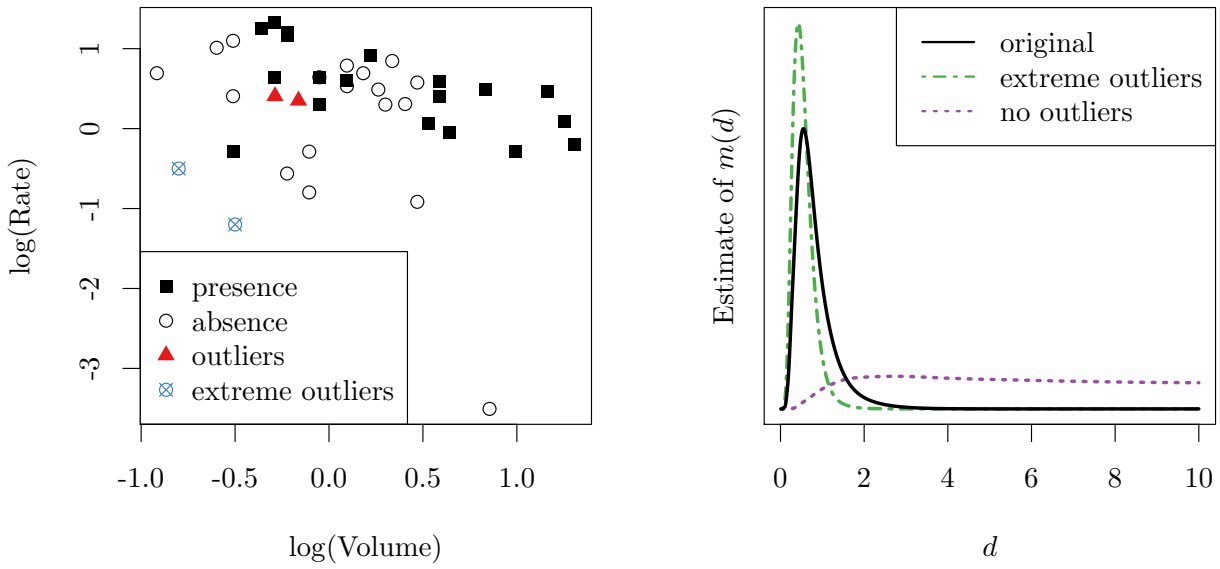

Figure S-1: Left panel shows the Finney (1947) data: circles denote absence ($Y = 0$) of vaso-constriction; squares denote its presence ($Y = 1$); triangles colored red denote the outliers in the original data set; and circles colored blue with an x mark denote artificially created more extreme outliers. Right panel shows the estimate of $m(\cdot)$ (up to a constant) for the three versions of the data set.

IWMDE approach which, as before, is nearly as good as Rao-Blackwellization. In principle, the Markov chain could be constructed by using this data augmentation strategy instead of using NUTS; however we found the mixing with NUTS to be substantially faster.

We now consider the vaso-constriction data set of Finney (1947). The data set is very simple—there are only two predictors and only 39 observations—so it does not make apparent the usefulness of our theoretical results. We use it however, because it has two influential outliers and is easy to visualize; thus, the effect that these outliers have on the results of our analysis is easy to understand. Moreover, it has been studied in several papers which deal with diagnostics for binary regression; therefore we can compare our results with those in the previous literature. The data come from a controlled study of the effect of taking a single deep breath of volume V at rate R, on presence of the reflex "vaso-constriction" in the skin of the digits. There are $m = 39$ observations, and we will follow the traditional analysis, in which the predictor is taken to be $x = (1, \log(V), \log(R))$. Pregibon (1981) established that observations 4 and 18 have very high influence in determining a logistic regression fit. The data are plotted on the left panel of Figure S-1, in which the outliers are plotted with different plotting symbols. For later use, we also plotted two additional points, which are the original outliers made more extreme. We will consider three versions of the Finney (1947) data: (i) a version with the two outliers removed, (ii) the original data set, and (iii) a version in which the outliers are removed and replaced with outliers which are more extreme. We do this in order to show trends in our analysis.

We implemented our procedure, running a Markov chain on (β, d) for 10,000 iterations, discarding

the first 5,000 to burn-in. (Note: (β, d) corresponds to (θ, h) in our earlier notation.) The right panel of Figure S-1 gives plots of the estimate of $m(\cdot)$ (up to a constant) over the range 0 to 10 for the three versions of the data set. Table S-1 gives more results from our analysis. Let $d_{\text{opt}} = \arg\max_d m(d)$ and $\hat{d}_{\text{opt}} = \arg\max_d \widehat{M}_n(d)$ ($\widehat{M}_n$ is defined prior to the statement of Theorem 1). Row 1 of the table gives $\hat{d}_{\text{opt}}$ and a 95% confidence interval for d_{opt} for the three versions of the data set. The interval is constructed via the method of batching (see the paragraph following the statement of Theorem 2). For the original version, the point estimate is 0.55, and the confidence interval is $(0.52, 0.58)$, ruling out the logistic and probit models. From row 1 of the table we see that as outliers are introduced and made more extreme, $\hat{d}_{\text{opt}}$ gets smaller, indicating that the empirical Bayes procedure is correctly choosing models which are increasingly robust against outliers.

We can evaluate the performance of the model that is based on a given value of d by calculating the mean squared error given by $\text{MSE}(d) = (1/m)\sum_{j=1}^{m}(Y_j - \widehat{Y}_{d,j})^2$. In this expression, $\widehat{Y}_{d,j}$ is the expected value under the posterior distribution, given d and the data, of the response of an individual with covariate x_j; that is, $\widehat{Y}_{d,j} = E_{\pi_{\beta \mid (d,y)}}(T_d(x_j^\top \beta))$. Rows 2 and 3 of Table S-1 show that the predictive power of the model that uses $d = \hat{d}_{\text{opt}}$ is greater than that of either the logistic or probit model when there exist outliers, with gains increasing as outliers are made more extreme. As expected, when there are no outliers, the predictive performance of the model reverts back to the performance of the logistic and probit models; this is supported by the right panel for Figure S-1, which shows that for the no-outlier case, the marginal likelihoods for $d = \hat{d}_{\text{opt}}$ and $d = 7$ (which approximates the logistic model) are not appreciably different.

	Outliers removed	Original data	Outliers made extreme
$\hat{d}_{\text{opt}}$ and CI for d_{opt}	2.70 $(1.15, 4.25)$	0.55 $(0.52, 0.58)$	0.43 $(0.41, 0.45)$
MSE ratio: logistic vs. EB	1.00	1.35	1.86
MSE ratio: probit vs. EB	0.99	1.40	1.94

Table S-1: Comparison of the model that uses $d = \hat{d}_{\text{opt}}$ ("EB") with the logistic and probit models for three versions of the Finney data: the version with the outliers removed, the original data set, and the version in which the outliers are removed and replaced with more extreme outliers.

Next, we use this example to illustrate that approaches based on importance sampling, such as serial tempering, umbrella sampling, or parallel tempering, will not succeed in this problem when the sample size is large (here, "sample size" refers to the data set, not the length of the Monte Carlo simulation). This is true even though in the present situation $\dim(d) = 1$, so the curse of dimensionality does not come in. To this end, we created a data set designed to mimic the original Finney data except that the data sample size is 5,000, and we did this as follows. (i) We generated $x_1^*, \ldots, x_{5,000}^*$ by sampling with replacement from the original x_j's. (ii) For $j = 1, \ldots, 5{,}000$, we generated $Y_{5,000}^*$ according to the robit model with $d = 0.55$, predictors $x_1^*, \ldots, x_{5,000}^*$, and β equal to the posterior mean under the fully-Bayes model. We then considered the simple importance sampling estimate

$$\widetilde{M}_n(d) = \frac{1}{n}\sum_{i=1}^{n} \frac{\ell_{d,y}(\beta_i)}{\ell_{d_1,y}(\beta_i)}, \tag{S-1.3}$$

where $\ell_{d,y}(\beta)$ is the likelihood of β under the robit model with parameter d, and the β_i's are a Markov chain with invariant distribution equal to $\pi_{\beta\,|\,(d_1,y)}$; see (1.2). This estimate gives a consistent approximation to $m(d)/m(d_1)$.

One should keep in mind that $\arg\max_d m(d)$ is not necessarily equal to 0.55 (it depends on the artificial data set and is unknown); however, because the data sample size is large, we expect it to be close to 0.55. In a preliminary experiment using the FBEB methodology with a Markov chain length of 2,000, we obtained $(0.593, 0.614)$ as a 95% confidence interval for $\arg\max_d m(d)$. The left panel of Figure S-2 shows this confidence interval (it is depicted by the two vertical dashed lines) and displays estimates of $m(d)$ using (S-1.3) with $d_1 = 0.4$ and $d_1 = 0.55$, and using the FBEB estimator, each of these three being computed for a Markov chain length of 2,000. Note that, although $d_1 = 0.4$ is quite close to the true maximum, the importance sampling estimator completely fails: for $d_1 = 0.4$, $\arg\max_d \widetilde{M}_n(d)$ is not even close to being in the confidence interval, and the relative likelihood of all points in the confidence interval is essentially 0. Furthermore, even when $d_1 = 0.55$, $\arg\max_d \widetilde{M}_n(d)$ is not in the confidence interval.

The right panel of Figure S-2 explains the cause of this failure. It plots the sequence

$$\omega_{in}(d; d_1) = \frac{\widetilde{\omega}_{in}(d; d_1)}{\sum_{i'=1}^{n} \widetilde{\omega}_{i'n}(d; d_1)} \qquad \text{where} \qquad \widetilde{\omega}_{in}(d; d_1) = \frac{\ell_{d,y}(\beta_i)}{\ell_{d_1,y}(\beta_i)}$$

for $d_1 = 0.55$. To get an interpretation of this sequence, note that the estimate of the quantity $I(d)$ in Section 2.2 based on a chain run at d_1 is given by $\sum_{i=1}^{n} \omega_{in}(d; d_1) g(\beta_i)$ (see (2.16)), and the $\omega_{in}(d; d_1)$'s are viewed as weights. The right panel of Figure S-2 shows that, even when d_1 is close to d, most of the weight comes from a very small number of samples from the posterior, suggesting a highly imprecise estimate of $m(d)/m(d_1)$. This confirms the statement made near the end of Section 2 that, in certain situations where the data sample size is large, unless d is very close to d_1, the importance sampling estimate is very unstable. The significance of this illustration is that it highlights a circular problem inherent with the importance sampling estimate: to estimate $\arg\max_d m(d)$, one has to have a good approximation to it in order to determine an acceptable value of d_1.

We conclude that the importance sampling estimate will not succeed when the sample size is large due to the fact that the "radius" around d_1 in which one can accurately estimate $m(d)/m(d_1)$ is very narrow. We also note that strategies such as serial tempering and umbrella sampling—which rely on selecting many different d_1's—will also fail unless the number of d_1's considered is exceedingly large. FBEB neatly bypasses this problem, producing an accurate estimate of $\arg\max_d m(d)$ as long as the posterior assigns mass near $\arg\max_d m(d)$ and the Markov chain we use mixes fast.

S-2 Proofs of Theorems 1–3 and of an Auxiliary Result

Proof of Theorem 1

We first prove (2.2). We consider the following notationally simpler but conceptually equivalent situation: (U, V) is a bivariate random vector, $G^{(v)}$ denotes the conditional distribution of U given that $V = v$, and $G^{(v)}$ has density $g^{(v)}$. Then (2.2) is essentially equivalent to the following:

$$g^{(V)} \text{ is an unbiased estimator of the density of } U. \tag{S-2.1}$$

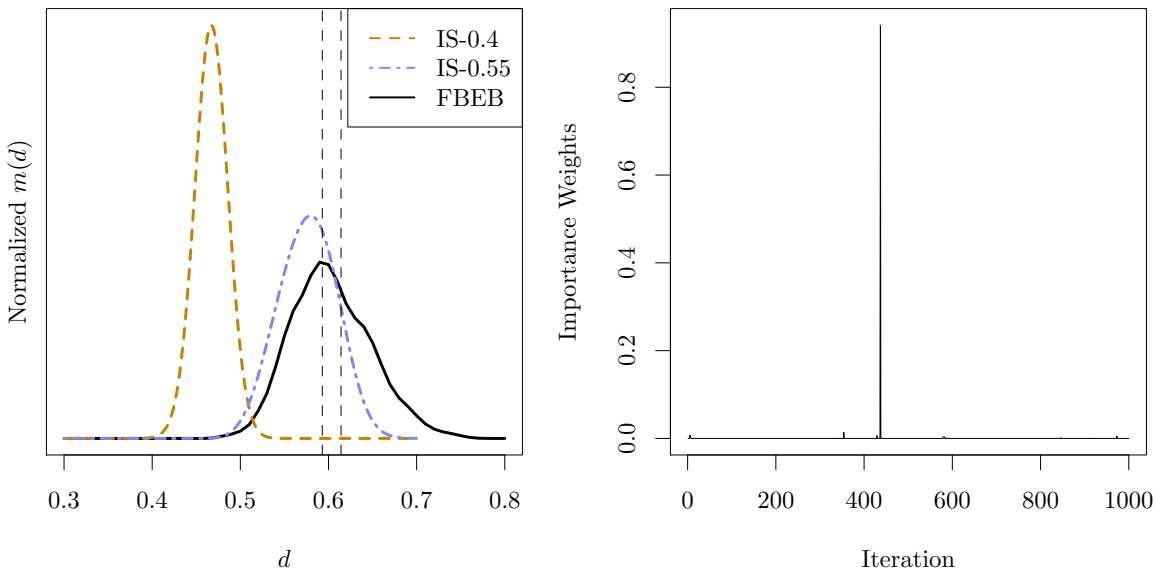

Figure S-2: Left panel: plots of the estimated marginal likelihood surface (normalized to integrate to 1) for the importance sampling estimator using $d_1 = 0.4$ and $d_1 = 0.55$, with the estimate provided by FBEB. Right panel: plot of the importance weights for the model $d = 0.55$ when $d_1 = 0.4$ is used for importance sampling.

A direct approach for proving (S-2.1) involves imposing differentiability conditions on $G^{(v)}$ so that $g^{(v)}$ can be viewed as the derivative of $G^{(v)}$, and laboriously justifying changing the order of limits and integrals. Actually, (2.2) is, strictly speaking, nonsense, in that densities are defined only almost everywhere (hence the word "essentially" above); on the other hand, (S-2.1) can be proved rigorously by showing that the function $g(\cdot) = E(g^{(V)}(\cdot))$ satisfies the defining property of the (marginal) density of U, namely that for any Borel set A, $\int_A g(u)\,du = P(U \in A)$. We prove this last statement by writing

$$\int_A g(u)\,du = \int_A E(g^{(V)}(u))\,du = E\left(\int_A g^{(V)}(u)\,du\right) = E(G^{(V)}\{A\}) = P(U \in A),$$

where the second equality is a consequence of Fubini's theorem, and the fourth is a consequence of the law of iterated expectation.

To prove Part 1, Note that

$$\widehat{M}_n(h) = \frac{(1/n)\sum_{i=1}^n f_h(\theta_i)}{\pi_h(h)} = \frac{\left(\sum_{r=1}^R S_{h,r}\right)/R}{\pi_h(h)\left(\sum_{r=1}^R N_r\right)/R},$$

where the first equality is from (2.1) and (2.10), and the second equality is from (2.12). The key condition for obtaining a Glivenko-Cantelli result for $\left(\sum_{r=1}^R S_{h,r}\right)/R$ is Condition C9. This Glivenko-Cantelli result states that

$$\sup_{h\in\mathcal{H}} \left|\left(\sum_{r=1}^R S_{h,r}\right)/R - \pi_{h\,|\,y}(h)E(N_1)\right| \xrightarrow{\text{a.s.}} 0,$$

where we have used the fact that $E(S_{h,r}) = E(f_h(\theta))E(N_1)$. Since π_h is continuous and positive on the compact set $\mathcal{H}$ (condition C2), π_h has a strictly positive lower bound on $\mathcal{H}$. Therefore,

$$\sup_{h\in\mathcal{H}} \left| \frac{\left(\sum_{r=1}^{R} S_{h,r}\right)/R}{\pi_h(h)\left(\sum_{r=1}^{R} N_r\right)/R} - \frac{\pi_{h\,|\,y}(h)}{\pi_h(h)} \right| \xrightarrow{\text{a.s.}} 0,$$

i.e. $\sup_{h\in\mathcal{H}} \left|\widehat{M}_n(h) - M(h)\right| \xrightarrow{\text{a.s.}} 0$.

Part 2 is an immediate consequence of the result below, which is a standard fact from analysis (and is easy to prove). □

Fact Suppose that H is a compact subset of Euclidean space, and let $g_n,\ n = 1, 2, \ldots$ and g be deterministic real-valued functions defined on H. Suppose further that g is continuous and has a unique maximizer, and that for each n the maximizer of g_n exists and is unique. If g_n converges to g uniformly on H, then the maximizer of g_n converges to the maximizer of g.

Proof of Theorem 2

The proof is similar to the proof of Part 1 of Theorem 4 of Doss and Park (2018). The difference between the current situation and the situation considered in Doss and Park (2018) is that they are considering $\arg\max_h E(f_h(\theta))$, where $f_h(\theta) = \pi_{\theta\,|\,(h,y)}(\theta)/\pi_{\theta\,|\,(h_1,y)}(\theta)$, for some fixed $h_1 \in \mathcal{H}$, and the expectation is with respect to $\pi_{\theta\,|\,(h_1,y)}$. In the present paper, we are working with the function $f_h(\theta) = \pi_{h\,|\,(\theta,y)}(h)$ and, moreover, the function of h whose argmax we are considering is $E(f_h(\theta)/\pi_h(h))$, where the expectation is with respect to $\pi_{\theta\,|\,y}$. Our conditions C5, C7–C10 correspond to conditions A3, A1, A2, A4, and A5 in Doss and Park (2018), in that order. Conditions C1, C3, C4, and C6 are stated in Doss and Park (2018) in the text prior to their Theorem 4. The proof of Part 1 of Theorem 4 of Doss and Park (2018) applies to Part 1 of our Theorem 2 with minor modifications, and Condition C2 is needed for these modifications. □

Proof of Theorem 3

Proof of Part 1 By Theorem 1, (2.3) holds uniformly in h, i.e.

$$\sup_h \left| \frac{1}{n}\sum_{i=1}^{n} \pi_{h\,|\,(\theta=\theta_i,y)}(h) - \pi_{h\,|\,y}(h) \right| \xrightarrow{\text{a.s.}} 0. \tag{S-2.2}$$

By replacing the function $f_h(\theta) = \pi_{h\,|\,(\theta,y)}(h)$ (see (2.10)) with the function $g(\theta)f_h(\theta)$, the same arguments used to show (S-2.2) can be used again to show that

$$\sup_h \left| \frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\pi_{h\,|\,(\theta=\theta_i,y)}(h) - N(h) \right| \xrightarrow{\text{a.s.}} 0,$$

where, recall, $N(h)$ is defined in (2.14d). Part 1 now follows from the definition of $\hat{I}_n$ given in (2.16), and the fact that $\pi_{h\,|\,y}(\cdot)$ is bounded away from 0 on $\mathcal{H}$ (by C1 and D1). It should be noted that Theorem 1 requires condition C9, which we are not making here. But this is because this condition is subsumed by the conditions that $E(S_{h_0,1}) < \infty$ for some $h_0 \in \mathcal{H}$ and that $E(\sup_h \|\nabla_h S_{h,1}\|) < \infty$, both of which are implied by the assumptions of Theorem 3.

Proof of Part 2 Before proceeding, we need to review the notion of Donsker class. Let $(\Omega, \mathcal{A}, P)$ be a probability space, and suppose $X_1, X_2, \ldots \overset{\text{iid}}{\sim} P$. For $f \in L_2(\Omega, \mathcal{A}, P)$, let $Y_n(f) = n^{1/2}\big(\big[(1/n)\sum_{i=1}^n f(X_i)\big] - \int f\, dP\big)$. A class of functions $\mathcal{F} \subset L_2(\Omega, \mathcal{A}, P)$ is called P-Donsker if the stochastic process $\{Y_n(f),\ f \in \mathcal{F}\}$ converges in distribution to a mean-zero Gaussian process indexed by $\mathcal{F}$. Here, convergence takes place in $l^\infty(\mathcal{F})$, the space of bounded functions from $\mathcal{V}$ to $\mathbb{R}$ equipped with the supremum norm. (For more formal definitions, see van der Vaart and Wellner (1996).) In our case, the probability space we're dealing with is $(\Theta, \mathcal{B}_\Theta, \pi_{\theta\,|\,y})$, where $\mathcal{B}_\Theta$ is the Borel σ-field on Θ, and the function class is $\mathcal{F} = \{f_h,\ h \in \mathcal{H}\}$, where f_h is defined by (2.10).

Suppose first that the θ_i's are iid. The condition $E\big(\sup_h \|\nabla_h f_h(\theta)\|^2\big) < \infty$ together with the condition that $E\big(f_{h_0}^2(\theta)\big) < \infty$ for some $h_0 \in \mathcal{H}$ implies that $E\big(\sup_h f_h^2(\theta)\big) < \infty$. By D2, the condition $E\big(\sup_h f_h^2(\theta)\big) < \infty$, and the bound on covering numbers for Euclidean classes given in Nolan and Pollard (1987, page 789), we see that the conditions of Pollard-Koltchinskii theorem (for a statement see, e.g., Theorem 8.19 of Kosorok (2008)) are satisfied, and we may conclude that the class $\mathcal{F}$ is $\pi_{\theta\,|\,y}$-Donsker. Now with probability one, $n^{1/2}\big(\big[(1/n)\sum_{i=1}^n f_h(\theta_i)\big] - \int f_h\, d\pi_{\theta\,|\,y}\big) \in C(\mathcal{H})$, so weak convergence actually takes place in the space $C(\mathcal{H})$ endowed with the sup norm topology (cf. van der Vaart and Wellner, 1996, Theorem 1.3.10). This proves (2.17) (for the iid case). Very similarly, let $\mathcal{G} = \{gf_h,\ h \in \mathcal{H}\}$. Under the same condition that for every θ, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$, and the condition that $E\big(\sup_h \|\nabla_h g(\theta) f_h(\theta)\|^2\big) < \infty$, the class $\mathcal{G}$ is $\pi_{\theta\,|\,y}$-Donsker. This proves (2.18) (for the iid case). Define the map $\Phi\colon C(\mathcal{H}) \times C(\mathcal{H}) \to C(\mathcal{H})$ by $(\Phi(x,y))(h) = x(h)/y(h)$. This map is Hadamard differentiable at the point $(N(\cdot), \pi_{h\,|\,y}(\cdot))$ (see van der Vaart and Wellner (1996, page 388) for a proof, and van der Vaart and Wellner (1996, Section 3.9.1) for a definition of Hadamard differentiability). The convergence statement (2.19) (for the iid case) now follows from the functional delta method (van der Vaart and Wellner, 1996, Theorem 3.9.4).

For the case where the θ_i's are a geometrically ergodic Markov chain, in essence we use (2.12) to translate results regarding averages of the independent random variables $S_{h,r}$, $r = 1, \ldots, R$ and $T_{h,r}$, $r = 1, \ldots, R$ to averages of the dependent variables $f(\theta_i)$, $i = 1, \ldots, n$ and $g(\theta_i)f(\theta_i)$, $i = 1, \ldots, n$. However, the definition of the Donsker classes is a bit delicate, so we need to be careful in how we define these. For the Markov chain case, the probability space is $(\Theta^\infty, \mathcal{B}_{\Theta^\infty}, \Pi)$, where Π is the distribution of the entire sequence $\theta_1, \theta_2, \ldots$. In the Athreya and Ney (1978) construction, the sequence $\theta_1, \theta_2, \ldots$ induces the regeneration sequence $\tau_0, \tau_1, \ldots$. Therefore, for any $h \in \mathcal{H}$, $S_{h,1}$ may be viewed as a function mapping Θ^∞ to $\mathbb{R}$, and the function class is taken to be $\mathscr{F} = \{S_{h,1},\ h \in \mathcal{H}\}$. Similarly, we consider the function class $\mathscr{G} = \{T_{h,1},\ h \in \mathcal{H}\}$. Now D6 and D7 take the place of the conditions $E\big(\sup_h \|\nabla_h f_h(\theta)\|^2\big) < \infty$ and $E\big(\sup_h \|\nabla_h g(\theta) f_h(\theta)\|^2\big) < \infty$, respectively, and we conclude that $\mathscr{F}$ and $\mathscr{G}$ are Π-Donsker, i.e. (2.17) and (2.18) hold for the case of a geometrically ergodic Markov chain. The proof that (2.17) and (2.18) imply (2.19) is identical to the proof for the iid case. □

Proof That $\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(\cdot)$ Is an Unbiased Estimate of $\pi_{h\,|\,y}(\cdot)$

Let $\Pi_{(\theta,h)\,|\,y}$ denote the probability measure associated with the density $\pi_{(\theta,h)\,|\,y}$ (note: we act as if the density is with respect to Lebesgue measure, but we do this strictly for notational convenience). We will

show that $E_{\pi_{(\theta,h)\,|\,y}}\big(\hat{\pi}^{\text{iwmde}}_{h\,|\,y}(\cdot)\big)$ satisfies the defining property of $\pi^{\text{iwmde}}_{h\,|\,y}(\cdot)$, namely that

$$\int_A E_{\pi_{(\theta,h)\,|\,y}}\left(w_\theta(h)\frac{\pi_{(\theta,h)\,|\,y}(\theta,h_*)}{\pi_{(\theta,h)\,|\,y}(\theta,h)}\right)dh_* = \Pi_{(\theta,h)\,|\,y}(\{h\in A\}) \qquad \text{for all Borel sets } A\subset\mathcal{H}.$$

For any Borel set $A\subset\mathcal{H}$, we have

$$\begin{aligned}
\int_A E_{\pi_{(\theta,h)\,|\,y}}\left(w_\theta(h)\frac{\pi_{(\theta,h)\,|\,y}(\theta,h_*)}{\pi_{(\theta,h)\,|\,y}(\theta,h)}\right)dh_* &= \int_A\int_\Theta\int_{\mathcal{H}} w_\theta(h)\pi_{(\theta,h)\,|\,y}(\theta,h_*)\,dh\,d\theta\,dh_* \\
&= \int_A\int_\Theta \pi_{(\theta,h)\,|\,y}(\theta,h_*)\left(\int_{\mathcal{H}} w_\theta(h)\,dh\right)d\theta\,dh_* \\
&= \int_A\int_\Theta \pi_{(\theta,h)\,|\,y}(\theta,h_*)\,d\theta\,dh_* \\
&= \Pi_{(\theta,h)\,|\,y}(\{h\in A\}). \qquad \square
\end{aligned}$$

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88** 669–679.

Athreya, K. B. and Ney, P. (1978). A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society* **245** 493–501.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P. and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software* **76**.

Doss, H. and Park, Y. (2018). An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. *The Annals of Statistics* **46** 1630–1663.

Finney, D. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34** 320–334.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics* **2** 1360–1383.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15** 1593–1623.

Juárez, M. A. and Steel, M. F. (2010). Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics* **28** 52–66.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Livingstone, S., Betancourt, M., Byrne, S. and Girolami, M. (2016). On the geometric ergodicity of Hamiltonian Monte Carlo. *arXiv preprint arXiv:1601.08057* .

Miller, J. W. (2019). Fast and accurate approximation of the full conditional for gamma shape parameters. *Journal of Computational and Graphical Statistics* **28** 476–480.

Mudholkar, G. S. and George, E. O. (1978). A remark on the shape of the logistic distribution. *Biometrika* **65** 667–668.

Nolan, D. and Pollard, D. (1987). U-Processes: Rates of convergence. *The Annals of Statistics* **15** 780–799.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics* **9** 705–724.

Roy, V. (2014). Efficient estimation of the link function parameter in a robust Bayesian binary regression model. *Computational Statistics & Data Analysis* **73** 87–102.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes, With Applications to Statistics*. Springer-Verlag, New York.