ESTIMATING STANDARD ERRORS FOR IMPORTANCE SAMPLING ESTIMATORS WITH MULTIPLE MARKOV CHAINS

Vivekananda Roy, Aixin Tan, and James M. Flegal

Iowa State University, University of Iowa, and University of California, Riverside

Supplementary Material

S1 Proof of Theorem 1

In the supplement, we denote Doss and Tan (2014) by D&T. The proof of the consistency of $\hat{\boldsymbol{d}}$ follows from D&T section A.1 and is omitted. Establishing a CLT for $\hat{\boldsymbol{d}}$ is analogous to section A.2 of D&T, but there are significant differences. Below we establish the CLT for $\hat{\boldsymbol{d}}$ and finally show that $\widehat{V}$ is a consistent estimator of V.

We begin by considering $n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0)$. As before, let ∇ represents the gradient operator. We expand $\nabla \ell_n$ at $\hat{\boldsymbol{\zeta}}$ around $\boldsymbol{\zeta}_0$, and using the appropriate scaling factor, we get

$$-n^{-1/2}\big(\nabla \ell_n(\hat{\boldsymbol{\zeta}}) - \nabla \ell_n(\boldsymbol{\zeta}_0)\big) = -n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*) n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0), \tag{S1.1}$$

where $\boldsymbol{\zeta}_*$ is between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Consider the left side of (S1.1), which is just $n^{1/2} n^{-1} \nabla \ell_n(\boldsymbol{\zeta}_0)$, since $\nabla \ell_n(\hat{\boldsymbol{\zeta}}) = 0$. There are several nontrivial components to the proof, so we first give an outline.

1. Following D&T we show that each element of the vector $n^{-1}\nabla \ell_n(\boldsymbol{\zeta}_0)$ can be represented as a linear combination of mean 0 averages of functions of the k chains.

2. Based on Step 1, applying CLT for each of the k Markov chain averages, we obtain a CLT for the scaled score vector. In particular, we show that $n^{1/2} n^{-1} \nabla \ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega)$, where Ω defined in (2.8) of the main text involves infinite sums of auto-covariances of each chain.

3. Following Geyer (1994) it can be shown that $-n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*) \xrightarrow{\text{a.s.}} B$ and that $\big(-n^{-1}\nabla^2 \ell_n(\boldsymbol{\zeta}_*)\big)^{\dagger} \xrightarrow{\text{a.s.}} B^{\dagger}$, where B is defined in (2.7).

4. We conclude that $n^{1/2}(\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, B^{\dagger}\Omega B^{\dagger})$.

5. Since $\boldsymbol{d} = g(\boldsymbol{\zeta}_0)$ and $\hat{\boldsymbol{d}} = g(\hat{\boldsymbol{\zeta}})$, where g is defined in (2.4), by the Delta method it follows that $n^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, V)$ where $V = D^\top B^\dagger \Omega B^\dagger D$.

We now provide the details.

1. Start by considering $n^{-1}\nabla \ell_n(\boldsymbol{\zeta}_0)$. For $r = 1, \ldots, k$, from D&T we have

$$\begin{aligned}
\frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= w_r \sum_{i=1}^{n_r} \big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0)\big) - \sum_{\substack{l=1 \\ l \neq r}}^{k} w_l \sum_{i=1}^{n_l} p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) \\
\text{(can be shown to)} \quad &= w_r \sum_{i=1}^{n_r} \Big(1 - p_r(X_i^{(r)}, \boldsymbol{\zeta}_0) - \big[1 - E_{\pi_r}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big]\Big) \qquad \text{(S1.2)} \\
&\quad - \sum_{\substack{l=1 \\ l \neq r}}^{k} w_l \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big].
\end{aligned}$$

That is, (S1.2) can be used to view $n^{-1}\partial \ell_n(\boldsymbol{\zeta}_0)/\partial \zeta_r$ as a linear combination of mean 0 averages of functions of the k chains.

2. Next, we need a CLT for the vector $\nabla \ell_n(\boldsymbol{\zeta}_0) = (\partial \ell_n(\boldsymbol{\zeta}_0)/\partial \zeta_1, \cdots, \partial \ell_n(\boldsymbol{\zeta}_0)/\partial \zeta_k)^T$, that is, to show that $n^{-1/2}\nabla \ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} N(0, \Omega)$ as $n \to \infty$. Note that,

$$\begin{aligned}
\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} &= -\frac{1}{\sqrt{n}} \sum_{l=1}^{k} w_l \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] \\
&= -\sum_{l=1}^{k} \sqrt{\frac{n}{n_l}} a_l \frac{1}{\sqrt{n_l}} \sum_{i=1}^{n_l} \big[p_r(X_i^{(l)}, \boldsymbol{\zeta}_0) - E_{\pi_l}\big(p_r(X, \boldsymbol{\zeta}_0)\big)\big] \\
&= -\sum_{l=1}^{k} \sqrt{n} a_l \bar{\bar{Y}}^{(r,l)},
\end{aligned}$$

where $\bar{\bar{Y}}^{(r,l)} := \frac{1}{n_l} \sum_{i=1}^{n_l} Y_i^{(r,l)}$ and $Y_i^{(r,l)}$ is as defined in (2.6). Since $p_r(x, \boldsymbol{\zeta}) \in (0, 1)$ for all x, r and $\boldsymbol{\zeta}$, we have $E_{\pi_l}\big(|p_r(X, \boldsymbol{\zeta}_0) - E_{\pi_l}(p_r(X, \boldsymbol{\zeta}_0))|^{2+\delta}\big) < \infty$ for any $\delta > 0$. Then since Φ_l is polynomially ergodic of order $m > 1$, we have asymptotic normality for the univariate quantities $\sqrt{n_l}\bar{\bar{Y}}^{(r,l)}$ (see e.g. Corollary 2 of Jones (2004)). Since $n_l/n \to s_l$ for $l = 1, \ldots, k$ and a_l's are known, by independence of the k chains, we conclude that

$$\frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_r} \xrightarrow{d} \mathcal{N}(0, \Omega_{rr}) \text{ as } n \to \infty,$$

where Ω is defined in (2.8). Next, we extend the component-wise CLT to a joint CLT.

Consider any $\boldsymbol{t} \in (t_1, \cdots, t_k) \in \mathbb{R}^k$, we have

$$\begin{aligned}
&t_1 \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_1} + \cdots + t_k \frac{1}{\sqrt{n}} \frac{\partial \ell_n(\boldsymbol{\zeta}_0)}{\partial \zeta_k} \\
&= -\sum_{l=1}^{k} \left(t_1 \sqrt{n} a_l \frac{\sum_{i=1}^{n_l} Y_i^{(1,l)}}{n_l} + \cdots + t_k \sqrt{n} a_l \frac{\sum_{i=1}^{n_l} Y_i^{(k,l)}}{n_l} \right) \\
&= -\sum_{l=1}^{k} \sqrt{\frac{n}{n_l}} a_l \frac{\sum_{i=1}^{n_l} \left(t_1 Y_i^{(1,l)} + \cdots + t_k Y_i^{(k,l)} \right)}{\sqrt{n_l}} \xrightarrow{d} \mathcal{N}(0, \boldsymbol{t}^T \Omega \boldsymbol{t}) \text{ as } n \to \infty.
\end{aligned}$$

Hence, the Cramér-Wold device implies the joint CLT,

$$n^{-1/2} \nabla \ell_n(\boldsymbol{\zeta}_0) \xrightarrow{d} \mathcal{N}(0, \Omega) \qquad \text{as } n \to \infty. \tag{S1.3}$$

Steps 3-5 are omitted since the derivations are basically the same as in D&T.

Next we provide a proof of the consistency of the estimate of the asymptotic covariance matrix V, that is, we show that $\widehat{V} \equiv \widehat{D}^\top \widehat{B}^\dagger \widehat{\Omega} \widehat{B}^\dagger \widehat{D} \xrightarrow{\text{a.s.}} V \equiv D^\top B^\dagger \Omega B^\dagger D$ as $n \to \infty$. Since $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ and $\hat{\boldsymbol{d}} \xrightarrow{\text{a.s.}} \boldsymbol{d}$, it implies that $\widehat{D} \xrightarrow{\text{a.s.}} D$. From D&T, we know that $\widehat{B} \xrightarrow{\text{a.s.}} B$ and using the spectral representation of $\widehat{B}$ and of B, it follows that $\widehat{B}^\dagger \xrightarrow{\text{a.s.}} B^\dagger$.

To complete the proof, we now show that $\widehat{\Omega} \xrightarrow{\text{a.s.}} \Omega$ where the BM estimator $\widehat{\Omega}$ is defined in (2.13). This will be proved in couple of steps. First, we consider a single chain Φ_l used to calculate k quantities and establish a multivariate CLT. We use the results in Vats et al. (2015) who obtain conditions for the nonoverlapping BM estimator to be strongly consistent in multivariate settings. Second, we combine results from the k independent chains. Finally, we show that $\widehat{\Omega}$ is a strongly consistent estimator of Ω.

Denote $\bar{\bar{Y}}^{(l)} = \left(\bar{\bar{Y}}^{(1,l)}, \bar{\bar{Y}}^{(2,l)}, \ldots, \bar{\bar{Y}}^{(k,l)} \right)^\top$. Similar to deriving (S1.3) via the Cramér-Wold device, we have the following joint CLT for W_l:, $\sqrt{n_l} \bar{\bar{Y}}^{(l)} \xrightarrow{d} \mathcal{N}(0, \Sigma^{(l)})$ as $n_l \to \infty$, where $\Sigma^{(l)}$ is a $k \times k$ covariance matrix with

$$\Sigma_{rs}^{(l)} = E_{\pi_l}\{Y_1^{(r,l)} Y_1^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_1^{(r,l)} Y_{1+i}^{(s,l)}\} + \sum_{i=1}^{\infty} E_{\pi_l}\{Y_{1+i}^{(r,l)} Y_1^{(s,l)}\}. \tag{S1.4}$$

The nonoverlapping BM estimator of $\Sigma^{(l)}$ is given in (2.10). We now prove the strong consistency of $\widehat{\Sigma}^{(l)}$. Note that $\widehat{\Sigma}^{(l)}$ is defined using the terms $\bar{Z}_m^{(r,l)}$'s which involve the random quantity $\hat{\boldsymbol{\zeta}}$. We define $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0)$ to be $\widehat{\Sigma}^{(l)}$ with $\boldsymbol{\zeta}_0$ substituted for $\hat{\boldsymbol{\zeta}}$, that is,

$$\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[\bar{Y}_m^{(l)} - \bar{\bar{Y}}^{(l)} \right] \left[\bar{Y}_m^{(l)} - \bar{\bar{Y}}^{(l)} \right]^\top \quad \text{for } l = 1, \ldots, k,$$

where $\bar{Y}_m^{(l)} = \left(\bar{Y}_m^{(1,l)}, \ldots, \bar{Y}_m^{(k,l)} \right)^\top$ with $\bar{Y}_m^{(r,l)} := \sum_{j=mb_l+1}^{(m+1)b_l} Y_j^{(r,l)} / b_l$. We prove $\widehat{\Sigma}^{(l)} \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ in two steps: (1) $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ and (2) $\widehat{\Sigma}^{(l)} - \widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} 0$. Strong consistency of the

multivariate BM estimator $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0)$ requires both $e_l \to \infty$ and $b_l \to \infty$. Since for all r, $E_{\pi_l}\big(|p_r(X,\boldsymbol{\zeta}_0) - E_{\pi_l}(p_r(X,\boldsymbol{\zeta}_0))|^{4+\delta}\big) < \infty$ for any $\delta > 0$, Φ_l is polynomially ergodic of order $m > 1$, and $b_l = \lfloor n_l^{\nu} \rfloor$ where $1 > \nu > 0$, it follows from Vats et al. (2015) that $\widehat{\Sigma}^{(l)}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} \Sigma^{(l)}$ as $n_l \to \infty$. We show $\widehat{\Sigma}^{(l)}_{rs} - \widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0) \xrightarrow{\text{a.s.}} 0$ where $\widehat{\Sigma}^{(l)}_{rs}$ and $\widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0)$ are the (r,s)th elements of the $k \times k$ matrices $\widehat{\Sigma}^{(l)}_{rs}$ and $\widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0)$ respectively. By the mean value theorem (in multiple variables), there exists $\boldsymbol{\zeta}^* = t\hat{\boldsymbol{\zeta}} + (1-t)\boldsymbol{\zeta}_0$ for some $t \in (0,1)$, such that

$$\widehat{\Sigma}^{(l)}_{rs} - \widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}_0) = \nabla \widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}^*) \cdot (\hat{\boldsymbol{\zeta}} - \boldsymbol{\zeta}_0), \tag{S1.5}$$

where $\cdot$ represents the dot product. Note that

$$\widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta}) = \frac{b_l}{e_l - 1} \sum_{m=0}^{e_l - 1} [\bar{Z}^{(r,l)}_m(\boldsymbol{\zeta}) - \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta})][\bar{Z}^{(s,l)}_m(\boldsymbol{\zeta}) - \bar{\bar{Z}}^{(s,l)}(\boldsymbol{\zeta})],$$

where $\bar{Z}^{(r,l)}_m(\boldsymbol{\zeta}) := \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X^{(l)}_j, \boldsymbol{\zeta})/b_l$ and $\bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta}) := \sum_{j=1}^{n_l} p_r(X^{(l)}_j, \boldsymbol{\zeta})/n_l$. Some calculations show that for $t \neq r$

$$\frac{\partial \bar{Z}^{(r,l)}_m(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_t} = -\frac{1}{b_l} \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X^{(l)}_j, \boldsymbol{\zeta}) p_t(X^{(l)}_j, \boldsymbol{\zeta})$$

and

$$\frac{\partial \bar{Z}^{(r,l)}_m(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_r} = \frac{1}{b_l} \sum_{j=mb_l+1}^{(m+1)b_l} p_r(X^{(l)}_j, \boldsymbol{\zeta})(1 - p_r(X^{(l)}_j, \boldsymbol{\zeta})).$$

We denote $\bar{U}^r_m := \bar{Z}^{(r,l)}_m(\boldsymbol{\zeta}) - E_{\pi_l}[p_r(X,\boldsymbol{\zeta})]$, $\bar{\bar{U}}^r := \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta}) - E_{\pi_l}[p_r(X,\boldsymbol{\zeta})]$, and similarly the centered versions of $\partial \bar{Z}^{(r,l)}_m(\boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t$ and $\partial \bar{\bar{Z}}^{(r,l)}(\boldsymbol{\zeta})/\partial \boldsymbol{\zeta}_t$ by $\bar{V}^{(r,t)}_m$ and $\bar{\bar{V}}^{(r,t)}$ respectively. Since $p_r(X,\boldsymbol{\zeta})$ is uniformly bounded by 1 and Φ_l is polynomially ergodic of order $m > 1$, there exist $\sigma^2_r, \tau^2_{r,t} < \infty$ such that $\sqrt{b_l}\bar{U}^r_m \xrightarrow{d} N(0,\sigma^2_r)$, $\sqrt{n_l}\bar{\bar{U}}^r \xrightarrow{d} N(0,\sigma^2_r)$, $\sqrt{b_l}\bar{V}^{(r,t)}_m \xrightarrow{d} N(0,\tau^2_{r,t})$, and $\sqrt{n_l}\bar{\bar{V}}^{(r,t)} \xrightarrow{d} N(0,\tau^2_{r,t})$. We have

$$\begin{aligned}
\frac{\partial \widehat{\Sigma}^{(l)}_{rs}(\boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}_t} &= \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} [\sqrt{b_l}(\bar{U}^r_m - \bar{\bar{U}}^r)\sqrt{b_l}(\bar{V}^{(s,t)}_m - \bar{\bar{V}}^{(s,t)}) + \sqrt{b_l}(\bar{V}^{(r,t)}_m - \bar{\bar{V}}^{(r,t)})\sqrt{b_l}(\bar{U}^s_m - \bar{\bar{U}}^s)] \\
&= \frac{1}{e_l - 1} \sum_{m=0}^{e_l-1} \left[\sqrt{b_l}\bar{U}^r_m \sqrt{b_l}\bar{V}^{(s,t)}_m + \sqrt{b_l}\bar{V}^{(r,t)}_m \sqrt{b_l}\bar{\bar{U}}^s_m\right] - \frac{1}{e_l - 1}\left[\sqrt{n_l}\bar{\bar{U}}^r \sqrt{n_l}\bar{\bar{V}}^{(s,t)} + \sqrt{n_l}\bar{\bar{V}}^{(r,t)} \sqrt{n_l}\bar{\bar{U}}^s\right].
\end{aligned}$$

It is easy to see that the negative term in the above expression goes to zero as $e_l \to \infty$. Further, since

$$\left|\sqrt{b_l}\bar{U}^r_m \sqrt{b_l}\bar{V}^{(s,t)}_m\right| \leq \frac{1}{2}\left[b_l(\bar{U}^r_m)^2\right] + \frac{1}{2}\left[b_l(\bar{V}^{(s,t)}_m)^2\right],$$

we have

$$\left|\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\sqrt{b_l}\bar{U}_m^r\sqrt{b_l}\bar{V}_m^{(s,t)}\right| \le \frac{1}{2}\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\left[b_l(\bar{U}_m^r)^2\right]+\frac{1}{2}\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\left[b_l(\bar{V}_m^{(s,t)})^2\right] \xrightarrow{\text{a.s.}} \frac{1}{2}\sigma_r^2+\frac{1}{2}\tau_{s,t}^2,$$

where the last step above is due to strong consistency of the BM estimators for the asymptotic variances of the sequences $\{p_r(X_j^{(l)},\boldsymbol{\zeta}), j=1,\cdots,n_l\}$ and $\{\partial p_s(X_j^{(l)},\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}_t, j=1,\cdots,n_l\}$ respectively. Similarly, we have

$$\left|\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\sqrt{b_l}\bar{V}_m^{(r,t)}\sqrt{b_l}\bar{U}_m^s\right| \le \frac{1}{2}\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\left[b_l(\bar{V}_m^{(r,t)})^2+\frac{1}{2}\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\left[b_l(\bar{U}_m^s)^2\right]\right] \xrightarrow{\text{a.s.}} \frac{1}{2}\tau_{r,t}^2+\frac{1}{2}\sigma_s^2.$$

Note that the terms $U_m^r V_m^{(r,t)}, \sigma_r^2, \tau_{r,t}^2$, etc, above actually depends on $\boldsymbol{\zeta}$, and we are indeed concerned with the case where $\boldsymbol{\zeta}$ takes on the value $\boldsymbol{\zeta}^*$, lying between $\hat{\boldsymbol{\zeta}}$ and $\boldsymbol{\zeta}_0$. Since, $\hat{\boldsymbol{\zeta}} \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$, $\boldsymbol{\zeta}^* \xrightarrow{\text{a.s.}} \boldsymbol{\zeta}_0$ as $n_l \to \infty$. Let $\|u\|$ denotes the L_1 norm of a vector $u \in \mathbb{R}^k$. So from (S1.5), and the fact that $\partial\widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}_t$ is bounded with probability one, we have

$$|\widehat{\Sigma}_{rs}^{(l)}-\widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}_0)| \le \max_{1\le t\le k}\left\{\left|\frac{\partial\widehat{\Sigma}_{rs}^{(l)}(\boldsymbol{\zeta}^*)}{\partial\boldsymbol{\zeta}_t}\right|\right\}\|\hat{\boldsymbol{\zeta}}-\boldsymbol{\zeta}_0\| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n\to\infty.$$

Since $\widehat{\Sigma}^{(l)} \xrightarrow{\text{a.s.}} \Sigma^{(l)}$, for $l=1,\ldots,k$, it follows that $\widehat{\Sigma} \xrightarrow{\text{a.s.}} \Sigma$ where $\widehat{\Sigma}$ is defined in (2.11) and Σ is the corresponding $k^2\times k^2$ covariance matrix, that is, Σ is a block diagonal matrix as $\widehat{\Sigma}$ with $\Sigma^{(l)}$ substituted for $\widehat{\Sigma}^{(l)}, l=1,\ldots,k$. Since $n_l/n \to s_l$ for $l=1,\ldots,k$, we have $A_n \to A_s$ as $n\to\infty$ where A_n is defined in (2.12) and

$$A_s = \left(-\sqrt{\frac{1}{s_1}}a_1 I_k \quad -\sqrt{\frac{1}{s_2}}a_2 I_k \quad \ldots \quad -\sqrt{\frac{1}{s_k}}a_k I_k\right).$$

Finally from (2.8) and (S1.4) we see that $\Omega = A_s\Sigma A_s^T$. So from (2.13) we have $\widehat{\Omega} \equiv A_n\widehat{\Sigma}A_n^T \xrightarrow{\text{a.s.}} A_s\Sigma A_s^T = \Omega$ as $n\to\infty$.

S2 Proof of Theorem 2

As in Buta and Doss (2011) we write

$$\sqrt{n}(\hat{u}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}})-u(\pi,\pi_1)) = \sqrt{n}(\hat{u}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}})-\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}))+\sqrt{n}(\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})-u(\pi,\pi_1)). \tag{S2.6}$$

First, consider the 2nd term, which involves randomness only from the 2nd stage. From (3.3) note that $\sum_{l=1}^k a_l E_{\pi_l} u(X;\boldsymbol{a},\boldsymbol{d}) = u(\pi,\pi_1)$. Then from (3.1) we have

$$\sqrt{n}(\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})-u(\pi,\pi_1)) = \sum_{l=1}^k a_l\sqrt{\frac{n}{n_l}}\frac{\sum_{i=1}^{n_l}(u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d})-E_{\pi_l}u(X;\boldsymbol{a},\boldsymbol{d}))}{\sqrt{n_l}}.$$

Since Φ_l is polynomially ergodic of order m and $E_{\pi_l}|u(X;\boldsymbol{a},\boldsymbol{d})|^{2+\delta}$ is finite where $m > 1+2/\delta$, it follows that $\sum_{i=1}^{n_l}(u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) - E_{\pi_l}u(X;\boldsymbol{a},\boldsymbol{d}))/\sqrt{n_l} \xrightarrow{d} N(0,\tau_l^2(\pi;\boldsymbol{a},\boldsymbol{d}))$ where $\tau_l^2(\pi;\boldsymbol{a},\boldsymbol{d})$ is defined in (3.4). As $n_l/n \to s_l$ and the Markov chains Φ_l's are independent, it follows that $\sqrt{n}(\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) - u(\pi,\pi_1)) \xrightarrow{d} N(0,\tau^2(\pi;\boldsymbol{a},\boldsymbol{d}))$.

Now we consider the 1st term in the right hand side of (S2.6). Letting $F(\boldsymbol{z}) = \hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{z})$, by Taylor series expansion of F about $\boldsymbol{d}$ we have

$$\sqrt{n}(F(\hat{\boldsymbol{d}}) - F(\boldsymbol{d})) = \sqrt{n}\nabla F(\boldsymbol{d})^\top(\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{\sqrt{n}}{2}(\hat{\boldsymbol{d}} - \boldsymbol{d})^\top\nabla^2 F(\boldsymbol{d}^*)(\hat{\boldsymbol{d}} - \boldsymbol{d}), \tag{S2.7}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$. Simple calculations show that

$$[\nabla F(\boldsymbol{d})]_{j-1} = \sum_{l=1}^{k}\frac{a_l}{n_l}\sum_{i=1}^{n_l}\frac{a_j\nu_j(X_i^{(l)})\nu(X_i^{(l)})}{(\sum_{s=1}^{k}a_s\nu_s(X_i^{(l)})/d_s)^2 d_j^2} \xrightarrow{\text{a.s.}} [c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \tag{S2.8}$$

where $[c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$ is defined in (3.5). We know that $n/N \to q$. Using similar arguments as in Buta and Doss (2011), it follows that $\nabla^2 F(\boldsymbol{d}^*)$ is bounded in probability. Thus from (S2.7) we have

$$\begin{aligned}\sqrt{n}(F(\hat{\boldsymbol{d}}) - F(\boldsymbol{d})) &= \sqrt{\frac{n}{N}}\nabla F(\boldsymbol{d})^\top\sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{1}{2\sqrt{N}}\sqrt{\frac{n}{N}}[\sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d})]^\top\nabla^2 F(\boldsymbol{d}^*)[\sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d})] \\ &= \sqrt{q}c(\pi;\boldsymbol{a},\boldsymbol{d})^\top\sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + o_p(1).\end{aligned}$$

Then Theorem 2 (1) follow from (S2.6) and the independence of the two stages of Markov chain sampling.

Next to prove Theorem 2 (2), note that, we already have a consistent BM estimator $\widehat{V}$ of V. From (S2.8), we have $[\hat{c}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} = [\nabla F(\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [c(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$. Applying mean value theorem on $[\nabla F(\boldsymbol{d})]_{j-1}$ and the fact that $\nabla^2 F(\boldsymbol{d}^*)$ is bounded in probability, it follows that $[\hat{c}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})]_{j-1} - [\hat{c}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} 0$. Writing $c(\pi;\boldsymbol{a},\boldsymbol{d})^\top V c(\pi;\boldsymbol{a},\boldsymbol{d})$ as $\sum_{i=1}^{k-1}\sum_{j=1}^{k-1}c_i V_{ij} c_j$, it then follows that $\hat{c}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})^\top\widehat{V}\hat{c}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} c(\pi;\boldsymbol{a},\boldsymbol{d})^\top V c(\pi;\boldsymbol{a},\boldsymbol{d})$.

We now show $\hat{\tau}_l^2(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})$ is a consistent estimator of $\tau_l^2(\pi;\boldsymbol{a},\boldsymbol{d})$ where τ_l^2 and $\hat{\tau}_l^2$ are defined in (3.4) and (3.7), respectively. Since the Markov chains $\{X_i^{(l)}\}_{i=1}^{n_l}$ are independent, it then follows that $\tau^2(\pi;\boldsymbol{a},\boldsymbol{d})$ is consistently estimated by $\hat{\tau}^2(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})$ completing the proof of Theorem 2 (2).

If $\boldsymbol{d}$ is known from the assumptions of Theorem 2 (2) and the results in Vats et al. (2015), we know that $\tau_l^2(\pi;\boldsymbol{a},\boldsymbol{d})$ is consistently estimated by its BM estimator $\hat{\tau}_l^2(\pi;\boldsymbol{a},\boldsymbol{d})$. Note that, $\hat{\tau}_l^2(\pi;\boldsymbol{a},\boldsymbol{d})$ is defined in terms of the quantities $u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d})$'s. We now show that $\hat{\tau}_l^2(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - \hat{\tau}_l^2(\pi;\boldsymbol{a},\boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

Denoting $\hat{\tau}_l^2(\pi;\boldsymbol{a},\boldsymbol{z})$ by $G(\boldsymbol{z})$, by the mean value theorem (in multiple variables), there exists $\boldsymbol{d}^* = t\hat{\boldsymbol{d}} + (1-t)\boldsymbol{d}$ for some $t \in (0,1)$, such that $G(\hat{\boldsymbol{d}}) - G(\boldsymbol{d}) = \nabla G(\boldsymbol{d}^*) \cdot (\hat{\boldsymbol{d}} - \boldsymbol{d})$. For any $j \in \{2, \cdots, k\}$, and $\boldsymbol{z} \in R^{+k-1}$,

$$\frac{\partial G(\boldsymbol{z})}{\partial z_j} = \frac{b_l}{e_l - 1}\left[\sum_{m=0}^{e_l-1} 2(\bar{u}_m(\boldsymbol{a},\boldsymbol{z}) - \bar{\bar{u}}(\boldsymbol{a},\boldsymbol{z}))\left(\frac{\partial \bar{u}_m(\boldsymbol{a},\boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}(\boldsymbol{a},\boldsymbol{z})}{\partial z_j}\right)\right] \tag{S2.9}$$

Let $\bar{W}_m := \bar{u}_m(\boldsymbol{a},\boldsymbol{z}) - E_{\pi_l}(u(X;\boldsymbol{a},\boldsymbol{z}))$ and $\bar{\bar{W}} := \bar{\bar{u}}(\boldsymbol{a},\boldsymbol{z}) - E_{\pi_l}(u(X;\boldsymbol{a},\boldsymbol{z}))$. Note that, there exists, $\sigma^2 < \infty$ such that $\sqrt{b_l}\bar{W}_m \xrightarrow{d} \mathrm{N}(0,\sigma^2)$, and $\sqrt{n_l}\bar{\bar{W}} \xrightarrow{d} \mathrm{N}(0,\sigma^2)$. Simple calculations show that

$$\frac{\partial \bar{u}_m(\boldsymbol{a},\boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2}\frac{1}{b_l}\sum_{i=mb_l+1}^{(m+1)b_l}\left[\frac{\nu(X_i^{(l)})\nu_j(X_i^{(l)})}{\left(\sum_s a_s\nu_s(X_i^{(l)})/z_s\right)^2}.\right]$$

Hence, letting $\alpha_j = E_{\pi_l}[\nu(X)\nu_j(X)/\left(\sum_s a_s\nu_s(X)/z_s\right)^2]$, we write

$$\frac{\partial \bar{u}_m(\boldsymbol{a},\boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}(\boldsymbol{a},\boldsymbol{z})}{\partial z_j} \equiv \frac{a_j}{z_j^2}\left\{\bar{Z}_{m,j}\right\} - \frac{a_j}{z_j^2}\left\{\bar{\bar{Z}}_j\right\},$$

where $\bar{Z}_{1,j} = (1/b_l)\sum_{i=1}^{b_l}[\nu(X_i^{(l)})\nu_j(X_i^{(l)})/\{\sum_s a_s\nu_s(X_i^{(l)})/z_s\}^2] - \alpha_j$ and $\bar{\bar{Z}}_j$ is similarly defined. Note that, there exists $\tau_j^2 < \infty$, such that $\sqrt{b_l}\bar{Z}_{m,j} \xrightarrow{d} \mathrm{N}(0,\tau_j^2)$, and $\sqrt{n_l}\bar{\bar{Z}}_j \xrightarrow{d} \mathrm{N}(0,\tau_j^2)$. From (S2.9) we have

$$\begin{aligned}
\frac{\partial G(\boldsymbol{z})}{\partial z_j} =& \frac{a_j}{z_j^2}\frac{2}{e_l-1}\sum_{m=0}^{e_l-1}\left[\sqrt{b_l}(\bar{W}_m - \bar{\bar{W}})\sqrt{b_l}\left(\bar{Z}_{m,j} - \bar{\bar{Z}}_j\right)\right] \\
=& \frac{a_j}{z_j^2}\frac{2}{e_l-1}\sum_{m=0}^{e_l-1}\left[\sqrt{b_l}\bar{W}_m\sqrt{b_l}\bar{Z}_{m,j}\right] \\
& - \frac{a_j}{z_j^2}2b_l\left[\bar{\bar{Z}}_j\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\bar{W}_m + \bar{\bar{W}}\frac{1}{e_l-1}\sum_{m=0}^{e_l-1}\bar{Z}_{m,j} - \frac{e_l}{e_l-1}\bar{\bar{W}}\bar{\bar{Z}}_j\right] \\
=& \frac{a_j}{z_j^2}\frac{2}{e_l-1}\sum_{m=0}^{e_l-1}\left[\sqrt{b_l}\bar{W}_m\sqrt{b_l}\bar{Z}_{m,j}\right] - \frac{a_j}{z_j^2}\frac{2}{e_l-1}\left[\sqrt{n_l}\bar{\bar{W}}\sqrt{n_l}\bar{\bar{Z}}_j\right].
\end{aligned}$$

Then using similar arguments as in the proof of Theorem 1, it can be shown that $\partial G(\boldsymbol{z})/\partial z_j$ is bounded with probability one. Then it follows that

$$|G(\hat{\boldsymbol{d}}) - G(\boldsymbol{d})| \leq \max_{1\leq j\leq k-1}\left\{\left|\frac{\partial G(\boldsymbol{d}^*)}{\partial z_j}\right|\right\}\|\hat{\boldsymbol{d}} - \boldsymbol{d}\| \xrightarrow{\text{a.s.}} 0.$$

S3 Proof of Theorem 3

As in the proof of Theorem 2 we write

$$\sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - E_\pi f) = \sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})) + \sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f). \tag{S3.10}$$

First, consider the 2nd term, which involves randomness only from the 2nd stage. Since

$$\hat{v} \xrightarrow{\text{a.s.}} \sum_{l=1}^{k} a_l E_{\pi_l} v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) = \int_{\mathsf{X}} \frac{f(x)\sum_{l=1}^{k} a_l \nu_l(x)/m_l}{\sum_{s=1}^{k} a_s \nu_s(x)/(m_s/m_1)} \nu(x)\,\mu(dx) = \frac{m}{m_1} E_\pi f,$$

we have $\sum_{l=1}^{k} a_l E_{\pi_l} v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) = E_\pi f u(\pi,\pi_1)$. Then from (3.1) we have

$$\sqrt{n}\begin{pmatrix} \hat{v}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f u(\pi,\pi_1) \\ \hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) - u(\pi,\pi_1) \end{pmatrix} = \sum_{l=1}^{k} a_l \sqrt{\frac{n}{n_l}} \frac{1}{\sqrt{n_l}} \sum_{i=1}^{n_l} \begin{pmatrix} v^{[f]}(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) - E_{\pi_l} v^{[f]}(X;\boldsymbol{a},\boldsymbol{d}) \\ u(X_i^{(l)};\boldsymbol{a},\boldsymbol{d}) - E_{\pi_l} u(X;\boldsymbol{a},\boldsymbol{d}) \end{pmatrix}. \tag{S3.11}$$

From the conditions of Theorem 3 and the fact that the Markov chains $\Phi_l, l = 1,\ldots,k$ are independent, it follows that the above vector (S3.11) converges in distribution to the bivariate normal distribution with mean 0 and covariance matrix $\Gamma(\pi;\boldsymbol{a},\boldsymbol{d})$ defined in (3.9). Then applying the Delta method to the function $g(x,y) = x/y$ we have a CLT for the ratio estimator $\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})$, that is, we have $\sqrt{n}(\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d}) - E_\pi f) \xrightarrow{d} N(0,\rho(\pi;\boldsymbol{a},\boldsymbol{d}))$ where $\rho(\pi;\boldsymbol{a},\boldsymbol{d}))$ is defined in (3.10).

Next letting $L(\boldsymbol{z}) = \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{z})$, by Taylor series expansion of L about $\boldsymbol{d}$ we have

$$\sqrt{n}(L(\hat{\boldsymbol{d}}) - L(\boldsymbol{d})) = \sqrt{n}\nabla L(\boldsymbol{d})^\top(\hat{\boldsymbol{d}} - \boldsymbol{d}) + \frac{\sqrt{n}}{2}(\hat{\boldsymbol{d}} - \boldsymbol{d})^\top \nabla^2 L(\boldsymbol{d}^*)(\hat{\boldsymbol{d}} - \boldsymbol{d}), \tag{S3.12}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$. Simple calculations show that

$$[\nabla L(\boldsymbol{d})]_{j-1} = [\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [e(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \tag{S3.13}$$

where $[e(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$ and $[\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$ are defined in (3.11) and (3.12) respectively. It can be shown that $\nabla^2 L(\boldsymbol{d}^*)$ is bounded in probability. Thus from (S3.12) we have $\sqrt{n}(L(\hat{\boldsymbol{d}}) - L(\boldsymbol{d})) = \sqrt{q}e(\pi;\boldsymbol{a},\boldsymbol{d})^\top\sqrt{N}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + o_p(1)$. Then Theorem 3 (1) follow from (S3.10) and the independence of the two stages of Markov chain sampling.

Next to prove Theorem 3 (2), note that, we already know that $\widehat{V}$ is a consistent BM estimator of V. From (S3.13), we have $[\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} [e(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1}$. Applying mean value theorem on $[\nabla L(\boldsymbol{d})]_{j-1}$ and the fact that $\nabla^2 L(\boldsymbol{d}^*)$ is bounded in probability, it follows that $[\hat{e}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}})]_{j-1} - [\hat{e}(\pi;\boldsymbol{a},\boldsymbol{d})]_{j-1} \xrightarrow{\text{a.s.}} 0$.

From (3.8) we know that $\hat{u}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} u(\pi,\pi_1)$. From (3.13) we know $\hat{\eta}^{[f]}(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} E_\pi f$. Since $\hat{v}^{[f]}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d}) = \hat{\eta}^{[f]}(\pi;\boldsymbol{a},\boldsymbol{d})\hat{u}(\pi,\pi_1;\boldsymbol{a},\boldsymbol{d})$, it follows that $\hat{v}^{[f]}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} E_\pi f u(\pi,\pi_1)$. Thus $\nabla h(\hat{v}^{[f]}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}}),\hat{u}(\pi,\pi_1;\boldsymbol{a},\hat{\boldsymbol{d}})) \xrightarrow{\text{a.s.}} \nabla h(E_\pi f u(\pi,\pi_1), u(\pi,\pi_1))$. Thus to prove Theorem 3 (2), we only need to show that $\widehat{\Gamma}_l(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) \xrightarrow{\text{a.s.}} \Gamma_l(\pi;\boldsymbol{a},\boldsymbol{d})$.

If $\boldsymbol{d}$ is known from the assumptions of Theorem 3 (2) and the results in Vats et al. (2015), we know that $\Gamma_l(\pi;\boldsymbol{a},\boldsymbol{d})$ is consistently estimated by its BM estimator $\widehat{\Gamma}_l(\pi;\boldsymbol{a},\boldsymbol{d})$. We now show that $\widehat{\Gamma}_l(\pi;\boldsymbol{a},\hat{\boldsymbol{d}}) - \widehat{\Gamma}_l(\pi;\boldsymbol{a},\boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

From Theorem 2 (2), we know that $\hat{\gamma}_l^{22}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{22}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$. We now show $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$. Letting $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{z})$ by $H(\boldsymbol{z})$, by the mean value theorem, there exists $\boldsymbol{d}^* = t\hat{\boldsymbol{d}} + (1-t)\boldsymbol{d}$ for some $t \in (0,1)$, such that $H(\hat{\boldsymbol{d}}) - H(\boldsymbol{d}) = \nabla H(\boldsymbol{d}^*) \cdot (\hat{\boldsymbol{d}} - \boldsymbol{d})$. For any $j \in \{2, \cdots, k\}$, and $\boldsymbol{z} \in R^{+k-1}$,

$$\frac{\partial H(\boldsymbol{z})}{\partial z_j} = \frac{b_l}{e_l - 1} \left[\sum_{m=0}^{e_l - 1} 2(\bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})) \left(\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \right) \right].$$

Let $\bar{W}_m^{[f]} := \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(v^{[f]}(X; \boldsymbol{a}, \boldsymbol{z}))$ and $\bar{\bar{W}}^{[f]} := \bar{\bar{v}}^{[f]}(\boldsymbol{a}, \boldsymbol{z}) - E_{\pi_l}(v^{[f]}(X; \boldsymbol{a}, \boldsymbol{z}))$. Note that, there exists, $\sigma_f^2 < \infty$ such that $\sqrt{b_l}\bar{W}_m^{[f]} \xrightarrow{d} \mathrm{N}(0, \sigma_f^2)$, and $\sqrt{n_l}\bar{\bar{W}}^{[f]} \xrightarrow{d} \mathrm{N}(0, \sigma_f^2)$. Simple calculations show that

$$\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2} \frac{1}{b_l} \sum_{i=mb_l+1}^{(m+1)b_l} \left[\frac{f(X_i^{(l)}) \nu(X_i^{(l)}) \nu_j(X_i^{(l)})}{\left(\sum_s a_s \nu_s(X_i^{(l)})/z_s \right)^2}. \right]$$

Hence, letting $\alpha_j^{[f]} = E_{\pi_l}[f(X)\nu(X)\nu_j(X)/\left(\sum_s a_s \nu_s(X)/z_s\right)^2]$, we write

$$\frac{\partial \bar{v}_m^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} - \frac{\partial \bar{\bar{u}}^{[f]}(\boldsymbol{a}, \boldsymbol{z})}{\partial z_j} \equiv \frac{a_j}{z_j^2} \left\{ \bar{Z}_{m,j}^{[f]} \right\} - \frac{a_j}{z_j^2} \left\{ \bar{\bar{Z}}_j^{[f]} \right\},$$

where $\bar{Z}_{1,j}^{[f]} = (1/b_l) \sum_{i=1}^{b_l} [f(X_i^{(l)}) \nu(X_i^{(l)}) \nu_j(X_i^{(l)}) / \{\sum_s a_s \nu_s(X_i^{(l)})/z_s\}^2] - \alpha_j^{[f]}$ and $\bar{\bar{Z}}_j^{[f]}$ is similarly defined. Note that, there exists $\tau_{j,f}^2 < \infty$, such that $\sqrt{b_l}\bar{Z}_{m,j} \xrightarrow{d} \mathrm{N}(0, \tau_{j,f}^2)$, and $\sqrt{n_l}\bar{\bar{Z}}_j \xrightarrow{d} \mathrm{N}(0, \tau_{j,f}^2)$. The rest of the proof is analogous to Theorem 2, in that we have

$$\frac{\partial H(\boldsymbol{z})}{\partial z_j} = \frac{a_j}{z_j^2} \frac{2}{e_l - 1} \sum_{m=0}^{e_l - 1} \left[\sqrt{b_l}\bar{W}_m^{[f]} \sqrt{b_l}\bar{Z}_{m,j}^{[f]} \right] - \frac{a_j}{z_j^2} \frac{2}{e_l - 1} \left[\sqrt{n_l}\bar{\bar{W}}^{[f]} \sqrt{n_l}\bar{\bar{Z}}_j^{[f]} \right].$$

Then it can be shown $\hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{11}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$ and finally $\hat{\gamma}_l^{12}(\pi; \boldsymbol{a}, \hat{\boldsymbol{d}}) - \hat{\gamma}_l^{12}(\pi; \boldsymbol{a}, \boldsymbol{d}) \xrightarrow{\text{a.s.}} 0$.

S4 Regeneration with general weights

Tan et al. (2015) provide a regeneration based central limit theorem (CLT) for the estimators $\hat{\eta}$ and $\hat{u}$ defined in 1.3 and 3.1 respectively in the main manuscript. In the case when $\boldsymbol{d}$ is unknown, they allow only a special choice for the weight vector, namely $\boldsymbol{a} = (1, \hat{\boldsymbol{d}})$ for their results to hold, where $\hat{\boldsymbol{d}}$ is the estimator of $\boldsymbol{d}$ based on the Stage 1 chains discussed in Section 2 of the main paper. In this section, we establish a regeneration based CLT for $\hat{\eta}$ and $\hat{u}$ with any choice of the weight vector $\boldsymbol{a}$.

We will refer to the following conditions.

A1 For each $l = 1, \ldots, k$, the Markov chain $\Phi_l = \{X_0^{(l)}, X_1^{(l)}, \ldots\}$ is geometrically ergodic and has π_l as its invariant density.

A2 Let $k_l : \mathsf{X} \times \mathsf{X} \to [0, \infty)$ be the Markov transition density for Φ_l, so that for any measurable set A we have $P\big(X_{n+1}^{(l)} \in A \,|\, X_n^{(l)} = x\big) = \int_A k_l(y|x)\mu(dy)$. Suppose that for each $l = 1, \ldots, k$, k_l satisfies the following *minorization condition*:

$$k_l(y|x) \geq s_l(x)\, q_l(y) \qquad \text{for all } x, y \in \mathsf{X}, \tag{S4.14}$$

where the function $s_l \colon \mathsf{X} \to [0, 1)$ with $E_{\pi_l} s_l > 0$, and q_l is a probability density function on X.

A3 Recall the functions $u(X; \boldsymbol{a}, \boldsymbol{d})$ and $v^{[f]}(X; \boldsymbol{a}, \boldsymbol{d})$ defined in (3.2) of our paper. There exists $\epsilon > 0$ such that $E_{\pi_l}|v^{[f]}(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ and $E_{\pi_l}|u(X; \boldsymbol{a}, \boldsymbol{d})|^{2+\epsilon}$ are finite.

A4 Suppose Φ_l is simulated for R_l regenerative tours for $l = 1, \ldots, k$. Assume $R_l/R_1 \to b_l \in (0, \infty)$ as $R_1 \to \infty$.

Following Tan et al. (2015), let the *regeneration times* for the l^{th} Markov chain be $\tau_0^{(l)} = 0, \tau_1^{(l)}, \tau_2^{(l)}, \ldots$. Accordingly, the chain Φ_l is broken up into "tours" $\big\{\big(X_{\tau_{t-1}^{(l)}}, \ldots, X_{\tau_t^{(l)}-1}\big),\ t = 1, 2, \ldots\big\}$ that are independent stochastic replicas of each other. Suppose we simulate R_l tours of the l^{th} Markov chain for $l = 1, \ldots, k$, so the length of the l^{th} chain is $n_l = \tau_{R_l}^{(l)}$. Also as in Tan et al. (2015), for $t = 1, 2, \ldots, R_l$ define

$$V_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} v^{[f]}(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \quad U_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} u(X_i^{(l)}; \boldsymbol{a}, \boldsymbol{d}), \quad \text{and} \quad T_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} 1 = \tau_t^{(l)} - \tau_{t-1}^{(l)}, \tag{S4.15}$$

where the sums range over the values of i that constitute the t^{th} tour.

Recall from Remark 3 in Section 3 of our paper, when $\boldsymbol{d}$ is unknown, we set $\boldsymbol{a} = \boldsymbol{w} * (1, \hat{\boldsymbol{d}})$ where $*$ denotes component-wise multiplication. That is, $(a_1, \ldots, a_k) = (w_1, w_2, \ldots, w_k) * (1, \hat{d}_2, \ldots, \hat{d}_k)$ for any pre-determined weight $\boldsymbol{w}$. With this choice, the expressions for u and $v^{[f]}$ in (3.2) become

$$u\big(x; \boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) = \frac{\nu(x)}{\sum_{l=1}^k w_l \nu_l(x)} \quad \text{and} \quad v^{[f]}\big(x; \boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) = \frac{f(x)\nu(x)}{\sum_{l=1}^k w_l \nu_l(x)}. \tag{S4.16}$$

The above quantities do not involve $\hat{\boldsymbol{d}}$, and consequently for each l, the triples $\big(V_t^{(l)}, U_t^{(l)}, T_t^{(l)}\big)$, $t = 0, 1, 2, \ldots$ defined in (S4.15) are iid, and we have independence across l's. The estimator for η reduces to

$$\begin{aligned}
\hat{\eta} = \hat{\eta}_{N,n}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) &= \sum_{l=1}^k \frac{w_l \hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k w_l \nu_s(X_i^{(l)})} \Big/ \sum_{l=1}^k \frac{w_l \hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k w_s \nu_s(X_i^{(l)})} \\
&= \sum_{l=1}^k \frac{w_l \hat{d}_l}{n_l} \sum_{t=1}^{R_l} V_t^{(l)} \Big/ \sum_{l=1}^k \frac{w_l \hat{d}_l}{n_l} \sum_{t=1}^{R_l} U_t^{(l)} \\
&= \sum_{l=1}^k w_l \hat{d}_l \frac{\bar{V}^{(l)}}{\bar{T}^{(l)}} \Big/ \sum_{l=1}^k w_l \hat{d}_l \frac{\bar{U}^{(l)}}{\bar{T}^{(l)}},
\end{aligned} \tag{S4.17}$$

where

$$U_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^k w_s \nu_s(X_i^{(l)})} \quad \text{and} \quad V_t^{(l)} = \sum_{i=\tau_{t-1}^{(l)}}^{\tau_t^{(l)}-1} \frac{f(X_i^{(l)})\nu(X_i^{(l)})}{\sum_{s=1}^k w_s \nu_s(X_i^{(l)})},$$

$\bar{T}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} T_t^{(l)}$ be the average tour length and, analogously, $\bar{V}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} V_t^{(l)}$ and

$\bar{U}^{(l)} = R_l^{-1} \sum_{t=1}^{R_l} U_t^{(l)}$. Similarly, the estimator for m/m_1 reduces to

$$\begin{aligned} \hat{u} = \hat{u}_{N,n}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) &= \sum_{l=1}^{k} \frac{w_l \hat{d}_l}{n_l} \sum_{i=1}^{n_l} \frac{\nu(X_i^{(l)})}{\sum_{s=1}^{k} w_s \nu_s(X_i^{(l)})} \\ &= \sum_{l=1}^{k} \frac{w_l \hat{d}_l}{n_l} \sum_{t=1}^{R_l} U_t^{(l)} = \sum_{l=1}^{k} w_l \hat{d}_l \frac{\bar{U}^{(l)}}{\bar{T}^{(l)}}. \end{aligned} \tag{S4.18}$$

Theorem 4 below gives the asymptotic distributions of $\hat{\eta}$ and $\hat{u}$. It extends Tan et al.'s (2015) Theorem 2 to the general choice of weight vector $\boldsymbol{a}$. To state the theorem, we first need to define some notation. Let $\tilde{M}$ and $\tilde{L}$ be the vectors of length $k-1$ for which the $(j-1)^{\text{th}}$ coordinates are, for $j = 2, \ldots, k$,

$$\begin{aligned} &\tilde{M}_{j-1} = w_j E_{\pi_j} u \text{ and} \\ &\tilde{L}_{j-1} = \frac{w_j E_{\pi_j} v^{[f]}}{\sum_{l=1}^{k} w_l d_l E_{\pi_l} u} - \frac{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} v^{[f]}\big)\big(w_j E_{\pi_j} u\big)}{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} u\big)^2}. \end{aligned} \tag{S4.19}$$

As in Tan et al. (2015), assume that in Stage 1, for $l = 1, \ldots, k$, the l^{th} chain has been run for ρ_l regenerations. So the length of the l^{th} chain, $N_l = T_1^{(l)} + \ldots + T_{\rho_l}^{(l)}$, is random. We assume that $\rho_1, \ldots, \rho_k \to \infty$ in such a way that $\rho_l/\rho_1 \to c_l \in (0, \infty)$, for $l = 1, \ldots, k$.

Theorem 4 *Suppose that for the Stage* 1 *chains, conditions A1 and A2 hold, and that for the Stage* 2 *chains, conditions A1–A4 hold. If* $\rho_1 \to \infty$ *and* $R_1 \to \infty$ *in such a way that* $R_1/\rho_1 \to q \in [0, \infty)$*, then*

$$R_1^{1/2}\big(\hat{u} - m/m_1\big) \xrightarrow{d} \mathcal{N}\big(0, q\tilde{M}^\top W \tilde{M} + \kappa^2\big)$$

and

$$R_1^{1/2}\big(\hat{\eta} - \eta\big) \xrightarrow{d} \mathcal{N}\big(0, q\tilde{L}^\top W \tilde{L} + \tau^2\big),$$

where $\tilde{M}$*,* $\tilde{L}$ *are given in equations* (S4.19)*,* W*,* κ^2 *and* τ^2 *are given in equations (2.15), (2.8), and (2.10) of Tan et al. (2015), respectively. In their (2.8) and (2.10),* $\boldsymbol{a}$ *is taken to be* $\boldsymbol{a} = \boldsymbol{w} * (1, \boldsymbol{d})$*. Furthermore, we can form strongly consistent estimates of the asymptotic variances if we use* $\widehat{W}$*,* $\hat{\kappa}^2$*, and* $\hat{\tau}^2$ *defined in (2.16) and (2.11) of Tan et al. (2015), respectively, and use the standard empirical estimates of* $\tilde{M}$ *and* $\tilde{L}$*.*

S4.1 Proof of Theorem 4

We first prove the CLT for $\hat{\eta}$. Note that

$$R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \eta\big] = R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \hat{\eta}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big)\big] + R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big) - \eta\big]. \tag{S4.20}$$

The second term on the right side of (S4.20) involves randomness coming only from Stage 2 sampling, and its distribution is given by Theorem 1 of Tan et al. (2015): it is asymptotically normal with mean 0 and variance τ^2. The first term involves randomness from both Stage 1 and Stage 2 sampling. However, as in the proofs of Theorem 2 and 3, we can show that for this term, the randomness from Stage 2 is asymptotically negligible, so that only Stage 1 sampling contributes to its asymptotic distribution. Finally, the asymptotic normality of the left side of (S4.20) follows since the two stages of sampling are independent. We now provide the details of the proof.

Consider the first term on the right side of (S4.20). Recall that if $\boldsymbol{a} = \boldsymbol{w} * (1, \boldsymbol{d})$, then

$$v^{[f]}(x) := v^{[f]}(x; \boldsymbol{a}, \boldsymbol{d}) = \frac{f(x)\nu(x)}{\sum_{l=1}^k w_l \nu_l(x)} \quad \text{and} \quad u(x) := u(x; \boldsymbol{a}, \boldsymbol{d}) = \frac{\nu(x)}{\sum_{l=1}^k w_l \nu_l(x)}.$$

With (S4.17) and (S4.18) in mind, define the function

$$A(\boldsymbol{z}) = \hat{\eta}\big(\boldsymbol{w} * (1, \boldsymbol{z}), \boldsymbol{z}\big) = \sum_{l=1}^k \frac{w_l z_l}{n_l} \sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)}) \Big/ \sum_{l=1}^k \frac{w_l z_l}{n_l} \sum_{i=1}^{n_l} u(X_i^{(l)})$$

for $\boldsymbol{z} = (z_2, \ldots, z_k)^\top$, with $z_l > 0$ for $l = 2, \ldots, k$, and $z_1 = 1$. Note that setting $\boldsymbol{z} = \boldsymbol{d}$ gives $A(\boldsymbol{d}) = \hat{\eta}(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d})$, and setting $\boldsymbol{z} = \hat{\boldsymbol{d}}$ gives $A(\hat{\boldsymbol{d}}) = \hat{\eta}(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}})$.

By a Taylor series expansion of A about $\boldsymbol{d}$ we get

$$\begin{aligned} R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w}*(1,\hat{\boldsymbol{d}}),\hat{\boldsymbol{d}}\big) - \hat{\eta}\big(\boldsymbol{w}*(1,\boldsymbol{d}),\boldsymbol{d}\big)\big] &= R_1^{1/2}\nabla A(\boldsymbol{d})^\top(\hat{\boldsymbol{d}}-\boldsymbol{d}) + \frac{R_1^{1/2}}{2}(\hat{\boldsymbol{d}}-\boldsymbol{d})^\top \nabla^2 A(\boldsymbol{d}^*)(\hat{\boldsymbol{d}}-\boldsymbol{d}) \\ &= R_1^{1/2}\nabla A(\boldsymbol{d})^\top(\hat{\boldsymbol{d}}-\boldsymbol{d}) + \frac{R_1^{1/2}}{2\rho_1}\big(\rho_1^{1/2}(\hat{\boldsymbol{d}}-\boldsymbol{d})\big)^\top \nabla^2 A(\boldsymbol{d}^*)\big(\rho_1^{1/2}(\hat{\boldsymbol{d}}-\boldsymbol{d})\big), \end{aligned}$$

where $\boldsymbol{d}^*$ is between $\boldsymbol{d}$ and $\hat{\boldsymbol{d}}$. As $R_1 \to \infty$, $n_l \to \infty$ for each l. We first show that the gradient $\nabla A(\boldsymbol{d})$ converges almost surely to a finite constant vector by proving that each one of its components, $[A(\boldsymbol{d})]_{j-1}$, $j = 2, \ldots, k$, converges almost surely as $R_1 \to \infty$. As $n_l \to \infty$ for $l = 1, \ldots, k$, for $j = 2, \ldots, k$, we have

$$\begin{aligned} [\nabla A(\boldsymbol{d})]_{j-1} &= \frac{(w_j/n_j)\sum_{i=1}^{n_j} v^{[f]}(X_i^{(j)})}{\sum_{l=1}^k (w_l d_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)})} \\ &\quad - \frac{\big(\sum_{l=1}^k (w_l d_l/n_l)\sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)})\big)\big((w_j/n_j)\sum_{i=1}^{n_j} u(X_i^{(j)})\big)}{\big(\sum_{l=1}^k (w_l d_l/n_l)\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2} \\ &\xrightarrow{\text{a.s.}} \frac{w_j E_{\pi_j} v^{[f]}}{\sum_{l=1}^k w_l d_l E_{\pi_l} u} - \frac{\big(\sum_{l=1}^k w_l d_l E_{\pi_l} v^{[f]}\big)\big(w_j E_{\pi_j} u\big)}{\big(\sum_{l=1}^k w_l d_l E_{\pi_l} u\big)^2}. \end{aligned}$$

The expression above corresponds to $\tilde{L}_{j-1}$, which is defined in (S4.19), and it is finite by assumption $A3$. Next, we show that the random Hessian matrix $\nabla^2 A(\boldsymbol{d}^*)$ is bounded in probability, i.e., each element of this matrix is $O_p(1)$. As $n_l \to \infty$ for $l = 1, \ldots, k$, for any $j, t \in \{2, \ldots, k\}, j \neq t$, we have

$$[\nabla^2 A(\boldsymbol{d}^*)]_{t-1,j-1} = -\frac{\big(\frac{w_j}{n_j}\sum_{i=1}^{n_j} v^{[f]}(X_i^{(j)})\big)\big(\frac{w_t}{n_t}\sum_{i=1}^{n_t} u(X_i^{(t)})\big)}{\big(\sum_{l=1}^{k}\frac{w_l d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2}$$

$$- \left(\frac{w_j}{n_j}\sum_{i=1}^{n_j} u(X_i^{(j)})\right)\left[\frac{\frac{w_t}{n_t}\sum_{i=1}^{n_t} v^{[f]}(X_i^{(t)})}{\big(\sum_{l=1}^{k}\frac{w_l d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^2} - 2\frac{\big(\sum_{l=1}^{k}\frac{w_l d_l^*}{n_l}\sum_{i=1}^{n_l} v^{[f]}(X_i^{(l)})\big)\big(\frac{w_t}{n_t}\sum_{i=1}^{n_t} u(X_i^{(t)})\big)}{\big(\sum_{l=1}^{k}\frac{w_l d_l^*}{n_l}\sum_{i=1}^{n_l} u(X_i^{(l)})\big)^3}\right]$$

$$\xrightarrow{\text{a.s.}} -\frac{(w_j E_{\pi_j} v^{[f]})(w_t E_{\pi_t} u)}{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} u\big)^2} - (w_j E_{\pi_j} u)\left[\frac{w_t E_{\pi_t} v^{[f]}}{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} u\big)^2} - 2\frac{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} v^{[f]}\big)(w_t E_{\pi_t} u)}{\big(\sum_{l=1}^{k} w_l d_l E_{\pi_l} u\big)^3}\right],$$

where the limits are also finite.

Now, we can rewrite (S4.20) as

$$\begin{aligned} R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \eta\big] &= (R_1/\rho_1)^{1/2}\nabla A(\boldsymbol{d})^\top \rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big) - \eta\big] \\ &\quad + \frac{1}{2\rho_1^{1/2}}(R_1/\rho_1)^{1/2}\big[\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d})\big]^\top \nabla^2 A(\boldsymbol{d}^*)\big[\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d})\big] \\ &= q^{1/2}[\nabla A(\boldsymbol{d})]^\top \rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) + R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big) - \eta\big] + o_p(1). \end{aligned}$$

Since from Tan et al. (2015) we have $\rho_1^{1/2}(\hat{\boldsymbol{d}} - \boldsymbol{d}) \xrightarrow{d} \mathcal{N}(0, W)$ and the two sampling stages are assumed to be independent, we conclude that

$$R_1^{1/2}\big[\hat{\eta}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \eta\big] \xrightarrow{d} \mathcal{N}\big(0, q\tilde{L}^\top W \tilde{L} + \tau^2\big).$$

The proof of the CLT for $\hat{u}$ is similar. As in (S4.20), we have

$$\begin{aligned} R_1^{1/2}\big[\hat{u}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - m/m_1\big] &= R_1^{1/2}\big[\hat{u}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \hat{u}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big)\big] \\ &\quad + R_1^{1/2}\big[\hat{u}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big) - m/m_1\big]. \end{aligned} \tag{S4.21}$$

The asymptotic distribution of the second term in (S4.21) is given in Tan et al.'s (2015) Theorem 1. The first term is linear in $\hat{\boldsymbol{d}} - \boldsymbol{d}$:

$$\hat{u}\big(\boldsymbol{w} * (1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - \hat{u}\big(\boldsymbol{w} * (1, \boldsymbol{d}), \boldsymbol{d}\big) = \sum_{j=2}^{k} w_j \left(\frac{1}{n_j}\sum_{i=1}^{n_j} u(X_i^{(j)})\right)(\hat{d}_j - d_j). \tag{S4.22}$$

For $j = 2, \ldots, k$, the coefficient of $(\hat{d}_j - d_j)$ in (S4.22) converges almost surely to $w_j E_{\pi_j} u$, which is the term $\tilde{M}_{j-1}$ defined in (S4.19).

Finally, from the independence of the two terms in (S4.21) we conclude that

$$R_1^{1/2}\big[\hat{u}\big((1, \hat{\boldsymbol{d}}), \hat{\boldsymbol{d}}\big) - m/m_1\big] \xrightarrow{d} \mathcal{N}\big(0, q\tilde{M}^\top W \tilde{M} + \kappa^2\big).$$

S5 Toy example

In this section, we follow up on the simulation studies that involve t distributions from Section 4 of the main paper to verify Theorems 1-3. We also discuss different weights in forming generalized IS estimators and their effects on estimates of expectations and ratios of normalizing constants.

Let $t_{r,\mu}$ denote the t-distribution with degree of freedom r and central parameter μ. We consider $\pi_1(\cdot)$ and $\pi_2(\cdot)$ as the density functions for a $t_{5,\mu_1=1}$ and $t_{5,\mu_2=0}$, respectively. For simplicity, let $\nu_i(\cdot) = \pi_i(\cdot)$ for $i = 1, 2$. Our plan is to first estimate the ratio between the two normalizing constants, $d = m_2/m_1$. Then we will study a sea of t-distributions $\Pi = \{t_{5,\mu} : \mu \in M\}$ where M is a fine grid over $[0, 1]$, say $M = \{0, .01, \cdots, .99, 1\}$. For each $\mu \in M$, we assume that $\nu_\mu(\cdot) = \pi_\mu(\cdot)$ and we estimate the ratio between its normalizing constant and m_1, denoted by $d_\mu := \frac{m_\mu}{m_1}$. We also estimate the expectation of each distribution in Π, denoted $E_{t_{5,\mu}}X$ or $E_\mu X$ for short. Clearly, the exact answers are $d = d_\mu = 1$ and $E_\mu X = \mu$ for any $\mu \in M$. Nevertheless, we follow the two-stage procedure from Sections 2 and 3 to generate Markov chains from π_1 and π_2 and build MCMC estimators from Theorems 1-3. The primary goal is to compare the performance of BM and RS estimators.

We draw iid samples from π_1 and Markov chain samples from π_2 using the so called independent Metropolis Hastings algorithm with proposal density $t_{5,1}$. For RS, we follow the idea of Mykland et al. (1995, Section 4.1) on constructing minorization conditions to identify regeneration times. Based on a carefully tuned minoration condition, the Markov chain for π_2 regenerates about every 3 iterations on average. In contrast, for users of the BM method proposed in this paper, no such theoretical development is needed. For $i = 1, 2$ we draw N_i observations from π_i in stage 1 and n_i observations from π_i in stage 2. We set $N_1 = N_2$ and $n_1 = n_2 = N_1/10 = N_2/10$. Recall the reason for smaller stage 2 sample sizes is due to computing cost. For completeness, note generating Markov chain samples using RS results in a random chain length so these chains were run in such a way that $N_1 \sim N_2$ and $n_1 \sim n_2$.

For estimators based on stage 1 samples, Theorem 1 allows any choice of weight, $\boldsymbol{a}^{[1]}$. For estimators based on stage 2 samples, Theorem 2 and 3 allow any choice of weight, $\boldsymbol{a}^{[2]}$, in constructing consistent BM estimators of the asymptotic variances. RS based estimators in stage 2 are calculated using Theorems stated in D&T and Tan et al. (2015) with a general weight choice noted in Remark 4. This is an important generalization in that now any non-negative numerical weight vector can be used. We discuss the choice of weights and their impact on the estimators later in this section.

The following details the simulation study presented in the main article. We consider

increasing sample sizes from $N_1 = 10^3$ to 10^5 in order to examine trace plots for BM and RS estimates. The two stage procedure is repeated 1000 times independently. The unknown true value of the asymptotic variance of $\hat{d}$ is estimated by its empirical asymptotic variance over the 1000 replications at $N_1 = 10^5$. We consider the naive weight, $\boldsymbol{a}^{[1]} = (0.5, 0.5)$, that is proportional to the sample sizes, and an alternative $\boldsymbol{a}^{[1]} = (0.82, 0.18)$ that weighs the iid sample more than the Markov chain sample. As illustrated in Figure 1 of the main article, both the BM and the RS estimates approach the empirical asymptotic variance as the sample size increases suggesting consistency. Similarly for stage 2, Figure 1 shows convergence of the BM and the RS estimates to the corresponding empirical asymptotic variances of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$. Plots for other $\mu \in M$ show similar results, but are not included here.

Overall, the simulation study suggests BM and RS methods provide consistent estimators for the true asymptotic variance. RS estimators enjoy smaller mean squared error in most cases. Nevertheless, when the number of regenerations is not great, BM estimators could be the more stable estimator. For example, in the top left panel of Figure 1, at stage 2 sample size $n_2 \approx n_1 = 100$, or about 35 regenerations for chain 2, the RS method substantially over-estimated the target in about 5% of the replications. Further, in the cases where regeneration is unavailable or the number of regenerations is extremely small, then BM would be the more viable estimator.

S5.1 Choice of stage 1 weights

For stage 1, we recommend obtaining a close-to-optimal weight $\hat{\boldsymbol{a}}^{[1,\text{opt}]}$ using a pilot study described in D&T. In short, one can generate samples of small size from π_1 and π_2, estimate $\hat{d}$ and its asymptotic variance based on Theorem 1 for a grid of weights, and then identify the weight that minimizes the estimated variance. With a small pilot study based on samples of size 1000 from both distributions, we obtained $\hat{\boldsymbol{a}}^{[1,\text{opt}]} = (0.82, 0.18)$. As depicted by the horizontal lines accross the pictures in Figure 1 of the main article, the asymptotic variance of the estimator $\hat{d}$ based on $\hat{\boldsymbol{a}}^{[1,\text{opt}]}$ is approximately 0.07, which is more than 30% smaller than of the estimator based on the naive choice $\boldsymbol{a}^{[1]} = (.5, .5)$. Note that the naive weight is proportional to the sample sizes from π_1 and π_2, which is asymptotically optimal if both samples were independent. However, since sample 2 is from a Markov chain sample, using a weight that appropriately favors the independent sample has lead to smaller error in the estimator. The gain in efficiency using a close-to-optimal weight will be more pronounced if the difference in the mixing rates of the two samples is larger.

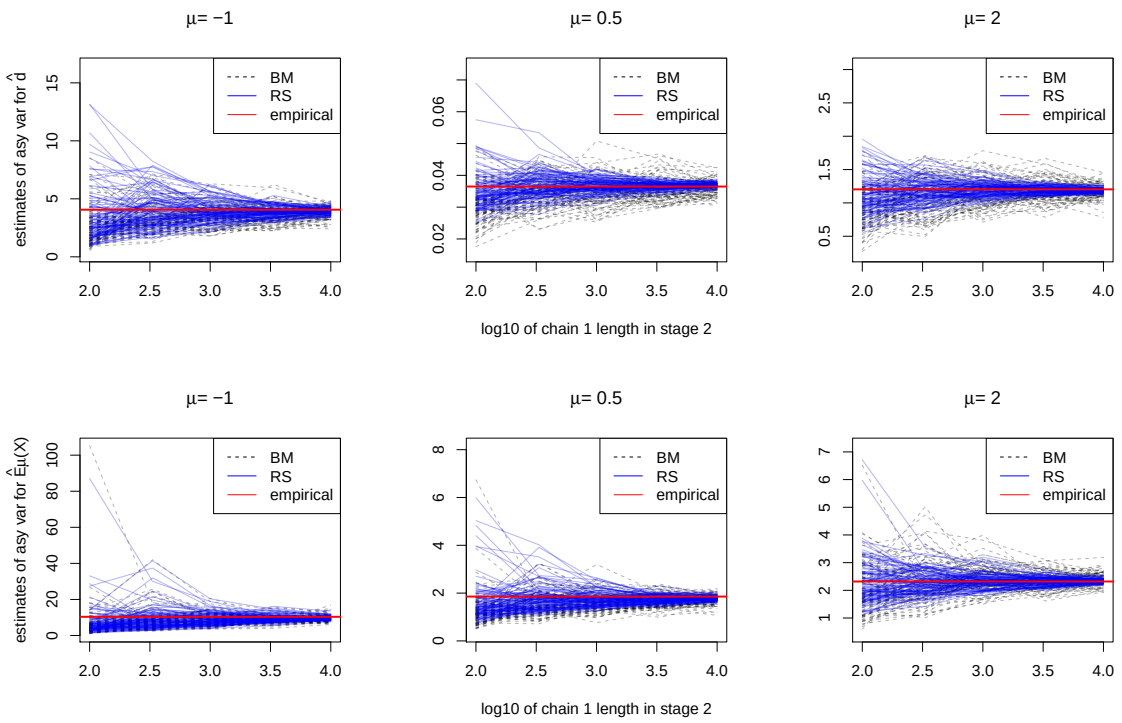

Figure 1: Estimates of the asymptotic variance of $\hat{d}_\mu$ (upper panels) and $\hat{E}_\mu(X)$ (lower panels) in stage 2, with naive weight $\boldsymbol{a}^{[2]} = (.5, .5)$.

S5.2 Choice of stage 2 weights

In stage 2, for each $\mu \in M$, the asymptotic variance of $\hat{d}_\mu$ and $\hat{E}_\mu(X)$ are minimized at different weights. Instead of searching for each of the $2|M|$ optimal weights in a pilot study, it is more practical to set sub-optimal weights using less costly strategies. Below, we perform a simulation study to examine three simple weighting strategies:

1. naive: $\boldsymbol{a}^{[2]} \propto (n_1, n_2)$,
2. inverse distance (inv-dist): $\boldsymbol{a}^{[2]}(\mu) \propto \left(\frac{n_1}{|\mu-\mu_1|}, \frac{n_2}{|\mu-\mu_2|}\right)$,
3. effective sample size (ess) by inverse distance (inv-dist): $\boldsymbol{a}^{[2]}(\mu) \propto \left(\frac{\text{ess}_1}{|\mu-\mu_1|}, \frac{\text{ess}_2}{|\mu-\mu_2|}\right)$.

Using each of the three strategies, we construct generalized IS estimators for d_μ and $E_\mu(X)$ for a grid of μ values between -1.5 and 4. Note that samples are drawn from two reference distributions indexed by $\mu = 1$ and $\mu = 0$ respectively. Hence our simulation study concerns

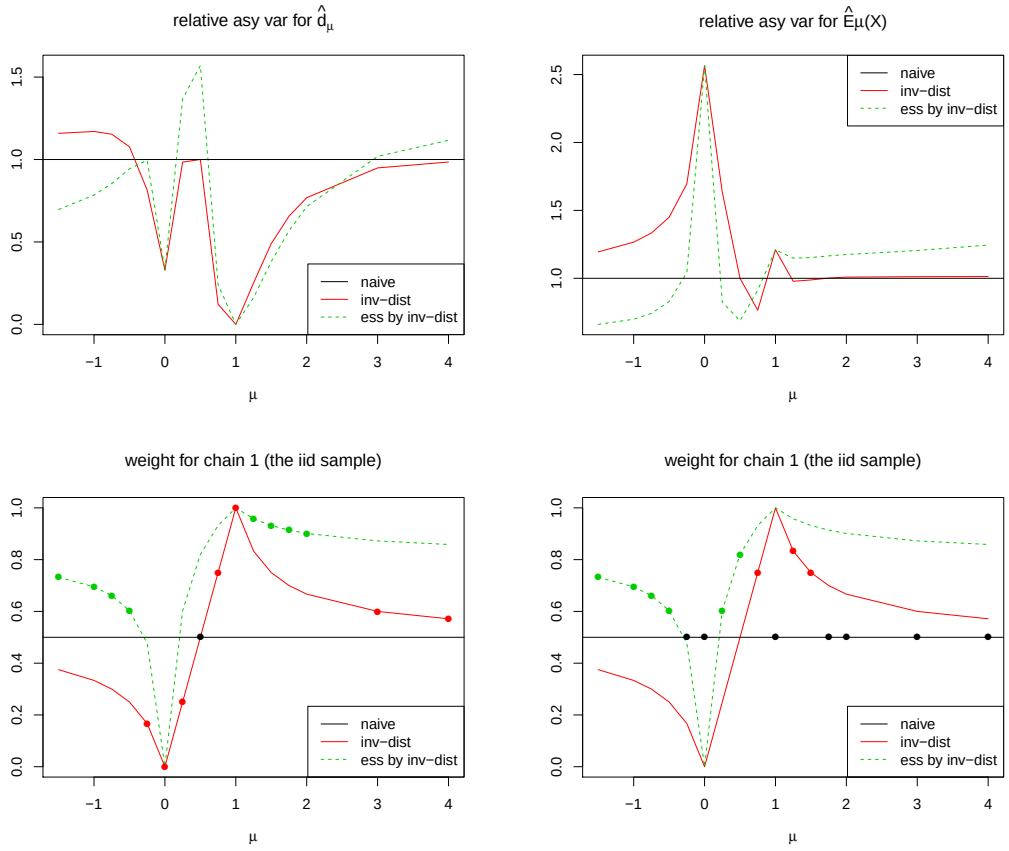

Figure 2: Comparisons of three weight strategies in terms of the asymptotic variance of the corresponding estimators $\hat{d}_\mu$ and $\hat{E}_\mu(X)$. The solid dots show which strategy achieves the smallest asymptotic variance among the three at any given μ (ties awarded to the more basic strategy).

both interpolation and extrapolation. A summary of their performance is provided in Figure 2, and detailed results for selected simulation setups are shown in Figures 1, 3, and 4 for strategies 1, 2, and 3, respectively. Figure 2 suggests that none of the three strategies is uniformly better than the others. In particular, we observe the following.

1. **For estimating** d_μ

   (a) For $\mu \in (0, 1)$, strategy 2 works the best.

   (b) For $\mu = 0$, strategies 2 and 3 work better than strategy 1. Indeed, both of them simply set their stage 2 estimates $\hat{d}_0$ to be the stage 1 estimate, $\hat{d}$. This would

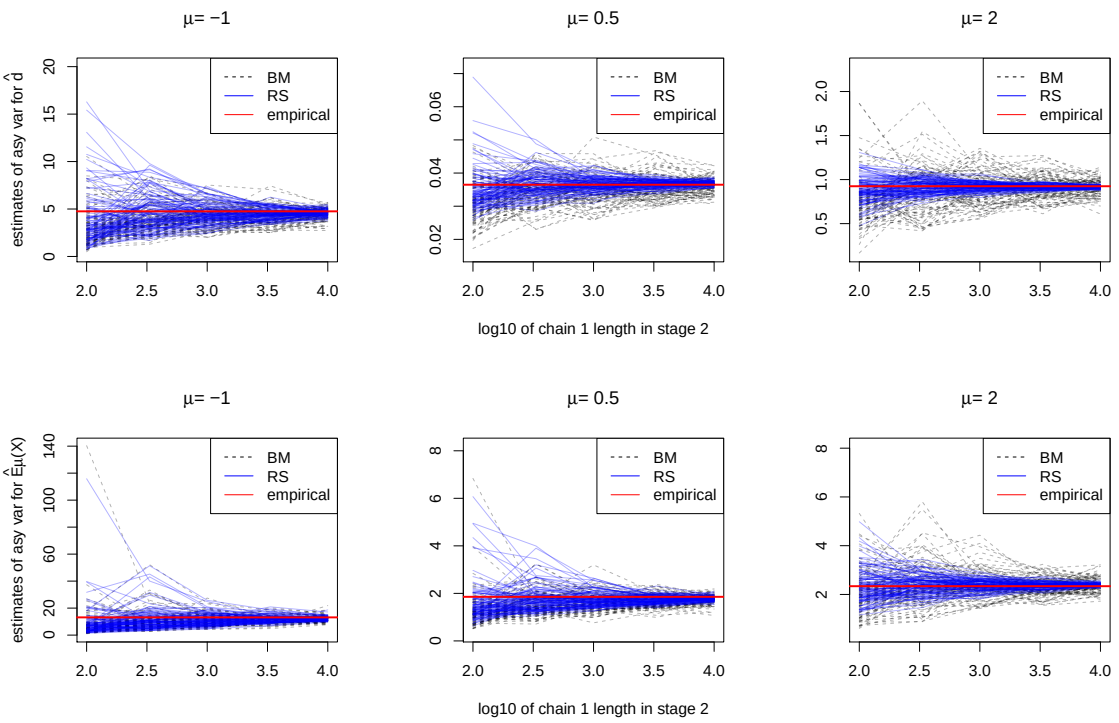

Figure 3: Estimates of the asymptotic variance of $\hat{d}_\mu$ (upper panels) and $\hat{E}_\mu(X)$ (lower panels) in stage 2, with weight $\boldsymbol{a}^{[2]}(\mu)$ chosen by strategy 2.

be a better choice than strategy 1 because in a two-step procedure, stage 1 chains are often much longer than stage 2 chains, and hence $\hat{d}$ is already a very accurate estimate for $d_0 = d$.

(c) For $\mu \notin [0, 1]$, strategies 2 and 3 generally lead to more stable estimates of d_μ. However, all strategies lead to very large asymptotic variances for $\mu < 0$. Hence, one needs to be mindful when doing extrapolation with IS estimators — always obtain an estimate of the standard error, or reconsider the placement of the reference points.

2. **For estimating** $E_\mu(X)$

(a) For $\mu \in (0, 1)$, strategy 2 works the best in general, while strategy 3 is very unstable.

(b) For either $\mu = 0$ or 1, strategy 2 and 3 are the same, and they only utilize the reference chain from μ. This was a wise choice for estimating d_μ as explained

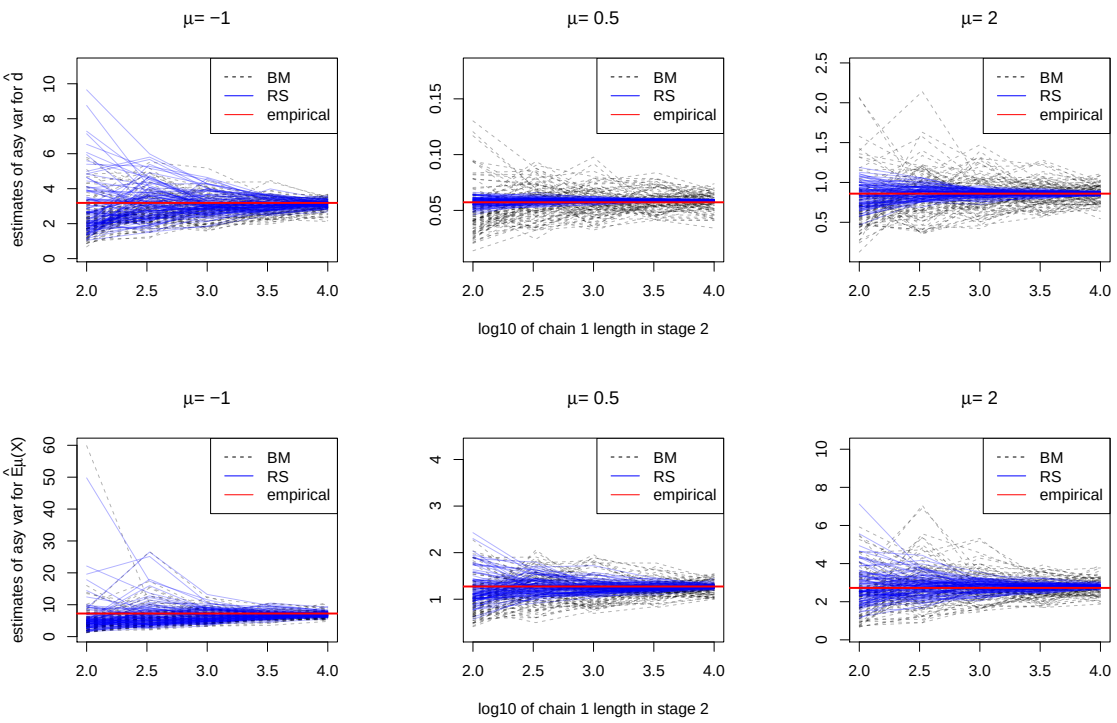

Figure 4: Estimates of the asymptotic variance of $\hat{d}_\mu$ (upper panels) and $\hat{E}_\mu(X)$ (lower panels) in stage 2, with naive weight $\boldsymbol{a}^{[2]}(\mu)$ chosen by strategy 3.

before, but not so for other quantities of interest.

(c) For $\mu \notin [0, 1]$, all strategies lead to fairly large asymptotic variances, especially for $\mu < 0$.

Overall in stage 2, strategy 2 has an advantage when the estimands are ratios between normalizing constants. However, when estimating $E_\mu(X)$, the situation is more complicated. Our impression is that assigning any extreme weight will lead to high variability in the estimator. So it is reasonable to simply use the naive weight, or other strategies that bound the weights away from 0 and 1.

S6 Bayesian variable selection models

Here, we consider a class of Bayesian variable selection (BVS) models for linear regression with independent normal priors on the regression coefficients. This model involves a 2-dimensional

prior hyperparameter that influences inference, yet no default choice guarantees good performance in practice. Hence, displaying the effect of different hyperparameter values on the posterior distribution would greatly benefit users of the model. When the number of predictors, q, is large, the computing is challenging. Our solution is to obtain MCMC samples for a small number of models with different hyperparameter values, based on which generalized IS estimates can be obtained for BFs and other posterior expectations for a large number of models. Again, an important problem in practice is how long the Markov chains need to be run? In this context, the only affordable method that we are aware of is to estimate the SE of these IS estimators using the proposed BM method.

As introduced by Mitchell and Beauchamp (1988), let $Y = (Y_1, \ldots, Y_m)^\top$ denote the vector of responses and $X_1, \ldots, X_q$ denote q potential predictors, each a vector of length m. The predictors are standardized, so that for $j = 1, \ldots, q$, $1_m^T X_j = 0$ and $X_j^T X_j = m$, where 1_m is the vector of m 1's. The BVS model is given by:

$$\text{given } \gamma, \sigma^2, \beta_0, \beta_\gamma, \qquad Y \sim \mathcal{N}_m(1_m\beta_0 + X_\gamma\beta_\gamma, \sigma^2 I), \tag{S6.1a}$$

$$\text{given } \gamma, \sigma^2, \beta_0, \qquad \beta_j \overset{\text{ind}}{\sim} \mathcal{N}\big(0, \frac{\gamma_j}{\lambda}\sigma^2\big) \text{ for } j = 1, \ldots, q, \tag{S6.1b}$$

$$\text{given } \gamma, \qquad (\sigma^2, \beta_0) \sim p(\beta_0, \sigma^2) \propto 1/\sigma^2, \tag{S6.1c}$$

$$\gamma \sim p(\gamma) = w^{q_\gamma}(1-w)^{q-q_\gamma}. \tag{S6.1d}$$

The binary vector $\gamma = (\gamma_1, \ldots, \gamma_q)^\top \in \{0,1\}^q$ identifies a subset of predictors, such that X_j is included in the model if and only if $\gamma_j = 1$, and $|\gamma| = \sum_{j=1}^q \gamma_j$ denotes the number of predictors included. So (S6.1a) says that each γ corresponds to a model given by $Y = 1_m\beta_0 + X_\gamma\beta_\gamma + \epsilon$, where X_γ is an $n \times |\gamma|$ sub-matrix of X that consists of predictors included by γ, β_γ is the vector that contains corresponding coefficients, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. It is sometimes more convenient to use the notation, $Y = X_{0\gamma}\beta_{0\gamma} + \epsilon$, where $X_{0\gamma}$ has one more column of 1's than X_γ and $\beta_{0\gamma}^T = (\beta_0, \beta_\gamma^T)$. Unknown parameters are $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ for which we set a hierarchical prior in (S6.1b) to (S6.1d). In (S6.1d), an independent Bernoulli prior is set for γ, where $w \in (0,1)$ is a hyperparameter that reflects the prior inclusion probability of each predictor. In (S6.1c), a non-informative prior is set for (σ^2, β_0). In (S6.1b), an independent normal prior is assigned to β_γ, where $\lambda > 0$ is a second hyperparameter, that controls the precision of the prior. Overall, θ is given an improper prior due to (S6.1c) but the posterior of θ is indeed proper.

One can actually integrate out $(\beta_\gamma, \beta_0, \sigma^2)$ and arrive at the following model with param-

eter γ only:

$$\begin{aligned} Y|\gamma \sim \ell_h(\gamma; Y) &= \int_{\mathbb{R}_+}\int_{\mathbb{R}}\int_{\mathbb{R}^{|\gamma|}} f(Y|\gamma,\sigma^2,\beta_0,\beta_\gamma) f(\beta_\gamma|\gamma,\sigma^2,\beta_0) f(\sigma^2,\beta_0) d\beta_\gamma d\beta_0 d\sigma^2 \\ &= c_m\, \lambda^{\frac{|\gamma|}{2}} \left|A_{0\gamma}\right|^{-\frac{1}{2}} \left[(Y-\overline{Y})^T(Y-\overline{Y}) - \widetilde{\beta}_\gamma^T A_{0\gamma} \widetilde{\beta}_\gamma\right]^{-(m-1)/2}, \qquad \text{(S6.2)} \\ \gamma \sim p_h(\gamma) &= w^{q_\gamma}(1-w)^{q-q_\gamma}. \end{aligned}$$

Here, c_m is a constant depending only on the sample size m. Further, $A_{0\gamma} = X_{0\gamma}^T X_{0\gamma} + \Lambda_{0\gamma}$, where $\Lambda_{0\gamma}$ is a diagonal matrix, the main diagonal of which is the $(1+|\gamma|)$-dimensional vector $(0, \lambda, \cdots, \lambda)$. Finally, $\widetilde{\beta}_\gamma = A_{0\gamma}^{-1} X_{0\gamma}^T Y$.

Using the model at (S6.2) requires specification of the hyperparameter $h = (w, \lambda)$. Smaller w values assign high prior probabilities to models with fewer predictors, and priors with smaller λ values allow selected predictors to have large coefficients. It is common to set $w = 0.5$ (a uniform prior on the model space) and $\lambda = 1$ (a unit information prior for uncorrelated predictors, see e.g. Kass and Raftery (1995)). One can also choose h adaptively, say according to the marginal likelihood $m_h = \sum_\gamma \ell_h(\gamma; Y) p_h(\gamma)$. A small value of m_h indicates that the prior p_h is not compatible with the observed data, while $h_{\text{EB}} = \arg\max m_h$ is defined to be the empirical Bayesian choice of h. The empirical Bayes idea has been successfully applied to various models with variable selection components (see e.g. George and Foster (2000); Yuan and Lin (2005)). However, we have not seen this idea being carried out for the model in (S6.1), except where $n = p$ and the design matrices are orthogonal (Johnstone and Silverman (2005); Clyde and George (2000)). Due to the improper prior in (S6.1d), m_h is not uniquely defined. Nevertheless, the Bayes factor among any two models, say $m_h/m_{h'}$, is well-defined because the same improper prior is assigned to the shared parameters of the two models (see e.g. Kass and Raftery (1995, sec.5) and Liang et al. (2008, sec.2)).

Here, we concentrate on two goals. The first is to evaluate $\{m_h/m_{h_1}, h \in \mathcal{H}\}$, the marginal likelihood of model h relative to a reference model h_1, which allows us to identify the empirical Bayesian choice of h. The second is to evaluate the posterior mean of the vector of coefficients β for each $h \in \mathcal{H}$, which we denote by $\boldsymbol{b}_h$. Predictions can then be made for new observations using $Y^{(\text{new})} = (x^{(\text{new})})^T \boldsymbol{b}_h$.

For model (S6.2) with a fixed h, a Metropolis Hastings random-swap algorithm (Clyde et al. (2011)) can be used to generate Markov chains of γ from its posterior distribution. In each iteration, with probability $\rho(\gamma)$, we propose flipping a random pair of 0 and 1 in γ, and with probability $1-\rho(\gamma)$, we propose changing γ_j to $1-\gamma_j$ for a random j while leaving other coordinates untouched. We set $\rho(\gamma) = 0$ when γ corresponds to the null model or the full model, and $\rho(\gamma) = .5$ otherwise. Finally, the proposal is accepted with an appropriate

probability. Since this Markov chain lies on a finite state space, it is uniformly ergodic and hence polynomially ergodic as well. Further, moment conditions in Theorems 2 and 3 are satisfied because they reduce to summations of 2^q terms, a large but finite number. To achieve the goal, we generate Markov chains of γ with respect to model (S6.2) at several h values that scatters in $\mathcal{H}$, from which we build generalized IS estimators, and estimate their standard errors.

S6.1 Cookie dough data

We demonstrate the aforementioned sensitivity analysis using the biscuit dough dataset (Osborne et al. (1984); Brown et al. (2001)). The dataset, available in the `R` package `ppls` (Kraemer and Boulesteix (2012)), contains a training set of 39 observations and a test set of 31 observations. These data were obtained from a near-infrared spectroscopy experiment that study the composition of biscuit dough pieces. For each biscuit, the reflectance spectrum is measured at 700 evenly spaced wavelengths. We use these measurements as covariates to predict the response variable, the percentage of water in each dough. We follow previous studies (Hans (2011)) and thin the spectral to $q = 50$ evenly spaced wavelengths.

Figure 5 provides a general picture of the sensitivity analysis. The left plots provide two ways to visualize estimates of the BFs. To form the plot, we took the 12 reference values of $h = (w, \lambda)$ to be such that $(w, -\log(\lambda)) \in \{0.1, 0.2, 0.3, 0.4\} \times \{1, 3, 5\}$. In stage 1 we ran each of the 12 Markov chains at the above values of h for 10^5 iterations to obtain $\hat{\boldsymbol{d}}$. In stage 2, we ran the same 12 Markov chains for $50,000$ iterations each, to form the estimates $\hat{u}_n$ over a fine grid that consists of 475 different h values, with the w component ranging from 0.05 to 0.5 in increments of 0.025 and the $-\log(\lambda)$ component ranging from 0 to 6 in increments of 0.25.

How trustworthy are these BF estimates? Their estimated standard errors are obtained using the BM method, based on Theorem 2. We choose to display the relative SE with respect to the BF estimates, as shown in the upper right panel of Figure 5. The relative SEs are smaller than or equal to 5%, and we believe the BF estimates are accurate enough. Finally, the lower-right panel of Figure 5 shows the prediction mean squared error (pmse) over the test set for all h.

Based on our estimation, the BF attains the maximum value 9.75 at $h_{\text{EB}} = (0.075, e^{-5})$. Recall when comparing any two models indexed by h and h' respectively, the BF between them is given by $\text{BF}_{h,h'} = \frac{m_h / m_{h_1}}{m_{h'} / m_{h_1}}$. Also, according to Jeffreys (1998) and Kass and Raftery (1995), the evidence for h over h' is considered to be strong only if $\text{BF}_{h,h'}$ is greater than 10

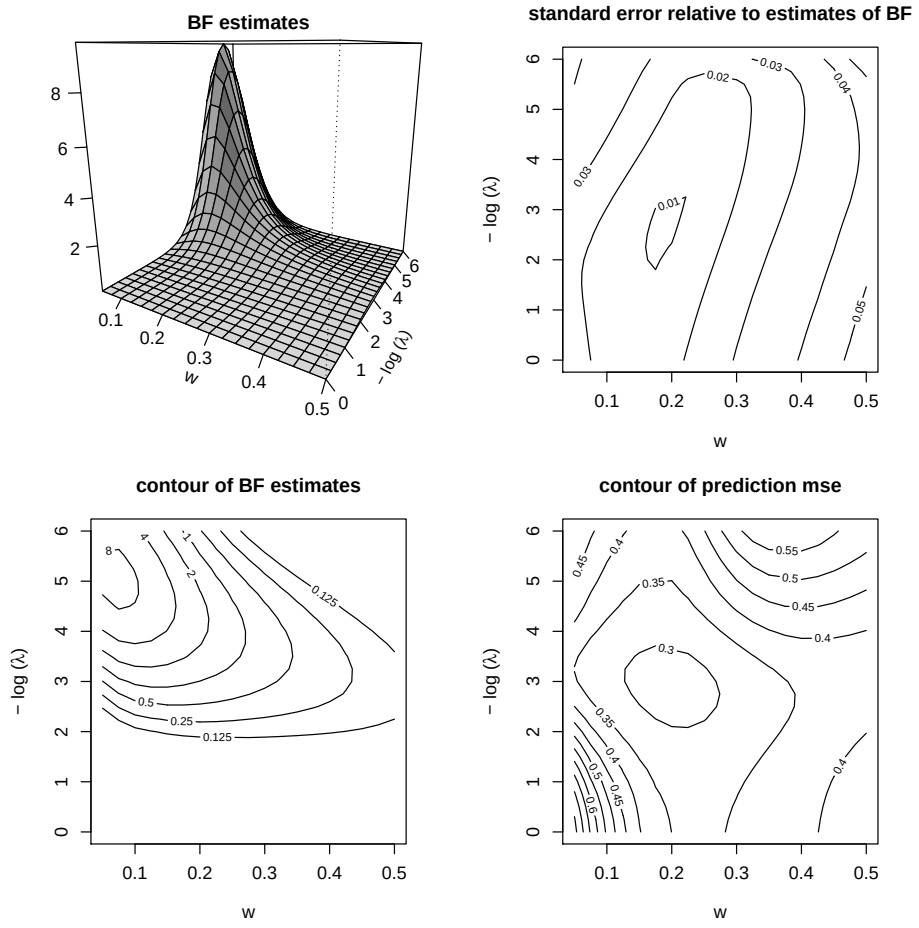

Figure 5: Left panels provide a surface plot and a contour plot for BF estimates. The upper-right panel displays standard errors with respect to the BF estimates. The lower-right panel shows pmse over the test set.

or 20. Hence, all h with BF over 1/10 or 1/20 times the maximum BF can be considered as reasonably well supported by the data as that of the empirical Bayesian choice. Comparing the lower two plots of Figure 5, we see that the set $A_c := \{h \in \mathcal{H} : \text{BF} > c\}$ for $c = 1$ and 0.5 do overlap with an area that corresponds to relatively small pmse. Outside $A_{0.5}$, a region that consists of larger w and smaller $-\log(\lambda)$ also enjoys small pmse values, at around 0.3 to 0.4. This region includes the common choice of $h_0 = (0.5, e^0)$. These suggest that h_{EB} and its vicinity might not be the only area of h that has good prediction performances.

To better compare the effect of $h_{\text{EB}} = (0.075, e^{-5})$ and the commonly used $h = (0.5, e^0)$,

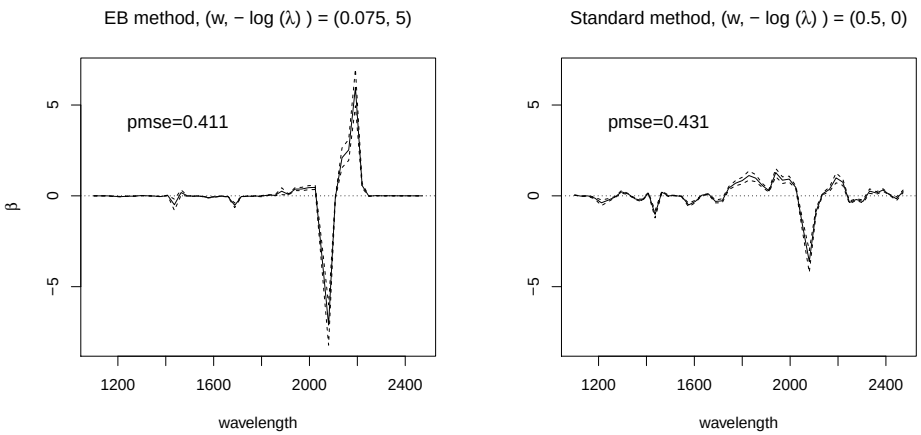

Figure 6: Estimated posterior mean of regression coefficients (with 95% point-wise confidence intervals) for the empirical Bayesian and the standard choice of h, respectively.

Figure 6 displays the estimated posterior mean of regression coefficients at both choices of h, together with the point-wise 95% confidence intervals for the posterior means. Due to the small size of w_{EB} and λ_{EB}, the empirical Bayesian method yields a model with a few covariates that have big coefficients. In comparison, the common choice has larger w and λ values, leading to a regression model that combines more covariates each having smaller effects. It turns out these two opposite strategy of modeling both predict the test dataset well, with pmse being 0.411 and 0.431 respectively. For comparison, pmses were calculated for several frequentist penalized linear regression methods with their respective penalty parameters chosen by 10-fold cross validation. The resulting pmses for the ridge, the lasso and the elastic net method are 4.675, 0.633 and 0.536, respectively.

The BM method for estimating SE is carried out above without the need of further user input. Theoretically, its competitor RS can be developed too, if enough regeneration times can be identified for each Markov chain. Recall that with the random-swap algorithm, each Markov chain lives on the discrete state space Γ of size 2^q. A naive way to introduce regeneration is to specify a single point γ_1, then each visit of the Markov chain to γ_1 marks a regeneration time. Note that the chance of visiting γ_1 converges to $\pi(\gamma_1)$, the posterior probability of γ_1. In our BVS model with $2^{50} \approx 1.1 \times 10^{15}$ states of γ, even $\max_{\gamma\in\Gamma} \pi(\gamma)$ could be very small. Take for example the Markov chain for the BVS model with $h = (w, \log(g)) = (.4, 1)$, the point with the highest frequency appeared only 8 out of a run of 10^4 iterations. And

that the waiting times between consecutive regenerations are highly variable, which ranges from less than ten iterations to a few thousand iterations. To obtain alternative schemes of identifying regeneration times, one can take the general minorization condition approach. It could potentially increase the chance of regeneration and reduce variability of the waiting times. Specifically, for any $\alpha \in \{1, 2, \cdots, 2^q\}$, one could define D_α to contain the α points with the highest posterior probabilities, and find $\epsilon_\alpha \in (0, 1]$ and a probability mass function $k_\alpha(\cdot)$ such that $p(\gamma'|\gamma) \geq \epsilon_\alpha I_{D_\alpha}(\gamma) k_\alpha(\gamma')$ for all $\gamma' \in \Gamma$. Note that as α increases, the chance of visiting D_α improves, but ϵ_α, the conditional rate of regeneration given the current state γ is in D_α, would decrease sharply. Finding a good α to maximize the overall chance of regeneration requires tuning that is specific for both the model specification h and the dataset. Even if we can find the optimal α for each Markov chain used in the example, it is unlikely that all of them would regenerate often enough for the RS estimator to be stable.

References

Brown, P. J., Fearn, T. and Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *J. Amer. Statist. Assoc.* **96** 398–408.

Buta, E. and Doss, H. (2011). Computational approaches for empirical Bayes methods and Bayesian sensitivity analysis. *Ann. Statist.* **39** 2658–2685.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 681–698.

Clyde, M. A., Ghosh, J. and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *J. Comput. Graph. Statist.* **20**.

Doss, H. and Tan, A. (2014). Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 683–712.

George, E. and Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731.

Geyer, C. J. (1994). Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Tech. Rep. 568r, Department of Statistics, University of Minnesota.

Hans, C. (2011). Elastic net regression modeling with the orthant normal prior. *J. Amer. Statist. Assoc.* **106** 1383–1393.

Jeffreys, H. (1998). *The theory of probability.* 3rd ed. Oxford.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Statist.* **33** 1700–1752.

Jones, G. L. (2004). On the Markov chain central limit theorem. *Probab. Surv.*, 1:299–320.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90** 773–795.

Kraemer, N. and Boulesteix, A. (2012). ppls: Penalized partial least squares, R package version 1.05.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103** 410–423.

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *J. Amer. Statist. Assoc.* **83** 1023–1036.

Mykland, P., Tierney, L. and Yu, B. (1995). Regeneration in Markov chain samplers. *J. Amer. Statist. Assoc.* **90** 233–41.

Osborne, B. G., Fearn, T., Miller, A. R. and Douglas, S. (1984). Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture* **35** 99–105.

Tan, A. and Doss, H. and Hobert, J. P. (2015). Honest importance sampling with multiple Markov chains. *J. Comput. Graph. Statist.*, **24** 792–826.

Vats, D., Flegal, J. M., and Jones, G. L.. (2015). Multivariate output analysis for Markov chain Monte Carlo. *ArXiv e-prints.*

Yuan, M. and Lin, Y. (2005). Efficient empirical Bayes variable selection and estimation in linear models. *J. Amer. Statist. Assoc.* **100** 1215–1225.