Electronic Journal of Statistics
Vol. 17 (2023) 895–947
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2118

Divide-and-conquer Bayesian inference in hidden Markov models*

Chunlei Wang and Sanvesh Srivastava†

Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa 52242, U.S.A.
e-mail: chunlei-wang@uiowa.edu; sanvesh-srivastava@uiowa.edu

Abstract: Divide-and-conquer Bayesian methods consist of three steps: dividing the data into smaller computationally manageable subsets, running a sampling algorithm in parallel on all the subsets, and combining parameter draws from all the subsets. These methods use the combined parameter draws for efficient posterior inference in massive data settings. Existing divide-and-conquer methods have a major limitation in that their first two steps assume that the observations are independent. We address this problem by developing a divide-and-conquer method for Bayesian inference in parametric hidden Markov models, where the state space is known and finite. Our main contributions are two-fold. First, we partition the data into smaller blocks of consecutive observations and modify the likelihood on every time block. For any time block, the posterior distribution computed using the modified likelihood is such that its variance has the same asymptotic order as that of the true posterior. Second, suppose the number of subsets is chosen appropriately depending on the mixing properties of the hidden Markov chain. In that case, we show that the subset posterior distributions defined using the modified likelihood are asymptotically normal as the subset sample size tends to infinity. This result facilitates using any existing combination algorithm in the third step. We show that the combined posterior distribution obtained using one such algorithm is close to the true posterior distribution in 1-Wasserstein distance under widely used regularity assumptions. Our numerical results show that the proposed method provides an accurate approximation of the true posterior distribution than its competitors in simulation studies and a real data analysis.

Keywords and phrases: Divide-and-conquer Bayesian inference, hidden Markov model, Monte Carlo, massive data.

Received May 2021.

1. Introduction

We focus on divide-and-conquer Bayesian inference in parametric hidden Markov models with a known and discrete state space, shortened as HMMs. Let $(Y_1, \ldots, Y_n)$ be the observed data from an HMM with parameter θ, K be the number of subsets, and $m = n/K$ be the subset sample size, where m is assumed to be an

*Chunlei Wang and Sanvesh Srivastava are partially supported by grants from the Office of Naval Research (ONR-BAA N000141812741) and the National Science Foundation (DMS-1854667/1854662).
†Corresponding Author.

integer for convenience. Bayesian inference in HMMs using Monte Carlo algorithms has been studied extensively (Cappé et al., 2006; Frühwirth-Schnatter, 2006). In massive data settings, however, posterior computations are inefficient due to multiple passes through the full data. The proposed method for divide-and-conquer inference in HMMs addresses this problem in three steps.

1. Partition the observed data sequence into K subsets
$$Y_{[0]} = \emptyset,\ Y_{[1]} = (Y_1, \ldots, Y_m),\ \ldots,\ Y_{[K]} = (Y_{(K-1)m+1}, \ldots, Y_n). \tag{1.1}$$
2. Given a prior density $\pi(\theta)$ and the modified conditional likelihood $p_\theta(Y_{[j]} \mid Y_{[j-1]})$, obtain θ draws on the subsets in parallel with posterior densities
$$\pi(\theta \mid Y_{[j]}, Y_{[j-1]}) \propto \pi(\theta)\{p_\theta(Y_{[j]} \mid Y_{[j-1]})\}^K, \quad j = 1, \ldots, K. \tag{1.2}$$
3. Combine posterior draws of θ from all the K subsets.

We study asymptotic properties of the posterior distributions estimated in the second and third steps, resulting in two related contributions. First, we show that mixing properties of the hidden Markov chain determine an appropriate choice of K and that $\{p_\theta(Y_{[j]} \mid Y_{[j-1]})\}^K$ in (1.2) accurately approximates the true conditional likelihood as m, K tend to infinity. Second, the combined posterior distribution estimated in the third step has similar asymptotic properties as the true posterior distribution under widely used regularity assumptions.

Two major groups of Markov chain Monte Carlo algorithms exist for posterior inference in HMMs, one based on data augmentation (Robert et al., 1999; Scott, 2002) and the other on the Metropolis-Hastings algorithm (Celeux et al., 2000; Robert et al., 2000; Cappé et al., 2003). The former and latter groups respectively draw and average over the hidden Markov chain in every iteration. The data augmentation-type algorithms use forward-backward (or Baum-Welch) recursions, but repeated passes through the full data are time-consuming in massive data settings (Scott, 2002; Rydén, 2008). Metropolis-Hastings-type algorithms bypass the last problem, but the proposal tuning required for the optimal performance outweighs their advantages in practice. A major reason for the popularity of these algorithms is the availability of efficient software, which faces computational bottlenecks in massive data applications (Rydén, 2008). Unfortunately, divide-and-conquer Bayesian methods cannot be used directly to solve the bottlenecks due to their focus on independent observations. Our main goal is to extend the existing divide-and-conquer toolbox to HMMs.

Online methods based on stochastic approximation have been widely used to address the inefficiency of data augmentation (Cappé and Moulines, 2009). Online Expectation Maximization (EM) has also been extended to HMMs and is efficient in massive data settings (Cappé, 2011; Le Corff and Fort, 2013; Kantas et al., 2015). Online EM has also been used to develop efficient variational Bayes algorithms, but such algorithms are known to underestimate posterior uncertainty (Foti et al., 2014; Giordano et al., 2018). Another developing area of research leverages piece-wise-deterministic Markov process (Goldman and Singh, 2021). These algorithms are scalable and guarantee convergence to a

general target, but they also require tuning for optimal performance in HMMs. Any such algorithm, which is properly tuned, can be embedded in the second step of the proposed method for drawing parameters on the subsets using the modified likelihood and posterior density defined in (1.2).

Particle Markov chain Monte Carlo methods are widely used for achieving efficient posterior inference in HMMs, but a majority of these methods use the full data sequence (Andrieu et al., 2010; Chopin and Singh, 2015; Aicher et al., 2019). Notable exceptions are the particle smoothing methods developed in Chan et al. (2016) and Ding and Gandy (2018). They divide the observed data sequence into smaller time blocks and estimate the parameters by applying a particle filter separately on each time block. More importantly, Chan et al. (2016) develop an unbiased estimate of the true likelihood and parallelize the sequential Monte Carlo algorithm over the time blocks for enhanced computational efficiency. These ideas are similar to the first two steps of our approach; however, the main focus of Chan et al. (2016) is in optimal parameter estimation, whereas our goal is to approximate the true posterior distribution of the HMM parameters using the divide-and-conquer technique and justify the asymptotic optimality of the estimated posterior.

A variety of divide-and-conquer methods exist for efficient posterior inference using Monte Carlo algorithms (Scott et al., 2016; Li et al., 2017; Minsker et al., 2017; Srivastava et al., 2018; Robert et al., 2018; Xue and Liang, 2019; Jordan et al., 2019; Wu and Robert, 2019; Shyamalkumar and Srivastava, 2022; Guhaniyogi et al., 2022). The combination steps in these methods are fairly general and their theoretical guarantees mainly rely on asymptotic normality of the subset posterior distributions. On the other hand, their first two steps assume that the observations are independent, so they are inapplicable to HMMs. The major difficulty lies in extending the second step to dependent observations, which defines the posterior density on any subset using a quasi-likelihood that raises the subset likelihood to a power of K. This modification compensates for the missing $(1 - 1/K)$-fraction of the full data on any subset (Minsker et al., 2014). The generalization of this idea to dependent data is unclear and we are unaware of any extensions for HMMs similar to (1.2).

Our first major contribution is the definition of subset posterior distribution for divide-and-conquer Bayesian inference in HMMs. The first step divides the observed data into K non-overlapping time blocks $Y_{[1]}, \ldots, Y_{[K]}$ defined in (1.1). For $j = 1, \ldots, K$, the second step defines the conditional likelihood of jth subset, which is denoted as $p_\theta(Y_{[j]} \mid Y_{[j-1]})$ in (1.2), by conditioning only on $Y_{[j-1]}$ instead of $Y_{[1]}, \ldots, Y_{[j-1]}$. The modified conditional likelihood is raised to a power of K that determines the quasi-likelihood for defining the (quasi) subset posterior density in (1.2). The quasi-likelihood replaces the true likelihood in the original Monte Carlo algorithm for posterior inference on a subset and is easily calculated using a simple modification of the forward-backward recursions. The implementation requires access to $(2/K)$-fraction of the full data, which is significantly more efficient than using the full data when $m \ll n$.

Our modified conditional likelihood on a subset is a type of composite likelihood specific to HMMs, where the composite marginal likelihood replaces the

true likelihood (Rydén, 1994, 1997; Andrieu et al., 2005; Pauli et al., 2011). The likelihood in (1.2) is a type of "fixed-lag conditional likelihood approximation", where the lag corresponds to the size of a time block. It also corresponds to an mth order antedependence model likelihood (Zimmerman and Núñez-Antón, 2009). The proposed method preserves the dependence on the historical data by computing the fixed-lag conditional likelihood using only two subsets as an alternative to the true conditional likelihood, which requires access to the full data. Based on this idea, variations of the proposed method simply replace $p_\theta(Y_{[j]} \mid Y_{[j-1]})$ in (1.2) by any composite likelihood in the second step. Specifically, if the subsets are large enough and the dependence is weak, then the marginal likelihood can be used in (1.2) for subset posterior computation. We choose $p_\theta(Y_{[j]} \mid Y_{[j-1]})$ as the conditional likelihood on subset j because it is readily computed using forward-backward recursions.

Our second major contribution is to show that the K subset posterior distributions defined in (1.2) are asymptotically normal as m and K tend to infinity. For independent observations, this is the first step in justifying asymptotic normality of the combined posterior distribution; see, for example, Li et al. (2017) and Xue and Liang (2019). This step is nontrivial in our case due to the dependence within and between the K time blocks. We show that if the growth of K is chosen appropriately depending on the mixing properties of the hidden Markov chain, then all the subset posterior distributions are asymptotically normal and their covariance matrices have the same asymptotic order as that of the true posterior distribution. Such results that quantify the growth of K on a mixing property are missing in the literature on divide-and-conquer Bayesian methods.

The asymptotic normality of subset posterior distributions simplifies the third step. First, it implies that any existing combination algorithm can be used to combine parameter draws from all the subsets. We choose the simplest one based on the Double Parallel Monte Carlo algorithm, where the combined parameter draw is a weighted average of the subset posterior draws (Xue and Liang, 2019). Second, the regularity assumptions required for justifying the asymptotic normality of subset posterior distributions extend naturally to the third step. Specifically, we need only an extra assumption on the weights used in the combination step for showing the asymptotic normality of the combined posterior distribution as m and K tend to infinity. Finally, posterior distribution estimated in the third step is called the *block filtered posterior distribution* due to the importance of prediction filter in the definition of subset posterior in (1.2).

The posterior distribution of parameter in HMMs is asymptotically normal under usual regularity assumptions (Bickel et al., 1998; De Gunst and Shcherbakova, 2008; Gassiat et al., 2014; Vernet, 2015b). We extend these results to divide-and-conquer Bayesian inference in that the block filtered posterior distribution is asymptotically normal if K and m are chosen properly depending on mixing property of the hidden Markov chain. Our regularity assumptions and convergence rates are natural extensions of existing results in that they reduce to those in De Gunst and Shcherbakova (2008) if $K = 1$ and to those in Li et al. (2017) if the data are independent. The additional set of regularity assumptions are required for guaranteeing the accuracy of the subset posterior distribution.

Finally, we identify conditions under which inference obtained using the true and block filtered posterior distributions agree asymptotically.

2. Divide-and-conquer Bayesian inference in HMMs

2.1. Setup and background

An HMM is a discrete-time stochastic process $\{(X_t, Y_t), t \geq 1\}$, where the Y-variables are observed and X-variables are latent. In our setup, HMMs satisfy three properties. First, $\{X_t\}$ is a time-homogeneous Markov chain with a finite state-space $\mathcal{X} = \{1, \ldots, S\}$, where S is known, and Y_t takes values in a general space $\mathcal{Y}$ for every t. Second, $\{Y_t, t \geq 1\}$ are conditionally independent given $\{X_t, t \geq 1\}$. Third, the distribution of Y_t given $\{X_u, u \geq 1\}$ equals the distribution of Y_t given X_t. The initial distribution and transition probability matrix of $\{X_t\}$ are denoted as $r(\cdot)$ and Q, respectively, where $r(\cdot)$ is stationary with respect to Q, $\mathbb{P}(X_1 = a) = r(a)$ for $a \in \mathcal{X}$, and

$$Q_{ab} = q(a, b), \quad q(a, b) = \mathbb{P}(X_{t+1} = b \mid X_t = a), \quad a, b \in \mathcal{X}, \quad t \geq 1. \tag{2.1}$$

The conditional distributions of Y_t given X_t $(t = 1, \ldots, n)$ belong to a common family. For any $x \in \mathcal{X}$, the conditional distribution of Y_t given $X_t = x$ is $G(\cdot \mid x)$ and has density $g(\cdot \mid x)$ with respect to a σ-finite measure μ on $\mathcal{Y}$.

We adopt a widely used setup for studying the asymptotic properties of HMMs (De Gunst and Shcherbakova, 2008; Bickel et al., 1998; Leroux, 1992). Let $\Theta \subset \mathbb{R}^d$ be the parameter space of fixed dimension $d \in \mathbb{N}$, the dependence of $r(\cdot)$, $q(\cdot, \cdot)$, $g(\cdot \mid x)$ on a parameter $\theta \in \Theta$ be denoted as $r_\theta(\cdot)$, $q_\theta(\cdot, \cdot)$, $g_\theta(\cdot \mid x)$, and $\theta_0 \in \Theta$ be the true parameter value. We use the shorthand Y_m^n for $(Y_m, \ldots, Y_n)$, where $1 \leq m \leq n$. Denote the sample size as n, the observed data as $(Y_1, \ldots, Y_n) \equiv Y_1^n$, the latent data as $(X_1, \ldots, X_n) \equiv X_1^n$, and the augmented data as (X_1^n, Y_1^n). The joint density of (X_1^n, Y_1^n) with respect to (counting measure)$^n \times \mu^n$ and marginal density of Y_1^n with respect to μ^n are

$$\begin{aligned} p_\theta(X_1^n, Y_1^n) &= r_\theta(X_1) \prod_{t=1}^{n-1} q_\theta(X_t, X_{t+1}) \prod_{t=1}^{n} g_\theta(Y_t \mid X_t), \\ p_\theta(Y_1^n) &= \sum_{X_1^n \in \mathcal{X}^n} p_\theta(X_1^n, Y_1^n), \quad \theta \in \Theta. \end{aligned} \tag{2.2}$$

The true density of the observed data is $p_{\theta_0}(Y_1^n)$ for some $\theta_0 \in \Theta$. Given a prior distribution on Θ with density π, many Monte Carlo algorithms are available for posterior inference on θ. When n is large, most of them are computationally expensive due to the repeated passes through the full data in every iteration.

2.2. Partitioning scheme

The first step in divide-and-conquer Bayesian inference partitions the full data into smaller subsets. A widespread practice is to randomly divide the samples

into K subsets, but it fails to preserve the dependence structure of HMM on any subset. We instead partition the full data into K non-overlapping blocks of consecutive observations starting from Y_1, where each block has size $m = n/K$ and m is assumed to be an integer for convenience. Denote the jth block of observation with sample size m as $Y_{(j-1)m+1}^{jm}$ and the latent X-variables as $X_{(j-1)m+1}^{jm}$ $(j = 1, \ldots, K)$. The K observation blocks constitute the K partitions of Y_1^n in that $Y_1^n = (Y_1^m, \ldots, Y_{(K-1)m+1}^n)$. The choice of K (or m) depends on n and the mixing properties of the hidden Markov chain. It plays a crucial in asymptotically valid inference on any subset relative to the true posterior distribution; see Theorem 3.1 in Section 3.2 for the details.

After partitioning the data, the likelihoods of θ on the first and jth subsets $(j = 2, \ldots, K)$ conditional on the preceding subsets are

$$\begin{aligned} p_\theta(Y_1^m) &= \sum_{X_1 \in \mathcal{X}} p_\theta(Y_1^m \mid X_1)\, p_\theta(X_1), \\ p_\theta(Y_{(j-1)m+1}^{jm} \mid Y_1^{(j-1)m}) &= \sum_{X_{(j-1)m+1}} p_\theta(Y_{(j-1)m+1}^{jm} \mid X_{(j-1)m+1}) p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m}), \end{aligned} \tag{2.3}$$

respectively, where $p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m})$ is the *prediction filter* of $X_{(j-1)m+1}$ given the previous $(j-1)$ subset blocks and $p_\theta(X_1)$ equals the stationary distribution $r_\theta(X_1)$. For every j, the prediction filter is computed in $O(n)$ operations using *forward filtering* (or Baum-Welch algorithm) if we have access to the full data. Unfortunately, no subset has access to the full data in the divide-and-conquer setup. Further, using $Y_{(j-1)m+1}^{jm}$ instead of Y_1^n for inference on θ implies that the subset j likelihood ignores the dependence on $Y_1^{(j-1)m}$. We next address both problems by modifying the subset j likelihood in (2.3) using the Kth power of one-block conditional likelihood $\{p_\theta(Y_{(j-1)m+1}^{jm} \mid Y_{(j-2)m+1}^{(j-1)m})\}^K$ that requires accessing only $(2/K)$-fraction of the full data.

2.3. Subset sampling scheme

In the divide-and-conquer setup, any chosen sampler requires modification of the subset likelihood because every subset posterior conditions on an $(1/K)$-fraction of the full data. A popular solution to this problem is raising the subset likelihood to the power of K. This is equivalent to replicating the data on a subset K-times, and it compensates for the missing fraction of data, resulting in uncertainty estimates of parameters that are asymptotically equivalent to those obtained using the true posterior distribution (Minsker et al., 2014). The likelihood modification is called *stochastic approximation* and has been widely used in parametric models for independent data (Srivastava et al., 2015; Li et al., 2017; Xue and Liang, 2019). The dependence induced by the hidden Markov chain implies that stochastic approximation developed for independent data is inapplicable for divide-and-conquer Bayesian inference in HMMs.

One of our main contributions is to extend stochastic approximation so that it accounts for dependence between subsets. The prediction filter $p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m})$ in (2.3) is geometrically ergodic and rapidly converges to its martingale limit as m tends to infinity (Bickel et al., 1998). Motivated by this result and fixed-lag smoothing methods, we define the jth approximate "prediction filter" by conditioning only on the $(j-1)$th subset; that is, $p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m})$ replaces $p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m})$ in (2.3). The likelihoods in (2.3) are accordingly modified using this approximate prediction filter to yield the one-block conditional likelihood of θ on subset j $(j = 2, \ldots, K)$ as

$$p_\theta(Y_{(j-1)m+1}^{jm} \mid Y_{(j-2)m+1}^{(j-1)m}) = \sum_{X_{(j-1)m+1} \in \mathcal{X}} p_\theta(Y_{(j-1)m+1}^{jm} \mid X_{(j-1)m+1}) \times p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m}). \quad (2.4)$$

To calculate the one-block conditional likelihood in (2.4), we use $p_\theta(Y_{(j-1)m+1}^{jm} \mid X_{(j-1)m+1})$ from (2.2) and estimate the prediction filter $p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m})$ using forward-backward recursions of the Baum–Welch algorithm. The one-block conditional likelihood accurately approximates the true conditional likelihood in (2.3) for every θ in total variation distance as m tends to infinity; see Proposition 3.1 in Section 3 for details. The one-block conditional likelihood in (2.4) is also efficiently computed in $O(m)$ operations using forward-backward recursions and requires access to only $2m$ observations instead of n required for computing the true likelihood in (2.3), leading to computational gains if $m \ll n$.

The jth subset posterior distribution is defined using the jth one-block conditional likelihood. Specifically, given a prior distribution Π on Θ with density π defined with respect to Lebesgue measure, the density of the posterior distribution of θ given the jth subset is defined as

$$\tilde{\pi}_m(\theta \mid Y_{(j-1)m+1}^{jm}) = \frac{e^{Kw_j(\theta)}\pi(\theta)}{\int_\Theta e^{Kw_j(\theta)}\pi(\theta)d\theta}, \quad (2.5)$$

where $w_1(\theta) = \log p_\theta(Y_1^m)$ and $w_j(\theta) = \log p_\theta(Y_{(j-1)m+1}^{jm} \mid Y_{(j-2)m+1}^{(j-1)m})$ for $j = 2, \ldots, K$. The quasi-likelihoods used in (2.5) are the Kth power of the one-block conditional likelihoods. They account for the dependence of a subset on their immediately preceding time block and are raised to a power of K to compensate for the missing $(1 - 1/K)$-fraction of the data. The latter idea is also used in the stochastic approximation for independent data. Indeed, if the Y_1^n are independent, then (2.5) recovers the definition of subset posterior distributions in divide-and-conquer Bayesian inference for independent data; therefore, our proposal in (2.5) generalizes stochastic approximation to dependent data. The jth one-block conditional likelihood is also a composite likelihood in that it approximates the true conditional likelihood in (2.3) without passing through the full data. Other variations of $\tilde{\pi}_m(\theta \mid Y_{(j-1)m+1}^{jm})$ are obtained by replacing $w_j(\theta)$ in (2.5) with other composite likelihoods that are used for inference in HMMs, including those in Rydén (1994, 1997); Andrieu et al. (2005).

The asymptotically normality of $\tilde{\pi}_m(\theta \mid Y_{(j-1)m+1}^{jm})$ in (2.5) does not follow from existing results. The one-block conditional likelihood approximates the true conditional likelihood in (2.4) and is raised by a power K in (2.5). Unlike the common asymptotic setup where $K = 1$ and $n = m$, K and m in our case simultaneously tend to infinity and $n = mK$ (De Gunst and Shcherbakova, 2008; Vernet, 2015b,a). A crucial step in establishing the asymptotic normality of the block filtered posterior distribution is to show that $\tilde{\pi}_m(\theta \mid Y_{(j-1)m+1}^{jm})$ is asymptotically normal. We establish that if the growth rate of K is chosen appropriately depending on the mixing properties of the prediction filter and K, m tend to infinity, then the subset posterior distributions are asymptotically normal with covariance matrices having the same asymptotic order as that of true posterior distribution; see Theorem 3.1 in Section 3 for details.

Any Monte Carlo algorithm can now be used for drawing θ using (2.5) on all the K subsets in parallel. Let $\theta_{(j)}^{(t)}$ be the tth draw of θ on subset j and T be the number of post-burn-in draws collected on any subset. Motivated from the asymptotic normality of subset posterior distributions, the next section develops a combination scheme that approximates the true posterior distribution using a mixture with K components, where the draws from the jth mixture component are obtained by appropriately transforming, centering, and scaling the subsets posterior draws $\{\theta_{(j)}^{(t)}\}_{t=1}^{T}$ for $j = 1, \ldots, K$.

2.4. Combination scheme

The final step of a divide-and-conquer Bayesian approach combines parameter draws from all the subsets. Our combination scheme is an extension of the Double-Parallel Monte Carlo algorithm (Xue and Liang, 2019), which is computationally simple. It uses a global centering vector $\tilde{\theta} \in \mathbb{R}^d$ and a $d \times d$ symmetric positive definite scaling matrix $\tilde{\Sigma}$. Let $\hat{\theta}_j$ be the maximum likelihood estimator and Σ_j be the posterior covariance matrix of θ on subset j. Then, we approximate the density of the true posterior distribution using

$$\tilde{\pi}_n(\theta \mid Y_1^n) = \frac{1}{K} \sum_{j=1}^{K} \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \tilde{\pi}_m \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} (\theta - \tilde{\theta}) + \hat{\theta}_j \mid Y_{(j-1)m+1}^{jm} \right\}, \quad \theta \in \Theta, \tag{2.6}$$

where $\tilde{\pi}_m(\theta | Y_{(j-1)m+1}^{jm})$ is the jth subset posterior density defined in (2.5) and $\tilde{\pi}_n(\theta \mid Y_1^n)$ is a mixture of the K subset posterior densities. The distribution that has density $\tilde{\pi}_n(\theta \mid Y_1^n)$ in (2.6) is called the *block filtered posterior* distribution. Our combination scheme in (2.6) depends on the choice of $(\tilde{\theta}, \tilde{\Sigma})$, so we denote it as COMB$(\tilde{\theta}, \tilde{\Sigma})$.

Let $\{\theta_{(j)}^{(t)}\}_{t=1}^{T}$ be the parameter draws on subset j for $j = 1, \ldots, K$. The application of COMB$(\tilde{\theta}, \tilde{\Sigma})$ has two steps and yields θ draws from block filtered posterior distribution as

$$\tilde{\theta}_{(j)}^{(t)} = \tilde{\theta} + \tilde{\Sigma}^{1/2} \left\{ \Sigma_j^{-1/2} (\theta_{(j)}^{(t)} - \hat{\theta}_j) \right\}, \quad j = 1, \ldots, K; \; t = 1, \ldots, T, \tag{2.7}$$

where the θ draws on subset j are centered and scaled by $(\hat{\theta}_j, \Sigma_j)$ for $j = 1, \ldots, K$ and then re-centered and re-scaled by $(\tilde{\theta}, \tilde{\Sigma})$. The parameter draws $\tilde{\theta}_{(j)}^{(t)}$ in (2.7) are used as the computationally efficient alternative to draws from the true posterior distribution for inference on θ. If the centering and scaling parameters $(\hat{\theta}_j, \Sigma_j)_{j=1}^K$ is unavailable, then we use their Monte Carlo estimates $(\mu_{\theta_{(j)}}, \Sigma_{\theta_{(j)}})_{j=1}^K$ for obtaining the θ draws as

$$\mu_{\theta_{(j)}} = \frac{1}{T}\sum_{t=1}^{T} \theta_{(j)}^{(t)}, \quad \Sigma_{\theta_{(j)}} = \frac{1}{T}\sum_{t=1}^{T} \left(\theta_{(j)}^{(t)} - \mu_{\theta_{(j)}}\right)\left(\theta_{(j)}^{(t)} - \mu_{\theta_{(j)}}\right)^{\mathrm{T}}, \quad j=1, \ldots, K, \tag{2.8}$$

which reduces to a version of the Double-Parallel Monte Carlo algorithm.

Many existing combination algorithms are special cases of (2.7) with different choices of $(\tilde{\theta}, \tilde{\Sigma})$. Let I_d be a $d \times d$ identity matrix and $A^{1/2}$ be the square root of a symmetric positive definite matrix A. Then, three candidates for $(\tilde{\theta}, \tilde{\Sigma})$ are

$$\begin{aligned} &(\overline{\mu}, I_{d\times d}), \quad (\overline{\mu}, \check{\Sigma}), \quad (\overline{\mu}, \overline{\Sigma}), \quad \overline{\mu} = \frac{1}{K}\sum_{j=1}^{K} \mu_{\theta_{(j)}}, \\ &\check{\Sigma}^{1/2} = \frac{1}{K}\sum_{j=1}^{K} \Sigma_j^{1/2}, \quad \overline{\Sigma} = \frac{1}{K}\sum_{j=1}^{K} \Sigma_j. \end{aligned} \tag{2.9}$$

They lead to three combination algorithms COMB$(\overline{\mu}, I_{d\times d})$, COMB$(\overline{\mu}, \check{\Sigma})$, and COMB$(\overline{\mu}, \overline{\Sigma})$, where the first two have been proposed in Xue and Liang (2019) and Srivastava and Xu (2021), respectively. In Section 4 of experiments, we draw θ from the block filtered posterior distribution using COMB$(\hat{\theta}, \overline{\Sigma})$ in an HMM with Gaussian emission densities, where $\hat{\theta}$ is the maximum likelihood estimator using the full data and EM (or Baum-Welch) algorithm.

3. Theoretical properties

3.1. Setup

We introduce some notation required for stating the assumptions of our theoretical setup. Given $n = mK$ observations $Y_1^n \in \mathcal{Y}^n$, the first subset log-likelihood is $L_m(\theta) = \log p_\theta(Y_1^m)$ and the jth subset log-likelihood is $L_{jm}(\theta) = \log p_\theta(Y_{(j-1)m+1}^{jm})$ for $\theta \in \Theta$, $j = 2, \ldots, K$, where p_θ is defined in (2.2) in terms of r_θ, Q_θ, and g_θ. Let $\theta_0 \in \Theta$ be the true parameter value, $P_{\theta_0}^{(n)}$ be the law of Y_1^n, $\{P_\theta^{(n)} : \theta \in \Theta\}$ be the family of probability measures on $\mathcal{Y}^n$ such that $P_\theta^{(n)}$ has density p_θ with respect to μ^n, and Π be the prior distribution on Θ with density $\pi(\theta)$ with respect to the Lebesgue measure on $\mathbb{R}^d$. Denote the Euclidean norm as $\|\cdot\|_2$, $P_{\theta_0}^{(n)}$ as $\mathbb{P}_0$, $P_\theta^{(n)}$ as $\mathbb{P}_\theta$, expectation with respect to $\mathbb{P}_0$ and $\mathbb{P}_\theta$ as $\mathbb{E}_0$ and $\mathbb{E}_\theta$, respectively, and the first, second, and third derivative of $L_{jm}(\theta)$ with respect to θ as $L'_{jm}(\theta)$, $L''_{jm}(\theta)$, and $L'''_{jm}(\theta)$, respectively, for $j = 1, \ldots, K$.

The theoretical guarantees are based on the following assumptions on a family of HMM models $\{P_\theta^{(n)}, \theta \in \Theta\}$ and the prior distribution Π.

(A_1) (Existence of an envelope function.) There exists a constant $\delta_0 > 0$, such that for any $m \geq 1$, $L_m(\theta)$ is three times differentiable with respect to θ in a neighborhood $B_{\delta_0}(\theta_0) = \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq \delta_0\}$. For any $(z_1, \ldots, z_m) \in \mathcal{Y}^m$, there exists an envelope function $M(z_1, \ldots, z_m)$ that satisfies

$$\sup_{\theta \in B_{\delta_0}(\theta_0)} \left| \frac{d \log p_\theta(z_1, \ldots, z_m)}{d\theta_{l_1}} \right| \leq M(z_1, \ldots, z_m),$$

$$\sup_{\theta \in B_{\delta_0}(\theta_0)} \left| \frac{d^2 \log p_\theta(z_1, \ldots, z_m)}{d\theta_{l_1} d\theta_{l_2}} \right| \leq M(z_1, \ldots, z_m),$$

$$\sup_{\theta \in B_{\delta_0}(\theta_0)} \left| \frac{d^3 \log p_\theta(z_1, \ldots, z_m)}{d\theta_{l_1} d\theta_{l_2} d\theta_{l_3}} \right| \leq M(z_1, \ldots, z_m), \quad l_1, l_2, l_3 = 1, \ldots, d,$$

and $\frac{1}{m} M(Y_1, \ldots, Y_m) \to \mathcal{C}_M$ $\mathbb{P}_0$-almost surely as $m \to \infty$ for a constant $\mathcal{C}_M > 0$.

(A_2) (Law of large numbers for the log-likelihood function.) For any $\theta \in \Theta$ and $j \in \{1, \ldots, K\}$, the normalized jth subset log-likelihood $m^{-1} L_{jm}(\theta) \to L(\theta)$ $\mathbb{P}_0$-almost surely as $m \to \infty$, where $L(\theta)$ is a continuous deterministic function with a unique global maximum at θ_0. The convergence is uniform over any compact subsets of Θ.

(A_3) (Central limit theorem for the score function at the true parameter.) For any $j \in \{1, \ldots, K\}$, the normalized score function evaluated at θ_0 on jth subset, $m^{-1/2} L'_{jm}(\theta_0)$, $\mathbb{P}_0$-weakly converges to $N_d(0, I_0)$ as $m \to \infty$, where the I_0 is non-singular and defined as the Fisher information matrix of the model evaluated at θ_0.

(A_4) (Law of large numbers for the observed information matrix.) If θ_m^* is any stochastic sequence in Θ such that $\theta_m^* \to \theta_0$ $\mathbb{P}_0$-almost surely as $m \to \infty$, then $-m^{-1} L''_{jm}(\theta_m^*) \to I_0$ in $\mathbb{P}_0$-probability as $m \to \infty$. Furthermore, assume that $-m^{-1} L''_{jm}(\theta)$ is positive definite with eigenvalues uniformly bounded from below and above by constants for all $\theta \in B_{\delta_0}(\theta_0)$, every $j = 1, \ldots, K$, all values of $Y_{(j-1)m+1}^{jm} \in \mathcal{Y}^m$, and sufficiently large m.

(A_5) The parameter space $\Theta \subset \mathbb{R}^d$ is compact with $\operatorname{diam}(\Theta) := \sup_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_2 < \infty$ and θ_0 as an interior point.

(A_6) The prior density $\pi(\theta)$ is positive and continuous for all $\theta \in \Theta$.

(A_7) For $\theta \in \Theta$, the transition probability matrix $\{Q_\theta(a, b)\}$ $(a, b \in \{1, \ldots, S\})$ is ergodic.

Our regularity assumptions are commonly used for proving Bernstein-von Mises theorem in parametric models. Assumptions (A_2)–(A_6) are identical to those used in Theorem 2.1 of De Gunst and Shcherbakova (2008) if we set $m = n$ with $K = 1$. Assumption (A_1) is slightly stronger than the requirement of continuous second derivatives of the log-likelihood function in De Gunst and Shcherbakova (2008). We require the existence of an envelope function to con-

trol the growth of the remained term in the third order Taylor expansion of $L_{jm}(\theta)$ for every j after stochastic approximation. For the same reasons, Li et al. (2017) employ an assumption that is equivalent to Assumption (A_1). Assumptions (A_5)–(A_6) ensure that the prior density $\pi(\theta)$ is well-behaved on compact set Θ. Assumptions (A_1)–(A_6) are the extensions of assumptions required for proving the Bernstein-von Mises theorem for independent and identically distributed data (Lehmann and Casella, 1998, Theorem 8.2, Chapter 6). Assumption (A_7) implies that for any $\theta \in \Theta$, there exists a positive integer $v \geq 1$ such that every element of the v-step transition matrix Q_θ^v is positive. We assume that $v = 1$ and $\epsilon := \inf_{a,b\in\mathcal{X},\theta\in\Theta} Q_\theta(a,b) > 0$. The ergodicity of $\{X_t\}$ in Assumption (A_7) implies that $\{Y_t\}$ is stationary and ergodic under $\mathbb{P}_\theta$ (Leroux, 1992, Lemma 1). The equivalents of Assumptions (A_1)–(A_7) are also used for proving the asymptotic normality of the maximum likelihood estimator of θ_0 in HMMs (Bickel et al., 1998).

The definition of jth subset ($j = 2, \ldots, K$) posterior distribution in (2.5) uses the jth quasi-likelihood with approximate prediction filter conditioning only on the $(j-1)$th subset. Due to this approximation, we need the following assumptions to guarantee the consistency of the jth subset posterior distribution defined using the jth quasi-likelihood.

(A_8) (i) For $\theta \in \Theta$

$$\max_{a\in\mathcal{X}} \int_{\mathcal{Y}} \left\{ \max_{b\in\mathcal{X}} |\log g_\theta(y|b)| \right\} g_{\theta_0}(y|a)\mu(dy) < \infty.$$

(ii) Let $h_\theta(y) = \max_{a,b\in\mathcal{X}} \frac{g_\theta(y|a)}{g_\theta(y|b)} \geq 1$. There exists a constant $M \geq 1$ such that $\mathbb{P}_0$-almost surely,

$$\sup_{\theta\in\Theta} h_\theta(Y_1) < M.$$

(iii) (Convergence of prediction filter.) For $\theta \in \Theta$, define $\mu_\theta(y) = \{1 + (S-1)\epsilon^{-2}h_\theta(y)\}^{-1}$ such that $\mathbb{P}_0$-almost surely,

$$\|p_\theta(X_{m+1}|Y_1^m) - r_\theta(X_{m+1})\|_{\mathrm{TV}} \leq S \prod_{i=2}^{m-1} \exp\{-2\mu_\theta(Y_i)\} \leq S\rho^{m-2},$$

where the mixing coefficient

$$\rho := e^{-\frac{2}{1+(S-1)\epsilon^{-2}M}} \tag{3.1}$$

and $\|d\mathbb{P} - d\mathbb{Q}\|_{\mathrm{TV}}$ is the total variation distance between measures $\mathbb{P}$ and $\mathbb{Q}$.

Assumptions (A_8)(i)–(iii) are based on Le Gland and Mevel (2000, Theorem 2.1). They are global over Θ and imply that the prediction filters, hence the subset log-likelihood functions, forget the condition of initial distribution exponentially fast. Among them, (A_8)(ii) is slightly stronger as we require $h_\theta(Y_1)$ to

be bounded $\mathbb{P}_0$-almost surely. Assumption (A_8) (iii) implies that $p_\theta(X_{m+1} \mid Y_1^m)$ converges to $r_\theta(X_{m+1})$ geometrically with mixing coefficient ρ, where the martingale convergence of $p_\theta(X_{m+1} \mid Y_1^m)$ and functions $\mu_\theta(Y_i)$ are discussed in Bickel et al. (1998, Lemma 5). Importantly, (iii) is stronger than the assumptions of Bickel et al. (1998) in that we assume that as m tends to infinity, the martingale limit of $p_\theta(X_{m+1} \mid Y_1^m)$ is r_θ instead of $p_\theta(\cdot \mid Y_1^\infty)$. The bound on the mixing coefficient ρ in (3.1) is not tight. Specifically, a referee has pointed out that when $\epsilon = 1/S$ or $M = 1$, the HMM sequence degenerates to an independent sequence but the bound in (3.1), hence ρ, fails to reflect this scenario.

The following proposition highlights the importance of Assumption (A_8). It shows that the jth approximate prediction filter converges to the true prediction filter $(j = 2, \ldots, K)$, which implies that the quasi-likelihood in (2.4) provides an accurate approximation of the true likelihood in (2.3) as m tends to infinity.

Proposition 3.1. *If Assumption (A_8) holds, then for $j = 2, \ldots, K$ and $\theta \in \Theta$, as $m \to \infty$,*

$$\|p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m}) - p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m})\|_{TV} \to 0$$

$\mathbb{P}_0$-almost surely.

3.2. Main results

We now present two theorems stating theoretical properties of the subset and block filtered posterior distributions. First, the following theorem specifies the asymptotic order of K depending on the mixing coefficient ρ and other regularity assumptions such that the limiting distributions of the K subset posterior distributions are normal as m and K tend to infinity.

Theorem 3.1. *Let the number of subsets K be a sequence $K_m \to \infty$ as $m \to \infty$ and $n = mK_m$. For $j = 1, \ldots, K_m$, denote the maximum likelihood estimator of θ as $\hat{\theta}_j$ that solves $L'_{jm}(\theta) = 0$, define the local parameter $h_j = \sqrt{n}(\theta - \hat{\theta}_j)$, and denote the distribution of h_j implied by the jth subset posterior density in (2.5) as $\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm})$. If Assumptions (A_1)–(A_7) hold, then*

(i) as $m \to \infty$,

$$\|\Pi_m^{h_1}(\cdot \mid Y_1^m) - N_d(0, I_0^{-1})\|_{TV} \to 0 \text{ in } \mathbb{P}_0\text{-probability}; \tag{3.2}$$

(ii) In additional, if Assumption (A_8) also holds and the number of subsets satisfies $K_m = o(\rho^{-m})$ for ρ defined in (3.1), then for $j \geq 2$, as $m \to \infty$,

$$\|\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d(0, I_0^{-1})\|_{TV} \to 0 \text{ in } \mathbb{P}_0\text{-probability}, \tag{3.3}$$

where $K_m = o(\rho^{-m})$ denotes the sequence satisfying $K_m \rho^m \to 0$ as $m \to \infty$ and $N_d(0, I_0^{-1})$ is a d-variate normal distribution with zero mean and the inverse of Fisher information matrix I_0^{-1} as the covariance matrix.

Theorem 3.1 shows that the credible intervals and confidence intervals based on the jth subset posterior density (2.5) and maximum likelihood estimator on jth subset, respectively, are asymptotically equivalent ($j = 1, \ldots, K$). We also establish that $K = o(\rho^{-m})$ is rate-optimal in that the limiting distribution of subset posterior distributions are normal. This result is required for establishing that the combined and true posterior distributions are asymptotically close in total variation distance. The upper bound on the growth of K immediately yields a lower bound on the growth of the subset size because $m = n/K$. For weakly dependent HMMs with $\rho \ll 1$, the number of K can be chosen to be large and m will be relatively small. On the other hand, for strongly dependent HMMs with $\rho \approx 1$, m has to be large to guarantee accurate estimation of the subset posterior distributions, which results in a relatively small K. The theoretical guidance for choosing $K = o(\rho^{-m})$ in Theorem (ii) assumes that n, m, K tend to infinity, but in practice unlimited computing resources are unavailable and K cannot exceed the total number processors used, say $K_{\max}$. In this case, the number of subsets is the minimum of $K_{\max}$ and $K = o(\rho^{-m})$.

The mixing coefficient ρ is unknown in practice and the relation $K = o(\rho^{-m})$ is asymptotic, implying that Theorem 3.1 (ii) cannot be used directly to choose K. This is a common problem in divide-and-conquer methods, where the choice of K depends on unknown parameters. To bypass this problem, we suggest a heuristic of choosing $K \in \{\log n, n^{1/4}, n^{1/3}, n^{1/2}\}$ and using the value with maximum accuracy. In our numerical results, we have observed that if $K \approx \log n$, then accuracy is very high but it comes at the cost of small run-time gains. The accuracy decreases slightly as K increases but the gain in run-times are significant. The optimal choice of K balances this accuracy-efficiency trade-off. If a crude estimate of ρ is available, then the theorem can be used with the heuristic. For a strongly dependent HMM sequence with $\rho \approx 1$, K cannot be too large; therefore, choosing $K = \log(n)$ is better. On the other hand, K can be chosen at a polynomial order of n for a weakly dependent HMM sequence. We evaluate this heuristic in Section 4.

Let $\tilde{\Pi}_m^{\theta}(\cdot \mid Y_{(j-1)m+1}^{jm})$ denote the jth subset posterior distribution of θ with density defined in (2.5). Then, the invariance of the total variation distance implies that as $m \to \infty$,

$$\|\tilde{\Pi}_m^{\theta}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d(\hat{\theta}_j, (nI_0)^{-1})\|_{TV} \to 0 \text{ in } \mathbb{P}_0\text{-probability for } j \geq 1. \tag{3.4}$$

If $K = 1$, Theorem 3.1 recovers the usual Bernstein-von Mises theorem for the true posterior distribution (De Gunst and Shcherbakova, 2008). Comparing the limiting distributions of subset posterior distribution and true posterior distribution, we observe that they are both asymptotically d-variate normal distributions with the same covariance matrix $(nI_0)^{-1}$, but with different maximum likelihood estimator as mean parameters. Based on this observation, the proposed combined posterior distribution, *block filtered posterior* distribution, is a mixture model of subset posterior distribution after appropriate re-scaling and re-centering.

Our next result establishes conditions under which the block filtered posterior distribution converges to the true posterior distribution at an optimal parametric rate as both m and K tend to infinity. The first step in this process is to show the asymptotic normality of the block filtered posterior distribution. To this end, we need the following assumption on the re-scaling matrix $\tilde{\Sigma}$ and subset posterior covariance matrices.

(A_9) Let Σ_j be the covariance matrix of jth subset posterior distribution of θ ($j = 1, \ldots, K$) and the rescaling matrix $\tilde{\Sigma}$ be a deterministic positive symmetric matrix. As $m \to \infty$,

$$\|(n\tilde{\Sigma})^{-1/2} - I_0^{1/2}\|_{\mathrm{F}}^2 \to 0, \quad \mathbb{E}_0 \|(n\Sigma_j)^{1/2} - I_0^{-1/2}\|_{\mathrm{F}}^2 \to 0, \tag{3.5}$$

uniformly for $j = 1, \ldots, K$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm.

Assumption (A_9) implies that $(n\Sigma_j)^{1/2}$ is close to $I_0^{-1/2}$ and that the square Frobenius norm of their difference converges to zero in $\mathbb{E}_0$-expectation. Furthermore, $(n\tilde{\Sigma})^{-1/2}$ is close to $I_0^{1/2}$ and the square Frobenius norm of their difference converges to zero. This ensures that the limiting covariance matrices of block filtered posterior and true posterior distributions coincide. For two integrable measures ν_1, ν_2 on Θ, the 1-Wasserstein distance is defined as $W_1(\nu_1, \nu_2) = \sup_{\|f\|_{\mathrm{Lip}} \le 1} \left\{ \int_\Theta f(\theta) d(\nu_1 - \nu_2)(\theta), f : \Theta \to \mathbb{R} \text{ is continuous} \right\}$. If we choose $\tilde{\Sigma}$ satisfying (3.5), the following theorem states that the block filtered posterior distribution is asymptotically normal and the 1-Wasserstein distance between the block filtered and true posterior distributions converges to 0 in $\mathbb{P}_0$-probability, with rate determined by the closeness between the re-centering vector and the maximum likelihood estimator of θ.

Theorem 3.2. *Let m, K_m satisfy the conditions in Theorem 3.1 and $n = mK_m$, $\tilde{\Pi}_n^\theta(\cdot \mid Y_1^n)$ be the block filtered posterior distribution of θ given Y_1^n with density defined in (2.7) based on the combination scheme COMB$(\tilde{\theta}, \tilde{\Sigma})$, and $\tilde{\Pi}_n^h(\cdot \mid Y_1^n)$ be the distribution of $h = \sqrt{n}(\theta - \tilde{\theta})$ implied by $\tilde{\Pi}_n^\theta(\cdot \mid Y_1^n)$. If Assumptions ($A_1$)–($A_9$) hold, then*

(*i*) *as $m \to \infty$,*

$$\|\tilde{\Pi}_n^h(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{TV} \to 0 \text{ in } \mathbb{P}_0\text{-probability.} \tag{3.6}$$

(*ii*) *Moreover, let $\hat{\theta}$ be the maximum likelihood estimator of θ that solves $L_n'(\theta) = 0$ and $\Pi_n^\theta(\cdot \mid Y_1^n)$ be the true posterior distribution of θ given Y_1^n. Then, up to a negligible term in $\mathbb{P}_0$-probability,*

$$n^{-1/2} diam(\Theta)^{-1} \Delta^2 \le \sqrt{n} W_1\{\tilde{\Pi}_n^\theta(\cdot \mid Y_1^n), \Pi_n^\theta(\cdot \mid Y_1^n)\} \le \Delta, \tag{3.7}$$

where $\Delta = \sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2$.

Theorem 3.2(*i*) shows that the block filtered posterior distribution defined by COMB$(\tilde{\theta}, \tilde{\Sigma})$ is asymptotically normal with mean parameter $\tilde{\theta}$ and covariance matrix $(nI_0)^{-1}$. Theorem 3.2(*ii*) specifies that the approximation accuracy of

block filtered posterior is determined by the asymptotic order of $\Delta = \sqrt{n}\|\tilde{\theta}-\hat{\theta}\|_2$. If $\Delta = o_{\mathbb{P}_0}(1)$, then the credible regions obtained using the block filtered and true posterior distributions match up to $o(n^{-1/2})$ in $\mathbb{P}_0$-probability as m, K tend to infinity, and the W_1 distance between them converges to zero in $\mathbb{P}_0$-probability at the optimal parametric rate $n^{1/2}$. The maximum likelihood estimator $\hat{\theta}$ in the theorem can be obtained efficiently using online EM (Cappé, 2011) or forward-backward recursions (Cappé et al., 2006). In our experiments, we compute $\hat{\theta}$ using the latter approach to attain the optimal parametric rate.

The optimal rate can also be achieved by setting $\tilde{\theta}$ to be any root-n consistent estimator of θ_0. This includes the maximum split data likelihood estimator of θ_0 in Rydén (1994) and related composite likelihood estimators of θ_0. If estimating a root-n consistent estimator of θ_0 is time consuming, then by setting $\tilde{\theta} = \hat{\theta}_j$ we have $\Delta = o_{\mathbb{P}_0}(\sqrt{\frac{n}{m}})$. In this case, Theorem 3.2(ii) shows that the W_1 distance between the block filtered posterior distribution and true posterior distribution converges to zero in $\mathbb{P}_0$-probability with sub-optimal rate $m^{1/2}$.

Consider the degenerate case in which elements of the hidden chain $\{X_t\}_{t=1}^n$ are independent. The complete data $\{X_t, Y_t\}_{t=1}^n$ are samples from a mixture model with S components, where X_t is the mixture component membership at time t. The total variance distance between prediction filters and initial distributions is exactly zero, which corresponds to a degenerate case of $\rho = 0$ in Assumption (A_8). In this case, the proof of Theorem 3.1 implies that the subset posterior inference using quasi-likelihood with prediction filter is equivalent to that using quasi-likelihood with initial distribution, and this equivalence removes the rate-optimal upper bound $K = o(\rho^{-m})$ in the dependent case, while having the same theoretical guarantees of block filtered posterior distribution. Furthermore, Theorem 3.2 recovers the existing results for divide-and-conquer Bayesian inference of independent data, where the number of subsets K and the size of subsets m tend to infinity. For the degenerate HMM, our theoretical setup of Theorem 3.2 imposes no restriction on the growth of K and m. This includes the case where $K = O(n^c)$ and $m = O(n^{1-c})$ for any $c \in (0, 1)$ as specified in Theorem 2 of Li et al. (2017).

4. Experiments

4.1. Setup

We compare the performance of block filtered posterior distribution with existing divide-and-conquer approaches to Bayesian inference. The true posterior distribution of θ, estimated using a Monte Carlo algorithm, serves as the benchmark in every comparison. We select Double Parallel Monte Carlo (Xue and Liang, 2019), Posterior Interval Estimation (Li et al., 2017), and Wasserstein Posterior (Srivastava et al., 2018) as our divide-and-conquer competitors. These competitors require modifications before they can be used for inference because their first two steps are designed for independent observations. The modified first step divides the observed data into K blocks of consecutive and non-overlapping observations. The modified second step uses the density in (2.5)

for drawing parameters on the subsets. Finally, the third step uses the original combination algorithms of the respective competitors. There is no software support for using Consensus Monte Carlo, so we excluded it in our comparisons (Scott et al., 2016). Following Theorem 3.2, the combination steps are specified by COMB$(\hat{\theta}, \overline{\Sigma})$ where the centering parameter $\hat{\theta}$ is the maximum likelihood estimator and is estimated using the EM (Baum-Welch) algorithm and the scaling matrix $\overline{\Sigma}$ is matrix barycenter defined in (2.9). All the sampling algorithms run for 10,000 iterations. The first 5000 samples are discarded as burn-in and the remaining chain is thinned by collecting every fifth sample. We have implemented all the algorithms in R and used an Oracle Grid Engine cluster with 2.6GHz 16 core compute nodes for all our experiments.

The accuracy metric for comparing the true posterior distribution and its approximation is based on the total variation distance. Let ξ be a one-dimensional parameter, $\pi(\xi \mid Y_1^n)$ be the true posterior density of ξ, and $\hat{\pi}(\xi \mid Y_1^n)$ be the density of its approximation. Following Li et al. (2017), the approximation accuracy of $\hat{\pi}(\xi \mid Y_1^n)$ is

$$\text{Acc}\left\{\hat{\pi}(\xi \mid Y_1^n)\right\} = 1 - \frac{1}{2}\int_{\mathbb{R}} \left|\hat{\pi}(\xi \mid Y_1^n) - \pi(\xi \mid Y_1^n)\right| d\xi. \tag{4.1}$$

The integral in (4.1) is the total variation distance between $\hat{\pi}(\xi \mid Y_1^n)$ and $\pi(\xi \mid Y_1^n)$. It is approximated numerically using the ξ draws from $\hat{\pi}(\xi \mid Y_1^n)$ and $\pi(\xi \mid Y_1^n)$, respectively, and kernel density estimation. For a $\theta = (\xi_1, \ldots, \xi_d)$, we extend (4.1) as the median of Acc $\{\hat{\pi}(\xi_i \mid Y_1^n)\}$ $(i = 1, \ldots d)$. If this metric is high, then $\hat{\pi}(\theta \mid Y_1^n)$ is an accurate estimator of $\pi(\theta \mid Y_1^n)$.

Hamiltonian Monte Carlo (HMC) algorithm, implemented using RStan (Carpenter et al., 2017), serves as the benchmark for comparing the accuracy and efficiency of the divide-and-conquer approaches, which are extensions of MCMC algorithms. The block filtered posterior's divide-and-conquer competitors are meant for independent data. If the accuracies of all the divide-and-conquer approaches are close to that of HMC, then block filtered posterior's practical gains are marginal relative to its competitors. The efficiency gains are compared in terms of run time. The lower the run time, the higher the efficiency gain. All the divide-and-conquer methods have similar run times, so the practical gains result from higher accuracies. We expect marginal practical gains for the block filtered posterior when K or S is small, or the dependence is relatively weak. In cases where either K or S is large, we expect the block filtered posterior's accuracy to be much closer to that of HMC and higher than that of its competitors. Due to the distributed computations, we expect the block-filtered posterior's run time to be shorter than that of HMC, indicating significant practical gains.

4.2. Simulated data analysis

Our simulation is based on an HMM with Gaussian emission densities, which has been used in Rydén (2008). Using the notation from Section 2.1, let $\mathcal{Y} = \mathbb{R}$, $g(\cdot \mid X = a) = \phi(\cdot; \mu_a, \sigma_a^2)$ be the density of Normal distribution with mean μ_a and

variance σ_a^2 $(a = 1, \ldots, S)$, $\theta = \{r_1, \ldots, r_S, Q_{11}, \ldots, Q_{SS}, \mu_1, \ldots, \mu_S, \sigma_1^2, \ldots, \sigma_S^2\}$, and $d = S^2 + 2S - 1$.

The prior distribution of θ is conjugate to the likelihood $p_\theta(X_1^n, Y_1^n)$ in (2.2). The initial distribution $r = (r_1, \ldots, r_S)$ and rows of the Markov transition matrix $\{(q_{a1}, \ldots, q_{aS})\}_{a=1}^S$ are given an independent Dirichlet prior $\mathrm{Dir}(1, \ldots, 1)$; the mean parameters $\mu_1, \ldots, \mu_S$ are given an independent Normal prior $N(\xi, \kappa^{-1})$ with $\xi = (\min y_i + \max y_i)/2$ and $\kappa = 1/(\max y_i - \min y_i)^2$; the variance parameters $\sigma_1^{-2}, \ldots, \sigma_S^{-2}$ are given an independent gamma prior with parameters $(1, 1)$. The division and subset posterior computation steps for the block filtered posterior distribution and its divide-and-conquer competitors are based on Sections 2.2 and 2.3. This example slightly violates our Assumption (A_8) (ii) but is particularly attractive because posterior computations on a subset using (2.5) are analytically tractable; see Appendix D.1 for greater details.

Simulation Study of Accuracy and Efficiency. This simulation study is aimed to analyze the performance of block filtered posterior and compare it with divide-and-conquer competitors. We set $S = 3$, $\mu_1 = -2$, $\mu_2 = 0$, $\mu_3 = 2$, and $\sigma_1 = \sigma_2 = \sigma_3 = 0.5$. Following Rydén (2008), Q is defined as $Q_{11} = 0.6$, $Q_{12} = 0.3$, $Q_{13} = 0.1$, $Q_{21} = 0.1$, $Q_{22} = 0.8$, $Q_{23} = 0.1$, $Q_{31} = 0.1$, $Q_{32} = 0.3$, and $Q_{33} = 0.6$. This choice of Q also determines the stationary distribution $r = (0.2, 0.6, 0.2)^{\mathrm{T}}$ because $r^{\mathrm{T}}Q = r^{\mathrm{T}}$. We replicate this simulation ten times.

We use three choices of K and n, respectively. For a given n, K is varied as $\log n$, $n^{1/4}$, and $n^{1/3}$, where the first and last two choices of K have been used to justify the asymptotic properties of Wasserstein Posterior and Posterior Interval Estimation, respectively. We vary n as $10^4, 10^5$, and 10^6. For a given K and n, the subset sample size is defined as $m = \lceil n/K \rceil$, where $\lceil x \rceil$ denotes the smallest integer z such that $z \geq x$. If $K = \log n$, then m is large, which implies that for $n = 10^6$, all the divide-and-conquer methods are slow due to a large m; on the other hand, they are most efficient when $K = n^{1/3}$.

The accuracy based on (4.1) of all the five methods are compared for posterior inference on $(\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$ (Tables 1). The divide-and-conquer competitors of the block filtered posterior have low accuracy for $n = 10^5, 10^6$ and any K. In comparison, the block filtered posterior distribution is the top performer for every choice of K and n and its accuracy is close to that of the Hamiltonian Monte Carlo. When n is large the block filtered posterior is almost ten times faster than Hamiltonian Monte Carlo and has better accuracy compared to divide-and-conquer competitors for all parameters. This shows that the block filtered posterior is an accurate and efficient algorithm for inference in HMMs under massive data settings; see Appendix D.2 for timing comparisons and accuracy for posterior inference on Q.

Simulation Study of Mixing Property. This simulation analyzes the impact of mixing coefficient ρ on the choice of K and accuracy of the block filtered posterior distribution. In the previous simulation study, we have $(S = 3, Q_\epsilon = 0.1, \mu_1 = -2, \mu_2 = 0, \mu_3 = 2)$. We consider three more HMMs with increasing value of mixing coefficients ρ in this simulation: $(S = 2, Q_\epsilon = 0.3, \mu_1 = -2, \mu_2 = 2)$, $(S = 5, Q_\epsilon = 0.05, \mu_1 = -4, \mu_2 = -2, \mu_3 = 0, \mu_4 = 2, \mu_5 = 4)$, and $(S = 7, Q_\epsilon = 0.05, \mu_1 = -8, \mu_2 = -4, \mu_3 = -2, \mu_4 = 0, \mu_5 = 2, \mu_6 = 4, \mu_7 = 8)$.

TABLE 1

Accuracy of the approximate posterior distributions of $(\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$. *The accuracy values are averaged over ten simulation replications and across all the parameter dimensions. The maximum Monte Carlo error for block filtered posterior distribution is 0.02*

n	K	BFP	WASP	DPMC	PIE	HMC
10^4	$\log n$	0.93	0.48	0.48	0.49	0.96
	$n^{1/4}$	0.93	0.48	0.48	0.48	
	$n^{1/3}$	0.92	0.37	0.36	0.37	
10^5	$\log n$	0.93	0.18	0.18	0.18	0.95
	$n^{1/4}$	0.93	0.18	0.18	0.18	
	$n^{1/3}$	0.92	0.18	0.18	0.18	
10^6	$\log n$	0.93	0.14	0.14	0.15	0.96
	$n^{1/4}$	0.93	0.12	0.12	0.13	
	$n^{1/3}$	0.93	0.14	0.14	0.14	

[1] BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

For every HMM, we set $\sigma_a = 0.5$ for $1 \leq a \leq S$ and vary n as 10^4 and 10^5. For a given n, K is varied as $\log n$, $n^{1/4}$, and $n^{1/3}$. We replicate this simulation ten times and benchmark the performance of divide-and-conquer approaches relative to Hamiltonian Monte Carlo.

The accuracy of block filtered posterior distribution for inference on the emission distribution parameters follows from the results in Theorem 3.1 (Tables 1 and 2). The dependence in HMMs increases with ρ. The HMMs with $S = 5, 7$ have $\rho \approx 1$, resulting in a relatively higher dependence than the other two HMMs with $S = 2, 3$. Our theoretical results imply that the accuracy of the block filtered posterior is fairly insensitive to the choice of K in the latter two HMMs than the former two. Indeed, Hamiltonian Monte Carlo and block filtered posterior have similar accuracy for every (n, K) combinations when $S = 2, 3$. On the other hand, the accuracy of block filtered posterior is smaller than that of Hamiltonian Monte Carlo and decreases with increasing K when $S = 5, 7$ due to the increased dependence. Our theoretical results also imply that the accuracy depends only on ρ and not on S. This is further confirmed by our empirical results in that the accuracy values for the HMMs with $S = 5, 7$ are very similar due to almost equal ρ values.

The block filtered posterior distribution outperforms its divide-and-conquer competitors for every ρ and K. The competing divide-and-conquer methods are developed for independent data, so their accuracy values are relatively large when $S = 2$, ρ is small, and the dependence is weak; however, their accuracy values quickly drop as dependence increases with ρ, reducing to 0 when $S = 5, 7$. On the other hand, the block filtered posterior distribution is designed for dependent data and maintains its superior performance over the divide-and-conquer competitors for every ρ. Relative to its divide-and-conquer competitors, the block filtered posterior distribution is also very robust to choice of K and its accuracy is impacted only when the subset size is very small (for example, $K = O(n^{1/3})$ and $n = 10^4$); see Appendix D.2 for accuracy for posterior inference

on Q. We conclude based on this simulation that the block filtered posterior distribution provides accurate posterior inference using the divide-and-conquer technique for a broad range of ρ values.

TABLE 2

Accuracy of the approximate posterior distributions of $(\mu_1, \ldots, \mu_S, \sigma_1, \ldots, \sigma_S)$. *The accuracy values are averaged over ten simulation replications and across all the parameter dimensions. The maximum Monte Carlo error for block filtered posterior distribution is 0.02.*

n	K	BFP	WASP	DPMC	PIE	HMC
$S = 2$						
10^4	$\log n$	0.97	0.64	0.64	0.65	0.95
	$n^{1/4}$	0.97	0.64	0.64	0.64	
	$n^{1/3}$	0.97	0.61	0.61	0.61	
10^5	$\log n$	0.97	0.46	0.46	0.46	0.95
	$n^{1/4}$	0.97	0.46	0.46	0.46	
	$n^{1/3}$	0.97	0.45	0.45	0.45	
$S = 5$						
10^4	$\log n$	0.83	0.00	0.00	0.00	0.95
	$n^{1/4}$	0.82	0.00	0.00	0.00	
	$n^{1/3}$	0.82	0.00	0.00	0.00	
10^5	$\log n$	0.82	0.00	0.00	0.00	0.95
	$n^{1/4}$	0.79	0.00	0.00	0.00	
	$n^{1/3}$	0.80	0.00	0.00	0.00	
$S = 7$						
10^4	$\log n$	0.80	0.00	0.00	0.00	0.93
	$n^{1/4}$	0.77	0.00	0.00	0.00	
	$n^{1/3}$	0.69	0.00	0.00	0.00	
10^5	$\log n$	0.82	0.00	0.00	0.00	0.94
	$n^{1/4}$	0.81	0.00	0.00	0.00	
	$n^{1/3}$	0.81	0.00	0.00	0.00	

BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

4.3. Real data analysis

We illustrate the application of block filtered posterior distribution in a real-life setting using the data of transaction information made through the purchase card (PCard) program in the state of Oklahoma. We download the data reported in fiscal year 2013 from https://data.ok.gov/dataset/purchase-card-pcard-fiscal-year-2013. The data contains information of transaction amount, purchase description, transaction date, and merchant category in the previous year (2012). To remove the edge effect at the end of the year, we select the observations from 2012-07-01 to 2012-12-20 and remove the purchase amount larger than \$1000 resulting in $n = 184,303$ observations. Due to the data collection approach, multiple observations on some days have "00:00:00" as their transaction times; therefore, we treat multiple transactions in a single day as consecutive observations from a discrete time series that is arranged

according to the order of transactions posted in the data set.

The purchase amount is the observed Y variable and the merchant category is the latent discrete X variable. We de-trend the time series using Tukey's running median smoother as implemented in the *smooth* R-function and de-seasonalize it using seasonal decomposition by loess as implemented in *stl* R-function. Motivated by an annual report of PCard program audit in April 2015 (McCoy et al., 2015), we model this data set using an HMM with $S = 3$ where the three categories correspond to (1) miscellaneous and speciality retail stores; (2) services, catalog merchant, IT, utilities; and (3) airline, hotel and others. Replicating the simulation setup, we use data augmentation and Hamiltonian Monte Carlo for drawing θ and choose $K = 10$ ($\log(n) \approx 12$), 20 ($n^{1/4} \approx 21$), and 50 ($n^{1/3} \approx 57$) for the divide-and-conquer methods. Unlike our simulation setup, we use Hamiltonian Monte Carlo for sampling on the subsets.

The results of real and simulated data analyses in this section and Section 4.2 agree closely (Table 3). For $K = 20, 50$, the block filtered posterior has better accuracy than its divide-and-conquer competitors for inference on the emission distribution parameters for every K, and its accuracy is comparable to that of Hamiltonian Monte Carlo. The accuracy estimates for posterior inference on Q are even better; see Appendix D.2. The block filtered posterior distribution is more efficient than data augmentation and Hamiltonian Monte Carlo for every K, and its efficiency increases with K (Table 4). Noticeably, when $K = 50$, the block filtered posterior is over 20 times faster than Hamiltonian Monte Carlo and 40 times faster than data augmentation. These observations are very similar to that of our simulation study, so we conclude that the block filtered posterior distribution provides an accurate and efficient alternative for principled Bayesian inference in modeling the PCard data.

TABLE 3

Accuracy of the approximate posterior distributions of $(\mu_1, \mu_2, \mu_3, \sigma_1, \sigma_2, \sigma_3)$. The entries in the table are the median of the accuracy values computed across the three dimensions. The maximum Monte Carlo error for the block filtered posterior distribution is 0.01

K	BFP	WASP	DPMC	PIE	HMC
$10(\log n)$	0.88	0.80	0.80	0.80	
$20(n^{1/4})$	0.87	0.70	0.70	0.71	0.90
$50(n^{1/3})$	0.82	0.52	0.52	0.51	

BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

TABLE 4

Time (in hours) for computing the posterior distributions of θ

DA	BFP			HMC
	$10(\log n)$	$20(n^{1/4})$	$50(n^{1/3})$	
26.89	2.73	1.49	0.67	17.42

DA, data augmentation; BFP, block filtered posterior distribution; HMC, Hamiltonian Monte Carlo.

5. Discussion

Our focus has been on developing a divide-and-conquer Monte Carlo algorithm for Bayesian analysis of parametric HMMs where the state space is known and discrete; however, our Monte Carlo algorithm and its theoretical properties can be extended to general state space models, where the state space is continuous and known. This requires minimal changes in the partitioning, subset sampling, and combining steps. Our theoretical results can also be extended to general state space models using the ideas of Jensen and Petersen (1999), who show that the theoretical results of Bickel et al. (1998) for general HMMs extend naturally to general state space models.

Appendix A: Proofs of Proposition 3.1 and technical lemmas

We begin with several important lemmas and the proof of Proposition 3.1.

A.1. Preliminaries

Let $\mathcal{P}(\mathcal{X})$ denote the space of all probability measure on $\mathcal{X}$ equipped with total variation distance. For any $\theta \in \Theta$, the prediction filter $p_t^\theta \in \mathcal{P}(\mathcal{X})$ is defined as

$$p_t^\theta = \left\{\mathbb{P}_\theta(X_t = 1 \mid Y_1^{t-1}), \ldots, \mathbb{P}_\theta(X_t = S \mid Y_1^{t-1})\right\}^{\mathrm{T}}, \quad t = 1, \ldots, n. \tag{A.1}$$

The Baum's forward equation implies that for $t = 1, \ldots, n-1$,

$$\begin{aligned} p_{t+1}^\theta &= \frac{\mathbb{P}_\theta(X_{t+1} = \cdot, Y_t \mid Y_1^{t-1})}{\mathbb{P}_\theta(Y_t \mid Y_1^{t-1})} = \frac{Q_\theta^{\mathrm{T}} G_\theta(Y_t) p_t^\theta}{g_\theta(Y_t)^{\mathrm{T}} p_t^\theta} \equiv f^\theta(Y_t, p_t^\theta), \\ g_\theta(Y_t) &= \{g_\theta(Y_t \mid X_t = 1), \ldots, g_\theta(Y_t \mid X_t = S)\}^{\mathrm{T}}, \end{aligned} \tag{A.2}$$

where $G_\theta(Y_t) = \mathrm{diag}\{g_\theta(Y_t)\}$ and Q_θ is the transition matrix of $\{X_t\}$ with $(Q_\theta)_{ab} = q_\theta(a, b)$. For $\theta \in \Theta$, f^θ is a measurable function on $\mathcal{Y} \times \mathcal{P}(\mathcal{X})$. Extending the definition of $f^\theta(Y_t, p_t^\theta)$ for lags $s = 0, 1, \ldots, t-1$, we recursively define p_{t+1}^θ for $s = 0, \ldots, t-1$ as

$$p_{t+1}^\theta = f_0^\theta(Y_t^t, p_t^\theta) = f_1^\theta(Y_{t-1}^t, p_{t-1}^\theta) = \cdots = f_s^\theta(Y_{t-s}^t, p_{t-s}^\theta) = \cdots = f_{t-1}^\theta(Y_1^t, p_1^\theta), \tag{A.3}$$

where $f_0^\theta = f^\theta$.

A.2. Technical lemmas

We state five technical lemmas that are used to prove Theorem 3.1 and Theorem 3.2 in the main manuscript. The first lemma is a restatement of Lemma 3.1 in De Gunst and Shcherbakova (2008), except the full data likelihood and n are replaced by the jth subset likelihood and m.

Lemma A.1 (De Gunst and Shcherbakova (2008), Lemma 3.1)**.** *For any $\delta > 0$ and $j \in \{1, \ldots, K\}$, there exists $\epsilon > 0$ such that*

$$\mathbb{P}_0 \left[\sup_{\|\theta - \theta_0\|_2 > \delta} \frac{1}{m} \{L_{jm}(\theta) - L_{jm}(\theta_0)\} \leq -\epsilon \right] \to 1 \text{ as } m \to \infty.$$

Lemma A.2 (Le Gland and Mevel (2000) Theorem 2.1)**.** *Under Assumption (A_8), for any $\theta \in B_{\delta_0}(\theta_0)$, any $p_l^\theta, \tilde{p}_l^\theta \in \mathcal{P}(\mathcal{X})$, any $l, s \geq 1$ and sequence $y_l, \ldots, y_{l+s} \in \mathcal{Y}^s$,*

$$\begin{aligned} &\|f_s^\theta\{y_{l+s}, \ldots, y_l, p_s^\theta\} - f_s^\theta\{y_{l+s}, \ldots, y_l, \tilde{p}_s^\theta\}\|_{TV} \\ &\quad \leq \epsilon_\theta^{-1} h_\theta(y_i) \prod_{i=l+1}^{l+s} \{1 - \epsilon_\theta h_\theta(y_i)^{-1}\} \|p_l^\theta - \tilde{p}_l^\theta\|_{TV}, \end{aligned}$$

where $\epsilon_\theta = \min_{a,b \in \mathcal{X}} q_\theta(a, b)$, $\rho_\theta(y) = \max_{a,b \in \mathcal{X}} \frac{g_\theta(y|a)}{g_\theta(y|b)}$ and $\{1 - \epsilon_\theta h_\theta(y_i)^{-1}\} \in (0, 1)$.

Lemma A.3 (Le Gland and Mevel (2000) Example 3.3)**.** *Under Assumption (A_8), for any $\theta \in B_{\delta_0}(\theta_0)$, any $y \in \mathcal{Y}$ and any $p, \tilde{p} \in \mathcal{P}(\mathcal{X})$,*

$$|\log\{\sum_{a \in \mathcal{X}} p(a) g_\theta(y|a)\} - \log\{\sum_{a \in \mathcal{X}} \tilde{p}(a) g_\theta(y|a)\}| \leq [h_\theta(y) - 1] \|p - \tilde{p}\|_{TV}.$$

Lemma A.4. *The log-likelihood function of θ given Y_{m+1}^{2m} and the conditional log-likelihood of θ obtained from the conditional density of Y_{m+1}^{2m} given Y_1^m can be expressed as*

$$\begin{aligned} \exp\{w_1(\theta)\} &= p_\theta(Y_{m+1}^{2m}) \\ &= \sum_{X_{m+1}^{2m}} \prod_{t=m+1}^{2m-1} r_\theta(X_{m+1}) q_\theta(X_t, X_{t+1}) \prod_{t=m+1}^{2m} g_\theta(Y_t \mid X_t) \\ &= \prod_{t=m+1}^{2m} \sum_{X_t} g_\theta(Y_t \mid X_t) p_t^\theta(X_t), \end{aligned} \tag{A.4}$$

$$\begin{aligned} \exp\{w_2(\theta)\} &= p_\theta(Y_{m+1}^{2m} \mid Y_1^m) \\ &= \sum_{X_{m+1}^{2m}} \prod_{t=m+1}^{2m-1} p_\theta(X_{m+1} \mid Y_1^m) q_\theta(X_t, X_{t+1}) \prod_{t=m+1}^{2m} g_\theta(Y_t \mid X_t) \\ &= \prod_{t=m+1}^{2m} \sum_{X_t} g_\theta(Y_t \mid X_t) \tilde{p}_t^\theta(X_t), \end{aligned} \tag{A.5}$$

where

$$p_{m+1}^\theta(X_{m+1}) = r_\theta(X_{m+1}), \quad \tilde{p}_{m+1}^\theta(X_{m+1}) = p_\theta(X_{m+1} \mid Y_1^m),$$

and for $t = m+2, \ldots, 2m$,

$$p_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, r_\theta(X_{m+1})\}, \quad \tilde{p}_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, p_\theta(X_{m+1} \mid Y_1^m)\},$$

where f_{t-m-2}^θ *is defined in* (A.3).

Proof of Lemma A.4. We first prove the following equality by induction:

$$\sum_{X_1^m} \prod_{t=1}^{m-1} r_\theta(X_1) q_\theta(X_t, X_{t+1}) \prod_{t=1}^{m} g_\theta(Y_t \mid X_t) = \prod_{t=1}^{m} \sum_{X_t} g_\theta(Y_t|X_t) p_t^\theta(X_t). \quad \text{(A.6)}$$

For $m = 2$, the left hand side of (A.6) is

$$\sum_{X_1, X_2} r_\theta(X_1) q_\theta(X_1, X_2) g_\theta(Y_1|X_1) g_\theta(Y_2|X_2)$$
$$= \sum_{X_2} g_\theta(Y_2|X_2) \sum_{X_1} r_\theta(X_1) q_\theta(X_1, X_2) g_\theta(Y_1|X_1).$$

And the right hand side of (A.6) is

$$\left\{\sum_{X_1} r_\theta(X_1) g_\theta(Y_1 \mid X_1)\right\}\left\{\sum_{X_2} g_\theta(Y_2 \mid X_2) p_\theta(X_2 \mid Y_1)\right\} = \sum_{X_2} g_\theta(Y_2 \mid X_2)$$
$$\times \sum_{X_1} p_\theta(X_2 \mid Y_1) r_\theta(X_1) g_\theta(Y_1 \mid X_1).$$

For fixed $X_2 = x_2 \in \mathcal{X}$, we have

$$\sum_{X_1} r_\theta(X_1) q_\theta(X_1, x_2) g_\theta(Y_1 \mid X_1) - \sum_{X_1} p_\theta(x_2 \mid Y_1) r_\theta(X_1) g_\theta(Y_1 \mid X_1)$$
$$= \sum_{X_1} r_\theta(X_1) q_\theta(X_1, x_2) g_\theta(Y_1 \mid X_1)$$
$$- \sum_{X_1} r_\theta(X_1) g_\theta(Y_1 \mid X_1) \frac{\sum_{a_1} q_\theta(a_1, x_2) g_\theta(Y_1 \mid a_1) r_\theta(a_1)}{\sum_{a_1} g_\theta(Y_1 \mid a_1) r_\theta(a_1)}$$
$$= \frac{1}{\sum_{a_1} g_\theta(Y_1 \mid a_1) r_\theta(a_1)}$$
$$\times \left\{\sum_{X_1} r_\theta(X_1) q_\theta(X_1, x_2) g_\theta(Y_1 \mid X_1) \sum_{a_1} g_\theta(Y_1 \mid a_1) r_\theta(a_1)\right.$$
$$\left. - \sum_{X_1} r_\theta(X_1) g_\theta(Y_1 \mid X_1) \sum_{a_1} q_\theta(a_1, x_2) g_\theta(Y_1 \mid a_1) r_\theta(a_1)\right\}$$
$$= 0.$$

Hence, we have

$$\sum_{X_2} g_\theta(Y_2 \mid X_2) \left\{\sum_{X_1} r_\theta(X_1) q_\theta(X_1, X_2) g_\theta(Y_1 \mid X_1)\right.$$

$$-\sum_{X_1} p_\theta(X_2 \mid Y_1) r_\theta(X_1) g_\theta(Y_1 \mid X_1) \Bigg\} = 0,$$

which proves (A.6) at $m = 2$. Suppose (A.6) holds for $m = k - 1 \geq 2$, we have

$$\sum_{X_1^{k-1}} \prod_{t=1}^{k-2} r_\theta(X_1) q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t | X_t) = \prod_{t=1}^{k-1} \sum_{X_t} g_\theta(Y_t \mid X_t) p_t^\theta(X_t).$$

For $m = k$, the left hand side of (A.6) can be expressed as

$$\sum_{X_1^k} \prod_{t=1}^{k-1} r_\theta(X_1) q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k} g_\theta(Y_t \mid X_t)$$

$$= \sum_{X_k} g_\theta(Y_k \mid X_k) \sum_{X_1^{k-1}} q(X_{k-1}, X_k) r_\theta(X_1) \prod_{t=1}^{k-2} q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t). \square$$

And the right hand side of (A.6) can be expressed as

$$\prod_{t=1}^{k} \sum_{X_t} g_\theta(Y_t \mid X_t) p_t^\theta(X_t)$$

$$= \left\{ \sum_{X_k} g_\theta(Y_k \mid X_k) p_\theta(X_k \mid Y_1^{k-1}) \right\} \left\{ \prod_{t=1}^{k-1} \sum_{X_t} g_\theta(Y_t \mid X_t) p_t^\theta(X_t) \right\}$$

$$= \left\{ \sum_{X_k} g_\theta(Y_k \mid X_k) p_\theta(X_k \mid Y_1^{k-1}) \right\}$$

$$\times \left\{ \sum_{X_1^{k-1}} \prod_{t=1}^{k-2} r_\theta(X_1) q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t) \right\},$$

where the last equality is due to the induction assumption at $m = k - 1$. And we have

$$p_\theta(X_k \mid Y_1^{k-1}) = \frac{\sum_{a_1^{k-1}} q_\theta(a_{k-1}, X_k) r_\theta(a_1) \prod_{t=1}^{k-2} q_\theta(a_t, a_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid a_t)}{\sum_{a_1^{k-1}} r_\theta(a_1) \prod_{t=1}^{k-2} q_\theta(a_t, a_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid a_t)}.$$

For fixed $X_k = x_k \in \mathcal{X}$, we have

$$\left\{ \sum_{X_1^{k-1}} q(X_{k-1}, x_k) r_\theta(X_1) \prod_{t=1}^{k-2} q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t) \right\}$$

$$\times \left\{ \sum_{a_1^{k-1}} r_\theta(a_1) \prod_{t=1}^{k-2} q_\theta(a_t, a_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid a_t) \right\}$$

$$= \left\{ \sum_{X_1^{k-1}} r_\theta(X_1) \prod_{t=1}^{k-2} q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t) \right\}$$
$$\times \left\{ \sum_{a_1^{k-1}} q_\theta(a_{k-1}, x_k) r_\theta(a_1) \prod_{t=1}^{k-2} q_\theta(a_t, a_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid a_t) \right\}.$$

Hence, we have

$$\sum_{X_1^{k-1}} q(X_{k-1}, X_k) r_\theta(X_1) \prod_{t=1}^{k-2} q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t)$$
$$= p_\theta(X_k \mid Y_1^{k-1}) \left\{ \sum_{X_1^{k-1}} \prod_{t=1}^{k-2} r_\theta(X_1) q_\theta(X_t, X_{t+1}) \prod_{t=1}^{k-1} g_\theta(Y_t \mid X_t) \right\},$$

which implies that (A.6) holds for $m = k$. Changing indices from $(1, \ldots, m)$ to $(m+1, \ldots, 2m)$, we prove equation (A.4).

By the recursive definition of prediction filter, we have for $t = m+2, \ldots, 2m$,

$$p_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, r_\theta(X_{m+1})\}, \quad \tilde{p}_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, p_\theta(X_{m+1} \mid Y_1^m)\}.$$

Given Y_1^m, replacing the stationary distribution $r_\theta(X_{m+1})$ by the prediction filter distribution $p_t(X_{m+1} \mid Y_1^m)$, we prove equation (A.5). □

Lemma A.5 (Devroye et al. (2018) Theorem 1.2)**.** *Suppose $d > 1$, let $\mu_1 \neq \mu_2 \in \mathbb{R}^d$ and let V_1, V_2 be two positive definite $d \times d$ matrices. Let $v = \mu_1 - \mu_2$ and N be an arbitrary $d \times d-1$ matrix with columns forming a basis for the subspace orthogonal to v. Define the function*

$$tv(\mu_1, V_1, \mu_2, V_2) = \max\{\frac{|v^{\mathrm{T}}(V_1 - V_2)v|}{v^{\mathrm{T}} V_1 v}, \frac{v^{\mathrm{T}} v}{\sqrt{v^{\mathrm{T}} V_1 v}}, \|(N^{\mathrm{T}} V_1 N)^{-1} N^{\mathrm{T}} V_2 N - I_{d-1}\|_F\}.$$

Then, we have

$$\frac{1}{200} \leq \frac{\|N_d(\mu_1, V_1) - N_d(\mu_2, V_2)\|_{TV}}{tv(\mu_1, V_1, \mu_2, V_2)} \leq \frac{9}{2}.$$

A.3. Proof of Proposition 3.1

Proof of Proposition 3.1. For $j = 2$, the total variation distance equals to zero. Based on (A.3), for $j = 3, \ldots, K$, we have the following expressions for the approximate prediction filter and the true prediction filter:

$$p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m}) = f_{m-1}^\theta\{Y_{(j-2)m+1}^{(j-1)m}, r_\theta(X_{(j-2)m+1})\}$$
$$p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m}) = f_{m-1}^\theta\{Y_{(j-2)m+1}^{(j-1)m}, p_\theta(X_{(j-2)m+1} \mid Y_1^{(j-2)m})\}, \tag{A.7}$$

where $r_\theta(\cdot) \in \mathcal{P}(\mathcal{X})$ is the stationary distribution of $\{X_t, t \geq 1\}$. Then for $\theta \in \Theta$, we have that with $\mathbb{P}_0$-probability almost 1,

$$\begin{aligned}
&\|p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m}) - p_\theta(X_{(j-1)m+1} \mid Y_1^{(j-1)m})\|_{\mathrm{TV}} \\
&\overset{(i)}{\leq} \epsilon_\theta^{-1} h_\theta(Y_{(j-2)m+1}) \prod_{i=(j-2)m+2}^{(j-1)m} [1 - \epsilon_\theta h_\theta(Y_i)^{-1}] \|r_\theta(X_{(j-2)m+1}) \\
&\qquad - p_\theta(X_{(j-2)m+1} \mid Y_1^{(j-2)m})\|_{\mathrm{TV}} \\
&\overset{(ii)}{\leq} \epsilon^{-1} M(1-\epsilon M^{-1})^{m-1} \sup_{\theta \in B_{\delta_0}(\theta_0)} \|r_\theta(X_{(j-2)m+1}) - p_\theta(X_{(j-2)m+1} \mid Y_1^{(j-2)m})\|_{\mathrm{TV}} \\
&\overset{(iii)}{\leq} \epsilon^{-1} M\{(1-\epsilon M^{-1})\}^{m-1} S \rho^{(j-2)m-2}, \qquad \text{(A.8)}
\end{aligned}$$

where (i) follows from Lemma A.2, (ii) follows from that $\epsilon = \inf_{\theta \in \Theta} \epsilon_\theta$, and from (A_8)(ii), $h_\theta(Y_i) < M$ for $\theta \in \Theta$ and $i = 1, \ldots, n$ $\mathbb{P}_0$-almost surely. And (iii) follows from assumption (A_8)(iii). Since $(1-\epsilon M^{-1}), \rho \in (0,1)$ $\mathbb{P}_0$-almost surely, as $m \to \infty$, the right side of (A.8) converges to zero $\mathbb{P}_0$-almost surely.

Appendix B: Proof of Theorem 3.1

We first prove Theorem 3.1(i) following the arguments used for proving Theorem 1 in Li et al. (2017) and Theorem 2.1 in De Gunst and Shcherbakova (2008). The proof requires some essential technical changes to account for the definition of subset posterior distributions using (quasi) conditional distribution likelihood in (2.5) in the main manuscript. Then, we prove Theorem 3.1(ii) using results from Theorem 2.1 in Le Gland and Mevel (2000).

B.1. Proof of Theorem 3.1(i)

Step 1: Show that $\hat{\theta}_1$ is a weakly consistent estimator of θ_0. Assumption (A_5) implies that θ_0 is in the interior of Θ, which is compact. Assumption (A_1) implies that $L'_{jm}(\theta)$ exists and is continuously differentiable in the compact neighborhood $B_{\delta_0}(\theta_0)$ of θ_0 for every $(Y_{(j-1)m+1}, \ldots, Y_{jm})$ such that the components of $L''_{jm}(\theta)$ are uniformly bounded by an envelope function $M(Y_{(j-1)m+1}, \ldots, Y_{jm})$. Furthermore, $\mathbb{E}_0\left\{L'_{jm}(\theta_0)\right\} = 0$, so there exists a root of the equation $L'_{jm}(\theta) = 0$ with large $\mathbb{P}_0$-probability in $B_{\delta_0}(\theta_0)$ neighborhood. Denote this root as $\hat{\theta}_j$. Then, the weak consistency of $\hat{\theta}_j$ is a consequence of Lemma A.1.

Step 2: Simplify the statement of Theorem 3.1 (i). For any $t \in \mathbb{R}^d$, define

$$\begin{aligned}
\theta_t &= \hat{\theta}_1 + \frac{t}{(Km)^{1/2}}, \qquad \text{(B.1)} \\
w(t) &= L_{1m}\left\{\hat{\theta}_1 + \frac{t}{(Km)^{1/2}}\right\} - L_{1m}(\hat{\theta}_1),
\end{aligned}$$

$$C_m = \int_{\mathbb{R}^d} e^{Kw(z)} \pi \left\{ \hat{\theta}_1 + \frac{z}{(Km)^{1/2}} \right\} dz,$$
$$g_m(t) = e^{Kw(t)} \pi \left\{ \hat{\theta}_1 + \frac{t}{(Km)^{1/2}} \right\} - \exp\left(-\frac{t^{\mathrm{T}} I_0 t}{2} \right) \pi(\theta_0).$$

For proving Theorem 3.1 (i), we need to show that the sequence of random variable

$$\zeta_m = \int_{\mathbb{R}^d} \left| \frac{e^{Kw(t)} \pi \left\{ \hat{\theta}_1 + \frac{t}{(Km)^{1/2}} \right\}}{C_m} - \frac{1}{(2\pi)^{d/2} \det(I_0)^{-1/2}} \exp\left(-\frac{1}{2} t^{\mathrm{T}} I_0 t \right) \right| dt \tag{B.2}$$

converges to 0 in $\mathbb{P}_0$-probability as $m \to \infty$. Using the definition of $g_m(t)$ in (B.1),

$$\zeta_m = \int_{\mathbb{R}^d} \left| \frac{g_m(t) + \exp\left(-\frac{t^{\mathrm{T}} I_0 t}{2} \right) \pi(\theta_0)}{C_m} - \frac{1}{(2\pi)^{d/2} \det(I_0)^{-1/2}} \exp\left(-\frac{1}{2} t^{\mathrm{T}} I_0 t \right) \right| dt,$$

and the triangle inequality implies that

$$\zeta_m \le \frac{1}{C_m} \int_{\mathbb{R}^d} |g_m(t)| dt + \left| \frac{(2\pi)^{d/2} \det(I_0)^{-1/2} \pi(\theta_0)}{C_m} - 1 \right| \times \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2} \det(I_0)^{-1/2}} \exp\left(-\frac{1}{2} t^{\mathrm{T}} I_0 t \right) dt. \tag{B.3}$$

The definition of C_m in (B.1) also implies that

$$\begin{aligned} C_m &= \int_{\mathbb{R}^d} \left\{ g_m(z) + \exp\left(-\frac{z^{\mathrm{T}} I_0 z}{2} \right) \pi(\theta_0) \right\} dz \\ &= \int_{\mathbb{R}^d} g_m(z) dz + (2\pi)^{d/2} \det(I_0)^{-1/2} \pi(\theta_0). \end{aligned} \tag{B.4}$$

If $\int_{\mathbb{R}^d} |g_m(z)| dz \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$, then (B.4) implies that

$$\left| C_m - (2\pi)^{d/2} \det(I_0)^{-1/2} \pi(\theta_0) \right| \to 0 \text{ in } \mathbb{P}_0\text{-probability as } m \to \infty,$$

which after an application of Slutsky's theorem implies that

$$\left| \frac{(2\pi)^{d/2} \det(I_0)^{-1/2} \pi(\theta_0)}{C_m} - 1 \right| \to 0 \text{ in } \mathbb{P}_0\text{-probability as } m \to \infty; \tag{B.5}$$

therefore, the second term on the right hand side in (B.3) converges to 0 in $\mathbb{P}_0$-probability as $m \to \infty$. This also shows that proving $\zeta_m \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ is equivalent to showing that $\int_{\mathbb{R}^d} |g_m(t)| dt \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$.

Step 3: Show that $\int_{\mathbb{R}^d} |g_m(t)|dt \to 0$ *in* $\mathbb{P}_0$*-probability as* $m \to \infty$. Consider a θ_m^* in the interior of Θ that is close to $\hat{\theta}_1$. The Taylor expansion of $L_{1m}(\theta_m^*)$ at $\hat{\theta}_1$ is

$$L_{1m}(\theta_m^*) = L_{1m}(\hat{\theta}_1) + (\theta_m^* - \hat{\theta}_1)^{\mathrm{T}} L'_{1m}(\hat{\theta}_1) + \frac{1}{2}(\theta_m^* - \hat{\theta}_1)^{\mathrm{T}} L''_{1m}(\overline{\theta}_m)(\theta_m^* - \hat{\theta}_1), \tag{B.6}$$

where $\overline{\theta}_m$ is between θ_m^* and $\hat{\theta}_1$ and $L'_{1m}(\hat{\theta}_1) = 0$ because $\hat{\theta}_1$ is the MLE of θ for the first subset. Define

$$\epsilon_m(\overline{\theta}_m) = \frac{I_0}{2} + \frac{L''_{1m}(\overline{\theta}_m)}{2m},$$

and rewrite B.6 as

$$L_{1m}(\theta_m^*) = L_{1m}(\hat{\theta}_1) - \frac{m}{2}(\theta_m^* - \hat{\theta}_1)^{\mathrm{T}} I_0(\theta_m^* - \hat{\theta}_1) + m(\theta_m^* - \hat{\theta}_1)^{\mathrm{T}} \epsilon_m(\overline{\theta}_m)(\theta_m^* - \hat{\theta}_1). \tag{B.7}$$

If θ_m^* is a (possibly stochastic) sequence converging to θ_0, then $\epsilon_m(\overline{\theta}_m) \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ due to Assumption (A_4).

We use this Taylor expansion in (B.7) to show that $\int_{\mathbb{R}^d} |g_m(t)|dt \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$. We start by partitioning the domain of the integral into three sets: $A_1 = \{t \in \mathbb{R}^d : \|t\|_2 \geq \delta_1(Km)^{1/2}\}$, $A_2 = \{t \in \mathbb{R}^d : \delta_2 \leq \|t\|_2 < \delta_1(Km)^{1/2}\}$ and $A_3 = \{t \in \mathbb{R}^d : \|t\|_2 < \delta_2\}$, where δ_1, δ_2 are positive quantities to be chosen later. Using the triangular inequality,

$$\int_{\mathbb{R}^d} |g_m(t)|dt \leq \int_{A_1} |g_m(t)|dt + \int_{A_2} |g_m(t)|dt + \int_{A_3} |g_m(t)|dt.$$

The proof is complete if $\int_{A_i} |g_m(t)|dt \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ for $i = 1, 2, 3$.

Step 3.1: Show that $\int_{A_1} |g_m(t)|dt \to 0$ *in* $\mathbb{P}_0$*-probability as* $m \to \infty$. Let $\theta_m^* = \hat{\theta}_1 + t/(Km)^{1/2}$ in (B.7) for some $t \in A_1$. Using Lemma A.1, there exists an ϵ_1 depending on δ_1 such that

$$L_{1m}\{\hat{\theta}_1 + t/(Km)^{1/2}\} - L_{1m}(\theta_0) \leq -m\epsilon_1 \tag{B.8}$$

with $\mathbb{P}_0$-probability approaching 1 as $m \to \infty$. Because $\hat{\theta}_1$ is a weakly consistent estimator θ_0 and $L_{1m}(\theta)$ is a continuous function of θ, $L_{1m}(\hat{\theta}_1) \to L_{1m}(\theta_0)$ with $\mathbb{P}_0$-probability approaching 1 as $m \to \infty$. Using this in (B.8), we have that for any $t \in A_1$ and $\delta_1 > 0$ there exists an ϵ_1 depending on δ_1 such that

$$w(t) = L_{1m}\{\hat{\theta}_1 + t/(Km)^{1/2}\} - L_{1m}(\hat{\theta}_1) \leq -m\epsilon_1 \tag{B.9}$$

with $\mathbb{P}_0$-probability approaching 1 as $m \to \infty$, where $w(t)$ is defined in (B.1); therefore, with $\mathbb{P}_0$-probability approaching 1 as $m \to \infty$,

$$\int_{A_1} |g_m(t)|dt \leq \int_{A_1} e^{Kw(t)} \pi\left\{\hat{\theta}_1 + \frac{t}{(Km)^{1/2}}\right\} dt + \int_{A_1} \exp\left(-\frac{t^{\mathrm{T}} I_0 t}{2}\right) \pi(\theta_0)dt$$

$$\begin{aligned}
&\overset{(i)}{\leq} \exp(-Km\epsilon_1)\int_{A_1}\pi\left\{\hat{\theta}_1+\frac{t}{(Km)^{1/2}}\right\}dt+\pi(\theta_0)\int_{A_1}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{2}\right)dt\\
&\overset{(ii)}{\leq} \exp(-Km\epsilon_1)(Km)^{d/2}\int_{\|\theta-\hat{\theta}_1\|_2\geq\delta_1}\pi(\theta)d\theta+\pi(\theta_0)\\
&\qquad\times\int_{\|t\|_2\geq\delta_1(Km)^{1/2}}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{2}\right)dt\\
&\overset{(iii)}{\to} 0,
\end{aligned}$$

where (i) follows from (B.9), (ii) follows by a change of variable from $t\in A_1$ to $\theta=\hat{\theta}_1+t/(Km)^{1/2}$ in the first term of the summation, and (iii) follows from two facts: $\int_{\|\theta-\hat{\theta}_1\|_2\geq\delta_1}\pi(\theta)d\theta$ is bounded because $\pi(\theta)$ is density function and $\int_{\|t\|_2\geq\delta_1(Km)^{1/2}}\exp(-t^{\mathrm{T}}I_0t/2)dt$ converges to 0 because the integrand equals an un-normalized multivariate Gaussian density, which implies that the un-normalized tail probability converges to 0 as $m\to\infty$.

Step 3.2: Show that $\int_{A_2}|g_m(t)|dt\to 0$ *in* $\mathbb{P}_0$*-probability as* $m\to\infty$. We first show that $|g_m(t)|$ is bounded above by an integrable function with large $\mathbb{P}_0$-probability as $m\to\infty$. The desired result follows from an appropriate choice of δ_1 and δ_2. Consider the Taylor expansion of $L_{1m}(\theta)$ at $\hat{\theta}_1$ and set $\theta=\hat{\theta}_1+t/(Km)^{1/2}$ any $t\in A_2$. Because $L_{1m}(\hat{\theta}_1)=0$,

$$\begin{aligned}
w(t)&=L_{1m}\{\hat{\theta}_1+t/(Km)^{1/2}\}-L_{1m}(\hat{\theta}_1)=\frac{1}{2K}t^{\mathrm{T}}\frac{L''_{1m}(\hat{\theta}_1)}{m}t+R_m(t),\\
R_m(t)&=\frac{1}{6(Km)^{3/2}}t^{\mathrm{T}}\left\{L'''_{1m}(\tilde{\theta})t\right\}t,
\end{aligned}\tag{B.10}$$

for any $t\in A_2$, where $L'''_{1m}(\theta)$ is a $d\times d\times d$ array and $\tilde{\theta}$ satisfies $\|\tilde{\theta}-\hat{\theta}_1\|_2\leq\|t\|_2/(Km)^{1/2}\leq\delta_1$, where the last inequality follows because $\delta_2\leq\|t\|_2\leq\delta_1(Km)^{1/2}$. Furthermore, $\|\tilde{\theta}-\theta_0\|_2\leq\|\tilde{\theta}-\hat{\theta}_1\|_2+\|\hat{\theta}_1-\theta_0\|_2\leq\delta_1+\|\hat{\theta}_1-\theta_0\|_2$. If we choose $\delta_1\leq\delta_0/3$ and m to be large enough so that $\|\hat{\theta}_1-\theta_0\|_2\leq\delta_0/3$ with $\mathbb{P}_0$-probability approaching 1, then $\|\tilde{\theta}-\theta_0\|_2<\delta_0$ for every $t\in A_2$ with $\mathbb{P}_0$-probability approaching 1 as $m\to\infty$. For a given $t\in A_2$,

$$\begin{aligned}
|R_m(t)|&\leq\frac{\|t\|_2^3}{6K^{3/2}m^{1/2}}\sum_{l_3=1}^{d}\sum_{l_2=1}^{d}\sum_{l_1=1}^{d}\left|\frac{1}{m}\{L'''_{1m}(\tilde{\theta})\}_{l_1l_2l_3}\right|\\
&\overset{(i)}{\leq}\frac{\|t\|_2^3}{6K^{3/2}m^{1/2}}\sum_{l_3=1}^{d}\sum_{l_2=1}^{d}\sum_{l_1=1}^{d}\frac{1}{m}\sup_{\theta\in B_{\delta_0}(\theta_0)}|\{L'''_{1m}(\theta)\}_{l_1l_2l_3}|\\
&\overset{(ii)}{\leq}\frac{d^3\|t\|_2^3}{6K^{3/2}m^{1/2}}\frac{1}{m}M(Y_1,\ldots,Y_m)\overset{(iii)}{\leq}\frac{d^3\delta_1}{6K}\|t\|_2^2\frac{1}{m}M(Y_1,\ldots,Y_m)\\
&\overset{(iv)}{\to}\frac{d^3\delta_1}{6K}\|t\|_2^2\mathcal{C}_M\quad\mathbb{P}_0\text{-almost surely;}
\end{aligned}\tag{B.11}$$

where (i) follows because $\tilde{\theta}\in B_{\delta_0}(\theta_0)$ with a large $\mathbb{P}_0$-probability as $m\to\infty$, (ii) follows from Assumption (A_1), (iii) follows because $\|t\|_2\leq\delta_1(Km)^{1/2}$ and

(iv) also follows from Assumption (A_1). By Assumption (A_4), eigenvalues of $-\frac{1}{m}L''_{1m}(\theta)$ are bounded from blow and above by constants for all $\theta \in B_{\delta_0}(\theta_0)$. Hence we choose

$$\delta_1 = \min\left[\frac{\delta_0}{3}, \frac{6\min_{\theta\in B_{\delta_0}(\theta_0)}\lambda_1\{-\frac{1}{m}L''_{1m}(\theta)\}}{4d^3\mathcal{C}_M}\right], \tag{B.12}$$

where $\lambda_1(B)$ denotes the smallest eigenvalue of matrix B. With this choice of δ_1, (B.10) and (B.11) imply that for every $t \in A_2$ and with $\mathbb{P}_0$-probability approaching 1 as $m \to \infty$,

$$\begin{aligned}
|KR_m(t)| &\le \min_{\theta\in B_{\delta_0}(\theta_0)}\lambda_1\{-L''_{1m}(\theta)\}\, t^{\mathrm{T}}t/(4m) \le -t^{\mathrm{T}}L''_{1m}(\hat{\theta}_1)t/(4m),\\
Kw(t) &\le \frac{1}{2}t^{\mathrm{T}}\frac{L''_{1m}(\hat{\theta}_1)}{m}t + K|R_m(t)|\\
&\le \frac{1}{2}t^{\mathrm{T}}\frac{L''_{1m}(\hat{\theta}_1)}{m}t - \frac{1}{4}t^{\mathrm{T}}\frac{L''_{1m}(\hat{\theta}_1)}{m}t = \frac{1}{4}t^{\mathrm{T}}\frac{L''_{1m}(\hat{\theta}_1)}{m}t.
\end{aligned} \tag{B.13}$$

We use (B.13) to find an upper bound for $|g(t)|$ that is integrable on A_2. By assumption (A_4), $-m^{-1}L''_{1m}(\hat{\theta}_1) \to I_0$ in $\mathbb{P}_0$-probability as $m \to \infty$. This implies that $\exp(Kw(t)) \le \exp(-t^{\mathrm{T}}I_0t/4)$ with a large $\mathbb{P}_0$-probability as $m \to \infty$; therefore, with a large $\mathbb{P}_0$-probability as $m \to \infty$,

$$\begin{aligned}
&\int_{A_2}|g_m(t)|dt\\
&\overset{(i)}{\le} \int_{A_2}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{4}\right)\pi\left\{\hat{\theta}_1 + \frac{t}{(Km)^{1/2}}\right\}dt + \int_{A_2}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{2}\right)\pi(\theta_0)dt\\
&\le \int_{A_2}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{4}\right)\pi\left\{\hat{\theta}_1 + \frac{t}{(Km)^{1/2}}\right\}dt + \int_{A_2}\exp\left(-\frac{t^{\mathrm{T}}I_0t}{4}\right)\pi(\theta_0)dt\\
&\overset{(ii)}{\le} \left\{\sup_{\theta\in\Theta}\pi(\theta)\right\}\int_{A_2}2\exp\left(-\frac{t^{\mathrm{T}}I_0t}{4}\right)dt \overset{(iii)}{<} \infty,
\end{aligned} \tag{B.14}$$

where (i) follows from triangular inequality and (B.13), (ii) is because $\theta_0, \hat{\theta}_1 + t/(Km)^{1/2} \in \Theta$, Θ is compact from Assumption (A_5), and $\pi(\cdot)$ is continuous on Θ from Assumption (A_6), and (iii) follows because the continuity of $\pi(\cdot)$ and the compactness of Θ imply that $\pi(\theta)$ is bounded for every $\theta \in \Theta$ and the integrand equals an un-normalized Gaussian density, which has a finite integral over $\mathbb{R}^d$. We now can choose δ_2 sufficiently large such that $\int_{A_2}|g_m(t)|dt$ is arbitrarily small in $\mathbb{P}_0$-probability.

Step 3.3: Show that $\int_{A_3}|g_m(t)|dt \to 0$ *in* $\mathbb{P}_0$*-probability as* $m \to \infty$. Using (B.11) and Assumption (A_1), we have that for any $\|t\|_2 < \delta_2$ and a sufficiently large m,

$$\sup_{\|t\|_2<\delta_2}|KR_m(t)| \le \sup_{\|t\|_2<\delta_2}\frac{d^3\|t\|_2^3}{6K^{1/2}m^{1/2}}\frac{1}{m}M(Y_1,\ldots,Y_m)$$

$$\leq \frac{d^3\delta_2^3}{6K^{1/2}m^{1/2}}\frac{1}{m}M(Y_1,\ldots,Y_m) \to 0, \tag{B.15}$$

where the convergence is almost surely in $\mathbb{P}_0$-probability. The Taylor expansion in (B.10) yields that as $m \to \infty$,

$$\sup_{t\in A_3}\exp\left\{\left|Kw(t)+\frac{1}{2}t^\top I_0 t\right|\right\} = \sup_{\|t\|_2<\delta_2}\exp\left\{\left|\frac{1}{2}t^\top\frac{L''_{1m}(\hat{\theta}_1)}{m}t+KR_m(t)+\frac{1}{2}t^\top I_0 t\right|\right\}$$
$$\leq \exp\left\{\frac{\delta_2^2}{2}\left\|\frac{L''_{1m}(\hat{\theta}_1)}{m}+I_0\right\|_{\text{op}}\right\}\exp\left\{\sup_{\|t\|_2<\delta_2}|KR_m(t)|\right\}$$
$$\overset{(i)}{\to} 1, \text{ in } \mathbb{P}_0\text{-probability}, \tag{B.16}$$

where $\|\cdot\|_{\text{op}}$ is the operator norm and (i) follows from $L''_{1m}(\hat{\theta}_1)/m + I_0 \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ using Assumption (A_4) and weak consistency of $\hat{\theta}_1$, and $\sup_{\|t\|_2<\delta_2}|KR_m(t)| \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ using (B.15). The weak consistency of $\hat{\theta}_1$ and continuity of $\pi(\cdot)$ from Assumption (A_6) imply that for any fixed $\delta_2 > 0$

$$\sup_{\|t\|_2\leq\delta_2}\left|\pi\left\{\hat{\theta}_1+\frac{t}{(Km)^{1/2}}\right\}-\pi(\theta_0)\right| \to 0 \text{ as } m\to\infty \tag{B.17}$$

in $\mathbb{P}_0$-probability. Combining (B.16) and (B.17), we get that

$$\int_{A_3}|g_m(t)|dt \leq C\int_{A_3}\sup_{\|t\|_2\leq\delta_2}\left[\exp\left\{Kw(t)+\frac{1}{2}t^\top I_0 t\right\}-1\right]\times$$
$$\sup_{\|t\|_2\leq\delta_2}\left|\pi\left\{\hat{\theta}_1+\frac{t}{(Km)^{1/2}}\right\}-\pi(\theta_0)\right|dt, \tag{B.18}$$

where C is a universal constant; therefore, $\int_{A_3}|g_m(t)|dt \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$ because the first and second terms in the integrand converge to 0 uniformly from (B.16) and (B.17).

Step 4: Show that $\int_{\mathbb{R}^d}|g_m(t)|dt \to 0$ *in* $\mathbb{P}_0$*-probability as* $m \to \infty$. Using steps 3.1 to 3.3, choose δ_1 as defined in (B.12) and δ_2 large enough such that $\int_{A_2}|g_m(t)|dt \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$. This implies using the triangle inequality that

$$\int_{\mathbb{R}^d}|g_m(t)|dt \leq \int_{A_1}|g_m(t)|dt+\int_{A_2}|g_m(t)|dt+\int_{A_3}|g_m(t)|dt \to 0$$

in $\mathbb{P}_0$-probability as $m \to \infty$. The proof is complete. □

B.2. Proof of Theorem 3.1(ii)

We first prove Theorem 3.1(ii) for the second subset posterior distribution and then prove for jth subset posterior distribution defined in (2.5) in the main manuscript $(j = 3,\ldots,K)$.

Step 1: Define two log-likelihood functions. We define the two log-likelihood functions of θ that correspond to two different marginal distributions of X_{m+1}. Let Y_1^m and Y_{m+1}^{2m} denote the first and second data subsets, respectively, $w_1(\theta)$ be the log-likelihood of θ given Y_{m+1}^{2m}, and $w_2(\theta)$ be the conditional log-likelihood of θ obtained from the conditional density of Y_{m+1}^{2m} given Y_1^m. Then,

$$\begin{aligned}&\exp\{w_1(\theta)\}\\&\quad= p_\theta(Y_{m+1}^{2m}) = \sum_{X_{m+1}^{2m}\in\mathcal{X}^m} r_\theta(X_{m+1}) \prod_{t=m+1}^{2m-1} q_\theta(X_t, X_{t+1}) \prod_{t=m+1}^{2m} g_\theta(Y_t \mid X_t),\end{aligned} \tag{B.19}$$

$$\begin{aligned}&\exp\{w_2(\theta)\}\\&\quad= p_\theta(Y_{m+1}^{2m} \mid Y_1^m)\\&\quad= \sum_{X_{m+1}^{2m}\in\mathcal{X}^m} p_\theta(X_{m+1} \mid Y_1^m) \prod_{t=m+1}^{2m-1} q_\theta(X_t, X_{t+1}) \prod_{t=m+1}^{2m} g_\theta(Y_t \mid X_t).\end{aligned} \tag{B.20}$$

If the marginal distribution of X_{m+1} is $r_\theta(\cdot)$, then $\exp\{w_1(\theta)\}$ is obtained by marginalizing over X_{m+1}^{2m} in the joint distribution of $(Y_{m+1}^{2m}, X_{m+1}^{2m})$. Similarly, we recover $\exp\{w_2(\theta)\}$ after marginalizing over X_{m+1}^{2m} from the conditional distribution of $(Y_{m+1}^{2m}, X_{m+1}^{2m})$ given Y_1^m. This is also equivalent to the marginal density of Y_{m+1}^{2m} obtained from the joint distribution of $(Y_{m+1}^{2m}, X_{m+1}^{2m})$, where the marginal density of X_{m+1} is set to be $p_\theta(X_{m+1} \in \cdot \mid Y_1^m)$. We re-express $\exp\{w_1(\theta)\}$ and $\exp\{w_2(\theta)\}$ in terms of the prediction filter in (A.3) using Lemma A.4 as

$$\begin{aligned}\exp\{w_1(\theta)\} &= \prod_{t=m+1}^{2m} \sum_{a=1}^{S} g_\theta(Y_t \mid X_t = a)\, p_t^\theta(a),\\ \exp\{w_2(\theta)\} &= \prod_{t=m+1}^{2m} \sum_{a=1}^{S} g_\theta(Y_t \mid X_t = a)\, \tilde{p}_t^\theta(a),\end{aligned} \tag{B.21}$$

where

$$p_{m+1}^\theta(X_{m+1}) = r_\theta(X_{m+1}), \quad \tilde{p}_{m+1}^\theta(X_{m+1}) = p_\theta(X_{m+1} \mid Y_1^m),$$

and for $t = m+2, \ldots, 2m$

$$p_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, r_\theta(X_{m+1})\}, \quad \tilde{p}_t^\theta(X_t) = f_{t-m-2}^\theta\{Y_{m+1}^{t-1}, p_\theta(X_{m+1} \mid Y_1^m)\},$$

where f_{t-m-2}^θ is defined in (A.3).

Step 2: Show that $\sup_{\theta\in\Theta} |w_1(\theta) - w_2(\theta)| \le \tilde{C}\rho^{m-2}$ $\mathbb{P}_0$*-almost surely for some constant* $\tilde{C}$.

For any $\theta \in \Theta$ we have that

$$|w_1(\theta) - w_2(\theta)|$$

$$
\begin{aligned}
&= \left| \sum_{t=m+1}^{2m} \left[\log \left\{ \sum_{a=1}^{S} p_t^\theta(a) g_\theta(Y_t \mid X_t = a) \right\} - \log \left\{ \sum_{a=1}^{S} \tilde{p}_t^\theta(a) g_\theta(Y_t \mid X_t = a) \right\} \right] \right| \\
&\overset{(i)}{\leq} \sum_{t=m+1}^{2m} \left| \log \left\{ \sum_{a=1}^{S} p_t^\theta(a) g_\theta(Y_t \mid X_t = a) \right\} - \log \left\{ \sum_{a=1}^{S} \tilde{p}_t^\theta(a) g_\theta(Y_t \mid X_t = a) \right\} \right| \\
&\overset{(ii)}{\leq} \sum_{t=m+1}^{2m} \{h_\theta(Y_t) - 1\} \, \| p_t^\theta - \tilde{p}_t^\theta \|_{\mathrm{TV}} \\
&= \{h_\theta(Y_{m+1}) - 1\} \, \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\quad + \{h_\theta(Y_{m+2}) - 1\} \, \| p_{m+2}^\theta - \tilde{p}_{m+2}^\theta \|_{\mathrm{TV}} \\
&\quad + \sum_{t=m+3}^{2m} \{h_\theta(Y_t) - 1\} \, \| p_t^\theta - \tilde{p}_t^\theta \|_{\mathrm{TV}} \\
&\overset{(iii)}{\leq} \{h_\theta(Y_{m+1}) - 1\} \, \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\quad + \{h_\theta(Y_{m+2}) - 1\} \, \epsilon_\theta^{-1} h_\theta(Y_{m+2}) \| r_\theta(X_{m+1}) p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\quad + \sum_{t=m+3}^{2m} \{h_\theta(Y_t) - 1\} \, \epsilon_\theta^{-1} h_\theta(Y_t) \prod_{i=m+2}^{t-1} \{1 - \epsilon_\theta h_\theta(Y_i)^{-1}\} \| \\
&\quad \times r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\overset{(iv)}{\leq} \{h_\theta(Y_{m+1}) - 1\} \, \epsilon_\theta^{-1} h_\theta(Y_{m+1}) \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\quad + \{h_\theta(Y_{m+2}) - 1\} \, \epsilon_\theta^{-1} h_\theta(Y_{m+2}) \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \\
&\quad + \sum_{t=m+3}^{2m} \{h_\theta(Y_t) - 1\} \, \epsilon_\theta^{-1} h_\theta(Y_t) \prod_{i=m+2}^{t-1} \{1 - \epsilon h_\theta(Y_i)^{-1}\} \| \\
&\quad \times r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}}, \qquad \text{(B.22)}
\end{aligned}
$$

where (i) follows from the triangle inequality, (ii) follows from Lemma A.3, (iii) follows from Lemma A.2, and (iv) follows from from Lemma A.2 because $\epsilon^{-1} h_\theta(Y_{m+1}) > 1$. Assumption (A_8) implies that $\sup_{\theta \in \Theta} h_\theta(Y_t) \leq M$ for $1 \leq t \leq n$ $\mathbb{P}_0$-almost surely. The right side of (B.22) is a decreasing function of ϵ_θ and $\epsilon = \inf_{\theta \in \Theta} \epsilon_\theta > 0$. Hence, with $\mathbb{P}_0$-probability almost 1, by taking $\sup_{\theta \in \Theta}$ over the right hand side of (B.22) we have that

$$
\begin{aligned}
\sup_{\theta \in \Theta} |w_1(\theta) - w_2(\theta)| &\leq \frac{1}{\epsilon} \sup_{\theta \in \Theta} \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \times \\
&\quad \sup_{\theta \in \Theta} \left[\{h_\theta(Y_{m+1}) - 1\} \, h_\theta(Y_{m+1}) + \{h_\theta(Y_{m+2}) - 1\} \, h_\theta(Y_{m+2}) \right. \\
&\quad \left. + \sum_{t=m+3}^{2m} \{h_\theta(Y_t) - 1\} \, \epsilon^{-1} h_\theta(Y_t) \prod_{i=m+2}^{t-1} \{1 - \epsilon h_\theta(Y_i)^{-1}\} \right] \\
&\leq \frac{1}{\epsilon} \sup_{\theta \in \Theta} \| r_\theta(X_{m+1}) - p_\theta(X_{m+1} \mid Y_1^m) \|_{\mathrm{TV}} \times
\end{aligned}
$$

$$\left\{2M(M-1)+\sum_{t=m+3}^{2m} M(M-1)\prod_{i=m+2}^{t-1}(1-\epsilon M^{-1})\right\}$$

$$\le \frac{1}{\epsilon}M(M-1)\sup_{\theta\in\Theta}\|r_\theta(X_{m+1})-p_\theta(X_{m+1}\mid Y_1^m)\|_{\mathrm{TV}}\left\{2+\sum_{t=1}^{\infty}(1-\epsilon M^{-1})^t\right\}$$

$$\le \frac{1}{\epsilon^2}M(M^2-1)\sup_{\theta\in\Theta}\|r_\theta(X_{m+1})-p_\theta(X_{m+1}\mid Y_1^m)\|_{\mathrm{TV}}$$

$$\overset{(i)}{\le} \frac{1}{\epsilon^2}M(M^2-1)S\rho^{m-2}=\tilde{C}\rho^{m-2}, \tag{B.23}$$

where (i) follows from assumption (A_8)(iii): $\sup_{\theta\in\Theta}\|r_\theta(X_{m+1})-p_\theta(X_{m+1}\mid Y_1^m)\|_{\mathrm{TV}}<S\rho^{m-2}$ $\mathbb{P}_0$-almost surely.

Step 3: Show that the total variation distance of subset quasi posterior distributions induced by likelihood $\exp\{w_1(\theta)\}$ *and* $\exp\{w_2(\theta)\}$ *tends to 0 in* $\mathbb{P}_0$*-probability as m tends to infinity.*

The definition of $w_1(\theta)$ implies that $w_1(\theta)=L_{2m}(\theta)$, where $L_{2m}(\theta)$ is the log-likelihood of θ on the second subset Y_{m+1}^{2m}. Let $\hat{\theta}_2$ denote the maximum likelihood estimator of θ that solves $L'_{2m}(\theta)=0$ or $w'_1(\theta)=0$. For large m, $\hat{\theta}_2\in B_{\delta_0}(\theta_0)$, hence the limiting set of local parameters $\lim_{m\to\infty}\{t=(Km)^{1/2}(\theta-\hat{\theta}_2):\theta\in\Theta\}=\mathbb{R}^d$. Then the densities of the second subset posterior distributions of $t=(Km)^{1/2}(\theta-\hat{\theta}_2)$ with likelihoods $\exp\{w_1(\hat{\theta}_2+t/(Km)^{1/2})\}$ and $\exp\{w_2(\hat{\theta}_2+t/(Km)^{1/2})\}$ are

$$\tilde{\pi}_{w_1}(t\mid Y_{m+1}^{2m})=C_{1m}^{-1}e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\},$$

$$\tilde{\pi}_{w_2}(t\mid Y_{m+1}^{2m})=C_{2m}^{-1}e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\},$$

$$C_{1m}=\int_{\mathbb{R}^d}e^{K\left[w_1\left\{\hat{\theta}_2+\frac{z}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2+\tfrac{z}{(Km)^{1/2}}\right\}dz,$$

$$C_{2m}=\int_{\mathbb{R}^d}e^{K\left[w_2\left\{\hat{\theta}_2+\frac{z}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2+\tfrac{z}{(Km)^{1/2}}\right\}dz$$

The total variation distance between subset posterior distributions of θ with the likelihoods $\exp\{w_1(\theta)\}$ and $\exp\{w_2(\theta)\}$ equals $\int_{\mathbb{R}^d}|\tilde{\pi}_{w_1}(t\mid Y_{m+1}^{2m})-\tilde{\pi}_{w_2}(t\mid Y_{m+1}^{2m})|dt$; therefore, we want to prove that

$$\int_{\mathbb{R}^d}\left|C_{1m}^{-1}e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}-C_{2m}^{-1}e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\right|\times$$
$$\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\}dt\to 0 \tag{B.24}$$

in $\mathbb{P}_0$-probability as $m\to\infty$. An application of triangular inequality implies that an upper bound for the integral in (B.24) is

$$C_{1m}^{-1}\int_{\mathbb{R}^d}\left|e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}-e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\right|$$

$$\times \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt$$
$$+ \left|C_{1m}^{-1}C_{2m} - 1\right| \int_{\mathbb{R}^d} C_{2m}^{-1}\, e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]} \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt, \tag{B.25}$$

and we will show that the upper bound in (B.25) tends to 0 in $\mathbb{P}_0$-probability as $m \to \infty$.

First, we show that C_{1m} is bounded away from 0 with a large $\mathbb{P}_0$-probability as $m \to \infty$. Because $w_1(\theta) = L_{2m}(\theta)$, we have that $w_1\left\{\hat{\theta}_2 + \frac{t}{(Km)^{1/2}}\right\} = L_{2m}\left\{\hat{\theta}_2 + \frac{t}{(Km)^{1/2}}\right\}$ and that

$$C_{1m} \to (2\pi)^{d/2} \det(I_0)^{-1/2} \pi(\theta_0) \tag{B.26}$$

in $\mathbb{P}_0$-probability as $m \to \infty$ based on the same arguments used to prove (B.5); therefore, C_{1m}^{-1} is finite with a large $\mathbb{P}_0$-probability as $m \to \infty$.

Second, we find a sufficient condition which guarantees that $|C_{1m}^{-1}C_{2m} - 1| \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$. We have that

$$\begin{aligned}
&|C_{2m} - C_{1m}| \\
&\quad = \left| \int_{\mathbb{R}^d} e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]} - e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]} \right. \\
&\qquad \left. \times \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt \right|, \\
&\quad \leq \int_{\mathbb{R}^d} \left| e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]} - e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]} \right| \\
&\qquad \times \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt
\end{aligned} \tag{B.27}$$

where the last term is the first integral in (B.25); therefore, it is sufficient to show that the first integral in (B.25) tends to 0 in $\mathbb{P}_0$-probability as $m \to \infty$ because C_{1m} is bounded away from zero with a large $\mathbb{P}_0$-probability as $m \to \infty$.

Thirdly, we find a sufficient condition which gaurantees that the second term of the upper bound in (B.25) tends to 0 in $\mathbb{P}_0$-probability as $m \to \infty$. If the first integral in (B.25) tends to 0 in $\mathbb{P}_0$-probability as $m \to \infty$, then $|C_{1m}^{-1}C_{2m} - 1| \to 0$ and $|C_{2m} - C_{1m}| \to 0$ in $\mathbb{P}_0$-probability as $m \to \infty$, and by triangle inequality we have that

$$\begin{aligned}
&\int_{\mathbb{R}^d} e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]} \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt \\
&\leq C_{1m} + \int_{\mathbb{R}^d} \left| e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]} - e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]} \right| \times
\end{aligned}$$

$$\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\}dt\to(2\pi)^{d/2}\det(I_0)^{-1/2}\pi(\theta_0) \tag{B.28}$$

in $\mathbb{P}_0$-probability as $m\to\infty$, which implies that the second integral in (B.25) is bounded in $\mathbb{P}_0$-probability as $m\to\infty$. By Slutsky's theorem we have that

$$\left|C_{1m}^{-1}C_{2m}-1\right|\int_{\mathbb{R}^d}C_{2m}^{-1}e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\}dt\to 0 \tag{B.29}$$

in $\mathbb{P}_0$-probability as $m\to\infty$; therefore, it is sufficient to show that in $\mathbb{P}_0$-probability as $m\to\infty$,

$$\int_{\mathbb{R}^d}\left|e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}-e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\right| \times\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\}dt\to 0. \tag{B.30}$$

Step 4: Show that the integral in (B.30) tends to 0 in $\mathbb{P}_0$-probability as $m\to\infty$.

For any $t\in\mathbb{R}^d$, let $\theta^*=\hat{\theta}_2+\frac{t}{(Km)^{1/2}}$. Using (B.23) in step 2, as $m\to\infty$, with $\mathbb{P}_0$ probability tending to 1,

$$\begin{aligned}&|Kw_1(\theta^*)-Kw_2(\theta^*)|\le K\sup_{\theta\in\Theta}|w_1(\theta)-w_2(\theta)|\\&\le\tilde{C}\rho^{-2}K\rho^m.\end{aligned} \tag{B.31}$$

If $K=o(\rho^{-m})$, as $m\to\infty$, $|Kw_1(\theta^*)-Kw_2(\theta^*)|\to 0$ with $\mathbb{P}_0$-probability tending to 1. Let $t=0$, we have that $\left|Kw_1(\hat{\theta}_2)-Kw_2(\hat{\theta}_2)\right|\to 0$ with $\mathbb{P}_0$-probability tending to 1. By triangle inequality, as $m\to\infty$, $\left|K[w_1(\theta^*)-w_1(\hat{\theta}_2)]-K[w_2(\theta^*)-w_2(\hat{\theta}_2)]\right|\to 0$ with $\mathbb{P}_0$-probability tending to 1. Then by continuous mapping theorem, we have that

$$\left|1-\exp\left(K[w_2(\theta^*)-w_2(\hat{\theta}_2)]-K[w_1(\theta^*)-w_1(\hat{\theta}_2)]\right)\right|\to 0 \tag{B.32}$$

with $\mathbb{P}_0$-probability tending to 1 as $m\to\infty$. The integral in (B.30) can be expressed as

$$\begin{aligned}&\int_{\mathbb{R}^d}\left|e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}-e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\right|\\&\quad\times\pi\left\{\hat{\theta}_2+\tfrac{t}{(Km)^{1/2}}\right\}dt\\&=\int_{\mathbb{R}^d}e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\\&\quad\times\left|1-\exp\left(K[w_2(\theta^*)-w_2(\hat{\theta}_2)]-K[w_1(\theta^*)-w_1(\hat{\theta}_2)]\right)\right|\end{aligned}$$

$$\times \qquad \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt, \tag{B.33}$$

and we know that $\int_{\mathbb{R}^d} e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2 + \frac{t}{(Km)^{1/2}}\right\} dt = C_{1m}$, which is bounded in $\mathbb{P}_0$-probability by (B.26). An application of the dominated convergence theorem implies that integral in (B.33) tends to 0 in $\mathbb{P}_0$-probability as $m \to \infty$.

Step 5: Show that Theorem 3.1(ii) holds for $j = 2$.

The proof of Theorem 3.1(i) in Section B.1 implies that as $m \to \infty$,

$$\int_{\mathbb{R}^d} \left| C_{1m}^{-1} e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]} \times \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2} e^{-\frac{1}{2}t^{\mathrm{T}} I_0 t}\right| dt \to 0 \tag{B.34}$$

in $\mathbb{P}_0$-probability. For $j = 2$, (3.3) reduces to

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left| C_{2m}^{-1} e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2} e^{-\frac{1}{2}t^{\mathrm{T}} I_0 t}\right| dt \\
&\le \int_{\mathbb{R}^d} \left| C_{2m}^{-1} e^{K\left[w_2\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_2\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} - C_{1m}^{-1} e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\right| \\
&\qquad \times \pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} dt \\
&\quad + \int_{\mathbb{R}^d} \left| C_{1m}^{-1} e^{K\left[w_1\left\{\hat{\theta}_2+\frac{t}{(Km)^{1/2}}\right\}-w_1\left\{\hat{\theta}_2\right\}\right]}\pi\left\{\hat{\theta}_2 + \tfrac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2} e^{-\frac{1}{2}t^{\mathrm{T}} I_0 t}\right| dt.
\end{aligned} \tag{B.35}$$

The first and second integrals on the right hand side of (B.35) tend to 0 in $\mathbb{P}_0$-probability as $m \to \infty$ using steps 3, 4 and (B.34), respectively. This proves (3.3) for $j = 2$

We conclude the proof by showing that Theorem 3.1(ii) holds for $j = 3, \ldots, K$.

Step 6: Prove Theorem 3.1 (ii) for jth subset quasi posterior distributions ($j = 3, \ldots, K$).

For any $\theta \in \Theta$ and $j = 3, \ldots, K$, use (B.19), (B.20) to define the jth subset likelihood functions that are equivalent to $\exp\{w_1(\theta)\}$ and $\exp\{w_2(\theta)\}$, respec-

tively, as

$$\exp\{w_1^j(\theta)\} = p_\theta(Y_{(j-1)m+1}^{jm}) = \prod_{t=(j-1)m+1}^{jm} \sum_{a=1}^{S} g_\theta(Y_t \mid X_t = a) p_t^\theta(a),$$

$$\exp\{w_2^j(\theta)\} = p_\theta(Y_{(j-1)m+1}^{jm} \mid Y_{(j-2)m+1}^{(j-1)m}) = \prod_{t=(j-1)m+1}^{jm} \sum_{a=1}^{S} g_\theta(Y_t \mid X_t = a) \tilde{p}_t^\theta(a),$$

where

$$p_{(j-1)m+1}^\theta(X_{(j-1)m+1}) = r_\theta(X_{(j-1)m+1}),$$
$$\tilde{p}_{(j-1)m+1}^\theta(X_{(j-1)m+1}) = p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m})$$

and for $t = (j-1)m+2, \ldots, jm$,

$$p_t^\theta(X_t) = f_{t-m-2}^\theta \left\{ Y_{(j-1)m+1}^{t-1}, r_\theta(X_{(j-1)m+1}) \right\},$$
$$\tilde{p}_t^\theta(X_t) = f_{t-m-2}^\theta \left\{ Y_{(j-1)m+1}^{t-1}, p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m}) \right\},$$

where f_{t-m-2}^θ is defined in (A.3). By Assumption (A_8), we have that with $\mathbb{P}_0$-probability almost 1,

$$\sup_{\theta \in B_{\delta_0}(\theta_0)} \|r_\theta(X_{(j-1)m+1}) - p_\theta(X_{(j-1)m+1} | Y_{(j-2)m+1}^{(j-1)m})\|_{\mathrm{TV}} < \rho^{m-2}.$$

For any $t \in \mathbb{R}^d$, let $\theta^* = \hat{\theta}_j + \frac{t}{(Km)^{1/2}}$. Using a similar argument of (B.23) in step 2, we have that

$$\begin{aligned}
&\left| K w_1^j(\theta^*) - K w_2^j(\theta^*) \right| \\
&\le \frac{1}{\epsilon^2} M(M^2-1) K \|r_{\theta^*}(X_{(j-1)m+1}) - p_{\theta^*}(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m})\|_{\mathrm{TV}} \\
&\le \frac{1}{\epsilon^2} M(M^2-1) K \sup_{\theta \in \Theta} \|r_\theta(X_{(j-1)m+1}) - p_\theta(X_{(j-1)m+1} \mid Y_{(j-2)m+1}^{(j-1)m})\|_{\mathrm{TV}} \\
&\le \epsilon^{-2} C \rho^{-2} M(M^2-1) K \rho^m. \qquad \text{(B.36)}
\end{aligned}$$

As $m \to \infty$, if $K = o(\rho^{-m})$, $\left| K w_1^j(\theta^*) - K w_2^j(\theta^*) \right| \to 0$ with $\mathbb{P}_0$-probability tending to 1. For $t \in \lim_{m\to\infty} \{t = (Km)^{1/2}(\theta - \hat{\theta}_j) : \theta \in \Theta\} = \mathbb{R}^d$, we define

$$\tilde{\pi}_{w_1}^j(t \mid Y_{m+1}^{2m}) = C_{1m}^{-1} e^{K\left[w_1^j\left\{\hat{\theta}_j + \frac{t}{(Km)^{1/2}}\right\} - w_1^j\left\{\hat{\theta}_j\right\}\right]} \pi\left\{\hat{\theta}_j + \tfrac{t}{(Km)^{1/2}}\right\},$$

$$\tilde{\pi}_{w_2^j}(t \mid Y_{m+1}^{2m}) = C_{2m}^{-1} e^{K\left[w_2^j\left\{\hat{\theta}_j + \frac{t}{(Km)^{1/2}}\right\} - w_2^j\left\{\hat{\theta}_j\right\}\right]} \pi\left\{\hat{\theta}_j + \tfrac{t}{(Km)^{1/2}}\right\},$$

$$C_{1m}^j = \int_{\mathbb{R}^d} e^{K\left[w_1^j\left\{\hat{\theta}_2 + \frac{z}{(Km)^{1/2}}\right\} - w_1^j\left\{\hat{\theta}_j\right\}\right]} \pi\left\{\hat{\theta}_j + \tfrac{z}{(Km)^{1/2}}\right\} dz, \quad \text{(B.37)}$$

$$C_{2m}^{j} = \int_{\mathbb{R}^d} e^{K\left[w_2^j\left\{\hat{\theta}_j+\frac{z}{(Km)^{1/2}}\right\}-w_2^j\left\{\hat{\theta}_j\right\}\right]}\pi\left\{\hat{\theta}_j+\frac{z}{(Km)^{1/2}}\right\}dz \quad \text{(B.38)}$$

Following a similar argument of step 3 and step 4, we can prove that as $m \to \infty$, with large $\mathbb{P}_0$-probability

$$C_{1m}^{j} \to (2\pi)^{d/2}\det(I_0)^{-1/2}\pi(\theta_0),$$
$$C_{2m}^{j} \to (2\pi)^{d/2}\det(I_0)^{-1/2}\pi(\theta_0),$$
$$\int_{\mathbb{R}^d}\left|(C_{1m}^{j})^{-1}\, e^{K\left[w_1^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_1^j\left\{\hat{\theta}_j\right\}\right]} - (C_{2m}^{j})^{-1}\, e^{K\left[w_2^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_2^j\left\{\hat{\theta}_j\right\}\right]}\right| \times \pi\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}dt \to 0. \quad \text{(B.39)}$$

By the proof of Theorem 3.1(i) in Section B.1 implies that as $m \to \infty$,

$$\int_{\mathbb{R}^d}\left|(C_{1m}^{j})^{-1}\, e^{K\left[w_1^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_1^j\left\{\hat{\theta}_j\right\}\right]}\pi\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2}e^{-\frac{1}{2}t^{\mathrm{T}}I_0 t}\right|dt \to 0 \quad \text{(B.40)}$$

in $\mathbb{P}_0$-probability.

Theorem 3.1 (ii) at j can be expressed as

$$\begin{aligned}
&\int_{\mathbb{R}^d}\left|(C_{2m}^{j})^{-1}\, e^{K\left[w_2^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_2^j\left\{\hat{\theta}_j\right\}\right]}\pi\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2}e^{-\frac{1}{2}t^{\mathrm{T}}I_0 t}\right|dt \\
&\le \int_{\mathbb{R}^d}\left|(C_{1m}^{j})^{-1}\, e^{K\left[w_1^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_1^j\left\{\hat{\theta}_j\right\}\right]} - (C_{2m}^{j})^{-1}\, e^{K\left[w_2^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_2^j\left\{\hat{\theta}_j\right\}\right]}\right| \times \pi\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}dt \\
&+ \int_{\mathbb{R}^d}\left|(C_{1m}^{j})^{-1}\, e^{K\left[w_1^j\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\}-w_1^j\left\{\hat{\theta}_j\right\}\right]}\pi\left\{\hat{\theta}_j+\frac{t}{(Km)^{1/2}}\right\} - (2\pi)^{-d/2}\det(I_0)^{1/2}e^{-\frac{1}{2}t^{\mathrm{T}}I_0 t}\right|dt. \quad \text{(B.41)}
\end{aligned}$$

The first and second integrals on the right hand side of (B.41) tend to 0 in $\mathbb{P}_0$-probability as $m \to \infty$ using (B.39) and (B.40), respectively. This proves Theorem 3.1 (ii) for jth subset quasi posterior distribution, which completes the proof. □

Appendix C: Proof of Theorem 3.2

C.1. Proof of Theorem 3.2(i)

Step 1: Find an upper bound for the total variation distance between $\tilde{\Pi}_n(\cdot \mid Y_1^n)$ *and d-variate normal distribution with mean zero and* I_0^{-1} *as the covariance matrix.*

For any $K \in \mathbb{N}$ and $j = 1, \ldots, K$, let $h_j = \sqrt{n}(\theta - \hat{\theta}_j)$ denote the jth local parameter, $\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm})$ denote the distribution of h_j implied by the jth subset posterior distribution $\tilde{\Pi}_m(\cdot \mid Y_{(j-1)m+1}^{jm})$ with density $\tilde{\pi}_m(\theta \mid Y_{(j-1)m+1}^{jm})$. Theorem 3.1 in the main manuscript implies that

$$\|\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} \to 0 \text{ in } \mathbb{P}_0\text{-probability as } m \to \infty, \quad j = 1, \ldots, K. \tag{C.1}$$

By combination scheme, the density of block filtered posterior distribution $\tilde{\Pi}_n(\theta \mid Y_1^n)$ is defined as

$$\tilde{\pi}_n(\theta \mid Y_1^n) = \frac{1}{K} \sum_{j=1}^{K} \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \tilde{\pi}_m \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} (\theta - \tilde{\theta}) + \hat{\theta}_j \mid Y_{(j-1)m+1}^{jm} \right\}, \quad \theta \in \Theta. \tag{C.2}$$

Let $\tilde{\Pi}_n^h(\cdot \mid Y_1^n)$ denote the distribution of $h = \sqrt{n}(\theta - \tilde{\theta})$ implied by block filtered posterior distribution $\tilde{\Pi}_n(\cdot \mid Y_1^n)$. Then the density of $\tilde{\Pi}_n^h(\cdot \mid Y_1^n)$ is given by

$$\begin{aligned} \tilde{\pi}_n^h(h \mid Y_1^n) &= \frac{1}{K} \sum_{j=1}^{K} \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \tilde{\pi}_m \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} \frac{h}{\sqrt{n}} + \hat{\theta}_j \mid Y_{(j-1)m+1}^{jm} \right\} n^{-d/2} \\ &= \frac{1}{K} \sum_{j=1}^{K} \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \pi_m^{h_j} \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} h \mid Y_{(j-1)m+1}^{jm} \right\}, \quad h \in \mathbb{R}^d, \end{aligned} \tag{C.3}$$

where $\pi_m^{h_j}(h \mid Y_{(j-1)m}^{jm})$ is the density of $\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m}^{jm})$.

Let $\phi(\theta; \mu_\theta, \Sigma)$ denote the density function of d-variate normal distribution with mean μ_θ and Σ as covariance matrix. The total variation distance between $\tilde{\Pi}_n^h(\cdot \mid Y_1^n)$ and $N_d(0, I_0^{-1})$ is upper bounded by

$$\begin{aligned} &\|\tilde{\Pi}_n^h(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} \\ &\overset{(i)}{=} \frac{1}{2} \int_{\mathbb{R}^d} \left| \frac{1}{K} \sum_{j=1}^{K} \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \pi_m^{h_j} \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} h \mid Y_{(j-1)m+1}^{jm} \right\} - \phi(h; 0, I_0^{-1}) \right| dh \\ &\overset{(ii)}{\leq} \frac{1}{2K} \sum_{j=1}^{K} \int_{\mathbb{R}^d} \left| \frac{\det(\Sigma_j)^{1/2}}{\det(\tilde{\Sigma})^{1/2}} \pi_m^{h_j} \left\{ \Sigma_j^{1/2} \tilde{\Sigma}^{-1/2} h \mid Y_{(j-1)m+1}^{jm} \right\} - \phi(h; 0, I_0^{-1}) \right| dh \end{aligned}$$

$$\overset{(iii)}{=} \frac{1}{2K}\sum_{j=1}^{K}\int_{\mathbb{R}^d}|\pi_m^{h_j}\left\{t \mid Y_{(j-1)m+1}^{jm}\right\} - \frac{\det(\tilde{\Sigma})^{1/2}}{\det(\Sigma_j)^{1/2}}\phi(\tilde{\Sigma}^{1/2}\Sigma_j^{-1/2}t;0,I_0^{-1})|dt$$

$$= \frac{1}{2K}\sum_{j=1}^{K}\int_{\mathbb{R}^d}|\pi_m^{h_j}\left\{t \mid Y_{(j-1)m+1}^{jm}\right\} - \phi(t;0,\Sigma_j^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2})|dt$$

$$= \frac{1}{K}\sum_{j=1}^{K}\left\{\|\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d(0,\Sigma_j^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2})\|_{\mathrm{TV}}\right\}$$

$$\overset{(iv)}{\leq} \frac{1}{K}\sum_{j=1}^{K}\left\{\|\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d(0,I_0^{-1})\|_{\mathrm{TV}}\right.$$
$$\left.+\|N_d(0,I_0^{-1}) - N_d\left(0,\Sigma_j^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2}\right))\|_{\mathrm{TV}}\right\}, \tag{C.4}$$

where (i) follows from the definition of total variation distance and (C.3), (ii) follows from triangle inequality, (iii) follows from the transformation $t = \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2}h$, and (iv) again follows from triangle inequality. The first term inside the summation of (C.4) converges to 0 in $\mathbb{P}_0$ probability by (C.1). The second term inside the summation of (C.4) is the total variation distance between two normal distribution with zero mean and different variance.

Step 2: Find an upper bound for the second term inside the summation of (C.4).

Let $N_d(\mu_1, V_1)$ and $N_d(\mu_2, V_2)$ be two d-variate normal distributions with mean μ_1, μ_2 and covariance matrix V_1, V_2, respectively. Let $v = \mu_1 - \mu_2$. By Lemma A.5, the total variation distance $\|N_d(\mu_1, V_1) - N_d(\mu_1, V_1)\|_{\mathrm{TV}}$ is upper bounded by:

$$\|N_d(\mu_1,V_1) - N_d(\mu_2,V_2)\|_{\mathrm{TV}}$$
$$\leq \frac{9}{2}\max\{\frac{|v^{\mathrm{T}}(V_1-V_2)v|}{v^{\mathrm{T}}V_1v}, \frac{v^{\mathrm{T}}v}{\sqrt{v^{\mathrm{T}}V_1v}}, \|(N^{\mathrm{T}}V_1N)^{-1}N^{\mathrm{T}}V_2N - I_{d-1}\|_F\}, \tag{C.5}$$

where N is a $d \times (d-1)$ matrix whose columns form a basis for the subspace orthogonal to v.

Let $A = (a_{ik})$ and $B = (b_{kj})$ be two matrix such that AB exists. In general, we have an upper bound for $\|AB\|_F$

$$\|AB\|_F = \left(\sum_i\sum_j|\sum_k a_{ik}b_{kj}|^2\right)^{1/2} \overset{(i)}{\leq} \left(\sum_i\sum_j\sum_k a_{ik}^2\sum_k b_{kj}^2\right)^{1/2}$$
$$= \left(\sum_{i,k}a_{ik}^2\sum_{j,k}b_{kj}^2\right)^{1/2} = \|A\|_F\|B\|_F, \tag{C.6}$$

where (i) follows from Cauchy-Schwarz inequality.

Denote $F_j = \Sigma_j^{1/2} - (nI_0)^{-1/2}$ and $\tilde{F} = \tilde{\Sigma}^{-1/2} - (nI_0)^{1/2}$. By assumption (A_9), we have

$$\begin{aligned}
\mathbb{E}_0\|\Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} - I_d\|_F^2 &= \mathbb{E}_0\,\|\{(nI_0)^{-1/2} + F_j\}\{(nI_0)^{1/2} + \tilde{F}\} - I_d\|_F^2 \\
&= \mathbb{E}_0\,\|n^{-1/2}I_0^{-1/2}\tilde{F} + n^{1/2}F_jI_0^{1/2} + F_j\tilde{F}\|_F^2 \\
&\overset{(i)}{\leq} 3n^{-1}\|I_0^{-1/2}\|_F^2\|\tilde{F}\|_F^2 + 3n\|I_0^{1/2}\|_F^2\,\mathbb{E}_0\,\|F_j\|_F^2 + 3\,\mathbb{E}_0\,n\|F_j\|_F^2 n^{-1}\|\tilde{F}\|_F^2 \\
&\overset{(ii)}{\to} 0, \qquad \text{(C.7)}
\end{aligned}$$

where (i) follows from (C.7), and (ii) follows (A_9) that $n^{-1}\|\tilde{F}\|_F^2, \mathbb{E}_0\,n\|F_j\|_F^2 \to 0$ and $\|I_0^{1/2}\|_F, \|I_0^{-1/2}\|_F < \infty$. Hence, for $j = 1, \ldots, K$, denote $M_j = \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} - I_d$. By Assumption (A_9) we have as $m \to \infty$,

$$\mathbb{E}_0\,\|M_j\|_F^2 \to 0, \quad \text{uniformly for } j = 1, \ldots, K. \qquad \text{(C.8)}$$

For $j = 1$, we consider the difference $\|N_d(0, I_0^{-1}) - N_d\big(0, \Sigma_1^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_1^{1/2}\big)\|_{\mathrm{TV}}$. And the corresponding vector v and matrix V_1, V_2 in (C.5) are

$$\begin{aligned}
&v = 0, \\
&V_2 - V_1 = \Sigma_1^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_1^{1/2} - I_0^{-1} = M_1I_0^{-1} + I_0^{-1}M_1^{\mathrm{T}} + M_1I_0^{-1}M_1^{\mathrm{T}}.
\end{aligned}$$

Since $v = 0$, the first two terms of the right hand side of (C.5) is zero. For the third term we have that

$$\begin{aligned}
&\|(N^{\mathrm{T}}V_1N)^{-1}N^{\mathrm{T}}V_2N - I_{d-1}\|_F \\
&\quad = \|(N^{\mathrm{T}}V_1N)^{-1}N^{\mathrm{T}}(V_2 - V_1)N\|_F \\
&\quad = \|(N^{\mathrm{T}}I_0^{-1}N)^{-1}N^{\mathrm{T}}(M_1I_0^{-1} + I_0^{-1}M_1^{\mathrm{T}} + M_1I_0^{-1}M_1^{\mathrm{T}})N\|_F. \qquad \text{(C.9)}
\end{aligned}$$

We have that

$$\begin{aligned}
&\|(N^{\mathrm{T}}I_0^{-1}N)^{-1}N^{\mathrm{T}}(M_1I_0^{-1} + I_0^{-1}M_1^{\mathrm{T}} + M_1I_0^{-1}M_1^{\mathrm{T}})N\|_F \\
&\qquad \overset{(i)}{\leq} \|I_0^{-1}\|_F\|M_1\|_F(2 + \|M\|_F)\|(N^{\mathrm{T}}I_0^{-1}N)^{-1}\|_F\|N\|_F^2 \\
&\qquad \overset{(ii)}{\leq} \|I_0^{-1}\|_F\overline{\lambda}(I_0)\|M_1\|_F(2 + \|M\|_F)(d-1)^{3/2}, \qquad \text{(C.10)}
\end{aligned}$$

where $\overline{\lambda}(I_0)$ denotes the largest singular value of I_0, (i) follows from (C.6), and (ii) follows from $\|(N^{\mathrm{T}}I_0^{-1}N)^{-1}\|_F \leq (d-1)^{1/2}\overline{\lambda}(I_0)$ and $\|N\|_F = \sqrt{d-1}$. Combining (C.5), (C.9), and (C.10), we have that

$$\|N_d(0, I_0^{-1}) - N_d\left(0, \Sigma_1^{1/2}\tilde{\Sigma}^{-1/2}I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_1^{1/2}\right))\|_{\mathrm{TV}} \leq \tilde{C}\|M_1\|_F(2 + \|M_1\|_F), \qquad \text{(C.11)}$$

where $\tilde{C} = \frac{9(d-1)^{3/2}\overline{\lambda}(I_0)\|I_0^{-1}\|_F}{2} < \infty$. By a simlar argument, we can show that for $j = 2, \ldots, K$,

$$\|N_d(0, I_0^{-1}) - N_d\left(0, \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} I_0^{-1} \tilde{\Sigma}^{-1/2}\Sigma_j^{1/2}\right))\|_{\mathrm{TV}} \leq \tilde{C}\|M_j\|_F(2 + \|M_j\|_F). \tag{C.12}$$

Step 3: Show that Theorem 3.2(i) holds for any sequence $K_m \to \infty$ as $m \to \infty$ and $K_m = o(\rho^{-m})$.

For any $K = 1, 2, \ldots$, we define a sequence of random variables $(D_{j,m}^K)_{m=1}^{\infty}$ as

$$D_{j,m}^K = \|\Pi_m^{h_j}(\cdot \mid Y_{(j-1)m+1}^{jm}) - N_d\left(0, I_0^{-1}\right)\|_{\mathrm{TV}}, \quad j = 1, \ldots, K, \;\; m = 1, 2, \ldots. \tag{C.13}$$

Since $D_{j.m}^K \in [0, 1]$ for any $1 \leq j \leq K$ and $K, m \in \mathbb{N}$, hence for any given sequence $K_m \in \mathbb{N}$ with $K_m \overset{m\to\infty}{\to} \infty$, we obtain that $\{D_{j,m}^{K_m}, j = 1, \ldots, K_m, m = 1, 2, \ldots\}$ is uniformly integrable. Together with $D_{j,m}^{K_m}$ converging to 0 in $\mathbb{P}_0$-probability we have that as $m \to \infty$,

$$\mathbb{E}_0\, D_{j,m}^{K_m} \to 0, \quad j = 1, \ldots, K_m. \tag{C.14}$$

Based on definition of subset quasi posterior distribution $\tilde{\Pi}_m(\cdot \mid Y_{(j-1)m+1}^{jm})$, the density of the first subset posterior $\tilde{\Pi}_m(\cdot \mid Y_1^m)$ is a function of Y_1^m, whereas the density of the $\tilde{\Pi}_m(\cdot \mid Y_{(j-1)m+1}^{jm})$ is a function of $Y_{(j-2)m+1}^{jm}$ for $j = 2, \ldots, K_m$. By assumption (A_7), $\{Y_t, t \geq 1\}$ is stationary. Hence we have that for any $K_m, m \geq 1$,

$$\mathbb{E}_0\, D_{1,m}^{K_m} \neq \mathbb{E}_0\, D_{2,m}^{K_m} = \cdots = \mathbb{E}_0\, D_{K_m,m}^{K_m}. \tag{C.15}$$

For any $\epsilon > 0$, we have that as $m \to \infty$,

$$\begin{aligned}
&\mathbb{P}_0\left[\|\tilde{\Pi}_n^h(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} > \epsilon\right] \\
&\overset{(i)}{\leq} \mathbb{P}_0\left[\frac{1}{K_m}\sum_{j=1}^{K_m}\left\{D_{j,m}^{K_m} + \|N_d(0, I_0^{-1}) - N_d(0, \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2})\|_{\mathrm{TV}}\right\} > \epsilon\right] \\
&\overset{(ii)}{\leq} \frac{1}{\epsilon K_m}\sum_{j=1}^{K_m}\mathbb{E}_0\left\{D_{j,m}^{K_m} + \|N_d(0, I_0^{-1}) - N_d(0, \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2})\|_{\mathrm{TV}}\right\} \\
&\overset{(iii)}{=} \frac{1}{\epsilon K_m}\mathbb{E}_0\, D_{1,m}^{K_m} + \frac{K_m - 1}{\epsilon K_m}\mathbb{E}_0\, D_{2,m}^{K_m} + \frac{1}{\epsilon K_m}\sum_{j=1}^{K_m}\mathbb{E}_0\,\|N_d(0, I_0^{-1}) \\
&\quad - N_d(0, \Sigma_j^{1/2}\tilde{\Sigma}^{-1/2} I_0^{-1}\tilde{\Sigma}^{-1/2}\Sigma_j^{1/2})\|_{\mathrm{TV}} \\
&\overset{(iv)}{\leq} \frac{1}{\epsilon K_m}\mathbb{E}_0\, D_{1,m}^{K_m} + \frac{K_m - 1}{\epsilon K_m}\mathbb{E}_0\, D_{2,m}^{K_m} + \frac{\tilde{C}}{\epsilon}\frac{1}{K_m}\sum_{j=1}^{K_m}\mathbb{E}_0\,\|M_j\|_F(2 + \|M_j\|_F)
\end{aligned} \tag{C.16}$$

$\overset{(v)}{\to} 0,$

where (i) follows from (C.4) in step 1, (ii) follows from Markov inequality, (iii) follows from (C.15), (iv) follows from (C.12) in step 2. The convergence in (v) follows from that the first term of (C.16) converges to 0 as $\frac{1}{K_m} \to 0$ and $\mathbb{E}_0 D_{1,m}^{K_m} \to 0$ as $m \to \infty$; the second term of (C.16) converges to 0 as $\frac{K_m - 1}{K_m} \to 1$ and $\mathbb{E}_0 D_{2,m}^{K_m} \to 0$ as $K_m \overset{m\to\infty}{\to} \infty$; for the third term of (C.16), by (C.8) and Jensen's inequality we have $\frac{1}{K_m} \sum_{j=1}^{K_m} \mathbb{E}_0 \|M_j\|_F (2 + \|M_j\|_F) \to 0$ as $m, K_m \to \infty$. And this completes the proof of Theorem 3.2 (i) for any given sequence $K_m \overset{m\to\infty}{\to} \infty$ with $K_m = o(\rho^{-m})$. □

C.2. Proof of Theorem 3.2(ii)

Let μ, ν be two probability measures on $(\Theta, \mathcal{F})$ where $\mathcal{F}$ is the Borel σ-field. The 1-Wasserstein distance is bounded by the total variation distance (Villani, 2008, Theorem 6.15)

$$W_1\{\mu, \nu\} \leq \operatorname{diam}(\Theta) \|\mu - \nu\|_{\mathrm{TV}},$$

where $\operatorname{diam}(\Theta) = \sup_{\theta_1, \theta_2 \in \Theta} \|\theta_1 - \theta_2\|_2 < \infty$ by the compactness of $\Theta \subset \mathbb{R}^d$. Hence, as $m \to \infty$ with $\mathbb{P}_0$-probability tending to 1,

$$\begin{aligned}
&\sqrt{n} W_1 \left\{ \tilde{\Pi}_n(\cdot \mid Y_1^n), N_d(\tilde{\theta}, (nI_0)^{-1}) \right\} \\
&\quad \overset{(i)}{=} W_1 \left\{ \tilde{\Pi}_n^{\tilde{h} = \sqrt{n}(\theta - \tilde{\theta})}(\cdot \mid Y_1^n), N_d(0, I_0^{-1}) \right\} \\
&\leq \operatorname{diam}(\Theta) \|\tilde{\Pi}_n^{\tilde{h}}(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} \overset{(ii)}{\to} 0, \\
&\sqrt{n} W_1 \left\{ \Pi_n(\cdot \mid Y_1^n), N_d(\hat{\theta}, (nI_0)^{-1}) \right\} \\
&\quad \overset{(i)}{=} W_1 \left\{ \Pi_n^{h = \sqrt{n}(\theta - \hat{\theta})}(\cdot \mid Y_1^n), N_d(0, I_0^{-1}) \right\} \\
&\leq \operatorname{diam}(\Theta) \|\Pi_n^{h}(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} \overset{(iii)}{\to} 0, \qquad \text{(C.17)}
\end{aligned}$$

where (i) follows from the rescaling property of W_1-distance, (ii) follows from Theorem 3.2(ii) and (iii) follows from Theorem 3.1 in the case $K = 1$. By Villani (2003), if Θ is bounded, we have the following relation between W_1 distance and W_2 distance. For two square integrable probability measures μ, ν on $(\Theta, \mathcal{F})$,

$$W_2\{\mu, \nu\}^2 \operatorname{diam}(\Theta)^{-1} \leq W_1\{\mu, \nu\} \leq W_2\{\mu, \nu\}, \qquad \text{(C.18)}$$

Furthermore, let μ be the probability measure induced by $N_d(m_1, \Sigma_1)$ and ν be the probability measure induced by $N_d(m_2, \Sigma_2)$, then $W_2(\mu, \nu)$ has a closed form (Olkin and Pukelsheim, 1982)

$$W_2\{N_d(m_1, \Sigma_1), N_d(m_1, \Sigma_1)\} = \|m_1 - m_2\|_2 + \|\Sigma_1^{1/2} - \Sigma_2^{1/2}\|_F. \qquad \text{(C.19)}$$

Now we consider the upper bound of Theorem 3.2 (ii):

$$\begin{aligned}
&\sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), \Pi_n(\cdot \mid Y_1^n)\right\} \\
&\overset{(i)}{\leq} \sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), N_d(\tilde{\theta}, (nI_0)^{-1})\right\} + \sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \\
&\qquad + \sqrt{n}W_1\left\{N_d(\tilde{\theta}, (nI_0)^{-1}), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \\
&\overset{(ii)}{\leq} \operatorname{diam}(\Theta)\left\{\|\tilde{\Pi}_n^{\tilde{h}}(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}} + \|\Pi_n^h(\cdot \mid Y_1^n), N_d(0, I_0^{-1})\|_{\mathrm{TV}}\right\} \\
&\qquad + \sqrt{n}W_1\left\{N_d(\tilde{\theta}, (nI_0)^{-1}), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \qquad \text{(C.20)}
\end{aligned}$$

where (i) follows from triangle inequality, (ii) follows from (C.17). The first term of the right hand side of (C.20) converges to 0 in $\mathbb{P}_0$-probability by (C.17). As for the second term, by (C.18) and (C.19), we have that

$$\operatorname{diam}(\Theta)^{-1}\sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2^2 \leq \sqrt{n}W_1\left\{N_d(\tilde{\theta}, (nI_0)^{-1}), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \leq \sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2. \qquad \text{(C.21)}$$

Hence up to a negligible term in $\mathbb{P}_0$-probability, we have

$$\sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), \Pi_n(\cdot \mid Y_1^n)\right\} \leq \sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2. \qquad \text{(C.22)}$$

For the lower bound of Theorem 3.2(ii), we have that

$$\begin{aligned}
&\sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), \Pi_n(\cdot \mid Y_1^n)\right\} \\
&\geq \sqrt{n}W_1\left\{N_d(\tilde{\theta}, (nI_0)^{-1}), N_d(\hat{\theta}, (nI_0)^{-1})\right\} - \sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), N_d(\tilde{\theta}, (nI_0)^{-1})\right) \\
&\qquad - \sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \\
&\geq \sqrt{n}W_1\left\{N_d(\tilde{\theta}, (nI_0)^{-1}), N_d(\hat{\theta}, (nI_0)^{-1})\right\} \\
&\qquad - \operatorname{diam}(\Theta)\left\{\|\tilde{\Pi}_n^{\tilde{h}}(\cdot \mid Y_1^n), N_d(0, I_0^{-1})\|_{\mathrm{TV}} + \|\Pi_n^h(\cdot \mid Y_1^n) - N_d(0, I_0^{-1})\|_{\mathrm{TV}}\right\}, \qquad \text{(C.23)}
\end{aligned}$$

where the second term of (C.23) converges to 0 in $\mathbb{P}_0$-probability and the first term is lower bounded by $\operatorname{diam}(\Theta)^{-1}\sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2^2$ by (C.21). Hence up to a negligible term in $\mathbb{P}_0$-probability, we have

$$\sqrt{n}W_1\left\{\tilde{\Pi}_n(\cdot \mid Y_1^n), \Pi_n(\cdot \mid Y_1^n)\right\} \geq \operatorname{diam}(\Theta)^{-1}\sqrt{n}\|\tilde{\theta} - \hat{\theta}\|_2^2. \qquad \text{(C.24)}$$

Combining (C.22) and (C.23) completes the proof of Theorem 3.2(ii). □

Appendix D: Further details of experiments

D.1. Illustrative example of simulations

We consider an illustrative example of HMM in Section 4.2 of the main manuscript. Using the notation from Section 2.1 in the main manuscript, let

$\mathcal{Y} = \mathbb{R}$, $g(\cdot \mid X = a) = \phi(\cdot; \mu_a, \sigma_a^2)$ be the density of normal distribution with mean μ_a and variance σ_a^2 $(a = 1, \ldots, S)$, $\theta = \{r, Q, \mu_1, \ldots, \mu_S, \sigma_1^2, \ldots, \sigma_S^2\} \subset \mathbb{R}^d$, and $d = S^2 + 2S - 1$. Our example is motivated from Rydén (2008), who uses a similar example with $S = 3$, $\sigma_1 = \ldots = \sigma_3 = \sigma$, and a specific choice of Q, which automatically determines r because $r^{\mathrm{T}}Q = r^{\mathrm{T}}$. Under this setup, posterior computations on subsets using the data augmentation based on (2.5) in the main manuscript are analytically tractable.

The prior distribution of θ is taken conjugate to the likelihood $p_\theta(X_1^n, Y_1^n)$ in (2.2) of the main manuscript. The initial distribution $r = (r_1, \ldots, r_S)$ and rows of the Markov transition matrix $\{(q_{a1}, \ldots, q_{aS})\}_{a=1}^S$ are given an independent Dirichlet prior $\mathrm{Dir}(1, \ldots, 1)$; the mean parameters $\mu_1, \ldots, \mu_S$ are given an independent normal prior $N(\xi, \kappa^{-1})$ with $\xi = (\min y_i + \max y_i)/2$ and $\kappa = 1/(\max y_i - \min y_i)^2$; the variance parameters $\sigma_1^{-2}, \ldots, \sigma_S^{-2}$ are given an independent gamma prior $\Gamma(1, 1)$.

For $j = 1, \ldots K$, the full conditional distributions of $\theta = \{r, Q, \mu_1, \ldots, \mu_S, \sigma_1^2, \ldots, \sigma_S^2\}$ after stochastic approximation given $(X_{(j-1)m+1}^{(jm)}, Y_{(j-1)m+1}^{(jm)})$ are as follows. The full conditional distribution of r is

$$(r_1, \ldots, r_S) \mid \cdots \sim \mathrm{Dir}\left(K \cdot I\{X_{(j-1)m+1} = 1\} + 1, \ldots, K \cdot I\{X_{(j-1)m+1} = S\} + 1\right) \tag{D.1}$$

where $\mid \cdots$ denotes the conditioning on the remaining random quantities and $I\{\cdot\}$ denotes the indicator function. The rows of transition matrix Q are conditional independent over $a = 1, \ldots, S$ and drawn as

$$(q_{a1}, \ldots, q_{aS}) \mid \cdots \sim \mathrm{Dir}\left(Kn_{a1} + 1, \ldots, Kn_{aS} + 1\right) \tag{D.2}$$

where n_{ab} is the cardinality of the set $\{(j-1)m+1 < i \leq jm \ : \ X_{i-1} = a, X_i = b\}$. The mean parameters are conditionally independent over $a = 1, \ldots, S$, and drawn as

$$\mu_a \mid \cdots \sim \mathrm{N}\left(\frac{KS_a + \kappa\xi\sigma^2}{Kn_a + \kappa\sigma^2}, \frac{\sigma^2}{Kn_a + \kappa\sigma^2}\right) \tag{D.3}$$

where $S_a = \sum_{X_i = a, (j-1)m+1 \leq i \leq jm} y_i$ and n_a is the cardinality of the set $\{(j-1)m + 1 \leq i \leq jm \ : \ X_i = a\}$. The variance parameters are conditionally independent over $a = 1, \ldots, S$ and drawn as

$$\sigma_a^{-2} \mid \cdots \sim \Gamma\left(1 + \frac{K}{2}n_a, 1 + \frac{K}{2}\sum_{i \in S_a}(y_i - \mu_a)^2\right). \tag{D.4}$$

Given the parameters and observed chain $\{Y_t\}$, the hidden Markov chain is updated as

$$\mathbb{P}(X_1 = a \mid \ldots) \propto \{r_a \phi(Y_1; \mu_a, \sigma_a^2)\, \mathbb{P}(Y_2^m \mid X_1 = a, \theta)\}^K, \tag{D.5}$$

and as

$$\mathbb{P}(X_i = b \mid X_{i-1} = a) \propto \{q_{ab} \phi(Y_i; \mu_b, \sigma_b^2)\, \mathbb{P}(Y_{i+1}^{jm} \mid X_i = b, \theta)\}^K, \tag{D.6}$$

for $1 \le j \le K, (j-1)m+1 < i \le jm$. The prediction filter is updated as

$$\begin{aligned}
&\mathbb{P}(X_{(j-1)m+1} = b \mid Y_{(j-2)m+1}^{(j-1)m}) \\
&\quad \propto \sum_a \{\mathbb{P}(X_{(j-1)m+1} = b \mid X_{(j-1)m} = a)\, \mathbb{P}_\theta(Y_{(j-2)m+1}^{(j-1)m}, X_{(j-1)m} = a)\}^K \\
&\quad = \sum_a \{q_{ab}\, \mathbb{P}_\theta(Y_{(j-2)m+1}^{(j-1)m}, X_{(j-1)m} = a)\}^K \qquad \text{(D.7)}
\end{aligned}$$

for $1 < j \le K$, where $\mathbb{P}_\theta(Y_{(j-2)m+1}^{(j-1)m}, X_{(j-1)m} = a)$ is calculated using forward-backward recursion.

The subset sampling scheme is carried out in parallel using Gibbs sampler that alternates between updating parameter θ and the hidden Markov chain X. On the first subset, we draw parameters by cycling through the sequence for the following six steps until the Markov chain convergent to the stationary distribution:

- (a_1) Draw $X_1, \ldots, X_m$ from $\{1, \ldots, S\}$ with equal weights as initial value for hidden Markov chain.
- (a_2) Update $(r_1, \ldots, r_S)$ according to (D.1).
- (a_3) Update Q by drawing $(q_{a1}, \ldots, q_{aS})$ according to (D.2) independently for $a = 1, \ldots, S$.
- (a_4) Update μ_a according to (D.3) independently for $a = 1, \ldots, K$.
- (a_5) Update σ_a^{-2} according to (D.4) independently for $a = 1, \ldots, K$.
- (a_6) Update $(X_1, \ldots, X_m)$ by drawing X_1 from (D.5) and $X_{i-1} \to X_i$ from (D.6) for $i = 2, \ldots, m$.

On jth subset ($j = 2, \ldots, K$), we use the prediction filter on the immediately preceding subset to approximate the initial distribution of the hidden Markov chain; therefore, we require access to $2m$ observations (jth and $(j-1)$th subsets), where the $(j-1)$th subset is used to estimate the prediction filter and the jth subset is used to sample posterior draws θ. For $j = 2, \ldots, K$, with access to $Y_{(j-2)m+1}^{(j-1)m}$ and $Y_{(j-1)m+1}^{jm}$, run the above Gibbs sampler (a_1)–(a_6) on $Y_{(j-2)m+1}^{(j-1)m}$ to get an estimation of prediction filter according to (D.7), and then run Gibbs sampler on $Y_{(j-1)m+1}^{jm}$ while setting $r = (r_1, \ldots, r_S)$ equal to the estimation of prediction filter obtained from $Y_{(j-2)m+1}^{(j-1)m}$.

- (b_1) Run the Gibbs sampler (a_1)–(a_6) with observations $Y_{(j-2)m+1}^{(j-1)m}$.
- (b_2) Calculate the prediction filter weights $(w_1, \ldots, w_S)$ according to (D.7) based on the MCMC estimation of parameters.
- (b_3) Set $(r_1, \ldots, r_S) = (w_1, \ldots, w_S)/\sum_a w_a$ as initial distribution for hidden Markov chain. Then draw $X_{(j-1)m+1}$ with weight $(r_1, \ldots, r_S)$ and $X_{(j-1)m+2}, \ldots, X_{gm}$ with equal weight from $\{1, \ldots, S\}$.
- (b_4) Run the Gibbs sampler (a_3)–(a_6) with observations $Y_{(j-1)m+1}^{cm}$.

If we set $K = 1$, then we get the same sampling scheme as those in Rydén (2008). Let $\{\theta_{(j)}^{(t)}\}_{t=1,j=1}^{T,K}$ be the posterior draws collected from subset posterior

sampling scheme. Then, the final step of divide-and-conquer method simply re-scales and re-centers the scaled and centered draws from subsets as follows:

(c_1) Calculate the maximum likelihood estimator $\hat{\theta}$ using the full data Y_1^n by Baum-Welch EM algorithm and the Monte Carlo estimate of subset posterior variance matrices $(\Sigma_{\theta_{(j)}})_{j=1}^K$ by

$$\mu_{\theta_{(j)}} = \frac{1}{T}\sum_{t=1}^{T}\theta_{(j)}^{(t)}, \quad \Sigma_{\theta_{(j)}} = \frac{1}{T}\sum_{t=1}^{T}\left(\theta_{(j)}^{(t)} - \mu_{\theta_{(j)}}\right)\left(\theta_{(j)}^{(t)} - \mu_{\theta_{(j)}}\right)^{\mathrm{T}},$$
$$j = 1, \ldots, K. \tag{D.8}$$

(c_2) Adopt $COMB(\tilde{\theta} = \hat{\theta}, \tilde{\Sigma} = \frac{1}{K}\sum_{j=1}^{K}\Sigma_{\theta_{(j)}})$ and then calculate the θ draws from block filtered posterior distribution according to the following

$$\tilde{\theta}_{(j)}^{(t)} = \tilde{\theta} + \tilde{\Sigma}^{1/2}\left\{\Sigma_{\theta_{(j)}}^{-1/2}(\theta_{(j)}^{(t)} - \hat{\theta}_j)\right\}, \quad j = 1, \ldots, K;\ t = 1, \ldots, T. \tag{D.9}$$

D.2. Time comparisons and accuracy for posterior inference on Q

Simulation Study of Accuracy and Efficiency. Following the discussion of the simulation study in the main manuscript of HMM with $(S = 3, Q_\epsilon = 0.1, \mu_1 = -2, \mu_2 = 0, \mu_3 = 2)$, we compare the accuracy of block filtered posterior distribution and its divide-and-conquer competitors for posterior inference on Q (Table 5). The divide-and-conquer competitors of block filtered posterior distribution have low accuracy for $n = 10^6$ and all K. In comparison, the block filtered posterior distribution has the highest accuracy for all (n, K) combinations, and its accuracy is close to the benchmark accuracy of the Hamiltonian Monte Carlo. These observations agree with that of the simulation study for posterior inference on the emission distribution parameters in the main manuscript.

The block filtered posterior distribution is about ten times faster than Hamiltonian Monte Carlo and about fifteen times faster than data augmentation when $n = 10^6$ and $K = n^{1/3}$ (Table 6). For all n, the running time of block filtered posterior distribution decreases as K increases from $\log n$ to $n^{1/3}$ because larger K results in smaller subsets size. The run-time of subset posterior sampling algorithms dominates the time required to combine the parameter draws from the subsets; therefore, the block filtered posterior distribution is more efficient than data augmentation and Hamiltonian Monte Carlo for all values of K when $n = 10^5$ and $n = 10^6$.

Simulation Study of Mixing Property. Following the discussion of the simulation study in the main manuscript of three more HMMs with $S = 2, 5, 7$, the accuracy of block filtered posterior distribution and its divide-and-conquer competitors for posterior inference on Q is compared in Tables 5 and 7. For HMMs with $S = 2, 3$, ρ is relatively small and the block filtered posterior distribution has similar accuracy compared to Hamiltonian Monte Carlo over all (n, K) combinations. On the other hand, HMMs with $S = 5, 7$ have $\rho \approx 1$ and the

TABLE 5

Accuracy of the approximate posterior distributions of Q. The accuracies are averaged over ten simulation replications and across all the dimensions of Q. The maximum Monte Carlo error for the block filtered posterior distribution is 0.02

n	K	BFP	WASP	DPMC	PIE	HMC
10^4	$\log n$	0.96	0.87	0.87	0.88	0.96
	$n^{1/4}$	0.96	0.87	0.87	0.87	
	$n^{1/3}$	0.93	0.58	0.57	0.58	
10^5	$\log n$	0.97	0.62	0.62	0.62	0.95
	$n^{1/4}$	0.97	0.63	0.63	0.63	
	$n^{1/3}$	0.97	0.66	0.66	0.66	
10^6	$\log n$	0.96	0.22	0.22	0.23	0.96
	$n^{1/4}$	0.97	0.22	0.22	0.23	
	$n^{1/3}$	0.96	0.25	0.25	0.25	

BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

TABLE 6

Time (in hours) for computing the posterior distributions of θ. The times are averaged over ten simulation replications

	DA	BFP			HMC
		$\log n$	$n^{1/4}$	$n^{1/3}$	
$n = 10^4$	1.36	1.88	0.73	0.33	0.68
$n = 10^5$	14.80	5.58	4.97	1.22	8.60
$n = 10^6$	131.13	59.76	21.16	8.83	91.89

DA, data augmentation; BFP, block filtered posterior distribution; HMC, Hamiltonian Monte Carlo.

block filtered posterior distribution has a smaller accuracy compared to Hamiltonian Monte Carlo. Furthermore, as K increases, the accuracy of block filtered posterior distribution decreases because of smaller subset size and increased dependence.

For HMMs with $S = 2, 3, 5, 7$ and every (n, K) combinations, the block filtered posterior distribution is the top performer compared to its divide-and-conquer competitors. For HMMs with $S = 2, 3$ and relatively weak dependence, the divide-and-conquer competitors have acceptable accuracy, but their accuracy values decrease quickly as the dependence ρ increases and drop to 0 for $S = 5, 7$. In contrast, the block filtered posterior distribution keeps its great performance on the accuracy for $S = 5, 7$ and all (n, K) combinations. These observations agree with that of the simulation study for posterior inference on the emission distribution parameters in the main manuscript.

Real Data Analysis. Following the discussion of the read data analysis in the main manuscript of HMM, we compare the accuracy of block filtered posterior distribution and its divide-and-conquer competitors for posterior inference on Q (Table 8). For $K = \log n, n^{1/4}$, and $n^{1/3}$, the block filtered posterior distribution outperforms its divide-and-conquer competitors, and its accuracy is close to that of Hamiltonian Monte Carlo.

TABLE 7

Accuracy of the approximate posterior distributions of Q. The accuracies are averaged over ten simulation replications and across all the dimensions of Q. The maximum Monte Carlo error for the block filtered posterior distribution is 0.02

n	K	BFP	WASP	DPMC	PIE	HMC
$S = 2$						
10^4	$\log n$	0.97	0.98	0.98	0.98	0.96
	$n^{1/4}$	0.97	0.98	0.98	0.99	
	$n^{1/3}$	0.97	0.96	0.96	0.96	
10^5	$\log n$	0.97	0.98	0.98	0.99	0.96
	$n^{1/4}$	0.97	0.98	0.98	0.99	
	$n^{1/3}$	0.97	0.98	0.98	0.98	
$S = 5$						
10^4	$\log n$	0.65	0.00	0.00	0.00	0.95
	$n^{1/4}$	0.66	0.00	0.00	0.00	
	$n^{1/3}$	0.62	0.00	0.00	0.00	
10^5	$\log n$	0.68	0.00	0.00	0.00	0.95
	$n^{1/4}$	0.65	0.00	0.00	0.00	
	$n^{1/3}$	0.43	0.00	0.00	0.00	
$S = 7$						
10^4	$\log n$	0.56	0.01	0.01	0.00	0.94
	$n^{1/4}$	0.57	0.00	0.01	0.00	
	$n^{1/3}$	0.56	0.00	0.00	0.00	
10^5	$\log n$	0.51	0.00	0.00	0.00	0.94
	$n^{1/4}$	0.54	0.00	0.00	0.00	
	$n^{1/3}$	0.55	0.00	0.00	0.00	

BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

TABLE 8

Accuracy of the approximate posterior distributions of Q. The entries in the table are the median of the accuracies computed across the Q dimensions. The maximum Monte Carlo error for the block filtered posterior distribution is 0.01

K	BFP	WASP	DPMC	PIE	HMC
$10(\log n)$	0.97	0.97	0.96	0.97	0.96
$20(n^{1/4})$	0.95	0.77	0.77	0.78	
$50(n^{1/3})$	0.93	0.43	0.38	0.43	

BFP, block filtered posterior distribution; WASP, Wasserstein posterior; DPMC, double parallel Monte Carlo; PIE, posterior interval estimation; HMC, Hamiltonian Monte Carlo.

References

Aicher, C., Y.-A. Ma, N. J. Foti, and E. B. Fox (2019). Stochastic gradient MCMC for state space models. *SIAM Journal on Mathematics of Data Science 1*(3), 555–587. MR4010763

Andrieu, C., A. Doucet, and R. Holenstein (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(3), 269–342. MR2758115

Andrieu, C., A. Doucet, and V. B. Tadic (2005). Online parameter estimation in general state-space models. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pp. 332–337. IEEE.

Bickel, P. J., Y. Ritov, and T. Rydén (1998). Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *The Annals of Statistics 26*(4), 1614–1635. MR1647705

Cappé, O. (2011). Online EM algorithm for hidden Markov models. *Journal of Computational and Graphical Statistics 20*(3), 728–749. MR2878999

Cappé, O. and E. Moulines (2009). Online Expectation-Maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(3), 593–613. MR2749909

Cappé, O., E. Moulines, and T. Rydén (2006). *Inference in hidden Markov models*. Springer Science & Business Media. MR2159833

Cappé, O., C. P. Robert, and T. Rydén (2003). Reversible jump, birth-and-death and more general continuous time Markov chain Monte Carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(3), 679–700. MR1998628

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software 76*(1).

Celeux, G., M. Hurn, and C. P. Robert (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association 95*(451), 957–970. MR1804450

Chan, H. P., C.-W. Heng, and A. Jasra (2016). Theory of segmented particle filters. *Advances in Applied Probability 48*(1), 69–87. MR3473568

Chopin, N. and S. S. Singh (2015). On particle Gibbs sampling. *Bernoulli 21*(3), 1855–1883. MR3352064

De Gunst, M. and O. Shcherbakova (2008). Asymptotic behavior of Bayes estimators for hidden Markov models with application to ion channels. *Mathematical Methods of Statistics 17*(4), 342–356. MR2483462

Devroye, L., A. Mehrabian, and T. Reddad (2018). The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*.

Ding, D. and A. Gandy (2018). Tree-based particle smoothing algorithms in a hidden Markov model. *arXiv preprint arXiv:1808.08400*.

Foti, N., J. Xu, D. Laird, and E. Fox (2014). Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems*, pp. 3599–3607.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media. MR2265601

Gassiat, E., J. Rousseau, et al. (2014). About the posterior distribution in hidden Markov models with unknown number of states. *Bernoulli 20*(4), 2039–2075. MR3263098

Giordano, R., T. Broderick, and M. I. Jordan (2018). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research 19*(1), 1981–2029. MR3874159

Goldman, J. V. and S. S. Singh (2021). Spatiotemporal blocking of the bouncy

particle sampler for efficient inference in state-space models. *Statistics and Computing 31*(5), 1–15. MR4307388

Guhaniyogi, R., C. Li, T. D. Savitsky, and S. Srivastava (2022). Distributed Bayesian varying coefficient modeling using a Gaussian process prior. *Journal of machine learning research 23*(84), 1–59.

Jensen, J. L. and N. V. Petersen (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *The Annals of Statistics 27*(2), 514–535. MR1714719

Jordan, M. I., J. D. Lee, and Y. Yang (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association 114*(526), 668–681. MR3963171

Kantas, N., A. Doucet, S. S. Singh, J. Maciejowski, and N. Chopin (2015). On particle methods for parameter estimation in state-space models. *Statistical Science 30*(3), 328–351. MR3383884

Le Corff, S. and G. Fort (2013). Online Expectation Maximization based algorithms for inference in hidden Markov models. *Electronic Journal of Statistics 7*, 763–792. MR3040559

Le Gland, F. and L. Mevel (2000). Exponential forgetting and geometric ergodicity in hidden Markov models. *Mathematics of Control, Signals and Systems 13*(1), 63–93. MR1742140

Lehmann, E. L. and G. Casella (1998). *Theory of point estimation*, Volume 31. Springer. MR1639875

Leroux, B. G. (1992). Maximum likelihood estimation for hidden Markov models. *Stochastic Processes and their Applications 40*(1), 127–143. MR1145463

Li, C., S. Srivastava, and D. B. Dunson (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika 104*(3), 665–680. MR3694589

McCoy, J., C. Allison, and Z. Ornelas (2015). Oklahoma supreme court-purchase card audit. https://www.ok.gov/dcs/searchdocs/app/manage_documents.php?att_id=16277. Online; Report Released 06/04/2015.

Minsker, S., S. Srivastava, L. Lin, and D. Dunson (2014). Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1656–1664.

Minsker, S., S. Srivastava, L. Lin, and D. B. Dunson (2017). Robust and scalable bayes via a median of subset posterior measures. *The Journal of Machine Learning Research 18*(1), 4488–4527. MR3763758

Olkin, I. and F. Pukelsheim (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications 48*, 257–263. MR0683223

Pauli, F., W. Racugno, and L. Ventura (2011). Bayesian composite marginal likelihoods. *Statistica Sinica*, 149–164. MR2796857

Robert, C. P., V. Elvira, N. Tawn, and C. Wu (2018). Accelerating MCMC algorithms. *Wiley Interdisciplinary Reviews: Computational Statistics 10*(5), e1435. MR3850448

Robert, C. P., T. Rydén, and D. M. Titterington (1999). Convergence controls for MCMC algorithms, with applications to hidden Markov chains. *Journal of Statistical Computation and Simulation 64*(4), 327–355. MR1748751

Robert, C. P., T. Ryden, and D. M. Titterington (2000). Bayesian inference in hidden Markov models through the reversible jump Markov chain Monte Carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(1), 57–75. MR1747395

Rydén, T. (1994). Consistent and asymptotically normal parameter estimates for hidden Markov models. *The Annals of Statistics*, 1884–1895. MR1329173

Rydén, T. (1997). On recursive estimation for hidden Markov models. *Stochastic Processes and their Applications 66*(1), 79–96. MR1431871

Rydén, T. (2008). EM versus Markov chain Monte Carlo for estimation of hidden Markov models: A computational perspective. *Bayesian Analysis 3*(4), 659–688. MR2469793

Scott, S. L. (2002). Bayesian methods for hidden Markov models: Recursive computing in the 21st century. *Journal of the American statistical Association 97*(457), 337–351. MR1963393

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. E. McCulloch (2016). Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management 11*(2), 78–88.

Shyamalkumar, N. D. and S. Srivastava (2022). An algorithm for distributed Bayesian inference. *Stat 11*(1), e432. MR4449228

Srivastava, S., V. Cevher, Q. Dinh, and D. Dunson (2015). WASP: Scalable Bayes via Barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 912–920.

Srivastava, S., C. Li, and D. B. Dunson (2018). Scalable Bayes via barycenter in Wasserstein space. *The Journal of Machine Learning Research 19*(1), 312–346. MR3862415

Srivastava, S. and Y. Xu (2021). Distributed Bayesian inference in linear mixed-effects models. *Journal of computational and graphical statistics 30*(3), 594–611. MR4313463

Vernet, E. (2015a). Non parametric hidden Markov models with finite state space: Posterior concentration rates. *arXiv preprint arXiv:1511.08624*. MR3331855

Vernet, E. (2015b). Posterior consistency for nonparametric hidden Markov models with finite state space. *Electronic Journal of Statistics 9*(1), 717–752. MR3331855

Villani, C. (2003). *Topics in Optimal Transportation*. Number 58. American Mathematical Soc. MR1964483

Villani, C. (2008). *Optimal Transport: Old and New*, Volume 338. Springer Science & Business Media. MR2459454

Wu, C. and C. P. Robert (2019). Parallelising MCMC via Random Forests. *arXiv preprint arXiv:1911.09698*.

Xue, J. and F. Liang (2019). Double-Parallel Monte Carlo for Bayesian analysis of big data. *Statistics and computing 29*(1), 23–32. MR3905537

Zimmerman, D. L. and V. Núñez-Antón (2009). *Antedependence models for longitudinal data*. Chapman and Hall/CRC. MR2640722