Journal of Machine Learning Research 19 (2018) 1-35 Submitted 02/17; Revised 03/18; Published 08/18

Scalable Bayes via Barycenter in Wasserstein Space

Sanvesh Srivastava SANVESH-SRIVASTAVA@UIOWA.EDU
Department of Statistics and Actuarial Science
University of Iowa
Iowa City, Iowa 52242, USA

Cheng Li STALIC@NUS.EDU.SG
Department of Statistics and Applied Probability
National University of Singapore
Singapore 117546, Singapore

David B. Dunson DUNSON@DUKE.EDU
Departments of Statistical Science, Mathematics, and ECE
Duke University
Durham, North Carolina 27708, USA

Editor: David Blei

Abstract

Divide-and-conquer based methods for Bayesian inference provide a general approach for tractable posterior inference when the sample size is large. These methods divide the data into smaller subsets, sample from the posterior distribution of parameters in parallel on all the subsets, and combine posterior samples from all the subsets to approximate the full data posterior distribution. The smaller size of any subset compared to the full data implies that posterior sampling on any subset is computationally more efficient than sampling from the true posterior distribution. Since the combination step takes negligible time relative to sampling, posterior computations can be scaled to massive data by dividing the full data into sufficiently large number of data subsets. One such approach relies on the geometry of posterior distributions estimated across different subsets and combines them through their barycenter in a Wasserstein space of probability measures. We provide theoretical guarantees on the accuracy of approximation that are valid in many applications. We show that the geometric method approximates the full data posterior distribution better than its competitors across diverse simulations and reproduces known results when applied to a movie ratings database.

Keywords: barycenter; big data; distributed Bayesian computations; empirical measures; linear programming; optimal transportation; Wasserstein distance; Wasserstein space.

1. Introduction

Developing efficient sampling algorithms is an active area of research motivated by tractable Bayesian inference in large sample settings. Sampling remains a primary tool for inference in Bayesian models, with Markov chain Monte Carlo (MCMC) and sequential Monte Carlo (SMC) providing two broad classes of algorithms that are routinely used. Most MCMC and SMC algorithms face problems in scaling up to massive data settings due to memory and computational bottlenecks that arise; this has motivated a rich literature in recent years

proposing a variety of strategies to enable better performance in such settings. Our focus is on proposing a very general divide-and-conquer technique, which is designed to combine results from any posterior sampling algorithm applied in parallel using subsets of the data.

Massive data pose major problems for existing sampling algorithms. First, if full data require multiple machines for storage, then a sampler has access to only a small fraction of the full data stored on the machine where it runs. Posterior sampling given the full data is expensive due to network latency and extensive communication among machines. Second, with sample size n, sampling in hierarchical Bayesian models requires generation of $O(n)$ latent variables, which becomes inefficient as n increases. Finally, even if full data are available to the sampler, sampling can be infeasible in practice because computation of Hessians and acceptance ratios can scale as $O(n^3)$ in some nonparametric models based on Gaussian process priors (Rasmussen and Williams, 2006). A variety of methods exist to address these issues using optimization and sampling.

Optimization-based methods for Bayesian inference obtain an analytic approximation of the full data posterior distribution. The two most common techniques are polynomial approximation (Rue et al., 2009) and projection of the full data posterior distribution on a class of distributions with analytically tractable posterior densities, which includes variational Bayes and expectation propagation (Wainwright and Jordan, 2008; Gelman et al., 2014). Both techniques estimate parameters of the approximate distribution using a variety of optimization algorithms (Tan and Nott, 2013; Kucukelbir et al., 2015; Rezende and Mohamed, 2015; Ranganath et al., 2016). Stochastic approximation significantly improves the efficiency of estimation by accessing the data in small batches and updating the parameter estimates sequentially (Broderick et al., 2013; Hoffman et al., 2013); however, optimization can be nontrivial for complex likelihoods frequently used in hierarchical models. Furthermore, variational Bayes and expectation propagation often have excellent predictive performance but can be highly biased in estimation of posterior uncertainty and dependence (Giordano et al., 2017).

There is extensive work in sampling-based methods for Bayesian inference. The three main techniques used are as follows. First, subsampling-based methods obtain posterior samples conditioned on a small fraction of the data (Maclaurin and Adams, 2015). Coupling of subsampling with modified Hamiltonian or Langevin dynamics improves posterior exploration and convergence to the stationary distribution (Welling and Teh, 2011; Ahn et al., 2012; Chen et al., 2014; Korattikara et al., 2014; Lan et al., 2014; Shahbaba et al., 2014); see Bardenet et al. (2017) for a review. Second, the exact transition kernel in posterior sampling is replaced by an approximation that significantly reduces the time required to finish an iteration of the sampler (Johndrow et al., 2015; Alquier et al., 2016). Finally, divide-and-conquer approaches first divide the data into smaller subsets and sample in parallel across subsets, and then combine the posterior samples from all the subsets. Our focus is on scalable Bayesian methods based on the divide-and-conquer technique. These methods have two subgroups that differ mainly in their sampling scheme for every subset and their method for combining posterior samples obtained from all the subsets.

The first subgroup modifies the prior to sample from the posterior distribution of the parameter conditioned on a data subset. Let k be the number of subsets, $\pi(\theta)$ be the prior density of parameter θ, and $l_i(\theta)$ be the likelihood for subset i ($i = 1, \ldots, k$). Samples from subset posterior distribution i are obtained using $l_i(\theta)$ and $\pi(\theta)^{1/k}$ as the likelihood

and prior. Consensus Monte Carlo combines subset posterior samples by averaging, which has been generalized in many ways (Rabinovich et al., 2015; Scott et al., 2016). This relies heavily on the normality assumption, which is relaxed using a combination based on kernel density estimation (Neiswanger et al., 2014). Both methods perform poorly if the supports of subset posteriors are different, which motivates the combination using the Weierstrass transform and random partition trees (Wang and Dunson, 2013; Wang et al., 2015). These methods offer simple approaches for combining samples from subset posterior distributions but have a major limitation that the sampling algorithm depends on the model parameterization.

The second subgroup modifies the subset likelihood to sample from a subset posterior distribution and combines samples from subset posterior distributions through their geometric center. These methods modify the likelihood to $l_i(\theta)^k$ and use prior $\pi(\theta)$ to sample from subset posterior distribution i $(i = 1, \ldots, k)$. M-Posterior combines subset posterior distributions through their median in the Wasserstein space of order 1 (Minsker et al., 2014, 2017). The robustness of the median implies that it could ignore valuable information in some subset posterior distributions, which motivates combination through the mean in the Wasserstein space of order 2 called Wasserstein Posterior (WASP) (Srivastava et al., 2015). The WASP approach strikes a balance between the generality of sampling and the efficiency of optimization. While WASP can be applied to any data or Bayesian model, its computations are developed for independent identically distributed (*iid*) data and its theoretical properties are unknown.

Our main goal is to study the theoretical properties of WASP and apply WASP in a variety of practical problems. The *iid* assumption of WASP rules out many important practical problems, including regression and classification, where the data are independent and non-identically distributed (*inid*). We relax this assumption and our theoretical results are applicable to *inid* data. Second, we show that if the number of subsets are chosen appropriately, then the WASP achieves almost the same rate of convergence as that of the full data posterior distribution. For linear models with error distribution in the location-scale family, we strengthen this result and show that the WASP and the full data posterior distribution have the same asymptotic mean and asymptotic variance. This implies that WASP can be used as an efficient alternative to the full data posterior distribution in massive data settings. Third, we show that the method for estimating WASP is independent of the form of the model, which implies that WASP is very general and can be easily used for estimating posterior summaries for any function of the model parameters. We emphasize that WASP is not a new sampling algorithm but a general approach to easily extend any existing sampling algorithms for massive data applications.

2. Preliminaries

2.1 Wasserstein Space, Wasserstein Distance, and Wasserstein Barycenter

We recall elementary properties and definitions related to the Wasserstein space of probability measures. Let (Θ, ρ) be a complete separable metric space and $\mathcal{P}(\Theta)$ be the space of

all probability measures on Θ. The Wasserstein space of order 2 is defined as

$$\mathcal{P}_2(\Theta) := \left\{ \mu \in \mathcal{P}(\Theta) : \int_\Theta \rho^2(\theta_0, \theta)\mu(d\theta) < \infty \right\}, \tag{1}$$

where $\theta_0 \in \Theta$ is arbitrary and $\mathcal{P}_2(\Theta)$ does not depend on the choice of θ_0. The space $\mathcal{P}_2(\Theta)$ is equipped with a natural distance between its elements. Let $\mu, \nu \in \mathcal{P}_2(\Theta)$ and $\Pi(\mu, \nu)$ be the set of all probability measures on $\Theta \times \Theta$ with marginals μ and ν, then the Wasserstein distance of order 2 between μ and ν is defined as

$$W_2(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Theta\times\Theta} \rho^2(x, y)\, d\pi(x, y) \right)^{\frac{1}{2}}. \tag{2}$$

In our applications ρ is the Euclidean metric and we refer to $\mathcal{P}_2(\Theta)$ and W_2 as the Wasserstein space and the Wasserstein distance without explicitly mentioning their order. If $\Pi_1, \ldots, \Pi_k$ are a collection of probability measures in $\mathcal{P}_2(\Theta)$, then their barycenter in $\mathcal{P}_2(\Theta)$ is defined as

$$\overline{\Pi} = \operatorname*{argmin}_{\Pi \in \mathcal{P}_2(\Theta)} \sum_{j=1}^{k} \frac{1}{k} W_2^2(\Pi, \Pi_j). \tag{3}$$

This generalizes the Euclidean barycenter, which is the sample mean, to $\mathcal{P}_2(\Theta)$ (Agueh and Carlier, 2011). The barycenter $\overline{\Pi}$ is analytically intractable, except in few special cases. Let $\delta_a(x) = 1$ if $a = x$ and 0 otherwise. If $X_{j1}, \ldots, X_{jm}$ are samples from Π_j $(j = 1, \ldots, k)$, then $\widehat{\Pi}_j(\cdot) = \sum_{i=1}^m \delta_{X_{ji}}(\cdot)/m$ is an empirical measure that approximates Π_j $(j = 1, \ldots, k)$. If $\overline{\Pi}$ is assumed to be an empirical measure, then the optimization problem in (3) reduces to a linear program; see Cuturi and Doucet (2014), Carlier et al. (2015), and Srivastava et al. (2015) for different algorithms to solve this linear program.

2.2 Stochastic Approximation and Subset Posterior Density

Consider a general set-up for *inid* data. Let $Y^{(n)} = (Y_1, \ldots, Y_n)$ be n observations and the distribution of Y_i is $P_{\theta,i}$, $i = 1, \ldots, n$, where θ lies in the parameter space $\Theta \subset \mathbb{R}^p$. Assume that $P_{\theta,i}$ has density $p_i(\cdot|\theta)$ with respect to the Lebesgue measure, so $dP_{\theta,i}(y_i) = p_i(y_i|\theta)dy_i$ and the likelihood given $Y^{(n)}$ is $l(\theta) = \prod_{i=1}^n p_i(y_i|\theta)$. Given a prior distribution Π on Θ that has density π with respect to the Lebesgue measure, the posterior density of θ given $Y^{(n)}$ using Bayes theorem is

$$\pi(\theta \mid Y^{(n)}) = \frac{\prod_{i=1}^n p_i(y_i \mid \theta)\pi(\theta)}{\int_\Theta \prod_{i=1}^n p_i(y_i \mid \theta)\pi(\theta)d\theta} = \frac{l(\theta)\pi(\theta)}{\int_\Theta l(\theta)\pi(\theta)d\theta}. \tag{4}$$

In most cases $\pi(\theta \mid Y^{(n)})$ is analytically intractable, and accurate approximations of $\pi(\theta \mid Y^{(n)})$ are obtained using Monte Carlo methods, such as importance sampling and MCMC, and deterministic approximations, such as Laplace's method and variational Bayes. For example, in the context of logistic regression, $P_{\theta,i}$ is the Bernoulli distribution with mean $1/\left\{1 + \exp(-x_i^T\theta)\right\}$, where x_i^T is the ith row of the design matrix $X \in \mathbb{R}^{n\times p}$ and $\Theta = \mathbb{R}^p$. The posterior density of θ is analytically intractable, and it is typical to rely on Gibbs

samplers based on data augmentation (Bishop, 2006). These samplers introduce latent variables $\{z_i, i = 1, \ldots, n\}$ and alternately sample the latent variables and the parameters from their full conditional distributions. Related algorithms are very common and are computationally prohibitive for large n because they require repeated passes through the whole data.

Divide-and-conquer-type methods resolve this problem by partitioning the data into smaller subsets. Let k be the number of subsets. The default strategy is to randomly allocate samples to subsets. Let $Y_{[j]} \equiv Y_j^{(m_j)} = (Y_{j1}, \ldots, Y_{jm_j})$ denote data on the jth subset, where m_j is the size of the jth subset and $\sum_{j=1}^k m_j = n$. We assume that $m_j = m$ $(j = 1, \ldots, k)$ for ease of presentation, so $n = km$, the likelihood given $Y_{[j]}$ is $l_j(\theta) = \prod_{i=1}^m p_{ji}(y_{ji}|\theta)$, and $l(\theta)$ in (4) equals $\prod_{j=1}^k l_j(\theta)$. Define subset posterior density j given $Y_{[j]}$ as

$$\pi_m(\theta \mid Y_{[j]}) = \frac{\{\prod_{i=1}^m p_{ji}(y_{ji}|\theta)\}^\gamma \pi(\theta)}{\int_\Theta \{\prod_{i=1}^m p_{ji}(y_{ji}|\theta)\}^\gamma \pi(\theta) d\theta} = \frac{l_j(\theta)^\gamma \pi(\theta)}{\int_\Theta l_j(\theta)^\gamma \pi(\theta) d\theta}, \tag{5}$$

where γ is a positive real number such that $g_1 \gamma m \leq n \leq g_2 \gamma m$ for some $g_1, g_2 > 0$. In the present context, we assume that $\gamma = k$ with $g_1 = g_2 = 1$ following Minsker et al. (2014); more general conditions on γ are defined later in Section 3.2. This modified form of subset posterior compensates for the fact that jth subset has access to only (m/n)-fraction of the full data and ensures that $\pi_m(\theta \mid Y_{[j]})$ and $\pi_n(\theta \mid Y^{(n)})$ in (4) have variances of the same order. Minsker et al. (2014) refer to this as *stochastic approximation* because raising $l_j(\theta)$ $(j = 1, \ldots, k)$ to the power γ is equivalent to replicating every X_{ji} $(i = 1, \ldots, m)$ γ-times so that $\pi_m(\theta \mid Y_{[j]})$ $(j = 1, \ldots, k)$ are noisy approximations of $\pi(\theta \mid Y^{(n)})$.

One advantage of using stochastic approximation to define $\pi_m(\theta \mid Y_{[j]})$ in (5) is that off-the-shelf sampling algorithms can be used directly even when the prior density is the form of a discrete mixture. Consider a simple example of univariate density estimation using Dirichlet process (DP) mixtures of Gaussians. Let X_i $(i = 1, \ldots, n)$ be *iid* samples from a distribution P_0 with density p_0. The data are randomly split into k subsets of equal size m. The truncated stick-breaking representation of DP implies that the prior distribution Π on $\mathcal{P}$ has a finite mixture representation, where $\mathcal{P}$ is the set of probability distributions that have a density. We show in the Appendix that modification of the likelihood using stochastic approximation leads to nearly identical subset and full data posterior computations.

Stochastic approximation does not add any extra burden to the computations required for sampling from the subset posterior distribution of θ conditioned on m observations. We raise the likelihood in every subset to the power γ. This is equivalent to replicating observations γ-times, which seems to offset the benefits of partitioning. However, the replication of observation is not required in implementation of the sampler; we simply modify the likelihood in the full data sampler by raising it to the power γ. For example, stochastic approximation is easily implemented using the `increment_log_prob` function in Stan (Stan Development Team, 2014). We provide more examples for a variety of models in Section 4.

A simple logistic regression example demonstrates that $\pi_m(\theta \mid Y_{[j]})$ in (5) is a noisy approximation of $\pi(\theta \mid Y^{(n)})$ in (4). We simulated data for logistic regression with $n = 10^5$, $p = 2$, $\theta = (-1, 1)^T$, and entries of X randomly set to ± 1 (Figure 1). We set $\gamma = k = 40$ and obtained samples of θ from $\pi(\theta \mid Y^{(n)})$ and from $\pi_m(\theta \mid Y_{[j]})$ $(j = 1, \ldots, k)$ using the Stan's HMC sampling algorithm. The contours for the subset and full data posterior densities

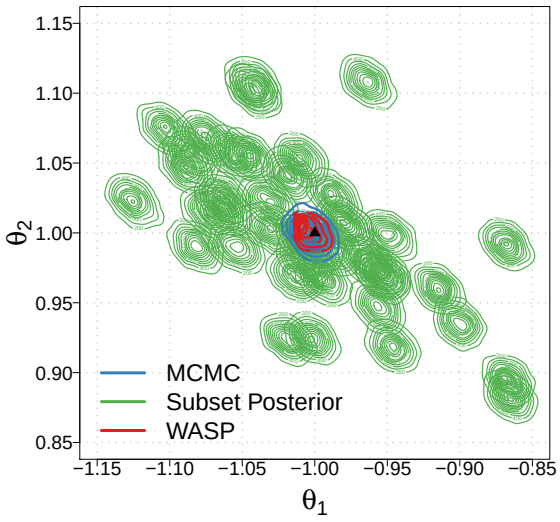

Figure 1: Binned kernel density estimates of full data posterior distribution, subset posterior distributions, and WASP for coefficients (θ_1, θ_2) in logistic regression. The x and y axes represent posterior samples for θ_1 and θ_2. The true values of θ_1 and θ_2 are -1 and 1 (black triangle).

are very similar, indicating all densities have similar spreads. We also notice that subset posteriors are noisy approximations of the full data posterior in that most of them have a bias and do not concentrate at the true θ.

3. Wasserstein Posterior (WASP): The General Framework

3.1 Definition and Estimation of the WASP

The WASP approach combines subset posterior distributions $\Pi_m(\cdot \mid Y_{[j]})$ $(j = 1, \ldots, k)$ through their barycenter in $\mathcal{P}_2(\Theta)$, where the density of $\Pi_m(\cdot \mid Y_{[j]})$ is $\pi_m(\cdot \mid Y_{[j]})$ in (5). The barycenter represents a geometric center of a collection of probability distributions that can be efficiently computed using a linear program. Motivated by this, Srivastava et al. (2015) proposed to combine a collection of subset posterior distributions through their barycenter in the Wasserstein space called *WASP*. Assuming that subset posterior distributions $\Pi_m(\cdot \mid Y_{[j]})$ $(j = 1, \ldots, k)$ have finite second moments, the WASP is defined using (3) as

$$\overline{\Pi}_n(\cdot \mid Y^{(n)}) = \underset{\Pi \in \mathcal{P}_2(\Theta)}{\operatorname{argmin}} \sum_{j=1}^{k} \frac{1}{k} W_2^2\{\Pi, \Pi_m(\cdot \mid Y_{[j]})\}. \tag{6}$$

Consider the following Gaussian example where the WASP is analytically tractable. Assume that the subset posterior distributions, $\Pi_1, \ldots, \Pi_k$, are Gaussian with means $\mu_1, \ldots, \mu_k$ and covariance matrices $\Sigma_1, \ldots, \Sigma_k$. If we fix ρ to be the Euclidean metric and $\Theta = \mathbb{R}^d$ in

Algorithm 1 Estimation of the WASP for $f(\theta)$ given samples of θ from k subset posteriors

Input: Samples from k subset posteriors, $\{\theta_{ji} : \theta_{ji} \sim \Pi_m(\cdot \mid Y_{[j]}),\ i = 1, \ldots, s_j,\ j = 1, \ldots, k\}$; mesh size $\epsilon > 0$.

Do:

1. Define $\phi_i^j = (\phi_{i1}^j, \ldots, \phi_{iq}^j) = f(\theta_{ji})$ $(i = 1, \ldots, s_j;\ j = 1, \ldots, k)$, the matrix of atoms of subset posterior j, $\Phi_j \in \mathbb{R}^{s_j \times q}$, with ϕ_i^j as row i $(i = 1, \ldots, s_j)$. For $r = 1, \ldots, q$, let $\phi_{\min} = (\phi_{\min\ 1}, \ldots, \phi_{\min\ q})$ with $\phi_{\min\ r} = \min_{j\ i} \phi_{ir}^j$, and $\phi_{\max} = (\phi_{\max\ 1}, \ldots, \phi_{\max\ q})$ with $\phi_{\max\ r} = \max_{j\ i} \phi_{ir}^j$.
2. Set the number of atoms in the empirical approximation for the WASP $g = g_1 \times \ldots \times g_q$, where $g_r = \left\lceil \frac{\phi_{\max\ r} - \phi_{\min\ r}}{\epsilon} \right\rceil$ $(r = 1, \ldots, q)$.
3. Define the matrix of WASP atoms $\overline{\Phi} \in \mathbb{R}^{g \times q}$ with rows formed by stacking vectors

$$\left\{ \phi_{\min\ 1} + \tfrac{i_1}{g_1} \left(\phi_{\max\ 1} - \phi_{\min\ 1} \right), \ldots, \phi_{\min\ q} + \tfrac{i_q}{g_q} \left(\phi_{\max\ q} - \phi_{\min\ q} \right) \right\}, \quad (i_r = 1, \ldots, g_r;\ r = 1, \ldots, q).$$

4. Set the distance matrix between the atoms of WASP and the jth subset posterior, $D_j \in \mathbb{R}_+^{g \times s_j}$, as

$$(D_j)_{uv} = \sum_{r=1}^{q} (\overline{\phi}_{ur} - \phi_{vr}^j)^2, \quad (u = 1, \ldots, g;\ v = 1, \ldots, s_j;\ j = 1, \ldots, k),$$

   where $\overline{\phi}_{ur}$ is the (u, r)-entry of $\overline{\Phi}$.
5. Estimate $\hat{a}_1, \ldots, \hat{a}_g$ by solving the linear program (42) in Appendix C.

Return: $\widehat{\overline{f \sharp \Pi}}(\cdot \mid Y^{(n)}) = \sum_{i=1}^{g} \hat{a}_i \delta_{\overline{\phi}_i}(\cdot)$, the atomic approximation of $\overline{f \sharp \Pi}_n(\cdot \mid Y^{(n)})$.

(2), then (3) implies that $\overline{\Pi}_n$ is Gaussian with mean $\overline{\mu}$ and covariance matrix $\overline{\Sigma}$, where

$$\overline{\mu} = \frac{1}{k} \sum_{j=1}^{k} \mu_j \text{ and } \overline{\Sigma} \text{ is such that } \frac{1}{k} \sum_{j=1}^{k} \left(\overline{\Sigma}^{1/2} \Sigma_j \overline{\Sigma}^{1/2} \right)^{1/2} = \overline{\Sigma}, \tag{7}$$

where $A^{1/2}$ is the symmetric square root of A (Agueh and Carlier, 2011). If θ is one dimensional, then (7) says that the standard deviation of WASP is the average of standard deviations of subset posteriors; therefore, the variance of WASP is typically about the same order as that of any subset posterior distribution. A similar relation also holds in higher dimensions and for a large class of posterior distributions, including elliptical distributions (Álvarez-Esteban et al., 2016).

The WASP is analytically tractable only in special cases, but it can be estimated using a linear program if the subset posterior distributions have an atomic form. Let $\{\theta_{j1}, \ldots, \theta_{jS}\}$ be the θ samples obtained from subset posterior density j in (6) using a sampling algorithm, including HMC, MCMC, SMC, or importance sampling. Approximate jth subset posterior distribution $\Pi_m(\cdot \mid Y_{[j]})$ using the empirical measure

$$\hat{\Pi}_m(\cdot \mid Y_{[j]}) = \sum_{i=1}^{S} \frac{1}{S} \delta_{\theta_{ji}}(\cdot) \quad (j = 1, \ldots, k). \tag{8}$$

Srivastava et al. (2015) approximate the WASP as

$$\hat{\overline{\Pi}}_n(\cdot \mid Y^{(n)}) = \sum_{j=1}^{k} \sum_{i=1}^{S} a_{ji} \delta_{\theta_{ji}}(\cdot), \quad 0 \le a_{ji} \le 1, \quad \sum_{j=1}^{k} \sum_{i=1}^{S} a_{ji} = 1, \tag{9}$$

where a_{ji} $(j = 1, \ldots, k;\ i = 1, \ldots, S)$ are unknown weights of the atoms. There are many specialized algorithms to estimate the WASP that exploit the structure of the linear program in (6) when $\Pi_m(\cdot \mid Y_{[j]})$ and $\overline{\Pi}_n(\cdot \mid Y^{(n)})$ are restricted to have atomic forms in (8) and (9), respectively; for example, Cuturi and Doucet (2014) extend the Sinkhorn algorithm using entropy-smoothed sub-gradient methods, Carlier et al. (2015) develop a non-smooth optimization algorithm, and Srivastava et al. (2015) propose an efficient linear program that exploits the sparsity of constraints to solve (6). A simple and efficient algorithm to find the WASP of a given function of parameters is summarized in Algorithm 1.

3.2 Theoretical Properties of the WASP

The WASP, denoted as $\overline{\Pi}_n$, replaces the full data posterior distribution, denoted as Π_n, for inference and prediction in massive data applications where n is large. In motivating applications, computation of Π_n is inefficient, and dividing the data into smaller subsets and performing posterior computations in parallel leads to massive speed-ups. A formal asymptotic justification for using $\overline{\Pi}_n$ to approximate Π_n would ideally show that the distance between $\overline{\Pi}_n$ and Π_n tends to 0 as the full data size n increases to infinity. We will illustrate this using a linear model example in Section 3.2.1, where we show that $n^{1/2} W_2(\overline{\Pi}_n, \Pi_n) \to 0$ as $n \to \infty$. Since both $\overline{\Pi}_n$ and Π_n have variances of order n^{-1}, our result implies that the mean and the variance of WASP match those of the full data posterior distribution.

A general theoretical justification for using $\overline{\Pi}_n$ in the place of Π_n for a multivariate θ given *inid* data is technically much more challenging. If the data are *iid* and θ is one-dimensional, then Li et al. (2017) proves that $n^{1/2} W_2(\overline{\Pi}_n, \Pi_n) \to 0$ as $n \to \infty$ for regular parametric models. The proof in Li et al. (2017) relies heavily on the Bernstein-von Mises theorem (BvM) for *iid* data and the one-dimensional quantile representation of Wasserstein distance. Unlike the *iid* case, a BvM-type theorem is generally unavailable if the data are *inid* or the model is non-regular (Ibragimov and Has' Minskii, 2013). In Section 3.2.2, we show that the WASP $\overline{\Pi}_n$ converges to the true parameter value at almost the same rate as Π_n when the number of subset k increases slowly with n. The previous theoretical justification of WASP in Srivastava et al. (2015) only includes posterior consistency under the stronger *iid* assumption without characterizing the convergence rate. Relaxing these limitations, we provide the convergence rate for the WASP in the *inid* case, including the convergence rate for WASP of general functionals of the original parameters.

3.2.1 Approximation Error of WASP for inid Data: Weighted Linear Model Example

We use a weighted linear model example to illustrate the theoretical approximation accuracy of WASP to the true posterior under the *inid* setup. For $i = 1, \ldots, n$, let y_i be a scalar response, x_i be a $p \times 1$ vector of predictors, and ϵ_i be the idiosyncratic error in y_i. Let $\theta = (\theta_1, \ldots, \theta_p)^T$ be the $p \times 1$ regression coefficients vector. Let $y = (y_1, \ldots, y_n)^T$, $X = [x_1, \ldots, x_n]^T$, and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ be the $n \times 1$ response vector, the $n \times p$ design matrix, and the $n \times 1$ error vector, respectively. If Σ is a known diagonal matrix with positive elements and $\mathrm{cov}(\epsilon) = \Sigma$, then the weighted linear regression model of y on X with a flat prior on θ assumes that

$$y = X\theta + \epsilon, \quad \epsilon \sim N_n(0, \Sigma), \quad \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2), \quad \pi(\theta) \propto 1, \tag{10}$$

where $\pi(\theta)$ is the flat prior on θ and $N_n(0, \Sigma)$ is a n-variate Gaussian distribution with $n \times 1$ mean 0 and covariance Σ. In this case, the data are *inid* since the distribution of y_i depends on the value of x_i. Since Σ is assumed to be known, the posterior distribution of θ is normal with mean $\mu = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y$ and covariance matrix $V = (X^T\Sigma^{-1}X)^{-1}$. Although the posterior of θ has a closed form in this example, the computational complexity of finding μ and V is $O(n^2)$, which becomes inefficient as the size of the data n increases.

The WASP of θ in (10) is analytically tractable. The computation of WASP has three steps. First, the training data are randomly split into k subsets. Let y_j, X_j, and Σ_j be the response vector, design matrix, and error covariance matrix specific to subset j ($j = 1, \ldots, k$). Second, we compute the subset posterior distributions after stochastic approximation on the k subsets in parallel as in (5) with $\gamma = k$. The jth subset posterior distribution of θ is $N_p(\mu_j, V_j)$, where $\mu_j = (X_j^T\Sigma_j^{-1}X_j)^{-1}X_j^T\Sigma_j^{-1}y_j$ and $V_j = k^{-1}(X_j^T\Sigma_j^{-1}X_j)^{-1}$. Third, (7) implies that the WASP of θ is also Gaussian with mean vector $\overline{\mu}$ and covariance matrix $\overline{V}$, where $\overline{\mu} = k^{-1}\sum_{j=1}^k \mu_j$ and $\overline{V}$ satisfies $\overline{V} = k^{-1}\sum_{j=1}^k (\overline{V}^{1/2}V_j\overline{V}^{1/2})^{1/2}$.

The WASP and full data posterior distributions lead to the same posterior inference on θ up to $o(n^{-1})$ terms. Let $\overline{\Pi}_n = N_p(\overline{\mu}, \overline{V})$ and $\Pi_n = N_p(\mu, V)$ be the WASP and full data posterior distributions for θ. Based on the divide-and-conquer technique, the computational complexity of $\overline{\Pi}_n$ is $O(km^2)$, which is smaller than that of Π_n by a factor of k. The true distribution of y, denoted as $P_{\theta_0}^{(n)}$, in (10) is $N_n(X\theta_0, \Sigma)$. If uncertainty quantification using $\overline{\Pi}_n$ and Π_n is the same, then it suffices to show that the difference in the second moments of $\overline{\Pi}_n$ and Π_n is $o(n^{-1})$ in $P_{\theta_0}^{(n)}$-probability because the variances $\overline{V}$ and V are both of order n^{-1}. This is equivalent to showing that the W_2 distance between $\overline{\Pi}_n$ and Π_n is $o(n^{-1})$ in $P_{\theta_0}^{(n)}$-probability, which is proved in the next theorem. In the statement of the theorem, we denote $A \prec B$ for positive definite matrices A and B if $B - A$ is also positive definite.

Theorem 1 *Assume that there exist* $a_n = o(1)$, $b_m = o(1)$ *such that* $\Omega_0 - a_nI_p \prec \frac{1}{n}X^T\Sigma^{-1}X \prec \Omega_0 + a_nI_p$ *and* $\Omega_0 - b_mI_p \prec \frac{1}{m}X_j^T\Sigma_j^{-1}X_j \prec \Omega_0 + b_mI_p$ *for all* $j = 1, \ldots, k$, *where* I_p, Ω_0 *are* $p \times p$ *identity and constant positive definite matrices. Then,*

$$E_{P_{\theta_0}^{(n)}}\|\overline{\mu} - \mu\|_2^2 = o\left(n^{-1}\right), \quad \text{tr}\left(\overline{V} - V\right) = o\left(n^{-1}\right), \quad E_{P_{\theta_0}^{(n)}}W_2^2(\overline{\Pi}_n, \Pi_n) = o\left(n^{-1}\right).$$

The proof of this theorem is in the appendix along with other proofs.

Theorem 1 shows that the uncertainty quantification of Π_n and $\overline{\Pi}_n$ are the same in $P_{\theta_0}^{(n)}$-probability for the data following the model in (10). Essentially, the WASP and the true posterior have the same posterior mean and posterior variance, and their differences are only in high order of the full data size n. Furthermore, Theorem 1 is valid for any block diagonal Σ as long as the data that belong to a particular diagonal block of Σ also belong to the same partition. In other words, Theorem 1 even holds for dependent data in which the dependence can be expressed as a block diagonal Σ in (10). Finally, Theorem 1 is in fact true for any error distribution satisfying $E(\epsilon) = 0$ and $\text{cov}(\epsilon) = \Sigma$, which includes the Gaussian distribution; see Definition 2.1 and Theorem 2.3 in Álvarez-Esteban et al. (2016).

3.2.2 General Convergence Rates of the WASP for inid Data

For general non-iid data, the standard Bayesian asymptotic theory for posterior convergence rates has been established in Ghosal and van der Vaart (2007), which also includes our *inid*

setup. We follow the theoretical framework of Ghosal and van der Vaart (2007) and develop the corresponding theory for divide-and-conquer Bayesian inference using the WASP.

We start with two definitions required to state the assumptions of our theoretical setup.

Definition 2 (Pseudo Hellinger distance) *The pseudo Hellinger distance between probability measures* $P_{\theta_1}^{(m)}, P_{\theta_2}^{(m)} \in \{\otimes_{i=1}^m P_{\theta,j,i} : \theta \in \Theta,\ dP_{\theta,j,i}(y) = p_{ji}(y \mid \theta)dy\}$ *is* $h_{mj}^2(\theta_1,\theta_2) = \frac{1}{m}\sum_{i=1}^m h^2\{p_{ji}(\cdot \mid \theta_1), p_{ji}(\cdot \mid \theta_2)\}$, *where* $h(p_1,p_2) = [\int\{\sqrt{p_1(y)} - \sqrt{p_2(y)}\}^2 dy]^{1/2}$ *is the Hellinger distance between two generic densities* p_1, p_2.

This definition generalizes the usual Hellinger distance to account for the *inid* data generating mechanism. The space $(\{\otimes_{i=1}^m P_{\theta,j,i} : \theta \in \Theta\}, h_{mj})$ is a metric space.

Definition 3 (Generalized bracketing entropy) *Let* Ξ *be a fixed subset of* Θ. *For an* m*-dimensional random vector* $Z = (Z_1,\ldots,Z_m)^T$, *denote its* L_q *norm as* $|Z|_q = \left[\frac{1}{m}\sum_{i=1}^m E\left(|Z_i|^q\right)\right]^{1/q}$ *and use* $\|Z\|$ *to represent* $|Z|_2$. *For a fixed* $j \in \{1,\ldots,k\}$, *let*

$$\mathcal{P}_j(\Xi) = \left\{\boldsymbol{p}_j(\boldsymbol{y}|\theta) = (p_{j1}(y_1|\theta),\ldots,p_{jm}(y_m|\theta))^T : \boldsymbol{y} = (y_1,\ldots,y_m)^T \in \otimes_{i=1}^m \mathcal{Y}_{ji}, \theta \in \Xi\right\}$$

be the class of m*-dimensional functions indexed by* θ. *For a given* $\delta > 0$, *let*

$$\begin{aligned}\mathcal{B}\left(\delta, \mathcal{P}_j(\Xi)\right) = \Big\{&[\boldsymbol{l}_s, \boldsymbol{u}_s] : \boldsymbol{l}_s(\boldsymbol{y}) = (l_{s1}(y_1),\ldots,l_{sm}(y_m))^T, \boldsymbol{u}_s(\boldsymbol{y}) = (u_{s1}(y_1),\ldots,u_{sm}(y_m))^T,\\ &\boldsymbol{y} = (y_1,\ldots,y_m)^T \in \otimes_{i=1}^m \mathcal{Y}_{ji}, s = 1,\ldots,N\Big\}\end{aligned}$$

be the generalized bracketing set of $\mathcal{P}_j(\Xi)$ *with cardinality* N, *such that for any* $\boldsymbol{p}_j(\boldsymbol{y}|\theta) \in \mathcal{P}_j(\Xi)$, *there exists a pair of functions* $[\boldsymbol{l}_s, \boldsymbol{u}_s] \in \mathcal{B}\left(\delta, \mathcal{P}_j(\Xi)\right)$, *such that*

$$\begin{aligned}&l_{si}(y_i) \le p_{ji}(y_i) \le u_{si}(y_i),\ \textit{for all } \boldsymbol{y} \in \otimes_{i=1}^m \mathcal{Y}_{ji},\ \textit{and all } i = 1,\ldots,m\\ &\textit{and } \|\sqrt{\boldsymbol{u}_s} - \sqrt{\boldsymbol{l}_s}\| \le \delta.\end{aligned}$$

The h_{mj}*-bracketing number of* $\mathcal{P}_j(\Xi)$, $N_{[]}\left(\delta, \mathcal{P}_j(\Xi), h_{mj}\right)$, *is defined as the smallest cardinality of the generalized bracketing set* $\mathcal{B}\left(\delta, \mathcal{P}_j(\Xi)\right)$. *The* h_{mj}*-bracketing entropy of* $\mathcal{P}_j(\Xi)$ *is defined as* $H_{[]}\left(\delta, \mathcal{P}_j(\Xi), h_{mj}\right) = \log\left(1 + N_{[]}\left(\delta, \mathcal{P}_j(\Xi), h_{mj}\right)\right)$.

Again, this definition generalizes the usual bracketing entropy to the *inid* cases. If the data are indeed *iid*, then Definition 3 coincides with that of the usual bracketing entropy.

Our theory for the convergence rate of WASP is built on the following assumptions.

(A1) Θ is a compact space in ρ metric, θ_0 is an interior point of Θ, and $g_1\gamma m \le n \le g_2\gamma m$ for some constants $g_1, g_2 > 0$.

(A2) For any $\theta, \theta' \in \Theta$ and $j = 1,\ldots,m$, there exist positive constants α and C_L such that $h_{mj}^2(\theta,\theta') \ge C_L\rho^{2\alpha}(\theta,\theta')$, where h_{mj}^2 is the pseudo Hellinger distance in Definition 2.

(A3) (Entropy Condition) There exist constants $D_1 > 0$, $0 < D_2 < D_1^2/2^{12}$, a function $\Psi(u, r) \geq 0$ that is nonincreasing in $u \in \mathbb{R}^+$ and nondecreasing in $r \in \mathbb{R}^+$, such that for all $j = 1, \ldots, k$, for any $u, r > 0$ and for all sufficiently large m,

$$H_{[]}\left(u, \{\boldsymbol{p}_j(\boldsymbol{y}|\theta) : \theta \in \Theta, h_{mj}(\theta, \theta_0) \leq r\}, h_{mj}\right) \leq \Psi(u, r) \text{ for all } j = 1, \ldots, k;$$
$$\text{and } \int_{D_1 r^2/2^{12}}^{D_1 r} \sqrt{\Psi(u, r)} du < D_2 \sqrt{m} r^2,$$

where $\boldsymbol{p}_j(\boldsymbol{y}|\theta) = \{p_{j1}(y_{j1} \mid \theta), \ldots, p_{jm}(y_{jm} \mid \theta)\}^T$ and $H_{[]}$ is the h_{mj}-bracketing entropy of the set $\{\boldsymbol{p}_j(\boldsymbol{y}|\theta) : \theta \in \Theta, h_{mj}(\theta, \theta_0) \leq r\}$ in Definition 3.

(A4) (Prior Thickness) There exist positive constants κ and c_π, such that uniformly over all $j = 1, \ldots, k$,

$$\Pi\left(\theta \in \Theta : \frac{1}{m} \sum_{i=1}^{m} E_{P_{\theta_0}} \exp\left(\kappa \log_+ \frac{p_{ji}(Y_{ji}|\theta_0)}{p_{ji}(Y_{ji}|\theta)}\right) - 1 \leq \frac{\log^2 m}{m}\right) \geq \exp(-c_\pi k \log^2 m)$$

where $\log_+ x = \max(\log x, 0)$ for $x > 0$.

(A5) The metric ρ satisfies $\rho(\sum_{i=1}^N w_i \theta_i, \theta') \leq \sum_{i=1}^N w_i \rho(\theta_i, \theta')$ for any $N \in \{1, 2, \ldots\}$, $\theta_1, \ldots, \theta_N, \theta' \in \Theta$ and non-negative weights $\sum_{i=1}^N w_i = 1$.

Our assumptions above are based on the standard assumptions in Bayesian asymptotic theory. Similar to Theorem 10 in Ghosal and van der Vaart (2007), we have assumed a compact support in (A1) and lower bounded pseudo Hellinger distance in (A2). Typically, $\alpha = 1$ for most regular models, such as generalized linear models. If the model is non-regular, then α can be less than 1; for example, the densities may have discontinuities depending on the parameter (Ibragimov and Has' Minskii, 2013, Chapters V, VI). Assumption (A3) parallels the entropy condition used in Theorem 1 of Wong and Shen (1995), which has been adapted here for the *inid* setup using the generalized bracketing entropy, and will simplify to a similar entropy condition to that in Theorem 1 of Wong and Shen (1995) if the data are *iid*. Assumption (A4) is crucial in providing a stronger control over the tail probability as the posterior probability mass moves away from the true parameter θ_0, typically with an exponentially decaying rate. The convexity property of ρ in (A5) is mainly used to establish an averaging inequality under W_2 distance and is satisfied by, for example, the Euclidean metric and L_q metric with $q \geq 1$.

The posterior risks of Π_n and $\overline{\Pi}_n$ in the ρ metric is directly related to the W_2 distance based on the ρ metric. If θ_0 denotes the true parameter value from which the data are generated, then the posterior risk of Π_n in the estimation of θ_0 is

$$\int_{\mathcal{Y}^{(n)}} \int_\Theta \rho^2(\theta, \theta_0) d\Pi_n(\theta \mid Y^{(n)}) dP_{\theta_0}^{(n)}(y_1, \ldots, y_n) = E_{P_{\theta_0}^{(n)}}\left[W_2^2\left\{\Pi_n(\cdot \mid Y^{(n)}), \delta_{\theta_0}(\cdot)\right\}\right]. \quad (11)$$

The classical result says that the posterior risk (11) in regular parametric models converges to zero at the n^{-1} rate under assumptions similar to (A2)–(A4), with m replaced by n (van der Vaart, 2000). The next theorem shows that the same posterior risk of the WASP converges at a similar rate to that of the true posterior Π_n, which mainly depends on the size of subsets m, and can be made close to the standard n^{-1} rate up to some logarithmic factors for regular parametric models.

Theorem 4 *If Assumptions (A1)-(A4) hold for the jth subset posterior* $\Pi_m(\cdot \mid Y_{[j]})$ *(*$j = 1, \ldots, k$*), then there exists a constants universal* $C_1 > 0$ *independent of* j*, such that as* $m \to \infty$*,*

$$E_{P_{\theta_0}^{(m)}} \left[W_2^2 \left\{ \Pi_m(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot) \right\} \right] \leq C_1 \left(\frac{\log^2 m}{m} \right)^{\frac{1}{\alpha}}, \quad j = 1, \ldots, k. \tag{12}$$

Additionally, if Assumption (A5) holds, then as $m \to \infty$*,*

$$E_{P_{\theta_0}^{(n)}} \left[W_2^2 \left\{ \overline{\Pi}_n(\cdot \mid Y^{(n)}), \delta_{\theta_0}(\cdot) \right\} \right] \leq C_1 \left(\frac{\log^2 m}{m} \right)^{\frac{1}{\alpha}}. \tag{13}$$

Theorem 4 proves posterior convergence in expectation, which is stronger than the commonly studied posterior convergence in probability. We present our results using the W_2 distance in order to account for the fact that the k subset posteriors sit on a common parameter space. Alternatively, from (11), the convergence rates in (12) and (13) are also the rates of posterior risks for the subset posterior distributions and the WASP. For regular models with $\alpha = 1$, if the number of subsets k increases slowly with n (e.g., $k = O(\log^c n)$ for some constant $c > 0$), then Theorem 4 implies that the WASP converges in W_2 distance at a near optimal convergence rate $O_p(n^{-1/2} \log^{c/2+1} n)$ to δ_{θ_0}. In this case, the standard parametric convergence rate of Π_n is $O_p(n^{-1/2})$, so the WASP attains the optimal convergence rate up to the $\log^{c/2+1} n$ factor. Equivalently, using (11), the posterior risk of the WASP converges to zero at the near optimal rate $O_p(n^{-1} \log^{c+2} n)$, compared to the $O_p(n^{-1})$ posterior risk of the true posterior Π_n.

In most applications, the interest also lies in functions of θ. Suppose $f : \Theta \mapsto \mathbb{R}^q$ is a function that maps θ to $\{f_1(\theta), \ldots, f_q(\theta)\}$, where $q \geq 1$ is a positive integer. A direct application of Lemma 8.5 in Bickel and Freedman (1981) gives the following corollary about the WASP of a function of θ. As long as the function is bounded almost linearly by the ρ metric in (1), its WASP possesses the same posterior convergence rate as in Theorem 4.

Corollary 5 *Suppose* $f(\cdot) = \{f_1(\cdot), \ldots, f_q(\cdot)\}$ *is a function that maps* $\Theta \mapsto \mathbb{R}^q$ *such that* $|f(\theta)|^2 = \sum_{i=1}^q \{f_i(\theta)\}^2 \leq C_f\{1 + \rho^2(\theta, \theta_0)\}$*, where* $C_f > 0$ *is a fixed constant. If the conditions in Theorem 4 hold and* $\overline{f\sharp\Pi}_n(\cdot \mid Y^{(n)})$ *represents the WASP of the subset posterior distributions for* $f(\theta)$*, then as* $m \to \infty$*,*

$$W_2 \left\{ \overline{f\sharp\Pi}_n(\cdot \mid Y^{(n)}), \delta_{f(\theta_0)}(\cdot) \right\} = O_{P_{\theta_0}^{(n)}} \left(\sqrt{\frac{\log^{2/\alpha} m}{m^{1/\alpha}}} \right).$$

Corollary 5 is very useful in applications because it says that the combination step in the WASP is independent of the model parametrization. Let $f\sharp\Pi_m(\cdot \mid Y_{[j]})$ be the jth subset posterior distribution for $f(\theta)$ $(j = 1, \ldots, k)$, then the WASP of k subset posterior distributions converges to $f(\theta_0)$ at the rate obtained in Theorem 4. In practice, we have S_j posterior samples of θ obtained from subset posterior j denoted as θ_{ji} $(i = 1, \ldots, s_j;$ $j = 1, \ldots, k)$. Algorithm 1 estimates an atomic approximation of $\overline{f\sharp\Pi}_n(\cdot \mid Y^{(n)})$, denoted as $\widehat{\overline{f\sharp\Pi}}_n(\cdot \mid Y^{(n)})$, based on the subset posterior samples $f(\theta_{ji})$ $(i = 1, \ldots, s_j;\ j = 1, \ldots, k)$.

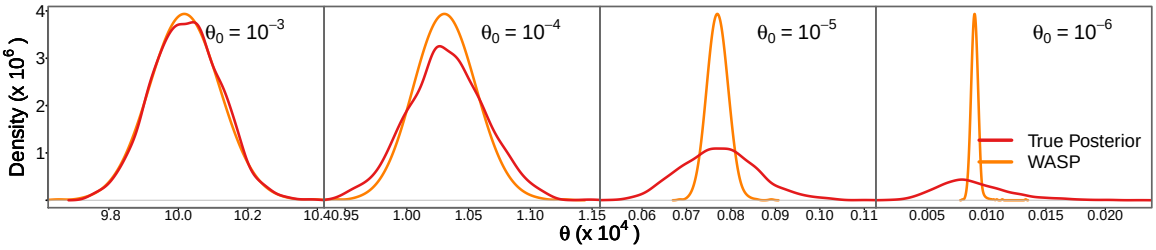

Figure 2: Kernel density estimates of the posterior densities of θ in the rare events example where assumption (A1) fails to hold for $\theta_0 = 10^{-5}, 10^{-6}$.

The atomic form of the WASP is supported on a grid with mesh-size ϵ estimated from the subset posterior samples of $f(\theta)$. Algorithm 1 estimates the weights of the atoms located on the grid by solving a discrete version of (6). The theoretical properties of discrete barycenters imply that $\widehat{f\sharp\Pi}_n(\cdot \mid Y^{(n)})$ is supported only on $O(k)$ elements of the grid; see Theorem 2 in Anderes et al. (2016). We exploit this sparsity by adapting the algorithm in Srivastava et al. (2015) and by using Gurobi (Gurobi Optimization Inc., 2014).

A key assumption in Theorem 4 and Corollary 5 is that the subset posterior distributions provide a noisy approximation of the full data posterior distribution. This is stated precisely in (12), which shows that the convergence rate of a subset posterior distribution in W_2 distance is obtained by using m as the sample size instead of n. If any of the assumptions (A1)–(A4) fail, then the subset posterior distributions may approximate the full data posterior distribution poorly, which could possibly lead to poor approximation quality for the WASP.

A simple example based on rare events demonstrates this phenomenon. Let $Y_1, \ldots, Y_n$ be *iid* Bernoulli random variables with unknown success probability $\theta \in (0, 1)$. The assumption (A1) is violated if the true parameter θ_0 is very close to 0; that is, observing 1 is a rare event. In our simulation example, we set $n = 10^7$ and $\theta_0 = 10^{-a}$ for $a = 3, 4, 5, 6$ so that as a increases, $s = \sum_{i=1}^{n} Y_i$ decreases and θ_0 gets closer and closer to the boundary of the parameter space. The standard Bayesian approach is to put Jefferys' prior Beta(0.5, 0.5) on θ and perform inference on θ using Beta($s + 0.5$, $n - s + 0.5$), which leads to a full data posterior that concentrates around the correct value of θ_0 even if θ_0 is small (Figure 2). However, if the data are randomly divided into $k = 100$ subsets, then a majority of the subsets contain only 0s as θ_0 decreases. As a result, a majority of the subset posterior distributions differ significantly in shape from the full data posterior distribution, leading to a failure of the WASP in approximating the full data posterior distribution because the assumption (A1) is severely violated for $\theta_0 = 10^{-5}, 10^{-6}$ (Figure 2).

4. Experiments

4.1 Setup

We compared WASP with consensus Monte Carlo (CMC) (Scott et al., 2016), semiparametric density product (SDP) (Neiswanger et al., 2014), and variational Bayes (VB). The sample sizes and the number of parameters in our experiments were chosen such that sam-

pling from the full data posterior distribution was computationally feasible. Every sampling algorithm ran for 10,000 iterations. We discarded the first 5,000 samples as burn-in and thinned the chain by collecting every fifth sample. Convergence of the chains to their stationary distributions was confirmed using trace plots. All experiments ran on an Oracle Grid Engine cluster with 2.6GHz 16 core compute nodes. Full data posterior computations were allotted memory resources of 64GB, and all other methods were allotted memory resources of 16GB.

The sampling algorithm for the full data posterior was modified to obtain samples from the subset posteriors in CMC, SDP, and WASP. The sampling algorithms for subset posteriors in CMC and SDP were the same and were based on Equation (2) in Scott et al. (2016). The sampling algorithm for subset posteriors in WASP was based on (5). Samples from the approximate posterior distributions of θ in CMC, SDP, and WASP were obtained in two steps. First, samples from subset posteriors of θ were obtained in parallel across k subsets. Second, the samples of θ from all the subsets were combined using implementations of CMC and SDP in `parallelMCMC` package (Miroshnikov and Conlon, 2014) and using Algorithm 1 for the WASP.

The full data posterior distribution obtained using MCMC served as the benchmark in all our comparisons. Let $\pi(\theta \mid Y^{(n)})$ be the density of the full data posterior distribution for θ estimated using sampling and $\hat{\pi}(\theta \mid Y^{(n)})$ be the density of an approximate posterior distribution for θ estimated using the WASP or its competitors. We used the following metric based on the total variation distance to compare the accuracy $\hat{\pi}(\theta \mid Y^{(n)})$ in approximating $\pi(\theta \mid Y^{(n)})$

$$\text{accuracy}\left\{\hat{\pi}(\theta \mid Y^{(n)})\right\} = 1 - \frac{1}{2}\int_{\Theta}\left|\hat{\pi}(\theta \mid Y^{(n)}) - \pi(\theta \mid Y^{(n)})\right| \mathrm{d}\,\theta. \tag{14}$$

The accuracy metric lies in $[0, 1]$ (Faes et al., 2012). The approximation of full data posterior density by $\hat{\pi}$ is poor or excellent if the accuracy metric is close to 0 or 1, respectively. In our experiments, we computed the kernel density estimates of $\hat{\pi}$ and π from the posterior samples of θ using R package `KernSmooth` (Wand, 2015) and calculated the integral in (14) using numerical approximation.

4.2 Simulated Data: Finite Mixture of Gaussians

Finite mixture of Gaussians are widely used for model-based classification, clustering, and density estimation (Fraley and Raftery, 2002). Let n, p, and L be the sample size, the dimension of observations, and the number of mixture components. If $\mathbf{y}_i \in \mathbb{R}^p$ is the ith observation ($i = 1, \ldots, n$), then the mixture of L Gaussians assumes that any $\mathbf{y} \in \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ is generated from the density

$$f_{\text{mix}}(\mathbf{y} \mid \theta) = \sum_{l=1}^{L} \pi_l \mathcal{N}_p(\mathbf{y} \mid \boldsymbol{\mu}_l, \Sigma_l), \tag{15}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$ lies in a $(L-1)$-simplex, μ_l and Σ_l ($l = 1, \ldots, L$) are the mean and covariance parameters of a p-variate Gaussian distribution, and $\theta = \{\boldsymbol{\pi}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L, \Sigma_1, \ldots, \Sigma_L\}$. We set $L = 2$ and $p = 2$ and simulated data from (15) using $\boldsymbol{\pi} = (0.3, 0.7)$, $\mu_1 = (1, 2)^T$, $\mu_2 = (7, 8)^T$, and $\Sigma_l = \Sigma$ ($l = 1, 2$), where $\Sigma_{12} = 0.5$, $\Sigma_{11} = 1$, and $\Sigma_{22} = 2$. We performed 10 simulation replications.

Table 1: Accuracies of the approximate posteriors for ρ_1, ρ_2, and $g_{0.05}(x)$ and $g_{0.95}(x)$ for $x \in \mathbb{R}$. The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior

	ρ_1		ρ_2		$g_{0.05}$		$g_{0.95}$	
VB	0.77 (0.31)		0.76 (0.29)		0.99 (0.00)		0.99 (0.00)	
	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$	$k = 5$	$k = 10$
CMC	0.97 (0.01)	0.96 (0.01)	0.96 (0.01)	0.96 (0.01)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
SDP	0.97 (0.01)	0.96 (0.01)	0.95 (0.01)	0.96 (0.01)	-	-	-	-
WASP	0.97 (0.01)	0.95 (0.01)	0.97 (0.01)	0.96 (0.01)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)

This simple example demonstrated the generality of WASP in estimating the posterior distribution of functions of θ as described in Corollary 5. We defined two nonlinear functions of θ as

$$\rho_l = (\Sigma_l)_{12} / \left\{(\Sigma_l)_{11}(\Sigma_l)_{22}\right\}^{1/2} \quad l = 1, 2, \quad g(x) = f_{\text{mix}}\{(x, x)^T\} \quad x \in \mathbb{R}, \tag{16}$$

where ρ_l is the correlation of lth mixture component and $g(x)$ is the value of density f_{mix} in (15) when $\mathbf{y} = (x, x)^T$. Our simulation setup implied that $\rho_1 = \rho_2$ and $g(x)$ was bimodal for $x \in \mathbb{R}$. We completed the hierarchical model in (15) by specifying independent conjugate priors on $\boldsymbol{\pi}$ and (μ_l, Σ_l) $(l = 1, 2)$ as

$$\boldsymbol{\pi} \sim \text{Dirichlet}(1/2, 1/2), \quad \boldsymbol{\mu}_l \mid \Sigma_l \sim \mathcal{N}_2(\mathbf{0}, 100\Sigma_l), \quad \Sigma_l \sim \text{Inverse-Wishart}(2, 4I_2), \tag{17}$$

where 2 is the prior degrees of freedom and $4I_p$ is the scale matrix of the Inverse-Wishart distribution. The posterior samples of θ were obtained using Gibbs sampling (Bishop, 2006), which were used to obtain posterior samples for ρ_1, ρ_2, and g.

We compared WASP with the posterior distributions estimated using CMC, Gibbs sampling, SDP, and VB. We used the VB algorithm developed in Bishop (2006). Two values of $k \in \{5, 10\}$ were used for CMC, SDP, and WASP and full data were partitioned into k subsets such that the mixture proportions were preserved in every subset. The approximate posterior distributions of ρ_1, ρ_2, and $g(x), x \in \mathbb{R}$, under each method were estimated using the subset posterior samples obtained after modifying the original Gibbs sampler. The sampling algorithm for WASP is described in the Supplementary Material.

We compared the accuracy (14) of CMC, SDP, VB, and WASP in approximating the full data posterior distributions of ρ_1, ρ_2, and point-wise 90% credible bands of $g(x)$ for $x \in \mathbb{R}$, denoted as $g_{0.05}(x)$ and $g_{0.95}(x)$. CMC, SDP, and WASP accurately approximated the full data posterior distributions of ρ_1 and ρ_2 for both ks, but VB underestimated the posterior uncertainty in ρ_1 and ρ_2. CMC, VB, and WASP were very accurate in estimating $g_{0.05}(x)$ and $g_{0.95}(x)$ for $x \in \mathbb{R}$, whereas the application of SDP failed due to a numerical error in matrix inversion (Table 1). This provides an empirical verification of Corollary 5, showing that the accuracy of the WASP was unaffected by the form of the parameters in the combination step. Theoretical guarantees similar to Corollary 5 were unavailable for CMC or SDP, but our numerical results illustrated that a similar result might also hold for these methods in mixture models.

4.3 Simulated Data: Linear Mixed Effects Model

Linear mixed effects models are extensively used in extending linear regression to account for longitudinal and nested dependence structures. Let n, s, and s_i be the sample size, total number of observations, and total number of observations for sample i $(i = 1, \ldots, n)$ so that $s = \sum_{i=1}^{n} s_i$. Suppose $X_i \in \mathbb{R}^{s_i \times p}$ and $Z_i \in \mathbb{R}^{s_i \times r}$ include predictors in the fixed and random effects components, respectively. Letting $\mathbf{y}_i \in \mathbb{R}^{s_i}$ be the response for sample i, the linear mixed effects model assumes that

$$\mathbf{y}_i \mid \boldsymbol{\beta}, \mathbf{u}_i, \tau^2 \sim \mathcal{N}_{s_i}(X_i\,\boldsymbol{\beta} + Z_i\,\mathbf{u}_i, \tau^2 I_{n_i}), \quad \mathbf{u}_i \sim \mathcal{N}_r(\mathbf{0}, \Sigma), \quad (i = 1, \ldots, n), \tag{18}$$

where $\mathbf{u}_i \in \mathbb{R}^r$ is the random effect for sample i with mean $\mathbf{0}$ and $r \times r$ covariance Σ, $\boldsymbol{\beta} \in \mathbb{R}^p$ denotes the fixed effects, and τ^2 is the error variance. The model parameters are $\theta = \{\boldsymbol{\beta}, \Sigma, \tau^2\}$.

We simulated data for a fixed n and s and varying p and r. We chose two values of $(p, r) \in \{(4, 3), (80, 6)\}$, fixed n and s to be 6000 and 100,000, and randomly assigned the s observations to n samples. The two choices of (p, r) ensured that the number of unknown parameters in $\boldsymbol{\beta}$ and Σ was 10 and 100 in the former and latter cases. The entries of X_i and Z_i were set to 1 or -1 with equal probability for every i. We fixed $\boldsymbol{\beta}$ entries as -2 and 2 alternately and $\tau^2 = 1$. The random effects covariance matrix $\Sigma = \text{diag}(\sqrt{1}, \ldots, \sqrt{r}) R \,\text{diag}(\sqrt{1}, \ldots, \sqrt{r})$, where diag($\mathbf{a}$) is a diagonal matrix with $\mathbf{a}$ along the diagonal and R is a correlation matrix with 1 along the diagonal. We set $R = R_1$ if $r = 3$ and $R = \text{bdiag}(R_1, R_1)$ if $r = 6$, where bdiag(A, B) is a block-diagonal matrix with A, B along the diagonal, $(R_1)_{ii} = 1$ $(i = 1, 2, 3)$, $R_{12} = -0.40$, $R_{13} = 0.30$, and $R_{23} = 0.001$. The matrix R_1 included negative, positive, and small to moderate strength correlations (Kim et al., 2013). We used this setup to simulate data from (18) and performed 10 replications.

We used the HMC algorithm in Stan for sampling from the full data and subset posterior distributions. The full data posterior computations were feasible for the chosen values of n and s and posterior samples were obtained after completing the hierarchical model in (18) by using the default weakly informative priors for $\boldsymbol{\beta}$, Σ, and τ^2 in Stan. Two values of $k \in \{10, 20\}$ were used for CMC, SDP, and WASP, and the n samples were randomly partitioned into k subsets. The sampling algorithms for subset posterior distributions for the three methods were implemented in Stan and posterior samples of θ were obtained in parallel across k subsets. This was followed by a combination step to estimate the approximate posterior distributions for the three methods. The sampling algorithm for WASP is described in the Supplementary Material. Stochastic gradient Langevin dynamics (SGLD; Welling and Teh 2011) has proven to be a successful stochastic version of MCMC in mixture and regression models but has not been extensively tested on linear mixed effects models in which multiple observations are available on a subject. We compared Stan's HMC and SGLD with batch sizes 2000, 4000, step sizes $10^{-4}, 10^{-5}$ and 10^4 iterations.

We compared the accuracy (14) of CMC, SDP, SGLD, VB, and WASP in approximating the marginal posterior distributions of fixed effects, variances and covariances of random effects, and the joint posterior distributions of three pairs of covariances of random effects. We used the streamlined algorithm (SA; Lee and Wand 2016) and automatic differentiation variational inference in Stan (ADVI; Kucukelbir et al. 2015) for estimating the VB posteriors for $\boldsymbol{\beta}$ and Σ . All methods except SGLD were significantly faster than the full data posterior

Table 2: Accuracies of the approximate posteriors for variances in (18). The accuracies are averaged over 10 simulation replications and across all diagonal elements of Σ. Monte Carlo errors are in parenthesis. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$r = 3$		$r = 6$	
ADVI	0.48 (0.31)		0.09 (0.23)	
SA	0.26 (0.19)		0.34 (0.22)	
SGLD (2000)	0.68 (0.08)		0.73 (0.08)	
SGLD (4000)	0.69 (0.09)		0.72 (0.08)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.93 (0.03)	0.91 (0.05)	0.89 (0.05)	0.80 (0.08)
SDP	0.92 (0.06)	0.86 (0.07)	0.84 (0.10)	0.77 (0.14)
WASP	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)

distribution, with SA being the fastest. CMC, SA, SDP, and WASP provided accurate approximations of the marginal posterior distributions of fixed effects and covariances of random effects. Unlike Stan's HMC, SGLD's performance was sensitive to the choices of step size and batch size. SGLD failed to converge for all batch sizes when the step size was 10^{-4}, and its accuracy increased with batch size. The performance of ADVI and SGLD deteriorated quickly as r increased from 3 to 6. The accuracy of CMC and SDP in approximating the marginal posterior distributions of variances of random effects depended on k and r. ADVI and SA provided a poor approximation for the posterior variances of random effects. In all these cases, the accuracy of WASP was stable for every k and r (Tables 2 and 3). All methods except SGLD showed similar accuracies in approximating the true joint posterior distributions of three pairs of covariances of random effects. The differences in accuracies of CMC, SA, SDP, and WASP for different values of k and r were due to the differences in numerical approximation of (14) (Tables 4 and 5 and Figures 3 and 4); see Table 1 in the Supplementary Material.

The accuracy of CMC, SDP, and WASP decreased when k increased from 10 to 20 because subset posterior distributions conditioned on a smaller fraction of the data. This provided an empirical verification of Theorem 4 for the WASP. Our numerical results illustrated that a similar result might also hold for CMC and SDP. The stable performance of WASP compared to that of CMC and SDP in the approximation of the posterior distributions of variances of random effects showed that the validity of the normal approximation for subset posterior distributions was crucial in obtaining accurate approximations of full data posterior using CMC and SDP. On the other hand, WASP results were free of any such assumptions and were valid for any nonlinear function of μ and Σ; see Corollary 5.

4.4 Simulated Data: Probablistic Parafac Model

We use probabilistic parafac model as a representative example for nonparametric density estimation using WASP. Probabilistic parafac is an approach for nonparametric Bayes modeling of joint dependence in multivariate categorical data (Dunson and Xing, 2009). Let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})$ be the data from sample i, where x_{ij} has d_j possible categorical values in $\{1, \ldots, d_j\}$ $(j = 1, \ldots, p)$. The hierarchical model for x_{ij} $(i = 1, \ldots, n;$

Table 3: Accuracies of the approximate posteriors for covariances in (18). The accuracies are averaged over 10 simulation replications and across all off-diagonal elements of Σ. Monte Carlo errors are in parenthesis. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$r = 3$		$r = 6$	
ADVI	0.69 (0.23)		0.49 (0.29)	
SA	0.94 (0.02)		0.94 (0.02)	
SGLD (2000)	0.07 (0.11)		0.13 (0.09)	
SGLD (4000)	0.07 (0.11)		0.12 (0.09)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.94 (0.03)	0.91 (0.05)	0.94 (0.03)	0.92 (0.05)
SDP	0.92 (0.04)	0.89 (0.06)	0.89 (0.07)	0.87 (0.10)
WASP	0.97 (0.01)	0.97 (0.01)	0.97 (0.01)	0.96 (0.01)

Table 4: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects when $r = 3$ in (18). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$		$(\sigma_{12}, \sigma_{23})$		$(\sigma_{13}, \sigma_{32})$	
ADVI	0.53 (0.28)		0.62 (0.14)		0.49 (0.25)	
SA	0.91 (0.01)		0.91 (0.01)		0.92 (0.01)	
SGLD (2000)	0.03 (0.01)		0.01 (0.00)		0.02 (0.01)	
SGLD (4000)	0.03 (0.01)		0.01 (0.00)		0.02 (0.01)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.88 (0.05)	0.79 (0.06)	0.88 (0.04)	0.82 (0.07)	0.91 (0.02)	0.85 (0.04)
SDP	0.90 (0.03)	0.89 (0.03)	0.90 (0.03)	0.87 (0.05)	0.92 (0.02)	0.89 (0.04)
WASP	0.93 (0.01)	0.94 (0.01)	0.93 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)

$j = 1, \ldots, p$) is

$$x_{ij} \mid \left(\psi_{h1}^{(j)}\right)_{h=1}^{\infty}, \ldots, \left(\psi_{hd_j}^{(j)}\right)_{h=1}^{\infty}, z_i \sim \text{Multinomial}(\{1, \ldots, d_j\}, \psi_{z_i 1}^{(j)}, \ldots, \psi_{z_i d_j}^{(j)}),$$
$$z_i \sim \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l)\delta_h \equiv \sum_{h=1}^{\infty} \nu_h \delta_h, \quad V_h \sim \text{Beta}(1, \alpha),$$
$$\psi_h^{(j)} \sim \text{Dirichlet}(a_{j1}, \ldots, a_{jd_j}), \quad \alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \quad (19)$$

where α has prior mean a_α / b_α. The hierarchical model for probabilistic parafac implies that

$$\text{pr}(x_{i1} = c_1, \ldots, x_{ij} = c_j, \ldots, x_{ip} = c_p) = \pi_{c_1, \ldots, c_p} = \sum_{h=1}^{\infty} \nu_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}. \quad (20)$$

The x_{ij}s are sampled independently given the latent class z_i and probability vectors $\boldsymbol{\psi}_h^{(j)}$ ($h = 1, \ldots, \infty$). The latent class for every sample is generated using the stick breaking representation of Dirichlet processes. The Gibbs sampling algorithm developed in Dunson and Xing (2009) is very slow even for moderate sample sizes. This example demonstrates that WASP can easily scale existing sampling algorithms to massive data, even when efficient VB alternatives are unavailable.

Table 5: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects when $r = 6$ in (18). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$		$(\sigma_{12}, \sigma_{23})$		$(\sigma_{13}, \sigma_{32})$	
ADVI	0.06 (0.16)		0.08 (0.22)		0.08 (0.17)	
SA	0.89 (0.02)		0.90 (0.02)		0.91 (0.02)	
SGLD (2000)	0.02 (0.01)		0.01 (0.01)		0.01 (0.01)	
SGLD (4000)	0.02 (0.01)		0.01 (0.01)		0.01 (0.01)	
	$k = 10$	$k = 20$	$k = 10$	$k = 20$	$k = 10$	$k = 20$
CMC	0.88 (0.05)	0.76 (0.10)	0.88 (0.04)	0.78 (0.07)	0.90 (0.04)	0.83 (0.07)
SDP	0.90 (0.03)	0.86 (0.05)	0.90 (0.04)	0.86 (0.04)	0.90 (0.03)	0.87 (0.04)
WASP	0.93 (0.02)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.01)	0.94 (0.02)

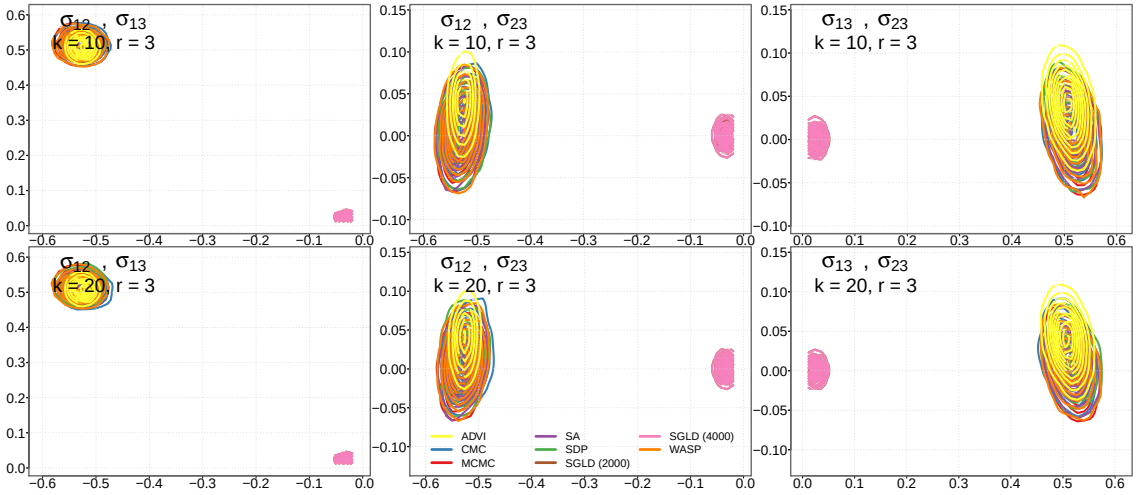

Figure 3: Kernel density estimates of the posterior densities of three covariance pairs when $r = 3$ in (18), where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. ADVI, automatic differentiation variational inference; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; MCMC, Markov chain Monte Carlo; SA, streamlined algorithm; SDP, semiparametric density product; WASP, Wasserstein posterior.

We followed the simulation setup in Dunson and Xing (2009), except with a much larger sample size. We fixed the sample size, number of dimensions, and number of categories in each dimension at $n = 10^5$, $p = 20$, and $d_j = 2$ $(j = 1, \ldots, p)$, respectively. These choices of n, p, and d_js ensured that computations for sampling from the full data posterior were tractable. Data were simulated as a mixture of two populations such that any sample belonged to the two populations with equal probability. The two categories in every dimension excluding $2, 4, 12$, and 14 were simulated from a discrete uniform in both populations. The dependence across dimensions $2, 4, 12$, and 14 was induced as follows. The probabilities π_2, π_4, π_{12}, and π_{14} were set to (0.20, 0.80), (0.25, 0.75), (0.80, 0.20), and (0.75, 0.25) in the first population and to (0.80, 0.20), (0.75, 0.25), (0.20, 0.80), and (0.25, 0.75) in the second population. The simulation setup was replicated 10 times.

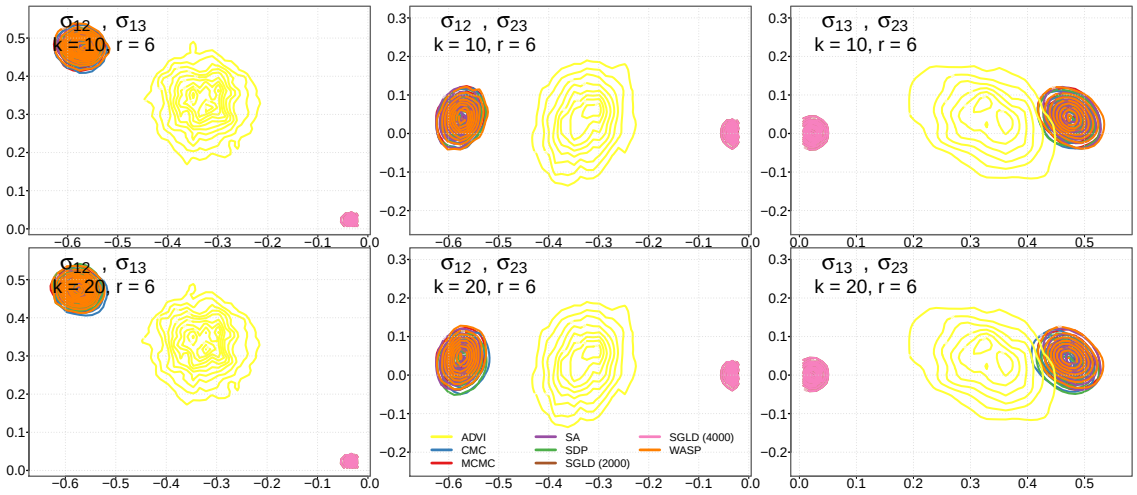

Figure 4: Kernel density estimates of the posterior densities of three covariance pairs when $r = 6$ in (18), where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. ADVI, automatic differentiation variational inference; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; MCMC, Markov chain Monte Carlo; SA, streamlined algorithm; SDP, semiparametric density product; WASP, Wasserstein posterior.

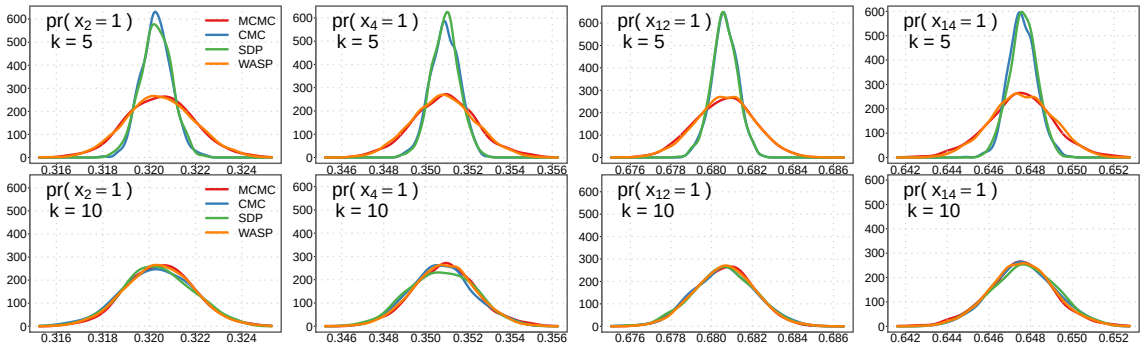

Figure 5: Kernel density estimates of the marginal posterior densities for dimensions 2, 4, 12, and 14. MCMC, Gibbs sampling algorithm of Dunson and Xing (2009); CMC, consensus Monte Carlo; SDP, semiparametric density product; VB, variational Bayes; WASP, Wasserstein posterior

We used CMC, SDP, and WASP to approximate the full data posterior distributions for $\text{pr}(x_i = 1)$, where $i \in \{2, 4, 12, 14\}$. Two values of $k \in \{5, 10\}$ were used for CMC, SDP, and WASP. The full data were randomly partitioned into k subsets and subset posterior samples for WASP were obtained after modifying the Gibbs sampling algorithm in Dunson and Xing (2009) using (5). Examples for the application of CMC and SDP were unavailable for Dirchlet process mixtures, and it was unclear how to raise the prior density to the power $1/k$ when the prior distribution has an atomic form similar to that in (19); therefore, we did not raise the prior to a power of $1/k$ for sampling from the subset posterior distributions

Table 6: Accuracies of the approximate marginal posterior distributions for dimensions 2, 4, 12, and 14 in (19). The accuracies are averaged over 10 simulation replications. Monte Carlo errors are in parenthesis. CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$k = 5$			$k = 10$		
	CMC	SDP	WASP	CMC	SDP	WASP
$\text{pr}(x_2 = 1)$	0.63 (0.02)	0.62 (0.02)	0.97 (0.01)	0.96 (0.02)	0.95 (0.01)	0.97 (0.01)
$\text{pr}(x_4 = 1)$	0.63 (0.02)	0.62 (0.02)	0.97 (0.01)	0.96 (0.01)	0.95 (0.02)	0.97 (0.01)
$\text{pr}(x_{12} = 1)$	0.62 (0.02)	0.62 (0.02)	0.97 (0.01)	0.95 (0.01)	0.96 (0.02)	0.97 (0.01)
$\text{pr}(x_{14} = 1)$	0.64 (0.01)	0.63 (0.01)	0.97 (0.01)	0.96 (0.02)	0.95 (0.02)	0.97 (0.01)

in CMC and SDP. The sampling algorithm for WASP based on stochastic approximation is summarized in the Supplementary Material. Subset posterior samples for $\text{pr}(x_2 = 1)$, $\text{pr}(x_4 = 1)$, $\text{pr}(x_{12} = 1)$, and $\text{pr}(x_{14} = 1)$ were combined to obtain their approximate posterior distributions using CMC, SDP, and WASP.

The accuracy (14) of CMC and SDP in approximating the full data marginal posterior distribution depended on k, with WASP outperforming CMC and SDP when $k = 5$ (Table 6). The approximate and full data posterior distributions were centered at the same value across all dimensions and replications, but the posterior densities for CMC and SDP were highly concentrated compared to the full data posterior density when $k = 5$ (Figure 5). The accuracy of WASP remained stable with varying k, providing an empirical verification of Theorem 4 in cases where our theory is not applicable. The time spent in combining subset posterior samples was negligible compared to the time spent in sampling; therefore, WASP could be used for data with much larger sample size by choosing k large enough such that sampling was efficient across all the data subsets.

4.5 Real Data: MovieLens Ratings Data

We used MovieLens data to illustrate the application of WASP to large-scale ratings data. MovieLens data are one of the largest publicly available ratings data with about 10 million ratings from about 72 thousand users of the MovieLens recommender system. Each observation in the database consists of a user, movie, rating of the movie from 0.5 to 5 in increments of 0.5, and the time of rating. Every movie is also classified into at least one of the 19 genres. We fit a linear mixed effects model (18) using movie- and user-specific information as predictors and the ratings as responses.

We generated three new predictors for accurate modeling of ratings following Perry (2017). First, movie genres were grouped into *movie categories* to reduce the number of genres from 19 to four: *Action* category included Action, Adventure, Fantasy, Horror, Sci-Fi, and Thriller genres; *Children* category included Animation and Children genres; *Comedy* category included Comedy genre; and *Drama* category included Crime, Documentary, Drama, Film-Noir, Musical, Mystery, Romance, War, and Western genres. If a movie belonged to multiple genres, then movie category scores were fractions proportional to the number of genres in the respective categories. Second, *popularity* predictor was defined as $\text{logit}\{(l+0.5)/(n+1.0)\}$, where l and n respectively were the number of users who liked and rated the movie in 30 most recent observations for the movie and $\text{logit}(x) = \log \frac{x}{1-x}$. Third, *previous* predictor was defined to be 1 if the user liked the previous movie and 0 otherwise.

Table 7: Accuracies of the approximate posteriors for variances in (18). The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	σ^2_{Action}	$\sigma^2_{\text{Children - Action}}$	$\sigma^2_{\text{Comedy - Action}}$	$\sigma^2_{\text{Drama - Action}}$	$\sigma^2_{\text{Popularity}}$	$\sigma^2_{\text{Previous}}$
ADVI	0.06 (0.14)	0.33 (0.30)	0.16 (0.23)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
SA	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
SGLD (2000)	0.10 (0.06)	0.06 (0.03)	0.05 (0.05)	0.08 (0.04)	0.10 (0.00)	0.10 (0.07)
SGLD (4000)	0.07 (0.06)	0.06 (0.03)	0.02 (0.06)	0.08 (0.04)	0.10 (0.00)	0.08 (0.07)
CMC	0.28 (0.13)	0.01 (0.01)	0.01 (0.01)	0.14 (0.09)	0.74 (0.10)	0.22 (0.10)
SDP	0.05 (0.03)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.35 (0.10)	0.03 (0.03)
WASP	0.92 (0.04)	0.93 (0.02)	0.87 (0.06)	0.85 (0.08)	0.92 (0.03)	0.93 (0.05)

We used *Action*, *Children − Action*, *Comedy − Action*, *Drama − Action*, *popularity*, and *previous* as the fixed and random effects in (18).

Following the setup in Section 4.3, we compared the performance of WASP with ADVI, CMC, SA, SGLD with batch sizes 2000, 4000, step size 10^{-5} and 10^4 iterations, and SDP using the full data posterior distribution as the benchmark. Sampling using the HMC algorithm in Stan was prohibitively slow for the full data posterior distribution, so we first randomly selected 5000 users and then randomly selected 20 ratings for every user. This resulted in a data set with 100,000 ratings. We randomly split the users into 10 training data sets such that ratings for any user belonged to the same training data set. To compute the approximate posteriors using CMC, SDP, and WASP, we set $k = 10$ and randomly partitioned the users into k subsets such that each subset contained all the ratings for a user. This setup was replicated for every training data.

WASP performed better than its competitors in approximating the full data posterior distributions for variances and covariances of the random effects. Similar to the simulation results in Section 4.3, ADVI, CMC, SA, SDP, and WASP were significantly faster than the full data posterior distribution, with SA being the fastest, and SGLD was the slowest. CMC, SDP, and WASP showed excellent performed in approximating the full data posterior distributions for the fixed effects. WASP outperformed its competitors in approximating the full data posterior distributions for variances, covariances, and pairs of covariances of the random effects (Tables 7, 8, and 9). ADVI, SA, and SGLD significantly under-performed in the estimation of the posterior distribution for the fixed effects and covariance matrix of the random effects. The accuracy of marginals in CMC and SDP depended on the magnitude of covariances, with both methods showing excellent accuracy for covariances with low magnitude. The accuracies of the two-dimensional joint distributions in CMC and SDP were poor because the full data posteriors concentrated at different locations (Figure 6). Except for the poor performance of CMC, SA, and SDP in approximating the posterior distribution of variances and covariances of the random effects, our real data results agreed with our simulation results. We concluded that WASP performed better than its competitors in MovieLens data analysis.

Table 8: Accuracies of the approximate posteriors for covariances in (18). The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. The subscripts $1, \ldots, 6$ are used for predictors *Action*, *Children* − *Action*, *Comedy* − *Action*, *Drama* − *Action*, *popularity*, and *previous*. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	σ_{12}	σ_{13}	σ_{14}	σ_{15}	σ_{16}	σ_{23}	σ_{24}	σ_{25}
ADVI	0.15 (0.30)	0.25 (0.26)	0.14 (0.16)	0.32 (0.12)	0.06 (0.09)	0.00 (0.00)	0.18 (0.20)	0.66 (0.15)
SA	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
SGLD (2000)	0.08 (0.03)	0.19 (0.10)	0.18 (0.08)	0.18 (0.11)	0.23 (0.09)	0.14 (0.00)	0.14 (0.01)	0.14 (0.10)
SGLD (4000)	0.08 (0.02)	0.16 (0.10)	0.14 (0.08)	0.12 (0.08)	0.20 (0.08)	0.14 (0.00)	0.13 (0.01)	0.11 (0.10)
CMC	0.06 (0.03)	0.16 (0.04)	0.18 (0.04)	0.83 (0.07)	0.33 (0.13)	0.01 (0.01)	0.07 (0.02)	0.80 (0.04)
SDP	0.01 (0.01)	0.08 (0.03)	0.07 (0.02)	0.75 (0.06)	0.14 (0.09)	0.00 (0.00)	0.02 (0.01)	0.73 (0.08)
WASP	0.95 (0.02)	0.91 (0.04)	0.91 (0.05)	0.94 (0.03)	0.90 (0.07)	0.89 (0.07)	0.85 (0.08)	0.93 (0.03)
	σ_{26}	σ_{34}	σ_{35}	σ_{36}	σ_{45}	σ_{46}	σ_{56}	
ADVI	0.47 (0.22)	0.50 (0.22)	0.64 (0.11)	0.62 (0.23)	0.64 (0.18)	0.49 (0.29)	0.42 (0.11)	
SA	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
SGLD (2000)	0.11 (0.10)	0.14 (0.10)	0.16 (0.07)	0.10 (0.11)	0.14 (0.12)	0.12 (0.10)	0.15 (0.09)	
SGLD (4000)	0.07 (0.10)	0.11 (0.09)	0.14 (0.07)	0.03 (0.11)	0.10 (0.11)	0.08 (0.10)	0.14 (0.08)	
CMC	0.66 (0.09)	0.65 (0.07)	0.76 (0.08)	0.71 (0.05)	0.82 (0.04)	0.61 (0.11)	0.55 (0.09)	
SDP	0.59 (0.11)	0.62 (0.06)	0.64 (0.09)	0.66 (0.08)	0.66 (0.09)	0.56 (0.14)	0.55 (0.13)	
WASP	0.91 (0.05)	0.94 (0.05)	0.93 (0.03)	0.91 (0.04)	0.93 (0.04)	0.93 (0.04)	0.94 (0.04)	

Table 9: Accuracies of the approximate two-dimensional joint posteriors for the covariances of random effects. The accuracies are averaged over 10 replications. Monte Carlo errors are in parenthesis. The subscripts $1, \ldots, 6$ are used for predictors *Action*, *Children* − *Action*, *Comedy* − *Action*, *Drama* − *Action*, *popularity*, and *previous*. ADVI, automatic differentiation variational inference; SA, streamlined algorithm; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; SDP, semiparametric density product; WASP, Wasserstein posterior

	$(\sigma_{12}, \sigma_{13})$	$(\sigma_{12}, \sigma_{14})$	$(\sigma_{12}, \sigma_{15})$	$(\sigma_{12}, \sigma_{16})$
ADVI	0.03 (0.06)	0.03 (0.07)	0.02 (0.06)	0.05 (0.11)
SA	0.18 (0.04)	0.22 (0.07)	0.31 (0.03)	0.31 (0.02)
SGLD (2000)	0.01 (0.02)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)
SGLD (4000)	0.01 (0.02)	0.01 (0.02)	0.01 (0.01)	0.01 (0.01)
CMC	0.05 (0.02)	0.04 (0.02)	0.06 (0.03)	0.05 (0.02)
SDP	0.05 (0.02)	0.04 (0.02)	0.06 (0.03)	0.05 (0.02)
WASP	0.88 (0.03)	0.88 (0.03)	0.88 (0.02)	0.86 (0.06)

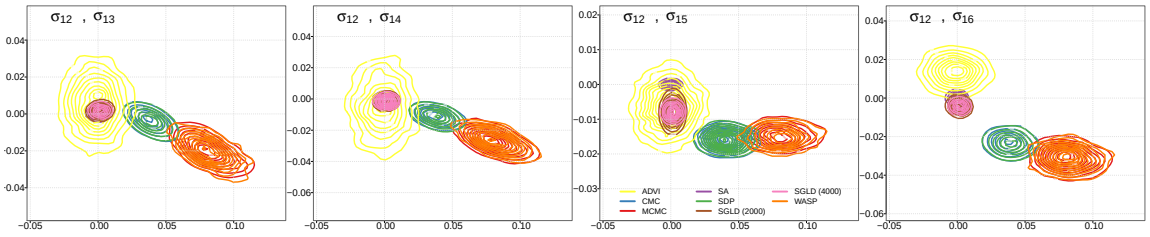

Figure 6: Kernel density estimates of the posterior densities of four covariance pairs, where σ_{ab}, σ_{cd} on every panel represents the two-dimensional posterior density of $(\sigma_{ab}, \sigma_{cd})$. ADVI, automatic differentiation variational inference; SGLD, stochastic gradient Langevin dynamics with batch size in parenthesis; CMC, consensus Monte Carlo; MCMC, Markov chain Monte Carlo; SA, streamlined algorithm; SDP, semiparametric density product; WASP, Wasserstein posterior.

5. Discussion

We have presented WASP as an approach for computationally efficient approximation of the posterior distributions of parameters and their functions when the sample size is large. WASP allows extensions of existing samplers to massive data with minimal modifications and is easily implemented using probabilistic programming languages, such as Stan. Theoretically, we have showed that the rate of convergence of WASP to the Dirac measure centered at the true parameter value in W_2 distance matches the optimal parametric rate up to a logarithmic factor if the number of subsets increases slowly with the size of the full data set. Empirically, we demonstrated that results from WASP and MCMC agree closely in several widely different examples, while WASP enables massive speed-ups in computational time.

We plan to explore several extensions of WASP in the future. First, the combination of subset posterior distributions using WASP and the proof of the convergence rate for the WASP in Theorem 4 are valid even if the data in different subsets are dependent; however, independence assumption within each subset is required in the proof of (12) in Theorem 4 and in our justification of stochastic approximation. Currently, it is unclear how to extend stochastic approximation to cases where the likelihood is unavailable in a product form. This extension in crucial for proper uncertainty quantification outside of settings in which the observations are conditionally independent given latent variables. Second, it is unclear how to optimally choose k in practice; larger k improves computational time when abundant processors are available but choosing k too large may lead to increasing statistical errors (refer to Theorem 4). Our numerical experiments show that the accuracy of WASP is robust to the choice of k if all the subset sizes are moderately large relative to the number of parameters. In addition, it is of interest to study more deeply the impact of the partitioning schemes and attempt to develop approaches that deal with not only large sample sizes but also high-dimensional data. A possibility in this regard is to combine WASP with approximate MCMC (Johndrow et al., 2015).

Acknowledgments

Volkan Cevher and Quoc Tran-Dinh proposed and implemented the linear program for calculating Wasserstein barycenter described in Srivastava et al. (2015). All experiments were based on a modified version of Tran-Dinh's Matlab and Gurobi code for estimating Wasserstein barycenter. Jack Baker provided extensive help in implementing the SGLD algorithm. The code used in the experiments is available at `https://github.com/blayes/WASP`. We thank the Associate Editor and two anonymous referees for their helpful comments that improved our paper. Cheng Li's work was partially supported by National University of Singapore start-up grant R155000172133.

Appendix A. Proofs of Theorems

A.1 Proof of Theorem 1

If $E_{P_{\theta_0}^{(n)}}$ represents the expectation with respect to $P_{\theta_0}^{(n)}$, then

$$E_{P_{\theta_0}^{(n)}}\left[W_2^2\{N_p(\mu,V),N_p(\overline{\mu},\overline{V})\}\right] = E_{P_{\theta_0}^{(n)}}\|\mu-\overline{\mu}\|_2^2 + \operatorname{tr}\left\{V+\overline{V}-2(\overline{V}^{1/2}V\overline{V}^{1/2})^{1/2}\right\}. \tag{21}$$

First, we find the asymptotic order of $E_{P_{\theta_0}^{(n)}}\|\mu_1-\mu_2\|_2^2$ in (21). Define

$$A=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}, \quad B=k^{-1}\left[(X_1^T\Sigma_1^{-1}X_1)^{-1}X_1^T\Sigma_1^{-1},\cdots,(X_k^T\Sigma_k^{-1}X_k)^{-1}X_k^T\Sigma_k^{-1}\right],$$

and $C=A-B$. After some algebra, we have that $AX=I_p$, $BX=I_p$, where I_p is a $p\times p$ identity matrix, and

$$\|\mu-\overline{\mu}\|_2^2=\|Cy\|_2^2, \quad E_{P_{\theta_0}^{(n)}}\|\mu-\overline{\mu}\|_2^2 = E_{P_{\theta_0}^{(n)}}(y^T)C^TCE_{P_{\theta_0}^{(n)}}(y)+\operatorname{tr}(C\Sigma C^T).$$

Since $E_{P_{\theta_0}^{(n)}}(y)=X\theta_0$ and $CX=AX-BX=I_p-I_p=0$, $E_{P_{\theta_0}^{(n)}}\|\mu-\overline{\mu}\|_2^2=\operatorname{tr}(C\Sigma C^T)$. Expanding $C\Sigma C^T$, we get

$$C=(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}-k^{-1}\left[(X_1^T\Sigma_1^{-1}X_1)^{-1}X_1^T\Sigma_1^{-1},\cdots,(X_k^T\Sigma_k^{-1}X_k)^{-1}X_k^T\Sigma_k^{-1}\right],$$

$$C^T=\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}-k^{-1}\begin{bmatrix}\Sigma_1^{-1}X_1(X_1^T\Sigma_1^{-1}X_1)^{-1}\\ \vdots \\ \Sigma_k^{-1}X_k(X_k^T\Sigma_k^{-1}X_k)^{-1}\end{bmatrix},$$

$$\operatorname{tr}(C\Sigma C^T)=\operatorname{tr}\{(X^T\Sigma X)^{-1}\}+k^{-2}\sum_{j=1}^{k}\operatorname{tr}\left\{(X_j^T\Sigma_jX_j)^{-1}\right\}-2\operatorname{tr}(D),$$

where

$$\begin{aligned}D&=k^{-1}\left[(X_1^T\Sigma_1X_1)^{-1}X_1^T,\cdots,(X_k^T\Sigma_kX_k)^{-1}X_k^T\right]\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}\\&=\left\{k^{-1}\sum_{j=1}^{k}(X_j^T\Sigma_j^{-1}X_j)^{-1}X_j^T\Sigma_j^{-1}X_j\right\}(X^T\Sigma^{-1}X)^{-1}=(X^T\Sigma^{-1}X)^{-1}\end{aligned}$$

because Σ is diagonal. We use the above display to obtain that

$$\begin{aligned}E_{P_{\theta_0}^{(n)}}\|\mu-\overline{\mu}\|_2^2=\operatorname{tr}(C\Sigma C^T)&=\frac{1}{k^2}\sum_{j=1}^{k}\operatorname{tr}\left\{(X_j^T\Sigma_j^{-1}X_j)^{-1}\right\}-\operatorname{tr}\left\{(X^T\Sigma^{-1}X)^{-1}\right\},\\&=\frac{1}{km}\operatorname{tr}\left\{\tfrac{1}{k}\sum_{j=1}^{k}\left(\tfrac{1}{m}X_j^T\Sigma_j^{-1}X_j\right)^{-1}\right\}-\frac{1}{n}\operatorname{tr}\left\{\left(\tfrac{1}{n}X^T\Sigma^{-1}X\right)^{-1}\right\}.\end{aligned}$$

Our assumptions and continuity of the matrix inverse for positive definite matrices imply that there are exist positive $a'_n = o(1)$, $b'_m = o(1)$, such that

$$\Omega_0^{-1} - a'_n I_p \prec \left(\tfrac{1}{n} X^T \Sigma^{-1} X\right)^{-1} \prec \Omega_0^{-1} + a'_n I_p,$$
$$\Omega_0^{-1} - b'_m I_p \prec \left(\tfrac{1}{m} X_j^T \Sigma_j^{-1} X_j\right)^{-1} \prec \Omega_0^{-1} + b'_m I_p.$$

This implies that the previous display reduces to

$$E_{P_{\theta_0}^{(n)}} \|\mu - \overline{\mu}\|_2^2 \leq p(b'_m + a'_n)/n = o(n^{-1}), \tag{22}$$

where the equality follow because p is fixed.

We now find the asymptotic order of the traces of the covariance matrices in (21). Following the same arguments used to derive (22), the full data and jth subset posterior covariance matrices satisfy

$$\frac{1}{n}\left(\Omega_0^{-1} - a'_n I_p\right) \prec V = \frac{1}{n}\left(\frac{1}{n} X^T \Sigma^{-1} X\right)^{-1} \prec \frac{1}{n}\left(\Omega_0^{-1} + a'_n I_p\right),$$
$$\frac{1}{n}\left(\Omega_0^{-1} - b'_m I_p\right) \prec V_j = \frac{1}{km}\left(\frac{1}{m} X_j^T \Sigma_j^{-1} X_j\right)^{-1} \prec \frac{1}{n}\left(\Omega_0^{-1} + b'_m I_p\right). \tag{23}$$

Let $M_j = \left\{\overline{V}^{1/2} \frac{1}{km}\left(\frac{1}{m} X_j^T \Sigma_j^{-1} X_j\right)^{-1} \overline{V}^{1/2}\right\}^{1/2}$. Then (23) implies that

$$-b'_m \overline{V} \prec n M_j^2 - \overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2} = n \overline{V}^{1/2}\left(V_j - n^{-1}\Omega_0^{-1}\right)\overline{V}^{1/2} \prec b'_m \overline{V}. \tag{24}$$

From the first inequality in (24), we have

$$\left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} \prec \left(n M_j^2 + b'_m \overline{V}\right)^{1/2} \prec n^{1/2} M_j + b'^{1/2}_m \overline{V}^{1/2}.$$

And similarly the second inequality in (24) implies that

$$n^{1/2} M_j \prec \left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2} + b'_m \overline{V}\right)^{1/2} \prec \left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} + b'^{1/2}_m \overline{V}^{1/2}.$$

Therefore

$$\left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} - b'^{1/2}_m \overline{V}^{1/2} \prec n^{1/2} M_j \prec \left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} + b'^{1/2}_m \overline{V}^{1/2}.$$

Using this relation and the definition of $\overline{V}$, we have that

$$\left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} - b'^{1/2}_m \overline{V}^{1/2} \prec n^{1/2} \overline{V} = \frac{1}{k} \sum_{j=1}^{k} n^{1/2} M_j \prec \left(\overline{V}^{1/2} \Omega_0^{-1} \overline{V}^{1/2}\right)^{1/2} + b'^{1/2}_m \overline{V}^{1/2}. \tag{25}$$

In (25), we take the square of $n^{1/2}\overline{V}$, apply the inequality $(A_1+A_2)^2 \prec 2(A_1^2+A_2^2)$ for two generic positive definite matrices A_1, A_2, and obtain that

$$n\overline{V}^2 \prec 2\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2} + 2b'_m\overline{V},$$
$$n\overline{V}^2 \succ \frac{1}{2}\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2} - b'_m\overline{V}.$$

Multiplying by $\overline{V}^{-1/2}$ on the left and right hand sides yields,

$$n\overline{V} \prec 2\Omega_0^{-1} + 2b'_m I_p,$$
$$n\overline{V} \succ \frac{1}{2}\Omega_0^{-1} - b'_m I_p. \tag{26}$$

Notice that $b'_m = o(1)$, Ω_0 is a constant positive definite matrix, and $\overline{V}$ is a positive definite matrix. Clearly, (26) forces $n\overline{V}$ to be an order-1 matrix. Now we take the square of $n^{1/2}\overline{V}$ in (25) again and obtain that

$$n\overline{V}^2 \prec \overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2} + b'_m\overline{V} + b'^{1/2}_m\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}\overline{V}^{1/2} + b'^{1/2}_m\overline{V}^{1/2}\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2},$$
$$n\overline{V}^2 \succ \overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2} + b'_m\overline{V} - b'^{1/2}_m\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}\overline{V}^{1/2} - b'^{1/2}_m\overline{V}^{1/2}\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}.$$

Multiplying by $\overline{V}^{-1/2}$ on the left and right hand sides yields,

$$n\overline{V} \prec \Omega_0^{-1} + b'_m I_p + b'^{1/2}_m\overline{V}^{-1/2}\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2} + b'^{1/2}_m\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}\overline{V}^{-1/2},$$
$$n\overline{V} \succ \Omega_0^{-1} + b'_m I_p - b'^{1/2}_m\overline{V}^{-1/2}\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2} - b'^{1/2}_m\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}\overline{V}^{-1/2}. \tag{27}$$

Since $n\overline{V}$ is an order-1 matrix, we have that $b'_m\overline{V}^{-1/2}\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2} = o(1)$, $b'_m\left(\overline{V}^{1/2}\Omega_0^{-1}\overline{V}^{1/2}\right)^{1/2}\overline{V}^{-1/2} = o(1)$. Hence (23) and (27) reduce to

$$\frac{1}{n}\{\Omega_0^{-1} - o(1)I_p\} \prec V_j \prec \frac{1}{n}\{\Omega_0^{-1} + o(1)I_p\}, \quad \frac{1}{n}(\Omega_0^{-1} - o(1)I_p) \prec \overline{V} \prec \frac{1}{n}\{\Omega_0^{-1} + o(1)I_p\}.$$

This implies that

$$\mathrm{tr}(\overline{V} - V) = o(n^{-1}), \tag{28}$$

where the last equality follows because p is fixed.

Finally, we find the asymptotic order of the variance term in (21). The display before (28) implies that for some positive $c_n = o(1)$,

$$\begin{aligned}
\overline{V}^{1/2}V\overline{V}^{1/2} &\prec \frac{1}{n^2}\{\Omega_0^{-1/2} + o(1)I_p\}\{\Omega_0^{-1} + o(1)I_p\}\{\Omega_0^{-1/2} + o(1)I_p\} \\
&\prec \frac{1}{n^2}[\Omega_0^{-2} + c_n I_p], \\
\overline{V}^{1/2}V\overline{V}^{1/2} &\succ \frac{1}{n^2}\{\Omega_0^{-1/2} - o(1)I_p\}\{\Omega_0^{-1} - o(1)I_p\}\{\Omega_0^{-1/2} - o(1)I_p\}
\end{aligned}$$

$$\succ \frac{1}{n^2}[\Omega_0^{-2} - c_n I_p].$$

Therefore, $\text{tr}\{(\overline{V}^{1/2} V \overline{V}^{1/2})^{1/2}\} = n^{-1}\,\text{tr}(\Omega_0^{-1}) + o(n^{-1})$ since p is fixed. Using this and (23) for the variance term in (21) gives

$$\begin{aligned}
&\text{tr}\left\{V + \overline{V} - 2\left(\overline{V}^{1/2} V \overline{V}^{1/2}\right)^{1/2}\right\} \\
&= \{n^{-1}\,\text{tr}(\Omega_0^{-1}) + o(n^{-1})\} + \{n^{-1}\,\text{tr}(\Omega_0^{-1}) + o(n^{-1})\} - \{2n^{-1}\,\text{tr}(\Omega_0^{-1}) + 2o(n^{-1})\} \\
&= o\left(n^{-1}\right). \qquad (29)
\end{aligned}$$

Combining the asymptotic expressions for the mean and variance terms in (22) and (29), (21) reduces to

$$E_{P_{\theta_0}^{(n)}}\left[W_2^2\left\{N(\overline{\mu}, \overline{V}), N(\mu, V)\right\}\right] = o\left(n^{-1}\right),$$

which completes the proof. □

A.2 Proof of Theorem 4

Let $\epsilon_m = \left(\frac{m}{\log^2 m}\right)^{-1/(2\alpha)}$. For ease of notation, in all the following proofs, we will sometimes write $p(y_{ji} \mid \theta) \equiv p_{ji}(y_{ji} \mid \theta)$.

Due to the compactness of Θ in (A1), we assume that $\rho(\theta, \theta_0) \leq M_0$ for a large finite constant M_0. We start with a decomposition of the W_2 distance from the jth subset posterior $\Pi_m(\cdot \mid Y_{[j]})$ to the Dirac measure at the true parameter θ_0:

$$\begin{aligned}
&E_{P_{\theta_0}} W_2^2\left(\Pi_m(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot)\right) = E_{P_{\theta_0}} \int_\Theta \rho^2(\theta, \theta_0) \Pi_m(d\theta \mid Y_{[j]}) \\
\leq &E_{P_{\theta_0}} \int_{\{\theta: \rho(\theta,\theta_0) \leq c_1\epsilon_m\}} \rho^2(\theta, \theta_0) \Pi_m(d\theta \mid Y_{[j]}) + E_{P_{\theta_0}} \int_{\{\theta: \rho(\theta,\theta_0) > c_1\epsilon_m\}} \rho^2(\theta, \theta_0) \Pi_m(d\theta \mid Y_{[j]}) \\
\leq &(c_1\epsilon_m)^2 + M_0^2 E_{P_{\theta_0}} \Pi_m\left(\rho(\theta, \theta_0) > c_1\epsilon_m \mid Y_{[j]}\right). \qquad (30)
\end{aligned}$$

We will choose the constant c_1 as $c_1 = \left(\frac{2r_1 g_2}{q_1 C_L}\right)^{1/(2\alpha)}$, where g_1, C_L, q_1, r_1 are the constants in (A1), (A2), and Lemma 5 and Lemma 6 in the Supplementary Material.

The following proofs are similar to the proofs of Theorem 1, 4, and 10 in Ghosal and van der Vaart (2007). The main difference is that our likelihood has been raised to the power γ. Using condition (A2), we can further replace the ρ metric by the pseudo Hellinger distance:

$$\begin{aligned}
&\Pi_m\left(\theta \in \Theta : \rho(\theta, \theta_0) > c_1\epsilon_m \mid Y_{[j]}\right) \\
\leq &\Pi_m\left(\theta \in \Theta : h_{mj}(P_{\theta,j}, P_{\theta_0,j}) > \sqrt{C_L}(c_1\epsilon_m)^\alpha \mid Y_{[j]}\right) \\
= &\int_{\left\{\theta \in \Theta: h_{mj}(\theta,\theta_0) > \sqrt{\frac{2r_1 g_2}{q_1}}\epsilon_m^\alpha\right\}} \frac{\prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)}\right]^\gamma \Pi(d\theta)}{\int_\Theta \prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)}\right]^\gamma \Pi(d\theta)}. \qquad (31)
\end{aligned}$$

For the denominator in (31), by Condition (A4) and Lemma 6, for m sufficiently large, with probability at least $1-\exp(-r_2 m\epsilon_m^{2\alpha})$

$$\int_\Theta \prod_{i=1}^m \frac{p(Y_{ji}|\theta)^\gamma}{p(Y_{ji}|\theta_0)^\gamma}\Pi(d\theta) > \exp(-r_1 n\epsilon_m^{2\alpha}). \tag{32}$$

For the numerator in (31), by Condition (A3) and Lemma 5, we set $\delta = \sqrt{2r_1 g_2/q_1}\epsilon_m^\alpha$ and obtain that with probability at least $1-4\exp\left(-\frac{2r_1 g_2 q_2}{q_1} m\epsilon_m^{2\alpha}\right) \geq 1-4\exp\left(-\frac{2r_1 q_2}{q_1} n\epsilon_m^{2\alpha}\right)$,

$$\sup_{\left\{\theta\in\Theta: h_{mj}(\theta,\theta_0)\geq\sqrt{2r_1 g_2/q_1}\epsilon_m^\alpha\right\}} \prod_{i=1}^m \left[\frac{p(Y_{ji}|\theta)}{p(Y_{ji}|\theta_0)}\right]^\gamma \leq \exp\left(-2r_1 g_2 m\epsilon_m^{2\alpha}\right) \leq \exp\left(-2r_1 n\epsilon_m^{2\alpha}\right) \tag{33}$$

Therefore, based on (31), (32), and (33), we obtain that with probability at least $1-4\exp\left(-\frac{2r_1 q_2}{q_1} n\epsilon_m^{2\alpha}\right) - \exp(-r_2 m\epsilon_m^{2\alpha})$,

$$\Pi_m\left(\theta\in\Theta : \rho(\theta,\theta_0) > c_1\epsilon_m \;\middle|\; Y_{[j]}\right) \leq \exp\left(-2r_1 n\epsilon_m^{2\alpha} + r_1 n\epsilon_m^{2\alpha}\right) \leq \exp\left(-r_1 n\epsilon_m^{2\alpha}\right).$$

Let A_{ϵ_m} be the event $\left\{\theta\in\Theta : \Pi\left(\theta\in\Theta : \rho(\theta,\theta_0) > c_1\epsilon_m \;\middle|\; Y_{[j]}\right) \leq \exp\left(-r_1 n\epsilon_m^{2\alpha}\right)\right\}$. Then we can bound the second term in (30) as

$$\begin{aligned}
&E_{P_{\theta_0}}\Pi_m\left(\rho(\theta,\theta_0) > c_1\epsilon_m \;\middle|\; Y_{[j]}\right)\\
&\leq E_{P_{\theta_0}}\left[I(A_{\epsilon_m})\Pi_m\left(\rho(\theta,\theta_0) > c_1\epsilon_m \;\middle|\; Y_{[j]}\right)\right] + E_{P_{\theta_0}}\left[I(A^c_{\epsilon_m})\Pi_m\left(\rho(\theta,\theta_0) > c_1\epsilon_m \;\middle|\; Y_{[j]}\right)\right]\\
&\leq \exp\left(-r_1 n\epsilon_m^{2\alpha}\right) + P^{(n)}_{\theta_0}(A^c_{\epsilon_m})\cdot 1\\
&\leq \exp\left(-r_1 n\epsilon_m^{2\alpha}\right) + 4\exp\left(-\frac{2r_1 q_2}{q_1} n\epsilon_m^{2\alpha}\right) + \exp(-r_2 m\epsilon_m^{2\alpha})\\
&\leq 6\exp\left(-c_2 m\epsilon_m^{2\alpha}\right),
\end{aligned}$$

for $c_2 = \min(r_1, r_2, 2r_1 q_2/q_1)$, as clearly the second term is dominating the other two given $m \lesssim n$.

Therefore, for (30), since $\epsilon_m = (m/\log^2 m)^{-1/(2\alpha)}$, as $m\to\infty$, an explicit bound will be

$$\begin{aligned}
&E_{P_{\theta_0}} W_2^2\left(\Pi_m(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot)\right) \leq c_1^2\frac{\log^{2/\alpha} m}{m^{1/\alpha}} + 6M_0^2\exp\left(-c_2\log^2 m\right)\\
&\leq c_1^2\frac{\log^{2/\alpha} m}{m^{1/\alpha}} + \frac{1}{m^{1+\frac{1}{\alpha}}} \leq C_1\frac{\log^{2/\alpha} m}{m^{1/\alpha}}
\end{aligned}$$

as m becomes sufficiently large, where the constant C_1 depends on α, c_1, c_2, which further depends on $g_1, g_2, q_1, q_2, r_1, r_2, C_L$. Since q_1, q_2 in Lemma 5 and r_1, r_2 in Lemma 6 depend on $g_1, g_2, D_1, D_2, \kappa, c_\pi$, it follows that C_1 depends on $g_1, g_2, C_L, D_1, D_2, \kappa, c_\pi$. □

Based on Lemma 7 in the Supplementary Material, if the assumption (A5) holds, then we have

$$E_{P^{(n)}_{\theta_0}}\left[W_2^2\left\{\overline{\Pi}_n(\cdot \mid Y^{(n)}), \delta_{\theta_0}(\cdot)\right\}\right] \leq E_{P^{(n)}_{\theta_0}}\left[\frac{1}{k}\sum_{j=1}^k W_2\left\{\Pi_m(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot)\right\}\right]^2$$

$$\leq \frac{1}{k}\sum_{j=1}^{k} E_{P_{\theta_0}^{(n)}} W_2^2\left\{\Pi_m(\cdot \mid Y_{[j]}), \delta_{\theta_0}(\cdot)\right\} \leq C_1 \frac{\log^{2/\alpha} m}{m^{1/\alpha}},$$

where the first inequality follows from Lemma 7 in the Supplementary Material, the second inequality follows from the Cauchy-Schwarz inequality, and the third inequality follows from the subset bound (12). □

Appendix B. Univariate Density Estimation

Let $X_1, \ldots, X_n$ be n copies of a scalar random variable X that follows probability distribution P_0 with density p_0. The full data are randomly split into k subsets and $X_{j1}, \ldots, X_{jm}$ represent the data on subset j ($j = 1, \ldots, k$). The hierarchical model for density estimation using the stick-breaking representation of Dirichlet process mixtures is

$$X_{ji} \mid z_{ji}, \{\mu_h\}_{h=1}^{\infty}, \{\sigma_h^2\}_{h=1}^{\infty} \sim \mathcal{N}(\mu_{z_{ji}}, \sigma_{z_{ji}}^2), \quad z_{ji} \sim \sum_{h=1}^{\infty} \nu_h \delta_h, \quad \nu_h = V_h \prod_{l<h}(1 - V_l), \quad V_h \mid \alpha \sim \text{Beta}(1, \alpha),$$
$$\alpha \sim \text{Gamma}(a_\alpha, b_\alpha), \quad \mu_h \mid \sigma_h^2 \sim \mathcal{N}(0, \sigma_h^2), \quad \sigma_h^2 \sim \text{Inverse-Gamma}(a_\sigma, b_\sigma), \tag{34}$$

where $a_\sigma > 2$ and Beta, Gamma, and Inverse-Gamma random variables have means $\frac{1}{1+\alpha}$, $\frac{a_\alpha}{b_\alpha}$, and $\frac{b_\sigma}{a_\sigma - 1}$ and variances $\frac{\alpha}{(1+\alpha)^2(2+\alpha)}$, $\frac{a_\alpha}{b_\alpha^2}$, and $\frac{b_\sigma^2}{(a_\sigma-1)^2(a_\sigma-2)}$. If l^* is the maximum number of atoms in the stick-breaking representation, then the prior density π is in the form a discrete mixture. We cannot use existing sampling algorithms directly if π is raised to a power of $1/k$, so it is unclear how to sample from the subset posterior density of competing approaches in Section 2.2.

We show that it is still possible to sample from the subset posterior density in (5) using data augmentation. Let L_j be the likelihood given $X_{j1}, \ldots, X_{jm}$ and latent variables $z_{j1}, \ldots, z_{jm}$ in (34), then

$$L_j(\{\mu_h\}_{h=1}^{l^*}, \{\sigma_h^2\}_{h=1}^{l^*}, \{\nu_h\}_{h=1}^{l^*}) = \prod_{h=1}^{l^*} (2\pi\sigma_h^2)^{-\frac{\sharp h_j}{2}} e^{-\frac{1}{2\sigma_h^2}\sum_{i=1}^{m} 1(z_{ji}=h)(x_{ji}-\mu_h)^2} \nu_h^{\sharp h_j}, \tag{35}$$

where $1(z_{ji} = h)$ is 1 if $z_{ji} = h$ and 0 otherwise and $\sharp h_j = \sum_{i=1}^{m} 1(z_{ji} = h)$. For stochastic approximation, we raise L_j in (35) to the power γ and obtain

$$L_j^{\gamma}(\{\mu_h\}_{h=1}^{l^*}, \{\sigma_h^2\}_{h=1}^{l^*}, \{\nu_h\}_{h=1}^{l^*}) = \prod_{h=1}^{l^*} (2\pi\sigma_h^2)^{-\frac{\gamma\sharp h_j}{2}} e^{-\frac{\gamma}{2\sigma_h^2}\sum_{i=1}^{m} 1(z_{ji}=h)(x_{ji}-\mu_h)^2} \nu_h^{\gamma\sharp h_j}. \tag{36}$$

Standard arguments imply that the analytic form of full conditional densities of parameters are

$$\mu_h \mid \text{rest} \propto e^{-\frac{\gamma\sharp h_j + 1}{2\sigma_h^2}\left(\mu_h^2 - 2\mu_h\gamma\frac{\sum_{i=1}^{m} 1(z_{ji}=h)x_{ji}}{\gamma\sharp h_j + 1}\right)},$$
$$\sigma_h^2 \mid \text{rest} \propto \sigma_h^{2\,-\frac{\gamma\sharp h_j}{2}} e^{-\frac{\gamma}{2\sigma_h^2}\sum_{i=1}^{m} 1(z_{ji}=h)(x_{ji}-\mu_h)^2} \sigma_h^{2\,-\frac{1}{2}} e^{-\frac{\mu_h^2}{2\sigma_h^2}} \sigma_h^{2\,-a_\sigma-1} e^{-\frac{b_\sigma}{\sigma_h^2}},$$
$$V_h \mid \text{rest} \propto V_h^{\gamma\sum_{i=1}^{m} 1(z_i=h)}(1-V_h)^{\gamma\sum_{i=1}^{m} 1(z_{ji}>h)}(1-V_h)^{\alpha-1},$$

$$\alpha \mid \text{rest} \propto \alpha^{a_\alpha - 1} e^{-b_\alpha \alpha} \alpha^{l^*} \prod_{h=1}^{l^*} (1 - V_d)^{\alpha - 1} \tag{37}$$

for $h = 1, \ldots, l^*$. Let

$$m_{jh} = \frac{\gamma \sum_{i=1}^{m} 1(z_{ji} = h) x_{ji}}{\gamma \sharp h_j + 1}, \quad v_{jh} = \frac{\sigma_h^2}{\gamma \sharp h_j + 1}, \tag{38}$$

$$a_{jh} = \frac{\gamma \sharp h_j + 1}{2} + a_\sigma, \quad b_{jh} = \frac{\gamma}{2} \sum_{i=1}^{m} 1(z_{ji} = h) \left(x_{ji} - \mu_h\right)^2 + \frac{\mu_h^2}{2} + b_\sigma \tag{39}$$

for $h = 1, \ldots, l^*$, then all full conditional densities are tractable in terms of standard distributions:

$$\begin{aligned}
&\mu_{jh} \mid \text{rest} \sim N(m_{jh}, v_{jh}), \quad \sigma_{jh}^2 \mid \text{rest} \sim \text{Inverse-Gamma}(a_{jh}, b_{jh}), \\
&V_{jh} \mid \text{rest} \sim \text{Beta}(1 + \gamma \sum_{i=1}^{m} 1(z_{ji} = h), \alpha + \gamma \sum_{i=1}^{m} 1(z_{ji} > h)), \\
&\alpha_{jh} \mid \text{rest} \sim \text{Gamma}(a_\alpha + l^*, b_\alpha - \sum_{h=1}^{l^*} \log(1 - V_{jh})).
\end{aligned} \tag{40}$$

Finally, posterior distribution of the latent variables is

$$z_{ji} \mid \text{rest} \sim \sum_{h=1}^{l^*} p_{jh} \delta_h, \quad p_{jh} = \frac{\nu_{jh} \mathcal{N}(\mu_{jh}, \sigma_{jh}^2)}{\sum_{\tilde{h}=1}^{l^*} \nu_{j\tilde{h}} \mathcal{N}(\mu_{j\tilde{h}}, \sigma_{j\tilde{h}}^2)}, \quad (i = 1, \ldots, m), \tag{41}$$

where $\nu_{jh} = V_{jh} \prod_{l<h} (1 - V_{jl})$ and $\mathcal{N}(m, v)$ is the Gaussian density with mean m and variance v.

Appendix C. Linear Program

$$\begin{aligned}
\underset{\mathbf{a}, T_1, \ldots, T_k}{\text{minimize}} \quad & \sum_{j=1}^{k} \text{trace}(T_j^T D_j) \\
\text{subject to} \quad & \\
& 0 \le a_i \le 1, \quad i = 1, \ldots, g, \\
& 0 \le (T_j)_{uv} \le 1, \quad u = 1, \ldots, g, \quad v = 1, \ldots, s_j, \quad j = 1, \ldots, k, \\
& \mathbf{1}^T \mathbf{a} = 1, \\
& T_j \mathbf{1}_{s_j} = \mathbf{a}, \quad j = 1, \ldots, k, \\
& T_j^T \mathbf{1}_s = \frac{\mathbf{1}_{s_j}}{s_j}, \quad j = 1, \ldots, k.
\end{aligned} \tag{42}$$

This linear program can be solved using a variety of linear programming solvers in `Matlab` or `R`, including the algorithms of Cuturi and Doucet (2014) and Srivastava et al. (2015).

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Sungjin Ahn, Anoop Korattikara, and Max Welling. Bayesian posterior sampling via stochastic gradient Fisher scoring. *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Pierre Alquier, Nial Friel, Richard Everitt, and Aidan Boland. Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels. *Statistics and Computing*, 26(1-2):29–47, 2016.

Pedro C Álvarez-Esteban, E del Barrio, JA Cuesta-Albertos, and C Matrán. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016.

Ethan Anderes, Steffen Borgwardt, and Jacob Miller. Discrete Wasserstein barycenters: optimal transport for discrete data. *Mathematical Methods of Operations Research*, 84 (2):389409, 2016. ISSN 1432-5217. doi: 10.1007/s00186-016-0549-x. URL `http://dx.doi.org/10.1007/s00186-016-0549-x`.

Rémi Bardenet, Arnaud Doucet, and Chris Holmes. On Markov chain Monte Carlo methods for tall data. *Journal of Machine Learning Research*, 18(47):1–43, 2017.

Peter J Bickel and David A Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981.

C. M. Bishop. *Pattern recognition and machine learning*, volume 4. Springer New York, 2006.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C Wilson, and Michael Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1727–1735, 2013.

Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *Proceedings of the 31st International Conference on Machine Learning, JMLR W&CP*, pages 685–693, 2014.

David B Dunson and Chuanhua Xing. Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association*, 104(487):1042–1051, 2009.

Christel Faes, John T Ormerod, and Matt P Wand. Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association*, 106(495):959–971, 2012.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

Andrew Gelman, Aki Vehtari, Pasi Jylänki, Christian Robert, Nicolas Chopin, and John P Cunningham. Expectation propagation as a way of life. *arXiv preprint arXiv:1412.4869*, 2014.

Subhashis Ghosal and Aad van der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Ryan Giordano, Tamara Broderick, and Michael I Jordan. Covariances, robustness, and variational bayes. *arXiv preprint arXiv:1709.02536*, 2017.

Gurobi Optimization Inc. *Gurobi Optimizer Reference Manual Version 6.0.0*, 2014.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, 2013.

Ildar Abdulovič Ibragimov and Rafail Zalmanovich Has' Minskii. *Statistical Estimation: Asymptotic Theory*, volume 16. Springer Science & Business Media, 2013.

James E. Johndrow, Jonathan C. Mattingly, Sayan Mukherjee, and David B. Dunson. Approximations of Markov chains and High-Dimensional Bayesian Inference. *arXiv preprint arXiv:1508.03387v1*, 2015.

Yoonsang Kim, Young-Ku Choi, and Sherry Emery. Logistic regression with multiple random effects: a simulation study of estimation methods and statistical packages. *The American Statistician*, 67(3):171–182, 2013.

Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. In *Proceedings of the 31st International Conference on Machine Learning*, page 181189, 2014.

Alp Kucukelbir, Rajesh Ranganath, Andrew Gelman, and David Blei. Automatic variational inference in Stan. In *Advances in Neural Information Processing Systems*, pages 568–576, 2015.

Shiwei Lan, Bo Zhou, and Babak Shahbaba. Spherical Hamiltonian Monte Carlo for constrained target distributions. In *JMLR workshop and conference proceedings*, volume 32, page 629. NIH Public Access, 2014.

Cathy Yuen Yi Lee and Matt P. Wand. Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal*, 58(4):868–895, 2016. ISSN 1521-4036. doi: 10.1002/bimj.201500007. URL `http://dx.doi.org/10.1002/bimj.201500007`.

Cheng Li, Sanvesh Srivastava, and David B Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104:665–680, 2017.

Dougal Maclaurin and Ryan Prescott Adams. Firefly Monte Carlo: Exact MCMC with Subsets of Data. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust Bayesian inference via the median posterior. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1656–1664, 2014.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable Bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18(1):4488–4527, 2017.

Alexey Miroshnikov and Erin Conlon. *parallelMCMCcombine: Methods for combining independent subset Markov chain Monte Carlo posterior samples to estimate a posterior density given the full data set*, 2014. URL `https://CRAN.R-project.org/package=parallelMCMCcombine`. R package version 1.0.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. In *Proceedings of the 30th International Conference on Uncertainty in Artificial Intelligence*, pages 623–632, 2014.

Patrick O Perry. Fast moment-based estimation for hierarchical models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):267–291, 2017.

Maxim Rabinovich, Elaine Angelino, and Michael I Jordan. Variational consensus Monte Carlo. In *Advances in Neural Information Processing Systems*, pages 1207–1215, 2015.

Rajesh Ranganath, Dustin Tran, and David Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. *MIT Press*, 2006.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1530–1538, 2015.

Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

Steven L. Scott, Alexander W. Blocker, Fernando V. Bonassi, Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. Bayes and big data: the consensus Monte Carlo algorithm. *International Journal of Management Science and Engineering Management*, 11(2):78–88, 2016.

Babak Shahbaba, Shiwei Lan, Wesley O Johnson, and Radford M Neal. Split Hamiltonian Monte Carlo. *Statistics and Computing*, 24(3):339–349, 2014.

Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. WASP: Scalable Bayes via barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 912–920, 2015.

Stan Development Team. Stan: A C++ library for probability and sampling, version 2.5.0, 2014. URL `http://mc-stan.org/`.

Linda SL Tan and David J Nott. Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188, 2013.

Aad W van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.*, 1:1–305, January 2008. doi: 10.1561/2200000001. URL `http://portal.acm.org/citation.cfm?id=1498840.1498841`.

Matt Wand. *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*, 2015. URL `http://CRAN.R-project.org/package=KernSmooth`. R package version 2.23-14.

Xiangyu Wang and David B Dunson. Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiangyu Wang, Fangjian Guo, Katherine A Heller, and David B Dunson. Parallelizing MCMC with random partition trees. In *Advances in Neural Information Processing Systems*, pages 451–459, 2015.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning*, pages 681–688, 2011.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362, 1995.