Robust and Scalable Variational Bayes

Carlos Misael Madrid Padilla CARLOS.MADRID@CIMAT.MX
Department of Statistics and Data Science
Washington University in St. Louis
St. Louis, MO 63130, USA

Shitao Fan SFAN211@UMD.EDU
Department of Mathematics,
University of Maryland
College Park, MD 20742-4015, USA

Lizhen Lin LIZHEN01@UMD.EDU
Department of Mathematics,
University of Maryland
College Park, MD 20742-4015, USA

Abstract

We propose a robust and scalable framework for variational Bayes (VB) that effectively handles outliers and contamination of arbitrary nature in large datasets. Our approach divides the dataset into disjoint subsets, computes the posterior for each subset, and applies VB approximation independently to these posteriors. The resulting variational posteriors with respect to the subsets are then aggregated using the geometric median of probability measures, computed with respect to the Wasserstein distance. This novel aggregation method yields the *Variational Median Posterior* (VM-Posterior) distribution. We rigorously demonstrate that the VM-Posterior preserves contraction properties akin to those of the true posterior, while accounting for approximation errors or the variational gap inherent in VB methods. We also provide provable robustness guarantee of the VM-Posterior. Furthermore, we establish a variational Bernstein–von Mises theorem for both multivariate Gaussian distributions with general covariance structures and the mean-field variational family. To facilitate practical implementation, we adapt existing algorithms for computing the VM-Posterior and evaluate its performance through extensive numerical experiments. The results highlight its robustness and scalability, making it a reliable tool for Bayesian inference in the presence of complex, contaminated datasets.

Keywords: Scalable Bayesian Inference, Robust Variational Bayes, Variational Median Posterior, Variational Bernstein-von Mises Theorem, Contraction Rates

1 Introduction

Bayesian inference is a foundational paradigm in statistics and machine learning, providing a rigorous framework for probabilistic modeling of unknown parameters and making predictions under uncertainty. Despite its theoretical appeal, practical application of Bayesian methods can face challenges due to the reliance on Markov Chain Monte Carlo (MCMC) methods for sampling the posterior distributions for inference, which is often computationally intractable, especially when dealing with large datasets and complicated models.

To address these limitations, variational Bayes (VB) has emerged as a computationally efficient alternative. VB approximates the posterior distribution by optimizing over a simpler family of distributions, effectively transforming the MCMC sampling problem into a tractable optimization problem. This shift significantly reduces computational costs, making VB particularly suitable for large-scale data applications. Theoretical guarantees for VB methods have also advanced recently

(see Wang and Blei (2019), Zhang and Gao (2020), Yang et al. (2020) and Ohn and Lin (2024)). These developments have further cemented VB as a practical and scalable solution that's backed up by theory. VB's computational efficiency and adaptability have driven its adoption across diverse fields such as natural language processing [e.g. Bowman et al. (2015), Li and Liang (2021)], Bayesian deep learning [e.g Nazaret and Blei (2022), Sen et al. (2024)], genomics and bioinformatics [e.g. Raj et al. (2014), Komodromos et al. (2022)], and healthcare analytics [e.g. Zabad et al. (2023), Koh et al. (2024)].

Despite the computational efficiency of VB, handling outliers and contamination remains a significant challenge in the VB framework. An outlier—following Box and Tiao (1968)—is an observation suspected of being partially or entirely irrelevant because it does not conform to the assumed stochastic model. Outliers and contaminations disrupt statistical inference, often introducing bias or compromising the accuracy of results. Traditional approaches to handling outliers, particularly in point estimation (e.g., Huber (2011) and Law (1986)), have achieved considerable success. However, Bayesian methods, including those within the VB framework, frequently rely on strong distributional assumptions or preprocessing techniques to mitigate the effects of outliers. For instance, Giordano et al. (2018), while not explicitly focuses on outliers, highlights how VB methods can misestimate variances and covariances when outliers are present, emphasizing the critical role of data-cleaning decisions. Similarly, Futami et al. (2018) utilizes a robust VB method where outliers are handled using a heavy-tailed distribution to directly account for anomalies in the data, improving the robustness of the inference. Additionally, Jin (2012) employs a heavy-tailed t-distribution in a hierarchical variational framework to solve inverse problems in the presence of outliers without requiring preprocessing steps, demonstrating robustness and convergence. Moreover, Christmas and Everson (2011) introduces a Bayesian autoregression model that uses Student-t distributed noise to manage outliers in time series data, enhancing performance without preprocessing. Some VB methods, such as Li et al. (2022), specifically handle outliers in applications like forward-looking imaging by using a Student-t distribution for non-Gaussian noise, although such approaches often lack theoretical guarantees. The reliance on such assumptions and preprocessing techniques can limit the flexibility of these methods, especially in large-scale applications. Additionally, not all robust VB techniques provide theoretical guarantees, underscoring the need for further advancements in this area.

In addition to the challenges faced by variational Bayes (VB) approaches, handling outliers and data contamination also poses significant difficulties in the broader context of standard Bayesian analysis. Many Bayesian methods address these issues by assuming specific noise distributions or relying on preprocessing steps to filter out anomalies. These strategies, while effective in certain specific settings, often compromise scalability and robustness, particularly when the true posterior distribution is intractable.

An notable exception to these limitations is the M-Posterior method proposed by Minsker et al. (2017). This approach partitions the data into non-overlapping subgroups, computes the posterior distribution for each subgroup independently, and combines the results by taking the median in the space of probability measures, using the Wasserstein distance. The method provides strong theoretical guarantees, ensuring robustness in the presence of contamination.

While the M-posterior method in Minsker et al. (2017) is provably robust, the need for MCMC sampling for each subset posterior can become computationally expensive, particularly for large datasets or high-dimensional problems where sampling from the true posterior is inherently challenging. We address these gaps by proposing a novel variational Bayes (VB) approach that eliminates the need for MCMC sampling for each subset posterior distribution while maintaining robustness and computational efficiency. Specifically, we introduce the Variational Median Posterior (VM-Posterior) method, which integrates the M-Posterior framework by Minsker et al. (2017) with

variational inference. The key innovation lies in combining the computational efficiency of VB with the robustness of the median aggregation step using the Wasserstein distance. The VM-Posterior method operates as follows:

- Data Partitioning and VB-approximation: The dataset is divided into disjoint subsets, and the posterior distribution with respect to each subset is approximated by a variational posterior.
- Robust Aggregation: The subset variational posteriors are combined using the geometric median in the space of probability measures, leveraging the Wasserstein distance to ensure robustness against outliers.

We will show that this approach effectively handles multiple outliers, regardless of their magnitude or nature, making it particularly suitable for large, real-world datasets often contaminated with anomalies. More specifically, our contributions can be summarized as follows:

- **Contraction and robustness Properties**: We develop a novel contraction analysis tailored to the Variational Median Posterior (VM-Posterior). Our approach involves first deriving the contraction rate of the original posterior to θ_0 with rate ϵ_l. Subsequently, we bound the additional error introduced by the variational approximation, i.e., *the variational approximation gap*, by an upper bound of order $l\epsilon_l^2$. Like the M-Posterior in Minsker et al. (2017), the VM-Posterior achieves this contraction rate even in the presence of outliers, demonstrating its robustness.

  This approach is valuable as it establishes the core contraction rate ϵ_l and quantifies the impact of the variational approximation through the variational gap, reflecting the trade-off between computational efficiency and fidelity to the true posterior. Bounding this gap at order $l\epsilon_l^2$ ensures the VM-Posterior maintains robustness while remaining computationally scalable. By carefully selecting the variational family, this method minimizes the gap, balancing accuracy and efficiency.

  Our work extends established frameworks, such as those by Zhang and Gao (2020) and Ohn and Lin (2024), to the VM-Posterior, showing for the first time that it preserves contraction properties while being robust to multiple outliers.

- **BVM (Berstein von-Mise theorem) for the VM-Posterior**:

  We analyze the asymptotic properties of the VM-Posterior for two variational families: mean-field class and Gaussian family with general covariance. In both cases, the VM-Posterior asymptotically follows a normal distribution centered at a robust estimator θ^*, which approximates the true parameter θ_0 with accuracy governed by the variational gap.

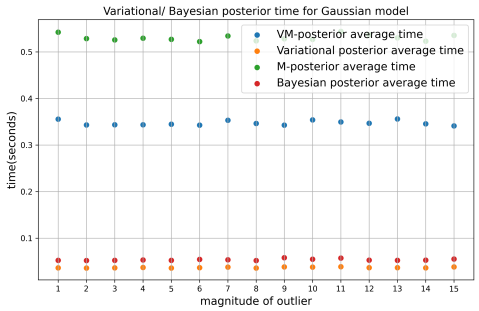
  The key distinction lies in the covariance structure of the limiting distribution: it is diagonal in the mean-field case (independent uncertainty) and full in the Gaussian case (parameter dependencies). In both cases, the covariance matrix is of the form $\frac{I^{-1}(\theta_0)}{n}$, where $I(\theta_0)$ is the Fisher information matrix.

  These results parallel the Bernstein–von Mises theorem, showing that the VM-Posterior retains the asymptotic normality of the true posterior, ensuring valid uncertainty quantification. This provides a solid theoretical foundation for using VM-Posterior Bayesian inference across both independent and dependent parameter models.

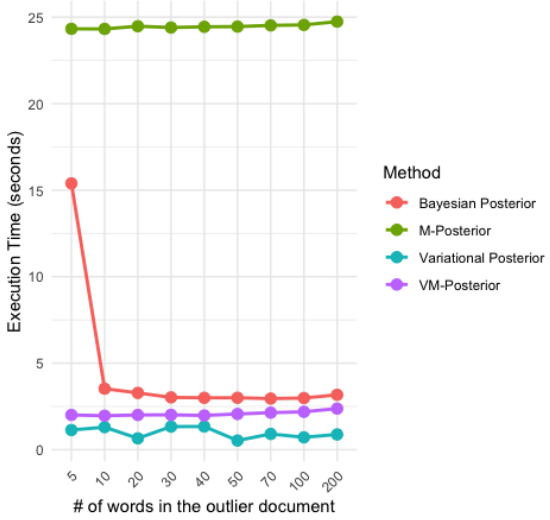
- **Practical Implications**: We conduct comprehensive numerical experiments on both simulated and real-world datasets to validate our theoretical findings. The experiments focus on

comparing the performance of our proposed VM-Posterior against the M-Posterior in terms of computational efficiency and robustness to outliers.

Figures 1a and 1b illustrate the computational advantages of the VM-Posterior, particularly in high-dimensional models and datasets with large outliers. The VM-Posterior demonstrates consistently lower computational times compared to the M-Posterior, which becomes significantly slower as outlier magnitude increases. These results highlight the VM-Posterior's practicality for large-scale Bayesian inference tasks where both robustness and computational efficiency are crucial.



(a) Computational Time Comparison for Gaussian model



(b) Computational time comparison for LDA model

Figure 1: Computational efficiency for VM-Posterior

Our further results in Section 6 also confirm the robustness of the VM-Posterior in terms of posterior coverage. The coverage closely matches expected levels, across various magnitudes of outliers and different significance levels. This robustness is as good as that of the M-Posterior, but the VM-Posterior stands out due to its significantly reduced computational cost. Moreover, variational Bayesian approach are often favored in many machine learning algorithm, where classic Bayesian methods may fail or computationally untractable.

1.1 Outline

The paper is organized as follows: Section 1.2 defines the notation used, while Section 2 introduces the methodology. Section 2.1 explains the model setup and data partition, followed by Section 2.2, which discusses likelihood power adjustment for partitioned posteriors. The variational approach, variational family, and contraction rates are covered in Section 2.3. Section 2.4 introduces the robust aggregation with respect to the Wasserstein distance leading to the VM-Posterior method. The theoretical foundations of our approach are established in Sections 3, with Section 4 exploring the Bernstein-von Mises theorem. Section 5 outlines the algorithms, including Section 5.1 for the variational approach, Section 5.2 for the geometric median, and Section 5.2.1, 5.2.2, and 5.3 for Gaussian models, Gaussian mixtures, and discrete distributions, respectively. Section 6 presents simulation studies, including evaluations on multivariate Gaussian models, Gaussian mixtures, and

latent Dirichlet allocation, as well as real data analysis in Section 6.4. The paper concludes in Section 7, summarizing findings and future research directions.

1.2 Notation

In what follows, $\|\cdot\|_2$ denotes the standard Euclidean distance in $\mathbb{R}^p$ and $\langle\cdot,\cdot\rangle_{\mathbb{R}^p}$ the associated dot product.

Given a totally bounded metric space $(\mathbb{Y}, d)$, the packing number $M(\varepsilon, \mathbb{Y}, d)$ is the maximal number N such that there exist N disjoint d-balls $B_1, \ldots, B_N$ of radius ε contained in $\mathbb{Y}$, i.e., $\bigcup_{j=1}^{N} B_j \subseteq \mathbb{Y}$.

Let $\{p_\theta, \theta \in \Theta\}$ be a family of probability density functions on $\mathbb{R}^p$. Let $l, u : \mathbb{R}^p \mapsto \mathbb{R}_+$ be two functions such that $l(x) \leq u(x)$ for every $x \in \mathbb{R}^p$ and $d^2(l,u) := \int_{\mathbb{R}^p}(\sqrt{u} - \sqrt{l})^2(x)dx < \infty$. A bracket $[l,u]$ consists of all functions $g : \mathbb{R}^p \mapsto \mathbb{R}$ such that $l(x) \leq g(x) \leq u(x)$ for all $x \in \mathbb{R}^p$. For $A \subseteq \Theta$, the bracketing number $N_{[]}(\varepsilon, A, d)$ is defined as the smallest number N such that there exist N brackets $[l_i, u_i], i = 1, \ldots, N$ satisfying $\{p_\theta, \theta \in A\} \subseteq \bigcup_{i=1}^{N} [l_i, u_i]$ and $d(l_i, u_i) \leq \varepsilon$ for all $1 \leq i \leq N$.

For $y \in \mathbb{Y}, \delta_y$ denotes the Dirac measure concentrated at y. In other words, for any Borel-measurable $B, \delta_y(B) = I\{y \in B\}$, where $I\{\cdot\}$ is the indicator function.

We will say that $k : \mathbb{Y} \times \mathbb{Y} \mapsto \mathbb{R}$ is a kernel if it is a symmetric, positive definite function. Assume that $(\mathbb{H}, \langle\cdot,\cdot\rangle_{\mathbb{H}})$ is a reproducing kernel Hilbert space (RKHS) of functions $f : \mathbb{Y} \mapsto \mathbb{R}$. Then k is a reproducing kernel for $\mathbb{H}$ if for any $f \in \mathbb{H}$ and $y \in \mathbb{Y}, \langle f, k(\cdot, y)\rangle_{\mathbb{H}} = f(y)$ (see Aronszajn, 1950 for details).

The Kullback-Leibler (KL) divergence between two probability density functions $p(x)$ and $q(x)$ is defined as:

$$\mathrm{KL}(p\|q) := \int_{\mathbb{R}^p} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx,$$

provided the support of $p(x)$ is contained within the support of $q(x)$. The KL divergence measures the difference between two probability distributions and is frequently used in variational inference to quantify how well the approximate distribution $q(x)$ matches the true posterior $p(x)$.

We use the notation P_0^l to denote both the expected value with respect to the random variables $(X_1, \ldots, X_l)$ that are i.i.d. under θ_0, and the probability measure associated with these random variables. In particular, $P_0^l(g) = \mathbb{E}_{P_0^l}[g(X_1, \ldots, X_l)]$ for any measurable function g, and for any measurable set $B \subseteq \mathbb{R}^{pl}$, $P_0^l(B)$ represents the probability of B under this measure.

2 The VM-Posterior

In this section, we present the VM-Posterior approach, a robust method for combining variational posteriors. This approach leverages either the *Wasserstein geometric median* or the *Wasserstein metric median* to construct a final posterior measure. These methods ensure robustness against outliers while maintaining computational efficiency, making the VM-Posterior a practical and reliable solution for large-scale Bayesian inference.

2.1 Model Setup and Data Partitioning

Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions over $\mathbb{R}^D$, where $\Theta \subset \mathbb{R}^D$ is the parameter space. For any $\theta \in \Theta$, P_θ is absolutely continuous with respect to the Lebesgue measure dx on $\mathbb{R}^D$, with $dP_\theta(\cdot) = p_\theta(\cdot)dx$. The space Θ is equipped with the *Hellinger metric*, defined as

$$\rho(\theta_1, \theta_2) := h(P_{\theta_1}, P_{\theta_2}), \tag{1}$$

where $h(\cdot,\cdot)$ denotes the Hellinger distance between probability distributions. We assume that (Θ,ρ) is a separable metric space.

In Bayesian inference, a prior distribution Π over Θ is specified, where Θ is equipped with the Borel σ-algebra induced by ρ. The prior Π encodes initial beliefs about the unknown true parameter θ_0 before any data is observed, and these beliefs are subsequently updated with the data to form a posterior distribution. Let $X_1,\ldots,X_n$ be i.i.d. $\mathbb{R}^D$-valued random vectors defined on a probability space $(\Omega,\mathcal{B},P)$, with unknown true distribution $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$. Given observed data $\mathbf{X}_n = \{X_1,\ldots,X_n\}$, the posterior distribution on Θ is defined as

$$\Pi_n(B \mid \mathbf{X}_n) := \frac{\int_B \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}$$

for all Borel measurable sets $B \subseteq \Theta$. Under general conditions, the posterior distribution Π_n is known to contract around the true parameter θ_0 as $n \to \infty$ with *contraction rates* ϵ_n (see Ghosal et al. (2000)), if for a suitable sequence $\epsilon_n \to 0$, the following holds:

$$\Pi_n\left(\theta \in \Theta : \rho\left(\theta,\theta_0\right) \geq M\varepsilon_n \mid \mathbf{X}_n\right) \to 0,$$

almost surely or in probability as n grows for an arbitrary $M \to \infty$, indicating that the posterior becomes increasingly concentrated around θ_0. However, the contraction property may be compromised in practical scenarios where the dataset $\mathbf{X}_n$ contains outliers of arbitrary nature and magnitude. For instance, even a single outlier may cause the posterior to deviate substantially from θ_0, disrupting the concentration behavior outlined above. Moreover, in scenarios where $\mathbf{X}_n$ is large, computing the full posterior Π_n is often computationally prohibitive due to memory or processing constraints, raising scalability issues.

To address both scalability and robustness, we propose to partition the sample into m disjoint groups $G_1,\ldots,G_m$, with each group containing at least $\lfloor n/m \rfloor$ observations. This is,

$$\mathbf{X}_n = \bigcup_{j=1}^m G_j, \quad G_i \cap G_l = \emptyset \text{ for } i \neq j, \quad |G_j| \geq \lfloor n/m \rfloor, \quad j = 1,\ldots,m, \quad 1 \leq m \leq \frac{n}{2}.$$

Our idea is to obtain a subset posterior with respect to the above subset data (after likelihood power adjustment described in the next subsection), provide an variational approximation to teach, which are then aggregated to obtain the final VM-posterior. This disjoint grouping of the dataset brings several significant advantages. First, it substantially improves computational efficiency by allowing posterior approximations to be computed in parallel across subsets, effectively addressing challenges posed by centralized processing (see, e.g., Wang et al. (2014), Wang et al. (2015)). Additionally, processing subsets independently reduces memory demands, which is crucial for large datasets requiring inference on manageable data chunks. This parallel strategy is foundational in modern large-scale Bayesian inference frameworks, enabling scalable computation, see for instance Srivastava et al. (2018) and Peruzzi and Dunson (2022).

2.2 Likelihood Power Adjustment

Partitioning offers important benefits for handling outliers and scaling large data. However, it can also inflate uncertainty around θ_0. This issue becomes more pronounced as the number of groups, m, increases. Specifically, consider the situation where $\theta \in \mathbb{R}$, and the Bernstein-von Mises theorem, see Van der Vaart (2000), holds. Under these conditions, each subset-based posterior, $\Pi_{|G_j|}(\cdot \mid G_j)$, defined as

$$\Pi_{|G_j|}\left(B \mid G_j\right) := \frac{\int_B \prod_{i\in G_j} p_\theta\left(X_i\right)\, d\Pi(\theta)}{\int_\Theta \prod_{i\in G_j} p_\theta\left(X_i\right)\, d\Pi(\theta)},$$

will be approximately normal, with an asymptotic covariance of $\frac{m}{n}I^{-1}(\theta_0)$, where $I(\theta_0)$ denotes the Fisher information. This covariance is larger than that of the posterior distribution based on the entire sample, which would asymptotically be $\frac{1}{n}I^{-1}(\theta_0)$. Consequently, the subset-based posteriors $\Pi_{|G_j|}$ may produce an artificially inflated uncertainty estimate.

To mitigate this issue, following Minsker et al. (2014), we apply a likelihood power adjustment to each subset-based posterior. Concretely, we modify each $\Pi_{|G_j|}$ by raising its likelihood to the power m. This adjusted posterior, named *stochastic approximation* and denoted as $\Pi_{n,m}^{|G_j|}(B) = \Pi_{n,m}(B \mid G_j)$, is given by

$$\Pi_{n,m}^{|G_j|}(B) := \frac{\int_B \left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\int_\Theta \left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}.$$

The raised likelihood $\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m$ can be interpreted as replicating each observation in G_j a total of m times. Applying the adjustment reduces inflated variance in subset-based posteriors, making each more representative of the full dataset. Each adjusted posterior $\Pi_{n,m}^{|G_j|}(\cdot \mid G_j)$ thus approximates the full posterior more closely than its unadjusted counterpart, as it behaves as if each data point in G_j is observed multiple times, creating a more accurate reflection of the information from the entire sample.

It turns out that the likelihood adjustment is essential for achieving realistic uncertainty across subsets. It aligns the subset-based posteriors with the full dataset's information content. By this calibration, we achieve a realistic assessment of parameter uncertainty. Notably, as highlighted by Srivastava et al. (2018), an advantage of the stochastic approximation approach is that it allows for the use of off-the-shelf sampling algorithms without additional computational load from actual data replication. In practice, the likelihood adjustment is often implemented by simply raising the likelihood to a power in full-data samplers. Combined with robust aggregation techniques, see Section 2.4, the adjustment enhances posterior predictive credible regions and posterior coverage. The process yields regions that better reflect the true underlying parameter distributions as supported by our numerical results, see Section 5.

2.3 Variational Posterior Computation and Contraction Rates

A key component in our approach is the use of variational inference to efficiently approximate $\Pi_{n,m}^{|G_j|}(\cdot \mid G_j)$, $j = 1, \ldots, m$, the subset-based posterior distributions after the likelihood power. This approximation is achieved by minimizing the Kullback-Leibler (KL) divergence between a simpler, variational candidate posterior Q and the target subset-based adjusted posterior $\Pi_{n,m}(\cdot \mid G_j)$,

$$\widehat{Q}_{n,m}^{(j)} = \arg\min_{Q\in\mathcal{Q}} \mathrm{KL}(Q, \Pi_{n,m}(\cdot \mid G_j)),$$

where $\mathcal{Q}$ denotes the variational family of candidate distributions. In the following, we refer to $\widehat{Q}_{n,m}^{(j)}$ as *subset-based variational posterior*. The choice of $\mathcal{Q}$ plays a critical role, as it determines the trade-off between computational efficiency and accuracy in the approximation. Among commonly used variational families, the mean-field family is particularly popular. This family assumes a fully factorized structure, allowing for computational efficiency. However, it may overlook dependencies between parameters. The mean-field family is defined as

$$\mathcal{Q}_{\mathrm{MF}} = \left\{ \prod_{j=1}^{d} q_j(\theta_j) \right\}.$$

Recent research has extended variational approaches beyond the mean-field by adopting more flexible families. These families are designed to capture complex posterior structures, accommodating dependencies that simpler models may overlook. Examples include Gaussian distributions with general covariance structures and mixtures of such distributions, both of which offer greater flexibility for modeling dependencies and multi-modal characteristics. The Gaussian family with general covariance structures, denoted as $\mathcal{Q}_{\mathrm{GG}}$, is defined by

$$\mathcal{Q}_{\mathrm{GG}} = \left\{ N(\mu, \Sigma) : \mu \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D\times D} \text{ is positive definite} \right\}.$$

The use of this family is supported by the Bernstein–von Mises theorem, which shows that posterior distributions converge to a Gaussian form in large-sample settings (see, for instance, Ray and Szabó (2022)). This theoretical basis underpins the effectiveness of the Gaussian family in approximating posteriors when sample sizes are large. However, for more complex finite-sample settings, Gaussian mixtures with general covariance matrices, such as

$$\mathcal{Q}_{\mathrm{GGM}} = \left\{ \sum_{i=1}^{s} \pi_i N(\mu_i, \Sigma_i) : \pi_i \geq 0, \sum_{i=1}^{s} \pi_i = 1, \mu_i \in \mathbb{R}^D, \Sigma_i \in \mathbb{R}^{D\times D} \right\}.$$

are particularly beneficial. These families effectively capture multi-modal or skewed posteriors, as emphasized by Zobay (2014), who explore the advantages of using mixtures of univariate Gaussians. Additionally, Lin et al. (2022) highlight the expressive capabilities of multivariate Gaussian mixtures, particularly with general covariance matrices, in handling intricate posterior structures.

In practice, selecting an appropriate variational family $\mathcal{Q}$ involves balancing computational efficiency with approximation quality. While the mean-field family $\mathcal{Q}_{\mathrm{MF}}$ offers computational efficiency for large-scale datasets, the Gaussian family $\mathcal{Q}_{\mathrm{GG}}$ provides a middle ground by capturing linear dependencies among parameters. Meanwhile, more complex families, such as Gaussian mixtures $\mathcal{Q}_{\mathrm{GGM}}$, are well-suited for accommodating multi-modal or skewed posteriors, making them advantageous for capturing non-Gaussian characteristics and nuanced dependencies.

Formally, we introduce the following condition on the prior and variational family, which is important in establishing the posterior contraction rates of the variational posterior distribution. This condition distinguishes itself from existing assumptions in the variational Bayes literature due to the power adjustment applied to the likelihood function, as described in Section 2.2. It reduces to a term that captures the KL divergence between the prior and a variational element Q and a term that accounts for the data-generating process.

Assumption 1 *(Prior and variational family) Let $l > 0$. Consider G_j with $l = |G_j|$, a partition of the observed data $\mathbf{X}_n = \{X_1, \ldots, X_n\}$. For each G_j, there exists a distribution $Q_{n,m}^{(j)^*} \in \mathcal{Q}$ such that*

$$\mathrm{KL}\left(Q_{n,m}^{(j)^*}, \Pi\right) + m Q_{n,m}^{(j)^*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)\right] \leq \mathfrak{c}_5^l l \left(\eta_{l,m} + \zeta_{l,m}\right)^2,$$

for some constant $\mathfrak{c}_5^l > 0$.

The term $\eta_{l,m}$ denotes the approximation error of the model and $\zeta_{l,m}$ the estimation error for model j. Consequently the combined term, or *oracle rate*,

$$\varepsilon_l := \left(\eta_{l,m} + \zeta_{l,m}\right). \tag{2}$$

The following proposition shows that Assumption 1 allows us to control the variational approximation gap $\mathrm{P}_0^{(l)}\left[\mathrm{KL}\left(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}\left(\cdot \mid G_j\right)\right)\right]$ to the original adjusted subset-based posterior $\Pi_{n,m}\left(\cdot \mid G_j\right)$.

Proposition 1 *(Variational approximation gap). Let $l > 0$. Consider G_j with $l = |G_j|$, a partition of the full data $X_1, \ldots, X_n$. For any $j \in \{1, ..., m\}$, we have that*

$$\mathrm{P}_0^{(l)} \left[\mathrm{KL} \left(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m} \left(\cdot \mid G_j \right) \right) \right] \leq \inf_{Q \in \mathcal{Q}} \left\{ \mathrm{KL} \left(Q, \Pi \right) + mQ \left[\mathrm{KL} \left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)} \right) \right] \right\}.$$

Further, suppose that Assumption 1 holds. Then

$$\mathrm{P}_0^{(l)} \left[\mathrm{KL} \left(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m} \left(\cdot \mid G_j \right) \right) \right] \leq \mathfrak{c}_5^l l \varepsilon_l^2 \tag{3}$$

holds for any $j \in \{1, ..., m\}$.

Proposition 1 establishes an upper bound on the variational approximation error between the estimated subset-based variational posterior $\widehat{Q}_{n,m}^{(j)}$ and the true posterior $\Pi_{n,m} \left(\cdot \mid G_j \right)$ under the data generating process $\mathrm{P}_0^{(l)}$, for each data subset. This result is significant as it underscores that, in line with the literature (see, for instance, Zhang and Gao (2020) and Ohn and Lin (2024)), when this gap is of the order $c_5^l l \varepsilon_l^2$, the subset-based variational posterior can achieve the same contraction rate, ϵ_l, as the true posterior, facilitating consistency in posterior estimation via classical change-of-measure arguments. This is shown in Theorem 3 below.

In preparation for formally establishing this key concentration result, we first revisit Theorem 1 in Wong et al. (1995), which plays a foundational role in our approach. For a set $A \subseteq \Theta$, the bracketing number $N_{[]}(u, A, d)$ is associated with the family $\{p_\theta, \theta \in A\}$, and is computed with respect to the distance

$$d(l, u) := \int_{\mathbb{R}^D} (\sqrt{l(x)} - \sqrt{u(x)})^2 \, dx.$$

The *bracketing entropy*, denoted $H_{[]}(u; A)$, is then given by

$$H_{[]}(u; A) := \log N_{[]}(u, A, d).$$

Additionally, we denote the "Hellinger ball" of radius r centered at θ_0 as

$$B(\theta_0, r) := \{\theta \in \Theta : h(P_\theta, P_{\theta_0}) \leq r\},$$

where $h(\cdot, \cdot)$ denotes the Hellinger distance. With these definitions in place, we now state the main result from Wong et al. (1995).

Theorem 2 *For a given constant $\zeta > 0$, there exist constants c_j for $j = 1, \ldots, 4$ such that if*

$$\int_{\zeta^2/2^8}^{\sqrt{2}\zeta} H_{[]}^{1/2} \left(\frac{u}{c_3}; B(\theta_0, \zeta\sqrt{2}) \right) du \leq c_4 \sqrt{l} \zeta^2,$$

then the probability bound

$$\Pr \left(\sup_{\theta : h(P_\theta, P_0) \geq \zeta} \prod_{j=1}^{l} \frac{p_\theta}{p_0}(X_j) \geq e^{-c_1 l \zeta^2} \right) \leq 4 e^{-c_2 l \zeta^2}$$

holds. In particular, these constants may be set as $c_1 = \frac{1}{24}$, $c_2 = \frac{4}{27} \cdot \frac{1}{1926}$, $c_3 = 10$, and $c_4 = \frac{(2/3)^{5/2}}{512}$.

Theorem 2 provides a bound on the probability that the likelihood ratio deviates significantly from an exponentially small value for values of θ lying outside a Hellinger ball centered at the true parameter. Specifically, Theorem 2 shows that the supremum of this likelihood ratio, taken over the set $\{\theta : h(P_\theta, P_0) \geq \zeta\}$, is exponentially small with high probability. This bound, which decays at a rate proportional to $l\zeta^2$, effectively limits the probability of substantial deviations from the true density p_0 when θ is sufficiently far from θ_0.

In common parametric scenarios where $\Theta \subseteq \mathbb{R}^p$, the bracketing entropy $H_{[]}(u; B(\theta_0, r))$ often satisfies the bound $H_{[]}(u; B(\theta_0, r)) \leq C_1 \log(C_2 r/u)$, making it possible to select a minimal value for ζ that satisfies the conditions of Theorem 2 with order $\zeta \approx \sqrt{\frac{1}{l}}$. In particular, it is easy to check via Theorem 2.7.11 of van der Vaart and Wellner (1996), that this is the case when the followings hold:

- **Local Lower Bound**: There exists $r_0 > 0$ such that

$$h(P_\theta, P_{\theta_0}) \geq K_1 \|\theta - \theta_0\|_2$$

  holds whenever $h(P_\theta, P_{\theta_0}) \leq r_0$.

- **Local Lipschitz Condition**: There exists $\alpha > 0$ such that for any $\theta_1, \theta_2 \in B(\theta_0, r_0)$,

$$|p_{\theta_1}(x) - p_{\theta_2}(x)| \leq F(x)\|\theta_1 - \theta_2\|_2^\alpha,$$

  where $\int_{\mathbb{R}^D} F(x)\, dx < \infty$.

In our setup, we utilize the result in Theorem 2 by applying it when $\zeta = \zeta_{l,m}$, with $\zeta = \zeta_{l,m}$ introduced in Assumption 1. Here, $\zeta_{l,m}$ represents the estimation error specific to each subset of the data partition, as previously detailed. By ensuring that $\zeta_{l,m}$ meets a minimum value, determined by the bracketing Hellinger entropy, we confirm that each subset-based posterior $\widehat{Q}_{n,m}^{(j)}$ remains concentrated around the true density p_0, even in partitioned data settings. This requirement is formalized in the following assumption.

Assumption 2 *Consider the partition $\{G_j\}_{j=1}^m$ of the sample $\{X_1, \ldots, X_n\}$ defined in Section 2.1, with $l = |G_j|$. Assume that the conditions of Theorem 2 hold with $\zeta := \zeta_{l,m}$ and subset-specific constants $\mathfrak{c}_1^l, \mathfrak{c}_2^l, \mathfrak{c}_3^l$, and $\mathfrak{c}_4^l$.*

By utilizing the entropy-based framework of Theorem 2 in Assumption 2, we effectively restrict the likelihood of large deviations from p_0, which allows us to arrive at Theorem 3, establishing the subset-based variational posterior contraction property.

Theorem 3 *Let $X_1, \ldots, X_n$ be i.i.d sampled from P_0. Consider $\{G_j\}_{j=1}^m$ the partition defined in Section 2.1. Assume the Oracle rate, defined in Equation (2), satisfies $l\varepsilon_l^2 \geq 1$, where $l = |G_j| = \lfloor \frac{n}{m} \rfloor$. Moreover, suppose Assumption 1 and 2 hold. Then, there exists a sufficiently large positive constant R such that, for any $j \in \{1, ..., m\}$*

$$\mathbb{E}_0\Big(\widehat{Q}_{n,m}^{(j)}\left(\rho(\theta, \theta_0) \geq R\varepsilon_l\right)\Big) \leq \exp(-\frac{\mathfrak{c}_1^l}{2}R^2 l\varepsilon_l^2) + 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2) + \frac{1}{l\varepsilon_l^2} + 3\mathfrak{c}_5^l\frac{1}{m}. \tag{4}$$

Theorem 3 establishes that the subset-based variational posterior $\widehat{Q}_{n,m}^{(j)}$ concentrates around the true parameter θ_0 with high probability. Overall, this result ensures that the subset-based variational posterior retains concentration properties similar to the full posterior.

2.4 Aggregation Step

In this section, we define and discuss two robust aggregation methods for combining subset-based variational posteriors defined Section 2.3: the Wasserstein geometric median (Q^*_{Geo}) and the Wasserstein metric median (Q^*_{Met}). This leads to our VM-posterior that is probably resistant to outliers, leveraging properties of the Wasserstein distance for robust and meaningful posterior aggregation while enabling practical computation.

Definition 4 *Let μ_1 and μ_2 be any Borel probability measures on the parameter space (Θ, ρ), with ρ defined in Equation (1). The Wasserstein distance between μ_1 and μ_2 is defined as*

$$d_{W_{1,\rho}}(\mu_1, \mu_2) = \inf_{\gamma \in \Pi(\mu_1, \mu_2)} \int_{\Theta \times \Theta} \rho(x, y) d\boldsymbol{\gamma}(x, y),$$

where $\Pi(\mu_1, \mu_2)$ is the collection of all joint probability measures on $\Theta \times \Theta$ with μ_1 and μ_1 as marginals. Specifically, for all subsets $U \subset \Theta$, we have $\gamma(U \times \Theta) = \mu_1(U)$ and $\gamma(\Theta \times U) = \mu_2(U)$.

The flexibility of the Wasserstein distance in handling probability distributions with potentially non-overlapping support makes it particularly well-suited for aggregating posteriors derived from distinct data partitions, where subtle yet important differences between subsets may exist.

In the following, to obtain a robust aggregation of the collection of subset-based variational posteriors $\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}$, we define the *Wasserstein geometric median*, $Q^*_{\text{Geo},}$, as follows,

$$Q^*_{\text{Geo}} = \text{med}_{\text{g}}\left(\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}\right) = \arg\min_{Q \in \mathcal{Q}} \sum_{j=1}^{m} d_{W_{1,\rho}}(Q, \widehat{Q}^{(j)}_{n,m}). \tag{5}$$

The Wasserstein geometric median Q^*_{Geo}, inspired by the concept introduced in Small (1990), can be seen as a natural extension of the univariate median to the space of probability measures over the parameter space Θ. In the univariate case, the median minimizes the sum of distances from itself to all other points, providing a central location that is robust to outliers. Similarly, Q^*_{Geo} is the measure that minimizes the sum of Wasserstein distances to all subset-based variational posteriors, resulting in a robust central representative of the distribution of subset-based variational posteriors. This formulation generalizes the robustness properties of the univariate median to higher-dimensional and probabilistic settings, allowing Q^*_{Geo} to to resist skewed or extreme outliers.

Another useful generalization of the univariate median is the Wasserstein metric median, which adapts the broader notion of metric medians studied in Hsu and Sabato (2016) to the context of Wasserstein distances and subset-based variational posteriors. Define B_* to be the $d_{W_{1,\rho}}$-ball of minimal radius such that it is centered at one of $\left\{\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}\right\}$ and contains at least half of these points. Then the *Wasserstein metric median* of $\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}$ is the center of B_*. In other words, let

$$\begin{aligned}\varepsilon_* := \inf\{\varepsilon > 0 : \exists j = j(\varepsilon) \in \{1, \ldots, m\} \text{ and } I(j) \subset \{1, \ldots, m\} \text{ such that} \\ |I(j)| > \frac{m}{2} \text{ and } \forall i \in I(j), d_{W_{1,\rho}}\left(\widehat{Q}^{(i)}_{n,m}, \widehat{Q}^{(j)}_{n,m}\right) \leq 2\varepsilon\Big\}\end{aligned}$$

$j_* := j(\varepsilon_*)$, where ties are broken arbitrarily, and set

$$Q^*_{\text{Met}} = \text{med}_0\left(\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}\right) := \widehat{Q}^{(j_*)}_{n,m}. \tag{6}$$

Note that Q^*_{Geo} and Q^*_{Met} are always probability measures. Indeed, we can demonstrate that there exists $\alpha_1 \geq 0, \ldots, \alpha_m \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ such that $Q^*_{\text{Geo}} = \sum_{j=1}^m \alpha_j \widehat{Q}^{(j)}_{n,m}$, and $Q^*_{\text{Met}} \in \left\{\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}\right\}$ by definition.

The metric median Q^*_{Met}, previously studied in Hsu and Sabato (2016), offers a robust aggregation of the subset-based variational posteriors by centering the posterior estimate within the smallest Wasserstein distance ball that contains at least half of the subset-based variational posteriors. By focusing on the subset-based variational posterior closest to the majority of others, this approach reduces the influence of extreme or outlying subset-based variational posteriors, making it particularly useful in situations where there are substantial variations between data partitions. Unlike the Wasserstein geometric median, which minimizes the overall sum of Wasserstein distances, the Wasserstein metric median is defined by proximity to the subset-based variational posteriors and effectively resists the impact of isolated but extreme variations.

3 Robustness of the VM-Posterior

In this section, we establish theoretical results that demonstrate the robustness of the VM-Posterior approach, introduced in Section 2. This robustness is shown by analyzing both the Wasserstein geometric median and the Wasserstein metric median, defined in Section 2.4, which together provide the foundation for robustly aggregating subset-based variational posteriors in the VM-Posterior construction.

The robustness of the Wasserstein geometric median and Wasserstein metric median can be precisely quantified through their concentration properties around the true Dirac measure, represented by the Dirac measure $\delta_0 = \delta_{\theta_0}$, under potential contamination. Both methods possess the desirable property of transforming a collection of independent, "weakly concentrated" estimators into a single estimator with markedly stronger concentration properties, effectively mitigating the influence of outliers or extreme subset deviations. We state this formally in the following theorems.

Theorem 5 *Consider the disjoint subsets $G_1, \ldots, G_m$ as defined in Section 2.1. Let $\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}$ denote the subset-based variational posteriors defined in Section 2.3. Denote by κ a constant satisfying $0 \leq \kappa < \frac{1}{3}$. Suppose $\epsilon > 0$ is such that, for each j in the range $1 \leq j \leq \lfloor (1-\kappa)m \rfloor + 1$, the inequality*

$$\Pr\left(d_{W_{1,\rho}}(\widehat{Q}^{(j)}_{n,m}, \delta_0) > \epsilon\right) \leq \frac{1}{7} \tag{7}$$

holds. Furthermore, let Q^*_{Geo} denote the Wasserstein geometric median as defined in Equation (5). Then, the following is satisfied,

$$\Pr\left(d_{W_{1,\rho}}(Q^*_{Geo}, \delta_0) > 1.52\epsilon\right) \leq \left[e^{(1-\kappa)\psi\left(\frac{3/7-\kappa}{1-\kappa}, \frac{1}{7}\right)}\right]^{-m}, \tag{8}$$

where the function $\psi(\alpha, q)$ is given by,

$$\psi(\alpha, q) = (1-\alpha)\log\frac{1-\alpha}{1-q} + \alpha\log\frac{\alpha}{q}. \tag{9}$$

Theorem 5 implies that the concentration of the Wasserstein geometric median Q^*_{Geo} of independent estimators around the "true" parameter θ_0 improves geometrically with the number m of such estimators. Additionally, the estimation rate remains preserved up to a constant factor. The strong concentration of Q^*_{Geo} in Equation (8) significantly enhances the weak concentration observed for individual subset-based variational posteriors in Equation (7). This improvement underscores the

robustness of Q^*_{Geo} in aggregating the information across subsets while maintaining a high degree of concentration around the true parameter.

The parameter κ plays a crucial role in maintaining robustness to outliers. Specifically, if the initial sample contains up to $\lfloor \kappa m \rfloor$ outliers (potentially of arbitrary nature), then no more than $\lfloor \kappa m \rfloor$ of the subset-based posteriors $\widehat{Q}^{(j)}_{n,m}$ can be impacted. Despite this, the Wasserstein geometric median remains close to the "true" delta measure δ_0 with high probability. This theorem demonstrates that even with some contamination among subsets, the Wasserstein geometric median retains its concentration properties, providing a measure of "robustness" that tolerates outliers while preserving a strong alignment with the "true" delta measure δ_0.

To further clarify, suppose $\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}$ are consistent estimators of Q_0 based on disjoint samples of size n/m. As $\frac{n}{m} \to \infty$, $\frac{\kappa m}{n} \to 0$, ensuring that the estimator Q^*_{Geo} remains consistent despite some contamination, handling a number of outliers that scales as $o(n)$. This result matches the best-case scenario for outlier robustness without imposing additional restrictions on the distribution or nature of the outliers. Furthermore, because the Wasserstein geometric median resides in the convex hull of the subset-based posteriors, it effectively "downweights" outlier observations, making the approach practical for large-scale settings.

Theorem 6 *Consider the disjoint subsets $G_1, \ldots, G_m$ as defined in Section 2.1. Let $\widehat{Q}^{(1)}_{n,m}, \ldots, \widehat{Q}^{(m)}_{n,m}$ denote the subset-based variational posteriors defined in Section 2.3. Denote by κ a constant satisfying $0 \leq \kappa < \frac{1}{3}$. Suppose $\epsilon > 0$ is such that, for each j in the range $1 \leq j \leq \lfloor (1-\kappa)m \rfloor + 1$, the inequality*

$$\Pr\left(d_{W_{1,\rho}}(\widehat{Q}^{(j)}_{n,m}, \delta_0) > \epsilon\right) \leq \frac{1}{4} \tag{10}$$

holds. Furthermore, let Q^*_{Met} denote the Wasserstein metric median as defined in Equation (6). Then, the following is satisfied,

$$\Pr\left(d_{W_{1,\rho}}(Q^*_{Met}, \delta_0) > 3\epsilon\right) \leq \left[e^{(1-\kappa)\psi\left(\frac{1/2-\kappa}{1-\kappa}, \frac{1}{4}\right)}\right]^{-m}. \tag{11}$$

Theorem 6 provides a robustness guarantee for the Wasserstein metric median Q^*_{Met}, analogous to the guarantee for the Wasserstein geometric median in Theorem 5. Defined by the smallest Wasserstein distance ball containing at least half of the subset-based variational posteriors, the metric median requires only pairwise distance information, making it especially practical in high-dimensional settings.

As with the geometric median, the parameter κ limits the impact of up to $\lfloor \kappa m \rfloor$ outliers, ensuring that Q^*_{Met} remains concentrated around the true delta measure δ_0 with high probability.

Overall, these two theorems illustrate an essential property of the median aggregation step, whereby a collection of weakly concentrated estimators is transformed into a single estimator with markedly stronger concentration around the true parameter θ_0. Both the Wasserstein geometric median and the Wasserstein metric median provide robust methods for aggregating subset-based variational posteriors, even under contaminations of arbitrary nature.

Next, we establish the weak concentration properties of each subset-based variational posterior, a critical component in proving the robustness of the Wasserstein geometric and metric medians as discussed earlier. Specifically, we develop concentration properties that satisfy the weak concentrations established in Equation (7) and Equation (10). This is provided in Theorem 7 which can be viewed as an adaptation of Theorem 3, applied to the Wasserstein distance $d_{W_{1,\rho}}(\widehat{Q}^{(j)}_{n,m}, \delta_0)$ rather than the closely related contraction rate concept of the posterior distribution itself. Here, the Wasserstein distance is specified in Definition 4 and evaluated with respect to the *Hellinger metric* $\rho(\cdot, \cdot)$, which introduced in Equation (1).

Theorem 7 *Let $X_1, \ldots, X_n$ be i.i.d sampled from P_0. Consider $\{G_j\}_{j=1}^m$ the partition defined in Section 2.1. Assume the Oracle rate, defined in Equation (2), satisfies $l\varepsilon_l^2 \geq 1$, where $l = |G_j| = \lfloor \frac{n}{m} \rfloor$. Moreover, suppose Assumption 1 and 2 hold. Then, there exists a sufficiently large positive constant R such that, for any $j \in \{1, ..., m\}$*

$$\begin{aligned} &\Pr\left(d_{W_{1,\rho}}(\widehat{Q}_{n,m}^{(j)}, \delta_0) \geq 2R\varepsilon_l + \exp(-ml\varepsilon_l^2)\right) \\ &\leq \exp(-\frac{\mathfrak{c}_1^l}{2}R^2 l\varepsilon_l^2) + 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2) + \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l\frac{1}{m} + \frac{\mathfrak{c}_5^l}{Rm\varepsilon_l}. \end{aligned} \tag{12}$$

This theorem establishes a probability bound for the Wasserstein distance between each subset-based variational posterior $\widehat{Q}_{n,m}^{(j)}$ and the Dirac measure δ_0, representing concentration around the true parameter θ_0. The bound shows that the Wasserstein distance is constrained by both the oracle rate ε_l and a term that decays exponentially with respect to m and l. This result highlights that subset-based variational posteriors, each operating on a reduced sample size, still achieve concentration properties that are reflective of the full posterior contraction rate, ε_l. Additionally, the probability bound comprises terms that diminish as m increases, supporting the inference that as we partition into more subsets, weak concentration remains achievable.

Theorem 7 directly connects to the robustness properties of the Wasserstein geometric median aggregation, Q_{Geo}^*, presented in Corollary 8.

Corollary 8 *Let $X_1, \ldots, X_n$ be i.i.d sampled from P_0. Assume $0 \leq \kappa < \frac{1}{3}$ and that the data set $\{X_1, \ldots, X_n\}$ contains at most $\lfloor \kappa m \rfloor$ outliers. Consider $\{G_j\}_{j=1}^m$ the partition defined in Section 2.1. Suppose Assumption 1 and 2 hold. Moreover, let the Oracle rate, defined in Equation (2), satisfies $l\varepsilon_l^2 \geq 1$, and for any $1 \leq j \leq \lfloor (1-\kappa)m \rfloor + 1$ we have that*

$$\exp(-\frac{\mathfrak{c}_1^l}{2}R^2 l\varepsilon_l^2) + 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2) + \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l\frac{1}{m} + \frac{\mathfrak{c}_5^l}{Rm\varepsilon_l} \leq \frac{1}{7},$$

where $l = |G_j| = \lfloor \frac{n}{m} \rfloor$, and R is a sufficiently large positive constant. Then,

$$\Pr(d_{W_{1,\rho}}(Q_{Geo}^*, \delta_0) \geq 1.52(2R\varepsilon_l + \exp(-ml\varepsilon_l^2))) \leq \left[e^{(1-\kappa)\psi\left(\frac{3/7-\kappa}{1-\kappa}, \frac{1}{7}\right)}\right]^{-m},$$

where the function ψ is defined in Equation (9).

Corollary 8 is a direct consequence of Theorem 5 and Theorem 7. Note that the weak concentration assumption in Equation (7) is implied by Equation (12). It is easy to see that a similar statement holds for Q_{Met}^* by applying Theorem 6 and Theorem 7.

These results confirm that, given sufficient sample size l and partition count m, the Wasserstein geometric median and Wasserstein metric median of the subset-based variational posteriors will remain concentrated around δ_0 with high probability. The probability bound demonstrates that the aggregated posterior Q_{Geo}^* and Q_{Met}^* are robust to potential contamination and computationally efficient, preserving the contraction rate even as m grows. This provides substantial support for the effectiveness of median-based aggregation in achieving a balance between computational efficiency and robustness in Bayesian inference, especially in large-scale settings. Notably, these results yield an exponential improvement in concentration compared to Theorem 7, underscoring the advantages of the VM-Posterior approach.

4 Asymptotic Normality and Confidence Bounds for VM-Posterior

In this section, we establish the asymptotic properties of the VM-Posterior, particularly its convergence to a normal distribution under the total variation (TV) distance. Additionally, we provide finite-sample confidence bounds for the estimated parameter, showcasing the robustness and reliability of the VM-Posterior in structured settings.

To facilitate this analysis, we focus on the asymptotic behavior of the Wasserstein metric median Q^*_{Met} by first specifying the variational family used in its construction. We consider two distinct families: the mean-field family,

$$\mathcal{Q}_{\text{MF}} = \left\{ \prod_{j=1}^{d} q_j(\theta_j) \right\},$$

and the Gaussian family,

$$\mathcal{Q}_{\text{GG}} = \left\{ N(\mu, \Sigma) : \mu \in \mathbb{R}^D, \Sigma \in \mathbb{R}^{D\times D} \text{ is positive definite} \right\}.$$

This distinction enables us to analyze the Wasserstein metric medians $Q^*_{\text{Met,MF}}$ and $Q^*_{\text{Met,GG}}$, where Q^*_{Met} is defined using either the mean-field or Gaussian family. For more information on these variational family specifications, refer to Section 2.3.

In the parametric setting, where the parameter space $\Theta \subseteq \mathbb{R}^p$, we demonstrate that as the sample size n grows, the VM-Posterior Q^*_{Met} converges to a normal distribution centered at a robust estimator θ^* of the true parameter θ_0. This convergence under the TV distance illustrates the VM-Posterior's behavior in finite-sample conditions and provides insight into its stability and robustness in the parametric case.

Furthermore, we establish a finite-sample bound for the estimator θ^*, confirming that it represents the center of a high-confidence region, with a convergence rate comparable to classical posterior distributions. These results underscore the reliability of the VM-Posterior in producing well-calibrated, robust estimates, even in large-scale settings and in the presence of potential contamination across data subsets.

To set the stage, we begin by analyzing each subset-based variational posterior $\{Q^{(j)}_{n,m}\}_{j=1}^m$ individually. For $\theta \in \Theta$, let

$$I(\theta) := \mathbb{E}_{\theta_0}\left[\frac{\partial}{\partial\theta} \log p_\theta(X) \left(\frac{\partial}{\partial\theta} \log p_\theta(X) \right)^T \right]$$

be the Fisher information matrix, which we assume is well-defined. We say that the family $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean (see Chapter 7 in Van der Vaart (2000) for details) if there exists a function $\dot{\ell}_{\theta_0} : \mathbb{R}^D \to \mathbb{R}^p$ such that

$$\int_{\mathbb{R}^D} \left(\sqrt{p_{\theta_0+h}} - \sqrt{p_{\theta_0}} - \frac{1}{2} h^T \dot{\ell}_{\theta_0} \sqrt{p_{\theta_0}} \right)^2 = o\left(\|h\|_2^2 \right)$$

as $h \to 0$. Generally, we have $\dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta} \log p_\theta(x)$. Using this framework, we define

$$\Delta_{l,\theta_0} := \frac{1}{\sqrt{l}} \sum_{j=1}^{l} I^{-1}(\theta_0) \dot{\ell}_{\theta_0}(X_j),$$

which will be instrumental in analyzing the distributional behavior of each subset-based variational posterior.

We now formalize these properties with the following theorem, which provides a convergence result for each subset-based variational posterior under the Gaussian and mean-field families.

Theorem 9 *Let $X_1, \ldots, X_n$ be i.i.d sampled from P_0. Consider $\{G_j\}_{j=1}^m$ the partition defined in Section 2.1. Denote by $\{Q_{n,m}^{(j),GG}\}_{j=1}^m$ the subset-based variational posteriors from Section 2.3, constructed using the Gaussian variational family $\mathcal{Q}_{GG}$. Then, for any $1 \leq j \leq m$, we have that*

$$\left\| Q_{n,m}^{(j),GG}(\cdot) - \mathcal{N}\left(\cdot; \theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}, \frac{I^{-1}(\theta_0)}{lm}\right)\right\|_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0.$$

The same result holds for $\{Q_{n,m}^{(j),MF}\}_{j=1}^m$ where $\{Q_{n,m}^{(j),MF}\}_{j=1}^m$ is constructed using the mean-field variational family $\mathcal{Q}_{MF}$. In this case, $I'^{-1}(\theta_0)$ replaces $I^{-1}(\theta_0)$, where $I'^{-1}(\theta_0)$ is diagonal and shares the same diagonal entries as $I^{-1}(\theta_0)$.

This theorem follows directly from Corollary 7 in Wang and Blei (2019) for the convergence of $\{Q_{n,m}^{(j),\mathrm{GG}}\}_{j=1}^m$ and Theorem 5 in Wang and Blei (2019) for the corresponding result on $\{Q_{n,m}^{(j),\mathrm{MF}}\}_{j=1}^m$, which lays the foundation for analyzing the aggregated posterior Q_{Met}^*.

To proceed with the analysis of the aggregated posterior Q_{Met}^*, we introduce a uniform integrability assumption for the collections $\{Q_{n,m}^{(j),\mathrm{GG}}\}_{j=1}^m$ and $\{Q_{n,m}^{(j),\mathrm{MF}}\}_{j=1}^m$. This condition ensures bounded second moments for each subset-based variational posterior, a necessary foundation for deriving asymptotic results on Q_{Met}^*.

Assumption 3 *We assume uniform integrability for the collections $\{Q_{n,m}^{(j),GG}\}_{j=1}^m$ and $\{Q_{n,m}^{(j),MF}\}_{j=1}^m$. Specifically, this requires that*

$$\sup_{1\leq j\leq m} \int_{\mathbb{R}^p} \|\theta\|_2^2 \, dQ_{n,m}^{(j),GG}(\theta) < \infty, \quad \text{and} \quad \sup_{1\leq j\leq m} \int_{\mathbb{R}^p} \|\theta\|_2^2 \, dQ_{n,m}^{(j),MF}(\theta) < \infty.$$

With Assumption 3 in place, we are now equipped to establish the asymptotic normality and finite-sample confidence bounds for the Wasserstein metric median Q_{Met}^* of the VM-Posterior.

Theorem 10 *Let $X_1, \ldots, X_n$ be i.i.d. samples from P_0.*

a) For any fixed $m \geq 1$, let $\{G_j\}_{j=1}^m$ be the partition defined in Section 2.1. Suppose Assumption 3 holds. Then, we have

$$\left\| Q_{Met,GG}^* - \mathcal{N}\left(\cdot; \theta_{GG}^*, \frac{1}{n} I^{-1}(\theta_0)\right)\right\|_{\mathrm{TV}} \to 0, \tag{13}$$

in P_{θ_0}-probability as $n \to \infty$, where θ_{GG}^* is the mean of $Q_{Met,GG}^*$. A similar result holds for $Q_{Met,MF}^*$, with mean θ_{MF}^* and $I'^{-1}(\theta_0)$ used in place of $I^{-1}(\theta_0)$, where $I'^{-1}(\theta_0)$ is a diagonal matrix matching the diagonal elements of $I^{-1}(\theta_0)$.

b) Let $\{G_j\}_{j=1}^m$ be the partition defined in Section 2.1. Assume $0 \leq \kappa < \frac{1}{3}$ and that the dataset $\{X_1, \ldots, X_n\}$ contains at most $\lfloor \kappa m \rfloor$ outliers. Suppose Assumptions 1 and 2 hold. Additionally, consider the Oracle rate defined in Equation (2) satisfy $l\varepsilon_l^2 \geq 1$, and $\sqrt{n} \leq m$, where $l = |G_j| = \lfloor \frac{n}{m} \rfloor$ and R is a sufficiently large positive constant.

Then,

$$\Pr\left(\|\theta_{GG}^* - \theta_0\|_2 \geq 3\left(2R\varepsilon_l + \exp(-ml\varepsilon_l^2)\right)\right) \leq \left[e^{(1-\kappa)\psi\left(\frac{1/2-\kappa}{1-\kappa}, \frac{1}{4}\right)}\right]^{-m},$$

with the function ψ is defined in Equation (9). A corresponding result holds for θ_{MF}^*.

Theorem 10 extends classical asymptotic results to the setting of the VM-Posterior, demonstrating that the Wasserstein metric median Q^*_{Met}, constructed from subset-based variational posteriors, converges to a normal distribution centered at a robust estimator θ^* of the true parameter θ_0. This is achieved under both Gaussian and mean-field variational families, where the precision matrices $I^{-1}(\theta_0)$ and $I'^{-1}(\theta_0)$ capture the effective sample size across the data subsets.

Part (a) of the theorem mirrors the classical Bernstein–von Mises (BvM) theorem, which traditionally establishes normality for Bayesian posteriors in large samples. However, Theorem 10 is significant because it applies to the median-aggregated posterior in a variational context, providing a robust, subset-based approach that retains asymptotic normality even in the presence of limited contamination. The results here indicate that Q^*_{Met} achieves a convergence rate analogous to that expected from a full-sample posterior, thus demonstrating its stability and reliability under variational approximations.

Part (b) provides a finite-sample confidence bound, distinguishing this result from traditional asymptotic BvM settings by quantifying the robustness of the VM-Posterior under finite sample sizes and potential contamination among subsets. This added robustness suggests that Q^*_{Met} can serve as a reliable posterior estimate in large-scale settings, where data contamination or high computational demands may challenge conventional Bayesian posteriors. As such, Theorem 10 reinforces the practical appeal of the VM-Posterior for real-world applications requiring scalable, robust Bayesian inference.

5 Computation Algorithms

In this section, we delve into the computation details of the variational approximation algorithm and explain how to compute the VM-posterior. The code for this paper is available at `https://github.com/waterism211/variational-median.git`.

5.1 Algorithm for the Variational Approximation

Variational approximations offer various techniques to approximate complex posterior distributions. Below, we outline two prominent methods that can be applied within the variational framework.

- **Mean-field Family:** The mean-field approximation simplifies computation by assuming that all variables are independent. We employ the widely-used Coordinate Ascent Variational Inference (CAVI) method as presented by Bishop (2006). CAVI iteratively optimizes each variational parameter while holding others fixed, which yields an efficient and tractable solution in many cases.

- **Gaussian Family with general covariance:** A flexible approach is to approximate the posterior distribution with a Gaussian family with a general covariance structure. This method extends beyond the mean-field assumption, allowing for dependencies between variables. Inspired by Mahdisoltani (2021), we implement Stochastic Variational Inference (SVI), where we approximate the posterior using stochastic optimization techniques. SVI is particularly useful in large datasets, as it optimizes variational parameters through mini-batch updates.

5.2 Algorithm to Compute the Geometric Median respect to Wasserstein distance

Following the approach from Judelo (2024), we compute the geometric median of a set of Gaussian distributions by utilizing linear programming in optimal transport. Additionally, we employ Weiszfeld's algorithm for the geometric median among discrete distributions.

5.2.1 Multivariate Gaussian Distributions

In the case where the distributions Q_j, $j = 1, \ldots, m$, are multivariate Gaussians $N(\mu_j, \Sigma_j)$, the geometric median, denoted as Q^*_{Geo}, is computed by an iterative algorithm that updates both the mean and covariance parameters iteratively, as described in Álvarez-Esteban et al. (2016).

Given covariance matrices Σ_j for each Gaussian distribution, the algorithm updates the covariance matrix S_n of the geometric median at each iteration n as follows:

$$S_{n+1} = \left(\sum_{j=1}^{m} \frac{1}{m} \left(S_n^{1/2} \Sigma_j S_n^{1/2} \right)^{1/2} \right) \tag{14}$$

The mean of Q^*_{Geo} is updated by computing the median of the means μ_j, which provides a robust central tendency measure across the distributions. This is described in Algorithm 1 below.

Algorithm 1: Geometry Median of Gaussian Distributions

Input : m Gaussian distributions Q_j with means μ_j and covariances Σ_j, initial covariance S_0, number of iterations N

Output: Geometry median Q^*_{Geo}

Calculate Mean: $\mu \leftarrow \text{Median}(\mu_1, \ldots, \mu_m)$

for *each iteration* n **do**

  Update Covariance: $S_{n+1} \leftarrow \left(\sum_{j=1}^{m} \frac{1}{m} \left(S_n^{1/2} \Sigma_j S_n^{1/2} \right)^{1/2} \right)$;

Q^*_{Geo} will be a Gaussian distribution with mean μ and covariance S_N.

5.2.2 Gaussian Mixture Models

For Gaussian mixture models (GMMs), where each distribution Q_j can be represented as a weighted sum of Gaussian components, we write:

$$Q_j = \sum_{k=1}^{K_j} \pi_j^k Q_j^k$$

where each Q_j^k is a multivariate Gaussian $N(\mu_j^k, \Sigma_j^k)$. The goal is to compute the geometric median Q^*_{Geo} over these mixture distributions. Since the median of GMMs are intractable (not necessarily a GMM), we will find an approximation of it. The solution is given by $Q^*_{\text{Geo}} = B\#\gamma^*$, where $B : (\mathbb{R}^d)^m \to \mathbb{R}^d$ maps $(x_1, \ldots, x_m)$ to $\sum \lambda_m x_j$. Here, γ^* represents an optimal multi-marginal transport plan.

The multi-marginal optimal transport problem for the geometric median is given by:

$$Q^*_{\text{Geo}}(Q_1, \ldots, Q_m) := \inf_{\gamma \in \Pi(Q_1, \ldots, Q_m) \cap GMM_{md}(\infty)} \int_{\mathbb{R}^{dm}} c(x_1, \ldots, x_m)\, d\gamma(x_1, \ldots, x_m)$$

where the cost function $c(x_0, \ldots, x_{m-1})$ is defined as:

$$c(x_1, \ldots, x_m) = \sum_{i=1}^{m} \frac{1}{m} \|x_i - B(x)\|^2 = \frac{1}{2m^2} \sum_{i,j=0}^{m-1} \|x_i - x_j\|^2$$

and $\Pi(Q_1, \ldots, Q_m)$ denotes the set of probability measures on $(\mathbb{R}^d)^m$ with $Q_1, Q_2, \ldots, Q_m$ as marginals, and $GMM_{md}(\infty)$ refers to the Gaussian mixture class.

To discretize the multi-marginal problem: The optimal solution γ^* can be expressed as a sum of the optimal transport plans between the Gaussian components of the GMMs:

$$\gamma^* = \sum_{\substack{1\leq k_1\leq K_1 \\ 1\leq k_m\leq K_m}} w^*_{k_1 k_2 \ldots k_m} \gamma^*_{k_1 k_2 \ldots k_m},$$

where $\gamma^*_{k_1 k_2 \ldots k_m}$ is the optimal multi-marginal plan between the Gaussian measures $Q_1^{k_0}, \ldots, Q_m^{k_m}$. Furthermore, w^* is the solution of the discrete optimization problem:

$$\min_{w \in \Pi(\pi_1,\ldots,\pi_m)} \sum_{k_1,\ldots,k_m}^{K_1,\ldots,K_m} w_{k_1\ldots k_m} Q^*_{\text{Geo}}(Q_1,\ldots,Q_m),$$

where $\Pi(\pi_1, \pi_2, \ldots, \pi_m)$ is the subset of tensors w in $\mathcal{M}_{K_1,K_2,\ldots,K_m}(\mathbb{R}^+)$ with $\pi_1, \pi_2, \ldots, \pi_m$ as discrete marginals.

In practice, we compute all the values $Q^*_{\text{Geo}}(Q_1, \ldots, Q_m)$ and the Gaussian plans $\gamma^*_{k_1 k_2 \ldots k_m}$ between the components of the GMM. We solve for the weights w^* using a linear programming solver (e.g., `linprog` from Scipy).

Compute Q^*_{Geo} from w^*: To solve the discrete multi-marginal problem, we use `linprog` from SciPy. The function `create_cost_matrix_from_gmm` in Judelo (2024) takes a list of GMM parameters, a vector of weights, and an integer N (corresponding to the number of iterations in the function `GaussianMedianW2`), and outputs the cost matrix C and the list of all Gaussian geometry medians between the Gaussian components of the GMM. The function `solveMMOT` in Judelo (2024) constructs matrices A and b to encode the constraints on w and uses `linprog` to solve:

$$\inf_{Aw=b} \langle C, w \rangle.$$

The resulting weights correspond to the mixture components of our Q^*_{Geo}. To summarize, the algorithm proceeds as follows:

Algorithm 2: Geometric Median among Many Gaussian Mixtures

Input : m Gaussian mixtures $Q_1, \ldots, Q_m$
Output: Geometric median Q^*_{Geo}
Create Cost Matrix:
Represent Q_j as a weighted sum of components: $Q_j = \sum_{k=1}^{K_j} \pi_j^k Q_j^k$. Use the function `create_cost_matrix_from_gmm` to compute the cost matrix C and the list of Gaussian geometry median:

$$C = Q^*_{\text{Geo}}(Q_1^{k_1}, \ldots, Q_m^{k_m})$$

Set up Linear Program:
Construct matrices A and b encoding the constraints on the weights w, where $w \in \Pi(\pi_1, \ldots, \pi_m)$ and $Aw = b$;
Solve MMOT Problem:
Use `linprog` to solve the linear programming problem:

$$\inf_{Aw=b} \langle C, w \rangle$$

5.3 Algorithm for Computing Geometric Median of General Distributions

Minsker et al. (2017) proposed an algorithm for computing the median of any posterior distribution with respect to the Reproducing Kernel Hilbert Space (RKHS) distance, which corresponds to the

median in Equation (5). An R package for RKHS with a radial basis function (RBF) kernel, developed by You (2021), facilitates the implementation of this algorithm. The general procedure is outlined as follows:

Algorithm 3: Geometric median of probability distributions via Weiszfeld's

Input : Discrete measures $Q_1, \ldots, Q_m$; The kernel $k(\cdot,\cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$; Threshold $\varepsilon > 0$

Output: Weights for the median respect to Discrete measure: $w_* := \left(w_1^{(t+1)}, \ldots, w_m^{(t+1)}\right)$

Initialize: Set $w_j^{(0)} := \frac{1}{m}, j = 1 \ldots m$; and $Q^{(0)} := \frac{1}{m}\sum_{j=1}^m Q_j$.

Starting from $t = 0$, for each $j = 1, \ldots, m$:

while $\left\|Q^{(t+1)} - Q^{(t)}\right\|_{\mathcal{F}_k} \leq \varepsilon$ **do**

1. Update $w_j^{(t+1)} = \frac{\left\|Q^{(t)} - Q_j\right\|_{\mathcal{F}_k}^{-1}}{\sum_{i=1}^m \left\|Q^{(t)} - Q_i\right\|_{\mathcal{F}_k}^{-1}}$;
2. Update $Q_t^{(t+1)} = \sum_{j=1}^m w_j^{(t+1)} Q_j$;

5.4 Likelihood Power Adjustment in Practice

Since we performed data-splitting and computed the geometric median for our VM-Posterior, regularization of the covariance term is necessary. In the variational approximation algorithm, there are two ways to address this:

- For the SVI algorithm, we optimize the Evidence Lower Bound (ELBO). To account for likelihood power adjustment, we optimize the following term:

$$E_{q(\theta)}\{m \cdot \log p(X|\theta) - \log \pi(\theta) - \log q(\theta)\}$$

  where m is the number of sub-datasets into which we split the data.

- Alternatively, we can adjust the covariance term of the VM-Posterior directly by dividing the covariance by $\sqrt{m}$.

6 Numerical studies

In this section, we compare the performance of our VM-Posterior with that of the variational posterior, M-posterior, and the original posterior distribution across the following models:

6.1 Multivariate Gaussian Models

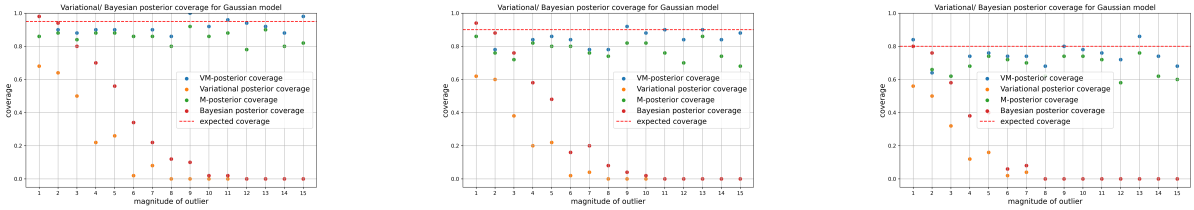
This section demonstrates with a simple example that the effect of the magnitude of an outlier on the variational posterior distribution of the mean parameter μ. We show that the VM-Posterior achieves robustness similar to the M-posterior in Minsker et al. (2017), with significantly faster computation. We consider the univariate Gaussian model $X \sim N(\mu, 1)$.

Firstly, we simulated 25 data sets, each containing 100 observations. Each data set $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,100})$ contained 99 independent observations from a standard Gaussian distribution $x_{i,j} \sim N(\mu = 2, 1)$ for $i = 1, \ldots, 25$ and $j = 1, \ldots, 99$. The last entry in each data set $x_{i,100}$ was an outlier, and its value increased linearly for $i = 1, \ldots, 15$ with $x_{i,100} = \max(\mathbf{x}_{i,1:99})$. The index of the outlier was unknown to the estimation algorithm, and we assumed that the variance of observations was known.

For computation, we performed both standard variational Bayes using `cmstan` with 2 chains and 1000 samples, and VM-posterior for the multivariate Gaussian model. We set the initial values

to the true value of μ and ran 1000 iterations. The VM-posterior approach proceeded as follows: For each data set $\mathbf{x}_i$, we randomly separated it into 10 groups, $G_1, \ldots, G_{10}$. We then computed $q(\cdot|G_1), \ldots, q(\cdot|G_{10})$ and used Algorithm 1 to find the geometric median of the multivariate Gaussians.

This method was applied to each data set $\mathbf{x}_1, \ldots, \mathbf{x}_{15}$ to analyze the impact of the outlier magnitude. We replicated each outlier level 50 times to assess whether the posterior credible interval contained the true value of μ. In Figure 2, we plot the coverage of different credible interval



(a) α=0.05 (b) α=0.1 (c) α=0.2

Figure 2: Posterior coverage for different levels of significance

levels—80%, 90%, and 95%. The magnitude of the outlier increases as i times the maximum data value. The standard Bayes and VB methods exhibit low coverage when $i = 2$, and they almost completely fail to cover the true parameter as the outlier magnitude increases further. In contrast, both the M-posterior and the VM-posterior maintain coverage close to the expected levels, regardless of the outlier magnitude, across all three significance levels. In Figure 3, we demonstrate

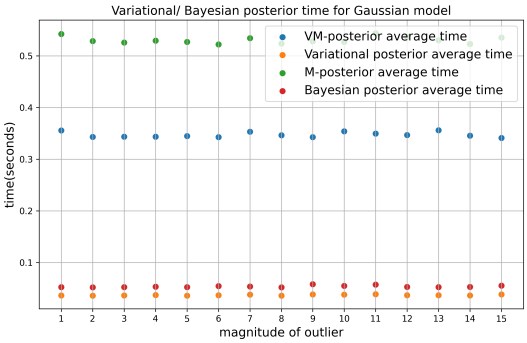


Figure 3: Posterior coverage computational cost

the computational advantages of the variational approach. The variational method is significantly faster than the standard Bayesian approach in our Gaussian examples. When applied to multiple data sets, the computational advantages of the variational approach become even more pronounced.

6.2 Posterior Predictive Density for Gaussian Mixture

In this section, we calculate the posterior predictive distribution using a mixture of Student's t-distributions from Bishop (2006):

$$p(\widehat{\mathbf{x}} \mid \mathbf{X}) = \frac{1}{\widehat{\alpha}} \sum_{k=1}^{K} \alpha_k \, \text{Student's-t} \left(\widehat{\mathbf{x}} \mid \mathbf{m}_k, \mathbf{L}_k, \nu_k + 1 - D \right)$$

where the k^{th} component has mean $\mathbf{m}_k$, and the precision matrix is given by

$$\mathbf{L}_k = \frac{(\nu_k + 1 - D)\,\beta_k}{(1+\beta_k)}\mathbf{W}_k$$

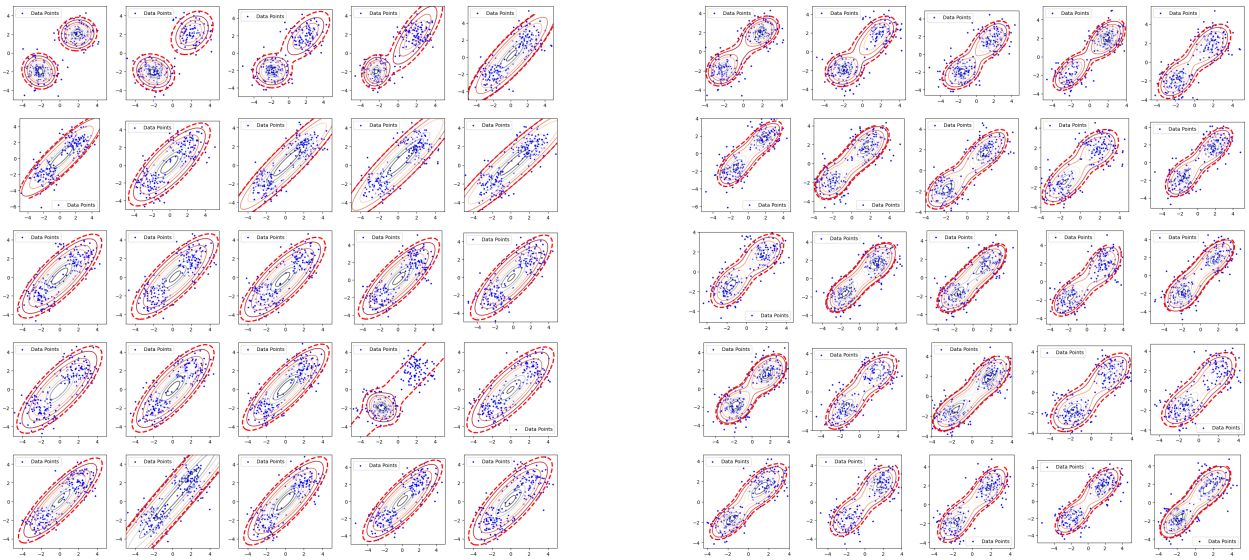
In our case, when the data set size N is large, the predictive distribution approximates a mixture of Gaussians.

Following the approach in the previous section, we simulate 25 data sets, each containing 200 observations. Each data set $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,200})$ consists of 199 independent observations from the Gaussian mixture distribution:

$$x_{i,j} \sim 0.5 * N(\mu = 2, 1) + 0.5 * N(\mu = 4, 1) \quad \text{for } i = 1, \ldots, 25 \text{ and } j = 1, \ldots, 199$$

The last entry in each data set, $x_{i,200}$, was an outlier with its value increasing linearly for $i = 1, \ldots, 25$ as $x_{i,200} = \max(|\mathbf{x}_{i,1:199}|)$. The outlier index was unknown to the estimation algorithm.

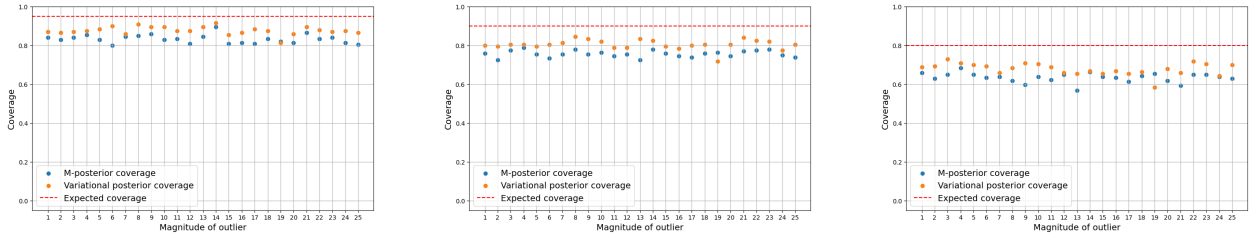
We compute the variational posterior predictive distribution using both the standard variational inference and the VM-posterior procedure. The standard variational posterior predictive distribution is calculated following Kapourani (2019). The VM-posterior is based on Algorithm 2 to obtain the geometric median. We evaluate the robustness of the two methods as we increase the magnitude of the outlier.



(a) Standard Variational Inference (b) Median Variational Inference

Figure 4: Variational inference for Guassian mixture with the increasing magnitude of the outlier

We plot the 95% posterior predictive credible region as the magnitude of the outlier increases. The outlier magnitude grows from left to right and from top to bottom. For better visualization, we omit plotting the outliers themselves. We observe that as the outlier magnitude increases, the standard method loses its ability to capture the mixture properties of the distribution. In contrast, the median method remains consistent regardless of the outlier magnitude. We also compared the posterior predictive coverage of the two methods in Figure 5. We counted the number of data points that lie within the predictive interval for each method. Although the median method achieves similar coverage to the standard method, the predictive region of the VM-Posterior has a smaller area and retains the mixture shape more accurately.



(a) α=0.05 (b) α=0.1 (c) α=0.2

Figure 5: Posterior predictive coverage for different levels of significance

6.3 Language Modeling: Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative model that uncovers hidden topics in a collection of documents based on the words within them.

Assume there are M documents, each containing N_m words, drawn from a vocabulary of size V. LDA assumes there are K latent topics, and each document is a mixture of these topics. For each document m, we represent the topic mixture with a vector θ_m, which has K elements (one for each topic). For each topic k, we represent the distribution over words with a vector ϕ_k, which has V elements (one for each word in the vocabulary).

The generative process is as follows:

$$\begin{aligned}
\theta_m &\sim p_\theta, \quad \text{for each document } m = 1, 2, \ldots, M, \\
\phi_k &\sim p_\phi, \quad \text{for each topic } k = 1, 2, \ldots, K, \\
z_{m,j} &\sim \text{Mult}(\theta_m), \quad \text{for each word } j = 1, 2, \ldots, N_m \text{ in document } m, \\
w_{m,j} &\sim \text{Mult}(\phi_{z_{m,j}}), \quad \text{for each word } j = 1, 2, \ldots, N_m \text{ in document } m.
\end{aligned}$$

Here θ_m is the topic distribution for document m, ϕ_k is the word distribution for topic k, $z_{m,j}$ is the topic assigned to the j-th word in document m, $w_{m,j}$ is the actual word at position j in document m, based on its assigned topic $z_{m,j}$.

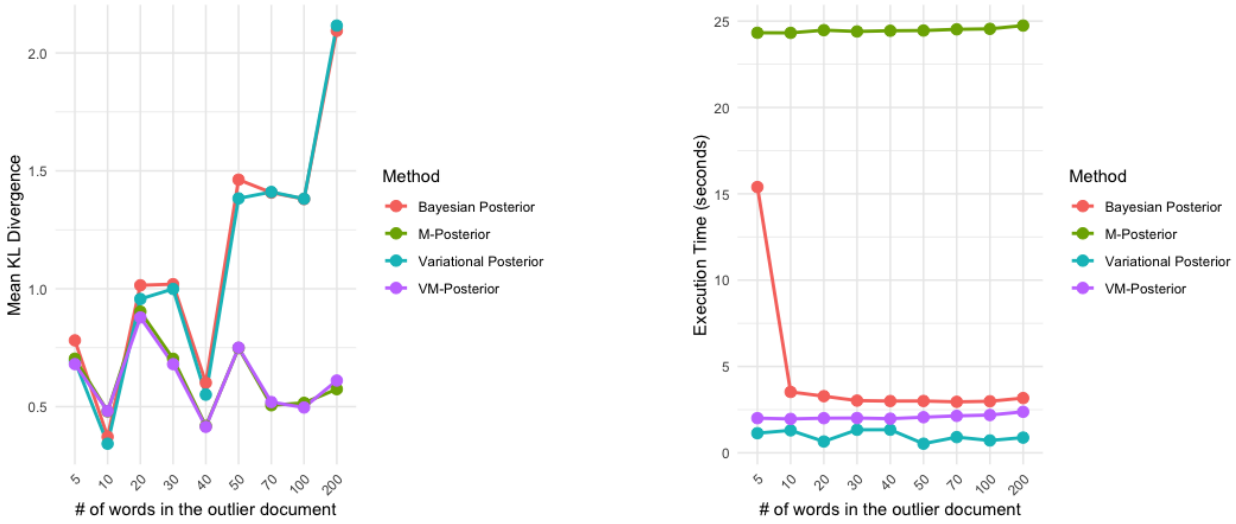
In simpler terms, the first two equations assign "priors" to the document-topic distributions θ_m and the topic-word distributions ϕ_k. The next two equations describe the process of selecting a topic for each word in a document and then choosing the word based on that topic.

In our simulation, we generate data from the LDA model with $M = 19$ documents, where each document contains $N_m \sim \text{Pois}(10)$ words. The words are drawn from a vocabulary of size $V = 4$, and there are $K = 2$ topics. Additionally, we include one "outlier" document with N words, using the same vocabulary. Our primary interest is in the global parameter ϕ_k.

We generate data using p_ϕ, a Dirichlet distribution with parameter $\beta = 1$, and p_θ, a Dirichlet distribution with parameter $\alpha = 2$. Thus, we have the true value of ϕ_k, denoted as ϕ_0.

To measure the performance of our simulation, we compute the KL divergence between $\Pi(\phi_k \mid \text{Documents})$ (the posterior distribution) and the true value ϕ_0. We use HMC for the Bayesian posterior, the MFVB for the variational posterior. For the M-posterior and VM-posterior, we utilize Algorithm 3.

In Figure 6, we fit the LDA model to our simulated data using the standard Bayesian method, M-Posterior, Variational Bayes, and VM-Posterior. We increase the number of words in the outlier document and calculate the mean KL divergence of ϕ. The results show that our VM-Posterior approach performs as well as the Bayesian median approach. Furthermore, our VM-Posterior based approach is significantly faster than the Bayesian approach while achieving similar performance.



(a) Mean KL Divergence Comparison

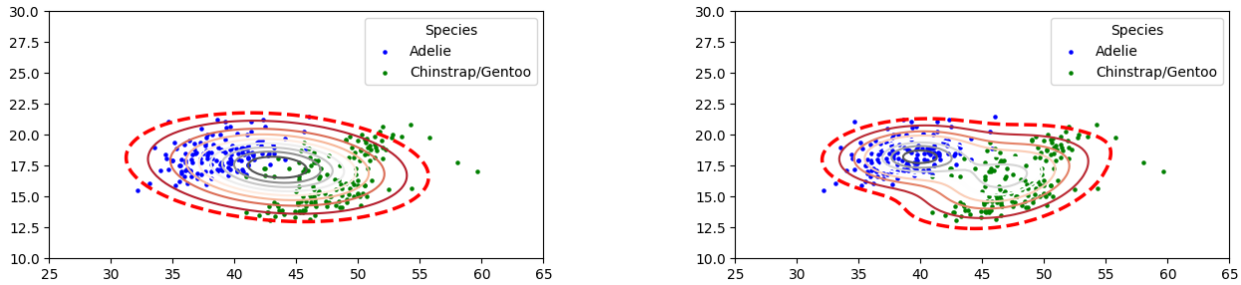
(b) Computational Time Comparison

Figure 6: Variational inference for LDA model

6.4 Real Data Analysis

In this section, we apply our methods to the "penguins" data set from Kaggle. This data set contains over 300 observations for three penguin species, with 6 covariates. We use the first 299 observations from the dataset and model it as a Gaussian mixture with 2 components for the covariates `culmen_length` and `culmen_depth`. We add an outlier with a value 5 times the largest in the dataset.

We fit both the standard variational posterior and the VM-Posterior to the penguins dataset. In Figure 7, we plot the dataset without the added outlier, with each species of penguin represented by a different color. When we fit a Gaussian mixture with 2 clusters using the standard variational approach (Figure 7a), we observe that the 95% credible interval lacks meaningful information about the mixture structure, treating the two species as a single cluster. In contrast, the VM-Posterior (Figure 7b) models the mixture data much better.



(a) Variational posterior

(b) VM-Posterior

Figure 7: Variational inference for penguins dataset with artificial outlier

7 Conclusion and Future Direction

This paper introduced the VM-Posterior, a novel approach for robust variational inference that combines computational efficiency with resistance to outliers and data contamination. The key innovation lies in using geometric or metric median aggregation in the Wasserstein distance framework during the variational posterior aggregation step. This approach not only enhances robustness but also improves concentration properties, as weak concentrations of subset-based variational posteriors are transformed into strong concentrations under the Wasserstein geometric or metric median. By leveraging this aggregation method, the VM-Posterior achieves significant improvements over traditional variational methods while maintaining comparable robustness to established techniques like the M-Posterior.

Through rigorous theoretical development and diverse experiments, we demonstrated the VM-Posterior's robustness, efficiency, and adaptability. Theoretical results established the VM-Posterior's strong concentration properties, asymptotic normality, and finite-sample guarantees, providing a strong foundation for its practical utility. The algorithmic framework, designed for mean-field and Gaussian variational families, was extended to handle general distributions using efficient numerical methods. Empirical results across synthetic and real-world datasets confirmed the VM-Posterior's superior performance in retaining credible coverage, preserving mixture structures, and resisting contamination effects, all while achieving computational efficiency that is crucial for large-scale data settings.

Building on these results, the framework showcased its versatility across multivariate Gaussian models, Gaussian mixtures, and the Latent Dirichlet Allocation (LDA) model. Real-world applicability was further validated using the penguins dataset, where the VM-Posterior demonstrated its ability to model complex data distributions while mitigating the effects of outliers. By balancing the strengths of variational inference and robust Bayesian estimation, the VM-Posterior positions itself as a powerful and practical tool for modern Bayesian inference challenges.

Future research may focus on extending the theoretical foundation of the VM-Posterior to justify its empirical success in Gaussian mixture models. While the robustness results developed here apply to general variational families, extending Corollary 7 in Wang and Blei (2019) to handle mixtures of Gaussians would establish strong asymptotic normality results for this specific family. Such advancements would solidify the numerical findings presented for Gaussian mixtures and provide a more comprehensive theoretical framework.

Another promising direction involves broadening the applicability of the VM-Posterior to more complex models and exploring its use in diverse machine learning and statistical settings. Applications in Bayesian deep learning, hierarchical models, or dynamic systems could showcase its robustness and computational efficiency in handling the challenges posed by high-dimensional, non-linear, and evolving data structures.

Appendix A. Proof of Theorem 5 and Theorem 6

Theorem 5 and Theorem 6 are an adaptation of Theorem 3.1 in Minsker (2015). Precisely Theorem 5 and Theorem 6 follows directly from the following Theorem in Minsker et al. (2017).

Theorem 11 **a**. *Assume that* $(\mathbb{H}, \|\cdot\|)$ *is a Hilbert space and* $\theta_0 \in \mathbb{H}$. *Let* $\hat{\theta}_1, \ldots, \hat{\theta}_m \in \mathbb{H}$ *be a collection of independent random variables. Let the constants* α, q, γ *be such that* $0 < q < \alpha < 1/2$, *and* $0 \leq \gamma < \frac{\alpha - q}{1-q}$. *Suppose* $\varepsilon > 0$ *is such that for all* $j, 1 \leq j \leq \lfloor (1-\gamma)m \rfloor + 1$,

$$\Pr\left(\left\|\hat{\theta}_j - \theta_0\right\| > \varepsilon\right) \leq q.$$

Let $\hat{\theta}_* = \text{med}_g\left(\hat{\theta}_1, \ldots, \hat{\theta}_m\right)$ *be the geometric median of* $\left\{\hat{\theta}_1, \ldots, \hat{\theta}_m\right\}$. *Then*

$$\Pr\left(\left\|\hat{\theta}_* - \theta_0\right\| > C_\alpha \varepsilon\right) \leq \left[e^{(1-\gamma)\psi\left(\frac{\alpha-\gamma}{1-\gamma}, q\right)}\right]^{-m}$$

where $C_\alpha = (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$.

b. *Assume that* $(\mathbb{Y}, d)$ *is a metric space and* $\theta_0 \in \mathbb{Y}$. *Let* $\hat{\theta}_1, \ldots, \hat{\theta}_m \in \mathbb{Y}$ *be a collection of independent random variables. Let the constants* q, γ *be such that* $0 < q < \frac{1}{2}$ *and* $0 \leq \gamma < \frac{1/2 - q}{1-q}$. *Suppose* $\varepsilon > 0$ *are such that for all* $j, 1 \leq j \leq \lfloor (1-\gamma)m \rfloor + 1$,

$$\Pr\left(d\left(\hat{\theta}_j, \theta_0\right) > \varepsilon\right) \leq q.$$

Let $\hat{\theta}_* = \text{med}_0\left(\hat{\theta}_1, \ldots, \hat{\theta}_m\right)$. *Then*

$$\Pr\left(d\left(\hat{\theta}_*, \theta_0\right) > 3\varepsilon\right) \leq e^{-m(1-\gamma)\psi\left(\frac{1/2-\gamma}{1-\gamma}, q\right)}$$

To get Theorem 5 in our paper, we take $q = \frac{1}{7}$ and $\alpha = \frac{3}{7}$ in part **(a)** of the previously theorem. Moreover, Theorem 6 is followed by part **(b)**, when considering $q = \frac{1}{4}$.

For completeness, we present the proof of the above Theorem. To this end, we make use of the following lemma (see lemma 2.1 in Minsker (2015)).

Lemma 12 *Let* $\mathbb{H}$ *be a Hilbert space,* $x_1, \ldots, x_m \in \mathbb{H}$ *and let* x_* *be their geometric median. Fix* $\alpha \in \left(0, \frac{1}{2}\right)$ *and assume that* $z \in \mathbb{H}$ *is such that* $\|x_* - z\| > C_\alpha r$, *where*

$$C_\alpha = (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$$

and $r > 0$. *Then there exists a subset* $J \subseteq \{1, \ldots, m\}$ *of cardinality* $|J| > \alpha m$ *such that for all* $j \in J, \|x_j - z\| > r$.

We now provide the proof of Theorem 11.

Proof Assume that event $\mathcal{E} := \left\{ \left\| \hat{\theta}_* - \theta_0 \right\| > C_\alpha \varepsilon \right\}$ occurs. Lemma 12 implies that there exists a subset $J \subseteq \{1, \ldots, m\}$ of cardinality $|J| \geq \alpha k$ such that $\left\| \hat{\theta}_j - \theta_0 \right\| > \varepsilon$ for all $j \in J$, hence

$$\Pr(\mathcal{E}) \leq \Pr\left(\sum_{j=1}^{m} I\left\{ \left\| \hat{\theta}_j - \theta_0 \right\| > \varepsilon \right\} > \alpha m \right) \leq \tag{15}$$

$$\Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{ \left\| \hat{\theta}_j - \theta_0 \right\| > \varepsilon \right\} > (\alpha - \gamma) m \frac{\lfloor (1-\gamma)m \rfloor + 1}{\lfloor (1-\gamma)m \rfloor + 1} \right) \leq \tag{16}$$

$$\Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{ \left\| \hat{\theta}_j - \theta_0 \right\| > \varepsilon \right\} > \frac{\alpha - \gamma}{1-\gamma} (\lfloor (1-\gamma)m \rfloor + 1) \right). \tag{17}$$

If W has Binomial distribution $W \sim B(\lfloor (1-\gamma)m \rfloor + 1, q)$, then

$$\Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{ \left\| \hat{\theta}_j - \theta_0 \right\| > \varepsilon \right\} > \frac{\alpha - \gamma}{1-\gamma} (\lfloor (1-\gamma)m \rfloor + 1) \right) \leq \tag{18}$$

$$\Pr\left(W > \frac{\alpha - \gamma}{1-\gamma} (\lfloor (1-\gamma)m \rfloor + 1) \right) \tag{19}$$

(see Lemma 23 in Lerasle and Oliveira (2011) for a rigorous proof of this fact). Chernoff bound (e.g., Proposition A.6.1 in van der Vaart and Wellner (1996)), together with an obvious bound $\lfloor (1-\gamma)m \rfloor + 1 > (1-\gamma)m$, implies that

$$\Pr\left(W > \frac{\alpha - \gamma}{1-\gamma} (\lfloor (1-\gamma)m \rfloor + 1) \right) \leq \exp\left(-m(1-\gamma) \psi\left(\frac{\alpha - \gamma}{1-\gamma}, q \right) \right).$$

To establish part **b**, we proceed as follows: let $\mathcal{E}_1$ be the event $\mathcal{E}_1 = \{$ more than a half of events $d\left(\hat{\theta}_j, \theta_0\right) \leq \varepsilon, j = 1 \ldots m$ occur $\}$. Assume that $\mathcal{E}_1$ occurs. Then we clearly have $\varepsilon_* \leq \varepsilon$, where ε_* is defined as

$$\varepsilon_* := \inf\{\varepsilon > 0 : \exists j = j(\varepsilon) \in \{1, \ldots, m\} \text{ and } I(j) \subset \{1, \ldots, m\} \text{ such that}$$
$$|I(j)| > \frac{m}{2} \text{ and } \forall i \in I(j), d\left(\hat{\theta}_i, \hat{\theta}_j\right) \leq 2\varepsilon \Big\}.$$

Indeed, for any $\theta_{j_1}, \theta_{j_2}$ such that $d\left(\hat{\theta}_{j_i}, \theta_0\right) \leq \varepsilon, i = 1, 2$, triangle inequality gives $d\left(\theta_{j_1}, \theta_{j_2}\right) \leq 2\varepsilon$. By the definition of $\hat{\theta}_*$, inequality $d\left(\hat{\theta}_*, \hat{\theta}_j\right) \leq 2\varepsilon_* \leq 2\varepsilon$ holds for at least a half of $\left\{\hat{\theta}_1, \ldots, \hat{\theta}_m\right\}$, hence, it holds for some $\hat{\theta}_{\tilde{j}}$ with $d\left(\hat{\theta}_{\tilde{j}}, \theta_0\right) \leq \varepsilon$. In turn, this implies (by triangle inequality) $d\left(\hat{\theta}_*, \theta_0\right) \leq 3\varepsilon$. We conclude that

$$\Pr\left(d\left(\hat{\theta}_*, \theta_0\right) > 3\varepsilon \right) \leq \Pr\left(\mathcal{E}_1\right)$$

The rest of the proof repeats the argument of part **a** since

$$\Pr\left(\mathcal{E}_1^c\right) = \Pr\left(\sum_{j=1}^{m} I\left\{ d\left(\hat{\theta}_j, \theta_0\right) > \varepsilon \right\} \geq \frac{m}{2} \right),$$

where $\mathcal{E}_1^c$ is the complement of $\mathcal{E}_1$. ■

Appendix B. Proof of Theorem 3

Proof To start, we consider the events

$$\mathbb{B}_l = \mathbb{B}_l\left(ml\varepsilon_l^2, Q_{n,m}^{(j)}, \Pi\right) := \Big\{ \int \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \geq \exp(-ml\varepsilon_l^2 - \mathrm{KL}(Q_{n,m}^{(j)}, \Pi))\Big\},$$

and

$$\mathbb{A}_l = \Big\{ \int_{\rho(\theta,\theta_0)>R\varepsilon_l} \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq e^{-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2}\Big\}.$$

Next, observe that

$$\begin{aligned}
&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\\
=&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l} + \widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l^c}\\
\leq&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l} + \mathbf{1}_{\mathbb{B}_l^c},
\end{aligned} \tag{20}$$

where the fact that $\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\leq 1$ has been used. Similarly, we have that

$$\begin{aligned}
&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{B_l}\\
=&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l^c}\\
\leq&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \mathbf{1}_{\mathbb{A}_l^c}.
\end{aligned} \tag{21}$$

Therefore, from Equation (20) and Equation (21),

$$\widehat{Q}_{n,m}^{(l)}(\rho(\theta,\theta_0)\geq R\varepsilon_l) \leq \widehat{Q}_{n,m}^{(l)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \mathbf{1}_{\mathbb{A}_l^c} + \mathbf{1}_{\mathbb{B}_l^c} \tag{22}$$

Now, we observe that using Lemma 13 with $v_l = ml\varepsilon_l^2$, we get that

$$\begin{aligned}
&\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\\
\leq&\frac{1}{v_l}\Big[\mathrm{KL}(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}^{|G_j|})\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\Big] + \exp(v_l)\Big[\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\Big].
\end{aligned} \tag{23}$$

By Equation (22) and Equation (23) it follows that,

$$\begin{aligned}
\mathbb{E}_0\Big(\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\Big) \leq& \mathbb{E}_0\left(\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\right)\\
&+ \mathbb{E}_0(\mathbf{1}_{\mathbb{A}_l^c}) + \mathbb{E}_0(\mathbf{1}_{\mathbb{B}_l^c}) + \mathbb{E}_0(\frac{1}{v_l}\Big[\mathrm{KL}(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}^{|G_j|})\Big])
\end{aligned}$$

Then by Markov's inequality,

$$\begin{aligned}
&\Pr\Big(\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\Big)\\
&\leq \exp(v_l)\mathbb{E}_0(\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}) + P(\mathbb{A}_l^c) + P(\mathbb{B}_l^c) + \frac{\mathbb{E}_0\Big[\mathrm{KL}(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}^{|G_j|})\Big]}{v_l}\\
&= I_1 + I_2 + I_3 + I_4.
\end{aligned} \tag{24}$$

Now, we bound each of the terms I_1, I_2, I_3 and I_4. For the term I_1, we notice that, on the event $\mathbb{A}_l\cap\mathbb{B}_l$ it is satisfied

$$\int_{\rho(\theta,\theta_0)>R\varepsilon_l} \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq \exp(-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2)$$

and,

$$\int \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \geq \exp(-ml\varepsilon_l^2 - \mathrm{KL}(Q_{n,m}^{(j)^*}, \Pi)).$$

In consequences,

$$\Pi_{n,m}(\rho(\theta,\theta_0) \geq R\varepsilon_l | G_j)\mathbf{1}_{\mathbb{B}_l \cap \mathbb{A}_l} \leq \exp(ml\varepsilon_l^2 + \mathrm{KL}(Q_{n,m}^{(j)^*}, \Pi) - \mathfrak{c}_1^l R^2 ml\varepsilon_l^2) \tag{25}$$

Therefore, by Equation (25), and by Assumption 1, we have that

$$I_1 \leq \exp(3ml\varepsilon_l^2 + \mathfrak{c}_5^l l\varepsilon_l^2 - \mathfrak{c}_1^l R^2 ml\varepsilon_l^2). \tag{26}$$

In the following line of arguments we proceed to bound I_2 and I_3. For I_2, Since if Theorem 2 holds with $\zeta := \varepsilon_l$, then it also holds with $\zeta := L\varepsilon_l$ for any $L \geq 1$, we have that by Assumption 2

$$\begin{aligned} P_0^{(l)}(\mathbb{A}_l) =& P_0^{(l)}\Big(\int_{\rho(\theta,\theta_0)>R\varepsilon} \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq \exp(-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2)\Big) \\ \geq& 1 - 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2). \end{aligned}$$

This is,

$$I_2 \leq 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2). \tag{27}$$

Finally, we bound the term I_3. For this purpose, we apply Lemma 13 with $\mathrm{F} = \log\Big(\Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m\Big)$, $Q_0 = Q_{n,m}^{(j)^*}$ and $\Pi_0 = \Pi$ to obtain

$$\log \int \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m \mathrm{d}\Pi(\boldsymbol{\theta}) \geq \int \log \left(\Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m \right) \mathrm{d}Q_{n,m}^{(j)^*}(\,\mathrm{d}\boldsymbol{\theta}) - \mathrm{KL}(Q_{n,m}^{(j)^*}, \Pi).$$

Using the definition of event $\mathbb{B}$, we have that

$$\mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) = \mathrm{P}_0^{(l)}\left(\log \int \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m \mathrm{d}\Pi(\boldsymbol{\theta}) < -ml\varepsilon_l^2 - \mathrm{KL}(Q_{n,m}^{(j)^*}, \Pi)\right),$$

from where,

$$\mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) \leq \mathrm{P}_0^{(l)}\left(- \int \log \left(\Big(\prod_{j\in G_j} \frac{p_0}{p_\theta}(X_j)\Big)^m \right) \mathrm{d}Q_{n,m}^{(j)^*}(\,\mathrm{d}\boldsymbol{\theta}) \leq -ml\varepsilon_l^2 \right).$$

Hence,

$$\begin{aligned} \mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) &\leq \mathrm{P}_0^{(l)}\left(\int 0 \vee \log \Big(\prod_{j\in G_j} \frac{p_0}{p_\theta}(X_j)\Big) \mathrm{d}Q_{n,m}^{(j)^*}(\boldsymbol{\theta}) \geq l\varepsilon_l^2 \right) \\ &\leq \frac{1}{l\varepsilon_l^2}\mathrm{P}_0^{(l)}\left[\int 0 \vee \log \Big(\prod_{j\in G_j} \frac{p_0}{p_\theta}(X_j)\Big) \mathrm{d}Q_{n,m}^{(j)^*}(\boldsymbol{\theta}) \right] \\ &\leq \frac{1}{l\varepsilon_l^2} Q_{n,m}^{(j)^*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right) + \sqrt{\frac{1}{2}\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)} \right] \\ &\leq \frac{1}{l\varepsilon_l^2}\left(2Q_{n,m}^{(j)^*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)\right] + 1\right). \end{aligned}$$

Here we use Markov's inequality for the fourth line and use Fubini's theorem and Lemma B.13 of Ghosal and Van der Vaart (2017) for the fifth line. For the last line, we use a simple inequality $z + \sqrt{z/2} \leq z + (z \vee 1) \leq 2z + 1$ for $z \geq 0$. Thus, by Assumption 1

$$I_3 = \mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) \leq \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l \frac{1}{m} \tag{28}$$

Finally, we analyze the term I_4. Using Proposition 1, we get that

$$I_4 = \frac{\mathbb{E}_0\left[\mathrm{KL}(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}^{|G_j|})\right]}{v_l} \leq \frac{\mathfrak{c}_5^l l\varepsilon_l^2}{ml\varepsilon_l^2} = \frac{\mathfrak{c}_5^l}{m}. \tag{29}$$

To conclude, we observe that by Equation (24), (26), (27), (28), and (29), we obtain that

$$\begin{aligned}\mathbb{E}_0\Big(\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0) \geq R\varepsilon_l)\Big) &\leq \exp(3ml\varepsilon_l^2 + \mathfrak{c}_5^l l\varepsilon_l^2 - \mathfrak{c}_1^l R^2 ml\varepsilon_l^2) \\ &\quad + 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2) + \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l\frac{1}{m} + \frac{\mathfrak{c}_5^l}{m}.\end{aligned}$$

Furthermore, using the fact that $l\varepsilon_l^2 \geq 1$ and considering R such that $R > \sqrt{\frac{6+2\mathfrak{c}_5^l}{\mathfrak{c}_1^l}}$, we get that

$$\leq \exp(-\frac{\mathfrak{c}_1^l}{2}R^2 l\varepsilon_l^2) + 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2) + \frac{1}{l\varepsilon_l^2} + 3\mathfrak{c}_5^l\frac{1}{m},$$

and the claim is followed. ■

Appendix C. Proof of Theorem 7

The proof of Theorem 7 follows a similar line of reasoning as the proof of Theorem 3, with minor modifications. For the sake of completeness, we provide the full details of the proof below, including steps that were previously analyzed in the proof of Theorem 3.

Proof By the definition of Wasserstein distance, we observe that

$$\begin{aligned} d_{W_1,\rho}(\widehat{Q}^{(j)}_{n,m},\delta_0) =& \int_\Theta \rho(\theta,\theta_0) d\widehat{Q}^{(j)}_{n,m}(\theta) \\ \leq& R\varepsilon_l + \int\limits_{\rho(\theta,\theta_0)\geq R\varepsilon_l} d\widehat{Q}^{(j)}_{n,m}(\theta) \\ =& R\varepsilon_l + \widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l). \end{aligned}$$

To obtain the result it remains to bound $\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)$. To this end, we consider the events

$$\mathbb{B}_l = \mathbb{B}_l\left(ml\varepsilon_l^2, Q^{(j)}_{n,m}, \Pi\right) := \Big\{ \int \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \geq \exp(-ml\varepsilon_l^2 - \mathrm{KL}(Q^{(j)}_{n,m},\Pi))\Big\},$$

and

$$\mathbb{A}_l = \Big\{ \int_{\rho(\theta,\theta_0)>R\varepsilon_l} \Big(\prod_{j\in G_j} \frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq e^{-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2}\Big\}.$$

Next, observe that

$$\begin{aligned} &\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l) \\ =&\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l} + \widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l^c} \\ \leq&\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l} + \mathbf{1}_{\mathbb{B}_l^c}, \end{aligned} \tag{30}$$

where the fact that $\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\leq 1$ has been used. Similarly, we have that

$$\begin{aligned} &\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{B_l} \\ =&\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l^c} \\ \leq&\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \mathbf{1}_{\mathbb{A}_l^c}. \end{aligned} \tag{31}$$

Therefore, from Equation (30) and Equation (31),

$$\widehat{Q}^{(l)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l) \leq \widehat{Q}^{(l)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} + \mathbf{1}_{\mathbb{A}_l^c} + \mathbf{1}_{\mathbb{B}_l^c} \tag{32}$$

Now, we observe that using Lemma 13 with $v_l = ml\varepsilon_l^2$, we get that

$$\begin{aligned} &\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} \\ \leq&\frac{1}{v_l}\Big[\mathrm{KL}(\widehat{Q}^{(j)}_{n,m},\Pi^{|G_j|}_{n,m})\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\Big] + \exp(v_l)\Big[\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l}\Big]. \end{aligned} \tag{33}$$

By Equation (32) and Equation (33) it follows that,

$$\begin{aligned} &\Pr\left(\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l) > \exp(-ml\varepsilon_l^2) + R\varepsilon_l\right) \\ &\leq \Pr(\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} > \exp(-ml\varepsilon_l^2 - v_l)) \\ &+ \Pr(\mathbf{1}_{\mathbb{A}_l^c}>0) + \Pr(\mathbf{1}_{\mathbb{B}_l^c}>0) + \Pr(\frac{1}{v_l}\Big[\mathrm{KL}(\widehat{Q}^{(j)}_{n,m},\Pi^{|G_j|}_{n,m})\Big] > R\varepsilon_l) \end{aligned}$$

Then by Markov's inequality,

$$\begin{aligned}
&\Pr\left(\widehat{Q}^{(j)}_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l) > \exp(-ml\varepsilon_l^2) + R\varepsilon_l\right)\\
&\leq \frac{\mathbb{E}_0(\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l})}{\exp(-ml\varepsilon_l^2 - v_l)} + P(\mathbf{1}_{\mathbb{A}_l^c} > 0) + P(\mathbf{1}_{\mathbb{B}_l^c} > 0) + \frac{\mathbb{E}_0\Big[\mathrm{KL}(\widehat{Q}^{(j)}_{n,m}, \Pi^{|G_j|}_{n,m})\Big]}{v_l R\varepsilon_l}\\
&= I_1 + I_2 + I_3 + I_4. \qquad (34)
\end{aligned}$$

Now, we bound each of the terms I_1, I_2, I_3 and I_4. For the term I_1, we notice that, on the event $\mathbb{A}_l \cap \mathbb{B}_l$ it is satisfied

$$\int_{\rho(\theta,\theta_0)>R\varepsilon_l}\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq \exp(-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2)$$

and,

$$\int\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \geq \exp(-ml\varepsilon_l^2 - \mathrm{KL}(Q^{(j)^*}_{n,m}, \Pi)).$$

In consequences,

$$\Pi_{n,m}(\rho(\theta,\theta_0)\geq R\varepsilon_l|G_j)\mathbf{1}_{\mathbb{B}_l\cap\mathbb{A}_l} \leq \exp(ml\varepsilon_l^2 + \mathrm{KL}(Q^{(j)^*}_{n,m}, \Pi) - \mathfrak{c}_1^l R^2 ml\varepsilon_l^2) \qquad (35)$$

Therefore, by Equation (35), and by Assumption 1, we have that

$$I_1 \leq \exp(3ml\varepsilon_l^2 + \mathfrak{c}_5^l l\varepsilon_l^2 - \mathfrak{c}_1^l R^2 ml\varepsilon_l^2). \qquad (36)$$

In the following line of arguments we proceed to bound I_2 and I_3. For I_2, Since if Theorem 2 holds with $\zeta := \varepsilon_l$, then it also holds with $\zeta := L\varepsilon_l$ for any $L \geq 1$, we have that by Assumption 2

$$\begin{aligned}
P_0^{(l)}(\mathbb{A}_l) =& P_0^{(l)}\Big(\int_{\rho(\theta,\theta_0)>R\varepsilon}\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m d\Pi(\theta) \leq \exp(-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2)\Big)\\
\geq& 1 - 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2).
\end{aligned}$$

This is,

$$I_2 \leq 4\exp(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2). \qquad (37)$$

Finally, we bound the term I_3. For this purpose, we apply Lemma 13 with $\mathrm{F} = \log\Big(\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m\Big)$, $Q_0 = Q^{(j)^*}_{n,m}$ and $\Pi_0 = \Pi$ to obtain

$$\log\int\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m \mathrm{d}\Pi(\boldsymbol{\theta}) \geq \int\log\left(\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m\right)\mathrm{d}Q^{(j)^*}_{n,m}(\,\mathrm{d}\boldsymbol{\theta}) - \mathrm{KL}(Q^{(j)^*}_{n,m}, \Pi).$$

Hence, we have

$$\begin{aligned}
\mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) &= \mathrm{P}_0^{(l)}\left(\log\int\Big(\prod_{j\in G_j}\frac{p_\theta}{p_0}(X_j)\Big)^m \mathrm{d}\Pi(\boldsymbol{\theta}) < -ml\varepsilon_l^2 - \mathrm{KL}(Q_{n,m}^{(j)*},\Pi)\right)\\
&\leq \mathrm{P}_0^{(l)}\left(-\int\log\left(\Big(\prod_{j\in G_j}\frac{p_0}{p_\theta}(X_j)\Big)^m\right)\mathrm{d}Q_{n,m}^{(j)*}(\,\mathrm{d}\boldsymbol{\theta}) \leq -ml\varepsilon_l^2\right)\\
&\leq \mathrm{P}_0^{(l)}\left(\int 0\vee\log\Big(\prod_{j\in G_j}\frac{p_0}{p_\theta}(X_j)\Big)\mathrm{d}Q_{n,m}^{(j)*}(\boldsymbol{\theta}) \geq l\varepsilon_l^2\right)\\
&\leq \frac{1}{l\varepsilon_l^2}\mathrm{P}_0^{(l)}\left[\int 0\vee\log\Big(\prod_{j\in G_j}\frac{p_0}{p_\theta}(X_j)\Big)\mathrm{d}Q_{n,m}^{(j)*}(\boldsymbol{\theta})\right]\\
&\leq \frac{1}{l\varepsilon_l^2}Q_{n,m}^{(j)*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)},\mathrm{P}_\theta^{(l)}\right)+\sqrt{\frac{1}{2}\mathrm{KL}\left(\mathrm{P}_0^{(l)},\mathrm{P}_\theta^{(l)}\right)}\right]\\
&\leq \frac{1}{l\varepsilon_l^2}\left(2Q_{n,m}^{(j)*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)},\mathrm{P}_\theta^{(l)}\right)\right]+1\right).
\end{aligned}$$

Here we use Markov's inequality for the fourth line and use Fubini's theorem and Lemma B.13 of Ghosal and Van der Vaart (2017) for the fifth line. For the last line, we use a simple inequality $z+\sqrt{z/2}\leq z+(z\vee 1)\leq 2z+1$ for $z\geq 0$. Thus, by Assumption 1

$$I_3 = \mathrm{P}_0^{(l)}\left(\mathbb{B}_l^C\right) \leq \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l\frac{1}{m} \tag{38}$$

Finally, we analyze the term I_4. Using Proposition 1, we get that

$$I_4 = \frac{\mathbb{E}_0\left[\mathrm{KL}(\widehat{Q}_{n,m}^{(j)},\Pi_{n,m}^{|G_j|})\right]}{v_l R\varepsilon_l} \leq \frac{\mathfrak{c}_5^l l\varepsilon_l^2}{ml\varepsilon_l^2 R\varepsilon_l} = \frac{\mathfrak{c}_5^l}{Rm\varepsilon_l}. \tag{39}$$

To conclude, we observe that by Equation (34), (36), (37), (38), and (39), we obtain that

$$\begin{aligned}
&\Pr\left(\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq R\varepsilon_l) > \exp(-ml\varepsilon_l^2)+R\varepsilon_l\right)\\
&\leq \exp(3ml\varepsilon_l^2+\mathfrak{c}_5^l l\varepsilon_l^2-\mathfrak{c}_1^l R^2 ml\varepsilon_l^2)\\
&+4\exp(-(\mathfrak{c}_2^l)^2R^2l\varepsilon_l^2)+\frac{1}{l\varepsilon_l^2}+2\mathfrak{c}_5^l\frac{1}{m}+\frac{\mathfrak{c}_5^l}{Rm\varepsilon_l}.
\end{aligned}$$

Furthermore, using the fact that $l\varepsilon_l^2\geq 1$ and considering R such that $R>\sqrt{\frac{6+2\mathfrak{c}_5^l}{\mathfrak{c}_1^l}}$, we get that

$$\begin{aligned}
&\Pr\left(\widehat{Q}_{n,m}^{(j)}(\rho(\theta,\theta_0)\geq A_l\varepsilon_l) > \exp(-ml\varepsilon_l^2)+l\varepsilon_l^2\right)\\
&\leq \exp(-\frac{\mathfrak{c}_1^l}{2}R^2l\varepsilon_l^2)+4\exp(-(\mathfrak{c}_2^l)^2R^2l\varepsilon_l^2)+\frac{1}{l\varepsilon_l^2}+2\mathfrak{c}_5^l\frac{1}{m}+\frac{\mathfrak{c}_5^l}{Rm\varepsilon_l},
\end{aligned} \tag{40}$$

and the claim is followed. ■

Appendix D. Proof of Theorem 10

Proof We now proceed to prove part *a*). To achieve this, we first analyze the Wasserstein metric median $Q^*_{\text{Met,GG}}$. By Theorem 9 and Assumption 3, convergence in total variation distance implies convergence of expectations in P_{θ_0}-probability. Specifically, we have,

$$\left\| \int_\Theta \theta \left(dQ^{(j),\text{GG}}_{n,m}(\theta) - dN\left(\theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}, \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)(\theta)\right)\right\|_2 \to 0 \quad \text{as } l \to \infty.$$

Next, note that the total variation distance between two Gaussian distributions $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, sharing the same covariance matrix, is bounded by a multiple of $\|\mu_1 - \mu_2\|_2$. Therefore, we can replace $\theta_0 + \frac{\Delta_{l,\theta_0}}{\sqrt{l}}$ in the above result with the mean,

$$\bar{\theta}^{\text{GG}}_{j,m}(G_j) := \int_\Theta \theta \, dQ^{(j),\text{GG}}_{n,m}(\theta).$$

This substitution allows us to reformulate the conclusion of Theorem 9 as,

$$\left\| Q^{(j),\text{GG}}_{n,m}(\cdot) - N\left(\bar{\theta}^{\text{GG}}_{j,m}(G_j), \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\text{TV}} \to 0 \quad \text{as } l \to \infty,$$

in P_{θ_0}-probability. Now, consider $m = \lfloor n/l \rfloor$ as fixed, and let $n, l \to \infty$. As before, let $G_1, \ldots, G_m$ denote disjoint groups of i.i.d. observations from P_{θ_0}, each of cardinality l. Recall that by the definition of $Q^*_{\text{Met,GG}}$ in Equation (6), $Q^*_{\text{Met,GG}} = Q^{(j_*),\text{GG}}_{n,m}$ for some $j_* \le m$, and its mean is given by $\theta^*_{\text{GG}} := \bar{\theta}^{\text{GG}}_{j_*,m}(G_{j_*})$. By definition, we have:

$$\begin{aligned} &\left\| Q^*_{\text{Met,GG}} - N\left(\theta^*_{\text{GG}}, \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\text{TV}} \\ &\le \max_{j=1,\ldots,m} \left\| Q^{(j),\text{GG}}_{n,m}(\cdot) - N\left(\bar{\theta}^{\text{GG}}_{j,m}(G_j), \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\text{TV}}. \end{aligned} \tag{41}$$

Since the right-hand side converges to zero as $l \to \infty$, we conclude:

$$\left\| Q^*_{\text{Met,GG}} - N\left(\theta^*_{\text{GG}}, \frac{1}{l\cdot m} I^{-1}(\theta_0)\right)\right\|_{\text{TV}} \to 0 \quad \text{as } n \to \infty.$$

This completes the proof for $Q^*_{\text{Met,GG}}$. The proof for $Q^*_{\text{Met,MF}}$ follows analogously by substituting $Q^{(j),\text{GG}}_{n,m}$ with $Q^{(j),\text{MF}}_{n,m}$, and replacing $I^{-1}(\theta_0)$ with $I'^{-1}(\theta_0)$, where $I'^{-1}(\theta_0)$ is a diagonal matrix matching the diagonal elements of $I^{-1}(\theta_0)$. All other steps remain identical, ensuring the same conclusion,

$$\left\| Q^*_{\text{Met,MF}} - N\left(\theta^*_{\text{MF}}, \frac{1}{l\cdot m} I'^{-1}(\theta_0)\right)\right\|_{\text{TV}} \to 0 \quad \text{as } n \to \infty.$$

This completes the proof of part *a*).

For part *b*), we proceed to analyze the mean of the Wasserstein metric median $Q^*_{\text{Met,GG}}$ under the assumptions of the theorem. By assumption, we have that ε_l satisfies $l\varepsilon_l^2 \ge 1$ and $\sqrt{n} \le m$. These conditions imply that, for any $1 \le j \le \lfloor (1-\kappa)m \rfloor + 1$, the following inequality holds,

$$\exp\left(-\frac{\mathfrak{c}_1^l}{2} R^2 l\varepsilon_l^2\right) + 4\exp\left(-(\mathfrak{c}_2^l)^2 R^2 l\varepsilon_l^2\right) + \frac{1}{l\varepsilon_l^2} + 2\mathfrak{c}_5^l \frac{1}{m} + \frac{\mathfrak{c}_5^l}{Rm\varepsilon_l} \le \frac{1}{4}.$$

Applying Theorem 7, this leads to the bound,

$$\Pr\left(d_{W_{1,\rho}}(\widehat{Q}_{n,m}^{(j)\mathrm{GG}}, \delta_0) \geq 2R\varepsilon_l + \exp(-ml\varepsilon_l^2)\right) \leq \frac{1}{4}.$$

Using Theorem 6, we then conclude:

$$\Pr\left(d_{W_{1,\rho}}(Q_{\mathrm{Met,GG}}^*, \delta_0) > 3\left(2R\varepsilon_l + \exp(-ml\varepsilon_l^2)\right)\right) \leq \left[e^{(1-\kappa)\psi\left(\frac{1/2-\kappa}{1-\kappa}, \frac{1}{4}\right)}\right]^{-m}.$$

This implies that the Wasserstein distance between $Q_{\mathrm{Met,GG}}^*$ and δ_0 satisfies,

$$d_{W_{1,\rho}}(Q_{\mathrm{Met,GG}}^*, \delta_0) \leq 3\left(2R\varepsilon_l + \exp(-ml\varepsilon_l^2)\right),$$

with probability at least $1 - \left[e^{(1-\kappa)\psi\left(\frac{1/2-\kappa}{1-\kappa}, \frac{1}{4}\right)}\right]^{-m}$. Finally, noting the relationship between the Wasserstein distance and the Euclidean norm, we have

$$\|\theta_{\mathrm{GG}}^* - \theta_0\|_2 \leq d_{W_{1,\rho}}(Q_{\mathrm{Met,GG}}^*, \delta_0).$$

This establishes the finite-sample confidence bound for $\|\theta_{\mathrm{GG}}^* - \theta_0\|_2$, concluding the proof for θ_{GG}^*. The proof for θ_{MF}^* is entirely analogous. This completes the proof of part *b*) and we conclude the result. ■

Appendix E. Auxiliary Lemmas

Lemma 13 *Let Θ be a measurable space. Then for any two distributions $Q_0, \Pi_0 \in \mathcal{P}(\Theta)$ and any measurable function $\mathrm{F} : \Theta \mapsto \mathbb{R}$,*

$$Q_0[\ \mathrm{F}] \leq \mathrm{KL}\left(Q_0, \Pi_0\right) + \log\left(\Pi_0\left[\mathrm{e}^{\mathrm{F}}\right]\right).$$

In particular, for any measurable subset $\Theta' \subset \Theta$ and positive constant $v > 0$,

$$Q_0\left(\Theta'\right) \leq \frac{1}{v}\left\{\mathrm{KL}\left(Q_0, \Pi_0\right) + \mathrm{e}^{v}\Pi_0\left(\Theta'\right)\right\}$$

Proof Consider the case Q_0 is not absolutely continuous with respect to Π_0 then $\mathrm{KL}\left(Q_0, \Pi_0\right) = \infty$, so the result trivially holds. Now assume otherwise. Otherwise, recall the following well-known duality formula (see Lemma 2.2 of Alquier and Ridgway (2020)),

$$\log\left(\Pi_0\left[\mathrm{e}^{\mathrm{F}}\right]\right) = \sup_{Q' \ll \Pi_0}\left[Q'[\mathrm{F}] - \mathrm{KL}\left(Q', \Pi_0\right)\right],$$

from where the first assertion is directly followed. For the proof of the last part of the lemma, we let $F(\boldsymbol{\theta}) := v\mathbf{1}\left(\boldsymbol{\theta} \in \Theta'\right)$. Then we have

$$\mathrm{e}^{v}\Pi_0\left(\Theta'\right) \geq \log\left(1 + \mathrm{e}^{v}\Pi_0\left(\Theta'\right)\right) \geq \log\left(\int \mathrm{e}^{\mathrm{F}(\boldsymbol{\theta})}\mathrm{d}\Pi_0(\boldsymbol{\theta})\right),$$

which completes the proof. ■

Appendix F. Proof of Proposition 1

Proof For any distribution $Q \in \mathcal{Q}$, we have the following series of inequalities

$$
\begin{aligned}
&\mathrm{P}_0^{(l)}[\mathrm{KL}\left(Q, \Pi_{n,m}\left(\cdot \mid G_n\right)\right)] \\
&\quad= \int \mathrm{P}_0^{(l)}\left[\log\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)\mathrm{d}Q(\boldsymbol{\theta})}{\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m \mathrm{d}\Pi(\boldsymbol{\theta})}\right)\right]\mathrm{d}Q(\boldsymbol{\theta}) \\
&\quad= \mathrm{KL}\left(Q,\Pi\right) + \int \mathrm{P}_0^{(l)}\left[\log\left(\frac{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}{\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m}\right)\right]\mathrm{d}Q(\boldsymbol{\theta}) \\
&\quad+ \mathrm{P}_0^{(l)}\left[\log\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}\right)\right] \\
&\quad= \mathrm{KL}\left(Q,\Pi\right) + m\int \mathrm{P}_0^{(l)}\left[\log\left(\frac{\prod_{i\in G_j} p_0(X_i)}{\prod_{i\in G_j} p_\theta(X_i)}\right)\right]\mathrm{d}Q(\boldsymbol{\theta}) \\
&\quad+ \mathrm{P}_0^{(l)}\left[\log\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}\right)\right] \\
&\quad= \mathrm{KL}\left(Q,\Pi\right) + mQ\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)\right] + \mathrm{P}_0^{(l)}\left[\log\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}\right)\right].
\end{aligned}
$$

Then, we observe that by Jensen Inequality

$$
\mathrm{P}_0^{(l)}\left[\log\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}\right)\right] \leq \log\left[\mathrm{P}_0^{(l)}\left(\frac{\int_{\Theta}\left(\prod_{i\in G_j} p_\theta(X_i)\right)^m d\Pi(\theta)}{\left(\prod_{i\in G_j} p_0(X_i)\right)^m}\right)\right] \leq 0,
$$

and in consequence,

$$
\mathrm{P}_0^{(l)}[\mathrm{KL}\left(Q, \Pi_{n,m}\left(\cdot \mid G_n\right)\right)] \leq \mathrm{KL}\left(Q,\Pi\right) + mQ\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)\right].
$$

Thus, by the definition of $\widehat{Q}_n$,

$$
\begin{aligned}
\mathrm{P}_0^{(l)}\left[\mathrm{KL}\left(\widehat{Q}_{n,m}^{(j)}, \Pi_{n,m}\left(\cdot \mid G_j\right)\right)\right] &= \mathrm{P}_0^{(l)}\left[\inf_{Q\in\mathcal{Q}} \mathrm{KL}\left(Q, \Pi_{n,m}\left(\cdot \mid G_j\right)\right)\right] \\
&\leq \inf_{Q\in\mathcal{Q}} \mathrm{P}_0^{(l)}\left[\mathrm{KL}\left(Q, \Pi_{n,m}\left(\cdot \mid G_j\right)\right)\right] \\
&\leq \inf_{Q\in\mathcal{Q}}\left\{\mathrm{KL}\left(Q,\Pi\right) + mQ\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)}, \mathrm{P}_\theta^{(l)}\right)\right]\right\},
\end{aligned}
$$

which proves the first desired result. The second assertion. follows from

$$\inf_{Q\in\mathcal{Q}}\left\{\mathrm{KL}\left(Q,\Pi\right)+mQ\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)},\mathrm{P}_\theta^{(l)}\right)\right]\right\}\leq \mathrm{KL}\left(Q_{n,m}^{(j)*},\Pi\right)+mQ_{n,m}^{(j)*}\left[\mathrm{KL}\left(\mathrm{P}_0^{(l)},\mathrm{P}_\theta^{(l)}\right)\right],$$

and Assumption 1. ■

References

P. Alquier and J. Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475 – 1497, 2020. doi: 10.1214/19-AOS1855. URL https://doi.org/10.1214/19-AOS1855.

P. C. Álvarez-Esteban, E. Del Barrio, J. 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.

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

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

G. E. Box and G. C. Tiao. A bayesian approach to some outlier problems. *Biometrika*, 55(1): 119–129, 1968.

J. Christmas and R. Everson. Robust autoregression: Student-t innovations using variational bayes. *IEEE Transactions on Signal Processing*, 59(1):48–57, 2011. doi: 10.1109/TSP.2010.2080271.

F. Futami, I. Sato, and M. Sugiyama. Variational inference based on robust divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 813–822. PMLR, 2018.

S. Ghosal and A. Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

S. Ghosal, J. K. Ghosh, and A. W. Van Der Vaart. Convergence rates of posterior distributions. *Annals of Statistics*, pages 500–531, 2000.

R. Giordano, T. Broderick, and M. I. Jordan. Covariances, robustness, and variational bayes. *Journal of machine learning research*, 19(51):1–49, 2018.

D. Hsu and S. Sabato. Loss minimization and parameter estimation with heavy tails. *Journal of Machine Learning Research*, 17(18):1–40, 2016. URL http://jmlr.org/papers/v17/14-273.html.

P. J. Huber. *Robust Statistics*, pages 1248–1251. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-04898-2. doi: 10.1007/978-3-642-04898-2_594. URL https://doi.org/10.1007/978-3-642-04898-2_594.

B. Jin. A variational bayesian method to inverse problems with impulsive noise. *Journal of Computational Physics*, 231(2):423–435, 2012. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2011.09.009. URL https://www.sciencedirect.com/science/article/pii/S0021999111005298.

Judelo. gmmot. https://github.com/judelo/gmmot, 2024. GitHub repository.

A. C. Kapourani. Variational mixture of gaussians. https://rpubs.com/cakapourani/variational_bayes_gmm, 2019. School of Informatics, University of Edinburgh, UK.

H. J. W. Koh, D. Gašević, D. Rankin, S. Heritier, M. Frydenberg, and S. Talic. Variational bayes machine learning for risk adjustment of general outcome indicators with examples in urology. *npj Digital Medicine*, 7(1):249, 2024.

M. Komodromos, E. O. Aboagye, M. Evangelou, S. Filippi, and K. Ray. Variational bayes for high-dimensional proportional hazards models with applications within gene expression. *Bioinformatics*, 38(16):3918–3926, 2022.

J. Law. Robust statistics—the approach based on influence functions, 1986.

M. Lerasle and R. I. Oliveira. Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*, 2011.

W. Li, M. Li, L. Zuo, H. Chen, and Y. Wu. Real aperture radar forward-looking imaging based on variational bayesian in presence of outliers. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–13, 2022. doi: 10.1109/TGRS.2022.3203807.

X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

L. Lin, W. Shi, J. Ye, and J. Li. Multisource single-cell data integration by maw barycenter for gaussian mixture models. *Biometrics*, 2022.

F. Mahdisoltani. Natural gradient variational inference with gaussian mixture models. *arXiv preprint arXiv:2111.08002*, 2021.

S. Minsker. Geometric median and robust estimation in banach spaces. 2015.

S. Minsker, S. Srivastava, L. Lin, and D. Dunson. Scalable and robust bayesian inference via the median posterior. In *International conference on machine learning*, pages 1656–1664. PMLR, 2014.

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

A. Nazaret and D. Blei. Variational inference for infinitely deep neural networks. In *International Conference on Machine Learning*, pages 16447–16461. PMLR, 2022.

I. Ohn and L. Lin. Adaptive variational bayes: Optimality, computation and applications. *The Annals of Statistics*, 52(1):335–363, 2024.

M. Peruzzi and D. B. Dunson. Spatial multivariate trees for big data bayesian regression. *Journal of Machine Learning Research*, 23(17):1–40, 2022.

A. Raj, M. Stephens, and J. K. Pritchard. faststructure: variational inference of population structure in large snp data sets. *Genetics*, 197(2):573–589, 2014.

K. Ray and B. Szabó. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281, 2022.

D. Sen, T. Papamarkou, and D. Dunson. Bayesian neural networks and dimensionality reduction. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*, pages 188–209. Chapman and Hall/CRC, 2024.

C. G. Small. A survey of multidimensional medians. *International Statistical Review/Revue Internationale de Statistique*, pages 263–277, 1990.

S. Srivastava, C. Li, and D. B. Dunson. Scalable bayes via barycenter in wasserstein space. *Journal of Machine Learning Research*, 19(8):1–35, 2018.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

J. Wang, Y. Tang, M. Nguyen, and I. Altintas. A scalable data science workflow approach for big data bayesian network learning. In *2014 IEEE/ACM International Symposium on Big Data Computing*, pages 16–25. IEEE, 2014.

X. Wang, F. Guo, K. A. Heller, and D. B. Dunson. Parallelizing mcmc with random partition trees. *Advances in neural information processing systems*, 28, 2015.

Y. Wang and D. M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

W. H. Wong, X. Shen, et al. Probability inequalities for likelihood ratios and convergence rates of sieve mles. *The Annals of Statistics*, 23(2):339–362, 1995.

Y. Yang, D. Pati, and A. Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2):886–905, 2020.

K. You. *SBmedian: Scalable Bayes with Median of Subset Posteriors*, 2021. URL https://CRAN.R-project.org/package=SBmedian. R package version 0.1.1.

S. Zabad, S. Gravel, and Y. Li. Fast and accurate bayesian polygenic risk modeling with variational inference. *The American Journal of Human Genetics*, 110(5):741–761, 2023.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207, 2020.

O. Zobay. Variational bayesian inference with gaussian-mixture approximations. 2014.