Summarizing Bayesian Nonparametric Mixture Posterior - Sliced Optimal Transport Metrics for Gaussian Mixtures

Khai Nguyen

Department of Statistics and Data Sciences, University of Texas at Austin

and

Peter Mueller

Department of Mathematics, University of Texas at Austin

November 26, 2024

Abstract

Existing methods to summarize posterior inference for mixture models focus on identifying a point estimate of the implied random partition for clustering, with density estimation as a secondary goal (Wade and Ghahramani, 2018; Dahl et al., 2022). We propose a novel approach for summarizing posterior inference in nonparametric Bayesian mixture models, prioritizing density estimation of the mixing measure (or mixture) as an inference target. One of the key features is the model-agnostic nature of the approach, which remains valid under arbitrarily complex dependence structures in the underlying sampling model. Using a decision-theoretic framework, our method identifies a point estimate by minimizing posterior expected loss. A loss function is defined as a discrepancy between mixing measures. Estimating the mixing measure implies inference on the mixture density and the random partition. Exploiting the discrete nature of the mixing measure, we use a version of sliced Wasserstein distance. We introduce two specific variants for Gaussian mixtures. The first, mixed sliced Wasserstein, applies generalized geodesic projections on the product of the Euclidean space and the manifold of symmetric positive definite matrices. The second, sliced mixture Wasserstein, leverages the linearity of Gaussian mixture measures for efficient projection.

Keywords: Random partition models; Density estimation; Cluster estimation.

1 Introduction

We propose an approach to summarize the posterior distribution on the random mixing measure in Bayesian nonparametric (BNP) mixture models (Ghosal and van der Vaart, 2017). By focusing on the mixing measure, the method fills a gap in the existing literature which focuses on summarizing the posterior distribution of implied random partitions of experimental units (Wade and Ghahramani, 2018; Dahl et al., 2022). The proposed approach requires only posterior Monte Carlo samples of random mixing measures as input and provides a well-defined summary of these measures. Importantly, the approach is model-agnostic, accommodating any prior distributions on the random mixing measures, including those with dependent structures such as dependent Dirichlet process (Quintana et al., 2022; MacEachern, 1999). Using the resulting point estimate of the mixing measure, if desired, one can derive an estimate of the density function and the partition.

Most approaches to posterior summarization for BNP mixture models assume a context of density estimation, conditioning on a sample $\{Y_1, \ldots, Y_n\}$ from a mixture. Latent indicators c_i, $i = 1, \ldots, n$ that link the data with atoms in the mixing measure define a random partition $\rho = \{S_1, \ldots, S_K\}$ of $[n] := \{1, \ldots, n\}$ into subsets S_k of matching c_i. Methods then aim to summarize $p(\rho \mid Y_1, \ldots, Y_n)$, or equivalently, $p(\boldsymbol{c} \mid Y_1, \ldots, Y_n)$ for $\boldsymbol{c} = (c_1, \ldots, c_n)$.

Proceeding under a decision-theoretic framework involves minimizing the posterior expectation of a chosen loss function to define the Bayes rule:

$$\hat{\rho}^\star = \arg \min_{\hat{\rho}} \mathbb{E}[\mathcal{L}(\rho, \hat{\rho}) \mid Y_1, \ldots, Y_n] \quad \text{or} \quad \hat{\boldsymbol{c}}^\star = \arg \min_{\hat{\boldsymbol{c}}} \mathbb{E}[\mathcal{L}(\boldsymbol{c}, \hat{\boldsymbol{c}}) \mid Y_1, \ldots, Y_n], \qquad (1)$$

where $\mathcal{L}$ is a loss function that can be expressed using both the partition or the cluster label notation. Lau and Green (2007) propose the use of Binder loss (Binder, 1978). Wade and Ghahramani (2018) use variational information (VI) loss (Meilă, 2007) as an

alternative. Other information-based distances include normalized variational information (NVI), information distance (ID), normalized information distance (NID) (Vinh et al., 2009), and one minus adjusted Rand index (omARI) (Rand, 1971). In addition to selecting the loss function, various search algorithms have been proposed to solve the optimization problem in (1), including binary integer programming (Lau and Green, 2007), greedy search based on the Hasse diagram (Wade and Ghahramani, 2018), the R & F algorithm (Rastelli and Friel, 2018), and the SALSO algorithm (Dahl et al., 2022).

Even when the primary inference target is a posterior summary of ρ, density estimation (for the mixture or the mixing measure) can still be reported in a second step. However, in applications where the density function plays a crucial role, such as anomaly detection and data generation, it may be more appropriate to focus directly on the density function. Consider then a general BNP mixture model: $Y_1, \ldots, Y_n \overset{i.i.d}{\sim} F, \quad F = f * G, \quad G \sim p(G),$ where f is a kernel and $*$ denotes the convolution operator, we propose to find a summary for $p(G|Y_1, \ldots, Y_n)$. From a point estimate $\widehat{G}$, one can directly obtain a point estimate of the density $\widehat{F}$ by convolving it with the density kernel f. And if desired, one can obtain a point estimate $\widehat{\rho}$ (or cluster labels $\widehat{c}$) induced by $\widehat{G}$. For the sake of easy exposition, in the upcoming discussion we assume i.i.d. sampling. But we do so without loss of generality. The only assumption is that posterior inference is provided as a posterior Monte Carlo sample of G. The details of the sampling model or prior model can be arbitrary.

We continue to use a decision-theoretic approach by minimizing a posterior expected loss. However, targeting G, we require a loss function that quantifies the discrepancy between two measures. Given the discrete nature of the random measure G, optimal transport distances (Villani, 2009; Peyré et al., 2019) are a natural choice in this context. Specifically, we choose sliced Wasserstein (SW) distance (Rabin et al., 2012; Bonneel et al., 2015) due to its computational and statistical scalability with respect to the number of support points. With a near-linear time complexity in the number of support points, the

SW distance facilitates efficient and accurate truncation of mixing measures.

For the common special case of Gaussian mixture models, we introduce two novel versions of SW, working with measures supported on the product of the Euclidean manifold and the manifold of symmetric positive definite (SPD) matrices. The first variant, called mixed SW (Mix-SW), uses generalized geodesic projection instead of the conventional linear projection used in standard SW. Consequently, Mix-SW preserves more geometric information compared to SW with a vectorization approach. The second variant, named sliced mixture Wasserstein (SMix-W), compares mixing measures by evaluating the induced mixture of Gaussian measures. By leveraging the linearity properties of mixture of Gaussian measures, SMix-W achieves a reduction in projection complexity compared to traditional SW while still being geometrically meaningful. Finally, we discuss basic properties of the proposed distances including boundedness and metricity.

The remainder of the article is organized as follows: In Section 2, we introduce the approach for summarizing the posterior of random mixing measures. Section 3 presents the two novel distances for Gaussian mixing measures, accompanied by a discussion of their theoretical and computational properties. In Section 4, we conduct an empirical analysis using simulated data and the Old Faithful Geyser dataset, employing a truncated Dirichlet Process Gaussian mixture model to assess proposed approach in both clustering and density estimation. Additional materials, including technical proofs, are provided in the appendices.

2   Point estimation of random mixing measures

In this section, we present a new approach for obtaining a point estimate of random mixing measures. As discussed, without loss of generality we consider the following Bayesian nonparametric (BNP) mixture model as an example: $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F, \quad F = f * G, \quad G \sim p(G)$, where f is a kernel and $p(G)$ denotes a prior on the random mixing measure G. Our

objective is to report a point estimate $\widehat{G}$. The point estimate can then be used for any downstream data analysis. We first define the problem in Section 2.1 and then discuss current options for optimal transport distances that can be used as the loss function in Section 2.2.

2.1 Problem Setup

Under a decision-theoretic framework, the point estimate of the mixing measure $\widehat{G}^\star$ minimizes posterior expected loss:

$$\widehat{G}^\star = \arg\min_{\widehat{G}} \mathbb{E}[\mathcal{D}(G, \widehat{G}) \mid Y_1, \ldots, Y_n], \tag{2}$$

where $\mathcal{D} : \mathcal{P}(\Theta) \times \mathcal{P}(\Theta) \to \mathbb{R}_+$ represents a distance on the space of measures supported on Θ, which is the support set of the mixing measure G. Since the posterior is usually intractable, the expectation in (2) is approximated using Markov chain Monte Carlo (MCMC) methods. Given M posterior Monte Carlo samples, $G_1, \ldots, G_M \sim p(G \mid Y_1, \ldots, Y_n)$, the optimization problem can be approximated as

$$\widehat{G}^\star = \arg\min_{\widehat{G}} \frac{1}{M} \sum_{m=1}^{M} \mathcal{D}(G_m, \widehat{G}). \tag{3}$$

We use a simple greedy procedure to solve the optimization problem in (3). Specifically, we construct an $M \times M$ matrix, where the entry at row i and column j represents the distance between the i-th and j-th posterior samples, i.e., $\mathcal{D}(G_i, G_j)$. We then identify the index $i^\star$ that minimizes the average distance across columns, $\frac{1}{M} \sum_{j=1}^{M} \mathcal{D}(G_i, G_j)$, and return $G_{i^\star}$ as the greedy solution. This is equivalent to selecting the best posterior sample from the posterior Monte Carlo sample.

In practice, truncating the mixing measure G can accelerate computation, improve convergence and simplify implementation (Ishwaran and James, 2001). Specifically, we can

use the following truncated version of G: $\bar{G} = \sum_{k=1}^{K} \alpha_k \delta_{\theta_k}$ with $0 < K < \infty$. Ideally, we want to choose the largest feasible value of K to minimize the approximation error introduced by truncation. We then approximate the distance between two measures by the distance between the truncated versions, i.e., $\mathcal{D}(G_1, G_2) \approx \mathcal{D}(\bar{G}_1, \bar{G}_2)$.

A point estimate of the mixing measure $\widehat{G}$, implies a point estimate of the density by convolution of $\widehat{G}$ with the kernel f, $\widehat{F} = f * \widehat{G}$. To estimate the partition, we determine the cluster membership indicator c_i for a data point i by maximum a posteriori (MAP) estimation:

$$\hat{c}_i = \arg \max_k p(c_i = k \mid \widehat{G}, x). \tag{4}$$

Here x generically denotes the observed data. When using truncation, i.e., $\widehat{G} \approx \sum_{k=1}^{K} \hat{w}_k \delta_{\hat{\theta}_k}$, this becomes:

$$\hat{c}_i = \arg \max_k p(c_i = k \mid \hat{w}_1, \ldots, \hat{w}_K, \hat{\theta}_1, \ldots, \hat{\theta}_K, x). \tag{5}$$

While there are alternative methods to obtain a point estimate $\hat{c}$ given the mixing measure, MAP is perhaps the most natural and computationally efficient approach.

In the optimization in (3), after truncation, we are left with two discrete mixing measures that may have disjoint supports. Traditional f-divergences (Ali and Silvey, 1966), such as the Kullback–Leibler (KL) divergence, Jensen–Shannon (JS) divergence, and others, cannot be used directly because they require access to the density ratio, which may be undefined in this context. Consequently, optimal transport metrics become a natural choice in this scenario. Next, we will brifely discuss currently available optimal transport distances.

2.2 Optimal Transport Distances for the loss function

Wasserstein distance. Let $G_1, G_2 \in \mathcal{P}(\Theta)$ and $d : \Theta \times \Theta \to \mathbb{R}^+$ be a ground metric. The Wasserstein-p ($p \geq 1$) distance (Villani, 2009) between two measures G_1 and G_2 is defined as follows:

$$W_d^p(G_1, G_2) = \inf_{\pi \in \Pi(G_1, G_2)} \int_{\Theta \times \Theta} d(x, y) \, d\pi(x, y), \tag{6}$$

where $\Pi(G_1, G_2) = \{\pi \in \mathcal{P}(\Theta \times \Theta) \mid \pi(A, \Theta) = G_1(A), \ \pi(\Theta, B) = G_2(B) \ \forall A, B \subset \Theta\}$ is the set of all transportation plans/couplings. When G_1 and G_2 are discrete measures, i.e., $G_1 = \sum_{i=1}^{K_1} \alpha_i \delta_{x_i}$ and $G_2 = \sum_{j=1}^{K_2} \beta_j \delta_{y_j}$, the Wasserstein distance can be rewritten as:

$$W_d^p(G_1, G_2) = \min_{\pi \in \Gamma(\alpha, \beta)} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} d(x_i, y_j) \pi_{ij}, \tag{7}$$

where the set of transportation plans becomes $\Gamma(\alpha, \beta) = \{\pi \in \mathbb{R}_+^{K_1 \times K_2} \mid \pi \mathbf{1} = \beta, \ \pi^\top \mathbf{1} = \alpha\}$. The computation of Wasserstein distance is often performed using linear programming (Peyré et al., 2019), with a time complexity of $\mathcal{O}((K_1 + K_2)^3 \log(K_1 + K_2))$. Alternatively, it can be approximated using the entropic regularization approach (Cuturi, 2013), which has a time complexity of $\mathcal{O}\left(K_1 K_2 \log(K_1 + K_2)/\epsilon\right)$, where $\epsilon > 0$ is the precision level. Consequently, using Wasserstein distance becomes impractical for large values of K_1 or K_2. Therefore, using the Wasserstein distance might in an undesirable way limit the truncation level K_1 and K_2.

Sliced Wasserstein (SW) Distance. SW distance exploits the availability of a closed-form expression for Wasserstein distance on the real line. Specifically, for two distributions $G_1, G_2 \in \mathcal{P}(\mathbb{R})$, the Wasserstein-$p$ distance with the ground metric defined as $d(x, y) =$

$|x - y|$ is expressed as follows (Peyré et al., 2019):

$$W_p^p(G_1, G_2) = W_d^p(G_1, G_2) = \int_0^1 |\text{CDF}_{G_1}^{-1}(t) - \text{CDF}_{G_2}^{-1}(t)|^p dt, \qquad (8)$$

where $\text{CDF}_{G_1}^{-1}$ and $\text{CDF}_{G_2}^{-1}$ denote the inverse cumulative distribution functions (quantile functions) of G_1 and G_2, respectively. When G_1 and G_2 are discrete measures i.e., $G_1 = \sum_{i=1}^{K_1} \alpha_i \delta_{x_i}$ and $G_2 = \sum_{j=1}^{K_1} \beta_j \delta_{y_j}$ (assumed that supports of two measures are sorted), the inverse CDFs can be defined as

$$\text{CDF}_{G_1}^{-1}(t) = \sum_{i=1}^{K_1} x_i I\{\sum_{j=1}^{i-1} \alpha_j \leq t \leq \sum_{j=1}^{i} \alpha_j\}, \quad \text{CDF}_{G_2}^{-1}(t) = \sum_{j=1}^{K_2} y_j I\{\sum_{i=1}^{j-1} \beta_i \leq t \leq \sum_{i=1}^{j} \beta_i\}$$

Solving (8) in this case has the time complexity of $\mathcal{O}((K_1 + K_2) \log(K_1 + K_2))$.

To leverage the closed-form solution of one-dimensional Wasserstein distance in high-dimensional settings, SW distance (Bonneel et al., 2015) is introduced. The central idea of SW is to randomly project two original measures onto two one-dimensional measures and then computes the expected value of the one-dimensional Wasserstein distance between the two projected measures. The conventional SW employs a linear projection, defined as $P_v(x) = \langle v, x \rangle$ for $v \in \mathbb{S}^{d-1}$, where v represents the projection direction. The sliced Wasserstein-p distance ($p \geq 1$) between G_1 and G_2, using the ground metric $d(x, y) = |x - y|$, is defined as follows:

$$SW_p^p(G_1, G_2) = \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(P_v \sharp G_1, P_v \sharp G_2)], \qquad (9)$$

where $P_v \sharp G_1$ and $P_v \sharp G_2$ are the push-forward measures of G_1 and G_2 through the function P_v, and $\mathcal{U}(\mathbb{S}^{d-1})$ denotes the uniform distribution over the unit hypersphere.

The expectation in sliced Wasserstein (SW) distance is intractable, thus Monte Carlo estimation is often employed to approximate SW. Specifically, let $v_1, \ldots, v_L \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}^{d-1})$

represent the projecting directions. The Monte Carlo estimation of SW is given by:

$$\widehat{SW}_p^p(G_1, G_2) = \frac{1}{L} \sum_{l=1}^{L} W_p^p(P_{v_l} \sharp G_1, P_{v_l} \sharp G_2). \tag{10}$$

The overall time complexity of SW is composed of the time required for sampling projecting directions, the time for applying the projection operator P_v, and the time for computing one-dimensional Wasserstein distances. When G_1 and G_2 are discrete measures supported by K_1 and K_2 atoms, respectively, and L Monte Carlo samples are used, the time complexity of SW is: $\mathcal{O}(Ld + Ld(K_1 + K_2) + L(K_1 + K_2)\log(K_1 + K_2)) = \mathcal{O}(Ld(K_1 + K_2) + L(K_1 + K_2)\log(K_1 + K_2))$, where $\mathcal{O}(Ld)$ accounts for sampling projecting directions, $\mathcal{O}(Ld(K_1 + K_2))$ is for the projection, and $\mathcal{O}(L(K_1 + K_2)\log(K_1 + K_2))$ is for computing L one-dimensional Wasserstein distances. Additionally, the projection complexity of SW is $\mathcal{O}(Ld)$, which corresponds to the memory required for storing the L projecting directions. We can see that SW is very scalable in terms of the numbers K_1 and K_2 of support points. It allows accurate truncation with large K_1 and K_2.

3 Sliced Optimal Transport Distances for Gaussian Mixing Measures

For Gaussian mixing measures G_1 and G_2, we use the parameter space $\Theta = \mathbb{R}^d \times S_d^{++}(\mathbb{R})$, where $S_d^{++}(\mathbb{R})$ is the manifold of all symmetric positive definite matrices. Conventional SW distance can not directly be applied in this context, as it is defined for measures on vector spaces. We first discuss an approach to apply SW using vectorization in Section 3.1. We then propose two novel variants of SW that preserve geometry. In Section 3.2 we start with a new variant of SW based on generalized geodesic projection onto the product of manifolds, which we call Mixed SW (Mix-SW). Finally, we introduce another variant of SW for finite Gaussian mixing measures by comparing their induced mixture measures,

which we call sliced mixture Wasserstein (SMix-W).

3.1   Vectorized Sliced Wasserstein

We are aiming to compare measures on the product of the Euclidean manifold and the manifold of symmetric positive definite (SPD) matrices, denoted as $\Theta = \mathbb{R}^d \times S_d^{++}(\mathbb{R})$, using the Sliced Wasserstein (SW) distance. However, SW is defined on vector spaces. A straightforward approach is to convert measures on $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$ into measures on a vector space. For any $\theta = (\mu, \Sigma) \in \mathbb{R}^d \times S_d^{++}(\mathbb{R})$, we can arrange the entries of Σ to obtain a vector representation, which can then be stacked with μ. For simplicity, we define the transformation $V(\theta) = (\mu, \Sigma^{(1)}, \ldots, \Sigma^{(d)})$, where $\Sigma^{(i)}$ is the i-th row of the matrix Σ. With this transformation, we can redefine SW distance as

$$SW_p^p(G_1, G_2) = \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{d(d+1)-1})}[W_p^p(P_{v\sharp}V\sharp G_1, P_{v\sharp}V\sharp G_2)], \tag{11}$$

for any $G_1, G_2 \in \mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$. Despite the appealing simplicity of this approach, there are two main complications to consider. The first is that vectorization destroys the geometry of the space, which may result in a distance that lacks geometric meaning. In contrast to the Wasserstein distance, where the ground metric can be flexibly designed, the ground metric in SW is constrained to exist in a one-dimensional space. The second issue pertains to the high-dimensional projection direction space, $\mathbb{S}^{d(d+1)-1}$, which may require increased computation and memory to achieve accurate approximations via Monte Carlo estimation. The time complexity of vectorized SW is given by $\mathcal{O}(Ld^2(K_1 + K_2) + L(K_1 + K_2)\log(K_1 + K_2))$ and the projection complexity is $\mathcal{O}(Ld^2)$, both of which are quadratic in dimensions.

3.2 Mixed Sliced Wasserstein

To address the loss of geometric information, we propose a new variant of SW distance using geodesic projection on the product manifolds $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$. In summary, we define a notion of projecting a support point $(\mu, \Sigma) \in \mathbb{R}^d \times S_d^{++}(\mathbb{R})$ onto a curve with the associated velocity vector $V_w = (w_1 A, w_2 A)$ in a way that the projection is easy to evaluate. From the new projection, the desired SW is defined as the expectation of the projected one-dimensional Wasserstein distance under the uniform-law of random curve parameters (v, A, w).

We begin by reviewing some basic definitions relevant to Riemannian manifolds, including the inner product, geodesics, length, geodesic distance, and the exponential map, as detailed in Appendix A.1. Additionally, we review the concept of geodesic projection (Bonet et al., 2024) and explore certain properties of the manifold of symmetric positive definite matrices $S_d^{++}(\mathbb{R})$ (Pennec et al., 2019).

Geodesic Projection Let γ be a curve on the manifold $\mathcal{M}$, and denote $\mathcal{A}$ as the set of all points belonging to that curve. The projection of a point $x \in \mathcal{M}$ onto the curve γ is defined as: $\tilde{P}_\gamma(x) = \arg\min_{y \in \mathcal{A}} d(x, y)$, where d is the geodesic distance. If we constrain γ to be a curve that passes through the origin (denoted as o) with unit velocity v (i.e., $\langle v, v \rangle_o = 1$), then we have $\Gamma = \{\exp_o(tv) \mid t \in \mathbb{R}\}$, where $\exp_o(\cdot)$ is the exponential map at the origin. The coordinates of the projection can be determined by solving:

$$P_{\gamma_{(o,v)}}(x) := P_v(x) = \arg\min_{t \in \mathbb{R}} d(x, \exp_o(tv)). \tag{12}$$

Product Manifold of $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$. From (Pennec et al., 2019), the origin of $S_d^{++}(\mathbb{R})$ is the identity matrix I, and the tangent space is the space of all symmetric matrices. The exponential map is given by $\exp_I(A) = \exp(A) = \sum_{n=0}^\infty \frac{A^n}{n!}$. While there are multiple geodesic distances on $S_d^{++}(\mathbb{R})$, we focus on the Log-Euclidean metric defined as: $d_{LE}(\Sigma_1, \Sigma_2) = \| \log \Sigma_1 - \log \Sigma_2 \|_F$, where $\log X = A$ if $\exp(A) = X$. Since X is a symmet-

ric positive definite matrix, we can use spectral decomposition: $X = Q\Lambda Q^T$, which gives us $\log X = Q\,\mathrm{diag}(\log\lambda_1,\ldots,\log\lambda_d)\,Q^T$. The origin of the manifold $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$ is $o = (0, I)$, where 0 is the d-dimensional zero vector. The exponential map in this manifold is defined as $\exp_o((\mu, \Sigma)) = (\mu, \exp(\Sigma))$. The geodesic distance of the product manifolds is defined as follows (Gu et al., 2019): $d((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + d_{LE}(\Sigma_1, \Sigma_2)^2}$.

Generalized Geodesic Projection. The geodesic projection and the corresponding SW distance for measures on $S_d^{++}(\mathbb{R})$ have been investigated in (Bonet et al., 2023). While the geodesic projection for the product of manifolds was introduced in (Bonet et al., 2024), it has not been explicitly derived for any specific case. In this work, we extend the notion of geodesic projection to a generalized geodesic projection by projecting onto a generalized curve with adjusted velocity vectors for each marginal manifold. This adjustment is essential for achieving the identity of indiscernibles in the subsequently defined SW metric. Furthermore, we demonstrate that the generalized geodesic projection has a closed-form expression on $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$.

Definition 1. *Given a product manifolds $\mathcal{M}_1 \times \mathcal{M}_2$ with the origin $o = (o_1, o_2)$, a generalized curve passing through the origin with the velocity vector $V_w = (w_1 v_1, w_2 v_2)$ with $\langle v_1, v_1\rangle_{o_1} = 1$, $\langle v_2, v_2\rangle_{o_2} = 1$, and $w_1^2 + w_2^2 = 1$, the generalized geodesic projection onto the generalized curve created by V_w is defined as:*

$$P_{V_w}(x) = \arg\min_{t\in\mathbb{R}} d(x, \exp_o(tV_w)), \tag{13}$$

where d is the geodesic distance on the product manifold.

Proposition 1. *The generalized geodesic projection onto a generalized curve passing through the origin with the velocity vector $V_w = (w_1 v, w_2 A)$ ($\|v\|_2^2 = 1, \|A\|_F^2 = 1, w_1^2 + w_2^2 = 1$) on the product manifold of $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$ has the following form: $P_{V_w}((\mu, \Sigma)) = w_1\langle\mu, v\rangle + w_2 Trace(A\log\Sigma)$.*

The proof of Proposition 1 is given in Appendix A.2. It turns out that the generalized geodesic projection is a weighted combination of the geodesic projection on the Euclidean manifold and the geodesic projection on the manifold $S_d^{++}(\mathbb{R})$.

Mixed Sliced Wasserstein. From the generalized geodesic projection, we now can define the mixed Sliced Wasserstein (Mix-SW) distance.

Definition 2. *Given two measures G_1 and G_2 belonging to $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$, $p \geq 1$, the Mixed Sliced Wasserstein distance is defined as follows:*

$$Mix\text{-}SW_p^p(G_1, G_2) = \mathbb{E}_{(w,v,A) \sim \mathcal{U}(\mathbb{S}) \otimes \mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))}[W_p^p(P_{V_w} \sharp G_1, P_{V_w} \sharp G_2)],$$

where $V_w = (w_1 v, w_2 A)$ and $\mathcal{U}(\mathcal{X})$ is the uniform distribution over the set $\mathcal{X}$ e.g., $\mathbb{S}, \mathbb{S}^{d-1}, S_d(\mathbb{R})$.

Mix-SW is similar to hierarchical hybrid sliced Wasserstein (H2SW) (Nguyen and Ho, 2024) i.e., they combine projections from marginals. However, Mix-SW comes from generalized geodesic projection for a specific product of manifold while H2SW comes from randomly combining two general types of Radon transforms. In addition, H2SW is only introduced for the product of the Euclidean manifold and the hypersphere.

Proposition 2. *If $\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_1, \Sigma_1), (\mu_0, \Sigma_0))^p dG_1(\mu_1, \Sigma_1) < \infty$ and $\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_0, \Sigma_0), (\mu_2, \Sigma_2))^p dG_2(\mu_1, \Sigma_1) < \infty$ for any $(\mu_0, \Sigma_0) \in \mathbb{R}^d \times S_d^{++}(\mathbb{R})$ with $d((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + \|\log \Sigma_1 - \log \Sigma_2\|_F^2}$, then $Mix\text{-}SW_p^p(G_1, G_2) < \infty$.*

We first show that Mix-$SW_p^p(G_1, G_2)$ is bounded as long as the expected geodesic distances with respect to G_1 and G_2 to any point (μ_0, Σ_0) are bounded. The proof of Proposition 2 is given in Appendix A.3. Next, we will show that Mix-SW is a valid metric.

Theorem 1. *Mixed Sliced Wasserstein is a valid metric on the space of measures which belong to $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$ and satisfy the constraint in Proposition 2.*

The proof of Theorem 1 is given in Appendix A.4 which extends the technique of the proofs in (Bonnotte, 2013; Nadjahi et al., 2020; Bonet et al., 2023) with the usage of the generalized geodesic projection.

Corollary 1. *By the identifiability of finite mixture of Gaussians (Proposition 2 in Yakowitz and Spragins (1968)), Mix-SW is also a metric on space of finite mixture of Gaussians.*

Corollary 1 suggests that Mix-SW can also be used to compare two finite mixtures of Gaussians created by two mixing measures on $\mathbb{R}^d \times S_d^{++}(\mathbb{R})$.

On the computational side, the expectation in Definition 2 is intractable. We therefore employ Monte Carlo estimation to approximate Mix-SW. Specifically, we sample $(w_1, v_1, A_1), \ldots, (w_L, v_L, A_L) \overset{\text{i.i.d}}{\sim} \mathcal{U}(\mathbb{S}) \otimes \mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))$. The Monte Carlo estimate of Mix-SW is then defined as:

$$\widehat{\text{Mix-}SW}_p^p(G_1, G_2) = \frac{1}{L} \sum_{l=1}^{L} W_p^p(P_{V_{w,l}} \sharp G_1, P_{V_{w,l}} \sharp G_2), \tag{14}$$

where $V_{w,l} = (w_{l1} v_l, w_{l2} A_l)$. When G_1 and G_2 are discrete measures with K_1 and K_2 supports, respectively, the time complexity of Mix-SW is $\mathcal{O}((K_1 + K_2 + L)d^3 + L(K_1 + K_2)d^2 + L(K_1+K_2)\log(K_1+K_2))$, which arises from sampling $A \sim \mathcal{U}(S_d(\mathbb{R}))$ (see Algorithm 1 in (Bonet et al., 2023)), computing the matrix logarithm, projecting the samples, and solving one-dimensional Wasserstein distances. The projection complexity of Mix-SW is $\mathcal{O}(Ld^2)$ since it requires storing L projection matrices $A_1, \ldots, A_L$.

3.3 Sliced Mixture Wasserstein

Mix-SW compares measures belonging to $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$. But it is not specifically designed for Gaussian mixing measures. To leverage the structure of Gaussian mixing measures, we introduce a variant called the sliced mixture Wasserstein (SMix-W) distance, which compares Gaussian mixing measures via their induced mixture of Gaussian mea-

sures. SMix-W is inspired by the Mixture Wasserstein distance (Delon and Desolneux, 2020), which is specifically for comparing mixtures of Gaussian measures.

Mixture Wasserstein distance. Given two discrete measures G_1 and G_2 belonging to $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$, and a Gaussian kernel $f(x \mid \mu, \Sigma)$, we define F_1 and F_2 as the corresponding mixtures of Gaussian measures, i.e., $F_1 = f * G_1$ and $F_2 = f * G_2$, where $*$ denotes the standard convolution operation. The Mixture Wasserstein (MW) distance (Delon and Desolneux, 2020) is defined as:

$$\text{MW}_2^2(F_1, F_2) = \inf_{\pi \in \Pi(F_1, F_2) \cap \text{GMM}_{2d}(\infty)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|_2^2 \, d\pi(x, y), \tag{15}$$

where $\text{GMM}_{2d}(\infty)$ denotes the set of all finite Gaussian mixture distributions in $2d$ dimensions. When $G_1 = \sum_{i=1}^{K_1} \alpha_i \delta_{(\mu_{1i}, \Sigma_{1i})}$ and $G_2 = \sum_{j=1}^{K_2} \beta_j \delta_{(\mu_{2j}, \Sigma_{2j})}$, the Mixture Wasserstein distance simplifies to: $\min_{\eta \in \Gamma(\alpha, \beta)} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \eta_{ij} W_2^2 \left(\mathcal{N}(\mu_{1i}, \Sigma_{1i}), \mathcal{N}(\mu_{2j}, \Sigma_{2j}) \right)$, Using the closed-form expression of the Wasserstein-2 distance between two Gaussian distributions, we can rewrite this as: $\min_{\eta \in \Gamma(\alpha, \beta)} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \eta_{ij} \left(\|\mu_{1i} - \mu_{2j}\|_2^2 + \text{Tr}(\Sigma_{1i}) + \text{Tr}(\Sigma_{2j}) \right.$
$\left. -\text{Tr}\left((\Sigma_{1i}^{1/2} \Sigma_{2j} \Sigma_{1i}^{1/2})^{1/2} \right) \right)$.

One-dimensional Mixture Wasserstein distance. In preparation of the upcoming definition of sliced MW by comparing one-dimensional projections of mixtures of normals, we note the special case of one-dimensional MW distance. When $G_1 = \sum_{i=1}^{K_1} \alpha_i \delta_{(\mu_{1i}, \sigma_{1i}^2)}$ and $G_2 = \sum_{j=1}^{K_2} \beta_j \delta_{(\mu_{2j}, \sigma_{2j}^2)}$ are one-dimensional mixtures of Gaussians, we have:

$$\text{MW}_2^2(F_1, F_2) = \min_{\gamma \in \Gamma(\alpha, \beta)} \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \gamma_{i,j} \left((\mu_{1i} - \mu_{2j})^2 + (\sigma_{1i} - \sigma_{2j})^2 \right) =$$

$$= W_2^2 \left((\text{Id}, \sqrt{\cdot}) \sharp G_1, (\text{Id}, \sqrt{\cdot}) \sharp G_2 \right), \tag{16}$$

which implies that the Mixture Wasserstein distance between one-dimensional Gaussian mixtures behaves like a two-dimensional Wasserstein-2 distance on the mixing measures, with a square root scaling applied to the variances. It is important to note that the one-dimensional MW distance does not offer computational advantages, as it is not equivalent to a one-dimensional Wasserstein distance.

Linear projection of mixture of Gaussians. When $G_1 = \sum_{i=1}^{K_1} \alpha_i \delta_{(\mu_{1i}, \Sigma_{1i})}$ and $G_2 = \sum_{j=1}^{K_2} \beta_j \delta_{(\mu_{2j}, \Sigma_{2j})}$, we have $F_1 = f * G_1 := \sum_{i=1}^{K_1} \alpha_i \mathcal{N}(\mu_{1i}, \Sigma_{1i})$ and $F_2 = f * G_2 := \sum_{j=1}^{K_2} \beta_j \mathcal{N}(\mu_{2j}, \Sigma_{2j})$. For a vector $v \in \mathbb{R}^d$ and $P_v(x) = \langle x, v \rangle$, we have $P_v \sharp F_1 := \sum_{i=1}^{K_1} \alpha_i \mathcal{N}(\langle v, \mu_{1i} \rangle, v^\top \Sigma_{1i} v)$ and $P_v \sharp F_2 := \sum_{j=1}^{K_2} \beta_j \mathcal{N}(\langle v, \mu_{2j} \rangle, v^\top \Sigma_{2j} v)$, which are two one-dimensional Gaussian mixtures with the mixing measures $P_v' \sharp G_1$ and $P_v' \sharp G_2$ with $P_v'(\mu, \Sigma) = (\langle v, \mu \rangle, v^\top \Sigma v)$.

Sliced Mixture Wasserstein. After applying linear projection to the mixture Gaussians (or P_v' on the Gaussian mixing measure), we can use one-dimensional MW in (16) to compare them i.e.,

$$\mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{d-1})}[\mathrm{MW}_2^2(P_v \sharp F_1, P_v \sharp F_2)] = \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{d-1})}[W_2^2((Id, \sqrt{}) \sharp P_v' \sharp G_1, (Id, \sqrt{}) \sharp P_v' \sharp G_2)]. \quad (17)$$

However, MW does not have a closed-form expression, as discussed. Since MW is equivalent to the Wasserstein-2 distance between mixing measures, we can replace it with the SW distance to achieve computational benefits, as SW is equivalent to the Wasserstein distance under a mild assumption (Bonnotte, 2013). This replacement leads to a novel variant of sliced Wasserstein distance for mixtures of Gaussians and their mixing measures.

Definition 3. *Given two finite discrete measures G_1 and G_2 belonging to $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$, $p \geq 1$, and $P_{v,w}(\mu, \Sigma) = w_1 \langle v, \mu \rangle + w_2 \log(\sqrt{v^\top \Sigma v})$, the sliced mixture Wasserstein (SMix-*

W) is defined as follows:

$$SMix\text{-}W_p^p(G_1, G_2) = \mathbb{E}_{(w,v)\sim\mathcal{U}(\mathbb{S})\otimes\mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(P_{v,w}\sharp G_1, P_{v,w}\sharp G_2)],$$

$\mathcal{U}(\mathcal{X})$ is the uniform distribution over the set $\mathcal{X}$ e.g., $\mathbb{S}, \mathbb{S}^{d-1}$.

We can rewrite

$$\text{SMix-}W_2^2(G_1, G_2) = \mathbb{E}_{v\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[SW_2^2\left((\text{Id}, \log\circ\sqrt{\cdot})\sharp P_v'\sharp G_1, (\text{Id}, \log\circ\sqrt{\cdot})\sharp P_v'\sharp G_2\right)\right]$$

which replaces MW in (17) with the SW distance, incorporating a logarithmic transformation to adjust the standard deviation as a geodesic projection. Compared to SW and Mix-SW, the projection space of SMix-W is smaller, specifically $\mathbb{S}^{d-1} \times \mathbb{S}$, as it utilizes a single projecting direction v for both the mean μ and the covariance matrix Σ.

Proposition 3. *If* $\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_1, \Sigma_1), (\mu_0, \Sigma_0))^p dG_1(\mu_1, \Sigma_1) < \infty$ *and* $\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_0, \Sigma_0), (\mu_2, \Sigma_2))^p dG_2(\mu_1, \Sigma_1) < \infty$ *for any* $(\mu_0, \Sigma_0) \in \mathbb{R}^d \times S_d^{++}(\mathbb{R}))$ *with* $d((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \sqrt{\|\mu_1 - \mu_2\|_2^2 + 0.25\log(\lambda_{max}(\Sigma_1, \Sigma_2))^2}$ *($\lambda_{max}(\Sigma_1, \Sigma_2)$ is the largest eigenvalue of the generalized problem $\Sigma_1 v = \lambda\Sigma_2 v$), then $SMix\text{-}W_p^p(G_1, G_2) < \infty$.*

The proof of Proposition 3 is given in Appendix A.5. After showing the Smix-W is bounded, we show that SMix-W is a valid metric for discrete measures on $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$.

Theorem 2. *SMix-W is a valid metric on the space of finite discrete measures on $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$ which satisfy the constraint in Proposition 3.*

The proof of Theorem 2 is given in Appendix A.6. It is worth noting that SMix-W is metric for only finite discrete measures on $\mathcal{P}(\mathbb{R}^d \times S_d^{++}(\mathbb{R}))$ since the proof of identity of indiscernibles of SMix-W relies on the identifiability of finite mixture of Gaussians (Proposition 2 in Yakowitz and Spragins (1968)). In addition, from the identifiability, SMix-W is also a metric between finite mixtures of Gaussians measures.

On the computational side, the expectation in Definition 3 is also intractable. As before, we use Monte Carlo estimation to approximate the value of SMix-W. In particular, we sample $(w_1, v_1), \ldots, (w_L, v_L) \overset{i.i.d.}{\sim} \mathcal{U}(\mathbb{S}) \otimes \mathcal{U}(\mathbb{S}^{d-1})$. The Monte Carlo estimation of SMix-W is defined as follows:

$$\widehat{\text{SMix-W}}_p^p(G_1, G_2) = \frac{1}{L} \sum_{l=1}^{L} W_p^p(P_{v_l, w_l} \sharp G_1, P_{v_l, w_l} \sharp G_2). \tag{18}$$

When G_1 and G_2 are discrete measures with K_1 and K_2 supports, respectively, the time complexity of SMix-W is $\mathcal{O}(L(K_1 + K_2 + 2)d^2 + L(K_1 + K_2) \log(K_1 + K_2))$, due to the computation of the projections and the solving of one-dimensional Wasserstein distances. The projection complexity of SMix-W is $\mathcal{O}(Ld)$, as it only requires storing L projections of $(w_1, v_1), \ldots, (w_L, v_L)$. We observe that SMix-W reduces the time complexity from $\mathcal{O}(d^3)$ of Mix-W to $\mathcal{O}(d^2)$ by exploiting the linearity of mixtures of Gaussians. Compared to vectorized SW and Mix-SW with $\mathcal{O}(d^2)$ in projection complexity, SMix-SW has a better projection complexity of $\mathcal{O}(d)$, making it more scalable with respect to the number of dimensions. Furthermore, a lower-dimensional projection space for SMix-W may lead to a reduced number of projections L required for a good approximation.

4   Empirical Analysis

We assess clustering and density estimation under two alternative summarization approaches: summarizing random partitions versus the proposed novel method of summarizing random mixing measures. For the first approach, we utilize the SALSO package (Dahl et al., 2022) with its greedy search algorithm to obtain point estimates of the random partition using Binder loss, VI loss, and omARI loss. For the proposed new methods, we employ vectorized SW, Mix-SW, and SMix-W (all are approximated with $L = 100$ projections) to obtain point estimates of the random mixing measures. For evaluation, we use Binder loss,

VI loss, and omARI loss to assess clustering performance, while we employ approximated Total Variation (TV) and approximated SW (with $L = 1000$ projections) computed on a grid over the data space to evaluate the density estimates.

Model and data. The proposed approach remains valid for any BNP mixture models with arbitrary prior structures and sampling models. For easier exposition and to facilitate the comparison we work with the conjugate truncated Dirichlet Process Gaussian mixture models (Ishwaran and James, 2001). We assume

$$\beta_1, \ldots, \beta_K \mid \alpha \sim Beta(1, \alpha), \qquad w_k = \beta_k \prod_{j=1}^{k-1}(1 - \beta_j), \tag{19}$$

$$z_i | w_1, \ldots, w_K \sim Multinomial(w_1, \ldots, w_K), \quad (\mu_i, \Sigma_i) \mid \mu_0, \lambda, \Psi, \nu \sim \mathcal{NIW}(\mu_0, \lambda, \Psi, \nu),$$

$$y_i \mid \mu_{1:K}, \Sigma_{1:K}, z_i = k \sim \mathcal{N}(\mu_k, \Sigma_k), \quad i = 1, \ldots, n,$$

where $K > 0$ is the truncation level, and $\mathcal{NIW}$ denotes the Normal Inverse Wishart distribution. The inference of the above model can be carried out efficiently by a Blocked Gibbs sampler (Ishwaran and James, 2001), which simulates from the joint distribution $p(\beta_{1:K}, \mu_{1:K}, \Sigma_{1:K}, z_{1:n}, y_{1:n})$. The blocked Gibbs sampler defines transition probabilities defined by sampling from the following complete conditional posterior distributions (1) $p(z_i = k \mid \beta_{1:K}, \mu_{1:K}, \Sigma_{1:K}, y_i) \propto w_k \mathcal{N}(y_i | \mu_k, \Sigma_k)$; (2) $\beta_k \sim Beta(1 + n_k, \alpha + \sum_{j>k} n_j)$; and (3) $\mu_k, \Sigma_k \sim \mathcal{NIW}\left(\frac{\lambda \mu_0 + n_k \bar{y}_k}{\lambda + n_k}, \lambda + n_k, \Psi + \sum_{i|z_i=k}(y_i - \bar{y}_k)(y_i - \bar{y}_k)^\top\right.$
$\left. + \frac{\lambda n_k}{\lambda + n_k}(\bar{y}_k - \mu_0)(\bar{y}_k - \mu_0)^\top, \nu + n_k\right),$

where n_k is the number of members of the k-th clusters given a partition, $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n} y_i \delta_{z_i=k}$ is the mean of cluster k-th. We implement inference for a simulated dataset in Section 4.1, and for the Old Faithful geyser dataset (Azzalini and Bowman, 1990) in Section 4.2.
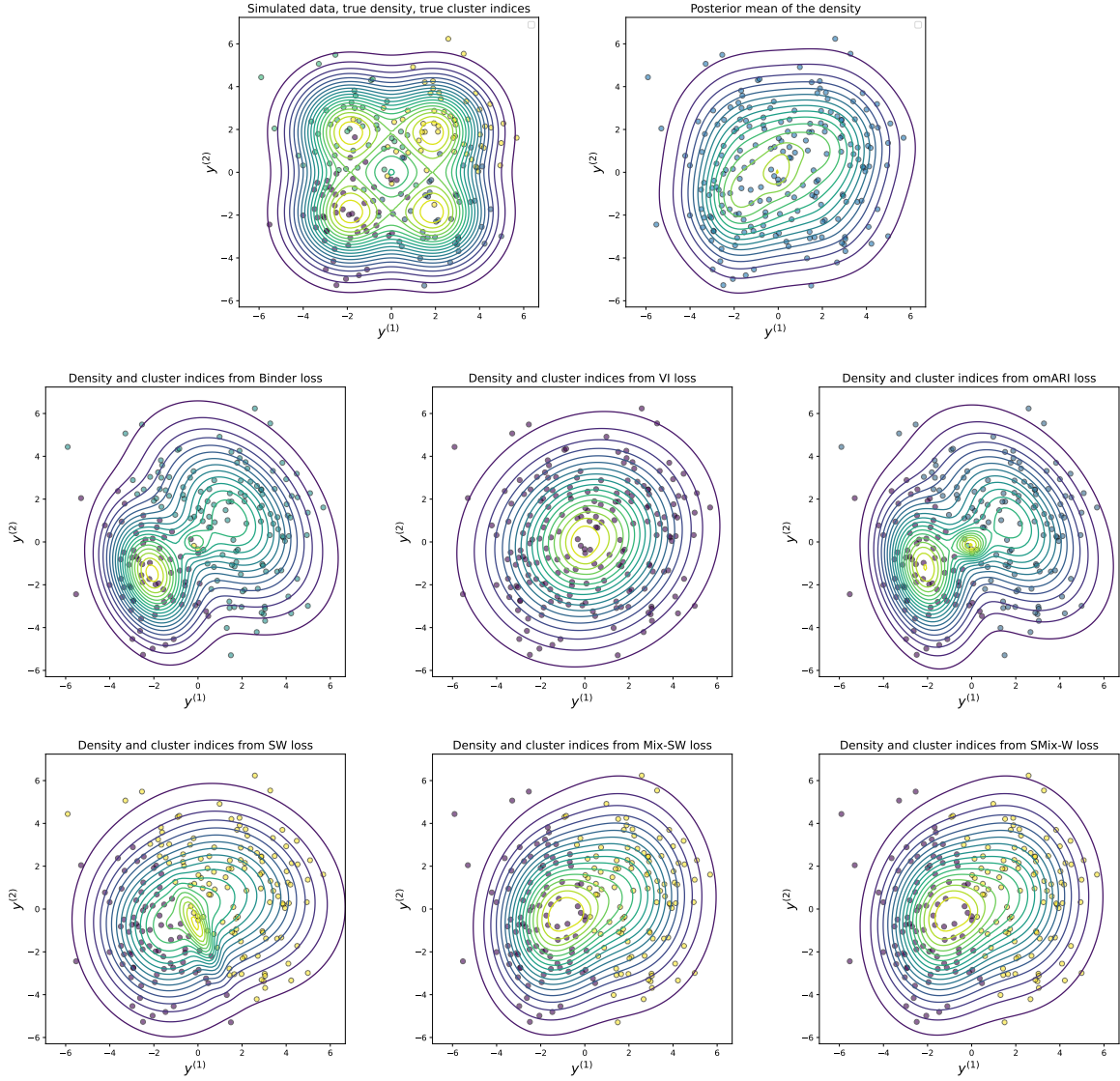
Figure 1: The figure shows the simulated data, the true generating density, the true cluster indices, the density and cluster indices of different loss functions in the two summarization approaches. For the prior approach of summarizing the partition (Binder loss, VI loss, and omARI loss), the posterior mean of the density conditioned on the reported point estimate of the partition is shown. For the proposed approach of summarizing the random mixing measure, the density is obtained via convoluting the point estimate of the mixing measure with the Gaussian density kernel, and the cluster indices are obtained via MAP given the the point estimate of the mixing measure.

4.1 Simulated data

Data generation and inference. Let $V = 1.5^2 \cdot I_2$. We sample 200 i.i.d data

$$y_i \overset{\text{i.i.d.}}{\sim} \frac{1}{4}\mathcal{N}\left((-2,-2),V)\right) + \frac{1}{4}\mathcal{N}\left((2,-2),V)\right) + \frac{1}{4}\mathcal{N}\left((-2,2),V\right) + \frac{1}{4}\mathcal{N}\left((2,2),V\right),$$

	k_n^*	$\mathbb{E}[B(\hat{z}, z)]$	$B(\hat{z}, z^\star)$	$\mathbb{E}[VI(\hat{z}, z)]$	$VI(\hat{z}, z^\star)$	$\mathbb{E}[omARI(\hat{z}, z)]$	$omARI(\hat{z}, z^\star)$
Binder	3	**0.3476**	0.3808	1.4548	2.0433	0.7022	0.7452
VI	1	0.4536	0.7478	**1.0211**	1.9936	1	1
omARI	4	0.3484	0.3536	1.5213	1.9641	**0.7**	0.7119
SW	2	0.3506	0.3724	1.4399	1.9719	0.7079	0.7324
Mix-SW	2	0.3529	**0.3338**	1.4322	**1.8838**	0.709	**0.6724**
SMix-W	2	0.3529	**0.3338**	1.4322	**1.8838**	0.709	**0.6724**

Table 1: The table shows the clustering performance on simulated data of different loss functions in the two summarization approaches. The columns from left to right are the number of unique clusters, expected Binder loss, Binder loss to the true cluster indices, expected VI loss, VI loss to the true cluster indices, expected omARI loss, and omARI loss to the true cluster indices. Lower losses are better.

	$\mathbb{E}[TV(\hat{F}, F))$	$TV(\hat{F}, F^\star)$	$\mathbb{E}[SW_2(\hat{F}, F))$	$SW_2(\hat{F}, F^\star)$
Binder	0.34195	0.4468	0.324	0.6278
VI	0.2745	0.4325	0.3112	0.6404
omARI	0.3423	0.4525	0.3264	0.5978
SW	0.2531	**0.3742**	0.2775	**0.488**
Mix-SW	**0.2453**	0.4143	**0.261**	0.5932
SMix-SW	**0.2453**	0.4143	**0.261**	0.5932

Table 2: The table shows the density estimation performance on simulated data of different loss functions in the two summarization approaches. he columns from left to right are expected Total Variation loss, Total Variation to the true density, expected SW_2 loss, SW_2 loss to the true density. Lower losses are better.

	$\mathbb{E}[SW_2(\hat{G}, G))$	$SW_2(\hat{G}, G^\star)$	$\mathbb{E}[\text{Mix-}SW_2(\hat{G}, G))$	$\text{Mix-}SW_2(\hat{G}, G^\star)$	$\mathbb{E}[\text{SMix-}W_2(\hat{G}, G))$	$\text{SMix-}W_2(\hat{G}, G^\star)$
SW	**1.3471**	**1.6524**	0.8207	1.0518	0.7464	0.9288
Mix-SW	1.406	1.6576	**0.8099**	**0.9898**	**0.7333**	**0.9122**
SMix-SW	1.406	1.6576	**0.8099**	**0.9898**	**0.7333**	**0.9122**

Table 3: The table shows the estimating the mixing measures performance of the proposed approach with SW, Mix-SW, and SMix-SW on the simulated data. The columns from left to right are expected SW_2 loss, SW_2 loss to the true mixing measure, expected Mix-SW_2 loss, Mix-SW_2 loss to the true mixing measure, expected SMix-W_2 loss ,and SMix-W_2 loss to the true mixing measure. Lower losses are better.

$i = 1, \ldots, n = 200$. We run 10000 blocked Gibbs sampler iterations (9000 burn-in iterations) with the following hyperparameters $\mu_0 = (0,0), \Psi = diag((1,1)), \lambda = 1, \nu = 4, \alpha = 1, K = 100$.

Figure 1 plots the simulated data, the true generating density, the true cluster indices, and the density and cluster indices obtained under different loss functions in the two summarization approaches. We evaluate the density on a 100×100 grid, with the range defined

by $(\min y_i - 1, \max y_i + 1)$. For the first approach, starting with a point estimate of the partition, we first determine $\widehat{\boldsymbol{c}}$ (using Binder, VI, or omARI loss), and then evaluate $\mathbb{E}\{F \mid \widehat{\boldsymbol{c}}\}$ by using 10 more iterations of the MCMC simulation to update F (freezing $\widehat{\boldsymbol{c}}$).

Clustering. Table 1 reports the expected losses and the relative loss, relative to the simulation truth, using Binder loss, VI loss, and omARI loss. We note that summarizing the random mixing measures yields comparable expected Binder loss and expected omARI loss, compared to the conventional approach of summarizing the partition first. In particular, summarizing the random mixing measures with SW results in only a 0.86% increase in expected Binder loss compared to the best expected Binder loss (using Binder loss in the optimization) and only a 1.13% increase in expected omARI loss compared to the best expected omARI loss (using omARI in the optimization). In this simulation, both Mix-SW and SMix-W yield the same point estimate of the random mixing measure. This is the case because we only search among the visited MCMC samples, as discussed in Section 2. They incur only a 1.53% increase in expected Binder loss relative to the best expected VI loss and only a 1.13% increase in expected omARI loss compared to the best expected omARI loss. The only loss for which the proposed approach leads to a considerable increase is VI loss, with an increase of about 40%. From a frequentist perspective, Mix-SW and SMix-W demonstrate the best clustering performance, as they have the smallest losses relative to the simulation truth. Overall, we conclude that with respect to the reported point estimate $\widehat{\boldsymbol{c}}$ the approaches that start with a point estimate of the random mixing measure perform comparable to approaches that start with the random partition.

Density estimation. Table 2 reports expected losses and relative losses, relative to the simulation truth using total variation loss and SW_2 loss. For the approachs that first summarize the partitions (based on Binder, VI, or omARI loss), the losses are averaged over 10 Monte Carlo samples of F (freezing the point estimate $\widehat{\boldsymbol{c}}$). Naturally, the proposed

	$k_n^\star$	$\mathbb{E}[B(\hat{z}, z)]$	$\mathbb{E}[VI(\hat{z}, z)]$	$\mathbb{E}[omARI(\hat{z}, z)]$
Binder	8	**0.0296**	**0.2588**	**0.0594**
VI	3	0.0333	0.246	0.0667
omARI	8	**0.0296**	**0.2588**	**0.0594**
SW	4	0.0303	0.2602	0.0607
Mix-SW	3	0.0306	0.2678	0.0614
SMix-W	3	0.0306	0.2678	0.0614

Table 4: The table assesses clustering for the Old Faithful dataset when using different loss functions in the two summarization approaches. The columns from left to right are the number of unique clusters, expected Binder loss, expected VI loss, and expected omARI loss. Lower losses are better.

	$\mathbb{E}[TV(\hat{F}, F))$	$\mathbb{E}[SW_2(\hat{F}, F))$
Binder	0.2128	1.0568
VI	0.2102	1.0301
omARI	0.2128	1.0568
SW	0.1901	0.868
Mix-SW	**0.1852**	**0.7768**
SMix-SW	**0.1852**	**0.7768**

Table 5: The table shows the density estimation performance on the Old Faithful dataset of different loss functions in the two summarization approaches. The columns from left to right are expected Total Variation loss and expected SW_2 loss. Lower losses are better.

approach of summarizing the mixing measures first results in lower losses. In particular, the best partition summarization method is 11.9% higher in expected total variation loss and 19.23% higher in expected SW loss compared to Mix-SW and SMix-W. Both Mix-SW and SMix-W outperform all partition focused approaches from a frequentist perspective in a comparison versus the known simulation truth. SW performs well in this, demonstrating expected losses nearly as good as those of SMix-W and Mix-SW.

Mixing measure estimation. Table 3 reports expected losses and relative losses, relative to the simulation truth, using SW_2 loss, Mix-SW_2 loss, and SMix-W_2. Not surprisingly, inference under each loss performs best when used for its intended evaluation.
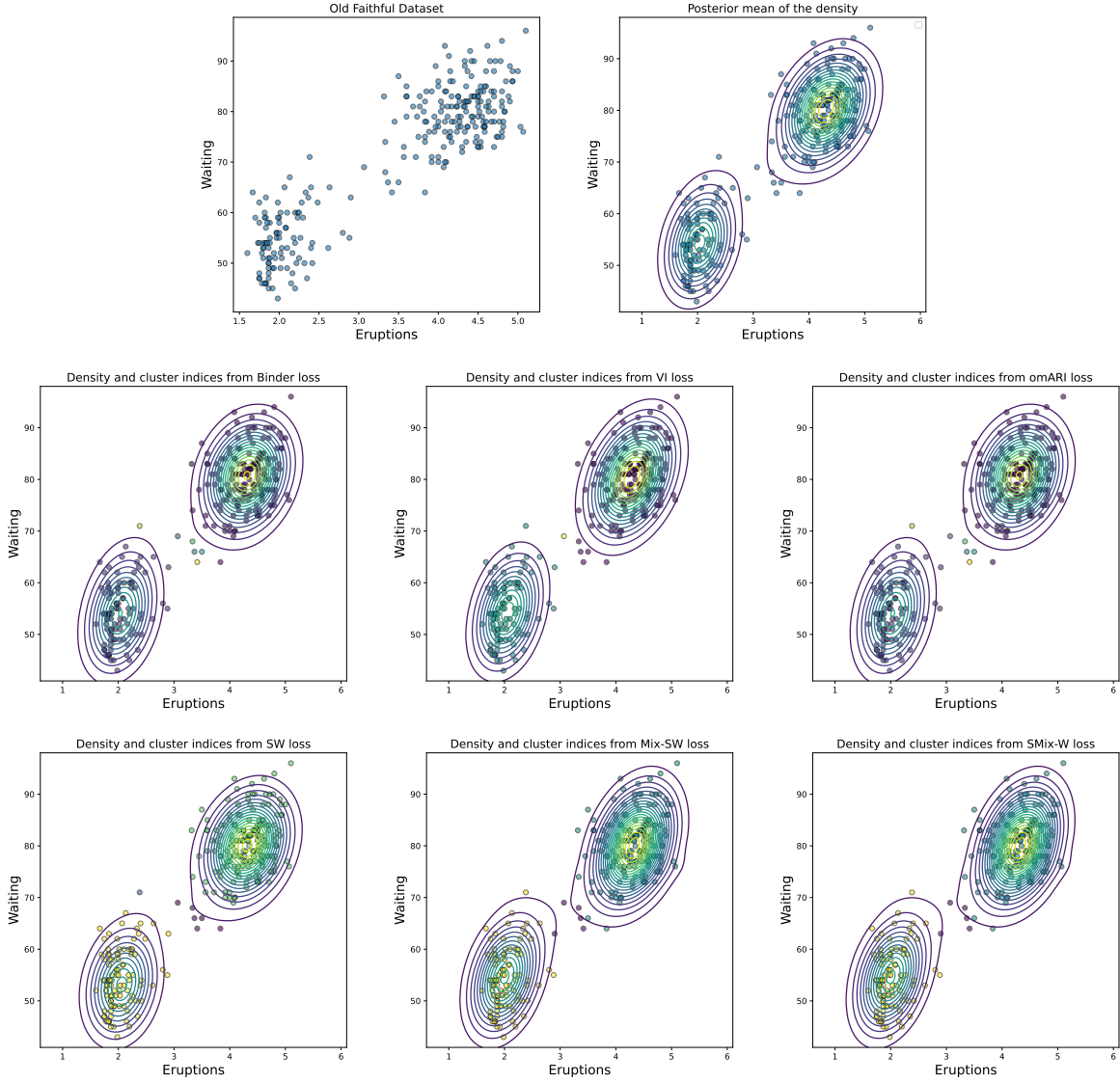
Figure 2: The figure shows the Old Faithful geyser data, the estimated density and partitions using different loss functions under the two summarization approaches. For the approaches targeting $\widehat{\boldsymbol{c}}$ first (Binder loss, VI loss, and omARI loss), the figure shows the estimated density conditional on $\widehat{\boldsymbol{c}}$. For the approaches targeting the mixing measure G first, the plots show the convolution with the Gaussian kernel, and the cluster indices are obtained as MAP given $\widehat{G}$.

	$\mathbb{E}[SW_2(\widehat{G}, G))$	$\mathbb{E}[\text{Mix-}SW_2(\widehat{G}, G))$	$\mathbb{E}[\text{SMix-}W_2(\widehat{G}, G))$
SW	**3.1211**	1.4876	1.4831
Mix-SW	3.7917	**1.3186**	**1.3091**
SMix-W	3.7917	**1.3186**	**1.3091**

Table 6: The table shows the estimating the mixing measures performance of the proposed approach with SW, Mix-SW, and SMix-SW on the Old Faithful data. The columns from left to right are expected SW_2 loss, expected Mix-SW_2 loss, and expected SMix-W_2 loss. Lower losses are better.

24

4.2 Old Faithful Geyser dataset

The Old Faithful geyser dataset contains 272 data samples in 2 dimensions. We run 10000 blocked Gibbs sampler iterations (9000 burn-in iterations) with the following hyperparameters $\mu_0 = (3, 70), \Psi = diag((4, 26)), \lambda = 1, \nu = 4, \alpha = 1, K = 100$. Figure 2 shows the data, along with the density estimate and estimated partition obtained under different loss functions using the two summarization approaches. We use the same visualization techniques for the density as described in Section 4.1.

Table 4 reports the expected losses using Binder loss, VI, and omARI loss. We note a similar overall pattern as in the simulation: summarizing the random mixing measures yields comparable expected Binder loss and expected omARI loss compared to partition-focused approaches. In particular, summarizing the random mixing measures with SW, Mix-SW, and SMix-SW results in only about a 3.38% increase in expected Binder loss compared to the best expected Binder loss, and only about a 3.37% increase in expected omARI loss compared to the best expected omARI loss. For VI loss, the proposed approach results in an increase of about 8.86% in expected VI loss compared to the best expected VI loss. Table 5 reports the expected losses using total variation and SW_2 loss for the density estimation. For the partition-focused approaches (using Binder, VI, or omARI loss), the losses are averaged over 10 Monte Carlo samples of F, as under the previously described simulation study. Again, we summarizing the mixing measure first naturally leads to better density estimates than summarizing partitions first. In particular, the best partition-focused method reports a 13.5% higher expected total variation loss and 19.23% higher expected SW loss compared to Mix-SW and SMix-SW. Furthermore, we find that SMix-SW and Mix-SW compare favorably to SW, highlighting the benefits of structured projection.

Finally, in Table 6 we report the expected losses under SW_2 loss, Mix-SW_2 loss, and SMix-W_2 for estimating the mixing measure. Again we ovserve similar patterns as in

the simulation study. Inference under each loss function performs best when used for its intended evaluation.

5   Conclusion

We present a new approach for Bayesian analysis of Bayesian Nonparametric (BNP) mixture models, focusing on summarizing the posterior distribution on the mixing measures. Our method minimizes the posterior expected loss using a discrepancy between measures as the loss function, utilizing the computationally scalable sliced Wasserstein distance. For Gaussian mixture models, we introduce two variants: mixed sliced Wasserstein (Mix-SW) and sliced mixture Wasserstein (SMix-W). Mix-SW uses generalized geodesic projection on the product of the Euclidean manifold and the manifold of symmetric positive definite matrices, providing a meaningful metric for comparing Gaussian mixing measures. SMix-W leverages the linearity of Gaussian mixtures for efficient projection. Empirical analyses show that our summarization approach yields more accurate density estimates while maintaining a good partition summary.

Limitations of the proposed approach include the restriction to truncated mixtures and the potential suboptimality of the reported point estimate of the mixing measures, as the solution is limited to the Monte Carlo set of posterior samples. The simple sliced Wasserstein distance might not be optimal for all mixture models; therefore, different variants should be designed to exploit specific geometry, as we do for Gaussian mixtures.

Future work will focus on refining the search algorithm to achieve better point estimation and extending the approach to more generalized Bayesian nonparametric (BNP) mixtures.

References

Ali, S. M. and Silvey, S. D. (1966), "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society: Series B (Methodological)*, 28, 131–142.

Azzalini, A. and Bowman, A. W. (1990), "A look at some data on the Old Faithful geyser," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39, 357–365.

Binder, D. A. (1978), "Bayesian cluster analysis," *Biometrika*, 65, 31–38.

Bonet, C., Drumetz, L., and Courty, N. (2024), "Sliced-Wasserstein Distances and Flows on Cartan-Hadamard Manifolds," *arXiv preprint arXiv:2403.06560*.

Bonet, C., Malézieux, B., Rakotomamonjy, A., Drumetz, L., Moreau, T., Kowalski, M., and Courty, N. (2023), "Sliced-Wasserstein on symmetric positive definite matrices for M/EEG signals," in *International Conference on Machine Learning*, PMLR.

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015), "Sliced and Radon Wasserstein Barycenters of Measures," *Journal of Mathematical Imaging and Vision*, 1, 22–45.

Bonnotte, N. (2013), *Unidimensional and evolution methods for optimal transportation*, Ph.D. thesis, Paris 11.

Cuturi, M. (2013), "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*.

Dahl, D. B., Johnson, D. J., and Müller, P. (2022), "Search algorithms and loss functions for Bayesian clustering," *Journal of Computational and Graphical Statistics*, 31, 1189–1201.

Delon, J. and Desolneux, A. (2020), "A Wasserstein-type distance in the space of Gaussian mixture models," *SIAM Journal on Imaging Sciences*, 13, 936–970.

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

Gu, A., Sala, F., Gunel, B., and Ré, C. (2019), "Learning mixed-curvature representations in products of model spaces," in *International conference on learning representations*, volume 5.

Ishwaran, H. and James, L. F. (2001), "Gibbs sampling methods for stick-breaking priors," *Journal of the American statistical Association*, 96, 161–173.

Lau, J. W. and Green, P. J. (2007), "Bayesian model-based clustering procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558.

MacEachern, S. N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science, Alexandria, VA: American Statistical Association.*.

Meilă, M. (2007), "Comparing clusterings—an information based distance," *Journal of multivariate analysis*, 98, 873–895.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. (2020), "Statistical and topological properties of sliced probability divergences," *Advances in Neural Information Processing Systems*, 33, 20802–20812.

Nguyen, K. and Ho, N. (2024), "Hierarchical Hybrid Sliced Wasserstein: A Scalable Metric for Heterogeneous Joint Distributions," *Advances in Neural Information Processing Systems*.

Paty, F.-P. and Cuturi, M. (2019), "Subspace Robust Wasserstein Distances," in *International Conference on Machine Learning*.

Pennec, X., Sommer, S., and Fletcher, T. (2019), *Riemannian geometric statistics in medical image analysis*, Academic Press.

Peyré, G., Cuturi, M., et al. (2019), "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, 11, 355–607.

Quintana, F. A., Müller, P., Jara, A., and MacEachern, S. N. (2022), "The dependent Dirichlet process and related models," *Statistical Science*, 37, 24–41.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2012), "Wasserstein barycenter and its application to texture mixing," in *Scale Space and Variational Methods in Computer Vision: Third International Conference, SSVM 2011, Ein-Gedi, Israel, May 29–June 2, 2011, Revised Selected Papers 3*, Springer.

Rand, W. M. (1971), "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, 66, 846–850.

Rastelli, R. and Friel, N. (2018), "Optimal Bayesian estimators for latent variable cluster models," *Statistics and Computing*, 28, 1169–1186.

Villani, C. (2009), *Optimal transport: old and new*, volume 338, Springer.

Vinh, N. X., Epps, J., and Bailey, J. (2009), "Information theoretic measures for clusterings comparison: is a correction for chance necessary?" in *Proceedings of the 26th annual international conference on machine learning*.

Wade, S. and Ghahramani, Z. (2018), "Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion)," *Bayesian Analysis*, 13, 559–626.

Yakowitz, S. J. and Spragins, J. D. (1968), "On the identifiability of finite mixtures," *The Annals of Mathematical Statistics*, 39, 209–214.

A Appendices

A.1 Review on Riemannian Manifolds

A Riemannian manifold $(\mathcal{M}, G)$ of dimension d is a space that behaves locally as a linear space diffeomorphic to $\mathbb{R}^d$, named a tangent space. For any $x \in \mathcal{M}$, the associated tangent space is defined as $T_x\mathcal{M}$ which supports an inner product $\langle \cdot, \cdot \rangle_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ i.e., $\langle u, v \rangle_x = u^\top G(x) v$. The joint space of the manifold and the tangent space is called the tangent bundle $T\mathcal{M} = \{x \in \mathcal{M}, v \in T_x\mathcal{M}\}$.

Geodesics. Given two points $x, y \in \mathcal{M}$, the smooth curve $\gamma : [0, 1] \to \mathcal{M}$ such as $\gamma(0) = x, \gamma(1) = y$ is called geodesic if it minimizes the length:

$$\mathcal{L}(\gamma) = \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}},$$

where $\gamma'(t)$ is the derivative of the curve $\gamma(t)$ with respective to t, which belongs to the tangent space $T_{\gamma(t)}\mathcal{M}$ for any $t \in [0, 1]$. The length of the geodesic line is the geodesic distance.

$$d(x, y) = \inf_{\gamma | \gamma(0) = x, \gamma(1) = y} \mathcal{L}(\gamma).$$

Exponential Map. Let $x \in \mathcal{M}$, for any $v \in T_x\mathcal{M}$, there exists a unique geodesic γ with $\gamma(0) = x$ and $\gamma'(0) = v$, denoted as $\gamma_{x,v}$. The exponential map $\exp : T\mathcal{M} \to \mathcal{M}$ maps $v \in T_x\mathcal{M}$ back to the manifold at the point reached by the geodesic $\gamma(1)$. In particular, we have the following definition of the exponential map:

$$\forall (x, v) \in T\mathcal{M}, \exp_x(v) = \gamma_{x,v}(1)$$

30

A.2   Proof of Proposition 1

By Definition 1, we have:

$$P_{V_w}((\mu, \Sigma)) = \arg\min_{t \in \mathbb{R}} d(x, \exp_0(tV_w)) = \arg\min_{t \in \mathbb{R}} d^2(x, \exp_0(tV_w))$$

$$= \arg\min_{t \in \mathbb{R}} \|\mu - tw_1 v\|_2^2 + \|\log\Sigma - \log\exp(tw_2 A)\|_F$$

$$= \arg\min_{t \in \mathbb{R}} \|\mu - tw_1 v\|_2^2 + \|\log\Sigma - tw_2 A\|_F$$

$$= \arg\min_{t \in \mathbb{R}} \|\mu\|_2^2 + w_1^2 t^2 - 2w_1 t\langle\mu, v\rangle + w_2^2 t^2 + Trace((log\Sigma)^2) - 2w_2 t Trace(A\log\Sigma)$$

$$= \arg\min_{t \in \mathbb{R}} \|\mu\|_2^2 + t^2 - 2w_1 t\langle\mu, v\rangle + Trace((log\Sigma)^2) - 2w_2 t Trace(A\log\Sigma)$$

$$:= \arg\min_{t \in \mathbb{R}} f(t)$$

Taking the derivative $\frac{d}{dt} f(t) = 2t - 2w_1\langle\mu, v\rangle - 2w_2 Trace(A\log\Sigma)$, then set it to 0. We obtain: $P_{V_w}(\theta) = t^\star = w_1\langle\mu, v\rangle + w_2 Trace(A\log\Sigma)$, which completes the proof.

A.3   Proof of Proposition 2

Since $P_{V_w}((\mu, \Sigma)) = w_1\langle\mu, v\rangle + w_2 Trace(A\log\Sigma)$ is a Borel meansurable, using Lemma 6 in (Paty and Cuturi, 2019), we have:

$$W_p^p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_2)] = \inf_{\pi_{V_w} \in \Pi(P_{V_w}\sharp G_1, P_{V_w}\sharp G_2)} \int_{\mathbb{R}\times\mathbb{R}} |x - y|^p d\pi_{V_w}(x, y)$$

$$= \inf_{\pi \in \Pi(G_1, G_2)} \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R}) \times \mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{V_w}(\mu_1, \Sigma_1) - P_{V_w}(\mu_2, \Sigma_2)|^p d\pi((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)).$$

Using the Minkowski's inequality, we have:

$$W_p^p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_2)] \leq \inf_{\pi \in \Pi(G_1,G_2))} \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R}) \times \mathbb{R}^d \times S_d^{++}(\mathbb{R})} 2^{p-1} \left(|P_{V_w}(\mu_1, \Sigma_1) - P_{V_w}(\mu_0, \Sigma_0)|^p \right.$$

$$+ |P_{V_w}(\mu_0, \Sigma_0) - P_{V_w}(\mu_2, \Sigma_2)|^p \right) d\pi_{V_w} d\pi((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$$

$$= 2^{p-1} \left(\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{V_w}(\mu_1, \Sigma_1) - P_{V_w}(\mu_0, \Sigma_0)|^p dG_1(\mu_1, \Sigma_1) \right.$$

$$+ \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{V_w}(\mu_0, \Sigma_0) - P_{V_w}(\mu_2, \Sigma_2)|^p dG_2(\mu_2, \Sigma_2) \right).$$

Moreover, from the Cauchy–Schwarz's inequality, we have:

$$|P_{V_w}(\mu_1, \Sigma_1) - P_{V_w}(\mu_0, \Sigma_0)|$$

$$= |w_1\langle \mu_1, v\rangle + w_2 Trace(A \log \Sigma_1) - w_1\langle \mu_0, v\rangle - w_2 Trace(A \log \Sigma_0)|$$

$$= |w_1\langle \mu_1 - \mu_0, v\rangle + w_2(Trace(A \log \Sigma_1) - Trace(A \log \Sigma_0))|$$

$$\leq \sqrt{w_1^2 + w_2^2}\sqrt{\langle \mu_1 - \mu_0, v\rangle^2 + (Trace(A(\log \Sigma_1 - \log \Sigma_0))^2}$$

$$= \sqrt{\langle \mu_1 - \mu_0, v\rangle^2 + (Trace(A(\log \Sigma_1 - \log \Sigma_0))^2}$$

$$\leq \sqrt{\|v\|_2^2\|\mu_1 - \mu_0\|_2^2 + \|A\|_F^2\|\log \Sigma_1 - \log \Sigma_0\|_F^2}$$

$$= \sqrt{\|\mu_1 - \mu_0\|_2^2 + \|\log \Sigma_1 - \log \Sigma_0\|_F^2} = d((\mu_1, \Sigma_1), (\mu_0, \Sigma_0)).$$

From the assumption, we get:

$$W_p^p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_2)] \leq 2^{p-1} \left(\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_1, \Sigma_1), (\mu_0, \Sigma_0))^p dG_1(\mu_1, \Sigma_1) \right.$$

$$+ \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_0, \Sigma_0), (\mu_2, \Sigma_2))^p dG_2(\mu_2, \Sigma_2) \right) < \infty,$$

which completes the proof.

A.4   Proof of Theorem 1

Symmetry and Non-negativity.   The symmetry and non-negativity of Mix-SW follows directly the symmetry and non-negativity of the Wasserstein distance (Peyré et al., 2019) since it is the expectation of projected Wasserstein distance.

Triangle Inequality.   Let consider three measure G_1, G_2, G_3, using the triangle inequality of Wasserstein distance, we have:

$$\text{Mix-}SW_p(G_1, G_2) = \left(\mathbb{E}_{(w,v,A)\sim\mathcal{U}(\mathbb{S})\otimes\mathcal{U}(\mathbb{S}^{d-1})\otimes\mathcal{U}(S_d(\mathbb{R}))}[W_p^p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_2)]| \right)^{\frac{1}{p}}$$

$$\leq \left(\mathbb{E}_{(w,v,A)\sim\mathcal{U}(\mathbb{S})\otimes\mathcal{U}(\mathbb{S}^{d-1})\otimes\mathcal{U}(S_d(\mathbb{R}))}[(W_p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_3) + W_p(P_{V_w}\sharp G_3, P_{V_w}\sharp G_2))^p]| \right)^{\frac{1}{p}} .$$

Using the Minkowski's inequality, we get:

$$\text{Mix-}SW_p(G_1, G_2) \leq \left(\mathbb{E}_{(w,v,A)\sim\mathcal{U}(\mathbb{S})\otimes\mathcal{U}(\mathbb{S}^{d-1})\otimes\mathcal{U}(S_d(\mathbb{R}))}[W_p^p(P_{V_w}\sharp G_1, P_{V_w}\sharp G_3)]| \right)^{\frac{1}{p}}$$

$$+ \left(\mathbb{E}_{(w,v,A)\sim\mathcal{U}(\mathbb{S})\otimes\mathcal{U}(\mathbb{S}^{d-1})\otimes\mathcal{U}(S_d(\mathbb{R}))}[W_p^p(P_{V_w}\sharp G_3, P_{V_w}\sharp G_2)]| \right)^{\frac{1}{p}}$$

$$= \text{Mix-}SW_p(G_1, G_3) + \text{Mix-}SW_p(G_3, G_2),$$

which completes the proof of triangle inequality.

Identity of indiscernibles.   When $G_1 = G_2$, we have directly Mix-$SW_p^p(G_1, G_2) = 0$. We now prove that if Mix-$SW_p^p(G_1, G_2) = 0$, we get $G_1 = G_2$. We first rewrite $P_{V_w}(\mu, \Sigma)$ as a composition of two function i.e., $P_{V_w}(\mu, \Sigma) = P_w \circ P_V(\mu, \Sigma)$. In particular, $P_V :$ $\mathbb{R}^d \times S_d^{++}(\mathbb{R}) \to \mathbb{R}^2$ i.e., $P_V(\mu, \Sigma) = (\langle \mu, v \rangle, Trace(A \log \Sigma))$ $(V = (v, A))$ and $P_w : \mathbb{R}^2 \to \mathbb{R}$

i.e., $P_w(x) = \langle w, x \rangle$. We then can rewrite mixed Sliced Wasserstein distance as:

$$\text{Mix-}SW_p^p(G_1, G_2) = \mathbb{E}_{(w,v,A) \sim \mathcal{U}(\mathbb{S}) \otimes \mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))} [W_p^p(P_{V_w} \sharp G_1, P_{V_w} \sharp G_2)]$$

$$= \mathbb{E}_{(v,A) \sim \mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))} \left[\mathbb{E}_{w \sim \mathcal{U}(\mathbb{S})} [W_p^p(P_w \sharp P_V \sharp G_1, P_w \sharp P_V \sharp G_2)] \right]$$

$$= \mathbb{E}_{(v,A) \sim \mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))} \left[SW_p^p(P_V \sharp G_1, P_V \sharp G_2) \right].$$

When $\text{Mix-}SW_p^p(G_1, G_2) = 0$, it means that $SW_p^p(P_V \sharp G_1, P_V \sharp G_2) = 0$ for $\mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))$-almost every (v, A). Using the identity of indiscernibles proptery of SW (Bonnotte, 2013), we have $P_V \sharp G_1 = P_V \sharp G_2$ for $\mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))$-almost every (v, A). Let denote $\mathcal{F}[P_V \sharp G_1]$ and $\mathcal{F}[P_V \sharp G_2]$ as the Fourier transform of G_1 and G_2 respectively, we have $\mathcal{F}[P_V \sharp G_1] = \mathcal{F}[P_V \sharp G_2]$ for $\mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))$-almost every (v, A). Moreover, for all $y \in \mathbb{R}^2$, we have:

$$\mathcal{F}[P_V \sharp G_1](y) = \int_{\mathbb{R}^2} e^{-2i\pi \langle y, x \rangle} d(P_V \sharp G_1)(x)$$

$$= \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} e^{-2i\pi(y_1 \langle v, \mu_1 \rangle + y_2 \langle A, \log \Sigma_1 \rangle_F)} dG_1(\mu_1, \Sigma_1)$$

$$= \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} e^{-2i\pi(\langle y_1 v, \mu_1 \rangle + \langle y_2 A, \log \Sigma_1 \rangle_F)} dG_1(\mu_1, \Sigma_1)$$

$$= \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} e^{-2i\pi(\langle y_1 v, \mu_1 \rangle + \langle y_2 A, S_1 \rangle_F)} d((Id, \log) \sharp G_1)(\mu_1, S_1)$$

$$= \mathcal{F}[(Id, \log) \sharp G_1](y_1 v, y_2 A).$$

Therefore, we obtain $\mathcal{F}[(Id, \log) \sharp G_1](y_1 v, y_2 A) = \mathcal{F}[(Id, \log) \sharp G_2](y_1 v, y_2 A)$ for $\mathcal{U}(\mathbb{S}^{d-1}) \otimes \mathcal{U}(S_d(\mathbb{R}))$-almost every (v, A). By injectivity of the Fourier transform, we get $(Id, \log) \sharp G_1 = (Id, \log) \sharp G_2$. Since the function $f(\mu, \Sigma) = (\mu, \log \Sigma)$ is injective i.e., $f^{-1}(\mu, \Sigma) = (\mu, \exp(\Sigma))$, we obtain $G_1 = G_2$, which completes the proof.

A.5 Proof of Proposition 3

Since $P_{v,w}(\mu, \Sigma) = w_1 \langle v, \mu \rangle + w_2 \log(\sqrt{v^\top \Sigma v})$ is a Borel meansurable, using Lemma 6 in (Paty and Cuturi, 2019), we have:

$$W_p^p(P_{v,w} \sharp G_1, P_{v,w} \sharp G_2)] = \inf_{\pi_{v,w} \in \Pi(P_{v,w} \sharp G_1, P_{v,w} \sharp G_2)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|^p d\pi_{v,w}(x, y)$$

$$= \inf_{\pi \in \Pi(G_1, G_2)} \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R}) \times \mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{v,w}(\mu_1, \Sigma_1) - P_{v,w}(\mu_2, \Sigma_2)|^p d\pi((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)).$$

Using the Minkowski's inequality, we have:

$$W_p^p(P_{v,w} \sharp G_1, P_{v,w} \sharp G_2)] \leq \inf_{\pi \in \Pi(G_1, G_2))} \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R}) \times \mathbb{R}^d \times S_d^{++}(\mathbb{R})} 2^{p-1} \left(|P_{v.w}(\mu_1, \Sigma_1) - P_{v.w}(\mu_0, \Sigma_0)|^p \right.$$

$$+ |P_{v.w}(\mu_0, \Sigma_0) - P_{v.w}(\mu_2, \Sigma_2)|^p \Big) \, d\pi_{V_w} d\pi((\mu_1, \Sigma_1), (\mu_2, \Sigma_2))$$

$$= 2^{p-1} \left(\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{v.w}(\mu_1, \Sigma_1) - P_{v.w}(\mu_0, \Sigma_0)|^p dG_1(\mu_1, \Sigma_1) \right.$$

$$+ \left. \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} |P_{v.w}(\mu_0, \Sigma_0) - P_{v.w}(\mu_2, \Sigma_2)|^p dG_2(\mu_2, \Sigma_2) \right).$$

Moreover, from the Cauchy–Schwarz's inequality, we have:

$$|P_{v.w}(\mu_1, \Sigma_1) - P_{v.w}(\mu_0, \Sigma_0)|$$

$$= |w_1 \langle v, \mu_1 \rangle + w_2 \log(\sqrt{v^\top \Sigma_1 v}) - w_1 \langle v, \mu_0 \rangle + w_2 \log(\sqrt{v^\top \Sigma_0 v})|$$

$$= \left| w_1 \langle \mu_1 - \mu_0, v \rangle + 0.5 w_2 \log \left(\frac{v^\top \Sigma_1 v}{v^\top \Sigma_0 v} \right) \right|$$

$$\leq \sqrt{w_1^2 + w_2^2} \sqrt{\langle \mu_1 - \mu_0, v \rangle^2 + 0.25 \log \left(\frac{v^\top \Sigma_1 v}{v^\top \Sigma_0 v} \right)^2}$$

$$= \sqrt{\langle \mu_1 - \mu_0, v \rangle^2 + 0.25 \log \left(\frac{v^\top \Sigma_1 v}{v^\top \Sigma_0 v} \right)^2}$$

$$\leq \sqrt{\|v\|_2^2 \|\mu_1 - \mu_0\|_2^2 + 0.25 \log \left(\max_v \frac{v^\top \Sigma_1 v}{v^\top \Sigma_0 v} \right)^2}$$

$$= \sqrt{\|\mu_1 - \mu_0\|_2^2 + 0.25 \log(\lambda_{max}(\Sigma_1, \Sigma_2))^2} = d((\mu_1, \Sigma_0), (\mu_0, \Sigma_0)),$$

where $\lambda_{max}(\Sigma_1, \Sigma_0)$ is the largest eigenvalue of the generalized problem $\Sigma_1 v = \lambda \Sigma_0 v$. From the assumption, we get:

$$W_p^p(P_{v.w} \sharp G_1, P_{v,w} \sharp G_2)] \leq 2^{p-1} \left(\int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_1, \Sigma_1), (\mu_0, \Sigma_0))^p dG_1(\mu_1, \Sigma_1) \right.$$

$$\left. + \int_{\mathbb{R}^d \times S_d^{++}(\mathbb{R})} d((\mu_0, \Sigma_0), (\mu_2, \Sigma_2))^p dG_2(\mu_2, \Sigma_2) \right) < \infty,$$

which completes the proof.

A.6 Proof of Theorem 2

Symmetry, Non-negativity, and Triangle Inequality. The symmetry, non-negativity, and triangle inequality of SMix-W can be obtained by following the proof for Mix-SW in Appendix A.4. In this section, we focus on the proof of identity of indiscernibles for SMix-W.

Identity of indiscernibles. When $G_1 = G_2$, we have directly SMix-$W_p^p(G_1, G_2) = 0$. We now prove that if SMix-$W_p^p(G_1, G_2) = 0$, we get $G_1 = G_2$. From the definition of SMix-W in Definition 3, we have:

$$\text{SMix-}W_p^p(G_1, G_2) = \mathbb{E}_{(w,v) \sim \mathcal{U}(\mathbb{S}) \otimes \mathcal{U}(\mathbb{S}^{d-1})}[W_p^p(P_{v,w} \sharp G_1, P_{v,w} \sharp G_2)]$$

$$= \mathbb{E}_{v \sim \mathcal{U}(\mathbb{S}^{d-1})}[SW_p^p((Id, \log \circ \sqrt{}) \sharp P_v' \sharp G_1, (Id, \log \circ \sqrt{}) \sharp P_v' \sharp G_2)].$$

When SMix-$W_p^p(G_1, G_2) = 0$, it implies $SW_p^p((Id, \log \circ \sqrt{}) \sharp P_v' \sharp G_1, (Id, \log \circ \sqrt{}) \sharp P_v' \sharp G_2) = 0$ for $\mathcal{U}(\mathbb{S}^{d-1})$-almost every v. Since $\log(x)$ is an injective function, it leads to the fact that $SW_p^p((Id, \sqrt{}) \sharp P_v' \sharp G_1, (Id, \sqrt{}) \sharp P_v' \sharp G_2) = 0$ for $\mathcal{U}(\mathbb{S}^{d-1})$-almost every v. By the identity of indiscernibles of SW (Bonnotte, 2013), we have $(Id, \sqrt{}) \sharp P_v' \sharp G_1 = (Id, \sqrt{}) \sharp P_v' \sharp G_2$ for $\mathcal{U}(\mathbb{S}^{d-1})$-almost every v with $P_v'(\mu, \Sigma) = (\langle v, \mu \rangle, v^\top \Sigma v)$. Since the square root function also injective on $\mathbb{R}^+$, we have $P_v' \sharp G_1 = P_v' \sharp G_2$ which is equivalent to $P_v \sharp F_1 = P_v \sharp F_2$ for $\mathcal{U}(\mathbb{S}^{d-1})$-

almost every v with $P_v(x) = \langle v, x \rangle$ and $F_1 = f * G_1$ and $F_2 = f * G_2$ (f is the Gaussian density kernel). Let denote $\mathcal{F}[P_v \sharp F_1]$ and $\mathcal{F}[P_v \sharp F_2]$ as the Fourier transform of F_1 and F_2 respectively, we have $\mathcal{F}[P_v \sharp F_1] = \mathcal{F}[P_v \sharp F_2]$ for $\mathcal{U}(\mathbb{S}^{d-1})$-almost every v. Moreover, for all $t \in \mathbb{R}$, we have:

$$
\begin{aligned}
\mathcal{F}[P_v \sharp F_1](t) &= \int_{\mathbb{R}^2} e^{-2i\pi t \epsilon} d(P_v \sharp F_1)(\epsilon) = \int_{\mathbb{R}^d} e^{-2i\pi t \langle v, x \rangle} dF_1(x) \\
&= \int_{\mathbb{R}^d} e^{-2i\pi \langle tv, x \rangle} dF_1(x) = \mathcal{F}[F_1](tv).
\end{aligned}
$$

Therefore, we get $\mathcal{F}[F_1](tv) = \mathcal{F}[F_2](tv)$ for $\mathcal{U}(\mathbb{S}^{d-1})$-almost every v. By the injectivity of Fourier Transform, we get $F_1 = F_2$ which leads to $G_1 = G_2$ due to the identifiability of finite mixture of Gaussians (Proposition 2 in (Yakowitz and Spragins, 1968)), which concludes the proof.