Gaussian Mixture Models in Hilbert Spaces via Kernel Methods

Daniel López-Montero[*]
Friedrich-Alexander-Universität Erlangen-Nürnberg
dani.lopez@fau.de

Antonio Álvarez-López[*]
Universidad Autónoma de Madrid
Nanyang Technological University
antonioalvarezlopez98@gmail.com

Marcos Matabuena[†]
Mohamed bin Zayed University of Artificial Intelligence
marcos.matabuena@mbzuai.ac.ae

Abstract

Modern datasets across many disciplines increasingly consist of time-evolving, potentially infinite-dimensional random objects, such as dynamic functional data, which are naturally modeled in Hilbert spaces. In these settings, characterizing probability measures, for example, through densities, can be ill-defined or technically challenging. Motivated by clustering applications, we propose a Gaussian mixture framework for Hilbert-space-valued data based on kernel mean embeddings and develop efficient optimization algorithms for estimation. We establish theoretical guarantees showing that the proposed algorithm is well-defined and that the model yields a dense class of approximations in infinite-dimensional spaces. We evaluate the framework through extensive experiments on diverse data geometries, including L^2-functional data and graph signals arising in modern medical applications.

1 Introduction

Technological progress across multiple domains is generating high-dimensional datasets that demand new analytical tools. Modern data structures, such as random functions, spatial fields, images, signals, graphs, or trajectories, are often more naturally represented as random objects in Hilbert spaces than as finite-dimensional Euclidean vectors in $\mathbb{R}^d$ (Figure 1). Analyzing the data directly using Hilbert-space methods preserves their intrinsic geometric structure and allows their full potential to be exploited [56, 28].

In many applications, observed data are heterogeneous and may arise from K latent populations, each characterized by a probability law P_i, with $i = 1, \ldots, K$. Mixture models [30] provide a flexible framework for representing such data-generating mechanisms and, under suitable regularity conditions, can approximate broad classes of probability measures. This paper proposes a mixture-based framework for probability measures on a separable Hilbert space $\mathcal{X}$, extending the classical idea of Gaussian mixture models (GMMs) for finite-dimensional data [24].

Extending GMMs beyond finite-dimensional spaces is challenging. Infinite-dimensional Hilbert spaces do not admit a Lebesgue measure, so densities cannot be defined in the standard way, and likelihood-based methods typically require the choice of a reference measure [21, 44]. We therefore adopt a density-free perspective: we fit Gaussian mixture models by minimizing the Maximum

[*]Equal contribution and shared first authorship.
[†]Senior author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

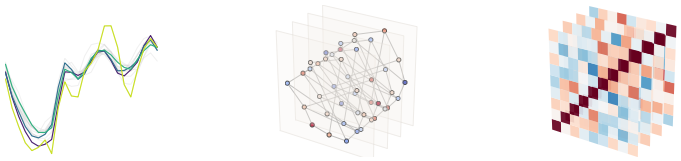


Figure 1: Representative structured data samples. **Left:** Hourly continuous glucose monitoring paths. **Center:** Signals on a fixed graph with varying node intensities. **Right:** Correlation matrices.

Mean Discrepancy (MMD) [35] between the empirical measure and the target mixture model via their kernel mean embeddings.

Model and objective. Let P be an unknown probability measure on a separable Hilbert space $\mathcal{X}$. A Gaussian measure on $\mathcal{X}$, denoted by $\mathcal{N}(m, \mathcal{K})$, is uniquely determined by a mean element $m \in \mathcal{X}$ and a covariance operator $\mathcal{K}\colon \mathcal{X} \to \mathcal{X}$, which generalizes the finite-dimensional covariance matrix. To ensure non-Dirac Gaussian components with finite total variance, $\mathcal{K}$ must be nonzero, self-adjoint, positive semidefinite, and trace-class.

For a fixed number of components $K \in \mathbb{N}$, we approximate P with a finite Gaussian mixture

$$Q_\theta := \sum_{k=1}^{K} \pi_k \mathcal{N}(m_k, \mathcal{K}_k), \qquad \pi_k \geq 0 \quad \forall k, \qquad \sum_{k=1}^{K} \pi_k = 1, \tag{1}$$

parameterized by $\theta := (\pi_k, m_k, \mathcal{K}_k)_{k=1}^{K} \in \Theta_K$, where Θ_K denotes the admissible parameter space.

We fit Q_θ to P by minimizing the Maximum Mean Discrepancy induced by a continuous positive-definite kernel $\kappa\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\mathrm{MMD}_\kappa^2(P, Q_\theta) = \mathbb{E}_{X,X' \sim P}[\kappa(X, X')] - 2\,\mathbb{E}_{X \sim P,\, Y \sim Q_\theta}[\kappa(X, Y)] + \mathbb{E}_{Y,Y' \sim Q_\theta}[\kappa(Y, Y')], \tag{2}$$

where $X, X' \sim P$ and $Y, Y' \sim Q_\theta$ are mutually independent. This leads to a density-free fitting criterion that depends only on kernel expectations. When the kernel κ is *characteristic*, the MMD is a metric on probability measures. Important examples include Gaussian radial kernels on $\mathbb{R}^d$ [63] and, more generally, on separable Hilbert spaces [74]. The formal definitions and statistical background are deferred to Section A in the supplementary material.

In practice, given $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$, we replace P by the empirical measure $P_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$ and solve the problem

$$\min_{\theta \in \Theta_K} \mathrm{MMD}_\kappa^2(P_n, Q_\theta). \tag{3}$$

Notation. Throughout, $\mathcal{X}$ denotes a separable real Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$, norm $\|\cdot\|_{\mathcal{X}}$, and identity operator $I_{\mathcal{X}}$. We let $\mathcal{L}(\mathcal{X})$ denote the space of bounded linear operators on $\mathcal{X}$. For a self-adjoint trace-class operator $T \in \mathcal{L}(\mathcal{X})$ with eigenvalues $(\lambda_j)_{j \geq 1}$, its Fredholm determinant is $\det_F(I_{\mathcal{X}} + T) := \prod_{j=1}^{\infty}(1 + \lambda_j)$. Let $\mathcal{P}(\mathcal{X})$ denote the set of Borel probability measures on $\mathcal{X}$, and for $p \geq 1$, let $\mathcal{P}_p(\mathcal{X})$ be the subset of measures with finite p-th moments. For finite-dimensional spaces, boldface lowercase letters ($\mathbf{x}$, $\mathbf{m}$) denote vectors and boldface uppercase letters ($\mathbf{K}$, $\mathbf{I}$) denote matrices. We write $\mathbf{1}$ for the all-ones vector and Δ^{K-1} for the probability simplex.

1.1 Contributions and paper organization

The main contributions of this paper are threefold:

- **Closed-form MMD fitting for Hilbert-space mixtures.** We derive exact closed-form expressions for $\mathrm{MMD}_\kappa^2(P_n, Q_\theta)$ for Gaussian radial basis and polynomial kernels, extending known $\mathbb{R}^d$ formulas to separable Hilbert spaces [2, 12]. Moreover, we prove that, for every $p \geq 1$, finite Gaussian mixtures of the form (1) are dense in $\mathcal{P}_p(\mathcal{X})$ under the Wasserstein distance, providing theoretical support for the approximation capacity of the proposed model.
- **Consistent finite-dimensional computation.** We introduce a spectral projection framework that reduces infinite-dimensional computation to standard linear algebra in $\mathbb{R}^M$. We prove that both

the projected MMD objective and the posterior probabilities converge to their exact infinite-dimensional counterparts as the truncation dimension M tends to infinity; see Propositions 1 and 2.

- **Empirical validation and digital health case study.** We evaluate the proposed method across diverse data structures, including random functions in L^2, spatial fields, rotations in $\mathrm{SO}(3)$, and graph signals, achieving competitive results against state-of-the-art approaches on established clustering benchmarks (Sections 4 and F). To further illustrate its utility on dynamic functional data, we analyze continuous glucose monitoring (CGM) data from a clinical trial to identify subgroups of individuals with heterogeneous responses to therapy.

All proofs and additional analytical results are deferred to the supplementary material.

Code and Lean 4 proofs: https://github.com/dani2442/rkhs-mixture-models/.

Related work

Functional data analysis and clustering. Functional data analysis (FDA) studies samples naturally viewed as functions, curves, or surfaces, typically using basis expansions, smoothing, and functional principal component analysis [56, 28]. Our spectral projection framework is related to Karhunen–Loève expansions for Gaussian measures [56, 28, 5], but uses projection as a computational approximation to the MMD objective. Model-based functional clustering methods, such as Gaussian mixtures, usually assume latent groups with group-specific low-dimensional structure [10, 19, 38, 33, 37]. Such likelihood-based approaches require a reference measure, which is delicate in infinite dimensions [9, 21, 23]. Distance-based methods [43, 29, 45] avoid this issue, but typically lack the explicit probabilistic and temporal interpretation that a generative model provides.

Kernel mean embeddings and MMD. Kernel mean embeddings represent probability measures as elements of a reproducing kernel Hilbert space (RKHS) [62, 48, 6], inducing the Maximum Mean Discrepancy (MMD) [35]. For characteristic kernels, MMD vanishes if and only if the two laws coincide [63, 60], making it a natural density-free fitting objective. MMD fitting has been used in generative modeling [42, 13], parametric inference [18, 69], kernel-based clustering [31, 46], distributional reinforcement learning [4], and temporal distributional learning [2]. These approaches often rely on closed-form kernel mean embeddings for Gaussian distributions under Gaussian and polynomial kernels [12]; related work also studies MMD testing on function spaces [74]. We extend these formulas and fitting methods from $\mathbb{R}^d$ to separable Hilbert spaces, using characteristic-kernel MMD to fit finite Gaussian mixtures with closed-form objective evaluation, approximation guarantees, and projection-consistent computation.

Mixture models. Gaussian mixture models are a classical tool for model-based clustering and density estimation [24, 30, 16]. Extensions beyond Gaussian components have been proposed [68, 65], and more recent work applies mixtures of Gaussian processes to functional data [70, 36, 47] and to more general infinite-dimensional settings [44]. These methods remain largely likelihood- or projected-density-based; in contrast, we estimate Hilbert-space Gaussian mixtures by minimizing MMD, giving a density-free fitting criterion that is more natural in infinite-dimensional regimes.

Digital health. Digital health technologies generate unprecedented amounts of near-continuous data, motivating dynamic, time-varying functional models [2, 25, 26]. Mixture models for these general data structures remain underdeveloped, despite their potential in clinical trials for identifying subgroups of individuals with heterogeneous treatment responses.

2 MMD fitting

To minimize (3), we evaluate kernel expectations under $\nu_k := \mathcal{N}(m_k, \mathcal{K}_k)$. Expanding the squared MMD between $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $Q_\theta = \sum_{k=1}^K \pi_k \nu_k$ yields

$$\begin{aligned}\mathrm{MMD}_\kappa^2(P_n, Q_\theta) = \frac{1}{n^2}\sum_{i,j=1}^n \kappa(X_i, X_j) - \frac{2}{n}\sum_{i=1}^n \sum_{k=1}^K \pi_k \,\mathbb{E}_{Y\sim\nu_k}[\kappa(X_i, Y)] \\ + \sum_{k,s=1}^K \pi_k \pi_s \,\mathbb{E}_{Y\sim\nu_k,\, Y'\sim\nu_s}[\kappa(Y, Y')].\end{aligned} \tag{4}$$

Because the first data–data term is constant in θ, optimization depends entirely on the data-to-component and component-to-component expectations:

$$J_{i,k} := \mathbb{E}_{Y\sim\nu_k}[\kappa(X_i, Y)], \qquad I_{k,s} := \mathbb{E}_{Y\sim\nu_k,\, Y'\sim\nu_s}[\kappa(Y, Y')]. \tag{5}$$

To make the objective a tractable, differentiable function of θ, we consider Gaussian and polynomial kernels, for which these expectations admit closed-form expressions (Section B):

$$\kappa_A(x, y) = \exp\left(-\frac{1}{2}\langle x - y, A(x - y)\rangle_{\mathcal{X}}\right) \qquad \text{and} \qquad \kappa(x, y) = (\langle x, y\rangle_{\mathcal{X}} + c)^p,$$

where $A \in \mathcal{L}(\mathcal{X})$ is bounded, self-adjoint, and positive semidefinite (often isotropic, $A = \sigma_\kappa^{-2} I_{\mathcal{X}}$), $c \geq 0$, and $p \in \mathbb{N}$. The kernel κ_A is characteristic whenever A is injective [74]. For the polynomial kernel of degree p, the objective depends on Q_θ only through its moment tensors up to order p, so fitting amounts to matching these moments.

2.1 Fitting and optimization

The parameters of the empirical objective (4) naturally split into two groups: mixture weights $\pi \in \Delta^{K-1}$ and component parameters $\{(m_k, \mathcal{K}_k)\}_{k=1}^K$. This enables an alternating scheme that combines a convex weight update with a nonconvex update of the means and covariances.

To make the dependence on π explicit, define the averaged cross-expectation vector $\mathbf{J} \in \mathbb{R}^K$ with entries $J_k := \frac{1}{n}\sum_{i=1}^n J_{i,k}$ and the Gram matrix $\mathbf{I} \in \mathbb{R}^{K\times K}$ with entries $I_{k,s}$. Since the data–data term is constant, minimizing (4) over π for fixed components reduces to the quadratic program

$$\min_{\pi\in\mathbb{R}^K} \; \pi^\top \mathbf{I}\,\pi - 2\mathbf{J}^\top\pi \qquad \text{s.t.} \qquad \pi \geq 0, \quad \mathbf{1}^\top\pi = 1. \tag{6}$$

Because $I_{k,s} = \langle \mu_{\nu_k}, \mu_{\nu_s}\rangle_{\mathcal{H}_\kappa}$ is the inner product of the kernel mean embeddings μ_{ν_k}, μ_{ν_s} in the RKHS $\mathcal{H}_\kappa$, $\mathbf{I}$ is a Gram matrix; hence it is positive semidefinite and the program is convex. Geometrically, the optimal weights $\pi^\star$ give the projection of the empirical mean embedding μ_{P_n} onto the convex hull of the component embeddings $\{\mu_{\nu_k}\}_{k=1}^K$.

Since joint optimization over the full parameter θ remains nonconvex, we employ an alternating minimization strategy, analogous to the expectation–maximization (EM) algorithm [24]. In each iteration, we first solve (6) for π with fixed components, and then fix π to update the means and covariances via a finite-dimensional gradient step (Section 2.2). Pseudocode and implementation details are deferred to Section B.2.

2.2 Numerical computation

For practical implementation, we project the Hilbert-space objects onto a finite-dimensional subspace. Fixing an orthonormal basis $(e_j)_{j\geq 1}$ of $\mathcal{X}$ (e.g., a cosine or data-driven Karhunen–Loève basis [56, 5]), we apply the orthogonal projector $\Pi_M : \mathcal{X} \to \mathcal{X}_M := \operatorname{span}\{e_1, \ldots, e_M\}$. This identifies $\mathcal{X}_M$ with $\mathbb{R}^M$, allowing each object to be represented by its coordinate vector.

For each component $\nu_k = \mathcal{N}(m_k, \mathcal{K}_k)$, let $\mathbf{m}_{k,M} \in \mathbb{R}^M$ and $\mathbf{K}_{k,M} \in \mathbb{R}^{M\times M}$ be the coordinate representations of the projected mean $\Pi_M m_k$ and covariance $\Pi_M \mathcal{K}_k \Pi_M$, and let $\mathbf{x}_{i,M} \in \mathbb{R}^M$ be the coordinates of $\Pi_M X_i$. Under this identification, the pushforward $\nu_{k,M} := \nu_k \circ \Pi_M^{-1}$ is the Gaussian $\mathcal{N}(\mathbf{m}_{k,M}, \mathbf{K}_{k,M})$ on $\mathbb{R}^M$. The projected mixture and empirical measure on $\mathbb{R}^M$ are

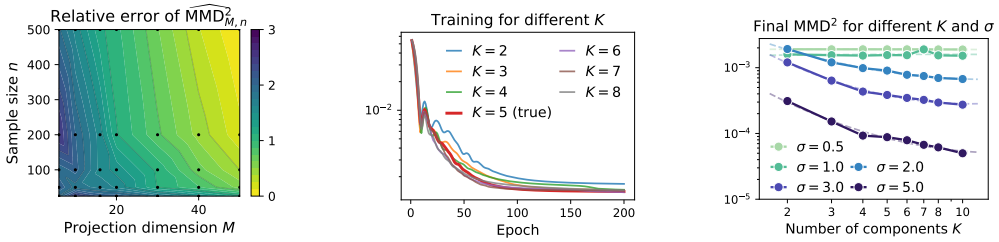
$$Q_{\theta,M} := \sum_{k=1}^K \pi_k \nu_{k,M}, \qquad P_{n,M} := \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{x}_{i,M}}.$$

The projected cross-expectations are

$$J_{i,k}^{(M)} := \mathbb{E}_{Y\sim\nu_k}\big[\kappa(\Pi_M X_i, \Pi_M Y)\big], \qquad I_{k,s}^{(M)} := \mathbb{E}_{Y\sim\nu_k,\, Y'\sim\nu_s}\big[\kappa(\Pi_M Y, \Pi_M Y')\big]. \tag{7}$$

Computationally, these expectations are evaluated exactly in $\mathbb{R}^M$ using the coordinate vectors $\mathbf{x}_{i,M}$ and the finite-dimensional Gaussian measures $\nu_{k,M}$. Explicit matrix formulas for (7) under Gaussian and polynomial kernels are deferred to Section B.1. The finite-dimensional counterpart of (4) is

$$\mathrm{MMD}_\kappa^2(P_{n,M}, Q_{\theta,M}) = \frac{1}{n^2}\sum_{i,j=1}^n \kappa(\Pi_M X_i, \Pi_M X_j) - \frac{2}{n}\sum_{i=1}^n\sum_{k=1}^K \pi_k J_{i,k}^{(M)} + \sum_{k,s=1}^K \pi_k\pi_s I_{k,s}^{(M)}. \tag{8}$$



(a) MMD^2 versus projection dimension M and sample size n.

(b) Training convergence for different numbers of components K.

(c) Final MMD^2 across component counts K and kernel bandwidths σ.

Figure 2: Numerical sensitivity of the MMD objective to various hyperparameters and sample sizes.

Disregarding the constant data–data term, the per-iteration cost of evaluating (8) is $O(K^2M^3 + nKM^2)$ for both the Gaussian radial and polynomial kernels. Moreover, if the projected covariances are assumed diagonal, the cost reduces to $O(K^2M + nKM)$, making the method scalable to large sample sizes and projection dimensions. CPU timings ($n = 10^5$, $K \leq 200$) and details on numerical conditioning appear in Sections H.2 and H.3.

Proposition 1 (Consistency of projected MMD)**.** *Let κ be a continuous kernel on $\mathcal{X}$. Assume there exist constants $C, q \geq 0$ such that*

$$|\kappa(x, y)| \leq C\big(1 + \|x\|_{\mathcal{X}}^q\big)\big(1 + \|y\|_{\mathcal{X}}^q\big) \qquad \textit{for all } x, y \in \mathcal{X}. \tag{9}$$

Then, the projected MMD objective converges to the exact MMD:

$$\mathrm{MMD}^2_\kappa(P_{n,M}, Q_{\theta,M}) \xrightarrow{M\to\infty} \mathrm{MMD}^2_\kappa(P_n, Q_\theta) \qquad \textit{a.s.} \tag{10}$$

Together with the known empirical stability of MMD that controls the error from replacing P by P_n (see [69, Proposition A.1]), Proposition 1 shows that the finite-dimensional surrogate (8) first approximates the Hilbert-space empirical criterion as $M \to \infty$, and then the population criterion as $n \to \infty$. This justifies (8) as a computationally tractable proxy for the infinite-dimensional MMD objective (see Figure 2 for an empirical illustration).

Posterior probabilities. The mixture admits the standard generative interpretation: draw a latent assignment $Z \in \{1, \ldots, K\}$ with prior $\mathbb{P}(Z = k) = \pi_k$, and then sample $X \mid \{Z = k\} \sim \nu_k$. The *posterior component probability*, or responsibility, is defined by the Radon–Nikodym derivative

$$\gamma_k(X) := \mathbb{P}(Z = k \mid X) = \frac{\mathrm{d}(\pi_k \nu_k)}{\mathrm{d}Q_\theta}(X), \qquad Q_\theta\text{-a.s.}, \tag{11}$$

which is well-defined because $\pi_k \nu_k \ll Q_\theta$ whenever $\pi_k > 0$. In the finite-dimensional setting, whenever all projected covariances $\mathbf{K}_{s,M}$ are nonsingular, each $\nu_{s,M}$ admits a Lebesgue density $p_{s,M}$ on $\mathbb{R}^M$ and (11) reduces to the familiar Bayes formula

$$\gamma_k^{(M)}(x) = \frac{\pi_k\, p_{k,M}(x)}{\sum_{s=1}^K \pi_s\, p_{s,M}(x)}. \tag{12}$$

As M grows, $\sigma(\Pi_M X)$ accumulates the information of $\sigma(X)$, yielding the following convergence.

Proposition 2. *Let (Z, X) be as above and let $\gamma_k^{(M)} := \frac{\mathrm{d}(\pi_k \nu_{k,M})}{\mathrm{d}Q_{\theta,M}}$. Then the projected responsibility satisfies the martingale property*

$$\gamma_k^{(M)}(\Pi_M X) = \mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \sigma(\Pi_M X)] \quad \textit{a.s.}, \tag{13}$$

and $\gamma_k^{(M)}(\Pi_M X) \to \gamma_k(X)$ as $M \to \infty$ both almost surely and in L^1.

The proofs of Propositions 1 and 2 are deferred to Section I.1 and Section I.2, respectively.

3 Approximation theory

In this section, we establish theoretical guarantees that justify the use of finite Gaussian mixtures in Hilbert spaces as a flexible approximation class.

Proposition 3 (Wasserstein and MMD density)**.** *Let $\mathcal{X} \neq \{0\}$ be a separable Hilbert space.*

(i) For every $p \geq 1$, finite mixtures of Gaussian measures on $\mathcal{X}$ of the form (1) are dense in $\mathcal{P}_p(\mathcal{X})$ with respect to the Wasserstein distance W_p.

(ii) Let κ be a continuous positive-definite kernel satisfying $\sup_{x\in\mathcal{X}} \kappa(x,x) \leq 1$. Then, for all $P \in \mathcal{P}(\mathcal{X})$ and all $K \geq 1$, there exists Q_K of the form (1) such that

$$\mathrm{MMD}_\kappa(P, Q_K) \leq \frac{2}{\sqrt{K}}.$$

Part (i) shows that the model class is expressive under Wasserstein metrics, which are widely used in mathematical statistics to study the convergence of estimators, resampling schemes such as the bootstrap, and related distributional properties [52]. This result also implies MMD density for characteristic bounded continuous kernels. Part (ii), however, applies to arbitrary probability measures in $\mathcal{P}(\mathcal{X})$ and provides an explicit dimension-free approximation rate for the MMD minimized by our algorithm. The proof of Proposition 3 is deferred to Section I.3.

4 Numerical experiments

The first goal of this section is to assess the empirical clustering performance of our method against established approaches in both finite- and infinite-dimensional regimes. We then focus on a digital health case study to illustrate the versatility of our framework for analyzing complex clinical biomarkers. In particular, we summarize glucose time series using different representations, including random functions and graphs, which can be viewed as dynamic functional objects. Additional numerical comparisons are reported in Section F.

4.1 Clustering benchmark

We compare the mixture model described in Section 2.2 against several baselines: (i) partitioning and centroid-based methods, including K-Medoids and K-Means [58, 59]; (ii) agglomerative hierarchical methods, including average-linkage, Ward, and BIRCH [49, 72, 75]; (iii) density-based methods, including DBSCAN and OPTICS [27, 3]; (iv) model-based methods, including Gaussian mixtures [30, 44]; and (v) graph-based and dimension-reduction methods, such as spectral clustering [50].

Datasets. We evaluate our approach using representative synthetic and real-world datasets.

- *Synthetic environments:* (i) $\mathbb{R}^d$: five standard `sklearn` clustering datasets with $n = 500$, $d = 2$, and $K \in \{2, 3\}$ [53]; and (ii) $\mathrm{SO}(3)$: ten synthetic datasets of rotations under an Euler-angle parameterization, projected onto Wigner-D matrix coefficients, with $n = 200$ and $K = 3$.
- *Real-world datasets:* (iii) L^2: six standard functional data benchmarks [56, 11, 28, 54, 51]; (iv) CGM: continuous glucose monitoring curves from $n = 98$ individuals with $K = 2$, detailed in Section 4.2; and (v) Molecular: MUTAG graph signals with $n = 188$ and $K = 2$ [22].

All implementation details and evaluation protocols are reported in Section G. Performance is evaluated using the Adjusted Rand Index (ARI), which compares unsupervised cluster assignments against ground-truth labels. We fix K to the known number of classes, as our goal here is to compare clustering accuracy.

Results. Although designed for general random objects in Hilbert spaces, our method is competitive on finite-dimensional $\mathbb{R}^d$ benchmarks and achieves the best performance on the CGM and $\mathrm{SO}(3)$ datasets, which fall outside the scope of most existing clustering methods. It also performs well in the infinite-dimensional L^2 setting. Overall, the proposed model combines strong empirical accuracy and computational efficiency with a unified probabilistic interpretation. This flexibility enables time-dependent interpretations such as the digital health case study below.

Table 1: Benchmark of clustering algorithms comparing the operating spaces; the computational complexity of training, inference, and memory as functions of the number of samples n and clusters K; and clustering performance (ARI, mean_{std}) across several datasets grouped by geometric space.

Method	Space	Train	Infer	Memory	$\mathbb{R}^d$	L^2	CGM	SO(3)	Molecular
MMD GMM (Gaussian)	Hilbert	$O(nK)$	$O(nK)$	$O(nK)$	$0.692_{0.379}$	$0.400_{0.202}$	$\mathbf{0.053}_{0.040}$	$\mathbf{0.867}_{0.140}$	$0.104_{0.032}$
MMD GMM (Polynomial)	Hilbert	$O(nK)$	$O(nK)$	$O(nK)$	$0.615_{0.334}$	$0.373_{0.132}$	$0.051_{0.022}$	$0.675_{0.217}$	$\mathbf{0.151}_{0.009}$
Projected GMM (Appendix D)	Hilbert	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.345_{0.268}$	$0.034_{0.000}$	$0.828_{0.172}$	$0.104_{0.059}$
K-Medoids [39, 58]	Metric	$O(n^2)$	$O(nK)$	$O(n^2)$	$0.569_{0.328}$	$\mathbf{0.483}_{0.192}$	$0.020_{0.000}$	$0.828_{0.109}$	$-0.005_{0.000}$
Hierarchical (Avg) [29, 41]	Metric	$O(n^2)$	—	$O(n^2)$	$0.555_{0.320}$	$0.211_{0.275}$	$0.002_{0.000}$	$0.490_{0.315}$	$0.081_{0.000}$
DBSCAN [27]	Metric	$O(n^2)$	—	$O(n^2)$	$0.844_{0.164}$	$0.000_{0.000}$	$0.000_{0.000}$	$0.627_{0.273}$	$0.074_{0.000}$
HDBSCAN [14]	Metric	$O(n^2)$	—	$O(n^2)$	$0.870_{0.171}$	$0.143_{0.217}$	$0.000_{0.000}$	$0.691_{0.227}$	$0.117_{0.000}$
K-Center [34]	Metric	$O(nK)$	$O(nK)$	$O(n)$	$0.392_{0.258}$	$0.162_{0.109}$	$0.002_{0.000}$	$0.188_{0.169}$	$0.074_{0.000}$
Kernel k-Groups [31]	Metric	$O(Kn^2)$	—	$O(n^2)$	$0.567_{0.328}$	$0.405_{0.202}$	$0.019_{0.000}$	$0.751_{0.189}$	$0.125_{0.000}$
GPmix [1]	L^2	$O(nK + n^3)$	—	$O(n^2)$	—	$0.364_{0.237}$	—	—	—
FDA K-Means [43]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.451_{0.222}$	$0.023_{0.000}$	—	—
FDA Fuzzy C-Means [7]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.390_{0.154}$	$0.020_{0.000}$	—	—
Funclust [37]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.429_{0.157}$	$0.015_{0.000}$	—	—
FunHDDC [10]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.420_{0.138}$	$0.013_{0.000}$	—	—
fclust [38]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.421_{0.240}$	$0.030_{0.000}$	—	—
K-Centres [19]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.401_{0.178}$	$0.001_{0.000}$	—	—
Curvclust [33]	L^2	$O(nK)$	$O(nK)$	$O(nK)$	—	$0.432_{0.191}$	$0.012_{0.000}$	—	—
MiniBatch KMeans [59]	$\mathbb{R}^d$	$O(nK)$	$O(nK)$	$O(n)$	$0.554_{0.327}$	—	—	—	—
Spectral [50]	$\mathbb{R}^d$	$O(n^3)$	—	$O(n^2)$	$\mathbf{0.949}_{0.074}$	—	—	—	—
Ward [72]	$\mathbb{R}^d$	$O(n^2)$	—	$O(n^2)$	$0.613_{0.330}$	—	—	—	—
BIRCH [75]	$\mathbb{R}^d$	$O(n)$	$O(nK)$	$O(n)$	$0.525_{0.303}$	—	—	—	—
Gaussian Mixture (EM) [30]	$\mathbb{R}^d$	$O(nK)$	$O(nK)$	$O(nK)$	$0.683_{0.387}$	—	—	—	—
Affinity Propagation [32]	$\mathbb{R}^d$	$O(n^2)$	$O(nK)$	$O(n^2)$	$0.489_{0.308}$	—	—	—	—
MeanShift [17]	$\mathbb{R}^d$	$O(n^2)$	$O(nK)$	$O(n)$	$0.564_{0.328}$	—	—	—	—
OPTICS [3]	$\mathbb{R}^d$	$O(n^2)$	—	$O(n)$	$0.809_{0.219}$	—	—	—	—

4.2 Case study: different representations of digital health time series

We now illustrate the proposed model in a digital health case study based on longitudinal continuous glucose monitoring (CGM) data from a pediatric clinical trial [71]. A central question in longitudinal studies is whether and how the distribution of individuals changes over time, for example in response to treatment [40, 20]. In the original study [71], the authors showed that a new insulin pump system improves several summary glycemic metrics after the intervention. Here, we examine these outcomes in a continuous-time formulation using dynamic functional representations of high-resolution glucose trajectories. Starting from multi-day glucose time series, we construct three daily representations: (i) functional observations in H^1; (ii) correlation matrices describing the dependence between glucose values at different hours of the day; and (iii) signals on a similarity graph between individuals. Similar representations have been considered as biomarkers in clinical fMRI studies [64, 73].

To analyze temporal changes across these three CGM representations, we extend (1) by allowing the mixture weights to vary over time while keeping the component means and covariance operators time-invariant. Our formulation is related to that of [2]; however, whereas that work models time-evolving distributions of finite-dimensional random vectors, our framework applies to general separable Hilbert spaces.

Background and data. Continuous glucose monitors record interstitial glucose concentration every 5 minutes, producing 288 measurements per day. The data come from the Pediatric Artificial Pancreas (PEDAP) randomized controlled trial [71], which evaluated a hybrid closed-loop insulin delivery system in young children with type 1 diabetes; the CGM records span more than six months. For each individual, a sliding window of W days with a stride of one day produces one observation per window. Within each window, the 288 time slots are averaged across valid days; slots with fewer than $\lceil 0.5\,W \rceil$ valid readings are linearly interpolated from neighboring slots. This yields one intraday glucose curve $f\colon [0, 24] \to \mathbb{R}$ per individual and window. After preprocessing, the dataset comprises 98 individuals, with 34 in the control group and 64 in the treatment group.

Time-varying mixture weights. Let $X_1, \ldots, X_n$ be i.i.d. copies of an $\mathcal{X}$-valued measurable stochastic process $X(t)$. For each t, denote by $P(t)$ the law of $X(t)$, and define the empirical

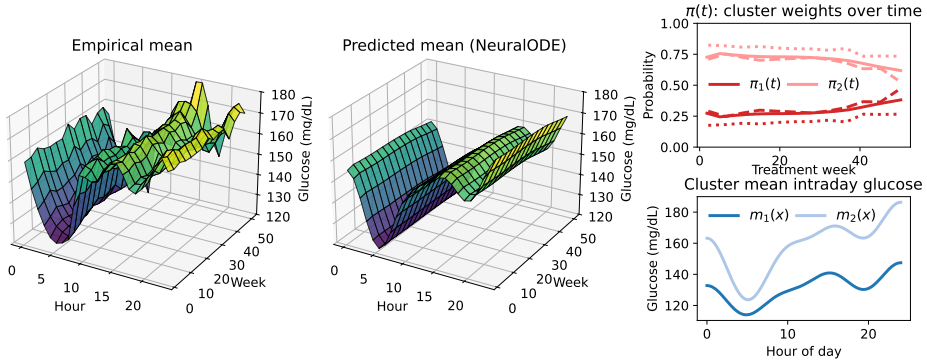


Figure 3: Temporal glucose mixture in H^1. **Left:** Empirical (**top**) and model-predicted (**bottom**) mean glucose surfaces. **Top right:** Global cluster weights $\pi_k(t)$ (solid lines) with group-level posteriors $\bar{\gamma}_{\text{ctrl}}(t)$ for control (dotted) and $\bar{\gamma}_{\text{treat}}(t)$ for treatment (dashed). **Bottom right:** Learned means.

distribution by $P_n(t) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(t)}$. We fit the model

$$Q_\theta(t) \coloneqq \sum_{k=1}^{K} \pi_k(t)\,\mathcal{N}(m_k, \mathcal{K}_k), \qquad \pi(t) \coloneqq \big(\pi_1(t), \ldots, \pi_K(t)\big) \in \Delta^{K-1}. \tag{14}$$

Conceptually, the shared components $(m_k, \mathcal{K}_k)$ capture stationary data archetypes, such as distinct physiological regimes or individual phenotypes, while the time-varying weight vector $\pi(t)$ describes population-level shifts between these archetypes over time.

Following the kernel k-groups framework [31], we select K by evaluating the within-cluster MMD across candidate values (see Section C). Model fitting minimizes the integrated MMD^2,

$$\min_{\theta(t)=\{\pi_k(t), m_k, \mathcal{K}_k\}_k} \frac{1}{T}\int_0^T \text{MMD}_\kappa^2\big(P_n(t), Q_\theta(t)\big)\,\mathrm{d}t. \tag{15}$$

We parameterize $\pi(t) = \text{softmax}\big(\mathbf{z}(t)\big)$, where the logit path $\mathbf{z}\colon [0,T] \to \mathbb{R}^K$ is the solution to a neural ODE [15], initialized at $\mathbf{z}(0) = \log \pi(0)$. This parameterization introduces a natural bias toward smooth temporal dynamics. Further details on density and consistency results, the projected $\mathbb{R}^M$ approximation of the integrated objective, and implementation specifics are deferred to Section C.

Group-level comparison. To compare the control and treatment arms, we compute each individual's posterior cluster membership $\gamma_k(X_i(t), t)$ at each time point and aggregate these memberships into group-level trajectories. Specifically, for each group $g \in \{\text{ctrl}, \text{treat}\}$ (control and treatment, respectively), we compute

$$\bar{\gamma}_g(t) = (\bar{\gamma}_{g,1}(t), \ldots, \bar{\gamma}_{g,K}(t)), \qquad \bar{\gamma}_{g,k}(t) \coloneqq \frac{1}{|G_g(t)|}\sum_{i \in G_g(t)} \gamma_k(X_i(t), t),$$

where $G_g(t)$ denotes the set of individuals in group g observed at time t. We then quantify the divergence between the study arms using the total variation distance $\text{TV}(t) = \frac{1}{2}\left\|\bar{\gamma}_{\text{treat}}(t) - \bar{\gamma}_{\text{ctrl}}(t)\right\|_1$. This metric ranges from 0 to 1, where 0 reflects identical group-level cluster usage and 1 indicates complete separation, providing an interpretable measure of treatment-induced distributional shifts over time.

4.2.1 Results

Intraday curves. We view each curve as an element of the Sobolev space $H^1(0, 24; \mathbb{R})$, project it onto a truncated cosine basis, and fit the model via (15) with $K = 2$ latent clusters. The two learned means contrast a flatter, well-regulated profile (m_1) with an elevated, variable profile (m_2); the empirical and model-predicted mean surfaces agree closely (Figure 3, left). While the control

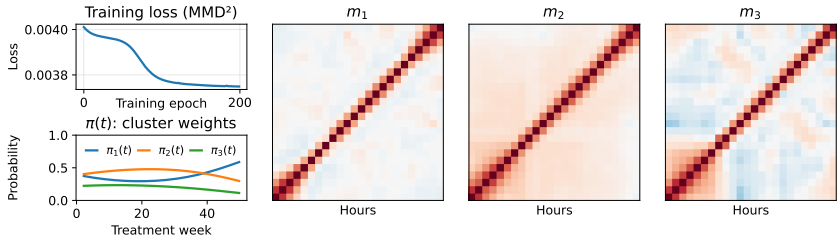


Figure 4: Correlation-based temporal mixture in $\mathrm{Sym}(24)$. **Left:** MMD^2 loss versus training epoch (**top**) and cluster weights $\pi_k(t)$ versus treatment week (**bottom**). **Right:** Learned mean matrices in $\mathrm{Sym}(24)$.

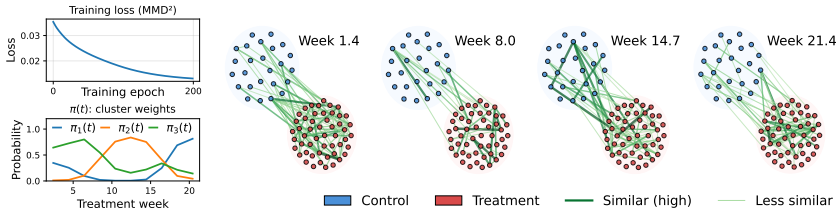


Figure 5: Temporal mixture on the individual similarity graph. **Left:** MMD^2 loss versus training epoch (**top**) and cluster weights $\pi_k(t)$ versus treatment week (**bottom**). **Right:** Network snapshots at four time points; node color distinguishes control (blue) from treatment (red), and edge width encodes similarity.

group remains approximately time-invariant, the treatment group gradually shifts toward m_1, with the between-group TV distance increasing from ≈ 0.11 at baseline to ≈ 0.19 at the end of the study, reflecting a progressive improvement in glycemic regulation.

Correlation matrices. Instead of clustering intraday curves, we cluster the running 24×24 *inter-hour correlation matrices* to capture how the *dependence structure* between hours of the day evolves over the study. Each sliding window ($W = 14$ days) is summarized into 24 hourly averages per day, from which the sample correlation matrix is estimated via linear shrinkage toward the identity ($\alpha = 0.1$) and projected onto the positive semidefinite cone. Since correlation matrices do not form a Hilbert space, we embed them in the ambient space $\mathrm{Sym}(24)$ of symmetric 24×24 matrices, equipped with the Frobenius inner product, and fit the temporal mixture model with $K = 3$ clusters. The learned clusters in Figure 4 identify three inter-hour dependence patterns: moderate local correlation (m_1), broadly strong correlation (m_2), and sharper early-hour dependence (m_3). Over time, membership shifts from m_3 toward the more regulated m_1 regime, consistent with improved glycemic regulation.

Individual similarity graph signals. We now move from hour- to individual-level structure. For each sliding window ($W = 21$ days), we compute pairwise H^1 similarities between individuals to obtain a time-varying similarity matrix $S(t) \in \mathbb{R}^{N \times N}$. A fixed individual graph is then built by thresholding the time-averaged similarity matrix at the 85th percentile, and the graph Laplacian eigenbasis (using $R = 4$ eigenvectors) is used to project each $S(t)$ onto graph signals. We then cluster these projected matrices over time with a $K = 3$ mixture model. As the trial progresses (Figure 5), the network structure evolves from diffuse inter-group connections toward denser within-group similarity in the treatment arm, consistent with individuals progressively achieving improved glycemic control.

5 Final remarks

The main contribution of this paper is a Gaussian mixture model for random elements in separable Hilbert spaces, estimated via an MMD-based optimization procedure. A key methodological and practical advantage of the proposed approach is that it bypasses the need for densities or dominating measures, which are generally intractable in infinite-dimensional settings. In addition, the frame-

work flexibly accommodates diverse data structures, is computationally efficient, and demonstrates competitive empirical performance.

Future work will focus on developing specialized temporal algorithms for specific Hilbert spaces, extending the framework to general metric spaces, and quantifying uncertainty in both parameter estimates and model predictions. On the theoretical side, building upon the methodological foundation established here, deriving convergence guarantees for the proposed algorithms, proving statistical consistency of the fitted parameters, and characterizing the identifiability of the mixture decomposition represent important avenues for future investigation.

Acknowledgments and Disclosure of Funding

A. Álvarez-López (AÁ) was supported by an FPU fellowship (FPU21/05673) from the Spanish Ministry of Universities. D. López-Montero (DL) was supported by the Alexander von Humboldt Professorship programme of the Alexander von Humboldt Foundation, funded by the German Federal Ministry of Education and Research (BMBF), at the Chair for Dynamics, Control, Machine Learning, and Numerics, Friedrich-Alexander-Universität Erlangen-Nürnberg. AÁ and DL were also supported by project PID2023-146872OB-I00 of MICIU, Spain.

The authors declare no competing interests.

References

[1] Emmanuel Akeweje and Mimi Zhang. "Learning Mixtures of Gaussian Processes through Random Projection". In: *Proceedings of the 41st International Conference on Machine Learning*. Vol. 235. Proceedings of Machine Learning Research. PMLR, 2024, pp. 720–739. URL: https://proceedings.mlr.press/v235/akeweje24a.html.

[2] Antonio Álvarez-López and Marcos Matabuena. *Continuous-Time Learning of Probability Distributions: A Case Study in a Digital Trial of Young Children with Type 1 Diabetes*. arXiv:2603.24427. 2026. arXiv: 2603.24427. URL: https://arxiv.org/abs/2603.24427.

[3] Mihael Ankerst et al. "OPTICS: Ordering Points to Identify the Clustering Structure". In: *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*. ACM, 1999, pp. 49–60. DOI: 10.1145/304182.304187.

[4] Mathis Antonetti, Henrique Donâncio, and Florence Forbes. "An Analysis of Distributional Reinforcement Learning with Gaussian Mixtures". In: *Transactions on Machine Learning Research* (2026). URL: https://openreview.net/forum?id=b4VgI1RTv8.

[5] Xavier Bay and Jean-Charles Croix. "Karhunen–Loève Decomposition of Gaussian Measures on Banach Spaces". In: *Probability and Mathematical Statistics* 39.2 (2019), pp. 279–297. DOI: 10.19195/0208-4147.39.2.3.

[6] Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Springer US, 2004. DOI: 10.1007/978-1-4419-9096-9.

[7] James C. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Springer US, 1981. DOI: 10.1007/978-1-4757-0450-1.

[8] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006.

[9] Vladimir I. Bogachev. *Gaussian Measures*. Vol. 62. Mathematical Surveys and Monographs. American Mathematical Society, 1998. DOI: 10.1090/surv/062.

[10] Charles Bouveyron and Julien Jacques. "Model-Based Clustering of Time Series in Group-Specific Functional Subspaces". In: *Advances in Data Analysis and Classification* 5.4 (2011), pp. 281–300. DOI: 10.1007/s11634-011-0095-6.

[11] Leo Breiman et al. *Classification and Regression Trees*. Routledge, 2017. DOI: 10.1201/9781315139470.

[12] François-Xavier Briol et al. "A Dictionary of Closed-Form Kernel Mean Embeddings". In: *Proceedings of the First International Conference on Probabilistic Numerics*. Vol. 271. Proceedings of Machine Learning Research. PMLR, 2025, pp. 84–94. URL: https://proceedings.mlr.press/v271/briol25a.html.

[13] François-Xavier Briol et al. *Statistical Inference for Generative Models with Maximum Mean Discrepancy*. arXiv:1906.05944. 2019. arXiv: 1906.05944. URL: https://arxiv.org/abs/1906.05944.

[14] Ricardo J. G. B. Campello et al. "Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection". In: *ACM Transactions on Knowledge Discovery from Data* 10.1 (2015), 5:1–5:51. DOI: 10.1145/2733381.

[15] Ricky T. Q. Chen et al. "Neural Ordinary Differential Equations". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc., 2018.

[16] Xin Chen and Anderson Ye Zhang. "Achieving Optimal Clustering in Gaussian Mixture Models with Anisotropic Covariance Structures". In: *Advances in Neural Information Processing Systems*. Vol. 37. Curran Associates, Inc., 2024. DOI: 10.52202/079017-3610.

[17] Yizong Cheng. "Mean Shift, Mode Seeking, and Clustering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.8 (1995), pp. 790–799. DOI: 10.1109/34.400568.

[18] Badr-Eddine Chérief-Abdellatif and Pierre Alquier. "Finite Sample Properties of Parametric MMD Estimation: Robustness to Misspecification and Dependence". In: *Bernoulli* 28.1 (2022), pp. 181–213. DOI: 10.3150/21-BEJ1338.

[19] Jeng-Min Chiou and Pai-Ling Li. "Functional Clustering and Identifying Substructures of Longitudinal Data". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 69.4 (2007), pp. 679–699. DOI: 10.1111/j.1467-9868.2007.00605.x.

[20] William Clarke and Boris Kovatchev. "Statistical Tools to Analyze Continuous Glucose Monitor Data". In: *Diabetes Technology & Therapeutics* 11.Supplement 1 (2009), S45–S54. DOI: 10.1089/dia.2008.0138.

[21] Giuseppe Da Prato and Jerzy Zabczyk. *Stochastic Equations in Infinite Dimensions*. 2nd ed. Vol. 152. Encyclopedia of Mathematics and Its Applications. Cambridge University Press, 2014. DOI: 10.1017/CBO9781107295513.

[22] Asim Kumar Debnath et al. "Structure-Activity Relationship of Mutagenic Aromatic and Heteroaromatic Nitro Compounds. Correlation with Molecular Orbital Energies and Hydrophobicity". In: *Journal of Medicinal Chemistry* 34.2 (1991), pp. 786–797. DOI: 10.1021/jm00106a046.

[23] Aurore Delaigle and Peter Hall. "Defining Probability Density for a Distribution of Random Functions". In: *The Annals of Statistics* 38.2 (2010), pp. 1171–1193. DOI: 10.1214/09-AOS741.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39.1 (1977), pp. 1–22. DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[25] Paromita Dubey and Hans-Georg Müller. "Functional Models for Time-Varying Random Objects". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 82.2 (2020), pp. 275–327. DOI: 10.1111/rssb.12337.

[26] Paromita Dubey and Hans-Georg Müller. "Modeling Time-Varying Random Objects and Dynamic Networks". In: *Journal of the American Statistical Association* 117.540 (2022), pp. 2252–2267. DOI: 10.1080/01621459.2021.1917416.

[27] Martin Ester et al. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. AAAI Press, 1996, pp. 226–231.

[28] Frédéric Ferraty and Philippe Vieu. *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Series in Statistics. Springer New York, 2006. DOI: 10.1007/0-387-36620-2.

[29] Laura Ferreira and David B. Hitchcock. "A Comparison of Hierarchical Methods for Clustering Functional Data". In: *Communications in Statistics - Simulation and Computation* 38.9 (2009), pp. 1925–1949. DOI: 10.1080/03610910903168603.

[30] Chris Fraley and Adrian E Raftery. "Model-Based Clustering, Discriminant Analysis, and Density Estimation". In: *Journal of the American Statistical Association* 97.458 (2002), pp. 611–631. DOI: 10.1198/016214502760047131.

[31] Guilherme França, Maria L. Rizzo, and Joshua T. Vogelstein. "Kernel K-Groups via Hartigan's Method". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2021), pp. 4411–4425. DOI: 10.1109/TPAMI.2020.2998120.

[32] Brendan J. Frey and Delbert Dueck. “Clustering by Passing Messages Between Data Points”. In: *Science* 315.5814 (2007), pp. 972–976. DOI: 10.1126/science.1136800.

[33] M. Giacofci et al. “Wavelet-Based Clustering for Mixed-Effects Functional Models in High Dimension”. In: *Biometrics* 69.1 (2013), pp. 31–40. DOI: 10.1111/j.1541-0420.2012.01828.x.

[34] Teofilo F. Gonzalez. “Clustering to Minimize the Maximum Intercluster Distance”. In: *Theoretical Computer Science* 38 (1985), pp. 293–306. DOI: 10.1016/0304-3975(85)90224-5.

[35] Arthur Gretton et al. “A Kernel Two-Sample Test”. In: *Journal of Machine Learning Research* 13.25 (2012), pp. 723–773.

[36] Mian Huang et al. “Estimating Mixture of Gaussian Processes by Kernel Smoothing”. In: *Journal of Business & Economic Statistics* 32.2 (2014), pp. 259–270. DOI: 10.1080/07350015.2013.868084.

[37] Julien Jacques and Cristian Preda. “Funclust: A Curves Clustering Method Using Functional Random Variables Density Approximation”. In: *Neurocomputing* 112 (2013), pp. 164–171. DOI: 10.1016/j.neucom.2012.11.042.

[38] Gareth M James and Catherine A Sugar. “Clustering for Sparsely Sampled Functional Data”. In: *Journal of the American Statistical Association* 98.462 (2003), pp. 397–408. DOI: 10.1198/016214503000189.

[39] Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley, 1990. DOI: 10.1002/9780470316801.

[40] Jan S Krouwer and George S Cembrowski. “A Review of Standards and Statistics Used to Describe Blood Glucose Monitor Performance”. In: *Journal of Diabetes Science and Technology* 4.1 (2010), pp. 75–83. DOI: 10.1177/193229681000400110.

[41] G. N. Lance and W. T. Williams. “A General Theory of Classificatory Sorting Strategies: 1. Hierarchical Systems”. In: *The Computer Journal* 9.4 (1967), pp. 373–380. DOI: 10.1093/comjnl/9.4.373.

[42] Yujia Li, Kevin Swersky, and Rich Zemel. “Generative Moment Matching Networks”. In: *Proceedings of the 32nd International Conference on Machine Learning*. Vol. 37. PMLR, 2015, pp. 1718–1727. URL: https://proceedings.mlr.press/v37/li15.html.

[43] María Luz López García, Ricardo García-Ródenas, and Antonia González Gómez. “K-Means Algorithms for Functional Data”. In: *Neurocomputing* 151 (2015), pp. 231–245. DOI: 10.1016/j.neucom.2014.09.048.

[44] Haoyu Lu, Junxiong Jia, and Deyu Meng. *Sequential Monte Carlo with Gaussian Mixture Approximation for Infinite-Dimensional Statistical Inverse Problems*. arXiv:2503.16028 (version 5, 2026). 2026. arXiv: 2503.16028. URL: https://arxiv.org/abs/2503.16028.

[45] Qianli Ma et al. “Learning Representations for Time Series Clustering”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[46] Marcos Matabuena et al. “Kernel Biclustering Algorithm in Hilbert Spaces”. In: *Advances in Data Analysis and Classification* 20.2 (2026), pp. 479–520. DOI: 10.1007/s11634-024-00619-0.

[47] Ian C. McDowell et al. “Clustering Gene Expression Time Series Data Using an Infinite Gaussian Process Mixture Model”. In: *PLOS Computational Biology* 14.1 (2018), e1005896. DOI: 10.1371/journal.pcbi.1005896.

[48] Krikamol Muandet et al. “Kernel Mean Embedding of Distributions: A Review and Beyond”. In: *Foundations and Trends in Machine Learning* 10.1–2 (2017), pp. 1–141. DOI: 10.1561/2200000060.

[49] Fionn Murtagh and Pedro Contreras. “Algorithms for Hierarchical Clustering: An Overview”. In: *WIREs Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97. DOI: 10.1002/widm.53.

[50] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. “On Spectral Clustering: Analysis and an Algorithm”. In: *Advances in Neural Information Processing Systems*. Vol. 14. MIT Press, 2001, pp. 849–856.

[51] Robert Thomas Olszewski. “Generalized Feature Extraction for Structural Pattern Recognition in Time-Series Data”. PhD thesis. Carnegie Mellon University, 2001.

[52] Victor M. Panaretos and Yoav Zemel. "Statistical Aspects of Wasserstein Distances". In: *Annual Review of Statistics and Its Application* 6 (2019), pp. 405–431. DOI: 10.1146/annurev-statistics-030718-104938.

[53] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.

[54] Cristian Preda, Quentin Grimonprez, and Vincent Vandewalle. "Categorical Functional Data Analysis. The cfda R Package". In: *Mathematics* 9.23 (2021), p. 3074. DOI: 10.3390/math9233074.

[55] Carlos Ramos-Carreño et al. "scikit-fda: A Python Package for Functional Data Analysis". In: *Journal of Statistical Software* 109.2 (2024), pp. 1–37. DOI: 10.18637/jss.v109.i02.

[56] J. O. Ramsay and B. W. Silverman. *Functional Data Analysis*. 2nd ed. Springer Series in Statistics. Springer New York, 2005. DOI: 10.1007/b98888.

[57] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006.

[58] Erich Schubert and Peter J. Rousseeuw. "Fast and Eager k-Medoids Clustering: $O(k)$ Runtime Improvement of the PAM, CLARA, and CLARANS Algorithms". In: *Information Systems* 101 (2021), p. 101804. DOI: 10.1016/j.is.2021.101804.

[59] D. Sculley. "Web-Scale k-Means Clustering". In: *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 1177–1178. DOI: 10.1145/1772690.1772862.

[60] Dino Sejdinovic et al. "Equivalence of Distance-Based and RKHS-based Statistics in Hypothesis Testing". In: *The Annals of Statistics* 41.5 (2013), pp. 2263–2291. DOI: 10.1214/13-AOS1140.

[61] Amir Shahroudy et al. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1010–1019. DOI: 10.1109/CVPR.2016.115.

[62] Alex Smola et al. "A Hilbert Space Embedding for Distributions". In: *Algorithmic Learning Theory*. Ed. by Marcus Hutter, Rocco A. Servedio, and Eiji Takimoto. Springer, 2007, pp. 13–31. DOI: 10.1007/978-3-540-75225-7_5.

[63] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. "Universality, Characteristic Kernels and RKHS Embedding of Measures". In: *Journal of Machine Learning Research* 12.70 (2011), pp. 2389–2410.

[64] Kaustubh Supekar et al. "Network Analysis of Intrinsic Functional Brain Connectivity in Alzheimer's Disease". In: *PLOS Computational Biology* 4.6 (2008), e1000100. DOI: 10.1371/journal.pcbi.1000100.

[65] Chengliang Tang et al. "Wasserstein Distributional Learning via Majorization-Minimization". In: *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*. Vol. 206. PMLR, 2023, pp. 10703–10731. URL: https://proceedings.mlr.press/v206/tang23b.html.

[66] Robert L. Thorndike. "Who Belongs in the Family?" In: *Psychometrika* 18.4 (1953), pp. 267–276. DOI: 10.1007/BF02289263.

[67] Robert Tibshirani, Guenther Walther, and Trevor Hastie. "Estimating the Number of Clusters in a Data Set via the Gap Statistic". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 63.2 (2001), pp. 411–423. DOI: 10.1111/1467-9868.00293.

[68] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[69] Ilya Tolstikhin, Bharath K. Sriperumbudur, and Krikamol Muandet. "Minimax Estimation of Kernel Mean Embeddings". In: *Journal of Machine Learning Research* 18.86 (2017), pp. 1–47.

[70] Volker Tresp. "Mixtures of Gaussian Processes". In: *Advances in Neural Information Processing Systems*. Vol. 13. MIT Press, 2000, pp. 654–660.

[71] R. Paul Wadwa et al. "Trial of hybrid closed-loop control in young children with type 1 diabetes". In: *New England Journal of Medicine* 388.11 (2023). For the PEDAP Trial Study Group, pp. 991–1001. DOI: 10.1056/NEJMoa2210834.

[72] Joe H. Ward Jr. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

[73] Tao Wu et al. “Parkinson’s Disease-Related Spatial Covariance Pattern Identified with Resting-State Functional MRI”. In: *Journal of Cerebral Blood Flow & Metabolism* 35.11 (2015), pp. 1764–1770. DOI: 10.1038/jcbfm.2015.118.

[74] George Wynne and Andrew B Duncan. “A Kernel Two-Sample Test for Functional Data”. In: *Journal of Machine Learning Research* 23.73 (2022), pp. 1–51.

[75] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. “BIRCH: An Efficient Data Clustering Method for Very Large Databases”. In: *ACM SIGMOD Record* 25.2 (1996), pp. 103–114. DOI: 10.1145/235968.233324.

Appendix organization. Section A provides background for Gaussian measures defined on Hilbert spaces and Maximum Mean Discrepancy, while Section B gives the closed-form MMD expressions and fitting algorithms. Sections C to E cover time-varying mixtures, projected likelihood estimation, and Banach-space extensions. Sections F to H present additional experiments, implementation details, and numerical diagnostics. Finally, Section I contains all proofs.

A Background

This section establishes the notation and foundational properties of Gaussian measures and kernel methods used throughout this work.

A.1 Gaussian measures on separable Hilbert spaces

Let $\mathcal{X}$ be a separable real Hilbert space equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{X})$. A Borel probability measure ν on $\mathcal{X}$ is said to be *Gaussian* if, for every continuous linear functional $f \in \mathcal{X}^*$, the pushforward measure $\nu \circ f^{-1}$ is a Gaussian measure on $\mathbb{R}$. By the Riesz representation theorem, we routinely identify $\mathcal{X}^*$ with $\mathcal{X}$.

A Gaussian measure ν is uniquely determined by its mean $m \in \mathcal{X}$ and its covariance operator $\mathcal{K} : \mathcal{X} \to \mathcal{X}$, which are defined via the relations

$$\langle m, h\rangle_{\mathcal{X}} = \int_{\mathcal{X}} \langle x, h\rangle_{\mathcal{X}} \, \mathrm{d}\nu(x), \quad h \in \mathcal{X},$$

and

$$\langle \mathcal{K}g, h\rangle_{\mathcal{X}} = \int_{\mathcal{X}} \langle x - m, g\rangle_{\mathcal{X}} \langle x - m, h\rangle_{\mathcal{X}} \, \mathrm{d}\nu(x), \quad g, h \in \mathcal{X}.$$

We denote such a measure by $\nu = \mathcal{N}(m, \mathcal{K})$. The covariance operator $\mathcal{K}$ is necessarily self-adjoint, positive semidefinite, and trace-class. The latter property implies that for any orthonormal basis $(e_i)_{i\geq 1}$ of $\mathcal{X}$, the trace is finite:

$$\operatorname{tr}(\mathcal{K}) = \sum_{i=1}^{\infty} \langle \mathcal{K}e_i, e_i\rangle_{\mathcal{X}} < \infty.$$

The case $\mathcal{K} = 0$ corresponds to the Dirac measure δ_m; this case is excluded from the model class in the main text. In particular, if $X \sim \mathcal{N}(m, \mathcal{K})$, its moments satisfy $\mathbb{E}\|X - m\|_{\mathcal{X}}^2 = \operatorname{tr}(\mathcal{K})$ and $\mathbb{E}\|X\|_{\mathcal{X}}^2 = \|m\|_{\mathcal{X}}^2 + \operatorname{tr}(\mathcal{K})$.

A.2 Kernel mean embeddings and Maximum Mean Discrepancy

Let $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a continuous positive-definite kernel with an associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}_\kappa$. Its canonical feature map $\Phi(x) = \kappa(x, \cdot)$ is continuous because

$$\|\Phi(x) - \Phi(y)\|_{\mathcal{H}_\kappa}^2 = \kappa(x, x) + \kappa(y, y) - 2\kappa(x, y).$$

Since $\mathcal{X}$ is separable, the RKHS is separable as well: the span of the features of a countable dense subset of $\mathcal{X}$ is dense in $\mathcal{H}_\kappa$. Thus Φ is strongly measurable. For probability measures $P, Q \in \mathcal{P}(\mathcal{X})$, we assume the following integrability condition holds:

$$\mathbb{E}_{x\sim P}\left[\sqrt{\kappa(x, x)}\right] < \infty, \quad \mathbb{E}_{y\sim Q}\left[\sqrt{\kappa(y, y)}\right] < \infty. \tag{16}$$

Under this assumption, the kernel mean embeddings

$$\mu_P \coloneqq \mathbb{E}_{x\sim P}[\kappa(x,\cdot)], \quad \mu_Q \coloneqq \mathbb{E}_{y\sim Q}[\kappa(y,\cdot)]$$

are well-defined as Bochner integrals in $\mathcal{H}_\kappa$, since $\|\Phi(x)\|_{\mathcal{H}_\kappa} = \sqrt{\kappa(x,x)}$. The squared Maximum Mean Discrepancy (MMD) between P and Q is defined as the squared distance between their embeddings:

$$\begin{aligned}\mathrm{MMD}^2_\kappa(P,Q) &\coloneqq \|\mu_P - \mu_Q\|^2_{\mathcal{H}_\kappa} \\ &= \mathbb{E}_{x,x'\sim P}[\kappa(x,x')] - 2\mathbb{E}_{x\sim P, y\sim Q}[\kappa(x,y)] + \mathbb{E}_{y,y'\sim Q}[\kappa(y,y')], \end{aligned} \tag{17}$$

where x, x' are i.i.d. samples from P and y, y' are i.i.d. samples from Q. Note that if κ is bounded, condition (16) is satisfied for all probability measures.

A kernel is called *characteristic* if the embedding map $P \mapsto \mu_P$ is injective. In this case, $\mathrm{MMD}_\kappa(P,Q) = 0$ if and only if $P = Q$, meaning the MMD induces a true metric on the space of probability measures.

A.3 Fourier interpretation of the Gaussian MMD

The Gaussian radial kernel, given by

$$\kappa(x,y) \coloneqq \exp\left(-\frac{\|x-y\|^2_{\mathcal{X}}}{2\sigma^2_\kappa}\right), \quad \sigma_\kappa > 0, \tag{18}$$

is both bounded and characteristic on separable Hilbert spaces [74].

To gain intuition into the geometry induced by this kernel, we briefly consider the finite-dimensional case $\mathcal{X} = \mathbb{R}^d$. For $P, Q \in \mathcal{P}(\mathbb{R}^d)$, let φ_P and φ_Q denote their respective characteristic functions, $\varphi_P(\omega) \coloneqq \int_{\mathbb{R}^d} e^{i\omega^\top x}\,\mathrm{d}P(x)$.

By Bochner's theorem, the shift-invariant kernel (18) admits a Fourier representation with a Gaussian spectral density ψ_{σ_κ}:

$$\kappa(x,y) = \frac{\sigma^d_\kappa}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\omega^\top(x-y)} e^{-\sigma^2_\kappa\|\omega\|^2/2}\,\mathrm{d}\omega.$$

By defining the signed measure $\mu \coloneqq P - Q$ and applying Fubini's theorem to the expanded MMD objective (17), we obtain:

$$\begin{aligned}\mathrm{MMD}^2_\kappa(P,Q) &= \iint_{\mathbb{R}^d\times\mathbb{R}^d} \kappa(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\ &= \frac{\sigma^d_\kappa}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|\int_{\mathbb{R}^d} e^{i\omega^\top x}\,\mathrm{d}\mu(x)\right|^2 e^{-\sigma^2_\kappa\|\omega\|^2/2}\,\mathrm{d}\omega \\ &= \frac{\sigma^d_\kappa}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \left|\varphi_P(\omega) - \varphi_Q(\omega)\right|^2 e^{-\sigma^2_\kappa\|\omega\|^2/2}\,\mathrm{d}\omega.\end{aligned}$$

This formulation reveals a frequency-domain perspective: the squared Gaussian MMD is a weighted L^2 distance between the characteristic functions of the two distributions. The Gaussian weight $e^{-\sigma^2_\kappa\|\omega\|^2/2}$ acts as a low-pass filter, effectively penalizing discrepancies in low-frequency modes (global structural differences) while naturally smoothing out high-frequency noise.

B Closed-form MMD expressions

To evaluate the cross-expectations $J_{i,k}$ and $I_{k,s}$ of (5) in the infinite-dimensional setting, we rely on the Fredholm determinant. Recall that for any trace-class operator $T \in \mathcal{L}(\mathcal{X})$, the determinant $\det_F(I_\mathcal{X} + T)$ is well-defined and serves as the natural generalization of the standard matrix determinant.

Proposition 4 (MMD for the Gaussian kernel)**.** *Let $A \in \mathcal{L}(\mathcal{X})$ be a self-adjoint, positive semidefinite operator, and consider the Gaussian kernel induced by A,*

$$\kappa_A(x, y) \coloneqq \exp\left(-\frac{1}{2}\langle x - y, A(x - y)\rangle_{\mathcal{X}}\right). \tag{19}$$

Then the expectations in (5) are given by

$$J_{i,k} = \frac{\exp\left(-\frac{1}{2}\langle X_i - m_k,\ A^{1/2}\left(I_{\mathcal{X}} + A^{1/2}\mathcal{K}_k A^{1/2}\right)^{-1} A^{1/2}(X_i - m_k)\rangle_{\mathcal{X}}\right)}{\det{}_F\left(I_{\mathcal{X}} + A^{1/2}\mathcal{K}_k A^{1/2}\right)^{1/2}}, \tag{20}$$

$$I_{k,s} = \frac{\exp\left(-\frac{1}{2}\langle m_k - m_s,\ A^{1/2}\left(I_{\mathcal{X}} + A^{1/2}(\mathcal{K}_k + \mathcal{K}_s)A^{1/2}\right)^{-1} A^{1/2}(m_k - m_s)\rangle_{\mathcal{X}}\right)}{\det{}_F\left(I_{\mathcal{X}} + A^{1/2}(\mathcal{K}_k + \mathcal{K}_s)A^{1/2}\right)^{1/2}}. \tag{21}$$

Proposition 5 (MMD for the polynomial kernel)**.** *Fix $p \in \mathbb{N}$, $c \geq 0$, and consider the kernel*

$$\kappa(x, y) \coloneqq (\langle x, y\rangle_{\mathcal{X}} + c)^p. \tag{22}$$

Then the expectations in (5) are given by

$$J_{i,k} = \sum_{t=0}^{\lfloor p/2 \rfloor} \frac{p!}{(p-2t)!\, 2^t\, t!} \langle X_i, \mathcal{K}_k X_i\rangle_{\mathcal{X}}^t \ (\langle X_i, m_k\rangle_{\mathcal{X}} + c)^{p-2t},$$

$$I_{k,s} = \sum_{r=0}^{p} \binom{p}{r} c^{p-r} \langle M_{r,k}, M_{r,s}\rangle_{\mathcal{X}^{\otimes r}}, \qquad \text{with } M_{r,k} \coloneqq \mathbb{E}_{y\sim\nu_k}[y^{\otimes r}] \in \mathcal{X}^{\otimes r}, \quad M_{0,k} = 1.$$

The formal proofs of Propositions 4 and 5 are given in Sections I.7 and I.8.

B.1 Numerical computations and algorithm description

Given an orthonormal basis $(e_r)_{r=1}^M$ of $\mathcal{X}_M$, we represent observations and means by their coordinate vectors

$$\mathbf{x}_i \coloneqq \left(\langle X_i, e_r\rangle_{\mathcal{X}}\right)_{r=1}^M \in \mathbb{R}^M, \qquad \mathbf{m}_k \coloneqq \left(\langle m_k, e_r\rangle_{\mathcal{X}}\right)_{r=1}^M \in \mathbb{R}^M, \tag{23}$$

and the covariance operator $\mathcal{K}_{k,M}$ by its matrix representation $\mathbf{K}_k \in \mathbb{R}^{M\times M}$ with entries

$$(\mathbf{K}_k)_{r\ell} \coloneqq \langle \mathcal{K}_{k,M} e_\ell, e_r\rangle_{\mathcal{X}} \qquad 1 \leq r, \ell \leq M. \tag{24}$$

In these coordinates, the pushforward measure $\nu_{k,M}$ becomes the multivariate normal $\mathcal{N}(\mathbf{m}_k, \mathbf{K}_k)$ on $\mathbb{R}^M$. Using these coordinate representations, we now provide explicit formulas for the finite-dimensional approximations.

Projected responsibilities. The pushforward $\nu_{k,M} = \nu_k \circ \Pi_M^{-1}$ is a multivariate normal distribution on $\mathbb{R}^M$,

$$\nu_{k,M} = \mathcal{N}(\mathbf{m}_k, \mathbf{K}_k).$$

If each $\mathbf{K}_k$ is positive definite, then each $\nu_{k,M}$ admits the Lebesgue density

$$p_k^{(M)}(\mathbf{x}) \coloneqq (2\pi)^{-M/2} \det(\mathbf{K}_k)^{-1/2} \exp\Big\{-\tfrac{1}{2}(\mathbf{x} - \mathbf{m}_k)^\top \mathbf{K}_k^{-1}(\mathbf{x} - \mathbf{m}_k)\Big\}.$$

In this case the projected responsibilities satisfy, for $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$,

$$\gamma_{i,k}^{(M)} \coloneqq \gamma_k^{(M)}(\mathbf{x}_i) = \frac{\pi_k p_k^{(M)}(\mathbf{x}_i)}{\sum_{s=1}^K \pi_s p_s^{(M)}(\mathbf{x}_i)}.$$

If some $\mathbf{K}_k$ are singular, the corresponding $\nu_{k,M}$ are supported on affine subspaces; the Radon–Nikodym definition of $\gamma_k^{(M)}$ still applies and one may compute densities on the supports or use a small ridge regularization.

1. Gaussian radial kernel (Proposition 4) Let $\mathbf{A} \in \mathbb{R}^{M\times M}$ denote the matrix representation of the compressed operator $A_M := \Pi_M A|_{\mathcal{X}_M} : \mathcal{X}_M \to \mathcal{X}_M$ in the basis $(e_r)_{r=1}^M$, with entries $\mathbf{A}_{r\ell} = \langle Ae_\ell, e_r\rangle_{\mathcal{X}}$. The closed forms of Proposition 4 become the finite-dimensional approximations

$$J_{i,k}^{(M)} := \det\big(I_M + \mathbf{A}^{1/2}\mathbf{K}_k\mathbf{A}^{1/2}\big)^{-1/2} \times \exp\Big\{-\tfrac{1}{2}(\mathbf{x}_i - \mathbf{m}_k)^\top \mathbf{A}^{1/2}\big(I_M + \mathbf{A}^{1/2}\mathbf{K}_k\mathbf{A}^{1/2}\big)^{-1}\mathbf{A}^{1/2}(\mathbf{x}_i - \mathbf{m}_k)\Big\},$$

$$I_{k,s}^{(M)} := \det\big(I_M + \mathbf{A}^{1/2}(\mathbf{K}_k + \mathbf{K}_s)\mathbf{A}^{1/2}\big)^{-1/2} \times \exp\Big\{-\tfrac{1}{2}(\mathbf{m}_k - \mathbf{m}_s)^\top \mathbf{A}^{1/2}\big(I_M + \mathbf{A}^{1/2}(\mathbf{K}_k + \mathbf{K}_s)\mathbf{A}^{1/2}\big)^{-1}\mathbf{A}^{1/2}(\mathbf{m}_k - \mathbf{m}_s)\Big\},$$

and we set

$$J_k^{(M)} := \frac{1}{n}\sum_{i=1}^n J_{i,k}^{(M)}, \qquad \mathbf{J}^{(M)} := (J_1^{(M)}, \dots, J_K^{(M)})^\top, \qquad \mathbf{I}^{(M)} := (I_{k,s}^{(M)})_{k,s=1}^K.$$

For the isotropic case $\mathbf{A} = \sigma_\kappa^{-2} I_M$, one writes $\alpha = -1/(2\sigma_\kappa^2)$ and recovers the equivalent forms $\det(I_M - 2\alpha\mathbf{K}_k)^{-1/2}$ and $(I_M - 2\alpha\mathbf{K}_k)^{-1}$ in the exponents.

For numerically stable log-domain evaluation, compute the symmetric eigendecompositions

$$\mathbf{A}^{1/2}\mathbf{K}_k\mathbf{A}^{1/2} = \mathbf{V}_k \mathrm{diag}(\lambda_{k,1}, \dots, \lambda_{k,M})\mathbf{V}_k^\top,$$

$$\mathbf{A}^{1/2}(\mathbf{K}_k + \mathbf{K}_s)\mathbf{A}^{1/2} = \mathbf{V}_{k,s}\mathrm{diag}(\lambda_{k,s,1}, \dots, \lambda_{k,s,M})\mathbf{V}_{k,s}^\top,$$

with orthogonal matrices $\mathbf{V}_k, \mathbf{V}_{k,s}$ and $\lambda_{\cdot,\cdot} \geq 0$. Note that $\det(I_M + \mathbf{A}^{1/2}\mathbf{K}_k\mathbf{A}^{1/2}) = \prod_r(1+\lambda_{k,r})$. Writing $\mathbf{z}_{i,k} := \mathbf{V}_k^\top \mathbf{A}^{1/2}(\mathbf{x}_i - \mathbf{m}_k)$ and $\mathbf{d}_{k,s} := \mathbf{V}_{k,s}^\top \mathbf{A}^{1/2}(\mathbf{m}_k - \mathbf{m}_s)$, we obtain the stable log-domain expressions

$$\log J_{i,k}^{(M)} = -\frac{1}{2}\sum_{r=1}^M \log(1+\lambda_{k,r}) \; - \; \frac{1}{2}\sum_{r=1}^M \frac{(\mathbf{z}_{i,k})_r^2}{1+\lambda_{k,r}},$$

$$\log I_{k,s}^{(M)} = -\frac{1}{2}\sum_{r=1}^M \log(1+\lambda_{k,s,r}) \; - \; \frac{1}{2}\sum_{r=1}^M \frac{(\mathbf{d}_{k,s})_r^2}{1+\lambda_{k,s,r}}.$$

When $\mathbf{A} = \sigma_\kappa^{-2} I_M$, the matrices $\mathbf{V}_k$ coincide with the eigenvectors of $\mathbf{K}_k$ and the expressions reduce to those with $\alpha = -1/(2\sigma_\kappa^2)$. To reduce cost, one may truncate these sums to $r \leq R$ (e.g. the leading R eigenvalues), i.e. replace the relevant matrices by their rank-R spectral approximations.

2. Polynomial kernel (Proposition 5) Fix an integer $p \geq 1$ and $c \geq 0$ and consider $\kappa(x,y) = (\langle x, y\rangle_{\mathcal{X}} + c)^p$. In the projected space, this becomes $\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + c)^p$ on $\mathbb{R}^M$.

Computing $J_{i,k}^{(M)}$. For $y \sim \nu_{k,M}$, define the scalars

$$\mu_{i,k}^{(M)} := \mathbf{x}_i^\top \mathbf{m}_k + c, \qquad v_{i,k}^{(M)} := \mathbf{x}_i^\top \mathbf{K}_k\, \mathbf{x}_i,$$

so that $\mathbf{x}_i^\top y + c$ is a one-dimensional Gaussian with mean $\mu_{i,k}^{(M)}$ and variance $v_{i,k}^{(M)}$. Then Proposition 5 yields the efficient closed form

$$J_{i,k}^{(M)} = \sum_{t=0}^{\lfloor p/2\rfloor} \frac{p!}{(p-2t)!\, 2^t\, t!} \big(v_{i,k}^{(M)}\big)^t \big(\mu_{i,k}^{(M)}\big)^{p-2t}.$$

Computing $I_{k,s}^{(M)}$. Let $y \sim \nu_{k,M}$ and $y' \sim \nu_{s,M}$ be independent in $\mathbb{R}^M$, and set $Z_{k,s} := y^\top y'$. Then

$$I_{k,s}^{(M)} = \sum_{r=0}^p \binom{p}{r} c^{p-r}\, \mathbb{E}\big[Z_{k,s}^r\big].$$

Equivalently, by Proposition 5,

$$I_{k,s}^{(M)} = \sum_{r=0}^{p} \binom{p}{r} c^{p-r} \left\langle \mathbf{M}_{r,k}^{(M)}, \mathbf{M}_{r,s}^{(M)} \right\rangle_{(\mathbb{R}^M)^{\otimes r}},$$

where

$$\mathbf{M}_{r,k}^{(M)} := \mathbb{E}_{y \sim \nu_{k,M}}[y^{\otimes r}] \in (\mathbb{R}^M)^{\otimes r}.$$

For small degrees, these moments admit simple closed forms in terms of $(\mathbf{m}_k, \mathbf{K}_k)$. In particular,

$$\begin{aligned}
\mathbb{E}[Z_{k,s}] &= \mathbf{m}_k^\top \mathbf{m}_s, \\
\mathbb{E}[Z_{k,s}^2] &= (\mathbf{m}_k^\top \mathbf{m}_s)^2 + \mathbf{m}_k^\top \mathbf{K}_s \, \mathbf{m}_k + \mathbf{m}_s^\top \mathbf{K}_k \, \mathbf{m}_s + \operatorname{tr}(\mathbf{K}_k \mathbf{K}_s), \\
\mathbb{E}[Z_{k,s}^3] &= (\mathbf{m}_k^\top \mathbf{m}_s)^3 + 3(\mathbf{m}_k^\top \mathbf{m}_s)\Big(\mathbf{m}_k^\top \mathbf{K}_s \, \mathbf{m}_k + \mathbf{m}_s^\top \mathbf{K}_k \, \mathbf{m}_s + \operatorname{tr}(\mathbf{K}_k \mathbf{K}_s)\Big) + 6 \, \mathbf{m}_k^\top \mathbf{K}_s \mathbf{K}_k \, \mathbf{m}_s.
\end{aligned}$$

For arbitrary degree p, the implementation computes these moments exactly using cumulants of a Gaussian quadratic form. Set

$$\mathbf{a} = \begin{pmatrix} \mathbf{m}_k \\ \mathbf{m}_s \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} 0 & \mathbf{K}_k \\ \mathbf{K}_s & 0 \end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix} 0 & I_M \\ I_M & 0 \end{pmatrix}.$$

Expanding the log moment-generating function

$$\log \mathbb{E}[e^{tZ_{k,s}}] = -\tfrac{1}{2} \log\det(I_{2M} - t\mathbf{B}) + \tfrac{t}{2} \mathbf{a}^\top \mathbf{F}(I_{2M} - t\mathbf{B})^{-1}\mathbf{a}$$

for t near zero gives the cumulants and raw-moment recurrence

$$\begin{aligned}
\chi_r &= \frac{(r-1)!}{2} \operatorname{tr}(\mathbf{B}^r) + \frac{r!}{2} \mathbf{a}^\top \mathbf{F} \mathbf{B}^{r-1} \mathbf{a}, \qquad r \geq 1, \\
u_0 &= 1, \qquad u_r = \sum_{j=1}^{r} \binom{r-1}{j-1} \chi_j u_{r-j} = \mathbb{E}[Z_{k,s}^r].
\end{aligned}$$

These identities also hold for singular covariances. The implementation uses M-dimensional blocks: odd trace terms vanish and $\operatorname{tr}(\mathbf{B}^{2j}) = 2\operatorname{tr}((\mathbf{K}_k \mathbf{K}_s)^j)$. Computing all cumulants and moments through degree p costs $O(pM^3 + p^2)$ per component pair.

B.2 Algorithm descriptions

We present pseudocode for the fitting procedure described in Section 2.1. The algorithm operates in the projected coordinate space $\mathbb{R}^M$ using the finite-dimensional quantities $(\mathbf{J}^{(M)}, \mathbf{I}^{(M)})$ from Section B.1, which we denote $(\mathbf{J}, \mathbf{I})$ for brevity.

Algorithm 1 alternates between the analytical weight update (6) and a gradient step on the component parameters $(\mathbf{m}_{k,M}, \mathbf{L}_k)$, with covariances parameterized through their Cholesky factors to enforce positive definiteness. The exact gradients are obtained from the differentiable closed forms of Propositions 4 and 5. An alternative is a fully gradient-based variant in which π is updated jointly with the components, either by projecting onto the simplex or by using logits $\ell \in \mathbb{R}^K$ with $\pi_k = \operatorname{softmax}(\ell)_k$.

C Time-varying mixtures

This appendix details the algorithmic implementation and theoretical results deferred from Section 4.2. We outline the projected $\mathbb{R}^M$ fitting procedure, and subsequently establish three guarantees: the consistency of the integrated empirical objective, the consistency of its projected $\mathbb{R}^M$ approximation, and the density of time-varying Gaussian mixtures with shared components.

Algorithm 2 implements the optimization of the time-varying objective (15) in $\mathbb{R}^M$. The component parameters $\{(\mathbf{m}_k, \mathbf{K}_k)\}_{k=1}^K$ are shared across all time slices, while the temporal variation is entirely captured by the weights $\pi(t)$, whose logits are governed by the neural ODE introduced in Section 4.2. Implementation details are summarized in Table 3.

Algorithm 1 Alternating optimization for projected MMD Gaussian mixtures

Require: Projected data $\{\mathbf{x}_{i,M}\}_{i=1}^{n} \subset \mathbb{R}^M$; components K; learning rate η; ridge $\varepsilon > 0$; iterations T

Ensure: Fitted weights $\hat{\pi}$ and components $\{(\hat{\mathbf{m}}_{k,M}, \hat{\mathbf{K}}_{k,M})\}_{k=1}^{K}$

$\{\mathbf{m}_{k,M}\}_{k=1}^{K} \leftarrow \textsc{KMeans}(\{\mathbf{x}_{i,M}\}_{i=1}^{n}, K)$ ▷ Initialize means from data

Assign $c_i \leftarrow \arg\min_k \|\mathbf{x}_{i,M} - \mathbf{m}_{k,M}\|$ and let $S_k = \{i : c_i = k\}$ ▷ Nearest-center partition

$\mathbf{K}_{k,M} \leftarrow \frac{1}{|S_k|-1} \sum_{i \in S_k} (\mathbf{x}_{i,M} - \mathbf{m}_{k,M})(\mathbf{x}_{i,M} - \mathbf{m}_{k,M})^\top + \varepsilon I_M$ ▷ Within-cluster covariance

$\mathbf{L}_k \leftarrow \mathrm{chol}(\mathbf{K}_{k,M})$ ▷ Cholesky parametrization ensures $\mathbf{K}_{k,M} \succ 0$

$\pi \leftarrow \mathbf{1}/K$ ▷ Uniform weight initialization

for $t = 1, \ldots, T$ **do**

  Compute $\mathbf{I}^{(M)} \in \mathbb{R}^{K \times K}$, $\mathbf{J}^{(M)} \in \mathbb{R}^K$ using Section B.1 ▷ Cross-expectations

  $\pi \leftarrow \arg\min_{\pi \in \Delta^{K-1}} \pi^\top \mathbf{I}^{(M)} \pi - 2(\mathbf{J}^{(M)})^\top \pi$ ▷ Exact QP update (6)

  $\mathcal{L}_M \leftarrow \pi^\top \mathbf{I}^{(M)} \pi - 2(\mathbf{J}^{(M)})^\top \pi$ ▷ Objective at fixed π

  $\mathbf{m}_{k,M} \leftarrow \mathbf{m}_{k,M} - \eta \nabla_{\mathbf{m}_{k,M}} \mathcal{L}_M$ ▷ Gradient step on means

  $\mathbf{L}_k \leftarrow \mathbf{L}_k - \eta \nabla_{\mathbf{L}_k} \mathcal{L}_M$ ▷ Gradient step on Cholesky factors

  $\mathbf{K}_{k,M} \leftarrow \mathbf{L}_k \mathbf{L}_k^\top + \varepsilon I_M$ ▷ Reconstruct PSD covariances

end for

return $\hat{\pi} \leftarrow \pi$ and $\{(\hat{\mathbf{m}}_{k,M}, \hat{\mathbf{K}}_{k,M})\} \leftarrow \{(\mathbf{m}_{k,M}, \mathbf{K}_{k,M})\}$

Algorithm 2 Time-varying MMD-based mixture fitting

Require: Projected temporal data $\{\mathbf{x}_{i,M}^{(l)}\}_{i=1}^{n_l} \subset \mathbb{R}^M$ across time slices $t_1, \ldots, t_L$; number of components K; learning rate η; epochs E; neural ODE parameters ϑ.

Ensure: Fitted parameters $\hat{\vartheta}$ and shared components $\{(\hat{\mathbf{m}}_k, \hat{\mathbf{K}}_k)\}_{k=1}^{K}$.

Initialize $\{(\mathbf{m}_k, \mathbf{K}_k)\}_{k=1}^{K}$ using the pooled data $\bigcup_{l=1}^{L} \{\mathbf{x}_{i,M}^{(l)}\}_{i=1}^{n_l}$.

Initialize ϑ.

for *epoch* $= 1, \ldots, E$ **do**

  Compute $\pi(t_l) \leftarrow \mathrm{softmax}(\mathbf{z}(t_l))$ for $l \in \{1, \ldots, L\}$ via the ODE.

  Compute $\mathbf{I}^{(M)} \in \mathbb{R}^{K \times K}$ via (7).

  for $l = 1, \ldots, L$ **do**

    Compute $\mathbf{J}_l^{(M)} \in \mathbb{R}^K$ where $[\mathbf{J}_l^{(M)}]_k \leftarrow \frac{1}{n_l} \sum_{i=1}^{n_l} J_{i,k}^{(M)}(t_l)$ via (7).

    $\mathcal{L}_l^{(M)} \leftarrow \pi(t_l)^\top \mathbf{I}^{(M)} \pi(t_l) - 2(\mathbf{J}_l^{(M)})^\top \pi(t_l)$. ▷ MMD^2 up to a constant

  end for

  $\mathcal{L} \leftarrow \frac{1}{L} \sum_{l=1}^{L} \mathcal{L}_l^{(M)}$. ▷ Approximate integrated loss up to a constant

  Update $(\vartheta, \{\mathbf{m}_k, \mathbf{K}_k\}) \leftarrow (\vartheta, \{\mathbf{m}_k, \mathbf{K}_k\}) - \eta \nabla \mathcal{L}$.

end for

return $\hat{\vartheta} \leftarrow \vartheta$, $\{(\hat{\mathbf{m}}_k, \hat{\mathbf{K}}_k)\}_{k=1}^{K} \leftarrow \{(\mathbf{m}_k, \mathbf{K}_k)\}_{k=1}^{K}$.

To evaluate the MMD efficiently within the algorithm, we rely on the finite-dimensional projections defined in (7). We separate the evaluation into terms that can be computed globally versus those computed per time slice. The matrix $\mathbf{I}^{(M)}$ depends only on the shared components and is computed once per epoch. In contrast, the vectors $\mathbf{J}_l^{(M)}$ evaluate the interaction between the shared components and the data specific to time slice t_l, requiring per-slice computation.

Motivated by the within-cluster RKHS dispersion criterion underlying kernel k-groups [31], we examine the MMD loss over candidate component counts as an elbow diagnostic [66, 67], supplemented by interpretability when the elbow is ambiguous.

Model selection. The kernel is fixed before selecting K and we use the Gaussian radial kernel by default, as it is characteristic, while the polynomial kernel is appropriate when the focus is on matching low-order moments. Comparisons across kernels require an external evaluation criterion, such as ARI when ground-truth labels are available.

Figure 6 shows the CGM loss values for $K = 1, \ldots, 6$. For each representation, we used the Gaussian radial kernel and initialized the model with k-means++. $K = 2$ for intraday curves was

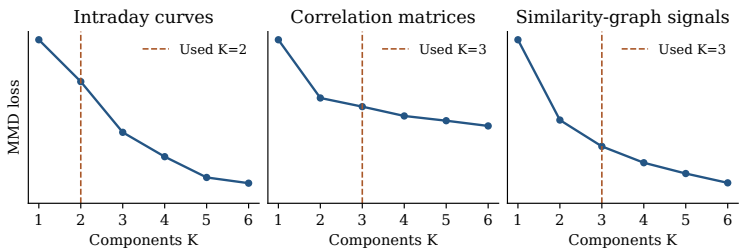


Figure 6: CGM model-selection: normalized MMD loss versus K for the three geometric representations. Dashed lines mark the component counts used in the case study.

guided by interpretability and treatment–placebo contrast, whereas $K = 3$ was used for correlation matrices and similarity-graph signals. Because group contrast informed the intraday choice, that contrast is an exploratory description rather than an independent validation of model selection.

Consistency guarantees

In parallel with the static framework of Section 3, we establish the theoretical foundations for the temporal mixture model. Unlike the static case, where sample consistency is standard, minimizing the integrated temporal objective requires controlling the time-varying empirical processes. We first establish this sample consistency, followed by the validity of the projected approximation and the expressiveness of the temporal model class.

Proposition 6 (Sample consistency of the integrated MMD)**.** *Let κ be a continuous positive-definite kernel on the separable Hilbert space $\mathcal{X}$. Let $X_1, \dots, X_n$ be i.i.d. copies of a jointly measurable $\mathcal{X}$-valued stochastic process $X\colon [0,T] \to \mathcal{X}$ with marginal laws $P(t)$, and let $P_n(t)$ be the corresponding empirical measure. Let $Q(t)$ be a family of probability measures on $\mathcal{X}$ such that its mean embedding trajectory $t \mapsto \mu_{Q(t)}$ is strongly measurable and belongs to $L^2(0,T;\mathcal{H}_\kappa)$. If $\int_0^T \mathbb{E}_{X\sim P(t)}[\kappa(X,X)]\,\mathrm{d}t < \infty$, we have*

$$\frac{1}{T}\int_0^T \mathrm{MMD}^2_\kappa\big(P_n(t), Q(t)\big)\,\mathrm{d}t \xrightarrow{n\to\infty} \frac{1}{T}\int_0^T \mathrm{MMD}^2_\kappa\big(P(t), Q(t)\big)\,\mathrm{d}t \qquad \textit{a.s.}$$

In particular, $\frac{1}{T}\int_0^T \mathrm{MMD}^2_\kappa\big(P_n(t), P(t)\big)\,\mathrm{d}t \to 0$ almost surely.

Proposition 7 (Projection consistency for temporal mixtures)**.** *Let $Q_\theta(t) = \sum_{k=1}^K \pi_k(t)\nu_k$ be a time-dependent Gaussian mixture where $\pi(t) \in \Delta^{K-1}$ and $t \mapsto \pi_k(t)$ is measurable for each k. Let $P_{n,M}(t)$ and $Q_{\theta,M}(t)$ be the projected measures on $\mathbb{R}^M$ defined in Section 2.2, and let κ_M denote the coordinate version of the projected kernel. Suppose κ is continuous and satisfies the polynomial growth condition (9) for some $q \geq 0$. If the sample paths satisfy $\int_0^T \|X_i(t)\|^{2q}_{\mathcal{X}}\,\mathrm{d}t < \infty$ almost surely for $i = 1, \dots, n$, then*

$$\frac{1}{T}\int_0^T \mathrm{MMD}^2_{\kappa_M}\big(P_{n,M}(t), Q_{\theta,M}(t)\big)\,\mathrm{d}t \xrightarrow{M\to\infty} \frac{1}{T}\int_0^T \mathrm{MMD}^2_\kappa\big(P_n(t), Q_\theta(t)\big)\,\mathrm{d}t \qquad \textit{a.s.}$$

Furthermore, for each fixed t, if $X(t) \sim Q_\theta(t)$, the projected responsibilities $\gamma_k^{(M)}(\Pi_M X(t), t)$ satisfy the martingale convergence property of Proposition 2 with weights $\pi(t)$.

Proposition 8 (Density of time-varying Gaussian mixtures)**.** *Assume $\mathcal{X} \neq \{0\}$, fix $p \geq 1$, and let $P\colon [0,T] \to \mathcal{P}_p(\mathcal{X})$ be continuous with respect to the Wasserstein distance W_p. Then, for every $\varepsilon > 0$, there exist $K \in \mathbb{N}$, shared Gaussian measures $\nu_1, \dots, \nu_K$ on $\mathcal{X}$ with nonzero covariance operator, and continuous functions $\pi_1, \dots, \pi_K\colon [0,T] \to [0,1]$ with $\sum_{k=1}^K \pi_k(t) = 1$, such that*

$$\sup_{t\in[0,T]} W_p\left(P(t), \sum_{k=1}^K \pi_k(t)\nu_k\right) < \varepsilon.$$

The proofs of Propositions 6, 7 and 8 are provided in Sections I.4 to I.6, respectively.

D Projected maximum likelihood estimation

Although this paper focuses on MMD-based fitting, we briefly describe the projected maximum-likelihood baseline used in Section 4; it is the only density-based competitor that is well-defined in the Hilbert-space setting of (1). As discussed in the introduction, likelihood-based fitting in infinite dimensions requires a common dominating measure. Here we use the Karhunen–Loève projector associated with the Gaussian prior

$$\rho := \mu_0 = \mathcal{N}(0, C)$$

on $\mathcal{X}$. Let $(\lambda_j, \phi_j)_{j\geq 1}$ be the eigenpairs of C ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots$, and fix M with $\lambda_M > 0$. Define

$$\mathcal{X}_M := \operatorname{span}\{\phi_1, \ldots, \phi_M\}, \qquad \Pi_M x := \sum_{j=1}^{M} \langle x, \phi_j\rangle_{\mathcal{X}}\, \phi_j, \qquad \Pi_M^{\perp} := I_{\mathcal{X}} - \Pi_M.$$

Under the decomposition $\mathcal{X} = \mathcal{X}_M \oplus \mathcal{X}_M^{\perp}$, the prior factorizes as

$$\rho = \rho_M \otimes \rho_M^{\perp}, \qquad \rho_M = \mathcal{N}(0, \Lambda_M), \qquad \Lambda_M := \operatorname{diag}(\lambda_1, \ldots, \lambda_M),$$

where $\rho_M^{\perp}$ is the law of $\Pi_M^{\perp} X$ for $X \sim \rho$. Following [44], we construct each component by altering only the first M coordinates. Given projected parameters $\mathbf{m}_k \in \mathbb{R}^M$ and $\mathbf{K}_k \in \mathbb{R}^{M\times M}$ with $\mathbf{K}_k \succ 0$, let

$$\nu_{k,M} := \mathcal{N}(\mathbf{m}_k, \mathbf{K}_k) \qquad \text{on } \mathbb{R}^M,$$

and lift it to

$$\nu_k := \nu_{k,M} \otimes \rho_M^{\perp} \qquad \text{on } \mathcal{X}.$$

Equivalently, $\nu_k = \mathcal{N}(m_k, \mathcal{K}_k)$, where

$$m_k = \sum_{j=1}^{M} (\mathbf{m}_k)_j\, \phi_j,$$

and $\mathcal{K}_k$ agrees with C on $\mathcal{X}_M^{\perp}$ while its restriction to $\mathcal{X}_M$ has matrix $\mathbf{K}_k$ in the basis $\{\phi_1, \ldots, \phi_M\}$. Since $\nu_{k,M}$ and ρ_M are nondegenerate Gaussians on $\mathbb{R}^M$, they are equivalent, and therefore $\nu_k \sim \rho$. Hence every mixture

$$Q := \sum_{k=1}^{K} \pi_k \nu_k, \qquad \pi_k > 0, \qquad \sum_{k=1}^{K} \pi_k = 1,$$

satisfies $Q \ll \rho$; indeed $Q \sim \rho$ because its density is strictly positive ρ-a.s.

For $x \in \mathcal{X}$, write

$$\mathbf{x} := \big(\langle x, \phi_1\rangle_{\mathcal{X}}, \ldots, \langle x, \phi_M\rangle_{\mathcal{X}}\big)^{\top}$$

and let

$$\varphi_M(\mathbf{x}; \mathbf{m}, \mathbf{K}) := (2\pi)^{-M/2} \det(\mathbf{K})^{-1/2} \exp\Big\{ -\tfrac{1}{2}(\mathbf{x}-\mathbf{m})^{\top}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\Big\}$$

denote the usual Gaussian density on $\mathbb{R}^M$. Then

$$p_k(x) := \frac{\mathrm{d}\nu_k}{\mathrm{d}\rho}(x) = \frac{\mathrm{d}\nu_{k,M}}{\mathrm{d}\rho_M}(\mathbf{x}) = \frac{\varphi_M(\mathbf{x}; \mathbf{m}_k, \mathbf{K}_k)}{\varphi_M(\mathbf{x}; 0, \Lambda_M)}$$

and therefore

$$p_k(x) = \det(\Lambda_M)^{1/2} \det(\mathbf{K}_k)^{-1/2} \exp\Big\{ -\tfrac{1}{2}(\mathbf{x}-\mathbf{m}_k)^{\top}\mathbf{K}_k^{-1}(\mathbf{x}-\mathbf{m}_k) + \tfrac{1}{2}\mathbf{x}^{\top}\Lambda_M^{-1}\mathbf{x}\Big\}.$$

Accordingly,

$$q(x) := \frac{\mathrm{d}Q}{\mathrm{d}\rho}(x) = \sum_{k=1}^{K} \pi_k\, p_k(x).$$

Since the factor $\varphi_M(\mathbf{x}; 0, \Lambda_M)^{-1}$ is common to every component, maximizing the ρ-log-likelihood is equivalent to maximizing the usual projected mixture log-likelihood

$$\ell_M(\pi, m, \mathcal{K}) := \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \pi_k\, \varphi_M(\mathbf{x}_i; \mathbf{m}_k, \mathbf{K}_k)\Big), \tag{25}$$

Algorithm 3 Projected EM for the Gaussian-mixture log-likelihood

Require: Projected data $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^M$; number of components K; ridge parameter $\varepsilon \geq 0$; EM iterations T_{EM}
Ensure: Fitted parameters $\hat{\pi}$, $\{(\hat{\mathbf{m}}_k, \hat{\mathbf{K}}_k)\}_{k=1}^K$
$\{\mathbf{m}_k\}_{k=1}^K \leftarrow \mathrm{KMEANS}(\{\mathbf{x}_i\}_{i=1}^n, K)$ ▷ Initialize means
$\mathbf{K}_k \leftarrow \mathbf{I}_M, \quad \pi_k \leftarrow 1/K$ ▷ Initialize covariances and weights
for $t = 1, \ldots, T_{\mathrm{EM}}$ **do**
  for $i = 1, \ldots, n$, $k = 1, \ldots, K$ **do**
    $r_{ik} \leftarrow \dfrac{\pi_k \, \varphi_M(\mathbf{x}_i; \mathbf{m}_k, \mathbf{K}_k)}{\sum_{s=1}^K \pi_s \, \varphi_M(\mathbf{x}_i; \mathbf{m}_s, \mathbf{K}_s)}$ ▷ E-step responsibilities
  end for
  for $k = 1, \ldots, K$ **do**
    $N_k \leftarrow \sum_{i=1}^n r_{ik}$ ▷ Effective cluster size
    $\pi_k \leftarrow N_k / n$ ▷ M-step: weights
    $\mathbf{m}_k \leftarrow \frac{1}{N_k} \sum_{i=1}^n r_{ik} \mathbf{x}_i$ ▷ M-step: means
    $\mathbf{K}_k \leftarrow \frac{1}{N_k} \sum_{i=1}^n r_{ik} (\mathbf{x}_i - \mathbf{m}_k)(\mathbf{x}_i - \mathbf{m}_k)^\top + \varepsilon \mathbf{I}_M$ ▷ M-step: ridge-regularized covariances
  end for
end for
return $\hat{\pi} \leftarrow \pi, \quad \{(\hat{\mathbf{m}}_k, \hat{\mathbf{K}}_k)\}_{k=1}^K \leftarrow \{(\mathbf{m}_k, \mathbf{K}_k)\}_{k=1}^K$

because

$$\sum_{i=1}^n \log q(X_i) = \ell_M(\pi, m, \mathcal{K}) - \sum_{i=1}^n \log \varphi_M(\mathbf{x}_i; 0, \Lambda_M),$$

and the second term is independent of the mixture parameters.

The EM algorithm [24] is therefore carried out in the projected coordinates $\mathbf{x}_i \in \mathbb{R}^M$. In the E-step,

$$r_{ik} = \frac{\pi_k \, p_k(X_i)}{\sum_{\ell=1}^K \pi_\ell \, p_\ell(X_i)} = \frac{\pi_k \, \varphi_M(\mathbf{x}_i; \mathbf{m}_k, \mathbf{K}_k)}{\sum_{\ell=1}^K \pi_\ell \, \varphi_M(\mathbf{x}_i; \mathbf{m}_\ell, \mathbf{K}_\ell)}.$$

For the unregularized full-covariance model, the M-step updates are

$$\pi_k^{\mathrm{new}} = \frac{1}{n} \sum_{i=1}^n r_{ik}, \qquad \mathbf{m}_k^{\mathrm{new}} = \frac{\sum_{i=1}^n r_{ik} \mathbf{x}_i}{\sum_{i=1}^n r_{ik}},$$

and

$$\mathbf{K}_k^{\mathrm{new}} = \frac{\sum_{i=1}^n r_{ik} (\mathbf{x}_i - \mathbf{m}_k^{\mathrm{new}})(\mathbf{x}_i - \mathbf{m}_k^{\mathrm{new}})^\top}{\sum_{i=1}^n r_{ik}}.$$

In practice, one stabilizes the covariance update by replacing $\mathbf{K}_k^{\mathrm{new}}$ with $\mathbf{K}_k^{\mathrm{new}} + \varepsilon I_M$ for some $\varepsilon > 0$. This is a regularized EM step rather than the exact M-step of the unpenalized likelihood, but it guarantees $\mathbf{K}_k \succ 0$ and therefore preserves equivalence with the reference measure after lifting. In the fixed-covariance benchmark variant used in our experiments (see Section 4), the same E-step is used, while only the weights and means are updated and the covariance matrix is held fixed. After each M-step, the updated parameters are lifted back to $\mathcal{X}$ by setting

$$m_k^{\mathrm{new}} = \sum_{j=1}^M (\mathbf{m}_k^{\mathrm{new}})_j \, \phi_j$$

and defining $\mathcal{K}_k^{\mathrm{new}}$ to agree with C on $\mathcal{X}_M^\perp$ and to have matrix $\mathbf{K}_k^{\mathrm{new}}$ on $\mathcal{X}_M$. Algorithm 3 summarizes this projected EM routine, with optional ridge regularization controlled by ε.

The next result shows that the projected likelihood ratios are consistent as $M \to \infty$.

Proposition 9 (Consistency of projected likelihood ratios)**.** *Let Q be a mixture (1) with $Q \ll \rho$, and let $q := dQ/d\rho$. For each M, set $\rho_M := \rho \circ \Pi_M^{-1}$, $Q_M := Q \circ \Pi_M^{-1}$, and define $q_M := dQ_M/d\rho_M$. Let X be an $\mathcal{X}$-valued random variable with distribution ρ and denote*

$$X^{(M)} := \Pi_M X.$$

Then $q_M(X^{(M)}) = \mathbb{E}_\rho[q(X) \mid X^{(M)}]$ *a.s. and* $q_M(X^{(M)}) \xrightarrow{M\to\infty} q(X)$ *a.s. and in* $L^1(\rho)$. *Consequently, for i.i.d.* $X_1, \dots, X_n \sim Q$,

$$\sum_{i=1}^{n} \log q_M(\Pi_M X_i) \longrightarrow \sum_{i=1}^{n} \log q(X_i) \qquad Q^n\text{-a.s.}$$

The proof is given in Section I.9.

E Extension to Banach spaces

This appendix outlines how our framework extends to real separable Banach spaces $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$. Let $\mathcal{X}^*$ be the continuous dual and $\langle x, f\rangle$ the dual pairing. A Borel probability measure ν on $\mathcal{X}$ is Gaussian if $\nu \circ f^{-1}$ is Gaussian for all $f \in \mathcal{X}^*$. If ν is centered, its characteristic functional is

$$\widehat{\nu}(f) := \int_{\mathcal{X}} e^{i\langle x, f\rangle} \, \mathrm{d}\nu(x) = \exp\Big(-\tfrac{1}{2} C_\nu(f, f)\Big), \qquad f \in \mathcal{X}^*,$$

where the covariance form $C_\nu : \mathcal{X}^* \times \mathcal{X}^* \to \mathbb{R}$ is given by

$$C_\nu(f, g) := \int_{\mathcal{X}} \langle x, f\rangle \, \langle x, g\rangle \, \mathrm{d}\nu(x).$$

By Fernique's theorem, ν has a well-defined covariance operator $R_\nu : \mathcal{X}^* \to \mathcal{X}$ given by $R_\nu f = \int_{\mathcal{X}} \langle x, f\rangle x \, \mathrm{d}\nu(x)$. The associated Cameron–Martin space $H(\nu)$ is the Hilbert completion of $R_\nu(\mathcal{X}^*)$ under the inner product $\langle R_\nu f, R_\nu g\rangle_{H(\nu)} = \langle R_\nu f, g\rangle$.

Karhunen–Loève decomposition. Assume that $H(\nu)$ is infinite-dimensional. Following Bay and Croix [5], one constructs vectors $x_n \in \mathcal{X}$, dual functionals $x_n^* \in \mathcal{X}^*$, and normalized coordinates

$$h_n := \sqrt{\lambda_n}\, x_n \in H(\nu), \qquad h_n^* := \lambda_n^{-1/2} x_n^* \in \mathcal{X}^*,$$

from successive maximizers of the Rayleigh quotient associated with the covariance operators of the residual Gaussian measures; see [5, Theorem 3.3]. The resulting coordinates satisfy $R_\nu h_n^* = h_n$, and the random variables $\langle \cdot, h_n^*\rangle$ are independent $\mathcal{N}(0, 1)$ under ν.

Corollary 10 (Banach-space Karhunen–Loève decomposition [5, Corollary 3.8])**.** *Let* ν *be a centered Gaussian measure on a real separable Banach space* $\mathcal{X}$, *assume* $H(\nu)$ *is infinite-dimensional, and let* $(\lambda_n, x_n, h_n, h_n^*)_{n\ge 0}$ *be constructed as above. Then*

$$x = \sum_{n=0}^{\infty} \langle x, h_n^*\rangle \, h_n$$

converges in $\mathcal{X}$ *for* ν*-a.e.* $x \in \mathcal{X}$, *and the coordinate maps* $x \mapsto \langle x, h_n^*\rangle$ *are independent* $\mathcal{N}(0, 1)$ *under* ν. *Equivalently, if* $(\xi_n)_{n\ge 0}$ *are i.i.d.* $\mathcal{N}(0, 1)$, *then*

$$\sum_{n=0}^{\infty} \sqrt{\lambda_n}\, \xi_n \, x_n$$

converges almost surely in $\mathcal{X}$ *and has law* ν.

In the Hilbert-space setting of the main text, we identify $\mathcal{X} \simeq \mathcal{X}^*$ via the Riesz representation theorem. Then R_ν becomes the usual covariance operator, and the above construction reduces to the classical spectral Karhunen–Loève expansion.

Applicability. The measure-theoretic components of our framework (including Gaussian mixtures, Radon–Nikodym responsibilities, and MMD weak density arguments) extend naturally to separable Banach spaces. While orthogonal projections are unavailable in this broader setting, one can substitute finite-rank continuous projections $P_M : \mathcal{X} \to \mathcal{X}$, such as the Bay–Croix projections from Corolary 10. However, we do not pursue this extension further, as the closed-form MMD expressions and projection algorithms developed in the main text rely intrinsically on the inner product structure of Hilbert spaces.

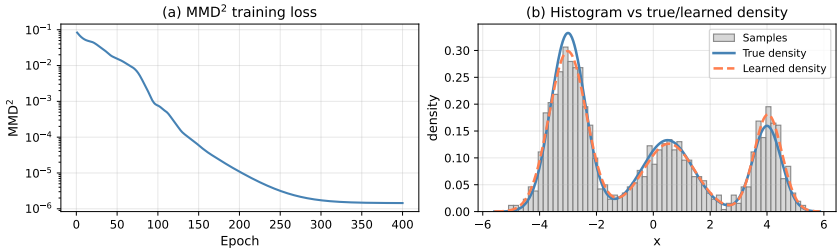


Figure 7: $\mathbb{R}^d$ Gaussian mixture recovery. **Left:** MMD^2 training loss. **Right:** Empirical histogram of the samples overlaid with the true (blue) and learned (dashed coral) densities.

F Synthetic experiments

We compare against four baseline families: (i) metric-space methods [58, 49, 27, 14, 34]; (ii) FDA-specific methods [55]; (iii) classical Euclidean methods restricted to $\mathbb{R}^d$; and (iv) a projected maximum-likelihood baseline inspired by [44] (see Section D). We report Adjusted Rand Index (ARI), train with Adam ($\mathrm{lr} = 0.1$) and k-means initialization, and run the experiments on CPU.

Model class. Across all domains, we fit the finite-dimensional mixture of Section 2.2 after fixing an orthonormal basis $(e_r)_{r=1}^M$ of $\mathcal{X}_M$. We use cosine bases for L^2/H^1, Wigner-D matrix coefficients for $\mathrm{SO}(3)$, Laplacian eigenvectors for graph signals, and the canonical basis for $\mathbb{R}^d$.

We evaluate the MMD-based Gaussian mixture fitting method on several Hilbert-space representations of increasing complexity. The first five experiments are controlled mixture-recovery tasks with known ground-truth parameters. The sixth experiment tests system identification for a Gaussian process induced by a linear SDE. The last two experiments apply the same fitting procedure to real molecular and skeleton-sequence data. In all cases, observations are represented in a finite-dimensional basis as in Section 2.2, and the model is fitted by minimizing the squared MMD objective with a Gaussian radial kernel. Final MMD values are reported as within-experiment diagnostics and are not meant to be compared directly across domains, since the coefficient dimension, scaling, and kernel bandwidth differ across settings.

F.1 Gaussian random variables in $\mathbb{R}^d$

As a sanity check, we first consider the finite-dimensional case $\mathcal{X} = \mathbb{R}^d$ with the standard inner product, where the model reduces to a classical Gaussian mixture model [8].

Experiment. We generate $n = 1500$ samples from a $K = 3$ Gaussian mixture on $\mathbb{R}$ with weights $\pi = (0.5, 0.3, 0.2)$, means $(-3.0,\ 0.5,\ 4.0)$, and standard deviations $(0.6,\ 0.9,\ 0.5)$. We then fit a $K = 3$ mixture by minimizing MMD^2. Figure 7 shows that the learned density closely matches the empirical histogram and the ground-truth density.

F.2 Functional data in $L^2(0, 1)$

We next consider functional data in $\mathcal{X} = L^2(0, 1; \mathbb{R}^d)$. Gaussian random elements in this space can be viewed as Gaussian processes with square-integrable sample paths, so the mixture model specializes to a mixture of Gaussian processes [57, 70, 2].

Experiment. We generate multivariate functional data in $\mathcal{X} = L^2(0, 1; \mathbb{R}^2)$ from $K = 5$ Gaussian components with weights $\pi = (0.30, 0.25, 0.20, 0.15, 0.10)$. Each component has a distinct mean function and diagonal covariance profile in coefficient space. We sample $n = 500$ trajectories and project them onto a cosine basis with $R = 15$ functions per spatial dimension, yielding coefficient vectors in $\mathbb{R}^{30}$. Figure 8 shows raw trajectories, recovered mean functions, recovered weights, and the training curve.

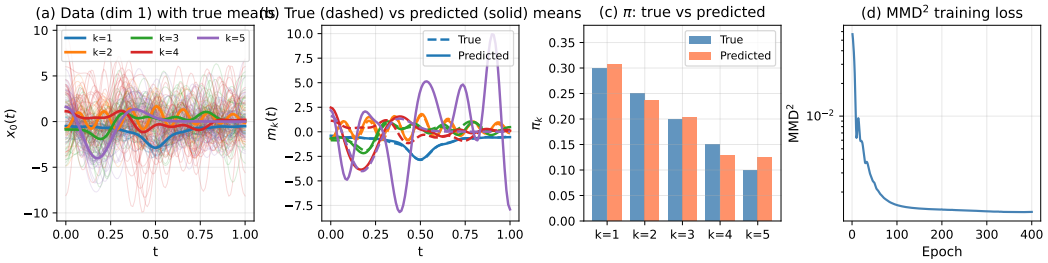


Figure 8: $L^2(0,1;\mathbb{R}^2)$ mixture with $K = 5$. **(a)** Raw trajectories (dimension 1) overlaid with true means. **(b)** True (dashed) versus predicted (solid) mean functions. **(c)** True versus predicted mixture weights. **(d)** MMD^2 training loss.

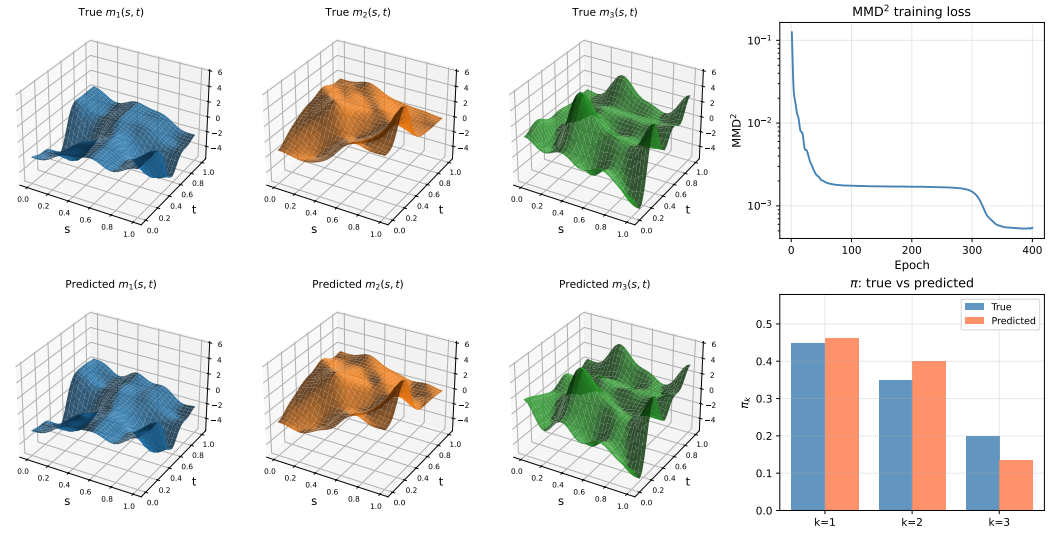


Figure 9: $L^2([0,1]^2)$ mixture with $K = 3$. **Left columns:** True mean surfaces $m_k(s,t)$ (**top row**) and predicted surfaces (**bottom row**), aligned by color. **Right column:** MMD^2 training loss (**top**) and true versus predicted mixture weights (**bottom**).

F.3 Functional data in $L^2([0,1]^2)$

The same construction extends to functions on two-dimensional domains by using tensor-product bases. Given an orthonormal basis $(\varphi_r)_{r\geq 0}$ of $L^2(0,1)$, the functions

$$\Phi_{m,n}(s,t) := \varphi_m(s)\varphi_n(t)$$

form an orthonormal basis of $L^2([0,1]^2)$. This setting covers image-like signals, spatial random fields, and spatio-temporal processes.

Experiment. We consider $\mathcal{X} = L^2([0,1]^2;\mathbb{R})$ with a tensor-product cosine basis, using $R_s = R_t = 8$ and hence $M = 64$ coefficients. We generate $n = 400$ samples from a $K = 3$ Gaussian mixture with weights $\pi = (0.45, 0.35, 0.20)$, where each component has a distinct 2D mean surface. Figure 9 compares the true and predicted mean surfaces and reports the training loss and recovered weights.

F.4 $L^2(\mathrm{SO}(3))$ rotation data

We also consider non-Euclidean domains equipped with an L^2 structure. For a compact group G with Haar probability measure μ, the space $L^2_\mu(G;\mathbb{R}^d)$ is a Hilbert space with inner product

$$\langle f, g\rangle_{\mathcal{X}} := \int_G f(x)^\top g(x)\,\mathrm{d}\mu(x).$$

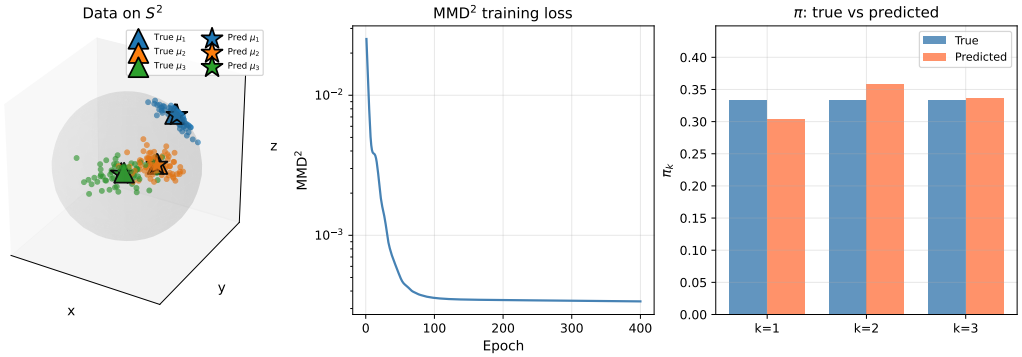


Figure 10: $L^2(\mathrm{SO}(3))$ mixture with $K = 3$. **Left:** Data projected on S^2 with true and learned component directions. **Middle:** MMD^2 training loss. **Right:** True versus predicted mixture weights.

For $G = \mathrm{SO}(3)$, the Peter–Weyl theorem provides an orthonormal basis in terms of Wigner D-matrices.

Experiment. We consider rotation data on $\mathrm{SO}(3)$, relevant for orientation estimation in robotics and molecular modeling. We generate $n = 200$ rotations from $K = 3$ concentrated components and encode them using real Wigner-basis features up to degree $L_{\max} = 3$, giving $M = 84$ coefficients. Figure 10 displays each rotation through the image of the reference direction $[1, 0, 0]$ on S^2, together with the true and learned component directions, the training loss, and the recovered weights.

F.5 Graph signals

We next test the method on graph-structured data, where the Hilbert-space geometry is induced by the graph Laplacian. Let $G = (V, E)$ be a finite weighted graph with Laplacian $L = D - W$. For $\alpha > 0$, we equip $\mathbb{R}^{|V|}$ with the inner product

$$\langle f, g \rangle_{\mathcal{X}} := f^\top (L + \alpha I) g.$$

This geometry gives larger norm to high-frequency graph components, while the finite-dimensional projection below restricts the representation to the leading Laplacian modes.

Experiment. We generate $n = 150$ graph signals on a random Erdős–Rényi graph with $|V| = 30$ nodes and edge probability 0.25. The signals are projected onto the first $M = 15$ Laplacian eigenvectors, using regularization $\alpha = 0.1$. The mixture has $K = 3$ components with uniform weights. Figure 11 shows the true and predicted mean graph signals, together with the training loss and recovered weights.

F.6 Linear SDE: system identification

As a parametric Gaussian-process example, we consider the linear stochastic differential equation

$$\mathrm{d}x(t) = \big(Ax(t) + Bu(t)\big)\,\mathrm{d}t + G\,\mathrm{d}W(t), \qquad t \in [a, b].$$

For Gaussian initial condition and deterministic input u, the solution is a Gaussian process. Its mean $m(t)$ and covariance $\Sigma(t) = \mathrm{Cov}[x(t), x(t)]$ satisfy

$$\dot{m} = Am + Bu, \qquad m(a) = m_0, \qquad \dot{\Sigma} = A\Sigma + \Sigma A^\top + GG^\top, \qquad \Sigma(a) = \Sigma_0.$$

The resulting projected mean and covariance can therefore be computed analytically in any chosen L^2 basis.

Experiment. We treat (A, B, G) as learnable parameters and identify them from $n_s = 20$ sampled trajectories of a 4-dimensional linear SDE over $[0, 5]$, driven by the input $u(t) = (\sin t, \cos 2t)$. The trajectories are projected onto an L^2 cosine basis with $R = 7$ functions per state dimension, giving $M = 28$ coefficients. We match the empirical coefficient distribution to the single Gaussian induced by the linear SDE. Figure 12 shows the sample paths, training loss, recovered mean function, and diagonal of the projected covariance.

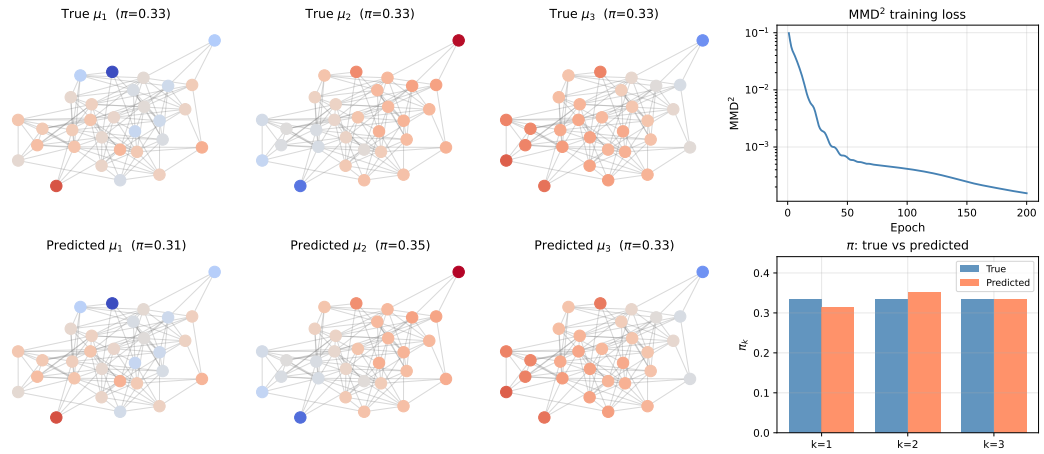


Figure 11: Graph-signal mixture with $K = 3$. **Left block:** True (**top**) versus predicted (**bottom**) mean signals per component, plotted on the shared Erdős–Rényi graph using a common colormap. **Right column:** MMD^2 training loss (**top**) and true versus predicted weights (**bottom**).

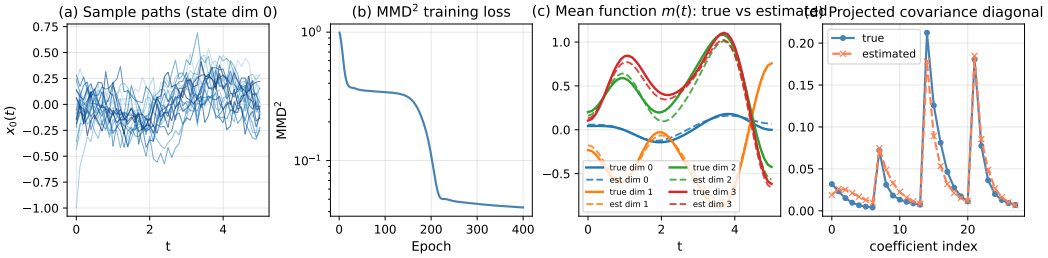


Figure 12: Linear SDE system identification via MMD. **(a)** Sample paths of the first state dimension. **(b)** MMD^2 training loss. **(c)** True versus estimated mean function $m(t)$ for every state dimension. **(d)** Diagonal of the projected covariance matrix.

F.7 3D molecular data (QM9)

We also test the method on real molecular structures using permutation-invariant graph fingerprints derived from 3D atomic coordinates.

Experiment. We sample $n = 500$ molecules from QM9, build WL-hash fingerprints of dimension 128 from distance-threshold molecular graphs, and project them onto the first $R = 64$ DCT coefficients. We fit a $K = 5$ Gaussian mixture with diagonal covariance. Figure 13 shows the three molecules closest to each learned component mean in coefficient space. The HOMO–LUMO gap and assignment responsibility are reported for each representative molecule.

F.8 NTU RGB+D skeleton sequences

Finally, we apply the method to human motion data, treating skeleton sequences as $\mathbb{R}^{75}$-valued paths corresponding to 25 joints with 3 coordinates each.

Experiment. We load $n = 150$ action sequences from the NTU RGB+D dataset [61], resample each sequence to 64 frames, subtract the mean trajectory, and project onto an L^2 cosine basis with $R = 10$. This gives coefficient vectors of dimension $M = 750$. We fit a $K = 5$ Gaussian mixture with diagonal covariance by minimizing MMD^2. Figure 14 shows the most representative skeleton sequences per component, selected as those closest to the learned component mean in coefficient space.

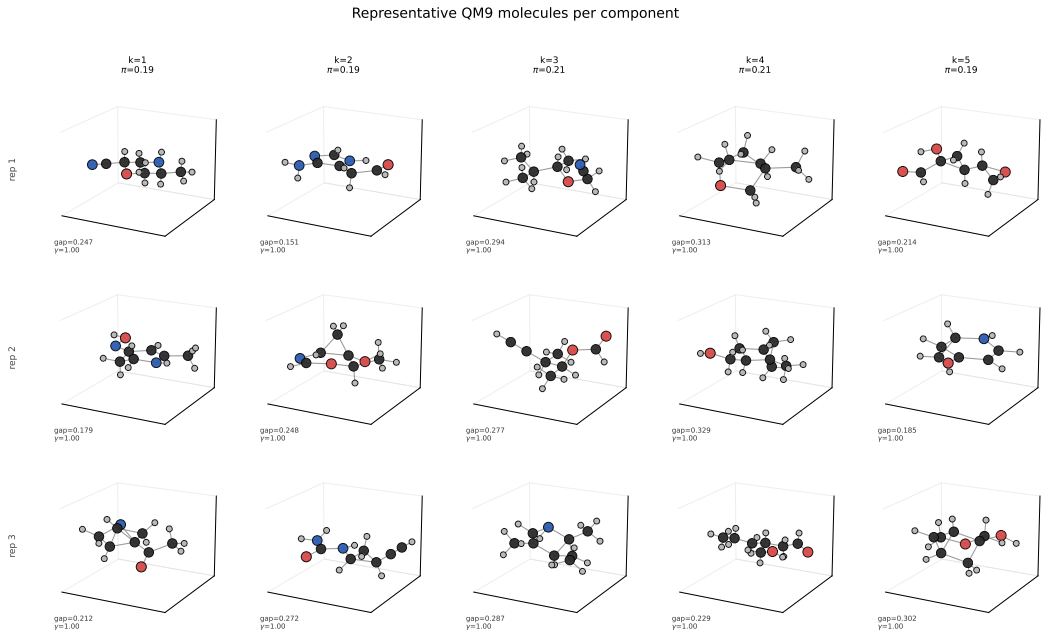


Figure 13: Representative QM9 molecules per component. Columns correspond to learned components $k = 1, \ldots, K$, and rows display the three closest molecules to each component mean in coefficient space. Each panel reports the HOMO-LUMO gap and assignment responsibility $\gamma_{i,k}$.

F.9 Summary

Table 3 summarizes the experimental settings and final MMD^2 values. The controlled experiments show that the method recovers the main mixture structure across Euclidean, functional, group-valued, and graph-based representations. The real-data experiments illustrate that the same coefficient-space fitting procedure also yields coherent representative components on molecular and skeleton-sequence data.

G Experimental details and compute resources

We record the implementation details used for the experiments in Sections 4 and F. Unless stated otherwise, all models are trained on projected coefficient vectors in $\mathbb{R}^M$, use `torch.float64`, use fixed random seeds, run on CPU, and initialize mixture means with k-means++.

G.1 Hilbert spaces and bases

Each experiment fits the projected mixture model of Section 2.2 after fixing a separable Hilbert space $\mathcal{X}$ and an orthonormal basis $(e_r)_{r\geq 1}$, which determines the truncated subspace $\mathcal{X}_M = \mathrm{span}\{e_1, \ldots, e_M\}$. We define here the Hilbert spaces and bases used in Tables 2 and 3, so that the table entries refer to this section instead of repeating the definitions.

(B1) Euclidean canonical basis. For $\mathcal{X} = \mathbb{R}^d$ we use the canonical basis $(e_r)_{r=1}^d$, so $M = d$.

(B2) L^2 cosine basis. For $\mathcal{X} = L^2(0, 1; \mathbb{R})$ with $\langle f, g\rangle = \int_0^1 f(t)g(t)\,\mathrm{d}t$, we use the orthonormal cosine basis

$$e_0(t) = 1, \qquad e_r(t) = \sqrt{2}\cos(\pi r t), \quad r \geq 1,$$

truncated at $r = R - 1$, so $M = R$. For vector-valued data $\mathcal{X} = L^2(0, 1; \mathbb{R}^d)$, we tensor with the canonical basis of $\mathbb{R}^d$, so $M = Rd$.

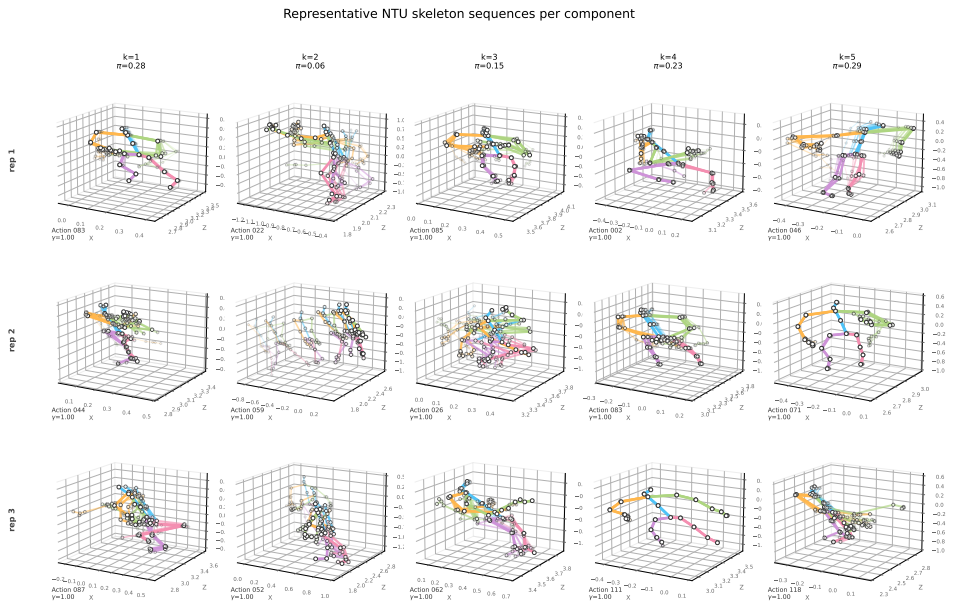


Figure 14: Most representative NTU skeleton sequences per component. Columns correspond to learned components $k = 1, \ldots, K$, and rows display the three closest sequences to each component mean. Each cell shows 6 evenly-spaced snapshots overlaid with increasing opacity, along with the action label and the assignment responsibility $\gamma_{i,k}$.

(B3) Tensor-product cosine basis on $[0,1]^2$. For $\mathcal{X} = L^2([0,1]^2;\mathbb{R})$ we use $\Phi_{m,n}(s,t) = e_m(s)\,e_n(t)$ with $0 \le m < R_s$, $0 \le n < R_t$, so $M = R_s R_t$.

(B4) H^1 cosine basis. For $\mathcal{X} = H^1(0,1;\mathbb{R})$ with $\langle f, g\rangle = \int_0^1 fg + \int_0^1 f'g'$, we use the H^1-rescaled cosines

$$\tilde{e}_r(t) = \frac{e_r(t)}{\sqrt{1+\pi^2 r^2}}, \qquad r = 0, \ldots, R_s - 1,$$

so $M = R_s$; the $\tilde{e}_r$ are orthonormal in H^1.

(B5) Real Wigner–D basis on $\mathrm{SO}(3)$. For $\mathcal{X} = L^2(\mathrm{SO}(3);\mathbb{R})$ under the Haar inner product, the Peter–Weyl theorem yields an orthonormal basis given by the (real) Wigner D-matrix entries,

$$e_{\ell,m,n}(R) = \sqrt{2\ell+1}\, D^{\ell}_{m,n}(R), \qquad \ell \ge 0,\ -\ell \le m, n \le \ell.$$

Truncating at $\ell \le L_{\max}$ gives $M = \sum_{\ell=0}^{L_{\max}} (2\ell+1)^2$; for $L_{\max} = 3$, $M = 84$.

(B6) Graph Laplacian eigenbasis. For a finite weighted graph $G = (V, E)$ with Laplacian $L = D - W$ and regularization $\alpha > 0$, we equip $\mathcal{X} = \mathbb{R}^{|V|}$ with $\langle f, g\rangle = f^\top (L + \alpha I) g$. Diagonalizing $L = U \Lambda U^\top$ with eigenpairs (λ_j, u_j), the rescaled eigenvectors $e_j = u_j/\sqrt{\lambda_j + \alpha}$ form an orthonormal basis of $\mathcal{X}$. We keep the M eigenvectors associated with the smallest eigenvalues.

(B7) Frobenius basis on $\mathrm{Sym}(d)$. For $\mathcal{X} = \mathrm{Sym}(d) \subset \mathbb{R}^{d\times d}$ with $\langle A, B\rangle = \mathrm{tr}(A^\top B)$, we use the canonical symmetric basis $\{E_{ii}\} \cup \{(E_{ij} + E_{ji})/\sqrt{2}\}_{i<j}$ (`vech` embedding), so $M = \binom{d+1}{2}$.

(B8) Weisfeiler–Lehman (WL) with kernel PCA / DCT. For molecular graphs (MUTAG, QM9) we compute WL hash fingerprints in $\mathbb{R}^D$, then project onto the leading M kernel-PCA directions (MUTAG) or DCT coefficients (QM9) to obtain coefficient vectors in $\mathbb{R}^M$. The basis is fixed in coefficient space and inherited from the projection.

Table 2: Benchmark hyperparameters and evaluation protocol. Bases (B1)–(B8) are defined in Section G.1. All MMD-GMM fits use Adam and k-means++ initialization; polynomial kernels use degree 2 and $c = 1$.

Experiment	Data and Runs	Representation and MMD-GMM training
$\mathbb{R}^d$ toy data	Five labeled `sklearn` datasets [53], $n = 500$, $d = 2$, $K \in \{2, 3\}$; seed 30, MMD seed 42.	*Representation:* canonical basis (B1) on standardized coordinates, $M = 2$. *Training:* 300 epochs, learning rate 0.05, full covariance; Gaussian bandwidth by 0.5 times the median pairwise-distance heuristic.
L^2 functional data	Growth and Canadian Weather [56], Waveform [11], Phoneme [28], Kneading [54], and ECG200 [51]; ARI averaged over the six datasets; seed 42.	*Representation:* L^2 cosine basis (B2), $R = 15$ ($R = 10$ for Waveform), coefficient normalization per dataset. *Training:* 400 epochs, learning rate 0.1, diagonal covariance; Gaussian bandwidth by median heuristic.
Glucodensity clustering	PEDAP CGM cohort [71, 2], 98 patients, $K = 2$; 5 MMD runs; seed 42.	*Representation:* H^1 cosine basis (B4) with $R_s = 8$ on overlapping 4-day sliding windows (stride 1 day, 16 time bins); patient-level features for our method. *Training:* 600 epochs, learning rate 0.01, diagonal covariance; cosine annealing to 10^{-4}; Gaussian bandwidth by median heuristic.
$\mathrm{SO}(3)$ rotations	10 generated datasets, $n = 200$, $K = 3$, noise concentration 8.0; seeds $42, \ldots, 51$.	*Representation:* real Wigner–D basis (B5) with $L_{\max} = 3$, $M = 84$. *Training:* 400 epochs, learning rate 0.1, diagonal covariance; Gaussian bandwidth by median heuristic.
MUTAG molecular graphs	MUTAG [22], $n = 188$, $K = 2, 5$ runs; seed 42.	*Representation:* WL fingerprints with kernel PCA (B8), fingerprint dimension 512, radius 3, reduced to $M = 32$. *Training:* 500 epochs, learning rate 0.05, 5 runs; Gaussian bandwidth selected from $\{0.1, 0.5, 1, 2, 4\}$ plus the median heuristic by a 60-epoch MMD proxy sweep.

G.2 Experimental setting

The clustering experiments are unsupervised: all methods are fit without access to labels, and labels are used only after fitting to compute Adjusted Rand Index (ARI). There are no train/test splits because the reported task is clustering or mixture recovery on a fixed sample; variability is instead measured across independently generated datasets, benchmark datasets, random restarts, or subsampled runs, as specified in Table 2.

The standalone mixture-recovery figures and case-study panels in Sections 4, 4.2 and F that are not part of the unified ARI benchmark follow the configurations summarized in Table 3.

G.3 Statistical reporting

All entries in Table 1 are reported as mean ARI with one standard deviation in subscript, computed directly from the list of ARI scores saved by each benchmark script. The standard deviation is a descriptive measure of run-to-run or dataset-to-dataset variability, not a standard error and not a formal confidence interval. The variability source is the labelled dataset collection for the $\mathbb{R}^d$ and L^2 columns, independent generated datasets for $\mathrm{SO}(3)$, and multiple runs for glucodensity and MUTAG. Deterministic single-dataset baselines have zero reported standard deviation when only one fitted value is produced.

G.4 Compute resources

All reported experiments were run on CPU only; no GPU was used or required. The local machine used for reproduction is an Ubuntu 24.04.4 LTS workstation equipped with two Intel Xeon E5-2650 v2 CPUs at 2.60 GHz (16 physical cores in total), 188 GiB of RAM, and a network-mounted filesystem. The software environment relies on `Python` 3.11.5 with the following core libraries: `PyTorch` 1.13.1, `NumPy` 1.26.4, `SciPy` 1.12.0, `scikit-learn` 1.5.2, `pandas` 1.5.3, `scikit-fda` 0.10.1, `torch-geometric` 2.7.0, `aeon` 0.11.1, `rdata` 1.0.0, `torchdiffeq` 0.2.5, and `torchsde` 0.2.6. The end-to-end reproduction pipeline runs as a single CPU-worker workflow and typically requires 2–4 wall-clock hours, depending on cached downloads and optional notebook regeneration.

H Additional comparisons and diagnostics

This section reports additional diagnostics: the stability of fitted partitions across initializations, the scaling of runtime and memory with the number of components, and the numerical stability of the closed-form expectations.

Table 3: Hyperparameters for the additional mixture-recovery and case-study figures. Bases (B1)–(B8) are defined in Section G.1.

Experiment	Basis	n	K	M	Epochs	Kernel/optimizer details
1D Gaussian mixture	(B1)	1500	3	1	400	Gaussian $\sigma = 1.0$, Adam learning rate 0.05.
$L^2(0,1;\mathbb{R}^2)$ recovery	(B2)	500	5	30	400	Cosine basis with $R = 15$ per coordinate; Gaussian $\sigma = 1.2$, Adam learning rate 0.1.
$L^2([0,1]^2)$ recovery	(B3)	400	3	64	400	Tensor-product cosine $R_s = R_t = 8$; Gaussian $\sigma = 2.0$, Adam learning rate 0.1.
$L^2(\mathrm{SO}(3))$ recovery	(B5)	200	3	84	400	Real Wigner–D, $L_{\max} = 3$; Gaussian $\sigma = 10.0$, Adam learning rate 0.05.
Graph-signal recovery	(B6)	150	3	15	200	Erdős–Rényi graph, $\alpha = 0.1$; Gaussian $\sigma = 2.0$, Adam learning rate 0.1.
Linear SDE identification	(B2)	20 paths	1	28	400	Cosine basis with $R = 7$ per state dim.; Gaussian bandwidth $2\,\mathrm{sd}(X)$, Adam learning rate 10^{-2}.
QM9 molecules	(B8)	500	5	64	150	WL fingerprints + first 64 DCT coefficients; Gaussian $\sigma = 2.0$, Adam learning rate 0.05.
NTU representatives	(B2)	150	5	750	250	Per-joint cosine basis $R = 10$ on $\mathbb{R}^{75}$-valued paths; Gaussian $\sigma = 2.0$, Adam learning rate 0.05.
CGM temporal H^1 mixture	(B4)	98 patients	2	8	400	Overlapping 4-day sliding windows with stride 1 day, 16 time bins, H^1 cosine basis with $R_s = 8$; 2-layer neural ODE weights, hidden width 64, RK4, Gaussian $\sigma = 0.8$, Adam learning rate 0.01.
CGM correlation mixture	(B7)	21,889 windows	3	300	200	Overlapping $W = 14$-day sliding windows with stride 1 day; each window gives a 24×24 inter-hour correlation matrix with linear shrinkage toward the identity (intensity $\alpha = 0.1$), `vech` embedding into $\mathrm{Sym}(24)$; 2-layer neural ODE logits, hidden width 64, RK4, Adam learning rate 10^{-2}.
CGM patient graph mixture	(B6)	76 patients	3	304	200	Overlapping 21-day sliding windows with stride 1 day, 10 time bins; pairwise H^1 similarities use $R_s = 16$ cosine modes; graph Laplacian basis rank $R = 4$; 2-layer neural ODE logits, hidden width 64, RK4, Adam learning rate 10^{-2}.

Table 4: Initialization stability: pairwise ARI (mean $\pm$ SD over ten partition pairs from five initializations). M denotes cosine modes per channel. The final row averages the six dataset means at each M.

Dataset	$M = 1$	$M = 5$	$M = 10$
Growth ($K = 2$)	0.359 ± 0.369	0.859 ± 0.079	0.974 ± 0.021
Canadian Weather ($K = 4$)	0.508 ± 0.145	0.716 ± 0.107	0.731 ± 0.189
Waveform ($K = 3$)	0.260 ± 0.254	0.910 ± 0.035	0.598 ± 0.168
Phoneme ($K = 5$)	0.606 ± 0.181	0.671 ± 0.140	0.687 ± 0.074
Kneading ($K = 3$)	0.337 ± 0.283	0.467 ± 0.152	0.633 ± 0.157
ECG200 ($K = 2$)	0.137 ± 0.201	0.458 ± 0.450	0.898 ± 0.059
Mean	0.368	0.680	0.754

H.1 Partition stability across initializations

To measure sensitivity to initialization and the effect of the projection dimension, we refit each model from five independent k-means++ initializations at each projection resolution $M \in \{1, 5, 10\}$ and compare the resulting partitions. The Gaussian bandwidth is the median of all pairwise distances between the standardized coefficient vectors.

Table 4 reports the mean and standard deviation of the ten pairwise ARIs among the five fitted partitions, rather than ARI against ground-truth labels. The ten pairs share partitions, so their standard deviation describes dispersion and is not a standard error. Agreement generally increases with M: the average across datasets rises steadily, and five of the six datasets become more stable at every step.

H.2 Scaling in the number of components

We fit balanced, well-separated Gaussian mixtures in $\mathbb{R}^{20}$ with $n = 10^5$ samples and an increasing number of components K, using five k-means++ initializations per K. All fits use the Gaussian

kernel with a median-heuristic bandwidth, diagonal covariances, and joint Adam updates in float64 on CPU, omitting the data–data constant. The cross-expectations $\mathbf{J}$ are evaluated as matrix products, which avoids forming an $n \times K \times M$ tensor and keeps memory at $O(nM + nK + K^2M)$.

Table 5: Component sweep at $n = 10^5$, $M = 20$: ARI over five initializations, median time per fit (initialization and optimization), and peak memory above the pre-data baseline.

Components K	5	10	20	50	100	200
ARI (mean ± SD)	1.000 ± 0.000	1.000 ± 0.000	0.977 ± 0.028	0.987 ± 0.017	0.979 ± 0.008	0.976 ± 0.004
Time/fit (s)	9.04	15.06	24.18	85.04	176.01	239.74
Memory (MiB)	185	188	355	295	568	983

Recovery remains near perfect across the sweep, while runtime and memory grow moderately with K and stay well within the capacity of a single workstation. These measurements show feasibility under controlled separation, not recovery guarantees for arbitrary large mixtures, and absolute timings depend on the execution environment.

H.3 Numerical stability

For the isotropic Gaussian kernel, the full-covariance expectations involve

$$A_k = I_M + \mathbf{K}_{k,M}/\sigma_\kappa^2, \qquad A_{k,s} = I_M + (\mathbf{K}_{k,M} + \mathbf{K}_{s,M})/\sigma_\kappa^2.$$

Positive semidefiniteness implies $\lambda_{\min}(A_k) \geq 1$ and

$$\operatorname{cond}_2(A_k) = \frac{1 + \lambda_{\max}(\mathbf{K}_{k,M})/\sigma_\kappa^2}{1 + \lambda_{\min}(\mathbf{K}_{k,M})/\sigma_\kappa^2} \leq 1 + \frac{\operatorname{tr}(\mathbf{K}_{k,M})}{\sigma_\kappa^2},$$

with the analogous bound for $A_{k,s}$. The identity shift prevents singularity, but does not guarantee good conditioning at small bandwidths. The spectral formulas in Section B.1 can be evaluated by eigendecomposition; the code instead uses Cholesky log-determinants and solves, without explicit inverses. These matrices are $M \times M$, not $n \times n$. Diagonal covariances require only elementwise operations in the optimization loop.

I Proofs

I.1 Proof of Proposition 1

Proof. Since $(e_j)_{j \geq 1}$ is an orthonormal basis of $\mathcal{X}$, the orthogonal projections Π_M converge strongly to the identity. That is, for every $x \in \mathcal{X}$, $\Pi_M x \to x$ in $\mathcal{X}$ as $M \to \infty$.

We first prove the almost sure convergence of $J_{i,k}^{(M)}$. Fix $i \in \{1, \ldots, n\}$ and $k \in \{1, \ldots, K\}$. Fix ω such that $x_i := X_i(\omega) \in \mathcal{X}$. Then $\Pi_M x_i \to x_i$ in $\mathcal{X}$. By the continuity of κ, for every $y \in \mathcal{X}$, we have $\kappa(\Pi_M x_i, \Pi_M y) \to \kappa(x_i, y)$. Furthermore, using the growth assumption (9) and the contraction property of the orthogonal projector ($\|\Pi_M x\|_{\mathcal{X}} \leq \|x\|_{\mathcal{X}}$), we can bound the kernel evaluation as:

$$|\kappa(\Pi_M x_i, \Pi_M y)| \leq C\big(1 + \|\Pi_M x_i\|_{\mathcal{X}}^q\big)\big(1 + \|\Pi_M y\|_{\mathcal{X}}^q\big) \leq C\big(1 + \|x_i\|_{\mathcal{X}}^q\big)\big(1 + \|y\|_{\mathcal{X}}^q\big).$$

Since Gaussian measures on separable Hilbert spaces have finite moments of all orders by Fernique's theorem, the right-hand side is ν_k-integrable with respect to y for our fixed x_i. Therefore, by the dominated convergence theorem,

$$J_{i,k}^{(M)}(\omega) = \int_{\mathcal{X}} \kappa(\Pi_M x_i, \Pi_M y)\, \mathrm{d}\nu_k(y) \to \int_{\mathcal{X}} \kappa(x_i, y)\, \mathrm{d}\nu_k(y) = J_{i,k}(\omega),$$

which establishes $J_{i,k}^{(M)} \to J_{i,k}$ almost surely.

Next, we establish the convergence of $I_{k,s}^{(M)}$. Fix $k, s \in \{1, \ldots, K\}$. By the continuity of κ, $\kappa(\Pi_M y, \Pi_M y') \to \kappa(y, y')$ for all $y, y' \in \mathcal{X}$. Similarly to the previous case, applying (9) and the contraction property yields:

$$|\kappa(\Pi_M y, \Pi_M y')| \leq C\big(1 + \|y\|_{\mathcal{X}}^q\big)\big(1 + \|y'\|_{\mathcal{X}}^q\big).$$

By Fernique's theorem, this dominating function is integrable with respect to the product measure $\nu_k \otimes \nu_s$. Invoking the dominated convergence theorem once more, we obtain

$$I_{k,s}^{(M)} = \int_{\mathcal{X}\times\mathcal{X}} \kappa(\Pi_M y, \Pi_M y')\, \mathrm{d}(\nu_k \otimes \nu_s)(y, y') \to \int_{\mathcal{X}\times\mathcal{X}} \kappa(y, y')\, \mathrm{d}(\nu_k \otimes \nu_s)(y, y') = I_{k,s}.$$

Finally, since $\kappa(\Pi_M X_i, \Pi_M X_j) \to \kappa(X_i, X_j)$ a.s. for all i, j, and the sums in (8) are finite with weights π_k independent of M, combining these results yields (10). □

I.2 Proof of Proposition 2

Proof. Fix $k \in \{1, \ldots, K\}$. We first show that the exact responsibility $\gamma_k(X)$ is a version of the conditional expectation $\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \sigma(X)]$. Let $Q_\theta = \sum_{s=1}^K \pi_s \nu_s$ be the marginal law of X. For any Borel set $A \in \mathcal{B}(\mathcal{X})$, the definition of $\gamma_k = \mathrm{d}(\pi_k \nu_k)/\,\mathrm{d}Q_\theta$ yields

$$\int_A \gamma_k(x)\, \mathrm{d}Q_\theta(x) = \pi_k \nu_k(A) = \mathbb{P}(X \in A,\ Z = k).$$

Equivalently, for every $A \in \mathcal{B}(\mathcal{X})$,

$$\mathbb{E}\big[\gamma_k(X)\mathbf{1}_{\{X\in A\}}\big] = \mathbb{E}\big[\mathbf{1}_{\{Z=k\}}\mathbf{1}_{\{X\in A\}}\big].$$

Since $\gamma_k(X)$ is $\sigma(X)$-measurable, this identity implies $\gamma_k(X) = \mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \sigma(X)]$ almost surely.

By an identical argument on the projected space, equipping $\mathcal{X}_M$ with the pushforward measures $\nu_{s,M}$ and $Q_{\theta,M}$, the projected derivative $\gamma_k^{(M)} = \mathrm{d}(\pi_k \nu_{k,M})/\,\mathrm{d}Q_{\theta,M}$ satisfies

$$\int_B \gamma_k^{(M)}(u)\, \mathrm{d}Q_{\theta,M}(u) = \mathbb{P}(X^{(M)} \in B,\ Z = k)$$

for any $B \in \mathcal{B}(\mathcal{X}_M)$. Equivalently, for every $B \in \mathcal{B}(\mathcal{X}_M)$,

$$\mathbb{E}\big[\gamma_k^{(M)}(X^{(M)})\, \mathbf{1}_{\{X^{(M)}\in B\}}\big] = \mathbb{E}\big[\mathbf{1}_{\{Z=k\}}\, \mathbf{1}_{\{X^{(M)}\in B\}}\big].$$

Since $\gamma_k^{(M)}(X^{(M)})$ is $\sigma(X^{(M)})$-measurable, this confirms that

$$\gamma_k^{(M)}(X^{(M)}) = \mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \sigma(X^{(M)})]$$

almost surely, establishing (13).

Next, define the filtration $\mathcal{F}_M := \sigma(X^{(M)})$. The property $\Pi_M = \Pi_M \Pi_{M+1}$ ensures that $X^{(M)} = \Pi_M X^{(M+1)}$, so $X^{(M)}$ is $\sigma(X^{(M+1)})$-measurable, meaning the filtration is increasing ($\mathcal{F}_M \subseteq \mathcal{F}_{M+1}$). Let $\mathcal{F}_\infty := \sigma(\cup_{M=1}^\infty \mathcal{F}_M)$. Since each projector Π_M is continuous, $X^{(M)}$ is $\sigma(X)$-measurable, yielding $\mathcal{F}_\infty \subseteq \sigma(X)$. Conversely, since $\Pi_M x \to x$, we have $X^{(M)}(\omega) \to X(\omega)$ pointwise in $\mathcal{X}$. Since $\mathcal{X}$ is a metric space, the pointwise limit of $\mathcal{F}_\infty$-measurable maps into $\mathcal{X}$ is $\mathcal{F}_\infty$-measurable. This forces $\sigma(X) \subseteq \mathcal{F}_\infty$, concluding that $\mathcal{F}_\infty = \sigma(X)$.

Finally, we apply Lévy's upward convergence theorem. Since the target $\mathbf{1}_{\{Z=k\}}$ is integrable, the martingale sequence $\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \mathcal{F}_M]$ converges almost surely and in L^1 to $\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \mathcal{F}_\infty]$. By our previous identities,

$$\mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \mathcal{F}_\infty] = \mathbb{E}[\mathbf{1}_{\{Z=k\}} \mid \sigma(X)] = \gamma_k(X) \qquad \text{a.s.}$$

Substituting these into the limit, we obtain, as $M \to \infty$,

$$\gamma_k^{(M)}(X^{(M)}) \to \gamma_k(X) \qquad \text{a.s. and in } L^1.$$

□

I.3 Proof of Proposition 3

Proof. Proof of (i).

Fix $p \geq 1$, $P \in \mathcal{P}_p(\mathcal{X})$, and $\varepsilon > 0$. Let $\mathcal{K}_0$ be any nonzero trace-class covariance operator on $\mathcal{X}$; for instance, $\mathcal{K}_0 = e \otimes e$ for a unit vector $e \in \mathcal{X}$, which exists because $\mathcal{X} \neq \{0\}$. It is enough to prove

density for mixtures with component covariances of the form $\sigma^2\mathcal{K}_0$, since these form a subclass of the mixtures in (1).

Because $\mathcal{X}$ is a separable Hilbert space, it is Polish. Hence finitely supported probability measures are dense in $\mathcal{P}_p(\mathcal{X})$ for W_p. Therefore, there exists $\nu = \sum_{i=1}^N w_i \delta_{x_i}$ such that $W_p(P, \nu) < \varepsilon/2$.

For $\sigma > 0$, define the Gaussian smoothing

$$\nu_\sigma := \sum_{i=1}^N w_i \mathcal{N}(x_i, \sigma^2 \mathcal{K}_0).$$

Let I be an index with $\mathbb{P}(I = i) = w_i$, and let $Z \sim \mathcal{N}(0, \mathcal{K}_0)$, independent of I. Then $X = x_I$ has law ν, while $Y = x_I + \sigma Z$ has law ν_σ. This coupling gives

$$W_p^p(\nu, \nu_\sigma) \le \mathbb{E}\|X - Y\|_{\mathcal{X}}^p = \sigma^p \mathbb{E}\|Z\|_{\mathcal{X}}^p.$$

Since $\mathcal{K}_0$ is trace-class, Fernique's theorem implies $\mathbb{E}\|Z\|_{\mathcal{X}}^p < \infty$. Hence $W_p(\nu, \nu_\sigma) \to 0$ as $\sigma \downarrow 0$. Choosing σ small enough yields $W_p(\nu, \nu_\sigma) < \varepsilon/2$, and the triangle inequality gives

$$W_p(P, \nu_\sigma) < \varepsilon.$$

Thus, finite Gaussian mixtures with nonzero covariance operators are W_p-dense in $\mathcal{P}_p(\mathcal{X})$.

Proof of (ii). Fix $P \in \mathcal{P}(\mathcal{X})$ and draw K i.i.d. samples $X_1, \dots, X_K \sim P$. Let the empirical measure be $\widehat{P}_K := \frac{1}{K}\sum_{k=1}^K \delta_{X_k}$. Since $\mathbb{E}[\kappa(X_k, \cdot)] = \mu_P$ and we assume $\sup_x \kappa(x, x) \le 1$, the expected squared MMD is bounded by the variance of the mean embedding:

$$\mathbb{E}\big[\,\mathrm{MMD}_\kappa^2(P, \widehat{P}_K)\big] = \frac{1}{K}\mathbb{E}\big[\|\kappa(X_1, \cdot) - \mu_P\|_{\mathcal{H}_\kappa}^2\big] \le \frac{1}{K}\mathbb{E}[\kappa(X_1, X_1)] \le \frac{1}{K}.$$

By the probabilistic method, there exists a specific realization $m_1, \dots, m_K \in \mathcal{X}$ such that the deterministic empirical measure $\mu_K := \frac{1}{K}\sum_{k=1}^K \delta_{m_k}$ satisfies $\mathrm{MMD}_\kappa(P, \mu_K) \le 1/\sqrt{K}$.

Fix any nonzero trace-class covariance operator $\mathcal{K}_0$ (such as $e \otimes e$ for a unit vector e, which exists because $\mathcal{X} \neq \{0\}$) and define $Q_{K,\sigma} := \frac{1}{K}\sum_{k=1}^K \mathcal{N}(m_k, \sigma^2\mathcal{K}_0)$. Then $Q_{K,\sigma}$ converges weakly to μ_K as $\sigma \downarrow 0$. Because κ is bounded and continuous, weak convergence implies convergence in MMD. Thus, we can choose $\sigma > 0$ small enough such that $\mathrm{MMD}_\kappa(Q_{K,\sigma}, \mu_K) \le 1/\sqrt{K}$.

Applying the triangle inequality yields

$$\mathrm{MMD}_\kappa(P, Q_{K,\sigma}) \le \mathrm{MMD}_\kappa(P, \mu_K) + \mathrm{MMD}_\kappa(\mu_K, Q_{K,\sigma}) \le \frac{2}{\sqrt{K}},$$

which completes the proof, as $Q_{K,\sigma}$ is a valid mixture of the form (1). □

I.4 Proof of Proposition 6

Proof. **Step 1.** Let $\Phi(x) = \kappa(x, \cdot) \in \mathcal{H}_\kappa$ be the canonical feature map. Because $\mathcal{X}$ is separable and κ is continuous, the reproducing kernel Hilbert space $\mathcal{H}_\kappa$ is separable. Define the Bochner space

$$\mathbb{H} := L^2(0, T; \mathcal{H}_\kappa).$$

For each sample path, define the random element $Y_i(t) := \Phi(X_i(t))$. Since Φ is continuous, $(\omega, t) \mapsto \Phi(X_i(t, \omega))$ is jointly measurable. Together with the separability of $\mathcal{H}_\kappa$ and the integrability estimate below, this defines a strongly measurable $\mathbb{H}$-valued random variable. By assumption, its squared norm is integrable:

$$\mathbb{E}[\|Y_1\|_{\mathbb{H}}^2] = \mathbb{E}\left[\int_0^T \|\Phi(X_1(t))\|_{\mathcal{H}_\kappa}^2 \,\mathrm{d}t\right] = \mathbb{E}\left[\int_0^T \kappa(X_1(t), X_1(t))\,\mathrm{d}t\right] < \infty.$$

Thus, $Y_1, \dots, Y_n$ are i.i.d. $\mathbb{H}$-valued random variables with finite first and second moments.

Step 2. By Jensen's inequality and Tonelli's theorem, the population mean embedding trajectory $\mu_P := (\mu_{P(t)})_{t\in[0,T]}$ satisfies $\int_0^T \|\mu_{P(t)}\|_{\mathcal{H}_\kappa}^2 \,\mathrm{d}t \le \mathbb{E}[\|Y_1\|_{\mathbb{H}}^2] < \infty$, so $\mu_P \in \mathbb{H}$. By the Bochner–Fubini theorem, μ_P is exactly the expected value $\mathbb{E}[Y_1]$ in $\mathbb{H}$. Similarly, the empirical mean embedding trajectory μ_{P_n} corresponds to the sample average:

$$\mu_{P_n} = \frac{1}{n}\sum_{i=1}^n Y_i \quad \text{in } \mathbb{H}.$$

By assumption, the target embedding trajectory $\mu_Q := (\mu_{Q(t)})_{t\in[0,T]}$ also belongs to $\mathbb{H}$.

Step 3. By the strong law of large numbers for Hilbert-space-valued random variables, the sample average converges to its expectation almost surely:

$$\mu_{P_n} \xrightarrow{n\to\infty} \mu_P \qquad \text{in } \mathbb{H} \text{ a.s.}$$

Since $\mu_{P_n} - \mu_Q \to \mu_P - \mu_Q$ in $\mathbb{H}$, the continuity of the squared norm gives $\|\mu_{P_n} - \mu_Q\|_{\mathbb{H}}^2 \to \|\mu_P - \mu_Q\|_{\mathbb{H}}^2$. Because the integrated squared MMD is precisely the squared distance in $\mathbb{H}$ scaled by $1/T$, this immediately implies that

$$\frac{1}{T}\int_0^T \mathrm{MMD}_\kappa^2\big(P_n(t), Q(t)\big)\,\mathrm{d}t \xrightarrow{n\to\infty} \frac{1}{T}\int_0^T \mathrm{MMD}_\kappa^2\big(P(t), Q(t)\big)\,\mathrm{d}t$$

almost surely. Finally, setting $Q(t) = P(t)$ directly yields $\frac{1}{T}\|\mu_{P_n} - \mu_P\|_{\mathbb{H}}^2 \to 0$ almost surely, completing the proof. □

I.5 Proof of Proposition 7

Proof. For each M, the identification of $\mathcal{X}_M$ with $\mathbb{R}^M$ via the basis $(e_r)_{r=1}^M$ maps the projected measures $P_{n,M}(t) := \frac{1}{n}\sum_{i=1}^n \delta_{\Pi_M X_i(t)}$ and $Q_{\theta,M}(t) := \sum_{k=1}^K \pi_k(t)\nu_{k,M}$ to their finite-dimensional coordinate representations $\mathbf{P}_{n,M}(t)$ and $\mathbf{Q}_{\theta,M}(t)$. By defining $\kappa_M(\mathbf{x},\mathbf{y}) := \kappa(\sum_{r=1}^M x_r e_r, \sum_{r=1}^M y_r e_r)$, we have

$$\mathrm{MMD}_{\kappa_M}^2\big(\mathbf{P}_{n,M}(t), \mathbf{Q}_{\theta,M}(t)\big) = \mathrm{MMD}_\kappa^2\big(P_{n,M}(t), Q_{\theta,M}(t)\big)$$

for every t. Expanding the exact and projected MMD objectives, we define the cross-expectations

$$J_{i,k}^{(M)}(t) := \mathbb{E}_{Y\sim\nu_k}\big[\kappa(\Pi_M X_i(t), \Pi_M Y)\big], \qquad J_{i,k}(t) := \mathbb{E}_{Y\sim\nu_k}\big[\kappa(X_i(t), Y)\big],$$
$$I_{k,s}^{(M)} := \mathbb{E}_{Y\sim\nu_k,\,Y'\sim\nu_s}[\kappa(\Pi_M Y, \Pi_M Y')], \qquad I_{k,s} := \mathbb{E}_{Y\sim\nu_k,\,Y'\sim\nu_s}[\kappa(Y, Y')].$$

Fix a sample path satisfying $\int_0^T \|X_i(t)\|_{\mathcal{X}}^{2q}\,\mathrm{d}t < \infty$ for every $i = 1,\dots,n$. For every t, $\Pi_M X_i(t) \to X_i(t)$ in $\mathcal{X}$. Hence, by the continuity of κ, the data–data interactions converge pointwise in t:

$$\kappa(\Pi_M X_i(t), \Pi_M X_j(t)) \to \kappa(X_i(t), X_j(t)).$$

For fixed t, the integrand in $J_{i,k}^{(M)}(t)$ satisfies $\kappa(\Pi_M X_i(t), \Pi_M Y) \to \kappa(X_i(t), Y)$ for every $Y \in \mathcal{X}$, and is bounded by $|\kappa(\Pi_M X_i(t), \Pi_M Y)| \le C(1 + \|X_i(t)\|_{\mathcal{X}}^q)(1 + \|Y\|_{\mathcal{X}}^q)$. Then dominated convergence applies because Gaussian measures on Hilbert spaces have finite q-moments, giving $J_{i,k}^{(M)}(t) \to J_{i,k}(t)$. The exact same argument applies to the component–component term, yielding $I_{k,s}^{(M)} \to I_{k,s}$.

We next justify passing the limit through the time integral using the Dominated Convergence Theorem. By the polynomial growth condition on κ and the contraction property of Π_M,

$$|\kappa(\Pi_M X_i(t), \Pi_M X_j(t))| \le C\big(1 + \|X_i(t)\|_{\mathcal{X}}^q\big)\big(1 + \|X_j(t)\|_{\mathcal{X}}^q\big).$$

This upper bound belongs to $L^1(0,T)$ by Cauchy–Schwarz and the sample path assumption. Likewise, if $Y_k \sim \nu_k$,

$$|J_{i,k}^{(M)}(t)| \le C\big(1 + \|X_i(t)\|_{\mathcal{X}}^q\big)\,\mathbb{E}\big[1 + \|Y_k\|_{\mathcal{X}}^q\big],$$

which is integrable on $[0,T]$. Finally, the component–component term is uniformly bounded by $C\,\mathbb{E}[1 + \|Y_k\|_{\mathcal{X}}^q]\,\mathbb{E}[1 + \|Y_s\|_{\mathcal{X}}^q]$, which is finite and independent of M.

Since $0 \le \pi_k(t)\pi_s(t) \le 1$, we can apply the Dominated Convergence Theorem to integrate term by term. Integrating the finite-dimensional MMD expansion

$$\mathrm{MMD}_\kappa^2\big(P_{n,M}(t), Q_{\theta,M}(t)\big) = \frac{1}{n^2}\sum_{i,j=1}^n \kappa(\Pi_M X_i(t), \Pi_M X_j(t)) - \frac{2}{n}\sum_{i=1}^n\sum_{k=1}^K \pi_k(t) J_{i,k}^{(M)}(t) + \sum_{k,s=1}^K \pi_k(t)\pi_s(t) I_{k,s}^{(M)}$$

over $[0, T]$ yields convergence for fixed n, K, and θ:

$$\int_0^T \mathrm{MMD}_\kappa^2\big(P_{n,M}(t), Q_{\theta,M}(t)\big)\,\mathrm{d}t \xrightarrow{M\to\infty} \int_0^T \mathrm{MMD}_\kappa^2\big(P_n(t), Q_\theta(t)\big)\,\mathrm{d}t \qquad \text{a.s.}$$

Dividing by T and using the coordinate identification κ_M completes the proof. □

I.6 Proof of Proposition 8

Proof. Fix $p \geq 1$, let $P\colon [0,T] \to \mathcal{P}_p(\mathcal{X})$ be continuous with respect to W_p, and fix $\varepsilon > 0$. Because $[0, T]$ is compact and P is continuous, the trajectory image $\mathcal{C} := P([0, T])$ is compact in the metric space $(\mathcal{P}_p(\mathcal{X}), W_p)$.

By Proposition 3, for every $\mu \in \mathcal{C}$, there exists a finite Gaussian mixture q_μ such that $W_p(\mu, q_\mu) < \varepsilon/4$. The open balls $U_\mu := B_{W_p}(\mu, \varepsilon/4) \cap \mathcal{C}$ form an open cover of $\mathcal{C}$. By compactness, we can extract a finite subcover centered at $\mu_1, \ldots, \mu_J \in \mathcal{C}$, with corresponding approximating mixtures $q_1, \ldots, q_J$.

Let $\{\lambda_j\}_{j=1}^J$ be a continuous partition of unity on $\mathcal{C}$ subordinate to this finite subcover. That is, each $\lambda_j\colon \mathcal{C} \to [0, 1]$ is continuous, $\sum_{j=1}^J \lambda_j(\mu) = 1$ for all $\mu \in \mathcal{C}$, and $\lambda_j(\mu) = 0$ whenever $\mu \notin U_j$.

Since each q_j is a finite Gaussian mixture, we can pool all distinct Gaussian components appearing across $q_1, \ldots, q_J$ into a single shared dictionary $\{\nu_1, \ldots, \nu_K\}$. We can then express each local mixture over this shared basis as

$$q_j = \sum_{k=1}^K a_{j,k}\nu_k, \qquad \text{where } a_{j,k} \geq 0 \text{ and } \sum_{k=1}^K a_{j,k} = 1.$$

We now construct the temporal weights by interpolating the coefficients:

$$\pi_k(t) := \sum_{j=1}^J \lambda_j(P(t))\, a_{j,k}, \qquad t \in [0,T], \quad k = 1, \ldots, K.$$

Because P and the partition functions λ_j are continuous, the trajectories $t \mapsto \pi_k(t)$ are continuous. Furthermore, $\pi_k(t) \geq 0$ and $\sum_{k=1}^K \pi_k(t) = 1$. The temporal mixture is therefore given by

$$Q(t) := \sum_{k=1}^K \pi_k(t)\nu_k = \sum_{j=1}^J \lambda_j(P(t))\, q_j.$$

To bound the distance between $P(t)$ and $Q(t)$, fix $t \in [0, T]$. We define the set of active indices at time t as $A(t) := \{j : \lambda_j(P(t)) > 0\}$. For each $j \in A(t)$, $P(t) \in U_j$, meaning $W_p(P(t), \mu_j) < \varepsilon/4$. By the triangle inequality,

$$W_p(P(t), q_j) \leq W_p(P(t), \mu_j) + W_p(\mu_j, q_j) < \varepsilon/2.$$

For each $j \in A(t)$, choose an optimal transport plan $\gamma_{j,t} \in \Pi(P(t), q_j)$, which satisfies $\int \|x - y\|_\mathcal{X}^p \,\mathrm{d}\gamma_{j,t}(x, y) = W_p^p(P(t), q_j) < (\varepsilon/2)^p$. We construct a mixed coupling:

$$\gamma_t := \sum_{j \in A(t)} \lambda_j(P(t))\, \gamma_{j,t}.$$

Because each $\gamma_{j,t}$ has first marginal $P(t)$ and second marginal q_j, the convex combination γ_t has first marginal $P(t)$ and second marginal $\sum_{j\in A(t)} \lambda_j(P(t))\, q_j = Q(t)$. Thus, $\gamma_t \in \Pi(P(t), Q(t))$.

Evaluating the transport cost of this coupling yields

$$\begin{aligned} W_p^p(P(t), Q(t)) &\leq \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_\mathcal{X}^p \,\mathrm{d}\gamma_t(x, y) \\ &= \sum_{j\in A(t)} \lambda_j(P(t)) \int_{\mathcal{X}\times\mathcal{X}} \|x - y\|_\mathcal{X}^p \,\mathrm{d}\gamma_{j,t}(x, y). \end{aligned}$$

Since the integrals are strictly bounded by $(\varepsilon/2)^p$ and the sum of the active weights is 1, we conclude

$$W_p^p(P(t), Q(t)) < \sum_{j \in A(t)} \lambda_j(P(t))\,(\varepsilon/2)^p = (\varepsilon/2)^p.$$

Hence $W_p(P(t), Q(t)) < \varepsilon/2$ for every $t \in [0, T]$, and therefore $\sup_{t\in[0,T]} W_p(P(t), Q(t)) \leq \varepsilon/2 < \varepsilon$, as required. □

I.7 Proof of Proposition 4

Proof. This proof is inspired by [74, Theorem 18]. Since A is bounded, self-adjoint, and positive semidefinite, its square root $T := A^{1/2}$ is a bounded, self-adjoint, positive semidefinite operator, and

$$\kappa_A(x, y) = \exp\left(-\frac{1}{2}\|Tx - Ty\|_{\mathcal{X}}^2\right).$$

By the stability of Gaussian measures under bounded linear maps, if $y \sim \nu_k = \mathcal{N}(m_k, \mathcal{K}_k)$, the transformed variable is distributed as $Ty \sim \mathcal{N}(Tm_k,\, T\mathcal{K}_kT)$. Because $\mathcal{K}_k$ is trace-class and T is bounded, the operator $T\mathcal{K}_kT$ remains trace-class. Thus, the general case reduces to evaluating the isotropic kernel $\kappa(x, y) = \exp(-\frac{1}{2}\|x - y\|_{\mathcal{X}}^2)$ on the transformed variables Tx and Ty.

To evaluate the isotropic case, we rewrite the data-to-model expectation by centering the integration variable. Let $h := X_i - m_k$ and write $y = m_k + \xi$, where $\xi \sim \mathcal{N}(0, \mathcal{K}_k)$. Since $X_i - y = h - \xi$, the expectation becomes:

$$J_{i,k} = \mathbb{E}_{y\sim\nu_k}\left[\exp\left(-\frac{1}{2}\|X_i - y\|_{\mathcal{X}}^2\right)\right] = \mathbb{E}_{\xi}\left[\exp\left(-\frac{1}{2}\|\xi - h\|_{\mathcal{X}}^2\right)\right].$$

By the spectral theorem for compact self-adjoint operators, there exists an orthonormal eigenbasis $(e_r)_{r\geq1}$ of $\mathcal{K}_k$ with corresponding eigenvalues $(\lambda_r)_{r\geq1}$. Expanding ξ and h in this basis yields

$$\xi = \sum_{r=1}^{\infty} \sqrt{\lambda_r} Z_r e_r, \qquad h = \sum_{r=1}^{\infty} h_r e_r,$$

where $Z_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$. For any $N \in \mathbb{N}$, define $\xi^{(N)} := \sum_{r=1}^{N} \sqrt{\lambda_r} Z_r e_r$ and $h^{(N)} := \sum_{r=1}^{N} h_r e_r$. By Parseval's identity, the squared norm of the difference is simply the limit of its partial sums:

$$\|\xi^{(N)} - h^{(N)}\|_{\mathcal{X}}^2 = \sum_{r=1}^{N} (\sqrt{\lambda_r} Z_r - h_r)^2 \xrightarrow{N\to\infty} \sum_{r=1}^{\infty} (\sqrt{\lambda_r} Z_r - h_r)^2 = \|\xi - h\|_{\mathcal{X}}^2 \quad \text{a.s.}$$

Since this convergence holds almost surely and $0 \leq \exp\left(-\frac{1}{2}\|\xi^{(N)} - h^{(N)}\|_{\mathcal{X}}^2\right) \leq 1$, the dominated convergence theorem allows us to compute the expectation as the limit of the finite-dimensional projections:

$$J_{i,k} = \lim_{N\to\infty} \mathbb{E}\left[\exp\left(-\frac{1}{2}\|\xi^{(N)} - h^{(N)}\|_{\mathcal{X}}^2\right)\right] = \lim_{N\to\infty} \prod_{r=1}^{N} \mathbb{E}\left[\exp\left(-\frac{1}{2}(\sqrt{\lambda_r} Z_r - h_r)^2\right)\right].$$

Expanding the square inside the exponential yields $-\frac{1}{2}(\sqrt{\lambda_r} Z_r - h_r)^2 = -\frac{\lambda_r}{2} Z_r^2 + h_r\sqrt{\lambda_r} Z_r - \frac{h_r^2}{2}$. Factoring out the constant term, we obtain:

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}(\sqrt{\lambda_r} Z_r - h_r)^2\right)\right] = e^{-h_r^2/2}\mathbb{E}\left[\exp\left(h_r\sqrt{\lambda_r} Z_r - \frac{\lambda_r}{2} Z_r^2\right)\right].$$

Using the Gaussian identity $\mathbb{E}[\exp(aZ + bZ^2)] = (1 - 2b)^{-1/2}\exp\left(\frac{a^2}{2(1-2b)}\right)$ for $Z \sim \mathcal{N}(0, 1)$, which holds if $2b < 1$, with $a = h_r\sqrt{\lambda_r}$ and $b = -\lambda_r/2$, each factor evaluates to:

$$e^{-h_r^2/2}(1 + \lambda_r)^{-1/2}\exp\left(\frac{h_r^2\lambda_r}{2(1 + \lambda_r)}\right) = (1 + \lambda_r)^{-1/2}\exp\left(-\frac{1}{2}\frac{h_r^2}{1 + \lambda_r}\right).$$

Substituting the one-dimensional evaluations back into the finite-dimensional product yields:

$$\mathbb{E}\left[\exp\left(-\frac{1}{2}\|\xi^{(N)} - h^{(N)}\|_{\mathcal{X}}^2\right)\right] = \left(\prod_{r=1}^{N}(1 + \lambda_r)^{-1/2}\right)\exp\left(-\frac{1}{2}\sum_{r=1}^{N}\frac{h_r^2}{1 + \lambda_r}\right).$$

Passing to the limit as $N \to \infty$, the infinite product converges to $\det_F(I_{\mathcal{X}} + \mathcal{K}_k)^{-1/2}$. Simultaneously, by the spectral decomposition of $(I_{\mathcal{X}} + \mathcal{K}_k)^{-1}$, the series in the exponent evaluates to

$$\sum_{r=1}^{\infty} \frac{h_r^2}{1+\lambda_r} = \langle h, (I_{\mathcal{X}} + \mathcal{K}_k)^{-1} h \rangle_{\mathcal{X}}.$$

Recalling that $h = X_i - m_k$, we conclude:

$$J_{i,k} = \det_F(I_{\mathcal{X}} + \mathcal{K}_k)^{-1/2} \exp\left(-\frac{1}{2}\langle X_i - m_k, (I_{\mathcal{X}} + \mathcal{K}_k)^{-1}(X_i - m_k)\rangle_{\mathcal{X}}\right).$$

Applying this isotropic solution to the transformed variables TX_i and $Ty \sim \mathcal{N}(Tm_k, T\mathcal{K}_k T)$, we obtain the intermediate form:

$$J_{i,k} = \det_F(I_{\mathcal{X}} + T\mathcal{K}_k T)^{-1/2} \exp\left(-\frac{1}{2}\langle T(X_i - m_k), (I_{\mathcal{X}} + T\mathcal{K}_k T)^{-1} T(X_i - m_k)\rangle_{\mathcal{X}}\right).$$

Substituting $T = A^{1/2}$ recovers Equation (20).

Finally, for the cross-expectation $I_{k,s}$, let $y \sim \nu_k$ and $y' \sim \nu_s$ be independent. Their difference $z := y - y'$ is distributed as $\mathcal{N}(m_k - m_s, \mathcal{K}_k + \mathcal{K}_s)$. Since $\kappa_A(y, y') = \kappa_A(z, 0)$, applying the formula for $J_{i,k}$ with $X_i = 0$ immediately yields:

$$I_{k,s} = \frac{\exp\left(-\frac{1}{2}\langle m_k - m_s,\, A^{1/2}\left(I_{\mathcal{X}} + A^{1/2}(\mathcal{K}_k + \mathcal{K}_s)A^{1/2}\right)^{-1} A^{1/2}(m_k - m_s)\rangle_{\mathcal{X}}\right)}{\det_F\left(I_{\mathcal{X}} + A^{1/2}(\mathcal{K}_k + \mathcal{K}_s)A^{1/2}\right)^{1/2}},$$

which completes the proof.

□

I.8 Proof of Proposition 5

Proof. Write $y = m_k + \xi$, where $\xi \sim \mathcal{N}(0, \mathcal{K}_k)$. Then

$$Z := \langle X_i, y\rangle_{\mathcal{X}} + c = \underbrace{\langle X_i, m_k\rangle_{\mathcal{X}} + c}_{\mu_{i,k}} + \langle X_i, \xi\rangle_{\mathcal{X}}$$

is a real Gaussian random variable with mean $\mu_{i,k} \in \mathbb{R}$ and variance

$$v_{i,k} := \mathbb{E}[\langle X_i, \xi\rangle_{\mathcal{X}}^2] = \langle X_i, \mathcal{K}_k X_i\rangle_{\mathcal{X}}.$$

Thus,

$$Z \overset{d}{=} \mu_{i,k} + \sqrt{v_{i,k}}\, G, \qquad G \sim \mathcal{N}(0,1),$$

and therefore

$$J_{i,k} = \mathbb{E}[Z^p] = \sum_{r=0}^{p} \binom{p}{r} \mu_{i,k}^{p-r} v_{i,k}^{r/2}\, \mathbb{E}[G^r].$$

Since $\mathbb{E}[G^r] = 0$ for odd r and $\mathbb{E}[G^{2t}] = \frac{(2t)!}{2^t t!}$, we obtain

$$J_{i,k} = \sum_{t=0}^{\lfloor p/2 \rfloor} \frac{p!}{(p-2t)!\, 2^t\, t!} \langle X_i, \mathcal{K}_k X_i\rangle_{\mathcal{X}}^{t} \left(\langle X_i, m_k\rangle_{\mathcal{X}} + c\right)^{p-2t}.$$

Now let $y \sim \nu_k$ and $y' \sim \nu_s$ be independent. By definition of the polynomial kernel,

$$I_{k,s} = \mathbb{E}\left[(\langle y, y'\rangle_{\mathcal{X}} + c)^p\right].$$

Applying the binomial theorem to expand the power, and using the linearity of expectation,

$$I_{k,s} = \mathbb{E}\left[\sum_{r=0}^{p} \binom{p}{r} c^{p-r} \langle y, y'\rangle_{\mathcal{X}}^r\right] = \sum_{r=0}^{p} \binom{p}{r} c^{p-r}\, \mathbb{E}\left[\langle y, y'\rangle_{\mathcal{X}}^r\right].$$

To evaluate this expectation, recall that $\mathcal{X}^{\otimes r}$ denotes the r-th tensor power of $\mathcal{X}$. For any vector $y \in \mathcal{X}$, its r-fold tensor product is $y^{\otimes r} := y \otimes \cdots \otimes y$. By the standard definition of the inner product on $\mathcal{X}^{\otimes r}$, we have the fundamental identity:

$$\langle y, y' \rangle_{\mathcal{X}}^r = \langle y^{\otimes r}, (y')^{\otimes r} \rangle_{\mathcal{X}^{\otimes r}},$$

where we adopt the convention $\mathcal{X}^{\otimes 0} = \mathbb{R}$ and $y^{\otimes 0} = 1$ for $r = 0$.

Since Gaussian measures on separable Hilbert spaces have finite moments of all orders, the Bochner moments

$$M_{r,k} := \mathbb{E}[y^{\otimes r}] \in \mathcal{X}^{\otimes r}, \qquad M_{r,s} := \mathbb{E}[(y')^{\otimes r}] \in \mathcal{X}^{\otimes r}$$

are well-defined. Using independence, the expectation of the inner product factors as the inner product of the expectations:

$$\mathbb{E}\big[\langle y^{\otimes r}, (y')^{\otimes r} \rangle_{\mathcal{X}^{\otimes r}}\big] = \big\langle \mathbb{E}[y^{\otimes r}], \mathbb{E}[(y')^{\otimes r}] \big\rangle_{\mathcal{X}^{\otimes r}} = \langle M_{r,k}, M_{r,s} \rangle_{\mathcal{X}^{\otimes r}}.$$

Substituting this back into the binomial expansion yields

$$I_{k,s} = \sum_{r=0}^{p} \binom{p}{r} c^{p-r} \, \langle M_{r,k}, M_{r,s} \rangle_{\mathcal{X}^{\otimes r}},$$

which completes the proof. □

I.9 Proof of Proposition 9

Proof. Let ρ be as in the proposition and define $q = dQ/d\rho \in L^1(\rho)$, so that $\int q \, \mathrm{d}\rho = Q(\mathcal{X}) = 1$. Since $Q \ll \rho$, the projected measure $Q_M = Q \circ \Pi_M^{-1}$ satisfies $Q_M \ll \rho_M$, so $q_M = dQ_M/d\rho_M$ is well-defined. Let X denote the identity random element on the probability space $(\mathcal{X}, \mathcal{B}(\mathcal{X}), \rho)$ and set $X^{(M)} := \Pi_M X$. Write $\mathcal{F}_M := \sigma(X^{(M)})$ and $\mathcal{F}_\infty := \sigma(\cup_{M \geq 1} \mathcal{F}_M)$. The same projection argument used in the proof of Proposition 2 gives $\mathcal{F}_\infty = \sigma(X) = \mathcal{B}(\mathcal{X})$.

Step 1 (identification as a conditional expectation). By definition, $q_M = dQ_M/d\rho_M$ and for each $B \in \mathcal{B}(\mathcal{X}_M)$,

$$\int_B q_M(u) \, \mathrm{d}\rho_M(u) = Q_M(B) = Q(\Pi_M^{-1}(B)) = \int_{\Pi_M^{-1}(B)} q(x) \, \mathrm{d}\rho(x).$$

Since $\rho_M = \rho \circ \Pi_M^{-1}$, the left-hand side equals

$$\int_{\mathcal{X}} \mathbf{1}_{\{X^{(M)} \in B\}} \, q_M(X^{(M)}) \, \mathrm{d}\rho.$$

Combining these identities, we obtain

$$\int_{\mathcal{X}} \mathbf{1}_{\{X^{(M)} \in B\}} \, q_M(X^{(M)}) \, \mathrm{d}\rho = \int_{\mathcal{X}} \mathbf{1}_{\{X^{(M)} \in B\}} \, q(X) \, \mathrm{d}\rho \qquad \forall B \in \mathcal{B}(\mathcal{X}_M),$$

so $q_M(X^{(M)})$ is a version of $\mathbb{E}_\rho[q(X) \mid \mathcal{F}_M]$. Equivalently, $q_M(\Pi_M x) = \mathbb{E}_\rho[q(X) \mid \mathcal{F}_M](x)$ for ρ-a.e. x.

Step 2 (martingale convergence). Since $(\mathcal{F}_M)_{M \geq 1}$ is an increasing filtration and $q(X) \in L^1(\rho)$, the martingale convergence theorem yields

$$\mathbb{E}_\rho[q(X) \mid \mathcal{F}_M] \longrightarrow \mathbb{E}_\rho[q(X) \mid \mathcal{F}_\infty] = q(X) \qquad \rho\text{-a.s. and in } L^1(\rho).$$

Combining with Step 1 shows $q_M(\Pi_M x) \to q(x)$ for ρ-a.e. x, and in $L^1(\rho)$.

Step 3 (log-likelihood convergence under Q). Since $Q \ll \rho$, the almost sure convergence in Step 2 also holds Q-a.s. Moreover, $q > 0$ Q-a.s. and $q_M > 0$ Q_M-a.s.; hence $q_M(\Pi_M X) > 0$ for Q-a.e. X, and $\log q(X)$ and $\log q_M(\Pi_M X)$ are well-defined Q-a.s. Therefore, for i.i.d. $X_1, \ldots, X_n \sim Q$,

$$\log q_M(\Pi_M X_i) \longrightarrow \log q(X_i) \qquad \text{a.s. for each } i,$$

and summing over $i = 1, \ldots, n$ yields the claimed convergence of the log-likelihoods. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the papers contributions, supported by theory and experiments, with assumptions and limitations clearly noted.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of our work are introduced in section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The complete list of assumptions and the proofs are introduced in Section I.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Complete methodological details are provided in Sections 4.1, 4.2 and F, covering both the case study and the synthetic experiments. Furthermore, Section G.2 fully discloses all training hyperparameters and configurations necessary for reproducibility. Specifically, Tables 2 and 3 exhaustively list the optimizers, learning rates, training epochs, random seeds, kernel bandwidth selection heuristics, data representations, and neural ODE architectures used across all reported experiments.

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code and scripts to reproduce all experimental results are provided in the supplementary material, which includes a `README.md` with exact execution commands. Complete details of the software environment and dependencies are specified in Section G.4. Information regarding data access, preprocessing, and generation is detailed in Section 4.2 for the clinical case study, and in Table 2 for all other public and synthetic benchmark datasets.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The main experimental setup is described in Sections 4 and 4.2. Full benchmark and case-study hyperparameters, seeds, optimizer choices, kernel bandwidth rules, and repetitions are specified in Section G.2. We also explicitly clarify in this appendix that train/test splits are not applicable due to the unsupervised nature of the clustering tasks, and detail our evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Table 1 reports mean Adjusted Rand Index (ARI) with one standard deviation as subscripts. In Section G.3, we clarify that these are descriptive standard deviations calculated directly from the distribution of scores across runs or datasets. We also detail the specific source of variability for each column (e.g., variation across multiple datasets, random generations, or multiple runs) and explain that deterministic, single-dataset baselines report a 0.000 standard deviation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Complete CPU-only hardware, memory, storage, software, and approximate end-to-end wall-clock time are provided in Section G.4; no GPU was used. Because the pipeline is extremely lightweight (2–4 hours on CPU), the total compute for the entire project, including preliminary runs, was minimal.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This work develops a general-purpose statistical framework and uses synthetic and previously collected research datasets, including publicly available, de-identified clinical-trial data. We did not recruit participants or conduct new clinical interventions. The clinical analyses are exploratory and are not intended to provide medical recommendations.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our method is a general-purpose tool for clustering structured data. As the CGM case study illustrates, it can summarize dense biomarker streams as time-varying mixtures, which may support more precise tracking of glycemic control and earlier detection of changes in disease progression or treatment response. When applied to health data, the usual privacy and validation safeguards for clustering sensitive records apply; we present the case-study findings as exploratory, not as clinical recommendations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All creators or original owners of assets used in the paper are properly credited, and the papers that produced the code package have been cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All assets utilized and created are synthetic or derived from open-source resources, requiring no consent from third parties.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This study does not involve crowdsourcing experiments or direct interaction with human subjects. It is based solely on the analysis of publicly available, anonymized data. Therefore, no participant instructions, screenshots, or compensation details are applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This study uses publicly available and anonymized data that does not involve interaction with human subjects or the collection of private or identifiable information. Therefore, Institutional Review Board (IRB) approval was not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: This work does not involve LLMs as an important, original, or non-standard component of its core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.