

Continuous Temporal Learning of Probability Distributions via Neural ODEs with Applications in Diabetes Digital Trials

Antonio Álvarez-López
Universidad Autónoma de Madrid
and
Marcos Matabuena
Harvard University

September 14, 2025

Abstract

Modeling the dynamics of probability distributions from time-dependent data samples is a fundamental problem in many fields, including digital health. The goal is to analyze how the distribution of a biomarker, such as glucose, changes over time and how these changes may reflect the progression of chronic diseases such as diabetes. We introduce a probabilistic model based on a Gaussian mixture that captures the evolution of a continuous-time stochastic process. Our approach combines a nonparametric estimate of the distribution, obtained with Maximum Mean Discrepancy (MMD), and a Neural Ordinary Differential Equation (Neural ODE) that governs the temporal evolution of the mixture weights. The model is highly interpretable, detects subtle distribution shifts, and remains computationally efficient. We illustrate the broad utility of our approach in a 26-week clinical trial that treats all continuous glucose monitoring (CGM) time series as the primary outcome. This method enables rigorous longitudinal comparisons between the treatment and control arms and yields characterizations that conventional summary-based clinical trials analytical methods typically do not capture.

Keywords: Neural ODEs; Gaussian mixture models; Maximum Mean Discrepancy; continuous glucose monitoring; distribution dynamics; longitudinal analysis

Table of contents

1	Introduction	2
2	Related Work and Preliminaries	6
3	Mathematical models	8
3.0.1	Discrete-time MMD fitting	10
3.0.2	Continuous-Time Weight Evolution via Neural ODEs	11
3.1	Theory	13
4	Simulations	16
4.1	Competitors	16
4.2	Results on synthetic data: low dimensions	19
4.3	Results on synthetic data: high dimensions	20
5	Digital health diabetes case study	20
5.1	Data description and scientific question	20
5.2	Modeling the marginal density of glucose time series	21
5.3	Bivariate analysis	25
6	Discussion	26
7	Computational Details	28
8	Proofs	29

1 Introduction

Characterizing the distribution function of a random variable X is a long-standing problem in statistics [Silverman \(2018\)](#) and machine learning [Bengio et al. \(2017\)](#). Today, it remains a central challenge in the era of generative AI. In large language models [Meskó & Topol](#)

(2023), for example, accurately modeling the probability distribution of textual data is essential for automatic text and report generation.

Beyond natural language processing, distribution estimation is equally important in clinical settings. In digital health, estimating the distribution of physiological time series data over specific time periods enables the construction of patient representations that capture their underlying physiological processes with high precision Matabuena et al. (2021a), Matabuena & Petersen (2023), Ghosal et al. (2023). Recent studies show that, when used properly, such representations can reveal clinically relevant patterns that traditional (non-digital) biomarkers do not detect Katta et al. (2024), Matabuena et al. (2024), Park et al. (2025).

We address the problem of estimating a CDF *continuously* over time: random samples are observed consecutively, and the goal is to learn how the dynamics of the underlying distribution evolve. Naively extending classical CDF or kernel density estimators (KDEs) Chacón & Duong (2018), Papamakarios et al. (2021) with a temporal dimension, or adopting normalized gradient flow techniques, often proves unsatisfactory—KDEs suffer from tuning parameter sensitivity and the curse of dimensionality, while gradient flow methods tend to lack interpretability. Statistical semiparametric approaches provide a partial solution to this problem. For example, time-varying models such as Generalized Additive Models for Location, Scale, and Shape (GAMLSS) alleviate some issues Rigby & Stasinopoulos (2005a), yet most implementations are constrained to scalar responses and can impose rigid functional forms. Recent multilevel functional frameworks based on functional-quantile representations Matabuena & Crainiceanu (2024) offer interpretability but rely on linear dynamics, which limits their ability to capture complex non-linear relationships and multivariate results.

To overcome these limitations, we propose a time-dependent Gaussian mixture model in which a Neural ODE governs the distribution shift smoothly over time, see Chen et al. (2018).

Notation and Problem Definition

Let $\mathcal{T} \subset \mathbb{R}_{\geq 0}$ be a continuous-time index set. For each $t \in \mathcal{T}$, the random variable $X(t) \in \mathbb{R}^d$ denotes the outcome of interest at time t . Its (cumulative) distribution function is

$$F(x, t) = \mathbb{P}(X(t) \leq x) = \int_{\prod_{i=1}^d (-\infty, x_i]} f(r, t) \, dr, \quad (1)$$

for $x = (x_1, \dots, x_d)^\top \in \mathbb{R}^d$, where the inequality is taken component-wise for $d > 1$. Although our target object is $F(\cdot, t)$, in practice we estimate the time-varying density $f(\cdot, t)$ and recover F via (1). We parameterize the density as a K -component Gaussian mixture

$$f(x, t) = \sum_{s=1}^K \alpha_s(t) \mathcal{N}(x | m_s, \Sigma_s),$$

where $m_s \in \mathbb{R}^d$ and $\Sigma_s \in \mathbb{R}^{d \times d}$ are the mean vector and positive definite covariance matrix of the s -th Gaussian component, and the weight vector $\alpha(t) = [\alpha_1(t), \dots, \alpha_K(t)]$ lies in the probability simplex:

$$\alpha(t) \in \Delta^{K-1} := \{ w \in \mathbb{R}^K : w_s \geq 0, \sum_{s=1}^K w_s = 1 \}. \quad (2)$$

We let $\alpha(t)$ vary continuously in time; its evolution is governed by a Neural ODE, endowing the model with universal approximation power while preserving interpretability.

Our model is motivated by the need to track the evolution of glucose distribution in a longitudinal diabetes trial [Battelino et al. \(2023\)](#), where glucose is continuously recorded using continuous glucose monitoring (CGM). In free-living conditions, individual glucose time series cannot be aligned directly, making the raw temporal stochastic process difficult to compare between participants [Ghosal & Matabuena \(2024\)](#), [Matabuena et al. \(2021a, 2024\)](#). Analyzing the time-varying probability distribution therefore provides a more natural and realistic biomarker to characterize the evolution of glucose metabolism [Katta et al. \(2024\)](#), [Park et al. \(2025\)](#), [Matabuena et al. \(2021b\)](#).

We emphasize that this approach conveys richer information than conventional CGM summary statistics —such as mean glucose level or time-in-range metrics—because it captures

the full spectrum of low, moderate and high glucose values simultaneously, which is not possible with traditional summaries [Matabuena et al. \(2021a\)](#), [Katta et al. \(2024\)](#).

Contributions

The main contributions of this paper are:

1. We propose a general framework for modeling the dynamics of multivariate continuous probability distributions in continuous time. Our algorithms scale to large multivariate settings and exploit the universal approximation power of mixture models driven by Neural ODEs.
2. We introduce a novel parameter estimation scheme for mixture models based on a Maximum Mean Discrepancy objective [Gretton et al. \(2012\)](#). This approach avoids the strong structural assumptions and serial correlation issues that affect maximum likelihood methods in this context. Moreover, defining the estimator as a V-statistic [Serfling \(2009\)](#) improves both the robustness and the statistical efficiency. To reduce computational cost, we derive a closed-form expression of the loss function.
3. We provide theoretical guarantees for the algorithm prior to the smoothing step. Specifically, we establish statistical consistency and convergence rates by empirical process analysis of the associated estimators [Kosorok \(2007\)](#). We also analyze the computational complexity of the post-smoothing stage, which explains the acceleration achieved by our implementation, based on massive univariate fitting followed by smoothing, relative to a joint optimization strategy.
4. We demonstrate the practical value of our methods in biomedical applications by analyzing continuous glucose monitoring data from longitudinal diabetes trials [Battelino et al. \(2023\)](#). Our approach offers a novel characterization of the glucose distributional dynamics that existing models cannot capture.

2 Related Work and Preliminaries

From a modeling perspective, the main novelty of our paper is the integration of mixture models with Neural ODEs to create a highly interpretable estimator for the dynamics of probability distributions generated by temporal stochastic processes. Our estimator relies on *Maximum Mean Discrepancy* as its divergence measure, with the aim of improving performance over classical likelihood-based methods. In what follows, we outline the core components of the framework and place it in the context of existing literature.

Maximum Mean Discrepancy (MMD)

To quantify the discrepancy between two probability distributions P and Q defined on a common measurable space \mathcal{X} , we use *Maximum Mean Discrepancy* (MMD) [Gretton et al. \(2012\)](#), [Muandet et al. \(2017\)](#). MMD is a semimetric that equals zero if and only if $P = Q$ provided the kernel is characteristic [Sriperumbudur et al. \(2011\)](#), [Sejdinovic et al. \(2013\)](#).

Let \mathcal{H} be a reproducing kernel Hilbert space (RKHS) with a positive definite kernel $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. Assume that $X \sim P$ and $Y \sim Q$ satisfy $\mathbb{E}[k(X, X)] < \infty$ and $\mathbb{E}[k(Y, Y)] < \infty$, so that all the expectations below are finite. Define *kernel mean embeddings* [Muandet et al. \(2017\)](#) by $\mu_P := \mathbb{E}_X[k(\cdot, X)] \in \mathcal{H}$ and $\mu_Q := \mathbb{E}_Y[k(\cdot, Y)] \in \mathcal{H}$. The squared MMD is then

$$\text{MMD}^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2 = \mathbb{E}_{X, X'}[k(X, X')] + \mathbb{E}_{Y, Y'}[k(Y, Y')] - 2 \mathbb{E}_{X, Y}[k(X, Y)] \geq 0,$$

where X' (resp. Y') is an independent copy of X (resp. Y).

Kernel choice. In our work, we employ the *Gaussian kernel*,

$$k(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \quad (3)$$

where $\sigma > 0$ is a bandwidth parameter that controls smoothness. The Gaussian kernel is *characteristic*; that is, the map $P \mapsto \mu_P$ is injective, so the kernel mean embedding uniquely characterizes the distribution P . Consequently, $\text{MMD}^2(P, Q) = 0$ if and only if $P = Q$.

In practice, the choice of σ strongly impacts the sensitivity of MMD. A standard heuristic sets σ to *median* of all pairwise Euclidean distances between samples—known as the median heuristic [Garreau et al. \(2017\)](#)—which balances sensitivity to both local and global structure in the data.

Intuition. MMD measures how far apart P and Q are in the feature space induced by k : it is the squared distance between their kernel mean embeddings, μ_P and μ_Q . In our setting, we use MMD to compare the empirically evolving time distribution of $X(t)$ with a parametric (or target) distribution, thereby tracking distributional changes over time.

Neural Ordinary Differential Equations (Neural ODEs)

Introduced in [E \(2017\)](#), [Haber & Ruthotto \(2017\)](#), [Chen et al. \(2018\)](#), neural ODEs replace discrete layers by the continuous evolution of a hidden state $\dot{z}(t) = f_\phi(z(t), t)$, $z(t_0) = z_0$, where $f_\phi : \mathbb{R}^d \times [t_0, t_1] \rightarrow \mathbb{R}^d$ is a learnable vector field usually parameterized by a multilayer perceptron. The value $z(t_1) = \text{ODESolve}(z_0, t_0, t_1, f_\phi)$ is then calculated using any standard numerical solver. Gradients with respect to ϕ are then obtained by the adjoint method, allowing constant-memory backpropagation [Massaroli et al. \(2020\)](#).

The continuous framework offers four key benefits: (i) it adapts computation to local dynamics; (ii) it is parameter efficient (a single vector field encodes arbitrary depth); (iii) it induces an (under mild conditions) invertible flow map; and (iv) it yields smooth latent trajectories. In our model, $\alpha(t)$ denotes the vector of mixture weights, so learning f_ϕ captures how the probabilities of the components evolve while respecting the simplex constraint.

Neural ODEs have been applied to distribution learning for generative modeling [Chen et al. \(2018\)](#), and recent work has established approximation and stability guarantees [Marzouk et al. \(2024\)](#), [Álvarez-López et al. \(2024\)](#). Their framework has also been extended to forecasting by modeling latent states as continuous-time trajectories [Kidger et al. \(2020\)](#), [Rubanova et al. \(2019\)](#), [Jia & Benson \(2019\)](#), a strategy that has been shown to be especially effective in biomedical applications [Qian et al. \(2021\)](#).

Distributional Data Analysis Distributional data analysis Ghosal et al. (2023), Szabó et al. (2016) is an emerging field that treats or aggregates probability distributions as random objects for unsupervised and supervised learning, for example, to predict clinical outcomes Matabuena et al. (2021b). Biomedical applications are its most prominent use case. In digital health, measurements collected by continuous glucose monitoring devices, accelerometers, or medical imaging modalities such as functional magnetic resonance imaging (fMRI) are now commonly represented through their empirical distributions, which serve as latent descriptions of the underlying physiological processes Ghosal et al. (2025), Ghosal & Matabuena (2024), Matabuena et al. (2024, 2022). In recent years, several regression frameworks have been proposed to represent predictors, responses, or both as probability density functions Ghosal et al. (2025), Ghosal & Matabuena (2024). Another research strand embeds probability distributions as random objects in metric spaces, for which dedicated statistical procedures have been developed (see, e.g. Lugosi & Matabuena (2024)). Despite this growing body of work, to the best of our knowledge, there is still no comprehensive analytical framework that handles moderate- to high-dimensional distributions both flexibly and robustly. The methods introduced here aim to fill that gap and provide a practical toolkit for distributional data analysis.

3 Mathematical models

In this section, we introduce our approach to learning probability distributions that evolve over time. Let $\mathcal{T} \subset \mathbb{R}$ and consider the family of probability density functions defined in $\mathcal{X} = \mathbb{R}^d$ that evolve in \mathcal{T} ,

$$\mathcal{D} := \{ f: (x, t) \in \mathcal{X} \times \mathcal{T} \mapsto f(x, t) \in \mathbb{R}_{\geq 0} \mid \int_{\mathcal{X}} f(x, t) dx = 1 \forall t \in \mathcal{T} \},$$

equipped with the norm

$$\|f\|_{L^1, \mathcal{T}} := \sup_{t \in \mathcal{T}} \|f(\cdot, t)\|_{L^1(\mathcal{X})}.$$

For practical analytical purposes, we restrict our attention to a subset of \mathcal{D} consisting of elements that vary smoothly with t , such as differentiable density functions.

A classical result (Wiener–Tauberian Theorem; [Wiener \(1932\)](#)) guarantees that the class of finite Gaussian mixtures,

$$\mathcal{M} = \left\{ x \mapsto \sum_{s=1}^K \alpha_s \mathcal{N}(x, m_s, \Sigma_s) \mid K \in \mathbb{N}, \alpha_s \geq 0, \sum_{s=1}^K \alpha_s = 1 \right\},$$

is dense in $(L^1(\mathcal{X}), \|\cdot\|_1)$. Consequently, for any $\varepsilon > 0$ and $f \in \mathcal{D}$ with some additional mild continuity/tightness assumptions on $t \mapsto f(\cdot, t)$ (see proposition 1), there exists

$$g(x, t) = \sum_{s=1}^K \alpha_s(t) \mathcal{N}(x \mid m_s, \Sigma_s) \in \mathcal{M},$$

with fixed means $\{m_s\}_{s=1}^K$ and covariances $\{\Sigma_s\}_{s=1}^K$ but time-varying weights $\{\alpha_s(t)\}_{s=1}^K$, such that

$$\sup_{t \in \mathcal{T}} \|f(\cdot, t) - g(\cdot, t)\|_{L^1(\mathcal{X})} < \varepsilon.$$

Only the local mixing weights $\alpha_s(t)$ vary in t , whereas m_s and Σ_s are shared globally. However, in practice we fix a moderate K for interpretability and select it by the application (cf. section 5). Let

$$\theta(t) = (\alpha_1(t), \dots, \alpha_K(t), m_1, \dots, m_K, \Sigma_1, \dots, \Sigma_K) \in \Delta^{K-1} \times \mathbb{R}^{Kd} \times \mathbb{R}^{\frac{Kd(d+1)}{2}} \subset \mathbb{R}^p$$

and thus $p = K \left(1 + d + \frac{d(d+1)}{2} \right)$. Assume the evolution of $\alpha(t) = (\alpha_1(t), \dots, \alpha_K(t))$ follows the Neural ODE

$$\dot{\alpha}(t) = f_\phi(\alpha(t), t), \quad \alpha(0) = \alpha_0,$$

where $f_\phi : \Delta^{K-1} \times [0, 1] \rightarrow \Delta^{K-1}$ is globally Lipschitz in α .

Denote the set of all valid parameters by

$$\Theta = \{(\alpha_s, m_s, \Sigma_s)_{s=1}^K \mid \alpha_s \in \Delta^{K-1}, m_s \in \mathbb{R}^d, \Sigma_s \succ 0\} \subset \mathbb{R}^p.$$

From an empirical standpoint, we observe data at the discrete time grid $\tau = \{t_0, \dots, t_m\} \subset \mathcal{T}$ ($|\tau| = m+1$). For each $t_i \in \tau$, we record $n_i \geq 1$ observations $X_{t_i, 1}, \dots, X_{t_i, n_i}$ drawn from the

target distribution $F(\cdot, t_i)$. Two regimes are relevant: (i) a *cross-sectional* setting in which the full collection $\{X_{t_i,j} : 0 \leq i \leq m, 1 \leq j \leq n_i\}$ is i.i.d.; (ii) a *longitudinal/time-series* setting in which temporal dependence may exist within or across the blocks indexed by t_i . In the second regime, maximum-likelihood estimation can be cumbersome and often lacks robustness to model misspecification Alquier & Gerber (2024). This motivates the minimum-MMD optimization approach introduced here.

3.0.1 Discrete-time MMD fitting

For each time $t_i \in \tau$, consider the empirical measure

$$F_{t_i, n_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{X_{t_i,j}}, \quad i = 0, \dots, m$$

where $\delta_{X_{t_i,j}}$ denotes the Dirac measure at the observation $X_{t_i,j}$. The discrete-time optimization minimizes the squared MMD between F_{t_i} and the kernel mean embedding of the Gaussian mixture density

$$f_i(x) = \sum_{s=1}^K w_s \mathcal{N}(x \mid m_s, \Sigma_s).$$

At each $t_i \in \tau$, we use a Gaussian kernel k_i such as (3), with $\sigma_i^2 \approx (\text{median}_{j \neq k} \|X_{t_i,j} - X_{t_i,k}\|)^2$ the median heuristic separately at each t_i . Thus,

$$\text{MMD}^2(F_{t_i, n_i}, Q) = \sum_{s=1}^K \sum_{r=1}^K w_s w_r I_{i,s,r} - \frac{2}{n_i} \sum_{s=1}^K \sum_{j=1}^{n_i} w_s J_{i,s,j} + \frac{1}{n_i^2} \sum_{j=1}^{n_i} \sum_{\ell=1}^{n_i} k_i(X_{t_i,j}, X_{t_i,\ell}).$$

The first two terms admit closed-form expressions:

$$I_{i,s,r} = \frac{(\sigma_i^2)^{d/2}}{\sqrt{\det(\Sigma_s + \Sigma_r + \sigma_i^2 \mathbf{Id})}} \exp\left(-\frac{1}{2}(m_s - m_r)^\top (\Sigma_s + \Sigma_r + \sigma_i^2 \mathbf{Id})^{-1}(m_s - m_r)\right),$$

$$J_{i,s,j} = \frac{(\sigma_i^2)^{d/2}}{\sqrt{\det(\Sigma_s + \sigma_i^2 \mathbf{Id})}} \exp\left(-\frac{1}{2}(X_{t_i,j} - m_s)^\top (\Sigma_s + \sigma_i^2 \mathbf{Id})^{-1}(X_{t_i,j} - m_s)\right),$$

while the last term is computed directly from the data. This MMD objective is minimized as follows.

Initialization. Run k -means clustering Jain (2010) on $\bigcup_{i,j} X_{t_i,j}$ to obtain the initial means $\{m_s\}_{s=1}^K$ as cluster centers; the initial covariances $\{\Sigma_s\}_{s=1}^K$ as the empirical covariances of each cluster; and the initial weights as $\{c_s/n\}_{s=1}^K$, where c_s is the number of points in the group s and $n = \sum_{i=0}^m n_i$.

Local update (E-step). For each t_i , update the weight vectors $\alpha_i = [\alpha_1(t_i), \dots, \alpha_K(t_i)]^\top \in \mathbb{R}^K$ as

$$\alpha_i = \underset{w \in \Delta^{K-1}}{\operatorname{argmin}} \{w^\top I_i w - 2w^\top J_i + \sum_{s=1}^K \lambda_s w_s^2\}, \quad (4)$$

where $I_i = (I_{i,s,r})_{s,r} \in \mathbb{R}^{K \times K}$; $J_i = (\sum_{j=1}^n J_{i,s,j})_s \in \mathbb{R}^K$; and $\lambda = (\lambda_s)_{s=1}^K \in \mathbb{R}^K$ are individual hyperparameters of ridge penalty.

Global update (M-step). Update m_s and Σ_s iteratively via (Adam) gradient descent on the objective.

Ridge penalty. We note that the loss function in (4) was equipped with an individual ridge penalty for each model weight, motivated by the favorable properties this type of penalization has shown in sparse variable selection Tucker et al. (2023), Bertsimas et al. (2020).

3.0.2 Continuous-Time Weight Evolution via Neural ODEs

Once all discrete-time local weights α_i have been fitted, the next step is to learn a continuous-time model for their evolution. We posit a Neural ODE

$$\frac{d\alpha(t)}{dt} = f_\phi(\alpha(t), t) \quad t \in \mathcal{T}, \quad (5)$$

which produces a trajectory $\alpha(t) \in \mathbb{R}^K$. We parameterize the field f_ϕ with a multilayer perceptron.

To ensure that $\alpha(t) \in \Delta^{K-1}$ for all $t \in \mathcal{T}$, i.e. satisfying $\sum_{s=1}^K \alpha_s(t) = 1$ and $\alpha_s(t) \geq 0$, we project

$$\alpha(t) \mapsto \alpha(t) = \frac{\alpha(t)}{\sum_{s=1}^K \alpha_s(t)}.$$

The ODE parameters ϕ are then optimized by minimizing the discrepancy between the integrated trajectory and the previously fitted weights α_i :

$$\mathcal{L}_{\text{NODE}}(\phi) = \sum_{i=0}^m \|\alpha(t_i; \phi) - \alpha_i\|_2^2 + \nu \|\phi\|_2^2, \quad (6)$$

where $\alpha(t_i)$ denotes the solution of (5) evaluated at time t_i , and $\nu \geq 0$ is a ridge hyperparameter.

Permutation symmetry. The permutation symmetry of Gaussian mixtures is solved in our workflow by the MMD fitting step. Because (m_s, Σ_s) are shared over time and kept fixed after the global fit, component labels are anchored: “component s ” denotes at every time (and, in the clinical study, across participants) the one centered at m_s with shape Σ_s . The NODE stage only evolves $\alpha_s(t)$, so trajectories cannot exchange labels, removing the permutation ambiguity without additional constraints.

Why MMD? (i) With a Gaussian kernel, the MMD between an empirical sample and a Gaussian mixture admits closed-form terms, delivering numerically stable and fast updates; (ii) the RKHS geometry gives well-posed objectives under characteristic kernels; and (iii) empirical studies show robustness to misspecification and adverse settings [Chérief-Abdellatif & Alquier \(2022\)](#), [Alquier & Gerber \(2024\)](#), [Gao et al. \(2021\)](#), [Alquier et al. \(2023\)](#). These properties make MMD a pragmatic choice for time-resolved density fitting.

Why a neural ODE? Our object of interest is the continuous-time distribution of CGM in free-living environments, where the measurements are irregular and not aligned between participants. Modeling the weight trajectories $\alpha(t)$ with an ODE yields a smooth and subject-invariant evolution that mitigates sensor noise and avoids grid alignment. Discrete sequence models [Wang & et al. \(2024\)](#), [Wu & et al. \(2023\)](#) are effective in forecasting raw traces but do not directly furnish the continuous-time distributional dynamics central to our aims.

3.1 Theory

We begin by introducing theoretical guarantees for our estimator of the model parameters based on the MMD before the smoothing step. Proofs are present in the last section.

First, we establish the universality of our model.

Proposition 1 (Universality). *Let $\{f(\cdot, t)\}_{t \in \mathcal{T}} \subset L^1(\mathbb{R}^d)$ be probability densities. Assume:*

1. *For every $\eta > 0$ there exists $R < \infty$ such that $\int_{\|x\| > R} f(x, t) dx < \eta$ for all $t \in \mathcal{T}$;*
2. $\lim_{|h| \rightarrow 0} \sup_{t \in \mathcal{T}} \|f(\cdot + h, t) - f(\cdot, t)\|_{L^1} = 0$.

Then, for every $\varepsilon > 0$ there exist $\sigma^2 > 0$ and a finite set of centres $\{\mu_s\}_{s=1}^K \subset \mathbb{R}^d$ such that, for each $t \in \mathcal{T}$, one can choose weights $\alpha(t) = (\alpha_1(t), \dots, \alpha_K(t)) \in \Delta^{K-1}$ with

$$\left\| f(\cdot, t) - \sum_{s=1}^K \alpha_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1} < \varepsilon,$$

where $\varphi_\sigma(x) = (2\pi\sigma^2)^{-d/2} \exp(-\|x\|^2/(2\sigma^2))$. If, in addition, $t \mapsto f(\cdot, t)$ is uniformly L^1 -continuous, the map $t \mapsto \alpha(t)$ can be chosen continuous.

Now we derive uniform convergence rates for the initial nonparametric fitting of the MMD model.

Theorem 1 (Uniform MMD Convergence Rate). *Let $\{F_t^\theta : t \in \mathcal{T}\}$ be a time-indexed family of mixture distributions on \mathbb{R}^d , each with K components parameterized by $\theta(t) \in \Theta \subset \mathbb{R}^p$. Assume that at each time t , we observe an i.i.d. sample of size n_t , and define $n_* = \min_{t \in \mathcal{T}} n_t$. Let F_{t, n_t}^θ denote the empirical estimator at time t based on the n_t samples of $\{F_t^\theta : t \in \mathcal{T}\}$, and define the minimum component separation at time t as*

$$\Delta(\theta(t)) = \min_{j \neq k} \|\theta_j(t) - \theta_k(t)\|.$$

Then:

1. Regular regime: *If there exists $\Delta > 0$ such that $\inf_{t \in \mathcal{T}} \Delta(\theta(t)) \geq \Delta$, then*

$$\sup_{t \in \mathcal{T}} \text{MMD}(F_{t, n_t}^\theta, F_t^\theta) = O_p(\Delta^{-2} n_*^{-1/2} \sqrt{\log n_*}).$$

2. Singular regime: *If there exists $t \in \mathcal{T}$ such that $\Delta(\theta(t)) = 0$, then*

$$\sup_{t \in \mathcal{T}} \text{MMD}(F_{t,n_t}^\theta, F_t^\theta) = O_p(n_*^{-1/4} \sqrt{\log n_*}).$$

In both regimes, the convergence rates are optimal up to logarithmic factors. An additional multiplicative constant of the order \sqrt{p} may appear depending on the dimension p of the parameter space.

Finally, we compare the computational cost of a “massive univariate” fitting procedure—where we independently fit at each time point $t_i \in \tau$ —with that of a global strategy that simultaneously optimizes the model at $m + 1$ time points. These results justify using the massive univariate fitting approach.

Proposition 2 (Comparison with Per-Time MMD Evaluation). *Let $\bar{n} = \max_i n_{t_i}$, and assume all $n_{t_i} \approx \bar{n}$, and that $n = \bar{n}$ in the joint formulation. Then:*

- **Per-time cost:**

$$\mathcal{O}((m+1)[K^2d^3 + \bar{n}Kd^2 + \bar{n}^2d])$$

- **Joint cost:**

$$\mathcal{O}(K^2d^3(m+1)^3 + \bar{n}Kd^2(m+1)^2 + \bar{n}^2d(m+1))$$

Remark 1 (Scope of statistical guarantees). *Our finite-sample rates quantify the error of the population minimizer and its empirical MMD counterpart in Step 1, under characteristic kernels and standard regularity. The alternating optimization and the subsequent ODE smoothing (Step 2) are used to enforce temporal coherence and interpretability; a full convergence analysis of EM-like routines and post-smoothing is beyond our scope. Thus, our rates refer to the oracle MMD estimator; the practical two-step procedure is designed so that smoothing acts as a stable refinement of a consistent first-stage estimate.*

Parameter estimation via MMD. Let Q_θ denote a K -component Gaussian mixture with parameter $\theta = (\alpha, m, \Sigma)$, and let

$$\theta_t^* \in \operatorname{argmin}_{\theta \in \Theta} \mathcal{R}_t(\theta) := \text{MMD}^2(P_t, Q_\theta), \quad \hat{\theta}_t \in \operatorname{argmin}_{\theta \in \Theta} \widehat{\mathcal{R}}_t(\theta) := \text{MMD}^2(\widehat{P}_t, Q_\theta).$$

We assume k is characteristic and the model is well-specified ($P_t = Q_{\theta_t^*}$) or misspecified with a unique population minimizer.

Theorem 2 (Rates for the MMD estimator of mixture parameters). *Fix t and suppose θ_t^* is identifiable up to label permutation, with Q_θ twice continuously differentiable in θ and the Jacobian $J_t := \nabla_\theta \mu_{Q_\theta}|_{\theta=\theta_t^*}$ having full column rank (regular regime). Then*

$$\|\hat{\theta}_t - \theta_t^*\| = O_p(n_t^{-1/2}),$$

and $\sqrt{n_t}(\hat{\theta}_t - \theta_t^*)$ asymptotically normal after an appropriate relabeling of components.

If, instead, θ_t^* is singular (e.g., at a component collision where J_t loses rank), and the first nonvanishing term in the local expansion of $\theta \mapsto \mu_{Q_\theta}$ around θ_t^* is quadratic, then

$$\|\hat{\theta}_t - \theta_t^*\| = O_p(n_t^{-1/4}).$$

The $n^{-1/4}$ rate is minimax optimal in the singularities of classical finite mixtures and extends to the MMD objective under the stated smoothness structure. Uniform statements on a finite grid τ follow by taking $n_* = \min_i n_{t_i}$ and applying a union bound.

Proof sketch. View $\widehat{\mathcal{R}}_t(\theta)$ as an M-estimation objective in a Hilbert space. In the regular regime, a second-order Taylor expansion of \mathcal{R}_t at θ_t^* yields strong local convexity with curvature $J_t^\top J_t$, while $\widehat{\mathcal{R}}_t - \mathcal{R}_t$ fluctuates at $n_t^{-1/2}$. Standard M-estimation theory then gives $\sqrt{n_t}$ -consistency and asymptotic normality. At singular points, J_t loses rank and the quadratic (Hessian) term may vanish in directions associated with collisions; the first nonzero term becomes quartic in the natural reparameterization of mixtures, leading to $n_t^{-1/4}$ rates (cf. classical finite-mixture singularity analyses). The same mechanism carries over because the MMD objective is a smooth functional of the mixture moments entering the Gaussian kernel integrals. \square

4 Simulations

To benchmark finite-sample performance against state-of-the-art methods, we conduct a simulation study. Unlike many competitors, our approach explicitly prioritizes interpretability through time-varying functions $\alpha_s(\cdot)$ for $s = 1, \dots, K$. The results show that this emphasis on interpretability does not compromise accuracy: our estimator closely approximates the underlying distribution function.

Let $\mathcal{T} = [0, 1]$. We fix a target population distribution f at each $t \in \mathcal{T}$ given by a mixture of three Gaussian components whose means and common variance evolve linearly in time:

$$f(x, t) = \frac{1}{3} \sum_{s=1}^3 \mathcal{N}(x; m_s(t), \sigma^2(t) \mathbf{Id}),$$

where

$$\sigma^2(t) = 1 + t, \quad t \in \mathcal{T},$$

and

$$m_1(t) = -2 + 20 \cdot t, \quad m_2(t) = 16 \cdot t, \quad m_3(t) = 5 + 6 \cdot t. \quad (7)$$

For dimensions $d \geq 2$ each mean vector $m_s(t)$ has identical coordinates given by (7), and the covariance operator remains isotropic as has identical coordinates given by the expressions above, and the covariance remains isotropic $\Sigma(t) = \sigma^2(t) \mathbf{Id}$. This design enables the simultaneous analysis of both multimodal and unimodal regimes over time, capturing different distributional shapes within a single simulation study. For each setting we generate $B = 100$ independent replicates. Performance is evaluated at $m = 11$ time points $t_i \in \mathcal{T} = [0, 1]$ (equally spaced unless stated otherwise). At each time t_i we draw n_t observations, with the sample size chosen from $n_t \in \{20, 50, 100, 200, 300, 500\}$.

4.1 Competitors

We compare our estimator with three competitor models in a low dimensional scenario $d = 1$ and a large dimensional scenario $d = 10$. The methods are the GAMLSS model

Rigby & Stasinopoulos (2005a) (used only for dimension $d = 1$), a conditional-in-time KDE Tsybakov & Tsybakov (2009), Chacón & Duong (2018), and a Masked Autoregressive Flow Papamakarios et al. (2021).

Although these approaches can account for changes over time, they generally do not introduce a single, interpretable function for continuous “shifts” in the distribution (like our $\alpha(t)$) and may be less transparent for complex, high-dimensional data. Brief overviews of each competing method are given below; all implementations use the same early-stopping rule and hyperparameter search.

Normalizing flows We study normalizing flow as an autoregressive time-series model, following Papamakarios et al. (2017). Let $x = (x_1, \dots, x_I) \in \mathbb{R}^I$ and factorize its joint density as

$$F(x) = \prod_{i=1}^I F(x_i | x_{<i}), \quad x_{<i} \equiv (x_1, \dots, x_{i-1}),$$

where each conditional is Gaussian,

$$F(x_i | x_{<i}) = \mathcal{N}(x_i | \mu_i, \exp(2\alpha_i)), \quad \mu_i = f_{\mu,i}(x_{<i}), \alpha_i = f_{\alpha,i}(x_{<i}).$$

Draw independent noise $u_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and set

$$x_i = \mu_i + \exp(\alpha_i) u_i,$$

which defines an invertible transformation $x = f(u)$ with $u = (u_1, \dots, u_I) \sim \mathcal{N}(0, I)$. The inverse recursion

$$u_i = (x_i - \mu_i) \exp(-\alpha_i)$$

is strictly lower-triangular, giving

$$|\det \nabla_x f^{-1}(x)| = \exp\left(-\sum_{i=1}^I \alpha_i\right).$$

Substituting $u = f^{-1}(x)$ and the Jacobian determinant into the change-of-variables formula yields an exact, tractable log-likelihood, confirming that this autoregressive model is pre-

cisely a normalizing flow. The official implementation of the Masked Autoregressive Flow (MAF) proposed by [Papamakarios et al. \(2017\)](#) is available at [this GitHub repository](#).

GAMLSS (Generalized Additive Models for Location, Scale, and Shape)

GAMLSS [Rigby & Stasinopoulos \(2005b\)](#) extends standard generalized additive models by allowing multiple parameters of a chosen distribution family (e.g., location, scale, skewness, and kurtosis) to depend on covariates—including time. Thus, one may write

$$Y(t) \sim D(\mu(t), \sigma(t), \nu(t), \tau(t)),$$

where D is a parametric family and each parameter (e.g., $\mu(t)$) is modeled via a smooth function of t . Although GAMLSS accommodates smoothly time-varying distributions, it typically assumes a single fixed family and may not readily capture abrupt or highly non-linear shifts across multiple dimensions.

In our application, we specialize D to the Gaussian (Normal) distribution, so that

$$Y(t) \sim \mathcal{N}(\mu(t), \sigma^2(t)),$$

with both the time-dependent mean $\mu(t)$ and standard deviation $\sigma(t)$ modeled using spline smoothers. We employ the official `gamlss` R package, which supports this and a rich variety of other distributions (e.g., Box–Cox, Student’s t , and other skewed or heavy-tailed laws) and enables model terms that include linear effects, spline-based smoothers, random effects, or spatial components. We further leverage its diagnostic tools, model-selection criteria (GAIC, BIC), and visualization routines.

Kernel Density Estimators Kernel density estimators (KDEs), see [Silverman \(1986\)](#), are non-parametric methods that approximate probability densities by centering a smooth kernel (e.g., Gaussian) at each data point and averaging. KDEs naturally adapt to complex distributions without assuming a fixed parametric form. Temporal dependence can be introduced by conditioning on time—fitting a separate KDE at each observed time point

and linearly interpolating between the resulting density estimates. While conceptually simple and intuitive, this approach can become computationally expensive as the numbers of time points and observations grow. In our experiments, we fit a conditional KDE at each time instant of the observed series and then linearly interpolate in time to obtain smooth, time-varying density estimates.

4.2 Results on synthetic data: low dimensions

In dimension $d = 1$, fig. 1 shows that our model quantitatively competitive with respect to the alternatives. The NFLOW competitor may achieve slightly lower L^2 -error, reflecting its higher expressivity under strong regularity, but our goal is to remain competitive while providing interpretable time-varying distribution with clinically meaningful mixture weights. We therefore accept a small loss in expressivity in exchange for transparency.

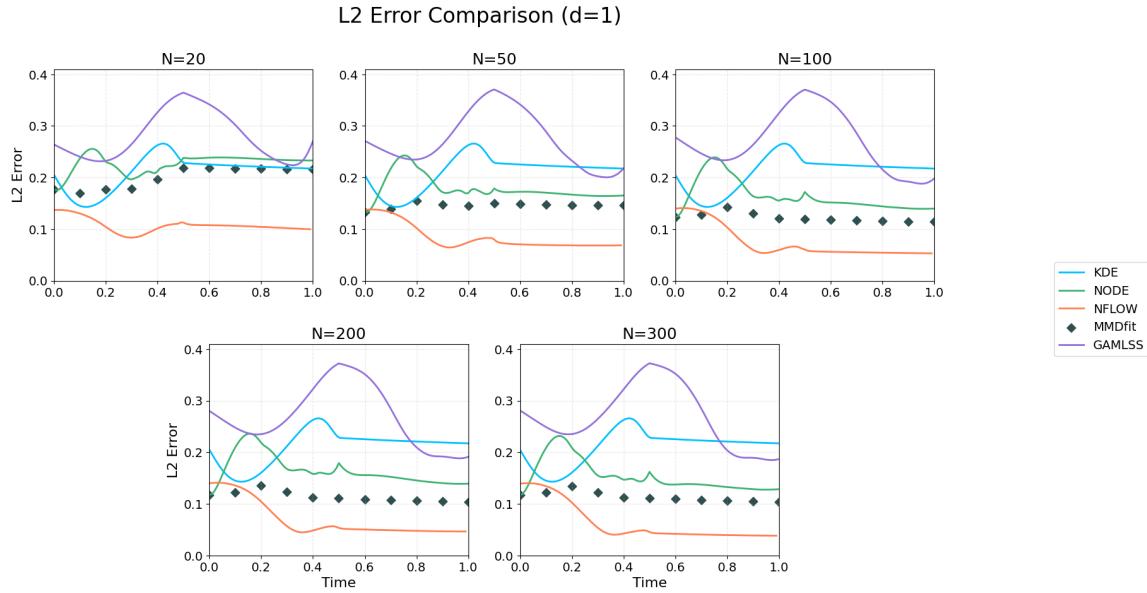


Figure 1: Pointwise L^2 -error over time in dimension $d = 1$. We compare NODE (our model), NFLOW, KDE and GAMLSS. Curves are averages over $B = 100$ seeds, measured against the ground truth of three Gaussians with time-dependent parameters. Errors for the discrete-time fitting are displayed as MMDFIT.

4.3 Results on synthetic data: high dimensions

In dimension $d = 10$, fig. 2 shows that our model outperforms the alternatives. In this setting, the method with the worst performance is KDE, which is widely recognized to be highly sensitive to dimensionality Tsybakov & Tsybakov (2009).

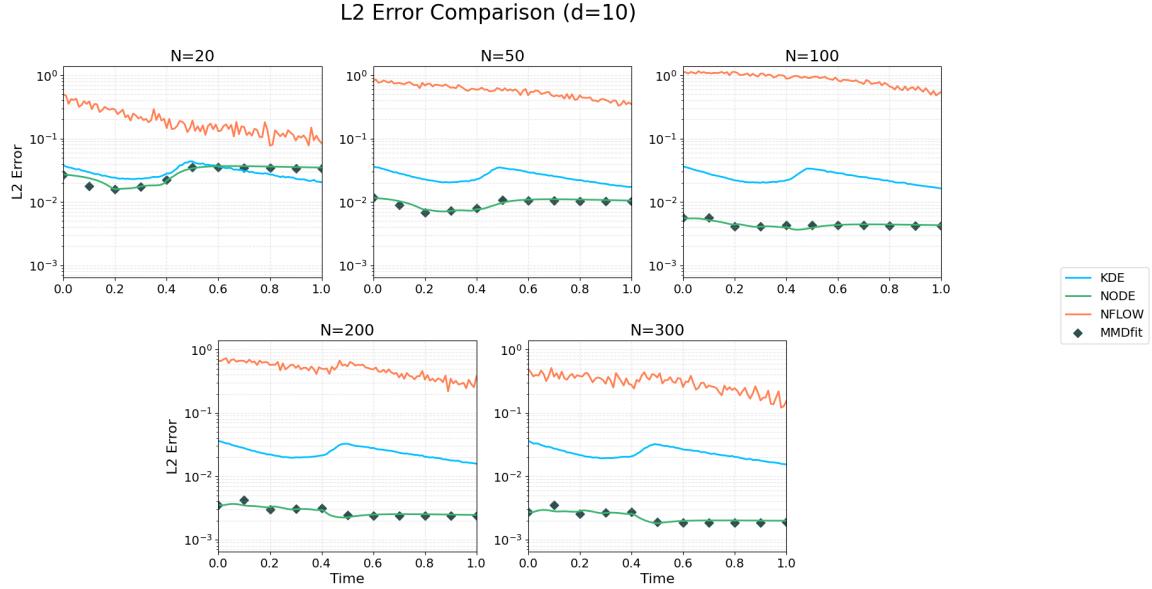


Figure 2: Pointwise L^2 -error over time. We compare NFLOW, KDE, and NODE (our model). Curves are averages over $B = 100$ simulations. Errors for the discrete-time fitting are displayed as MMDFIT.

5 Digital health diabetes case study

From a distributional data analysis perspective, this study aims to demonstrate the explainability advantages of our novel probabilistic framework for detecting distributional shifts in longitudinal clinical data.

5.1 Data description and scientific question

Our study is motivated by the Juvenile Diabetes Research Foundation (JDRF) Continuous Glucose Monitoring randomised controlled trial *Continuous Glucose Monitoring and*

Intensive Treatment of Type 1 Diabetes (2008), Group (2009).¹ A total of 451 participants (adults and children) with type 1 diabetes mellitus (T1DM) were randomised to either a *treatment* arm—receiving continuous glucose monitoring (CGM) data and education—or a *control* arm relying on standard self-monitoring without CGM feedback.

For data-quality reasons, we restrict our analysis to the 443 individuals with a maximum monitoring period of 26 weeks. Descriptive statistics are reported in the Supplementary Material. Our primary clinical question is whether continuous glucose monitor (CGM) use (*treatment*) versus usual care (*control*) leads to different individual longitudinal glucose trajectories over the course of the trial. The continuous-time nature of our model is ideally suited to address this question because it allows us to characterize how the entire glucose distribution evolves over the full follow-up period.

5.2 Modeling the marginal density of glucose time series

Let $Y_i(t)$ denote the CGM reading for participant i th at continuous time $t \in \mathcal{T}$. In practice, for a total number of n participants, we focus on a time-discrete process with weekly observations $t \in \mathcal{T}_i$. Within the interval $(t - 1, t]$ we collect n_{it} CGM readings. Our goal is to estimate the time-varying distribution function

$$F_i(x, t) = \mathbb{P}(Y_i(t) \leq x).$$

For computational and comparison purposes we assume that, for all $i = 1, \dots, n$, the observed data can be approximated enough well by the mixture density model

$$f_i(x, t) = \sum_{s=1}^K \alpha_{is}(t) \mathcal{N}(m_s, \Sigma_s),$$

where $\mathcal{N}(m_s, \Sigma_s)$ denotes a Gaussian distribution with mean m_s and scalar variance Σ_s , appropriate for modeling the univariate biomarker. To ensure comparability, we impose the constraint that the Gaussian parameters (m_s, Σ_s) of the mixture components are shared

¹Publicly available at <https://public.jaeb.org/datasets/diabetes>.

globally across all individuals; only the time-dependent mixing weights $\alpha_{is}(t)$ vary by participant.

Choice of the number of components. We set $K = 3$ to encode clinically meaningful phenotypes of glycaemic control (good, intermediate, poor). The same three components are used for all participants to enable cross-subject comparability; accordingly, $(m_s, \Sigma_s)_{s=1}^3$ are shared across individuals, whereas only the time-varying mixture weights $\alpha_{is}(t)$ are subject-specific. Interpreting the trajectories of $\alpha_{is}(t)$ within these three reference groups yields a clinically transparent summary of distributional dynamics.

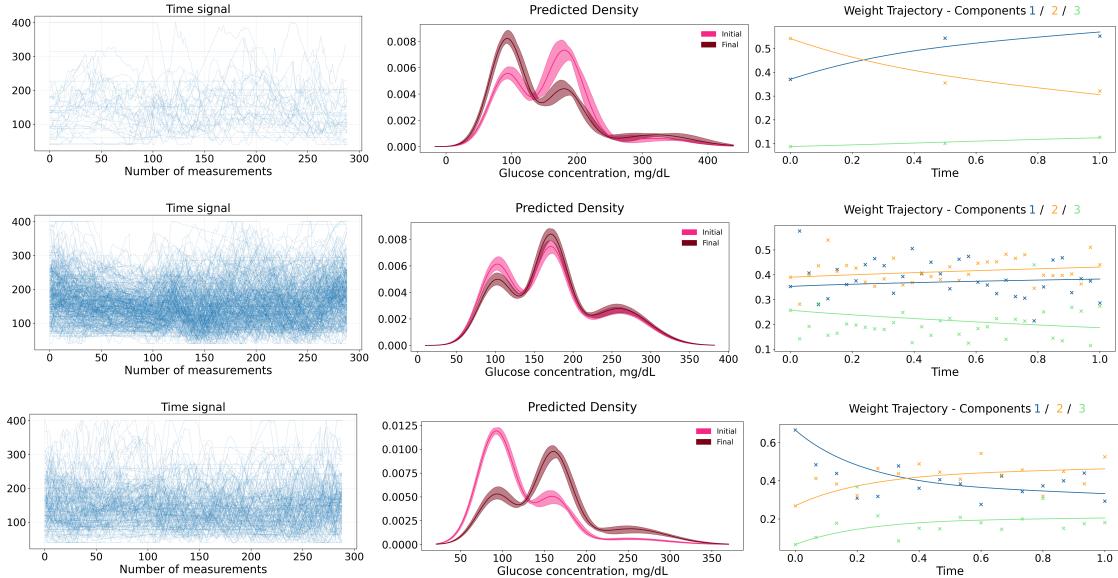


Figure 3: Three representative subjects (IDs 13, 62, and 377, one one each row). Left: the patient’s full CGM time series. Middle: initial (pink) and final (dark red) MMD-fitted densities with 95% bootstrap confidence bands. Right: weight trajectories $\alpha_s(t)$ from the neural ODE (Step 2); endpoints match the middle-column fits.

Table 1 reports the global mean and variance of each Gaussian component. From a clinical perspective, these parameters were selected to represent three clearly differentiated glycaemic states: (i) well-controlled diabetes condition, (ii) suboptimal diabetes control, and (iii) poorly controlled diabetes. By computing, at every time, the proportion of participants

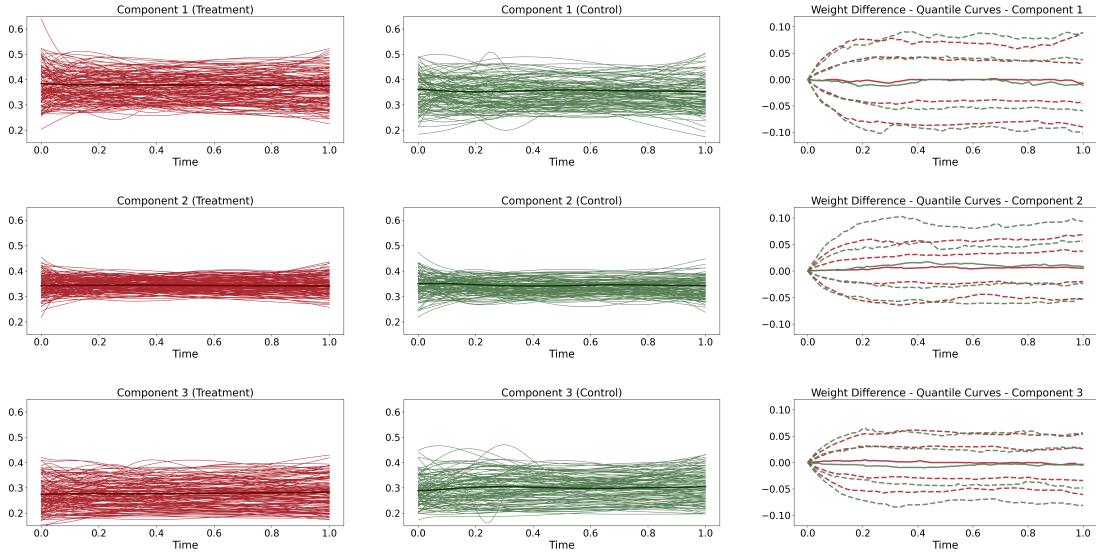


Figure 4: Component-wise weight dynamics by arm (components $s = 1, 2, 3$ in rows from top to bottom). Columns 1-2: individual curves $\alpha_s(t)$ (thin) and mean (thick) for treatment (left) and control (right) groups. Column 3: $(0.1, 0.25, 0.5, 0.75, 0.9)$ -quantile trajectories of the process Z_{is} defined in (8) across subjects for each of the two groups and by arm.

falling within these three archetypes [Cutler & Breiman \(1994\)](#) (or phenotypes), we obtain a dynamic, standardized index that summarizes the longitudinal evolution of glycaemic profiles for every subject in the trial.

Figure 3 shows the three representative day-to-day CGM trajectories together with their estimated marginal densities, highlighting the intrinsic complexity of the continuous-time series. In the left panel we observe substantial intra-day variability, suggesting that an analysis at the day level is advisable. To make subjects with different visit schedules comparable, time was rescaled to the unit interval $[0, 1]$. We then focus on the subject-specific weight functions $\alpha_{is}(t)$.

For the three selected participants, the first subject shows an improvement (density shifts left and weight α_{i1} increases), the second remains relatively stable, and the third worsens (density shifts right and weight α_{i1} decreases). Importantly, some patients deteriorate over time even in the treatment arm, consistent with previous findings that only a subset of

Table 1: Calibrated global statistics for univariate and bivariate data.

Univariate ($d = 1$): Fitted global means and variances			
Parameter	$s = 1$	$s = 2$	$s = 3$
μ_s	97.81	167.84	259.37
σ_s^2	537.72	501.04	1737.04
Bivariate ($d = 2$): Fitted global means and covariance matrices			
Parameter	$s = 1$	$s = 2$	$s = 3$
μ_s	(97.81, 0.58)	(167.66, 0.21)	(258.87, -0.059)
Σ_s	$\begin{bmatrix} 537.75 & -27.01 \\ -27.01 & 124.26 \end{bmatrix}$	$\begin{bmatrix} 501.01 & 1.41 \\ 1.41 & 63.62 \end{bmatrix}$	$\begin{bmatrix} 1737.04 & -18.21 \\ -18.21 & 208.75 \end{bmatrix}$

this cohort responds to the intervention. We also observe that certain participants exhibit apparent improvements in CGM metrics in the control group; however, these changes may reflect unrecorded external interventions—such as additional insulin dosing—rather than the effect of the study treatment. Moreover, despite randomization, baseline characteristics are not perfectly balanced between control and treatment groups (e.g. the treatment arm began with better glycaemic control) [Matabuena & Crainiceanu \(2024\)](#), [Group et al. \(2010\)](#).

Figure 4 contrasts the estimated mixture weights for the treatment and control arms. Participants in the treatment group devote a larger share of time to phenotype 1 across the entire follow-up, reflecting superior glycaemic control at all assessment points, including baseline. To investigate individual temporal dynamics, we examined the centered trajectories.

$$Z_{is}(t) = \alpha_{is}(t) - \alpha_{is}(0), \quad s = 1, 2, 3 \text{ and } i = 1 \dots, n. \quad (8)$$

For each arm, we plot the selected quantiles of the $Z_{is}(\cdot)$ process. For phenotype 1, the quantile curves are nearly identical—once good control is reached, further gains are limited.

For phenotype 3 (the worst glycaemic control), roughly 10% of control participants spend at least 5% of their time in this state. Because these participants start from a poorer baseline, they exhibit larger absolute reductions—improvements that are inherently easier to achieve than those seen in the treatment arm. We summarize the main results below.

- i) Roughly 10% of the treated participants increase their time in phenotype 1 within the first month.
- ii) Only a minority of patients, both improving and worsening, show time-varying changes in weights α_{is} ; improvements, when present, typically require at least two weeks and then stabilize or reverse after four months in the absence of intervention.
- iii) Continuous-time mixed-density models capture dynamic real-time fluctuations that static or discretized analyses miss.
- iv) Weight trajectories α_{is} provide a clinically interpretable metric to monitor individual response in CGM-based trials.

5.3 Bivariate analysis

We extend our univariate density analysis of clinical CGM data to the bivariate setting. The ten-day time windows considered here are identical to those in the main paper.

For each participant i and each ten-day interval t , we build the bivariate sample

$$(G_{it\ell}, S_{it\ell}), \quad \ell = 1, \dots, n_{it},$$

where $G_{it\ell}$ is the raw glucose reading, and

$$S_{it\ell} = \frac{G_{it\ell} - G_{it,(\ell-1)}}{\Delta t}$$

is its finite-difference rate of change (first temporal derivative), with Δt denoting the CGM sampling interval.

Bivariate mixture representation. Within each window, we model the joint density of (G, S) by the dynamic mixture

$$f_i(g, s; t) = \sum_{j=1}^K \alpha_{ij}(t) \mathcal{N}(\mu_j, \Sigma_j)(g, s),$$

where $\mu_j \in \mathbb{R}^2$ and $\Sigma_j \in \mathbb{R}^{2 \times 2}$ are shared across all participants, and the weight trajectories $\alpha_{ij}(t)$ capture subject-specific temporal dynamics. Again, we fix $F = 3$ and fit the global means $\mu_{s=1}^3$ and covariance matrices $\Sigma_{s=1}^3$ with the same three representative patients, see table 1.

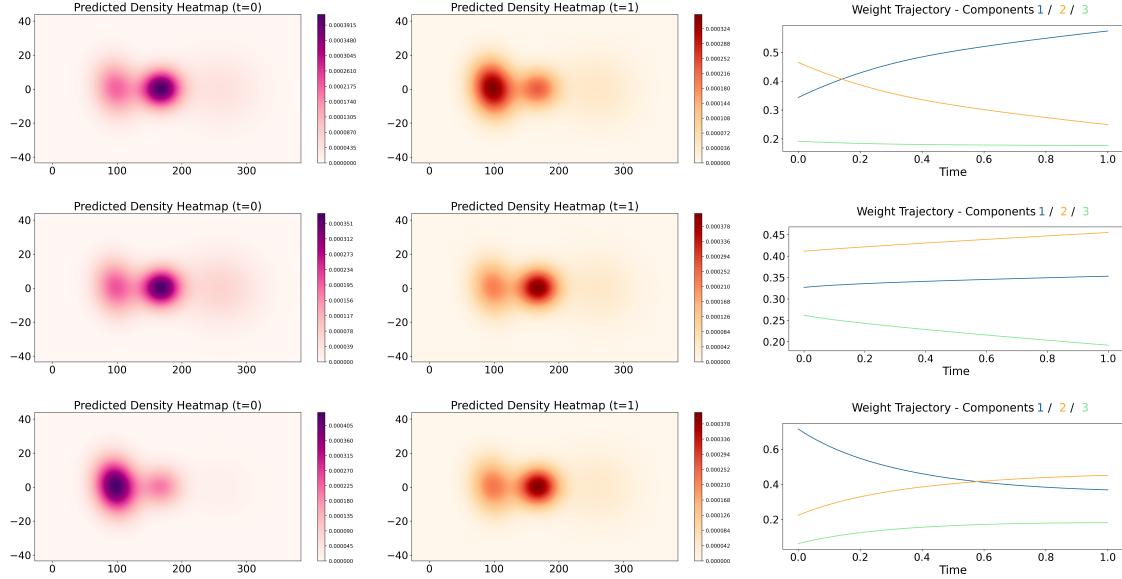


Figure 5: Three representative subjects (IDs 13, 62, and 377; one per row). Left: initial fitted bivariate density; Middle: final fitted bivariate density; Right: weight trajectories $\alpha_s(t)$ from the neural ODE model, whose endpoints match the middle-column fits.

Figure 6 compares the observed versus predicted values of Z_i ; points lie close to the identity line, underscoring the benefit of incorporating temporal dynamics via the mixture weights.

The interpretation of the process $Z_{ij}(t) = \alpha_{ij}(t) - \alpha_{ij}(0)$ and the resulting conclusions are analogous to those in the main article.

6 Discussion

We introduce an interpretable framework for estimating the dynamics of time-indexed probability distributions. In a comprehensive simulation study, the proposed method achieves a performance comparable to or exceeding state-of-the-art approaches, such as normalizing

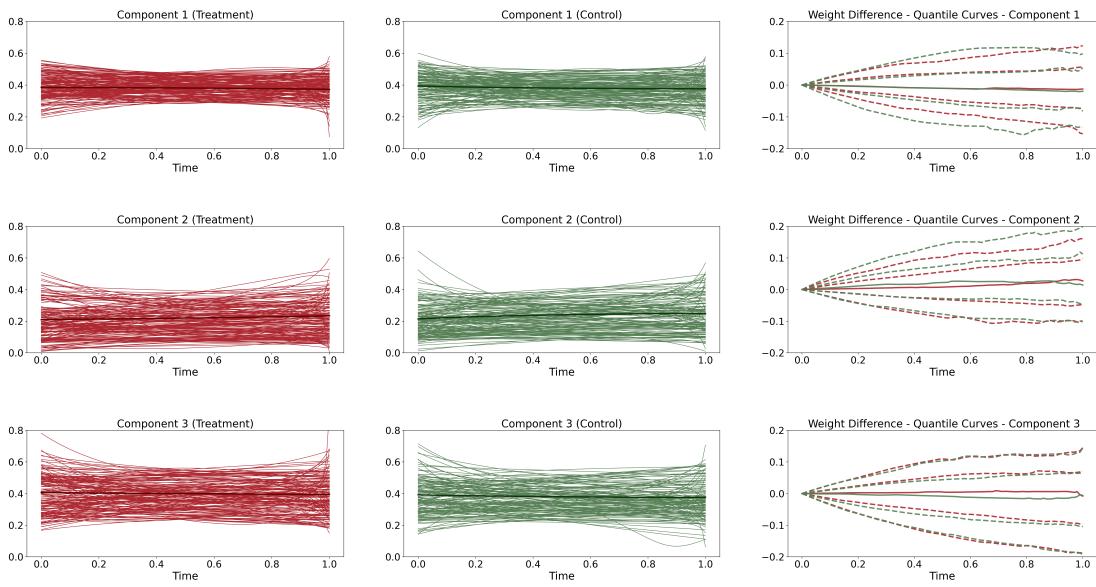


Figure 6: Component-wise weight dynamics on the full medical dataset after extending to a bivariate analysis by including the temporal derivative of glucose. Columns 1–2: individual curves $\alpha_s(t)$ (thin) and mean (thick) for treatment (left) and control (right). Column 3: $(0.1, 0.25, 0.5, 0.75, 0.9)$ -quantile trajectories of $Z_{is} = \alpha_{is}(t) - \alpha_{is}(0)$ across subjects.

flows, while offering faster computation and new insights into temporal evolution. Theoretical guarantees further substantiate these empirical findings. In a clinical case study using CGM data, our approach reveals temporal patterns that conventional statistical methods and standard machine learning techniques fail to capture.

As future work, we plan to: (i) develop distributed, online variants of the algorithm to enable large-scale, real-time analysis (e.g., motivated by the need to analyze massive cohorts as [All of Us Research Program Investigators \(2019\)](#)); (ii) integrate our dynamic density estimates into a model-free conformal prediction framework with finite-sample validity for metric-space-valued objects [Lugosi & Matabuena \(2024\)](#); (iii) extend the methods to vector-valued or functional biomarkers in richly multivariate longitudinal datasets (e.g., UCI studies with > 50 concurrent biomarkers); (iv) design new control systems driven by CGM insulin pumps based on this framework and Neural ODE control theory [E \(2017\)](#); and (v) generalize the models to time series whose units of interest lie in a separable Hilbert space [Dubey & Müller \(2020\)](#).

7 Computational Details

All experiments were executed on Ubuntu 20.04 LTS using eight NVIDIA GeForce RTX 3080 GPUs (10 GB each) on nodes with dual Intel Xeon Gold 6226R CPUs (2×32 cores @ 2.9 GHz), 384 GB RAM, and a 3.8 TB local SSD. The software environment was built with the following.

- Python 3.12.9 (Conda)
- CUDA 12.9 (Driver 575.51.03)
- PyTorch 2.2.0 (with `torchdiffeq` 0.2.3, `torchcde` 0.2.5)
- NumPy 1.26.4, SciPy 1.14.1, scikit-learn 1.6.1

8 Proofs

Proof of proposition 1. Fix $\varepsilon > 0$. By uniform tightness, choose $R < \infty$ such that

$$\int_{\|x\|>R} f(x, t) dx < \varepsilon/4 \text{ for all } t, \text{ and write } B_R = \{x \in \mathbb{R}^d : \|x\| \leq R\}.$$

Let φ_σ be a Gaussian mollifier. Using the approximate-identity representation and the change of variables $h = \sigma u$,

$$\begin{aligned} \|f(\cdot, t) - f(\cdot, t) * \varphi_\sigma\|_{L^1} &= \left\| \int_{\mathbb{R}^d} (f(\cdot, t) - f(\cdot - h, t)) \varphi_\sigma(h) dh \right\|_{L^1} \\ &\leq \int \varphi_1(u) \|f(\cdot, t) - f(\cdot - \sigma u, t)\|_{L^1} du. \end{aligned}$$

Taking \sup_t and using uniform translation-equicontinuity plus dominated convergence, we get $\sup_t \|f(\cdot, t) - f(\cdot, t) * \varphi_\sigma\|_{L^1} < \varepsilon/4$ for all σ small enough. Partition B_R into hypercubes $\{C_s\}_{s=1}^K$ of mesh size h and let μ_s be the centre of C_s . For each t , let

$$\tilde{\alpha}_s(t) = \int_{C_s} f(y, t) dy, \quad Z_t = \sum_{s=1}^K \tilde{\alpha}_s(t) = \int_{B_R} f(y, t) dy \in [1 - \varepsilon/4, 1],$$

and define $\alpha_s(t) = \tilde{\alpha}_s(t)/Z_t$, so that $\alpha(t) \in \Delta^{K-1}$. Writing $g_t = f(\cdot, t) * \varphi_\sigma$, we have

$$\begin{aligned} \left\| g_t - \sum_{s=1}^K \tilde{\alpha}_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1} &\leq \left\| \int_{B_R^c} f(y, t) \varphi_\sigma(\cdot - y) dy \right\|_{L^1} \\ &\quad + \left\| \sum_{s=1}^K \int_{C_s} f(y, t) (\varphi_\sigma(\cdot - y) - \varphi_\sigma(\cdot - \mu_s)) dy \right\|_{L^1}. \end{aligned}$$

The first term equals $1 - Z_t \leq \varepsilon/4$. Since φ_σ is Lipschitz in L^1 with $\|\varphi_\sigma(\cdot - y) - \varphi_\sigma(\cdot - \mu_s)\|_{L^1} \leq \|\nabla \varphi_\sigma\|_{L^1} \|y - \mu_s\| \leq C_d h/\sigma$, we obtain

$$\sup_t \left\| g_t - \sum_{s=1}^K \tilde{\alpha}_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1} \leq \frac{\varepsilon}{4} + C_d \frac{h}{\sigma} \sup_t \int_{B_R} f(y, t) dy \leq \frac{\varepsilon}{4} + C_d \frac{h}{\sigma}.$$

Moreover,

$$\left\| \sum_{s=1}^K \tilde{\alpha}_s(t) \varphi_\sigma(\cdot - \mu_s) - \sum_{s=1}^K \alpha_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1} = |Z_t - 1| \leq \varepsilon/4.$$

Finally, by the triangle inequality,

$$\left\| f(\cdot, t) - \sum_{s=1}^K \alpha_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1} \leq \underbrace{\|f - g_t\|_{L^1}}_{\leq \varepsilon/4} + \underbrace{\left\| g_t - \sum_{s=1}^K \tilde{\alpha}_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1}}_{\leq \varepsilon/4 + C_d h/\sigma} + \underbrace{\left\| \sum_{s=1}^K \tilde{\alpha}_s(t) \varphi_\sigma(\cdot - \mu_s) - \sum_{s=1}^K \alpha_s(t) \varphi_\sigma(\cdot - \mu_s) \right\|_{L^1}}_{\leq \varepsilon/4}.$$

Choosing $h \leq \sigma \varepsilon / (4C_d)$ gives the claim.

If, in addition, $t \mapsto f(\cdot, t)$ is uniformly L^1 -continuous, then $t \mapsto \tilde{\alpha}_s(t) = \int_{C_s} f(y, t) dy$ is continuous for each s . Since $Z_t = \sum_s \tilde{\alpha}_s(t) \in [1 - \varepsilon/4, 1]$ is continuous and bounded away from 0, the normalized weights $\alpha_s(t) = \tilde{\alpha}_s(t)/Z_t$ are continuous. Hence $t \mapsto \alpha(t)$ is continuous. \square

Proof of theorem 1. **Step 1: Pointwise concentration.** Since $k(x, \cdot) \in \mathcal{H}$ is bounded by 1, the centred feature map

$$Z_{t,i} = k(X_{t,i}, \cdot) - \mathbb{E}[k(X_{t,i}, \cdot)]$$

is sub-Gaussian in \mathcal{H} . By Pinelis' inequality

$$\|F_{t,n_t}^\theta - F_t^\theta\|_{\mathcal{H}} = O_p(n_t^{-1/2}).$$

In regime (S), local singularity of order $n_t^{-1/2}$ yields $O_p(n_t^{-1/4})$. For any bounded kernel with $\sup_x k(x, x) \leq K$ we have the sharp inequality $\text{MMD}(P, Q) \leq \sqrt{2K} \|P - Q\|_{\text{TV}}$, so minimax bounds in total variation automatically carry over to MMD. We must note that, when two mixture components become indistinguishable, the optimal rate in TV (and Hellinger) deteriorates to $n^{-1/4}$.

Step 2: Uniform control. Cover $[0, 1]$ by an n_*^{-1} -net of size $O(n_*)$. Lipschitz continuity of F_t^θ in \mathcal{H} implies a sub-Gaussian process indexed by this net. Applying Massart's maximal inequality and a Borell–TIS bound gives the uniform $O_p(\cdot)$ rates above. \square

Proof of proposition 2. We analyze the three terms in the MMD^2 decomposition separately, in the joint space $\mathbb{R}^{d(m+1)}$:

(1) Mixture–Mixture Term:

$$\sum_{s=1}^K \sum_{\ell=1}^K \alpha_s \alpha_\ell \int \int k(x, y) \mathcal{N}(x \mid m_s, \Sigma_s) \mathcal{N}(y \mid m_\ell, \Sigma_\ell) dx dy$$

This expression has a closed-form formula due to the product of Gaussians and the Gaussian kernel. Each pair (k, ℓ) requires:

- Computing a determinant and inverse of a $d(m+1) \times d(m+1)$ matrix, each of cost $\mathcal{O}((d(m+1))^3)$.
- Evaluating a quadratic form of cost $\mathcal{O}((d(m+1))^2)$, which is negligible compared to the matrix inverse.

There are K^2 such terms.

Thus, the total cost is:

$$\mathcal{O}(K^2(d(m+1))^3).$$

(2) Sample–Mixture Term:

$$\sum_{j=1}^n \sum_{s=1}^K \alpha_s \int k(X_j, y) \mathcal{N}(y | m_s, \Sigma_s) dy = \sum_{j,s} \alpha_s \mathcal{N}(X_j | m_s, \Sigma_s + \Gamma)$$

Each evaluation is a multivariate normal PDF at a single point:

- Cost of computing the density: $\mathcal{O}((d(m+1))^2)$.

There are nK such evaluations.

Thus, the total cost is:

$$\mathcal{O}(nK(d(m+1))^2).$$

(3) Sample–Sample Term:

$$\frac{1}{n^2} \sum_{j=1}^n \sum_{r=1}^n k(X_j, X_r)$$

Each kernel evaluation involves computing a squared Euclidean norm in $\mathbb{R}^{d(m+1)}$:

- Cost per evaluation: $\mathcal{O}(d(m+1))$

There are n^2 such terms.

Thus, the total cost is:

$$\mathcal{O}(n^2 d(m+1)).$$

Adding up the three costs yields the total complexity:

$$\mathcal{O}(K^2(d(m+1))^3 + nK(d(m+1))^2 + n^2d(m+1)).$$

□

References

- Alquier, P., Chérief-Abdellatif, B.-E., Derumigny, A. & Fermanian, J.-D. (2023), ‘Estimation of copulas via maximum mean discrepancy’, *JASA* **118**(543), 1997–2012.
- Alquier, P. & Gerber, M. (2024), ‘Universal robust regression via maximum mean discrepancy’, *Biometrika* **111**(1), 71–92.
- Álvarez-López, A., Geshkovski, B. & Ruiz-Balet, D. (2024), ‘Constructive approximate transport maps with normalizing flows’, *Applied Mathematics & Optimization*.
- Battelino, T., Alexander, C. M., Amiel, S. A., Arreaza-Rubin, G., Beck, R. W., Bergenstal, R. M., Buckingham, B. A., Carroll, J., Ceriello, A., Chow, E. et al. (2023), ‘Continuous glucose monitoring and metrics for clinical trials: an international consensus statement’, *The lancet Diabetes & endocrinology* **11**(1), 42–57.
- Bengio, Y., Goodfellow, I., Courville, A. et al. (2017), *Deep learning*, Vol. 1, MIT press Cambridge, MA, USA.
- Bertsimas, D., Pauphilet, J. & Van Parys, B. (2020), ‘Sparse regression’, *Statistical Science* **35**(4), 555–578.
- Chacón, J. E. & Duong, T. (2018), *Multivariate kernel smoothing and its applications*, CRC Press.
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J. & Duvenaud, D. K. (2018), Neural Ordinary Differential Equations, in ‘Advances in Neural Information Processing Systems’, Vol. 31, Curran Associates, Inc.
- Chérief-Abdellatif, B.-E. & Alquier, P. (2022), ‘Finite-sample properties of parametric mmd estimation: Robustness to misspecification and dependence’, *Bernoulli* **28**(1), 181–213.
- Continuous Glucose Monitoring and Intensive Treatment of Type 1 Diabetes* (2008), *New England Journal of Medicine* **359**(14), 1464–1476. PMID: 18779236.

- Cutler, A. & Breiman, L. (1994), ‘Archetypal analysis’, *Technometrics* **36**(4), 338–347.
- Dubey, P. & Müller, H.-G. (2020), ‘Functional models for time-varying random objects’, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **82**(2), 275–327.
- E, W. (2017), ‘A Proposal on Machine Learning via Dynamical Systems’, *Communications in Mathematics and Statistics* **5**(1), 1–11.
- Gao, R., Liu, F., Zhang, J., Han, B., Liu, T., Niu, G. & Sugiyama, M. (2021), Maximum mean discrepancy test is aware of adversarial attacks, in M. Meila & T. Zhang, eds, ‘Proceedings of the 38th International Conference on Machine Learning (ICML 2021)’, ML Research Press, pp. 3564–3575.
- Garreau, D., Jitkrittum, W. & Kanagawa, M. (2017), ‘Large sample analysis of the median heuristic’, *arXiv preprint arXiv:1707.07269*.
- Ghosal, R., Ghosh, S. K., Schrack, J. A. & Zipunnikov, V. (2025), ‘Distributional outcome regression via quantile functions and its application to modelling continuously monitored heart rate and physical activity’, *Journal of the American Statistical Association* (just-accepted), 1–20.
- Ghosal, R. & Matabuena, M. (2024), ‘Multivariate scalar on multidimensional distribution regression with application to modeling the association between physical activity and cognitive functions’, *Biometrical Journal* **66**(7), e202400042.
- Ghosal, R., Varma, V. R., Wolfson, D., Hillel, I., Urbanek, J., Hausdorff, J. M., Watts, A. & Zipunnikov, V. (2023), ‘Distributional data analysis via quantile functions and its application to modeling digital biomarkers of gait in alzheimer’s disease’, *Biostatistics* **24**(3), 539–561.
- Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012), ‘A kernel two-sample test’, *The Journal of Machine Learning Research* **13**(1), 723–773.

Group, J. D. R. F. C. G. M. S. (2009), ‘The effect of continuous glucose monitoring in

well-controlled type 1 diabetes’, *Diabetes care* **32**(8), 1378–1383.

Group, J. D. R. F. C. G. M. S. et al. (2010), ‘Effectiveness of continuous glucose monitoring

in a clinical care environment: evidence from the juvenile diabetes research foundation

continuous glucose monitoring (jdrf-cgm) trial’, *Diabetes Care* **33**(1), 17–22.

Haber, E. & Ruthotto, L. (2017), ‘Stable architectures for deep neural networks’, *Inverse*

Problems **34**(1), 014004.

Jain, A. K. (2010), ‘Data clustering: 50 years beyond k-means’, *Pattern recognition letters*

31(8), 651–666.

Jia, J. & Benson, A. R. (2019), Neural jump stochastic differential equations, *in* H. Wallach,

H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox & R. Garnett, eds, ‘Advances in

Neural Information Processing Systems’, Vol. 32, Curran Associates, Inc.

Katta, S., Parikh, H., Rudin, C. & Volfovsky, A. (2024), Interpretable causal inference

for analyzing wearable, sensor, and distributional data, *in* ‘International Conference on

Artificial Intelligence and Statistics’, PMLR, pp. 3340–3348.

Kidger, P., Morrill, J., Foster, J. & Lyons, T. (2020), Neural controlled differential equations

for irregular time series, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin,

eds, ‘Advances in Neural Information Processing Systems’, Vol. 33, Curran Associates,

Inc., pp. 6696–6707.

Kosorok, M. R. (2007), *Introduction to empirical processes and semiparametric inference*,

Springer Science & Business Media.

Lugosi, G. & Matabuena, M. (2024), ‘Uncertainty quantification in metric spaces’, *arXiv*

preprint arXiv:2405.05110.

Marzouk, Y., Ren, Z., Wang, S. & Zech, J. (2024), ‘Distribution learning via neural differ-

ential equations: a nonparametric statistical perspective’, *J. Mach. Learn. Res.* **25**(1).

Massaroli, S., Poli, M., Park, J., Yamashita, A. & Asama, H. (2020), Dissecting neural odes, in H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin, eds, ‘Advances in Neural Information Processing Systems’, Vol. 33, Curran Associates, Inc., pp. 3952–3963.

Matabuena, M. & Crainiceanu, C. M. (2024), ‘Multilevel functional distributional models with application to continuous glucose monitoring in diabetes clinical trials’, *arXiv preprint arXiv:2403.10514*.

Matabuena, M., Félix, P., Hammouri, Z. A. A., Mota, J. & del Pozo Cruz, B. (2022), ‘Physical activity phenotypes and mortality in older adults: a novel distributional data analysis of accelerometry in the nhanes’, *Aging Clinical and Experimental Research* **34**(12), 3107–3114.

Matabuena, M., Ghosal, R., Aguilar, J. E., Wagner, R., Merino, C. F., Castro, J. S., Zipunnikov, V., Onnela, J.-P. & Gude, F. (2024), ‘Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics’, *arXiv preprint arXiv:2410.00912*.

Matabuena, M. & Petersen, A. (2023), ‘Distributional data analysis of accelerometer data from the nhanes database using nonparametric survey regression models’, *Journal of the Royal Statistical Society Series C: Applied Statistics* **72**(2), 294–313.

Matabuena, M., Petersen, A., Vidal, J. C. & Gude, F. (2021a), ‘Glucodensities: A new representation of glucose profiles using distributional data analysis’, *Statistical methods in medical research* **30**(6), 1445–1464.

Matabuena, M., Petersen, A., Vidal, J. C. & Gude, F. (2021b), ‘Glucodensities: a new representation of glucose profiles using distributional data analysis’, *Statistical methods in medical research*.

Meskó, B. & Topol, E. J. (2023), ‘The imperative for regulatory oversight of large language models (or generative ai) in healthcare’, *NPJ digital medicine* **6**(1), 120.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B. et al. (2017), ‘Kernel mean

embedding of distributions: A review and beyond’, *Foundations and Trends® in Machine Learning* **10**(1-2), 1–141.

of Us Research Program Investigators, A. (2019), ‘The “all of us” research program’, *New England Journal of Medicine* **381**(7), 668–676.

Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S. & Lakshminarayanan, B. (2021), ‘Normalizing flows for probabilistic modeling and inference’, *Journal of Machine Learning Research* **22**(57), 1–64.

Papamakarios, G., Pavlakou, T. & Murray, I. (2017), Masked autoregressive flow for density estimation, in ‘Advances in Neural Information Processing Systems’, Vol. 30, Curran Associates, Inc.

Park, J., Kok, N. & Gaynanova, I. (2025), ‘Beyond fixed thresholds: optimizing summaries of wearable device data via piecewise linearization of quantile functions’, *arXiv preprint arXiv:2501.11777*.

Qian, Z., Zame, W., Fleuren, L., Elbers, P. & van der Schaar, M. (2021), ‘Integrating expert odes into neural odes: pharmacology and disease progression’, *Advances in Neural Information Processing Systems* **34**, 11364–11383.

Rigby, R. A. & Stasinopoulos, D. M. (2005a), ‘Generalized additive models for location, scale and shape’, *Journal of the Royal Statistical Society Series C: Applied Statistics* **54**(3), 507–554.

Rigby, R. A. & Stasinopoulos, D. M. (2005b), ‘Generalized additive models for location, scale and shape’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54**(3), 507–554.

Rubanova, Y., Chen, R. T. Q. & Duvenaud, D. K. (2019), Latent ordinary differential equations for irregularly-sampled time series, in ‘Advances in Neural Information Processing Systems’, Vol. 32, Curran Associates, Inc.

- Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. (2013), ‘Equivalence of distance-based and RKHS-based statistics in hypothesis testing’, *The Annals of Statistics* pp. 2263–2291.
- Serfling, R. J. (2009), *Approximation theorems of mathematical statistics*, John Wiley & Sons.
- Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Vol. 26, CRC press.
- Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, Routledge.
- Sriperumbudur, B. K., Fukumizu, K. & Lanckriet, G. R. (2011), ‘Universality, characteristic kernels and rkhs embedding of measures.’, *Journal of Machine Learning Research* **12**(7).
- Szabó, Z., Sriperumbudur, B. K., Póczos, B. & Gretton, A. (2016), ‘Learning theory for distribution regression’, *Journal of Machine Learning Research* **17**(152), 1–40.
- Tsybakov, A. B. & Tsybakov, A. B. (2009), ‘Nonparametric estimators’, *Introduction to Nonparametric Estimation* pp. 1–76.
- Tucker, D. C., Wu, Y. & Müller, H.-G. (2023), ‘Variable selection for global fréchet regression’, *Journal of the American Statistical Association* **118**(542), 1023–1037.
- Wang, S. & et al. (2024), Timemixer: Decomposable multiscale mixing for time series forecasting, *in ‘ICLR’*.
- Wiener, N. (1932), ‘Tauberian theorems’, *Annals of Mathematics* **33**(1), 1–100.
- Wu, H. & et al. (2023), Timesnet: Temporal 2d-variation modeling for general time series analysis, *in ‘ICLR’*.