UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ECONOMIA

MATHEUS VIZZOTTO DOS SANTOS

PREVENDO FUNÇÕES DE DENSIDADE DE PROBABILIDADE

Porto Alegre
2025

MATHEUS VIZZOTTO DOS SANTOS

PREVENDO FUNÇÕES DE DENSIDADE DE PROBABILIDADE

Projeto de dissertação apresentado ao Programa de Pós-Graduação em Economia da Faculdade de Ciências Econômicas da UFRGS, como requisito parcial para a obtenção do título de Mestre em Economia, com ênfase em Economia Aplicada.

Orientador: Prof. Dr. Flávio A. Ziegelmann

Coorientador: Prof. Dr. Eduardo de Oliveira Horta

Porto Alegre

2025

Ficha catalográfica

Resumo

A Análise de Dados Funcionais (Functional Data Analysis – FDA) tem emergido como um campo em rápida evolução, estendendo os métodos estatísticos clássicos para dados representados por funções. Nesse contexto, a análise de séries temporais também pode ser generalizada ao tratar cada observação como uma função, em vez de um escalar ou vetor. Este trabalho foca na previsão de uma classe específica de objetos funcionais: as funções densidade de probabilidade (pdfs). Um dos principais desafios nesse cenário decorre do fato de que as pdfs não formam um espaço vetorial, mas sim um subconjunto convexo de um, o que torna as técnicas padrão de séries temporais funcionais inaplicáveis. Para lidar com isso, exploramos uma abordagem de transformação que mapeia as pdfs para um espaço funcional mais apropriado, permitindo a aplicação de métodos existentes. A eficácia dessa abordagem é ilustrada por meio de uma aplicação em dados financeiros de alta frequência.

Palavras-chaves: Análise de dados funcionais. Séries temporais funcionais. Funções de densidade de probabilidade. Projeção. Expansão de Karhunen-Loève.

Abstract

Functional Data Analysis (FDA) has emerged as a rapidly evolving field, extending classical statistical methods to data represented by functions. In this context, time series analysis can also be generalized by treating each observation as a function rather than a scalar or vector. This work focuses on forecasting a specific class of functional objects: probability density functions (pdfs). A key challenge in this setting arises from the fact that pdfs do not form a vector space, but instead reside in a convex subset of one, rendering standard functional time series techniques inapplicable. To address this, we explore a transformation approach that maps pdfs into a more suitable functional space, enabling the application of existing methods. The effectiveness of this approach is illustrated through an application in high frequency financial data.

Keywords: Functional data analysis. Functional time series. Probability density functions. Forecasting. Karhunen-Loève expansion.

Sumário

1 Introduction

1.1 Density estimation

High-frequency trading (HFT) represents a significant evolution in modern financial markets, characterized by the execution of large volumes of trades at extremely high speeds, often within microseconds. In such an environment, the traditional assumptions of financial econometrics—such as normally distributed returns or independent and identically distributed (i.i.d.) increments—frequently break down. As a result, the *specification of the probability density function (PDF)* governing price changes or returns becomes a foundational aspect of modeling, forecasting, and executing trades in HFT systems.

PDF specification allows practitioners to move beyond point forecasts and assess the full range of possible future price realizations. This is particularly important in high-frequency contexts, where price changes are often small but frequent, and the tail behavior of the distribution can have outsized impacts on profitability and risk. Accurate modeling of PDFs aids in several key HFT tasks, including *order placement*, *market making*, *liquidity provision*, and *statistical arbitrage.*

Empirical studies have shown that the distributions of high-frequency returns exhibit *heavy tails*, *volatility clustering*, and *non-Gaussianity*, especially over very short horizons Cont (2001). Mischaracterizing these features can result in substantial model risk, leading to incorrect probability estimates and suboptimal execution. For instance, assuming normality in return distributions can underestimate the likelihood of large price swings, increasing exposure to adverse selection or sudden liquidity shocks.

Various models have been proposed to better capture the observed dynamics of high-frequency data. Nonparametric and semiparametric methods, such as *kernel* *density estimation* and *mixture models*, offer flexibility in modeling the empirical PDF without overly restrictive distributional assumptions Fan e Yao (2003). On the parametric side, models based on *generalized hyperbolic distributions* Prause (1999), *α-stable distributions* Nolan (2003), and *autoregressive conditional duration (ACD)* models Engle e Russell (1998) have been shown to provide better fits to high-frequency financial time series.

In practical HFT systems, the real-time estimation of these PDFs is often embedded in algorithmic decision engines. For example, *limit order placement algorithms* rely on an accurate forecast of the short-term price movement distribution to maximize expected fill rates while minimizing adverse selection risk Cartea, Jaimungal e Penalva (2015). Similarly, *market-making strategies* use estimates of the conditional PDF to dynamically adjust bid-ask spreads based on the predicted volatility and direction of price changes.

Thus, the specification of the probability density function is not merely a statistical exercise, but a key driver of performance in high-frequency trading. It directly informs the risk-reward trade-offs of algorithmic strategies and serves as a bridge between quantitative modeling and microstructural market dynamics.

1.2 Compositional Data Analysis

Compositional data (CoDa) are multivariate observations conveying relative information, typically represented as vectors with strictly positive components summing to a constant, usually one or 100% (AITCHISON, 1982). Such data arise naturally in diverse disciplines, including geology (e.g., mineral compositions), economics (e.g., market shares), biology (e.g., proportions of species in ecological samples), and medicine (e.g., time-use or microbiome data).

Classical multivariate statistical techniques often fail to appropriately handle the specific properties of compositional data due to the constant-sum constraint and the inherent relative scale of the data. As a consequence, applying standard techniques directly to raw compositional data can lead to misleading results (PAWLOWSKY-GLAHN; EGOZCUE; TOLOSANA-DELGADO, 2015).

Aitchison's pioneering work (AITCHISON, 1986) laid the foundation for the modern statistical treatment of compositional data. He introduced the use of log-ratio transformations, such as the centered log-ratio (clr), additive log-ratio (alr), and isometric log-ratio (ilr) transformations, to enable the application of standard statistical tools in an appropriate transformed space. These transformations map the data from the simplex (the sample space of compositions) to real Euclidean space, facilitating analysis while preserving the essential relative information.

The simplex, denoted as $\mathcal{S}^D = \left\{\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D_{>0} : \sum_{i=1}^D x_i = \kappa\right\}$, where κ is a positive constant (typically 1 or 100), serves as the sample space for compositional data (EGOZCUE et al., 2003). A key aspect of CoDa is the use of the Aitchison geometry on the simplex, which redefines operations such as perturbation (compositionally meaningful addition) and powering (compositionally meaningful scalar multiplication).

In recent years, CoDa methodology has seen significant advancements, particularly in its integration with functional data analysis, machine learning, and Bayesian inference (BOOGAART; TOLOSANA-DELGADO, 2013; GREENACRE, 2018). These developments have broadened the applicability of CoDa tools to more complex data structures, such as longitudinal compositional data or high-dimensional microbiome datasets.

This paper introduces the theoretical basis and methodological framework of CoDa, with emphasis on log-ratio transformations and the geometry of the simplex, providing a foundation for the application and development of compositional techniques in applied scientific research.

An important and growing extension of CoDa is its adaptation to probability density functions (PDFs), which share a key compositional property: they are non-negative and integrate to one. This extension is formally developed within the framework of *Bayes spaces*, where PDFs are treated as infinite-dimensional compositional objects (EGOZCUE; DÍAZ-BARRERO; PAWLOWSKY-GLAHN, 2006). The centered log-ratio (clr) transformation is generalized to functions, allowing PDFs to be analyzed in a Hilbert space endowed with the Aitchison geometry. For a density $f(x)$ defined on a compact support I, the clr transform is given by

$$\operatorname{clr}(f)(x) = \log(f(x)) - \frac{1}{|I|}\int_I \log(f(t))\,dt,$$

which maps f into a real-valued function with zero integral.

This perspective enables the application of functional principal component analysis (fPCA), clustering, and regression directly to probability densities, and has found applications in various domains including electricity demand modeling (DELICADO; EGOZCUE, 2011), Bayesian model assessment, and compositional inference in medical statistics (TALSKA et al., 2018).

This paper introduces the theoretical basis and methodological framework of CoDa, with emphasis on log-ratio transformations, the geometry of the simplex, and their extension to the analysis of probability density functions.

1.3 A glimpse into Functional Data Analysis

In the realm of statistics and data science, *Multivariate Data Analysis* (MDA) encompasses a collection of techniques designed to analyze data that arises from more than one variable. Unlike univariate or bivariate methods, MDA seeks to explore the structure and relationships that exist in datasets with multiple interdependent measurements, enabling a more comprehensive understanding of complex phenomena. These methods are particularly valuable in fields such as psychology, biology, finance, marketing, and social sciences, where multivariate observations are the norm rather than the exception.

The foundation of MDA lies in the recognition that many real-world processes are inherently multidimensional. For example, in market research, a consumer's preferences might be influenced by price, quality, brand perception, and peer opinion simultaneously. Analyzing each of these dimensions independently would obscure the relationships among them. MDA techniques, such as *Principal Component Analysis (PCA)*, *Factor Analysis*, *Discriminant Analysis*, *Cluster Analysis*, and *Canonical Correlation*, enable researchers to reduce dimensionality, classify observations, detect latent structures, and model the joint distribution of variables.

The formal development of MDA began in the early 20th century, parallel to advancements in linear algebra and matrix theory. One of the earliest and most influential contributions was *Principal Component Analysis* formulated by Pearson (1901), which aimed to reduce the dimensionality of a dataset while preserving as much variance as possible. Later, Hotelling (1936) introduced *Canonical Correlation Analysis*, establishing a framework to examine relationships between two sets of variables.

Another key milestone was the introduction of *Discriminant Analysis* by Fisher (1936), originally applied to distinguish between species of iris flowers using several morphological measurements. Fisher's Linear Discriminant Analysis (LDA) laid the groundwork for modern supervised classification techniques. These early methods were implemented manually or with the help of mechanical calculators, and only became widespread with the advent of digital computers in the mid-20th century.

By the 1960s and 1970s, multivariate analysis had become a core statistical tool, with widespread adoption in psychology, sociology, and economics. Seminal textbooks, such as Anderson (1958) and Tatsuoka (1971), codified the theory and practice of MDA. Software implementations began emerging in statistical packages like SAS, SPSS, and later R, which enabled more complex analyses to be conducted more efficiently and at scale.

Today, multivariate analysis is a foundational aspect of data science, with modern extensions including machine learning models, multivariate time series, and high-dimensional data exploration. The ability to extract meaning from multiple correlated variables remains crucial across disciplines, underscoring the enduring value and relevance of the early contributions to multivariate data analysis.

Functional Data Analysis (FDA) is a statistical framework for analyzing data that can be represented by functions, curves, or trajectories over a continuum such as time, space, or frequency. Unlike traditional multivariate analysis, which handles data as finite-dimensional vectors, FDA treats each observation as a function, often lying in an infinite-dimensional Hilbert space. This perspective is especially useful for studying processes that evolve continuously, such as temperature records, financial intraday returns, electroencephalogram (EEG) signals, or movement trajectories.

The emergence of FDA stems from the realization that many scientific and engineering datasets are best understood when their underlying smooth structure is preserved rather than discretized. For example, in biomedical sciences, a patient's heart rate over time is more naturally modeled as a curve rather than a collection of individual measurements. Functional data techniques provide tools for smoothing, registering, comparing, and modeling such curves Ramsay e Silverman (2005).

The foundational developments in FDA began in the late 20th century, notably with the pioneering work of Ramsay (1982), who introduced spline smoothing techniques for curve estimation. This was followed by a broader formalization of FDA as a distinct statistical discipline in the early 1990s and 2000s. In their influential texts, Ramsay e Silverman (2005) and Ferraty (2006) developed a unified theory that encompasses functional principal component analysis (FPCA), functional regression, and clustering of functional observations.

One of the earliest practical applications of FDA was in growth curve analysis, where children's height measurements taken at different ages were analyzed as smooth trajectories Ramsay e Dalzell (1991). Since then, FDA has seen widespread use in meteorology, chemometrics, biomechanics, and econometrics. Modern extensions now integrate FDA with machine learning, time series models, and high-dimensional statistics.

The core challenge in FDA lies in adapting classical statistical techniques to infinite-dimensional spaces. This requires tools from functional analysis, such as basis function expansions (e.g., splines, Fourier, wavelets), and the use of inner product structures for defining distances and covariances between functions. These methods enable dimension reduction (via FPCA), classification, hypothesis testing, and regression in the functional domain.

With the rise of high-frequency and longitudinal data in numerous fields, FDA continues to grow in relevance, offering a mathematically rich and practically effective framework for analyzing complex, smooth data structures.

Functional data analysis (FDA) is a branch of statistics that deals with data providing information about curves, surfaces or anything else varying over a continuum. In this context, a *functional time series* (FTS) is a sequence of random functions indexed by time. Each observation in the series is a function, typically lying in an infinite-dimensional function space, such as $L^2(\mathcal{T})$ for some compact interval $\mathcal{T} \subset \mathbb{R}$.

Let $\mathcal{H}$ be a separable Hilbert space, typically taken to be $\mathcal{H} = L^2(\mathcal{T})$, the space of square-integrable functions on a compact interval $\mathcal{T} \subset \mathbb{R}$, equipped with the inner product

$$\langle f, g \rangle = \int_{\mathcal{T}} f(t)g(t)\,dt,$$

and the associated norm $\|f\| = \sqrt{\langle f, f \rangle}$.

A *functional time series* is a sequence of $\mathcal{H}$-valued random variables $\{X_t\}_{t\in\mathbb{Z}}$, where each X_t is a random element of $\mathcal{H}$, i.e.,

$$X_t : \Omega \to \mathcal{H}, \quad t \in \mathbb{Z}.$$

A functional time series $\{X_t\}_{t\in\mathbb{Z}}$ is said to be **Mean-square continuous** if $\mathbb{E}\|X_t\|^2 < \infty$ for all t; **Second-order stationary** if the mean function $\mu(t) := \mathbb{E}[X_t]$ is constant over time and the autocovariance operator

$$\Gamma_h = \operatorname{Cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu) \otimes (X_t - \mu)]$$

depends only on the lag h, where $\otimes$ denotes the tensor product.

Functional Time Series (FTS) analysis is a modern statistical framework developed to handle data that are naturally viewed as functions observed over time. Unlike traditional time series models that deal with scalar or finite-dimensional vector observations, FTS methods treat each observation as a real-valued function, typically defined on a compact interval. This functional approach allows for the modeling of complex dynamic phenomena where each data point is an entire curve, such as daily temperature curves, intraday financial returns, or spectrometric curves.

A foundational treatment of linear models for functional data was provided by Bosq (2000), who developed autoregressive models in a Hilbert space setting, laying the groundwork for many later developments in the field. His approach enabled the extension of classical time series concepts like stationarity and autocorrelation to the infinite-dimensional setting.

Subsequent research has addressed various aspects of FTS, such as model assessment and estimation. Hall e Vial (2006) introduced diagnostic tools and inference procedures for assessing the adequacy of functional time series models, emphasizing the importance of model checking in high-dimensional settings. Their work underscored the challenges posed by the infinite-dimensional nature of the data and proposed practical solutions for effective model validation.

Bathia, Yao e Ziegelmann (2010) focused on identifying and estimating finite-dimensional dynamic structures in functional time series, addressing the issue of dimensionality reduction while preserving temporal dependence. Their methodology enables the recovery of dynamic factors that drive the functional observations over time, thus facilitating interpretable modeling and forecasting.

Aue, Norinho e Hörmann (2015) advanced the field further by developing predictive methods for FTS, including linear prediction theory and associated estimation techniques. Their work provided both theoretical guarantees and practical algorithms for functional time series forecasting, which is a central goal in many applications.

Together, these contributions form a robust theoretical and methodological foundation for the analysis and prediction of functional data observed over time, making Functional Time Series a vibrant and evolving area of research in modern statistics.

2 Goal

The main goal of this work is to assess the viability of forecasting functional time series that inherently carry relative information. Specifically, the functional observations are probability density functions, which are subject to the following constraints by definition:

Proposition: Let $f_X(x)$ denote the density function of a continuous random variable X, defined on the probability space $(\Omega, \mathbb{F}, \mathbb{P})$. Then $f_X(x)$ satisfies

1. $f_X(x) \geq 0$, for all x in $\mathbb{R}$
2. $\int_{-\infty}^{\infty} f(w)dw = 1$

Namely, the objectives are:

1. Perform a data analysis step of the time series subject to the study;
2. Evaluate the decomposition of objects;
3. Find the best time series model fitting to the principal component scores;
4. Obtain a set of forecast values for the probability density functions;
5. Compare the accuracy of the model with other state-of-the-art methods.

3 Literature Review

Aitchison Geometry

Let S^D denote the *D-part simplex*, defined as:

$$S^D = \left\{ \mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D : x_i > 0 \text{ for all } i, \ \sum_{i=1}^{D} x_i = 1 \right\}.$$

The **Aitchison geometry** on S^D is a vector space structure defined by the following components:

- **Perturbation (Composition Addition)**: For $\mathbf{x}, \mathbf{y} \in S^D$, their perturbation is defined as:

  $$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D),$$

  where $\mathcal{C}$ is the closure operator:

  $$\mathcal{C}(\mathbf{z}) = \left(\frac{z_1}{\sum_{i=1}^{D} z_i}, \ldots, \frac{z_D}{\sum_{i=1}^{D} z_i} \right).$$

- **Powering (Scalar Multiplication)**: For $\alpha \in \mathbb{R}$ and $\mathbf{x} \in S^D$,

  $$\alpha \odot \mathbf{x} = \mathcal{C}(x_1^{\alpha}, \ldots, x_D^{\alpha}).$$

- **Aitchison Inner Product**: For $\mathbf{x}, \mathbf{y} \in S^D$, define

  $$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log\left(\frac{x_i}{x_j}\right) \log\left(\frac{y_i}{y_j}\right).$$

- **Aitchison Norm**: The norm induced by the inner product is:

  $$\|\mathbf{x}\|_A = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_A}.$$

- **Aitchison Distance**: The distance between two compositions $\mathbf{x}, \mathbf{y} \in S^D$ is:

  $$d_A(\mathbf{x}, \mathbf{y}) = \|\mathrm{clr}(\mathbf{x}) - \mathrm{clr}(\mathbf{y})\|_2,$$

where $\mathrm{clr}(\mathbf{x})$ is the *centered log-ratio transformation*:

$$\mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{g(\mathbf{x})}, \ldots, \log \frac{x_D}{g(\mathbf{x})}\right), \quad g(\mathbf{x}) = \left(\prod_{i=1}^{D} x_i\right)^{1/D}.$$

This structure turns S^D into a real Hilbert space under the operations $\oplus$, $\odot$ and the inner product $\langle \cdot, \cdot \rangle_A$.

Aue, Norinho e Hörmann (2015),Bathia, Yao e Ziegelmann (2010), Benko, Härdle e Kneip (2009), Besse e Ramsay (1986), Bosq (2000), Dabo-Niang et al. (2008), Dauxois, Pousse e Romain (1982), Ferraty e Vieu (2003), Hall e Vial (2006), Ferraty (2006), Horta e Ziegelmann (2018), Hron et al. (2016), Müller, Dacorogna e Pictet (1998), Petersen e Müller (2016), Ramsay e Dalzell (1991), Ramsay e Silverman (2002), Ramsay e Silverman (2005), Ramsay, Hooker e Graves (2009)

4 Framework

First, we might define some useful concepts used in most work about functional time series.

Definition 1. *Let $X(t)$, $t \in \mathcal{T} \subset \mathbb{R}$, be a square-integrable stochastic process with mean function $\mu(t) = \mathbb{E}[X(t)]$ and covariance function*

$$C(s,t) = Cov(X(s), X(t)) = \mathbb{E}[(X(s) - \mu(s))(X(t) - \mu(t))]. \tag{1}$$

Then, if $C(s,t)$ is continuous and positive semi-definite, the Karhunen–Loève Expansion of $X(t)$ is given by

$$X(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k \phi_k(t), \tag{2}$$

where $\{\phi_k(t)\}_{k=1}^{\infty}$ are the orthonormal eigenfunctions of the covariance operator associated with $C(s,t)$; $\{\xi_k\}_{k=1}^{\infty}$ are uncorrelated random variables with zero mean and variances equal to the corresponding eigenvalues λ_k; and $\mathbb{E}[\xi_k \xi_j] = \lambda_k \delta_{kj}$, with δ_{kj} being the Kronecker delta.

Definition 2. *Let $\mathcal{H}$ be a separable Hilbert space taken to be $\mathcal{H} = L^2(\mathcal{T})$, that is, the space of square-integrable functions on a compact interval $\mathcal{T} \subset \mathbb{R}$, equipped with the inner product*

$$\langle f, g \rangle = \int_{\mathcal{T}} f(t) g(t) \, dt, \tag{3}$$

and the associated norm $\|f\| = \sqrt{\langle f, f \rangle}$.

A functional time series *is a sequence of $\mathcal{H}$-valued random variables $\{X_t\}_{t \in \mathbb{Z}}$, where each X_t is a random element of $\mathcal{H}$, i.e., $X_t : \Omega \to \mathcal{H}, t \in \mathbb{Z}$.*

If we consider an observed functional time series object Y_t, we define

$$Y_t(u) = X_t(u) + \varepsilon_t(u), \quad u \in \mathcal{T}, \quad t = 1, \ldots, n, \tag{4}$$

where the noise term $\varepsilon_t(u)$ is originated from experimental error and numerical rounding in discrete data treatment.

Now, we may ask ourselves how to deal with this type of data. In Bosq (2000), we can find a *functional autoregressive* (FAR) approach for time series forecasting, and this has long been the main method used in research because of the lack of other techniques. Nevertheless, the work of Aue, Norinho e Hörmann (2015) proposes a simplification of functional time series prediction by reducing it to a multivariate forecasting problem, thereby allowing the use of well-established tools, in contrast with the methodology of the FAR(p) model. The proposed algorithm consists of three steps: first, a number d of principal components is selected to retain $(\alpha \cdot 100)\%$ of the variance of the original data; then, given a forecast horizon h, a VAR(p) model is fitted to the principal components, and an h-step-ahead forecast is computed; finally, the multivariate forecasts are transformed back to the original functional space via a truncated Karhunen–Loève representation. It is also shown that the one-step-ahead forecast from a VAR(1) model in the second step is asymptotically equivalent to that of a FAR(1) model, which simplifies the forecasting task. Another important contribution of the paper is the proposal of a fully automatic and joint procedure for selecting the model order p and the number of components d through the minimization of a functional final prediction error (fFPE) criterion given by

$$fFPE(p,d) = \frac{n+pd}{n-pd}\mathrm{tr}(\hat{\Sigma}_Z) + \sum_{l>d} \hat{\lambda}_l, \tag{5}$$

which makes the proposed methodology entirely data-driven. The possibility of including exogenous variables in the model is also supported without major theoretical complications. Finally, simulation studies and applications to real data compare the performance of the new methodology with that of Hyndman e Ullah (2007), which carries out forecasting by treating the principal component scores as univariate time series, and Bosq (2000), using the autoregressive order selection criterion proposed by Kokoszka e Reimherr (2013). In both settings, the new method outperformed the alternatives. We can therefore conclude that this is a useful solution for the problem at hand.

Bathia, Yao e Ziegelmann (2010) propose a way to identify the dimensionality of these objects while modeling the serial dependence of the time series.

But when we're dealing with probability functions, we cannot use standard tools since the space they lie in is not a vector space. To overcome this, Hron et al. (2016) proposed a transformation into a Bayes space $\mathcal{B}^2$ of functional compositions.

Petersen e Müller (2016), also considering the inherent constraints of densities, thought of a mapping into into a Hilbert space through a continuous and invertible map.

Theorem 1. *This is the first theorem.*

Lemma 1. *This is a lemma that follows the theorem.*

Definition 3. *This is a definition related to the previous results.*

Definition 4. *In the theory of random processes, a sequence $\{X_n\}_{n=1}^{\infty}$ is said to be ψ-mixing if the dependence between past and future events decreases as they become further apart in time, according to a specific mixing coefficient.*

*Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$. The sequence is called ψ-*mixing *if there exists a function $\psi(n)$ such that for any two σ-algebras $\mathcal{F}_a^b = \sigma(X_a, X_{a+1}, \ldots, X_b)$ and $\mathcal{F}_c^d = \sigma(X_c, X_{c+1}, \ldots, X_d)$ with $a \leq b < c \leq d$, the following holds:*

$$\psi(n) = \sup_{A \in \mathcal{F}_1^k, B \in \mathcal{F}_{k+n}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

where $\psi(n) \to 0$ as $n \to \infty$.

The sequence is said to be ψ-mixing if $\psi(n) \to 0$ as $n \to \infty$. This condition implies that the events in the distant past and the far future become asymptotically independent.

Definition 5. *Let $\mathcal{X}$ be a domain and $h_0(x)$ a reference probability density function on $\mathcal{X}$. The* Bayes space *$B^2(\mathcal{X}, h_0)$ is defined as the space of all functions $h(x) > 0$ such that:*

$$\log \frac{h(x)}{h_0(x)} \in L^2(\mathcal{X}),$$

where $L^2(\mathcal{X})$ denotes the space of square-integrable functions on $\mathcal{X}$. The inner product between two elements $h_1(x), h_2(x) \in B^2(\mathcal{X}, h_0)$ is given by:

$$\langle h_1, h_2 \rangle_{B^2} = \int_{\mathcal{X}} \log \frac{h_1(x)}{h_0(x)} \log \frac{h_2(x)}{h_0(x)} h_0(x) dx.$$

The associated norm is:

$$\|h\|_{B^2} = \left(\int_{\mathcal{X}} \left(\log \frac{h(x)}{h_0(x)} \right)^2 h_0(x) dx \right)^{\frac{1}{2}}.$$

The Wasserstein metric, also known as the Earth Mover's Distance (EMD), is a distance function defined between probability distributions on a given metric space. It arises naturally in optimal transport theory, where the goal is to quantify the "cost"of transporting mass from one distribution to another.

Let $(\mathcal{X}, d)$ be a Polish metric space (i.e., a complete separable metric space), and let $\mathcal{P}_p(\mathcal{X})$ denote the space of Borel probability measures on $\mathcal{X}$ with finite p-th moment, defined as:

$$\mathcal{P}_p(\mathcal{X}) = \left\{ \mu \in \mathcal{P}(\mathcal{X}) \;\middle|\; \int_{\mathcal{X}} d(x_0, x)^p \, d\mu(x) < \infty \text{ for some } x_0 \in \mathcal{X} \right\}.$$

For two probability measures $\mu, \nu \in \mathcal{P}_p(\mathcal{X})$, the p-Wasserstein distance between them is defined as:

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\pi(x, y) \right)^{1/p},$$

where $\Pi(\mu, \nu)$ denotes the set of all couplings of μ and ν, i.e., all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ and ν.

The Wasserstein metric has gained significant attention in statistics, machine learning, and functional data analysis due to its meaningful geometric structure and robustness to small perturbations in distributions. In particular, it provides a powerful tool for comparing empirical distributions and studying convergence properties in probabilistic settings.

5 Data analysis

6 Results

Referências Bibliográficas

AITCHISON, J. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, Wiley, v. 44, n. 2, p. 139–160, 1982. Citado na página 7.

AITCHISON, J. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall, 1986. Citado na página 8.

ANDERSON, T. W. *An Introduction to Multivariate Statistical Analysis*. [S.l.]: Wiley, 1958. Citado na página 11.

AUE, A.; NORINHO, D. D.; HÖRMANN, S. On the prediction of stationary functional time series. *Journal of the American Statistical Association*, Taylor & Francis, v. 110, n. 509, p. 378–392, 2015. Citado 3 vezes nas páginas 14, 17 e 19.

BATHIA, N.; YAO, Q.; ZIEGELMANN, F. Identifying the finite dimensionality of curve time series. 2010. Citado 3 vezes nas páginas 14, 17 e 20.

BENKO, M.; HÄRDLE, W.; KNEIP, A. Common functional principal components. 2009. Citado na página 17.

BESSE, P.; RAMSAY, J. O. Principal components analysis of sampled functions. *Psychometrika*, Springer-Verlag, v. 51, n. 2, p. 285–311, 1986. Citado na página 17.

BOOGAART, K. G. Van den; TOLOSANA-DELGADO, R. Analyzing compositional data with r. *Springer*, 2013. Citado na página 8.

BOSQ, D. *Linear processes in function spaces: theory and applications*. [S.l.]: Springer Science & Business Media, 2000. v. 149. Citado 3 vezes nas páginas 13, 17 e 19.

CARTEA, ; JAIMUNGAL, S.; PENALVA, J. *Algorithmic and high-frequency trading*. [S.l.]: Cambridge University Press, 2015. Citado na página 7.

CONT, R. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, Taylor & Francis, v. 1, n. 2, p. 223–236, 2001. Citado na página 6.

DABO-NIANG, S. et al. Functional linear regression with functional response: application to prediction of electricity consumption. In: SPRINGER. *Functional and Operatorial Statistics*. [S.l.], 2008. p. 23–29. Citado na página 17.

DAUXOIS, J.; POUSSE, A.; ROMAIN, Y. Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *Journal of multivariate analysis*, Elsevier, v. 12, n. 1, p. 136–154, 1982. Citado na página 17.

DELICADO, P.; EGOZCUE, J. J. Compositional functional data analysis with application to time of use electricity data. *Statistical Modelling*, SAGE, v. 11, n. 6, p. 535–552, 2011. Citado na página 9.

EGOZCUE, J. J.; DÍAZ-BARRERO, J. L.; PAWLOWSKY-GLAHN, V. Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica, English Series*, Springer, v. 22, n. 4, p. 1175–1182, 2006. Citado na página 9.

EGOZCUE, J. J. et al. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, Springer, v. 35, n. 3, p. 279–300, 2003. Citado na página 8.

ENGLE, R. F.; RUSSELL, J. R. Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, Wiley Online Library, v. 66, n. 5, p. 1127–1162, 1998. Citado na página 7.

FAN, J.; YAO, Q. *Nonlinear time series: Nonparametric and parametric methods*. [S.l.]: Springer Science & Business Media, 2003. Citado na página 7.

FERRATY, F. *Nonparametric functional data analysis*. [S.l.]: Springer, 2006. Citado 2 vezes nas páginas 12 e 17.

FERRATY, F.; VIEU, P. Curves discrimination: a nonparametric functional approach. *Computational Statistics & Data Analysis*, Elsevier, v. 44, n. 1-2, p. 161–173, 2003. Citado na página 17.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, Wiley Online Library, v. 7, n. 2, p. 179–188, 1936. Citado na página 10.

GREENACRE, M. Compositional data analysis: recent advances. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley, v. 10, n. 5, p. e1441, 2018. Citado na página 8.

HALL, P.; VIAL, C. Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 68, n. 4, p. 689–705, 2006. Citado 2 vezes nas páginas 14 e 17.

HORTA, E.; ZIEGELMANN, F. Dynamics of financial returns densities: A functional approach applied to the bovespa intraday index. *International Journal of Forecasting*, Elsevier, v. 34, n. 1, p. 75–88, 2018. Citado na página 17.

HOTELLING, H. Relations between two sets of variates. *Biometrika*, Oxford University Press, v. 28, n. 3/4, p. 321–377, 1936. Citado na página 10.

HRON, K. et al. Simplicial principal component analysis for density functions in bayes spaces. *Computational Statistics & Data Analysis*, Elsevier, v. 94, p. 330–350, 2016. Citado 2 vezes nas páginas 17 e 20.

HYNDMAN, R. J.; ULLAH, M. S. Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, Elsevier, v. 51, n. 10, p. 4942–4956, 2007. Citado na página 19.

KOKOSZKA, P.; REIMHERR, M. Determining the order of the functional autoregressive model. *Journal of Time Series Analysis*, Wiley Online Library, v. 34, n. 1, p. 116–129, 2013. Citado na página 20.

MÜLLER, U. A.; DACOROGNA, M. M.; PICTET, O. V. Heavy tails in high-frequency financial data. *A practical guide to heavy tails: Statistical techniques and applications*, Boston, MA: Birkhäuser Boston, Inc, p. 55–78, 1998. Citado na página 17.

NOLAN, J. P. Modeling financial data with stable distributions. *Handbook of Heavy Tailed Distributions in Finance*, Elsevier, p. 105–130, 2003. Citado na página 7.

PAWLOWSKY-GLAHN, V.; EGOZCUE, J. J.; TOLOSANA-DELGADO, R. *Modelling and Analysis of Compositional Data*. [S.l.]: John Wiley & Sons, 2015. Citado na página 8.

PEARSON, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, Taylor & Francis, v. 2, n. 11, p. 559–572, 1901. Citado na página 10.

PETERSEN, A.; MÜLLER, H.-G. Functional data analysis for density functions by transformation to a hilbert space. 2016. Citado 2 vezes nas páginas 17 e 20.

PRAUSE, K. *The generalized hyperbolic model: Estimation, financial derivatives, and risk measures*. Tese (Doutorado) — University of Freiburg, 1999. Citado na página 7.

RAMSAY, J.; SILVERMAN, B. *Functional Data Analysis*. Springer, 2005. (Springer Series in Statistics). ISBN 9780387400808. Disponível em: ⟨https://books.google.com.br/books?id=mU3dop5wY_4C⟩. Citado 3 vezes nas páginas 11, 12 e 17.

RAMSAY, J. O. Fitting functional models to data. *Psychometrika*, Springer, v. 47, n. 3, p. 365–386, 1982. Citado na página 11.

RAMSAY, J. O.; DALZELL, C. Some tools for functional data analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Oxford University Press, v. 53, n. 3, p. 539–561, 1991. Citado 2 vezes nas páginas 12 e 17.

RAMSAY, J. O.; HOOKER, G.; GRAVES, S. *Functional Data Analysis with R and MATLAB*. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2009. ISBN 0387981845. Citado na página 17.

RAMSAY, J. O.; SILVERMAN, B. W. *Applied functional data analysis: methods and case studies*. [S.l.]: Springer, 2002. Citado na página 17.

TALSKA, R. et al. Principal component analysis for densities in bayes spaces. *Statistical Modelling*, SAGE, v. 18, n. 6, p. 512–534, 2018. Citado na página 9.

TATSUOKA, M. M. *Multivariate Analysis: Techniques for Educational and Psychological Research*. [S.l.]: Macmillan, 1971. Citado na página 11.

7 Apêndice

7.1 Figuras

7.2 Tabelas