DSSR

Discussion Paper No. 146

Density-valued ARMA models by spline mixtures

Yasumasa Matsuda and Rei Iwafuchi

June 23, 2025

Data Science and Service Research
Discussion Paper

Center for Data Science and Service Research
Graduate School of Economic and Management
Tohoku University
27-1 Kawauchi, Aobaku
Sendai 980-8576, JAPAN

Density-valued ARMA models by spline mixtures

Yasumasa Matsuda and Rei Iwafuchi*

June 23, 2025

Abstract

This paper proposes a novel framework for modeling time series of probability density functions by extending auto-regressive moving average (ARMA) models to density-valued data. The method is based on a transformation approach, wherein each density function on a compact domain $[0,1]^d$ is approximated by a B-spline mixture representation. Through generalized logit and softmax mappings, the space of density functions is transformed into an unconstrained Euclidean space, enabling the application of classical time series techniques. We define ARMA-type dynamics in the transformed space. Estimation is carried out via least squares for density-valued AR models and Whittle likelihood for ARMA models, with asymptotic normality derived under the joint divergence of the time horizon and basis dimension. The proposed methodology is applied to spatio-temporal human population data in Tokyo, where meaningful temporal structures in the distributional dynamics are successfully captured.

Keywords:B-spline basis; Density-valued time series; Logit function; Mixtures; Sieve MLE; Softmax function;
MOS subject classification: 62M10; 62M30.

1 Introduction

This paper focuses on modeling the dynamic behavior of density functions defined on the unit cube $I_d = [0,1]^d$. In many real-world applications, par-

*Graduate School of Economics and Management, Tohoku University

ticularly those involving two indexing dimensions such as time and space, it is desirable to analyze the variation of distributions over time. For example, spatio-temporal data may naturally be regarded as a time series of density functions. Figure 1 in Section 6 illustrates this concept using human population density in central Tokyo observed at three different time points, demonstrating the notion of density-valued time series.

The space of density functions on I_d is not a linear space in the traditional sense. Although it can be endowed with a metric structure (e.g., Hellinger distance), operations such as the addition or scalar multiplication of densities generally fall outside the space of densities, as they may violate key constraints:

$$\forall x \in I_d = [0,1]^d,\ f(x) \geq 0, \quad \int_{I_d} f(x)\, dx = 1.$$

Consequently, conventional techniques from functional data analysis cannot be directly applied to density functions. This motivates the development of alternative approaches that preserve the fundamental structure of densities while enabling dynamic modeling.

Several approaches to modeling density functions have been proposed in the literature. A comprehensive overview is provided by Petersen et al. (2022), who distinguish between transformation-based and object-oriented methods. Transformation-based methods involve mapping density functions into a Hilbert space, allowing for the application of linear tools. Notable examples include the log-hazard and log-quantile density transformations introduced by Petersen and Müller (2016). The key idea is to establish an explicit bijection between the space of densities and the representation space.

In contrast, object-oriented approaches employ geometric frameworks, such as those using the Wasserstein metric (Panaretos and Zemel, 2020) or Fisher–Rao metric (Srivastava et al., 2007), to endow the space of densities with a Riemannian manifold structure. This allows for defining models via tangent spaces associated with chosen metrics. Recent developments in this category include the Wasserstein autoregressive model proposed by Zhang

et al. (2022), which defines temporal dependence of densities through optimal transport maps and establishes a theoretical foundation for stationarity and inference in the Wasserstein space.

Our proposed methodology falls under the transformation-based framework. Specifically, we approximate a given density function f on I_d using a normalized J-term B-spline basis expansion:

$$f(s) = \sum_{j=1}^{J} \phi_j \tilde{B}_j(s), \quad s \in [0,1]^d, \quad \phi_j \geq 0, \quad \sum_{j=1}^{J} \phi_j = 1,$$

where $\tilde{B}_j(s)$ denotes the normalized j-th B-spline basis function. The vector of coefficients $\phi = (\phi_1, \ldots, \phi_J)$ lies in the standard probability simplex:

$$\mathscr{H}_J = \left\{ \phi \in \mathbb{R}^J \;\middle|\; \phi_j \geq 0,\ \sum_{j=1}^{J} \phi_j = 1 \right\}.$$

To handle potential numerical issues caused by boundary values (e.g., zeros), we extend this to a relaxed simplex:

$$\tilde{\mathscr{H}}_J = \left\{ \phi \in \mathbb{R}^J \;\middle|\; \phi_j \geq -1,\ \sum_{j=1}^{J} \phi_j = 1 \right\}.$$

We then construct a bijective mapping from $\tilde{\mathscr{H}}_J$ to $\mathbb{R}^{J-1}$ using generalized logit or softmax-like transformations. This representation space enables the use of conventional statistical modeling tools such as autoregressive moving average (ARMA) models. We refer to the resulting models as density-valued ARMA models.

In practice, density functions are rarely observed directly. Instead, we typically observe independent and identically distributed (i.i.d.) samples $s_1, \ldots, s_n \in I_d$ drawn from an unknown density. To estimate the B-spline coefficients ϕ_j from the sample, we adopt a sieve maximum likelihood estimation (MLE) framework (Chen, 2007). This can be interpreted as a form of nonparametric MLE over a sieve space. Following the general theory of

Shen and Wong (1994), we derive sufficient conditions under which our sieve MLE is consistent and attains optimal convergence rates.

Assuming observations $\{s_{it}\}_{i=1}^{n}$ for time points $t = 1, \ldots, T$, we estimate the coefficient vector $\phi_t \in \mathscr{H}_J$ at each time t using sieve MLE. We then fit ARMA models (Box et al., 1994) to the resulting sequence $\{\phi_t\}_{t=1}^{T}$ in the transformed representation space $\mathbb{R}^{J-1}$. Parameter estimation is conducted via standard methods such as least squares for AR models and Whittle likelihood for ARMA models, as detailed in the classical reference by Brockwell and Davis (1991). We also analyze the asymptotic behavior of the estimators under joint divergence of T and J (i.e., as both the number of time points and the number of basis functions increase).

The remainder of the paper is organized as follows. Section 2 introduces the B-spline mixture framework and the sieve MLE for density estimation. Section 3 develops the transformation between the B-spline coefficient space and the representation space. Section 4 proposes ARMA models in the transformed space. Section 5 provides an empirical analysis of human population density data. Section 6 concludes.

2 B-spline mixtures

2.1 B-spline basis expansions on $[0, 1]^d$

B-spline functions are piecewise polynomials that form a basis on $I_d = [0, 1]^d$. They can approximate functions in $L^p(I_d)$ for $p \geq 1$ under the L^p metric using piecewise polynomials. We briefly summarize the univariate case ($d = 1$) and extend it to the multivariate case, following the treatment by Schumaker (2007).

Let us define B-spline basis functions of order r. Let k be a positive integer and $t_1 < \cdots < t_k$ be distinct real numbers chosen such that the interval $[t_1, t_k]$ contains $[0, 1]$, with knots satisfying $t_{r+1} = 0$ and $t_{k-r} = 1$. A spline function on $I_1 = [0, 1]$ of order r with knots $0 = t_{r+1} < \cdots <$

$t_{k-r-1} < t_{k-r} = 1$ is expressed as

$$\sum_{i=1}^{k-r-1} \phi_i B_{i,r}(s), \quad s \in [0,1],$$

where the basis functions $B_{i,r}(s)$ are defined recursively (Theorem 4.15 in Schumaker (2007)). Specifically, initializing for $q = 1, \ldots, k-1$ by

$$B_{q,0}(s) = \begin{cases} 1, & t_q \le s < t_{q+1}, \\ 0, & \text{otherwise}, \end{cases}$$

we define for $p = 1, \ldots, r$ and $q = 1, \ldots, k-p-1$,

$$B_{q,p}(s) = \frac{s - t_q}{t_{q+p} - t_q} B_{q,p-1}(s) + \frac{t_{q+p+1} - s}{t_{q+p+1} - t_{q+1}} B_{q+1,p-1}(s).$$

Thus, the set $\{B_{j,r}(s)\}_{j=1}^{k-r-1}$ forms a B-spline basis of order r, consisting of piecewise polynomials of degree r that are $(r-1)$-times continuously differentiable on I_1. In other words, this basis spans a linear space of dimension $k-r-1$ determined by the knots $0 = t_{r+1} < \cdots < t_{k-r} = 1$.

We extend this construction to the multivariate case on $I_d = [0,1]^d$ via the tensor product. For each coordinate axis $i = 1, \ldots, d$, we fix knots $0 = t_{i,r+1} < \cdots < t_{i,k-r} = 1$, which may differ across dimensions, and construct the univariate B-spline bases $B_{j,r}^{(i)}(s_i)$. For $s = (s_1, \ldots, s_d) \in I_d$, define the tensor product basis by

$$B_{i_1,\ldots,i_d,r}(s) := \prod_{m=1}^{d} B_{i_m,r}^{(m)}(s_m).$$

A d-variate spline function of order r on I_d is then represented by

$$\sum_{i_1=1}^{k-r-1} \cdots \sum_{i_d=1}^{k-r-1} \phi_{i_1,\ldots,i_d} B_{i_1,\ldots,i_d,r}(s),$$

which we abbreviate by rearranging multi-indices into a single index $j =$

$1, \ldots, J$ with $J = (k - r - 1)^d$, as

$$\sum_{j=1}^{J} \phi_j B_{j,r}^d(s), \quad s \in [0,1]^d. \tag{1}$$

2.2 Maximum likelihood estimation for spline mixtures

We consider spline mixtures to model density functions on $I_d = [0,1]^d$. Utilizing non-negativity of B-spline basis functions, we employ mixtures of the normalized B-spline basis to express density functions in a nonparametric way.

Let $f_0(s)$ be an unknown density on $I_d = [0,1]^d$. Suppose we observe independent and identically distributed (iid) samples $s_1, s_2, \ldots, s_n \in I_d$ drawn from $f_0(s)$. Denote by $\tilde{B}_{j,r}^d(s), j = 1, \ldots, J$, the normalized B-spline basis functions of order r on I_d, satisfying

$$\int_{I_d} \tilde{B}_{j,r}^d(s) ds = 1, \quad j = 1, \ldots, J.$$

We approximate f_0 by a B-spline mixture,

$$f(\theta, s) = \sum_{j=1}^{J} \theta_j \tilde{B}_{j,r}^d(s), \tag{2}$$

where the coefficients satisfy $\theta_j \geq 0$ and $\sum_{j=1}^{J} \theta_j = 1$ to ensure $f(\theta, \cdot)$ is a valid density. The number of basis functions $J = J_n$ may depend on the sample size n and is assumed to increase as $n \to \infty$.

The log-likelihood function for the observed data is given by

$$\tilde{Q}_n(\theta) = \sum_{i=1}^{n} \log \left\{ \sum_{j=1}^{J} \theta_j \tilde{B}_{j,r}^d(s_i) \right\} - \lambda \left(\sum_{j=1}^{J} \theta_j - 1 \right), \tag{3}$$

where λ is the Lagrange multiplier enforcing the summation constraint on θ. Maximizing this likelihood yields the maximum likelihood estimator (MLE)

$\hat{\theta}_{ML}$, which satisfies the iterative update:

$$\theta_p^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \frac{\theta_p^{(m)} \tilde{B}_{p,r}^d(s_i)}{\sum_{j=1}^{J} \theta_j^{(m)} \tilde{B}_{j,r}^d(s_i)}, \quad p = 1, \ldots, J,$$

for iteration $m = 0, 1, 2, \ldots$ until convergence. This procedure is a special case of the Expectation-Maximization (EM) algorithm commonly used in mixture models (e.g., McLachlan and Peel, 2007).

2.3 Consistency

Our estimation problem fits into the sieve M-estimation framework described in Chen (2007). We consider i.i.d. observations $s_i, i = 1, \ldots, n$ from an unknown, but smooth density function $f_0 : [0,1]^d \to \mathbb{R}$. We estimate f_0 by maximizing the log-likelihood in (3) over a sieve space constructed using tensor-product B-spline basis functions on I_d introduced in (1) as $J = J_n \to \infty$.

Let $\{\mathcal{S}_n\}$ be the sequence of sieve spaces of B-splines defined as

$$\mathcal{S}_n = \left\{ \sum_{j=1}^{J_n} \phi_j \tilde{B}_{j,r}^d(s) : \phi_j \geq 0, \sum_{j=1}^{J_n} \phi_j = 1 \right\}, \tag{4}$$

where $\tilde{B}_{j,r}^d$ is the tensor-product of the normalized B-spline basis of order r on $I_d = [0,1]^d$ with $J_n = (k_n - r - 1)^d$. Then the maximum likelihood estimator maximizing (3) corresponds to the sieve estimator defined by

$$\hat{f}_n \coloneqq \arg\max_{f \in \mathcal{S}_n} \hat{Q}_n(f),$$

where

$$\hat{Q}_n(f) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \log f(s_i).$$

We now state the assumptions under which the consistency of the estimator $\hat{f}_n$ can be established.

Assumption 1 (Smoothness of the true density)**.** *The true density f_0 belongs to the Hölder class of α-smooth functions $\mathcal{H}^\alpha([0,1]^d)$, that is, f_0 is m times differentiable and its m-th derivative $D^m f_0$ is Hölder continuous with $0 < \gamma < 1$ and $\alpha = m + \gamma$.*

Assumption 2 (Spline order)**.** *The order r of the B-spline basis satisfies $r \geq \lfloor \alpha \rfloor$.*

Assumption 3 (Boundedness)**.** *The true density function f_0 is bounded away from zero and infinity; that is, there exist constants $K_1, K_2 \in \mathbb{R}$ such that*

$$0 < K_1 \leq \inf_{s \in [0,1]^d} f_0(s) \leq \sup_{s \in [0,1]^d} f_0(s) \leq K_2 < \infty. \tag{5}$$

It is well known that the space spanned by tensor-product B-spline basis functions is dense in the Hölder class $\mathcal{H}^\alpha([0,1]^d)$ under the L^p norm for any $1 \leq p \leq \infty$; see Schumaker (2007).

Theorem 1. *Under Assumptions 1–3, for $J_n \to \infty$,*

$$\|\hat{f}_n - f_0\|_\infty = o_p(1).$$

The proof is deferred to Section A.1 and follows the general sieve M-estimation argument in Chen (2007).

We now derive the convergence rate for $\hat{f}_n$ in the L^∞ norm. The rate is obtained by applying Theorem 3.2 in Chen (2007) and using the metric entropy evaluation of B-spline sieve spaces from Chen and Shen (1998).

Theorem 2. *Under Assumptions 1–3, and for*

$$J_n \asymp \left(\frac{\log n}{n} \right)^{-\frac{d}{2\alpha + d}},$$

we have

$$\|\hat{f}_n - f_0\|_\infty = O_p\left(\left(\frac{\log n}{n} \right)^{\frac{\alpha}{2\alpha + d}} \right).$$

The proof is deferred to Section A.2 and also follows the general sieve M-estimation argument in Chen (2007). It follows from Theorem 2 that the

estimator $\hat{f}_n$ achieves the optimal uniform convergence rate in the sense of Stone (1982).

3 Vector space construction for density functions

We consider the space of density functions on I_d defined via spline mixtures as in (2), denoted by

$$\mathscr{H}_J = \{p = (p_1, \ldots, p_J) : \forall j,\ p_j > 0,\ p_1 + \cdots + p_J = 1\}. \tag{6}$$

To introduce a regression framework on $\mathscr{H}_J$, we require well-defined operations of addition, scalar multiplication, and an inner product to endow $\mathscr{H}_J$ with a vector space structure. We aim to construct a bijective mapping from $\mathscr{H}_J$ to $\mathbb{R}^{J-1}$, where such linear operations and inner products are naturally defined.

3.1 Generalized Logit and Softmax Functions

The logit function is widely used in multiclass logistic regression. It transforms elements of $\mathscr{H}_J$ into $\mathbb{R}^{J-1}$ by selecting one coordinate as a reference (typically the first). For $(p_1, \ldots, p_J) \in \mathscr{H}_J$ with p_1 as the base, the transformation is given by

$$\operatorname{logit}(p) = \left(\log\left(\frac{p_2}{p_1}\right), \ldots, \log\left(\frac{p_J}{p_1}\right)\right), \tag{7}$$

which maps into $\mathbb{R}^{J-1}$. Its inverse is the softmax function, defined for $x = (x_2, \ldots, x_J) \in \mathbb{R}^{J-1}$ by

$$\operatorname{softmax}(x) = \left(\frac{1}{1 + \sum_{j=2}^{J} \exp(x_j)}, \frac{\exp(x_2)}{1 + \sum_{j=2}^{J} \exp(x_j)}, \ldots, \frac{\exp(x_J)}{1 + \sum_{j=2}^{J} \exp(x_j)}\right),$$

which belongs to $\mathscr{H}_J$. The pair of functions of the logit and softmax define a bijection between $\mathscr{H}_J$ and $\mathbb{R}^{J-1}$.

However, in density estimation, it often happens that some coefficients p_j

approach or equal zero, which causes instability in the logit transformation. To address this issue, we extend $\mathscr{H}_J$ to allow negative values:

$$\tilde{\mathscr{H}_J} = \{p = (p_1, \ldots, p_J) : \forall j,\ p_j > -1,\ p_1 + \cdots + p_J = 1\}, \tag{8}$$

and define a generalized logit function:

$$\widetilde{\text{logit}}(p) = \left(\log\left(\frac{1+p_2}{1+p_1}\right), \ldots, \log\left(\frac{1+p_J}{1+p_1}\right)\right), \tag{9}$$

with the corresponding generalized softmax function for $x = (x_2, \ldots, x_J) \in \mathbb{R}^{J-1}$ given by

$$\widetilde{\text{softmax}}(x) =$$
$$\left(\frac{J+1}{1+\sum_{j=2}^{J}\exp(x_j)} - 1, \frac{(J+1)\exp(x_2)}{1+\sum_{j=2}^{J}\exp(x_j)} - 1, \ldots, \frac{(J+1)\exp(x_J)}{1+\sum_{j=2}^{J}\exp(x_j)} - 1\right).$$

In summary, the logit and softmax functions yield a bijection between $\mathscr{H}_J$ and $\mathbb{R}^{J-1}$, while their generalized versions do so for $\tilde{\mathscr{H}_J}$ and $\mathbb{R}^{J-1}$. These define a bijection between $\tilde{\mathscr{H}_J}$ and $\mathbb{R}^{J-1}$. We shall employ the transformed space of $\tilde{\mathscr{H}_J}$ by the generalized logit as the representation space for density functions.

3.2 Inner product space of density functions

We now define suitable linear operations and an inner product to endow $\tilde{\mathscr{H}_J}$ with the structure of a Hilbert space. Note that $\tilde{\mathscr{H}_J}$ as defined in (8) is not closed under standard addition or scalar multiplication.

Our key idea is to define operations via their representations in $\mathbb{R}^{J-1}$. For $f, g \in \tilde{\mathscr{H}_J}$, we define:

$$f \oplus g := \widetilde{\text{softmax}}(\widetilde{\text{logit}}(f) + \widetilde{\text{logit}}(g)),$$
$$\alpha \otimes f := \widetilde{\text{softmax}}(\alpha \cdot \widetilde{\text{logit}}(f)),$$

and the inner product as

$$(f, g) := \left(\widetilde{\mathrm{logit}}(f), \widetilde{\mathrm{logit}}(g)\right)_H, \tag{10}$$

where for column vectors $\tilde{f} := \widetilde{\mathrm{logit}}(f)$ and $\tilde{g} := \widetilde{\mathrm{logit}}(g)$, the inner product is given by

$$\tilde{f}^\top H_{J-1} \tilde{g},$$

with H_{J-1} being a $(J-1) \times (J-1)$ positive definite matrix that defines a metric on $\mathbb{R}^{J-1}$.

To construct H_{J-1}, let e_i denote the i-th unit vector in $\mathbb{R}^{J-1}$, and define $\hat{e}_i := \widetilde{\mathrm{softmax}}(e_i) \in \tilde{\mathscr{H}}_J$. The corresponding spline density on $[0,1]^d$ is given by

$$\hat{e}_i(s) = \sum_{k=1}^{J} \hat{e}_{ik} \tilde{B}^d_{k,r}(s),$$

where $\tilde{B}^d_{k,r}$ denotes the normalized k-th B-spline basis function of order r on $[0,1]^d$. We define the (i,j)-th element of H_{J-1} by the standard L^2 inner product:

$$H^{(ij)}_{J-1} = \int_{[0,1]^d} \hat{e}_i(s) \hat{e}_j(s)\, ds, \tag{11}$$

which can be numerically evaluated using quadrature methods for B-spline functions. Note that H_{J-1} is generally non-diagonal due to the overlapped supports between B-spline basis functions.

4 Density-valued ARMA models

4.1 Definition

We define an auto-regressive and moving average (ARMA) model for time series of density functions. For a density function on $I_d = [0,1]^d$, we ap-

proximate it with an element in $\tilde{\mathscr{H}}_J$, where addition, subtraction, scalar multiplication, and inner product are defined. Here, J is assumed to be sufficiently large to ensure good approximation accuracy. Note that the positive definite matrix H_{J-1} used to define the inner product is the known constant given in (11).

In this section, we suppose to observe directly $\tilde{\mathscr{H}}_J$-valued time series as density-valued time series, ignoring fitting errors to estimate B-spline mixtures from discrete observations in practical situations.

For $\tilde{\mathscr{H}}_J$-valued time series $\{y_t \in \tilde{\mathscr{H}}_J, t = 1, 2, \ldots\}$, we define the auto-covariance function by identifying $\{y_t\}$ on the transformed space by the generalized logit. Namely, for $\tilde{y}_t = \widetilde{\mathrm{logit}}(y_t) \in \mathbb{R}^{J-1}$, the auto-covariance function is defined by

$$\mathrm{Cov}(y_t, y_{t-k}) := \mathbb{E}(\tilde{y}_t - \tilde{\mu}_t)' H_{J-1} (\tilde{y}_{t-k} - \tilde{\mu}_{t-k}), \tag{12}$$

where $\tilde{\mu}_t = \mathbb{E}\tilde{y}_t$ is the mean function. When both of mean and covariance functions do not depend on time, we say that $\{y_t\}$ is a $\tilde{\mathscr{H}}_J$-valued stationary process.

We define ARMA models for $\tilde{\mathscr{H}}_J$-valued stationary time series $\{y_t\}$ by those for transformed series by the generalized logit $\{\tilde{y}_t = \widetilde{\mathrm{logit}}(y_t) \in \mathbb{R}^{J-1}\}$ as

$$\tilde{y}_t = \tilde{\alpha} + \phi_1 \tilde{y}_{t-1} + \cdots + \phi_p \tilde{y}_{t-p} + \tilde{\varepsilon}_t + \theta_1 \tilde{\varepsilon}_{t-1} + \cdots + \theta_q \tilde{\varepsilon}_{t-q}, \tag{13}$$

where $\tilde{\varepsilon}_t$ are iid $(J-1)$-variate random vectors with mean zero and covariance matrix Σ_{J-1}. Note that $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q) \in \mathbb{R}^{p+q}$ is ARMA parameter vector and $\tilde{\alpha} = (\tilde{\alpha}_1, \ldots, \tilde{\alpha}_{J-1})$ is $(J-1)$-variate constant. We assume the following conditions for Σ_{J-1} and H_{J-1}:

$$\begin{aligned} J^{-1} \operatorname{tr}(H_{J-1}\Sigma_{J-1}) &\to C_1, \\ J^{-1} \operatorname{tr}(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1}) &\to C_2, \end{aligned} \tag{14}$$

as $J \to \infty$.

The ARMA model defined in (13) has a stationary auto-covariance func-

tion and, consequently, a spectral density function obtained as the inverse Fourier transform of the auto-covariance. Using (12), the auto-covariance of y_t becomes

$$R(k;\beta) = \mathbb{E}(\tilde{y}_t - \tilde{\mu})' H_{J-1}(\tilde{y}_{t-k} - \tilde{\mu}) = \gamma_k(\beta)\,\mathrm{tr}(H_{J-1}\Sigma_{J-1}),$$
$$\tilde{\mu} = \frac{\tilde{\alpha}}{1 - \phi_1 - \cdots - \phi_p},$$

where $\gamma_k(\beta)$ is the autocovariance function of a standard real-valued ARMA process driven by iid error with mean 0 and variance 1. Thus, the spectral density function is

$$\begin{aligned} f(\lambda;\beta) &= \frac{1}{2\pi}\sum_{k=-\infty}^{\infty} R(k;\beta)e^{-ik\lambda}, \quad \lambda \in [-\pi,\pi] \\ &= \frac{\mathrm{tr}(H_{J-1}\Sigma_{J-1})}{2\pi}\left|\frac{1+\sum_{j=1}^{q}\theta_j e^{-ij\lambda}}{1-\sum_{j=1}^{p}\phi_j e^{-ij\lambda}}\right|^2 = \frac{\mathrm{tr}(H_{J-1}\Sigma_{J-1})}{2\pi} g(\lambda;\beta), \end{aligned} \tag{15}$$

say.

Finally, the notions of causality and invertibility for the $\tilde{\mathscr{H}}_J$-valued ARMA models follow standard definitions used for real-valued ARMA models (Brockwell and Davis, 1991, Sec. 3.1). The necessary and sufficient condition is that the AR and MA polynomials

$$\phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p,$$
$$\theta(z) = 1 + \theta_1 z + \cdots + \theta_q z^q,$$

have no roots inside the unit circle. In subsequent inference procedures, we assume the parameter space for (13) is

$$\begin{aligned} C = \{\beta \in \mathbb{R}^{p+q} : \phi(z)\theta(z) \neq 0 \text{ for } |z| \leq 1, \phi_p \neq 0, \theta_q \neq 0, \text{ and} \\ \phi(\cdot) \text{ and } \theta(\cdot) \text{ has no common zeros}\}, \end{aligned}$$

so as to guarantee identifiability of the ARMA parameters.

4.2 Inference for density-valued AR models

Suppose we observe $\tilde{y}_1, \ldots, \tilde{y}_T \in \tilde{\mathbb{R}}^{J-1}$ from auto-regressive (AR) models of order p (i.e., when $q = 0$) in equation (13). We consider the estimation method and its asymptotic properties when both T and J diverge jointly.

Least squares estimation is applied to estimate the AR model parameters. Let $\phi = (\phi_1, \ldots, \phi_p)'$ and $\tilde{x}_t = (\tilde{y}_{t-1}, \ldots, \tilde{y}_{t-p})'$. The least squares estimator is obtained by minimizing

$$Q(\alpha, \phi) = \sum_{t=p+1}^{T} \|\tilde{y}_t - \tilde{\alpha} - \tilde{x}_t \phi\|_H^2 ,$$

under the metric defined in equation (10). For the mean-adjusted series

$$\bar{y}_t = \tilde{y}_t - \frac{1}{T-p} \sum_{t=p+1}^{T} \tilde{y}_t, \quad \bar{x}_t = \tilde{x}_t - \frac{1}{T-p} \sum_{t=p+1}^{T} \tilde{x}_t,$$

we obtain the least squares estimator

$$\hat{\phi}_J = \left\{ \sum_{t=p+1}^{T} \bar{x}_t' H_{J-1} \bar{x}_t \right\}^{-1} \left\{ \sum_{t=p+1}^{T} \bar{x}_t' H_{J-1} \bar{y}_t \right\}. \tag{16}$$

We now show the asymptotic normality of $\hat{\phi}_J$ as both T and J tend to infinity. Let ϕ_0 be the true parameter value and $\gamma(k; \phi_0)$ be the autocovariance function of the real-valued AR(p) model driven by iid errors with mean 0 and variance 1.

Theorem 3. *Let $\tilde{y}_t$ follow a causal AR(p) model as in (13) with $q = 0$, driven by iid errors with mean 0 and covariance matrix Σ_{J-1}. Under condition (14), as $T, J \to \infty$ jointly,*

$$\sqrt{TJ} \left(\hat{\phi}_J - \phi_0 \right) \to N \left(0, \tau_0 \Gamma_p^{-1} \right),$$

in distribution, where $\Gamma_p = [\gamma(i-j;\phi_0)]_{i,j=1}^p$ *and*

$$\tau_0 = \lim_{J\to\infty} \frac{J^{-1} tr(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1})}{\{J^{-1} tr(H_{J-1}\Sigma_{J-1})\}^2}. \tag{17}$$

The asymptotic variance of $\hat{\phi}_J$ *is consistently estimated by*

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1}, \tag{18}$$

where

$$\begin{aligned}
\hat{A} &= \sum_{t=p+1}^{T} \bar{x}_t' H_{J-1} \bar{x}_t, \\
\hat{u}_t &= \bar{y}_t - \bar{x}_t' \hat{\phi}_J, \\
\hat{B} &= \sum_{t=p+1}^{T} \bar{x}_t' H_{J-1} \hat{u}_t \hat{u}_t' H_{J-1} \bar{x}_t.
\end{aligned}$$

We extend the standard procedure for identifying the AR order p based on the partial autocovariance function (PACF) to the setting of density-valued time series. Define the sample autocorrelation function (ACF) by

$$\hat{\rho}(k) = \hat{R}(k)/\hat{R}(0), \quad k = 0, 1, 2, \ldots,$$

where

$$\hat{R}(k) = \frac{1}{T}\sum_{t=k+1}^{T} \left(\tilde{y}_t - \frac{1}{T}\sum_{s=1}^{T}\tilde{y}_s\right)' H_{J-1} \left(\tilde{y}_{t-k} - \frac{1}{T}\sum_{s=1}^{T}\tilde{y}_s\right). \tag{19}$$

The sample PACF of order k is defined as the k-th element of the least squares estimator $\hat{\phi}_J$ when fitting an AR(k) model.

Corollary 1. *Let* $\{\tilde{y}_t\}$ *be iid with mean* μ *and covariance matrix* Σ_{J-1}. *Under condition (14), as* $T, J \to \infty$ *jointly,*

$$\sqrt{TJ}\hat{\rho}(k) \to N(0, \tau_0), \quad \sqrt{TJ}\hat{\phi}(k) \to N(0, \tau_0),$$

in distribution, where τ_0 is given in (17). The asymptotic variance is consistently estimated by

$$\hat{\kappa} = T^{-1} \frac{tr\left(H_{J-1}\hat{\Sigma}_{J-1}H_{J-1}\hat{\Sigma}_{J-1}\right)}{\left\{tr(H_{J-1}\hat{\Sigma}_{J-1})\right\}^2},$$

where

$$\hat{\Sigma}_{J-1} = \frac{1}{T}\sum_{t=1}^{T}\left(\tilde{y}_t - \frac{1}{T}\sum_{s=1}^{T}\tilde{y}_s\right)\left(\tilde{y}_t - \frac{1}{T}\sum_{s=1}^{T}\tilde{y}_s\right)'.$$

It follows that the AR order can be identified by checking whether the sample PACF values are greater than $1.96\hat{\kappa}$, based on a 5% significance level.

4.3 Inference for density-valued ARMA models

We consider the estimation of ARMA models using observed samples $\tilde{y}_1, \ldots, \tilde{y}_T \in \mathbb{R}^{J-1}$ in equation (13). Least squares estimation cannot be directly applied in the ARMA setting. Instead, we employ the Whittle likelihood estimation method introduced by Brockwell and Davis (1991), omitting the discussion of maximum likelihood and least squares estimators in (Brockwell and Davis, 1991, Sec. 8.7) for simplicity.

Define the discrete Fourier transform and periodogram at the Fourier frequencies $\omega_j = 2\pi j/T,\ j = 1, \ldots, [T/2]$, as

$$d_y(\omega_j) := \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{y}_t e^{-i\omega_j t},$$
$$I_y(\omega_j) := d_y(\omega_j)' H_{J-1}\overline{d_y(\omega_j)}.$$

We estimate the ARMA parameter $\beta = (\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ by minimizing the Whittle likelihood function:

$$l_T(\beta) = \frac{1}{TJ}\sum_{j=1}^{[T/2]}\frac{I_y(\omega_j)}{g(\omega_j;\beta)}, \tag{20}$$

where $g(\omega;\beta)$ is defined in equation (15). Note that the zero frequency is omitted from this definition to avoid the effect of non-zero means.

We now present the asymptotic properties of the Whittle estimator under the assumption that $\tilde{\alpha}$ and Σ_{J-1} are nuisance parameters.

Theorem 4. *Let $\hat{\beta} \in C$ be the minimizer of $l_T(\beta)$ in equation (20), and suppose $\tilde{y}_t$ follows a causal and invertible ARMA model with true parameter $\beta_0 \in C$. Under the condition (14), as $T, J \to \infty$ jointly,*

$$\sqrt{TJ}(\hat{\beta} - \beta_0) \to N\left(0, \tau_0 W^{-1}(\beta_0)\right),$$

where

$$W(\beta_0) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\left[\frac{\partial \log g(\lambda;\beta_0)}{\partial\beta}\right]\left[\frac{\partial \log g(\lambda;\beta_0)}{\partial\beta}\right]' d\lambda,$$
$$\tau_0 = \lim_{J\to\infty}\frac{J^{-1}tr\left(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1}\right)}{\left\{J^{-1}tr(H_{J-1}\Sigma_{J-1})\right\}^2}.$$

The asymptotic variance of $\hat{\beta}$ is consistently estimated by

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1},$$

where

$$\hat{A} = \sum_{j=1}^{[T/2]} I_y(\omega_j)\left[\frac{\partial^2 g^{-1}(\omega_j;\hat{\beta})}{\partial\beta\partial\beta'}\right],$$
$$\hat{B} = \sum_{j=1}^{[T/2]} |I_y(\omega_j)|^2\left[\frac{\partial g^{-1}(\omega_j;\hat{\beta})}{\partial\beta}\right]\left[\frac{\partial g^{-1}(\omega_j;\hat{\beta})}{\partial\beta}\right]'.$$

5 Application to population data

This section analyzes human population density in Tokyo. After fitting spline mixtures in (2) by maximizing the likelihood in (3), we obtain a dataset expressed in the form of $\mathscr{H}_J$ as in (6). Transforming $\mathscr{H}_J$ into $\mathbb{R}_{J-1}$ by the generalized logit transformation in (9), we apply the time series methods

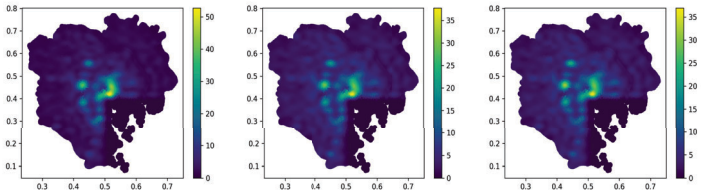

Figure 1: Estimated densities in April 2020 (left), April 2021 (middle), and April 2022 (right), for monthly averaged human population in the Tokyo 23 central districts at 15:00.

introduced in Section 4.

NTT DoCoMo, a Japanese mobile phone company, provides spatio-temporal data counting the number of people in 500-meter meshes every hour throughout Japan since 2016. We collected population data over 6,100 meshes in Tokyo (a 102 km $\times$ 74 km area including the 23 central districts) at 15:00 for each day. Averaging these daily measurements over each month, we constructed a monthly dataset spanning 96 months from January 2016 to December 2023.

Converting the longitude and latitude coordinates of the 6,100 meshes into two-dimensional Cartesian coordinates (using the transformation formula from the Geospatial Information Authority of Japan), we treat the resulting population data as spatio-temporal data on a two-dimensional Euclidean space. For analysis, we focus on the Tokyo 23 districts and normalize the spatial domain to lie within the unit square $[0,1]^2$.

We applied cubic B-spline mixtures over $[0,1]^2$ to estimate monthly densities in the Tokyo 23 districts using $J = 819$ basis functions. Figure 1 shows the estimated densities in April in 2020, 2021 and 2022. As a result, we obtained density functions $y_t \in \mathscr{H}_J$, $t = 1, \ldots, 96$, with $J = 819$. We then analyzed the transformed series $\tilde{y}_t = \widetilde{\text{logit}}(y_t)$ by generalized logit function in (9), applying the ARMA methodology from Section 4.

First, we computed the sample autocorrelation (ACF) and partial autocorrelation (PACF) functions defined in (19), including critical values at

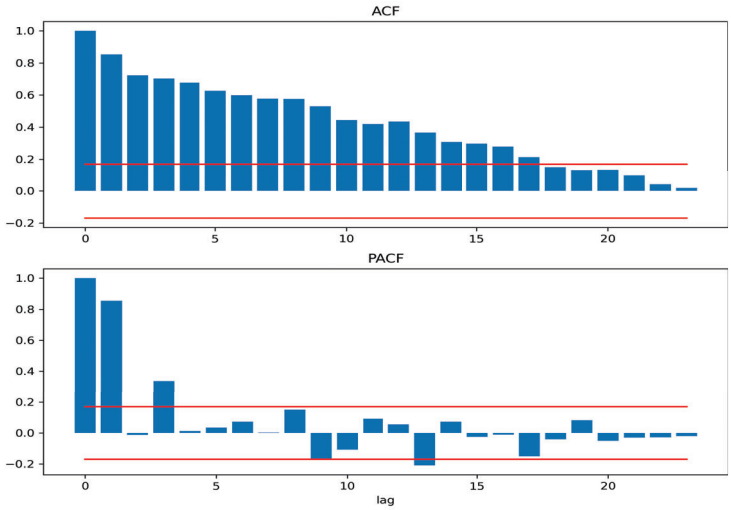

Figure 2: Sample autocorrelation functions (top) and partial autocorrelation functions (bottom) for the monthly averaged density-valued time series of human population in the Tokyo 23 central districts at 15:00 from Jan. 2016 to Dec. 2023.

the 5% significance level under the null hypothesis of independence, to determine the appropriate ARMA order. As shown in Figure 2, the PACF suggests that an AR(3) model is suitable, with significant partial autocorrelations up to lag 3. The ACF also shows positive values up to lag 17, consistent with AR(3) behavior.

The AR(3) parameters for the transformed series $\tilde{y}_t = \widetilde{\text{logit}}(y_t) \in \mathbb{R}^{J-1}$, estimated using (16), are:

$$\begin{aligned} \tilde{y}_t =& 0.867\tilde{y}_{t-1} - 0.303\tilde{y}_{t-2} + 0.337\tilde{y}_{t-3} + \tilde{\varepsilon}_t, \\ & (0.125) \qquad (0.121) \qquad (0.090), \\ R^2 =& 0.763. \end{aligned}$$

For comparison, the traditional estimation obtained just by fitting OLS directly to $y_t \in \mathscr{H}_J$ under the usual Euclidean distance without preserving the density structure yields:

$$\begin{aligned} y_t =& 0.830y_{t-1} - 0.242y_{t-2} + 0.316y_{t-3} + \varepsilon_t, \\ & (0.118) \qquad (0.106) \qquad (0.084), \\ R^2 =& 0.764. \end{aligned}$$

These results are similar, though the latter does not ensure that both sides of the equation are valid density functions. This suggests that our proposed density-valued time series framework appropriately accounts for serial correlation with preserving the density structure under the practically reasonable metric of H_{J-1}.

6 Conclusion

This paper has proposed a dynamic modeling framework for density-valued time series on the unit cube $I_d = [0,1]^d$ using a transformation-based approach. By representing density functions as B-spline mixtures, we construct a bijective mapping between the space of density functions and a Euclidean representation space via generalized logit and softmax transformations. This

approach enables us to define and estimate auto-regressive and moving average (ARMA) models for density-valued data using well-established tools from time series analysis.

A key advantage of our method is that it allows the application of linear operations and statistical inference in the transformed space, while ensuring the resulting densities remain valid under the inverse transformation. Although the inclusion of negative components in the mixture representation may appear counterintuitive, it enables computational tractability and interpretability, especially in contexts such as finance, where negative weights can reflect short positions in portfolios.

We developed estimation procedures based on least squares for AR models and Whittle likelihood for ARMA models, and established their asymptotic properties when both the time dimension T and the basis dimension J tend to infinity jointly. The effectiveness of the proposed method was demonstrated through an empirical application to human population density data in Tokyo, where the model successfully captured meaningful temporal dynamics while preserving the structural properties of densities.

Future work includes extending the framework to nonstationary settings, incorporating exogenous covariates, and exploring high-dimensional spatial domains or irregular domains beyond the unit cube. Moreover, formalizing the theoretical properties under estimation error of the initial density function (i.e., when plug-in estimates from observed data are used) is an important direction for making the method more robust in practice.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 23K21647.

A Proofs

A.1 Proof of Theorem 1

Let us begin to prepare some notation required in the proof. We define the true criterion function Q, the empirical criterion function $\hat{Q}_n$, and the sieve MLE $\hat{f}_n$ as follows:

$$\begin{aligned} Q(f) &\coloneqq \mathbb{E}[\log f(s_i)], \\ \hat{Q}_n(f) &\coloneqq \frac{1}{n}\sum_{i=1}^{n} \log f(s_i), \\ \hat{f}_n &\coloneqq \arg\max_{f \in \mathcal{S}_n} \hat{Q}_n(f). \end{aligned}$$

Define the space to which the true density function f_0 belongs by

$$\mathcal{S} := \left\{ f \in \mathcal{H}^{\alpha}([0,1]^d) : f(x) \geq 0, \int_{[0,1]^d} f(x)dx = 1 \right\},$$

and the spline sieve space by

$$\mathcal{S}_n = \left\{ \sum_{i=1}^{J_n} \phi_i \tilde{B}_{i,r}^{d}(s) : \boldsymbol{\phi} = (\phi_1, \ldots, \phi_J) \in \Delta^{J_n} \right\},$$

where

$$\Delta^{J_n} = \left\{ \boldsymbol{\phi} \in \mathbb{R}^{J_n} : \phi_i \geq 0, \sum_{i=i}^{J_n} \phi_i = 1 \right\}.$$

We restrict $\mathcal{S}$ together with the sieve space $\mathcal{S}_n$ in (4), in accordance with Assumption 3, as follows:

$$\mathcal{F} := \mathcal{S} \cap \mathcal{F}_{[K_1,K_2]}, \quad \mathcal{F}_n := \mathcal{S}_n \cap \mathcal{F}_{[K_1,K_2]},$$

where

$$\mathcal{F}_{[K_1,K_2]} := \left\{ f : [0,1]^d \to \mathbb{R}_+ \;\middle|\; K_1 \leq f(s) \leq K_2 \text{ for all } s \in [0,1]^d \right\}.$$

We now prove Theorem 1. The claim follows from Theorem 3.1 in Chen (2007), once we verify that all required conditions are satisfied. Specifically: we confirm Conditions 3.1, 3.2, 3.3, 3.4 and 3.5(i) of Theorem 3.1 in Chen (2007) by Lemmas 3, 2, 4, 1 and 5, respectively, which are shown in the followings.

Lemma 1. *The sieve spaces, $\mathcal{F}_k$, are compact under the distance $d(f_1, f_2) = \|f_1 - f_2\|_\infty$.*

Proof. Δ^{J_k} is a closed and bounded subset of $\mathbb{R}^{J_k}$, hence compact by Heine-Borel theorem. Define

$$\alpha : \Delta^{J_k} \longrightarrow C([0,1]^d), \qquad \alpha(\boldsymbol{\phi}) = \sum_{j=1}^{J_k} \phi_j \, \tilde{B}_j, \tag{21}$$

and equip $C([0,1]^d)$ with the supremum norm $\|f\|_\infty = \sup_{s\in[0,1]^d} |f(s)|$. For any $\boldsymbol{\phi}, \boldsymbol{\psi} \in \Delta^{J_k}$,

$$\|\alpha(\boldsymbol{\phi}) - \alpha(\boldsymbol{\psi})\|_\infty = \sup_{s\in[0,1]^d} \Big| \sum_{j=1}^{J_k} (\phi_j - \psi_j) \tilde{B}_j(s) \Big| \tag{22}$$

$$\leq \Big(\max_{1\leq j\leq J_k} \|\tilde{B}_j\|_\infty \Big) \, \|\boldsymbol{\phi} - \boldsymbol{\psi}\|_1, \tag{23}$$

hence α is Lipschitz, and in particular continuous. Since Δ^{J_k} is compact and α is continuous, $\mathcal{S}_k = \alpha(\Delta^{J_k})$ is compact in $C([0,1]^d)$ (supremum-norm topology).

For each fixed s, the evaluation map $\mathrm{ev}_s : \mathcal{F}_{[K_1,K_2]} \to \mathbb{R},\ f \mapsto f(s)$ is continuous under $\|\cdot\|_\infty$. Since $[K_1, K_2] \subset \mathbb{R}$ is closed and inverse images of closed sets under a continuous map are closed, each set $\mathrm{ev}_s^{-1}([K_1, K_2])$ is closed in the space of d-dimensional functions. Hence $\mathcal{F}_{[K_1,K_2]} = \bigcap_{s\in[0,1]^d} \mathrm{ev}_s^{-1}([K_1, K_2])$ is a intersection of closed sets, and therefore closed.

A closed subset of a compact space is compact. Thus $\mathcal{F}_k = \mathcal{S}_k \cap \mathcal{F}_{[K_1,K_2]}$ is compact in the supremum-norm topology. □

Lemma 2. *$\mathcal{F}_k \subseteq \mathcal{F}_{k+1} \subseteq \mathcal{F}$ for all $k \geq 1$; and there exists a sequence $\pi_k f_0 \in \mathcal{F}_k$ such that $d(\pi_k f_0, f_0) \to 0$ as $k \to \infty$.*

Proof. The sequence of sieve spaces $\{\mathcal{F}_n\}$ is nested, i.e., $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$ for all k. This follows from the construction of $\mathcal{F}_k$ as the span of tensor-product B-spline basis functions with increasing numbers of knots. Since adding more knots to the B-spline system strictly enlarges the span of the basis without discarding existing basis functions, the resulting spline spaces are automatically nested.

By Assumption 1, $f_0 \in \mathcal{H}^\alpha([0,1]^d)$. It is well known that, under the mesh ratio condition, the space of tensor-product B-spline functions of order $r \geq \alpha$ is dense in $\mathcal{H}^\alpha([0,1]^d)$ with respect to the L^∞-norm (see Schumaker (2007)). Hence, there exists a sequence $\hat{f}_n \in \text{span}\{\tilde{B}_{1,r}, \ldots, \tilde{B}_{J_n,r}\}$ such that $\|\hat{f}_n - f_0\|_\infty \to 0$.

Although $\hat{f}_n$ may not integrate to one, one can normalize it via $\tilde{f}_n(x) = \frac{\hat{f}_n(x)}{\int \hat{f}_n(u)du}$, and this operation preserves convergence in L^2, provided the denominator converges to 1, which is ensured as $\tilde{f}_n \to f_0$ in L^∞ and f_0 is a density.

Furthermore, since f_0 satisfies $m \leq f_0(x) \leq M$ and the sieve space $\mathcal{F}_n$ is restricted to functions satisfying the same bounds, the boundedness constraint does not interfere with the L^∞-density property. The class of functions in $\mathcal{F}_n$ remains sufficiently rich to approximate f_0 in L^∞ norm. □

Lemma 3. *(i) $Q(f)$ is continuous at f_0 in $\mathcal{F}$, $Q(f_0) > -\infty$; and (ii) for all $\varepsilon > 0$, $Q(f_0) > \sup_{\{f \in \mathcal{F}: \|f_0 - f\|_\infty \geq \varepsilon\}} Q(f)$.*

Proof. (i) Under Assumption 3, we have

$$\begin{aligned}
|Q(f) - Q(f_0)| &= \mathbb{E}\left[\log \frac{f(s)}{f_0(s)}\right] \\
&\leq \mathbb{E}\left[\frac{f(s)}{f_0(s)} - 1\right] \\
&\leq \int_{[0,1]^d} \frac{|f(s) - f_0(s)|}{f_0(s)} f_0(s) ds \\
&\leq \int_{[0,1]^d} |f(s) - f_0(s)| ds \\
&\leq \|f - g\|_\infty.
\end{aligned}$$

Then, $Q(f)$ is Lipschitz continuous on $\mathcal{F}$ at f_0. The first inequality follows from $\log x \leq x - 1$. It is obvious that $Q(f_0) \geq \log K_2 > 0$ under Assumption 3.

(ii) From the definition of the Kullback-Leibler divergence, we have $Q(f_0) - Q(f) = \mathrm{KL}(f_0||f) \geq 0$. Suppose, to the contrary, that $Q(f_0) = \sup_{\{f \in \mathcal{F}: \|f_0 - f\|_\infty \geq \varepsilon\}} Q(f)$. But if $Q(f_n) \to Q(f_0)$, it follows that $\mathrm{KL}(f_0||f_n) \to 0$, which in turn implies that $f_n \to f_0$ almost everywhere. However, this contradicts $\|f_0 - f\|_\infty \geq \varepsilon$ since both f_0 and f are continuous. □

Lemma 4. *For each $k \geq 1$, (i) $\hat{Q}_n(f)$ is a measurable function of the data $\{s_i\}_{i=1}^n$ for all $f \in \mathcal{F}_k$; and (ii) for any data $\{s_i\}_{i=1}^n$, $\hat{Q}_n(f)$ is upper semicontinuous on $\mathcal{F}_k$ under the metric $d(f_1, f_2) = \|f_1 - f_2\|_\infty$.*

Proof. (i) Fix any $k \geq 1$ and any $f \in \mathcal{F}_k$. Since each s_i is fixed, the map $(s_1, \ldots, s_n) \mapsto \frac{1}{n} \sum_{i=1}^n \log f(s_i)$ is measurable in the data. Hence, $\hat{Q}_n(f)$ is a measurable function of the data.

(ii) For each s_i, the map $f \mapsto f(s_i)$ is continuous with respect to the L^∞-norm on $\mathcal{F}_k$, as shown previously. Moreover, the logarithm function $x \mapsto \log x$ is continuous on $[K_1, K_2]$, so the composition $f \mapsto \log f(s_i)$ is continuous on $\mathcal{F}_k$. Finally, $\hat{Q}_n(f)$ is the finite average of the continuous maps $f \mapsto \log f(s_i)$, and thus $\hat{Q}_n$ is continuous on $\mathcal{F}_k$ with respect to $d(f_1, f_2) = \|f_1 - f_2\|_\infty$. Since continuous functions are trivially upper semicontinuous, the claim follows. □

Lemma 5. *For all $k \geq 1$, $\operatorname{plim\,sup}_{n \to \infty} \sup_{f \in \mathcal{F}_k} |\hat{Q}_n(f) - Q(f)| = 0$*

Proof. Fix any $k \geq 1$. The map $\theta \mapsto f_\theta$ is continuous and the simplex Δ^{J_k} is compact in $\mathbb{R}^{p_k^d}$; hence the sieve $\mathcal{F}_k = \{f_\theta : \theta \in \Delta^{J_k}\}$ is compact in $L^\infty([0,1]^d)$.

Because $f_\theta(s) \in [K_1, K_2]$ for all $s \in [0,1]^d$ and $\theta \in \Delta^{J_k}$,

$$\log m \;\leq\; \log f_\theta(s) \;\leq\; \log M, \qquad \forall\, s,\ \theta.$$

Define the envelope $G(s) := \max\{|\log m|, |\log M|\}$. Then $|\log f_\theta(s)| \leq G(s) < \infty$ and $G \in L^2(P_{f_0})$.

For any $\theta, \theta' \in \Delta^{J_k}$ and $s \in [0,1]^d$,

$$\big|\log f_\theta(s) - \log f_{\theta'}(s)\big| \;=\; \Big|\log \frac{f_\theta(s)}{f_{\theta'}(s)}\Big| \;\le\; \frac{1}{m}\,\big|f_\theta(s) - f_{\theta'}(s)\big| \;\le\; \frac{\|\tilde{B}\|_\infty}{m}\,\|\theta - \theta'\|_1,$$

where $\|\tilde{B}\|_\infty := \max_j \sup_s \tilde{B}_j(s)$. Hence $\theta \mapsto \log f_\theta$ is Lipschitz into $(L^\infty, \|\cdot\|_\infty)$, and

$$N\big(\varepsilon, \{\log f_\theta\}, \|\cdot\|_\infty\big) \;\lesssim\; \varepsilon^{-(p_k^d - 1)}.$$

Consequently,

$$\int_0^{\|G\|_2} \sqrt{\log N_{[]}\big(u, \log \mathcal{F}_k, L^2(P_{f_0})\big)}\, du < \infty,$$

so $\{\log f_\theta : \theta \in \Delta^{J_k}\}$ is P_{f_0}-Glivenko–Cantelli (van der Vaart & Wellner, 1996, Thm 2.4.1).

Therefore

$$\sup_{\theta \in \Delta^{J_k}} \Big|\frac{1}{n}\sum_{i=1}^n \log f_\theta(s_i) - \mathbb{E}_{f_0}\big[\log f_\theta(s)\big]\Big| \;\xrightarrow{\text{a.s.}}\; 0, \qquad n \to \infty,$$

and hence also in probability. Since $\hat{Q}_n(f) = n^{-1}\sum_{i=1}^n \log f(s_i)$ and $Q(f) = \mathbb{E}_{f_0}[\log f(s)]$, this yields

$$\operatorname{plim}_{n\to\infty} \sup_{f \in \mathcal{F}_k} \big|\hat{Q}_n(f) - Q(f)\big| = 0. \qquad \square$$

A.2 Proof of Theorem 2

We can prove this by checking Conditions 3.7 and 3.8 of Theorem 3.2 in Chen (2007), which are confirmed by Lemmas 6 and 7, respectively.

Lemma 6. *There is $C_1 > 0$ such that for all small $\varepsilon > 0$,*

$$\sup_{\{f \in \mathcal{F}_n : \|f_0 - f\|_\infty \le \varepsilon\}} \mathit{Var}(\log f(s_i) - \log f_0(s_i)) \le C_1 \varepsilon^2.$$

Proof. Let $s_i \sim f_0$, then

$$\begin{aligned}
\mathrm{Var}(\log f(s_i) - \log f_0(s_i)) &= \mathbb{E}\left(\log \frac{f(s_i)}{f_0(s_i)}\right)^2 - \mathbb{E}^2\left(\log \frac{f(s_i)}{f_0(s_i)}\right) \\
&\leq \mathbb{E}\left(\log \frac{f(s_i)}{f_0(s_i)}\right)^2 \\
&= \mathbb{E}\left(\log\left\{1 + \left(\frac{f(s_i)}{f_0(s_i)} - 1\right)\right\}\right)^2 \\
&\leq \mathbb{E}\left(\frac{f(s_i)}{f_0(s_i)} - 1\right)^2 \\
&= \mathbb{E}\left\{\frac{1}{f_0^2(s_i)}\left(f(s_i) - f_0(s_i)\right)^2\right\} \\
&\leq \frac{1}{K_1^2}\int_{[0,1]^d} f_0(s_i)\left(f(s_i) - f_0(s_i)\right)^2 ds \\
&\leq \frac{K_2}{K_1^2}\|f(s_i) - f_0(s_i)\|_\infty^2.
\end{aligned}$$

Thus, the result holds with $C_1 = \frac{K_2}{K_1^2}$. □

Lemma 7. *For any $\delta > 0$, there exists a constant $\rho \in (0, 2)$ such that*

$$\sup_{\{f \in \mathcal{F}_n : \|f_0 - f\|_\infty \leq \delta\}} |\log f(s_i) - \log f_0(s_i)| \leq \delta^\rho U(s_i),$$

with $\mathbb{E}[U^\gamma(s_i)] \leq C_2$ for some $\gamma \geq 2$.

Proof. Let $s_i \sim f_0$, then

$$\begin{aligned}
|\log f(s_i) - \log f_0(s_i)| &= \left|\log \frac{f(s_i)}{f_0(s_i)}\right| \\
&= \left|\log\left\{1 + \frac{f(s_i) - f_0(s_i)}{f_0(s_i)}\right\}\right| \\
&\leq \log\left\{1 + \frac{|f(s_i) - f_0(s_i)|}{f_0(s_i)}\right\}
\end{aligned}$$

Using the inequality, $\frac{x}{x+1} \leq \log(1+x) \leq x$ for all $x > 0$, we obtain

$$\begin{aligned}
\log\left\{1 + \frac{|f(s_i) - f_0(s_i)|}{f_0(s_i)}\right\} &\leq \max\left\{\frac{|f(s_i) - f_0(s_i)|}{f_0(s_i)}, \frac{|f(s_i) - f_0(s_i)|}{f(s_i)}\right\} \\
&\leq \frac{1}{K_1}|f(s_i) - f_0(s_i)| \\
&\leq \frac{1}{K_1}\|f - f_0\|_\infty
\end{aligned}$$

Thus, the result holds with $\rho = 1$ and $U(s_i) = \frac{1}{K_1}$. □

A.3 Proof of Theorem 3

Proof. In order for $\hat{\phi}_J$ to converge in distribution jointly as $T, J \to \infty$, we apply Proposition 6.3.9 of Brockwell and Davis (1991). Namely, we show that $\hat{\phi}_J$ converges in distribution as $T \to \infty$ for each finite J and that the limit distribution for each J converges in distribution as $J \to \infty$.

Observe that

$$\bar{x}_t := \tilde{x}_t - \frac{1}{T}\sum_{t=1}^{T}\tilde{x}_t = (\tilde{x}_t - \tilde{\mu}) - \left(\frac{1}{T}\sum_{t=1}^{T}\tilde{x}_t - \tilde{\mu}\right).$$

Since the second part in the last term is $O_p(T^{-1/2})$ in combinations with the condition, we have, for each $J = 1, 2, \ldots,$

$$\begin{aligned}
&\sqrt{TJ}\left(\hat{\phi}_J - \phi_0\right) = O_p\left(T^{-1/2}\right) \\
&+ \left\{\frac{1}{TJ}\sum_{t=1}^{T-p}(\tilde{x}_t - \tilde{\mu})' H_{J-1}(\tilde{x}_t - \tilde{\mu})\right\}^{-1} \left\{\frac{1}{\sqrt{TJ}}\sum_{t=1}^{T-p}(\tilde{x}_t - \tilde{\mu})' H_{J-1}\tilde{\varepsilon}_t\right\}.
\end{aligned}$$

The first prat in the second term converges in probability to

$$J^{-1} tr\left(H_{J-1}\Sigma_{J-1}\right)\Gamma_p,$$

while the second part converges in distribution to

$$N\left(0, J^{-1} tr\left(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1}\right)\Gamma_p\right),$$

as $T \to \infty$ for each $J = 1, 2, \ldots$. Applying Proposition 6.3.9 of Brockwell and Davis (1991), we have the result as T, J jointly diverge by checking that the asymptotic variance converges as $J \to \infty$. □

A.4 Proof of Theorem 4

We prove it in the same way with Theorem 3 by Proposition 6.3.9 of Brockwell and Davis (1991).

Lemma 8. *For every $\beta \in C$,*

$$l_T(\beta) \to \frac{J^{-1} tr(H_{J-1}\Sigma_{J-1})}{2\pi} \int_{-\pi}^{\pi} \frac{g(\lambda;\beta_0)}{g(\lambda;\beta)} d\lambda,$$

uniformly in $\beta \in \bar{C}$ in probability, as $T \to \infty$ for each $J = 1, 2, \ldots$.

Proof. It can be proved basically by following Proposition 10.8.2 of Brockwell and Davis (1991). □

Lemma 9. *Let $\hat{\beta}$ be the estimator in C which minimizes (20), where $\tilde{y}_t$ follows an ARMA model in (13) with a true parameter $\beta_0 \in C$. Then*

$$\hat{\beta} \to \beta_0,$$

in probability, as $T \to \infty$ for each $J = 1, 2, \ldots$.

Proof. The probability limit function of $l_T(\beta)$ evaluated in Lemma 8 attains the minimum at $\beta = \beta_0$ by Proposition 10.8.1 of Brockwell and Davis (1991). Since the convergence in Lemma 8 is uniform in C, the consistency follows. □

Now we are in the position to prove Theorem 4.

Proof. The Taylor expansion of $\partial l_T(\beta_0)/\partial\beta$ about $\beta = \hat{\beta}$ can be written as

$$\begin{aligned}\sqrt{TJ}\frac{\partial l_T(\beta_0)}{\partial\beta} &= \sqrt{TJ}\frac{\partial l_T(\hat{\beta})}{\partial\beta} + \sqrt{TJ}\frac{\partial^2 l_T(\tilde{\beta})}{\partial\beta\partial\beta'}(\beta_0 - \hat{\beta}) \\ &= -\sqrt{TJ}\frac{\partial^2 l_T(\tilde{\beta})}{\partial\beta\partial\beta'}(\hat{\beta} - \beta_0)\end{aligned}$$

for some $\tilde{\beta} \in C$ satisfying $||\tilde{\beta} - \hat{\beta}|| < ||\hat{\beta} - \beta_0||$. Lemmas 8 and 9 can be used to establish

$$\begin{aligned}\frac{\partial^2 l_T(\tilde{\beta})}{\partial\beta\partial\beta'} \to & \frac{J^{-1}tr(H_{J-1}\Sigma_{J-1})}{2\pi}\int_{-\pi}^{\pi} g(\lambda;\beta_0)\frac{\partial^2 g^{-1}(\lambda;\beta_0)}{\partial\beta\partial\beta'}d\lambda \\ & = 2W(\beta_0)J^{-1}tr(H_{J-1}\Sigma_{J-1})\end{aligned}$$

in probability as $T \to \infty$. Hence it sufices to show that

$$\sqrt{TJ}\frac{\partial l_T(\beta_0)}{\partial\beta} \to N\left(0, 4W(\beta_0)J^{-1}tr(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1})\right),$$

as $T \to \infty$ for each $J = 1, 2, \ldots$. Applying Propositions 10.8.5 and 10.8.6 of Brockwell and Davis (1991), we can show that, for $\eta(\omega) = \partial g^{-1}(\omega;\beta)/\partial\beta$,

$$E\left|\frac{1}{\sqrt{TJ}}\sum_j \{I(\omega_j)\eta(\omega_j) - g(\omega_j;\beta_0)I_\varepsilon(\omega_j)\eta(\omega_j)\}\right| \to 0,$$

and

$$\begin{aligned}&\frac{1}{\sqrt{TJ}}\sum_j g(\omega_j;\beta_0)I_\varepsilon(\omega_j)\eta(\omega_j) \to \\ &\qquad N\left(0, J^{-1}\frac{tr\left(H_{J-1}\Sigma_{J-1}H_{J-1}\Sigma_{J-1}\right)}{\pi}\int_{-\pi}^{\pi}\eta^2(\lambda)g^2(\lambda;\beta_0)d\lambda\right),\end{aligned}$$

in distribution as $T \to \infty$, respectively. Finally, applying Proposition 6.3.9 of Brockwell and Davis (1991), we complete the proof by checking that the asymptotic variance converges as $J \to \infty$. □

References

Box, G. E., G. M. Jenkins, and G. C. Reinsel (1994). *Time Series Analysis: Forecasting and Control*. Prentice Hall.

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. Springer.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics 6*, 5549–5632.

Chen, X. and X. Shen (1998). Sieve extremum estimates for weakly dependent data. *Econometric theory 14*(4), 553–575.

McLachlan, G. J. and D. Peel (2007). *Finite Mixture Models* (1st ed.). Wiley-Interscience.

Panaretos, V. M. and Y. Zemel (2020). *An Invitation to Statistics in Wasserstein Space*. Springer Series in Statistics. Springer.

Petersen, A. and H.-G. Müller (2016). Functional data analysis of density functions by transformation to a hilbert space. *The Annals of Statistics 44*(1), 183–218.

Petersen, A., C. Zhang, and P. Kokoszka (2022). Modeling probability density functions as data objects. *Econometrics and Statistics 21*, 159–178.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory* (3rd ed.). Cambridge University Press.

Shen, X. and W. H. Wong (1994). Convergence rate of sieve estimates. *The Annals of Statistics 22*(2), 580–615.

Srivastava, A., E. Klassen, S. Joshi, and I. Jermyn (2007). On the analysis of shape data. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*(4), 728–742.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Zhang, C., P. Kokoszka, and A. Petersen (2022). Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis 43*(1), 30–52.