Bayesian Analysis (2023) **18**, Number 2, pp. 579–605

Bayesian Spatial Homogeneity Pursuit of Functional Data: An Application to the U.S. Income Distribution*

Guanyu Hu[†], Junxian Geng[‡], Yishu Xue[§,‖], and Huiyan Sang[¶]

Abstract. An income distribution describes how an entity's total wealth is distributed amongst its population. A problem of interest to regional economics researchers is to understand the spatial homogeneity of income distributions among different regions. In economics, the Lorenz curve is a well-known functional representation of income distribution. In this article, we propose a mixture of finite mixtures (MFM) model as well as a Markov random field constrained mixture of finite mixtures (MRFC-MFM) model in the context of spatial functional data analysis to capture spatial homogeneity of Lorenz curves. We design efficient Markov chain Monte Carlo (MCMC) algorithms to simultaneously infer the posterior distributions of the number of clusters and the clustering configuration of spatial functional data. Extensive simulation studies are carried out to show the effectiveness of the proposed methods compared with existing methods. We apply the proposed spatial functional clustering method to state level income Lorenz curves from the American Community Survey Public Use Microdata Sample (PUMS) data. The results reveal a number of important clustering patterns of state-level income distributions across the US.

Keywords: Lorenz curve, Markov random field, mixture of finite mixtures, spatial functional data clustering.

MSC2020 subject classifications: Primary 62p20; secondary 91b72.

1 Introduction

Our study is motivated by an American Community Survey Public Use Microdata Sample (PUMS) data that contains incomes of United States (US) households in year 2017, which can be accessed via the PUMS data registry (https://www.census.gov/programs-surveys/acs/data/pums.html). Incomes of households and their states of residence are recorded. Our primary goal is to cluster the state level Income Distributions (ID; O'sullivan and Sheffrin, 2007), i.e., how a state's total wealth is distributed

*The research of Huiyan Sang was partially supported by NSF grant no. NSF DMS-1854655.

[†]Department of Statistics, University of Missouri Columbia, Columbia, MO 65211, gh7mr@missouri.edu

[‡]Boehringer-Ingelheim Pharmaceutical Inc, 900 Ridgebury Rd, Ridgefield, CT 06877, gengjunxianjohn@gmail.com

[§]Department of Statistics, University of Connecticut, Storrs, CT 06269, yishu.xue@uconn.edu

[¶]Department of Statistics, Texas A&M University, College Station, TX, 77843, huiyan@stat.tamu.edu

[‖]Corresponding author.

 https://doi.org/10.1214/22-BA1320

amongst its population. In order to avoid confusion between the economics term "Income Distribution" and the density or cumulative distribution of household income, we use ID to represent this particular economics term in the rest of the paper. The ID has been a central concern of economic theory since the time of classical economists such as Adam Smith and David Ricardo. While economists have been conventionally concerned with the relationship between the factors of production, land, labor, and capital for ID, modern economists now focus more on income inequality. Particularly, a balance between income inequality and economic growth is a desired goal for policy makers. Capturing the homogeneity pattern of state level IDs is of great research interest in economic studies, as it will improve understanding of income inequality among different regions within a country, and provide policy makers with a reference point for implementing differentiated policies for the identified regions. In macroeconomics, most governments want to obtain an equitable (fair) distribution of income, which is a crucial element of a functioning democratic society (Mankiw, 2014). To achieve this goal, the distribution of income or wealth in an economy is represented by a Lorenz curve (Lorenz, 1905), which is a function showing the proportion of total income received by the bottom $100p\%$ ($p \in [0, 1]$) of the population. The Gini coefficient, which is derived from the Lorenz curve, is a frequently used indicator of income inequality (Gini, 1997), and it has been widely adopted by numerous international organizations, such as the United Nations and World Bank, to analyze income inequalities between countries and regions. It is, on the other hand, a non-unique scalar summary of the statistical dispersion of ID, as two Lorenz curves can assume different shapes while still yielding the same Gini value. The Gini index is defined as twice the area between the 45-degree line and the Lorenz curve, which is insensitive to the changes in the Lorenz curve's shape. Similarly, the Hoover index (Hoover, 1936) is also derived from the Lorenz curve, and exhibits the same lack of uniqueness.

Thus far, many methods for modeling Lorenz curves have been proposed, either directly or indirectly through the modeling of statistical distribution functions of household income. In general, popular parametric methods for modeling the density of personal incomes rely on heavy tail distributions, including the Pareto (Pareto, 1964), log-normal (Gibrat, 1931), Weibull (Bartels and Van Metelen, 1975), gamma (Bartels and Van Metelen, 1975), and generalized beta distributions (McDonald, 1984; McDonald and Xu, 1995). Nonparametric methods include the commonly used empirical Lorenz curve estimation method and several other extensions that introduce various smoothing techniques (Ryu and Slottje, 1996; Cowell and Victoria-Feser, 2008). Most of these existing methods only focus on modeling a univariate personal ID. There is a need for the development of spatial functional data analysis techniques to jointly model Lorenz curves across counties or states in economic studies. Without spatial homogeneity pattern detection, each state needs to make its own policy, which could be a waste of public resources, while with a few clusters of states, only a policy for each cluster is needed.

There are several major challenges in developing clustering algorithms for spatial functional data. First, spatial functional data such as state-level Lorenz curves often exhibit strong location-related patterns. It is necessary to incorporate such spatial structure into spatial functional data clustering algorithms. Nevertheless, most existing functional clustering algorithms are designed under the assumption that the observed

functions are i.i.d curves (e.g., see a review paper by Jacques and Preda, 2014). These methods can be broadly divided into three categories: two-stage methods that reduce the dimension by using basis representations before applying clustering approaches, nonparametric methods that define specific dissimilarities among functions followed by heuristic or geometric procedures-based clustering algorithms such as K-means, and model-based methods that specify clustering models such as mixture of Gaussian for basis coefficients. Recently, several works have been proposed to extend these functional clustering algorithms to the spatial context. Romano et al. (2011) and Giraldo et al. (2012) followed the second path to define dissimilarities among spatial functions based on spatial variograms and cross-variograms. Jiang and Serban (2012) took the third path to model cluster memberships using an auto-regressive Markov random field, and introduce spatially dependent random errors in the conditional model for functions.

Second, certain spatial contiguous constraints on the clustering configuration are desired to facilitate interpretations in the spatial context. In other words, a local cluster is expected to include spatially connected components with flexible shapes and sizes. In addition, in many economics applications, this spatial contiguity constraint may not dominate the clustering configuration globally, in the sense that two clusters that are spatially disconnected may still belong to the same cluster. For example, despite the distance between them, the New England area and California may share similar demographic information. Although a large body of model based spatial clustering approaches have been proposed in various spatial contexts, to the best of our knowledge, there is still a lack of clustering methods that allow for both locally spatially contiguous clusters and globally discontiguous clusters. Existing Bayesian spatial clustering methods based on mixture models, for example, the finite mixture model used in the aforementioned spatial functional clustering algorithm, can introduce spatial dependence in cluster memberships but do not fully guarantee spatial contiguity (Jiang and Serban, 2012). Suarez et al. (2016) used conditionally independent Dirichlet process priors to cluster each signal coefficient in a multiresolution wavelet basis. Among the methods that ensure spatial contiguity, some impose certain constraints on cluster shapes (Knorr-Held and Raßer, 2000; Kim et al., 2005; Lee et al., 2017), while others do not allow for globally discontinuous clusters (Li and Sang, 2019; Zhang et al., 2022).

Finally, how to determine the number of clusters is an important consideration in clustering. Most existing methods, such as Heaton et al. (2017), require the number of clusters to be specified first. Dirichlet Process mixture models (DPM) have grown in popularity in Bayesian statistics due to their flexibility in allowing for an unknown number of clusters. Recently, Miller and Harrison (2018) proved that DPM can produce an inconsistent estimate of the number of clusters, and proposed a mixture of finite mixtures model to resolve the issue while inheriting many appealing mathematical and computational properties of DPM. However, because this does not take into account any spatial information, their method may not be sufficient for spatial clustering.

In this article, we develop a new Bayesian nonparametric method that combines the ideas of Markov random field models and mixture of finite mixtures models to leverage geographical information to address these challenges when analyzing spatial income Lorenz curves. A distinction of the method is its ability to capture both locally spatially

contiguous clusters and globally discontiguous clusters. Furthermore, it employs an efficient Markov chain Monte Carlo (MCMC) algorithm to estimate the number of clusters and the clustering configuration simultaneously while avoiding complex reversible jump MCMC or allocation samplers. We apply this new Bayesian nonparametric clustering model to the analysis of the US state level household income Lorenz curves. In particular, we use a similarity measure among functional curves based on the inner product matrix under elastic shape analysis (Srivastava and Klassen, 2016), which has a nice invariance property under shape-preserving transformations. The findings of real-world data analysis reveal intriguing clustering patterns of IDs across states, which provide important information for studying regional income inequalities.

The remainder of this paper is structured as follows. Section 2 provides a detailed introduction to the motivating PUMS data. Section 3.1 provides a brief overview of elastic shape analysis of functions, followed by a discussion of nonparametric Bayesian clustering methods in Section 3.2. In Section 3.3, we describe the proposed Markov random field constrained mixture of finite mixtures prior model and introduce our functional data clustering model. Section 4 introduces Bayesian inference, which includes the MCMC sampling algorithm, the model selection criterion for parameter tuning, post-MCMC inference, and convergence diagnostic criteria. Sections 5 and 6 present a simulation and a case study using the PUMS data, respectively. Section 7 brings the paper to a close with some conclusions and discussions.

2 Motivating Data

Our motivating data comes from the PUMS data registry's 2018 submission. The incomes of US households and the states in which they reside are recorded for the 50 states plus Washington, DC. In the remainder of this paper, we will refer to them as "51 states". The Lorenz curve is a functional representation of income or wealth distribution that reflects inequality in wealth distribution. It, specifically, assumes that the household income x follows a cumulative distribution function (CDF) $F(x)$ with a corresponding probability density function $f(x)$. Let $Q(p) = F^{-1}(p)$ be the inverse CDF defined as $Q(p) = \inf\{y : F(y) \geq p\}$. The Lorenz curve is defined as

$$L(p) = \frac{1}{\mu} \int_0^p Q(t)\mathrm{d}t, \quad \text{for } 0 \leq p \leq 1,$$

where $\mu = \int_0^1 Q(t)\mathrm{d}t$. The Lorenz curve, when plotted in a graph, always starts at $(0, 0)$ and ends at $(1, 1)$, and measures for the poorest for $100p\%$ of households, what percentage $100L\%$ of total income they have.

In practice, the empirical Lorenz curve can be constructed from data in a similar fashion as constructing the empirical CDF. Define for state i

$$\widehat{L}_i(p_k) = \sum_{j=1}^{k} y_{i,(j)} / \sum_{j=1}^{T} y_{i,(j)},$$

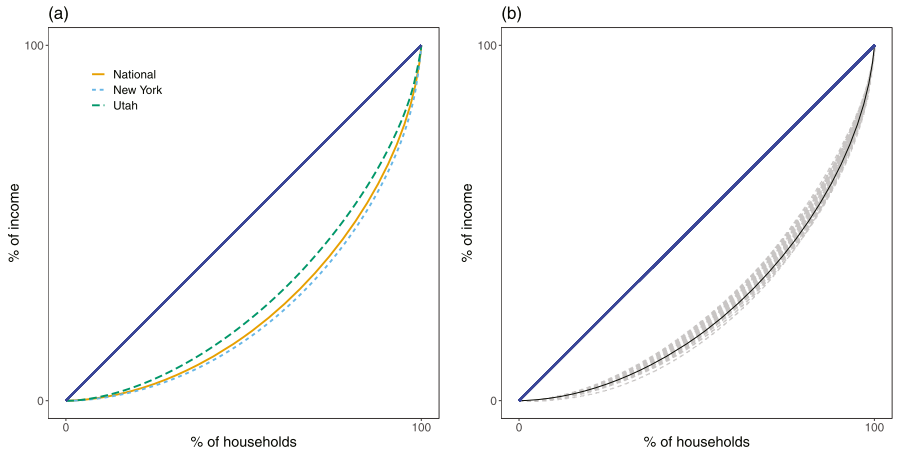

Figure 1: (a) Lorenz curves calculated based on the PUMS 2017 Household Income data on the national level and for two selected states; (b) Lorenz curves for all US states illustrated in dashed lines and the national curve in solid line.

where $p_k = k/n$, for $k = 1, \ldots, n$, and $y_{i,(j)}$ is the j-th order statistic observed in state i. Under mild regularity conditions, $\widehat{L}_i$ converges uniformly for $p \in [0, 1]$ to L_i with probability 1 (Gastwirth, 1972). As a derived measure, the Gini index is defined as two times the area between the Lorenz curve and the 45 degree line of equality from $(0, 0)$ to $(1, 1)$.

Figure 1 illustrates Lorenz curves computed with 2017 US household income data. The Lorenz curve calculated at the national level using all observations is plotted as a solid line, with a corresponding Gini coefficient of 0.4804. However, a closer examination of the state-level Lorenz curves reveals that the IDs do vary between states. Figure 1(a) also illustrates the Lorenz curves for two selected states, Utah and New York. While Utah's curve is above the national curve, indicating greater equality, New York's curve is below, suggesting a greater divide between rich and poor. In Figure 1(b), Lorenz curves for all US states are plotted together alongside the national curve, forming a "cloud" rather than being similar to each other. Lorenz curves' ability to describe income inequalities is clearly demonstrated here.

In addition to the Lorenz curves, descriptive statistics such as the Gini coefficient and state median income are presented in Figure 2. Utah has the lowest income inequality with a Gini coefficient of 0.423, while Washington, DC has the highest income inequality with a Gini coefficient of 0.512. Washington, DC also has the highest median income, $90,000, while Mississippi has the lowest, $43,500.

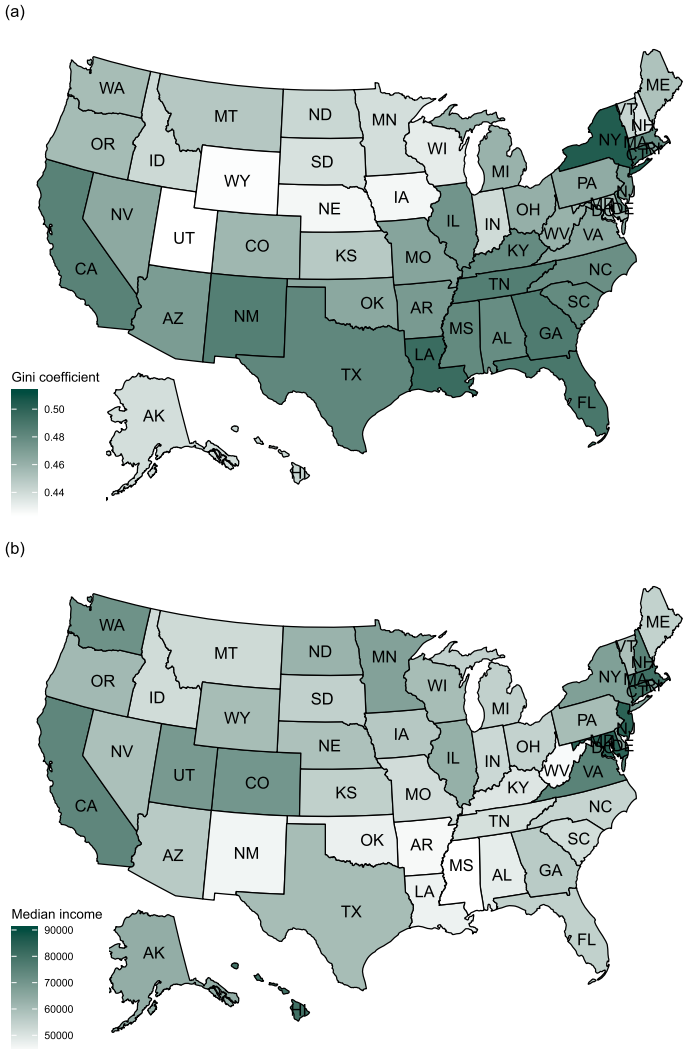

Figure 2: Descriptive statistics for PUMS data on the US map to visualize economic inequality: (a) Gini coefficient; (b) state median income.

3 Methodology

In this section, we treat the state-level Lorenz curves as spatial functional data. We begin by discussing the functional representation of ID and the shape-based similarity measure between two IDs. Next, a nonparametric Bayesian approach based on the similarity measure is introduced for functional data clustering. Additionally, a Markov random fields constraint mixture of finite mixtures model (MRFC-MFM) is proposed

incorporating spatial constraints into the clustering prior. The hierarchical model under MRFC-MFM is presented at the end of this section.

3.1 Functional Representation of Income Distribution

We begin the section by reviewing functional data shape analysis. To cluster functional data, proper metrics need to be defined for quantifying similarity between functional curves. Functional data has four critical characteristics: quantity, frequency, similarity, and smoothness. Commonly used distance measures, such as the Euclidean distance, are no longer suitable for assessing similarities between functions. In this article, we consider the inner product matrix calculated using a specific representation of curves called the square-root velocity function (SRVF; Srivastava et al., 2010). This inner product matrix is a summary statistic that encapsulates information about the similarity between curves for subsequent clustering analysis. It places a greater emphasis on the distinctions in the shape of functions. By focusing on shapes, one is more concerned with the numbers of and relative heights of peaks and valleys in a curve than with their precise locations. This feature is more appropriate for assessing the variations in IDs between regions, as exact locations or mean shifts have a smaller effect on ID inequality.

The SRVF of an absolutely continuous function $f(t) : [0,1] \to \mathcal{R}^p$ is defined as:

$$q(t) = \text{sign}\big(f'(t)\big)\sqrt{|f'(t)|}, \tag{3.1}$$

where $f'(t)$ is the derivative of f. It can be seen that the SRVF is a curve of unit length. There are a number of advantages to employing the SRVF to analyze functional data. First, the scaling, rotation and re-parameterization variabilities still remain based on the SRVF. In addition, the elastic metric is invariant to function reparameterization. The SRVF represents unit-length curves as a unit hypersphere in the Hilbert manifold. The SRVF for a given function can be obtained in R using the `f_to_srvf()` function provided by the **fdasrvf** package (Tucker, 2019). For given functions f_1 and f_2 which belong to $\mathcal{F} = \{f : [0,1] \to \mathcal{R}^p : f \text{ is absolutely continuous}\}$ and their corresponding SRVFs, q_1 and q_2, the inner product is defined based on the definition in Zhang et al. (2015) as follows:

$$S_{f_1,f_2} = \sup_{\gamma\in\Gamma, O\in SO(p)} \langle q_1, (q_2, (O,\gamma))\rangle\,, \tag{3.2}$$

where $SO(p)$ is the collection of orthogonal $p \times p$, i.e. p-dimensional rotation matrices, and Γ represents the set of all orientation-preserving diffeomorphisms over the domain [0,1]. The notation (O,γ) denotes a joint action of the rotation and reparameterization operations, and $(q_2,(O,\gamma))$ here represents a particular reparameterization and rotation of q_2. The maximization over $SO(p)$ and Γ can be performed iteratively as in Srivastava et al. (2010). The operation $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathbb{L}^2([0,1],\mathbb{R}^p)$: $\langle v,q\rangle = \int_0^1 \langle v(t),q(t)\rangle \mathrm{d}t$. The value of this integral ranges from -1 to 1, achieving the value -1 if $v = -q$ and 1 if $v = q$. The inner product of two functions can be calculated using the algorithm in Tucker et al. (2013). Computation in R is carried out with the `trapz()` function in package **pracma** (Borchers, 2019). Given $f_1,\ldots,f_n$, the $n \times n$ pairwise inner product matrix $\boldsymbol{S}$ can be calculated using the definition in (3.2), and **fdasrvf** and **pracma**.

3.2 Mixture of Finite Mixtures for Distances Between Functional Data

Next, we introduce nonparametric Bayesian methods to capture spatial homogeneity of functional data. We start with a Fisher's Z-transformation of the inner product matrix $\boldsymbol{S}$ to transform each entry $\boldsymbol{S}_{f_i,f_j} \in (-1,1)$ to the real line. The transformed inner product matrix is denoted as $\boldsymbol{\mathcal{S}}$, with each entry being

$$\mathcal{S}_{ij} = \frac{1}{2}\log\left(\frac{1+\boldsymbol{S}_{f_i,f_j}}{1-\boldsymbol{S}_{f_i,f_j}}\right).$$

The larger $\mathcal{S}_{ij}$ is, the closer f_i and f_j are. We further assume that

$$\begin{aligned} &\mathcal{S}_{ij} \mid \boldsymbol{\mu}, \boldsymbol{\tau}, k \sim \mathrm{N}(\mu_{ij}, \tau_{ij}^{-1}), \quad \mu_{ij} = U_{z_i z_j}, \\ &\tau_{ij} = T_{z_i z_j}, \quad 1 \le i \le j \le n, \end{aligned} \tag{3.3}$$

where k is the number of true underlying clusters, N() denotes the normal distribution, $z_i \in \{1,\ldots,k\}$ denotes the cluster membership of the i-th curve; $\boldsymbol{U} = [U_{rs}] \in (-\infty,+\infty)^{k\times k}$ and $\boldsymbol{T} = [T_{rs}] \in (0,+\infty)^{k\times k}$ are symmetric matrices, with $U_{rs} = U_{sr}$ indicating the mean closeness of any function f_i in cluster r and any function f_j in cluster s, and $T_{rs} = T_{sr}$ indicating the precision of closeness between any function f_i in cluster r and any function f_j in cluster s. Note that in the above formulation, only the upper triangle of matrix $\boldsymbol{\mathcal{S}}$ is modeled, including the diagonal.

Let $\mathcal{Z}_{n,k} = \left\{(z_1,\ldots,z_n) : z_i \in \{1,\ldots,k\}, 1 \le i \le n\right\}$ denote all possible partitions of n nodes into k clusters. Given $z \in \mathcal{Z}_{n,k}$, let $\boldsymbol{\mathcal{S}}_{[\boldsymbol{rs}]}$ denote the $n_r \times n_s$ sub-matrix of $\boldsymbol{\mathcal{S}}$ consisting of entries $\mathcal{S}_{ij}$ with $z_i = r$ and $z_j = s$. Following the common practice for stochastic block models (SBM; Holland et al., 1983), independence between entries of $\boldsymbol{\mathcal{S}}$, or edges, is assumed. The joint likelihood of $\boldsymbol{\mathcal{S}}$ under model (3.3) can be expressed as

$$\begin{aligned} P(\boldsymbol{\mathcal{S}} \mid \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}, k) &= \prod_{1\le r\le s\le k} P(\boldsymbol{\mathcal{S}}_{[\boldsymbol{rs}]} \mid \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}), \\ P(\boldsymbol{\mathcal{S}}_{[\boldsymbol{rs}]} \mid \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}) &= \prod_{1\le i<j\le n: z_i=r, z_j=s} \frac{1}{\sqrt{2\pi T_{rs}^{-1}}} \exp\left\{-\frac{T_{rs}(\mathcal{S}_{ij}-U_{rs})^2}{2}\right\}. \end{aligned} \tag{3.4}$$

A common Bayesian specification when k is given can be completed by assigning independent priors to $\boldsymbol{z}$, $\boldsymbol{U}$ and $\boldsymbol{T}$, and it can be easily incorporated into the framework of a finite mixture model. As k is unknown, a frequent technique is to introduce Dirichlet process mixture prior models (Antoniak, 1974) as follows:

$$\boldsymbol{\mathcal{S}}_i \sim F(\cdot, \boldsymbol{\theta}_i), \quad \boldsymbol{\theta}_i \sim G(\cdot), \quad G \sim DP(\alpha G_0), \tag{3.5}$$

where $\boldsymbol{\mathcal{S}}_{\boldsymbol{i}} = (\mathcal{S}_{i1}, \mathcal{S}_{i2}, \ldots, \mathcal{S}_{in})$, $\boldsymbol{\theta}_i = (\boldsymbol{\theta}_{i1}, \boldsymbol{\theta}_{i2}, \ldots, \boldsymbol{\theta}_{in})$ and $\boldsymbol{\theta}_{ij} = (\mu_{ij}, \tau_{ij})$.

The Dirichlet process (DP) is parameterized by a base measure G_0 and a concentration parameter α. If a set of values of $\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_n$ are drawn from G, a conditional prior

can be obtained by integration (Blackwell and MacQueen, 1973):

$$p(\boldsymbol{\theta}_{n+1} \mid \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \frac{1}{n+\alpha} \sum_{i=1}^{n} \delta_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_{n+1}) + \frac{\alpha}{n+\alpha} G_0(\boldsymbol{\theta}_{n+1}). \tag{3.6}$$

Here, $\delta_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_j) = I(\boldsymbol{\theta}_j = \boldsymbol{\theta}_i)$ is the distribution concentrated at a single point $\boldsymbol{\theta}_i$. Equivalent models can also be obtained by introducing cluster membership z_i's and letting the unknown number of clusters K go to infinity (Neal, 2000):

$$\begin{aligned} \boldsymbol{S}_i \mid z_i, \boldsymbol{\theta}^* &\sim F(\boldsymbol{\theta}^*_{z_i}), \\ z_i \mid \boldsymbol{\pi} &\sim \text{Discrete}(\pi_1, \ldots, \pi_K), \\ \boldsymbol{\theta}^*_c &\sim G_0, \\ \boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K), \end{aligned} \tag{3.7}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$. For each cluster c, the parameters $\boldsymbol{\theta}^*_c$ determine the cluster specific distribution $F(\cdot \mid \boldsymbol{\theta}^*_c)$.

By integrating out the mixing proportions $\boldsymbol{\pi}$, we can obtain the prior distribution of $(z_1, z_2, \ldots, z_n)$ that enables automatic inference on the number of clusters k. This is commonly known as the Chinese restaurant process (CRP; Aldous, 1985; Pitman, 1995; Neal, 2000). With the popular Chinese restaurant metaphor, z_i, $i = 2, \ldots, n$ are defined using the following conditional distribution (Pólya urn scheme, Blackwell and MacQueen, 1973):

$$P(z_i = c \mid z_1, \ldots, z_{i-1}) \propto \begin{cases} |c|, & \text{at an existing table labeled } c \\ \alpha, & \text{if c is a new table} \end{cases}, \tag{3.8}$$

where $|c|$ is the size of cluster c.

While the CRP has a very attractive feature of simultaneous estimation of the number of clusters and the cluster configuration, a significant shortcoming of this model was recently discovered. Miller and Harrison (2018) proved that the CRP produces extraneous clusters in the posterior, resulting in inconsistent estimation of the *number of clusters* even when the sample size approaches infinity. To address this issue, they proposed a modification to the CRP, known as the mixture of finite mixtures (MFM) model:

$$\begin{aligned} k &\sim p(\cdot), \quad (\pi_1, \ldots, \pi_k) \mid k \sim \text{Dirichlet}(\gamma, \ldots, \gamma), \\ z_i \mid k, \boldsymbol{\pi} &\sim \sum_{h=1}^{k} \pi_h \delta_h, \quad i = 1, \ldots, n, \end{aligned} \tag{3.9}$$

where $p(\cdot)$ is a proper probability mass function (p.m.f.) on $\{1, 2, \ldots, \}$ and δ_h is a point-mass at h. Compared to the CRP, introduction of new tables is slowed down by the factor $V_n(w+1)/V_n(w)$, allowing for model-based pruning of tiny superfluous clusters. The coefficient $V_n(w)$ is precomputed as:

$$V_n(w) = \sum_{k=1}^{+\infty} \frac{k_{(w)}}{(\gamma k)^{(n)}} p(k),$$

where $k_{(w)} = k(k-1)\dots(k-w+1)$, and $(\gamma k)^{(n)} = \gamma k(\gamma k+1)\dots(\gamma k+n-1)$. By convention, $x^{(0)} = 1$ and $x_{(0)} = 1$.

The conditional prior of $\boldsymbol{\theta}$ under the MFM is as follows:

$$P(\boldsymbol{\theta}_{n+1} \mid \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n) \propto \sum_{i=1}^{w} (n_i + \gamma)\delta_{\boldsymbol{\theta}_i^*} + \frac{V_n(w+1)}{V_n(w)}\gamma G_0(\boldsymbol{\theta}_{n+1}), \tag{3.10}$$

where $\boldsymbol{\theta}_1^*, \dots, \boldsymbol{\theta}_w^*$ are the distinct values taken by $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$, and w is the number of existing clusters. The cluster membership z_i, for $i = 2, \dots, n$, in (3.9) can be defined in a Pólya urn scheme similar to the CRP:

$$P(z_i = c \mid z_1, \dots, z_{i-1}) \propto \begin{cases} |c| + \gamma, & \text{at an existing table labeled } c \\ V_n(w+1)/V_n(w)\gamma, & \text{if } c \text{ is a new table} \end{cases}, \tag{3.11}$$

where w is the number of existing clusters.

Adapting MFM to our model setting for functional clustering, the model and prior can be expressed hierarchically as:

$$\begin{aligned}
&k \sim p(\cdot), \quad \text{where } p(\cdot) \text{ is a p.m.f. on } \{1,2,\dots\}, \\
&T_{rs} = T_{sr} \overset{\text{ind}}{\sim} \text{Gamma}(\alpha, \beta), \quad r, s = 1, \dots, k, \\
&U_{rs} = U_{sr} \overset{\text{ind}}{\sim} \text{N}(\mu_0, k_0^{-1} T_{rs}^{-1}), \quad r, s = 1, \dots, k, \\
&\text{pr}(z_i = j \mid \boldsymbol{\pi}, k) = \pi_j, \quad j = 1, \dots, k, \quad i = 1, \dots, n, \\
&\boldsymbol{\pi} \mid k \sim \text{Dirichlet}(\gamma, \dots, \gamma), \\
&\mathcal{S}_{ij} \mid \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}, k \overset{\text{ind}}{\sim} \text{N}(\mu_{ij}, \tau_{ij}^{-1}), \quad \mu_{ij} = U_{z_i z_j}, \quad \tau_{ij} = T_{z_i z_j}, \quad 1 \le i < j \le n.
\end{aligned} \tag{3.12}$$

We assume $p(\cdot)$ is a Poisson(1) distribution truncated to be positive through the rest of the paper, which has been proved by Miller and Harrison (2018) and Geng et al. (2019) to guarantee consistency for the mixing distribution and the number of clusters. We refer to the hierarchical model in (3.12) as MFM-fCluster.

3.3 Markov Random Field Constrained MFM in Functional Data

A possible weakness of MFM for spatial functional data is its failure to account for spatial structure or dependence, i.e., MFM ignores the spatial smoothness of a map, and hence the resulting clustering scheme does not comply with any spatial constraint, making it susceptible to noise in the data. This disadvantage can be overcome by introducing spatial coupling between adjacent features. Using a Markov random field prior in spatial statistical modeling is a classical Bayesian approach widely used in image segmentation problems (Geman and Geman, 1984). In this section, we use a similar idea of combining Markov random fields with MFM to introduce spatial constraints for clustering.

The Markov random field (MRF; Orbanz and Buhmann, 2008) provides a convenient approach to address the difficult problem of modeling a collection of dependent random

variables (Winkler, 2012). The dependence structure of different variables can be represented by a graph, with vertices representing random variables and an edge connecting two vertices indicating statistical dependence, which can conveniently introduce the spatial smoothness for both Gaussian and non-Gaussian data (e.g., mixture data). Interactions between variables are constrained to a small group that are typically assumed to be spatially closer, in order to reduce the complexity of the problem. The neighborhood dependence structure of a MRF is encoded by a weighted graph $\mathcal{N} = (V_{\mathcal{N}}, E_{\mathcal{N}}, W_{\mathcal{N}})$ in space, with vertices $V_{\mathcal{N}} = (v_1, \ldots, v_n)$ representing random variables at n spatial locations, $E_{\mathcal{N}}$ denoting a set of edges representing statistical dependence among vertices, and $W_{\mathcal{N}}$ denoting the edge weights representing the magnitudes of dependence.

The MRF for a collection of random variables $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ on a graph $\mathcal{N}$ has a valid joint distribution $M(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) := \frac{1}{Z_H}\exp\{-H(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)\}$, with H being the cost function with the following form

$$H(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) := \sum_{A \in \mathcal{C}_N} H_A(\boldsymbol{\theta}_A), \tag{3.13}$$

where $\mathcal{C}_N$ denotes the set of all cliques in $\mathcal{N}$, each term H_A is a non-negative function over the variables in clique A, and Z_H is a normalization term. By the Hammersley-Clifford Theorem, the corresponding conditional distributions enjoy the Markov property, i.e., $M(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}) = M(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{\partial(i)})$, where $\partial(i) := \{j \mid (i,j) \in E_{\mathcal{N}}\}$ denotes the set of neighbors of variable i. Considering only pairwise interactions, we model the conditional cost functions as

$$H(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}) := -\lambda \sum_{l \in \partial(i)} I(\boldsymbol{\theta}_l = \boldsymbol{\theta}_i) = -\lambda \sum_{l \in \partial(i)} I(z_l = z_i), \tag{3.14}$$

where $\lambda \in \mathbb{R}^+$ is a parameter controlling the magnitude of spatial smoothness, with larger values inducing stronger spatial smoothing. It can be seen that the function takes a value in $\{0, -\lambda\}$.

The Markov random field constrained MFM (MRFC-MFM) consists of an interaction term modeled by an MRF cost function that captures spatial interactions among vertices and a vertex-wise term modeled by an MFM. The resulting model defines a valid MRF distribution Π, which can be written as

$$\Pi(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) \propto P(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) M(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) \tag{3.15}$$

with $P(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ defined by the conditional distributions in (3.10) and $M(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ by the MRF model using (3.14) as the conditional cost function. As demonstrated in Theorem 3.1 below, this constrained model has a critical property: the MRF constraints affect only the finite component of the MFM model. The proof is deferred to Appendix 1.

Theorem 3.1. *Let $n_k^{(-i)}$ denote the size of the k-th cluster excluding θ_i, K^* denote the number of clusters excluding the i-th observation, and assume $H(\theta_i \mid \theta_{-i})$ is a valid MRF conditional cost function. The conditional distribution of a MRFC-MFM takes the form*

$$\Pi(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i}) \propto \sum_{k=1}^{K^*} (n_k^{(-i)} + \gamma) \frac{1}{Z_H} \exp(-H(\boldsymbol{\theta}_i \mid \boldsymbol{\theta}_{-i})) \delta_{\boldsymbol{\theta}_k^*}(\boldsymbol{\theta}_i) + \frac{V_n(K^*+1)}{V_n(K^*)} \frac{\gamma}{Z_H} G_0(\boldsymbol{\theta}_i).$$

An immediate corollary of Theorem 3.1 is the following Pólya urn scheme. Let z_i, $i = 1, \ldots, n$, denote the cluster memberships.

Corollary 1. *Suppose the conclusion of Theorem 3.1 holds. Then,*

$$\Pi(z_i = c \mid \boldsymbol{z}_{-i}) \propto \begin{cases} [|c| + \gamma] \exp[\lambda \sum_{l \in \partial(i)} I(z_l = z_i)], & \text{at an existing table labeled } c \\ V_n(K^* + 1)/V_n(K^*)\gamma, & \text{if } c \text{ is a new table} \end{cases},$$

where $\boldsymbol{z}_{-i} = \boldsymbol{z} \backslash \{z_i\}$*, i.e., all elements of* $\boldsymbol{z}$ *except for* z_i*.*

The above scheme offers an intuitive interpretation of MRFC-MFM again using the Chinese restaurant metaphor: the probability of customer i sitting at a table depends not only on the number of other customers already seated at that table, but also on the number of other customers that have spatial ties to the i-th customer. The parameter λ controls the strength of spatial ties, and ultimately, the number of clusters. The greater the value for λ, the stronger the spatial smoothing effect and the fewer clusters. This can be clearly observed in the simulation results presented in the sensitivity analysis section of the supplemental material (Hu et al., 2022). In particular, the MFM model developed in Miller and Harrison (2018) can be viewed as a special case of MRFC-MFM when $\lambda = 0$. We use the notation MRFC-MFM(λ, G_0) to represent the MRFC-MFM prior with smoothness parameter λ and base distribution G_0. The Markov random field constraint-mixture of finite mixture-functional clustering method (MRFC-MFM-fCluster) can be hierarchically written as

$$\begin{aligned} &\boldsymbol{U}, \boldsymbol{T}, \boldsymbol{z}, k \sim \text{MRFC-MFM}(\lambda, G_0), \\ &\mathcal{S}_{ij} \mid \boldsymbol{z}, \boldsymbol{U}, \boldsymbol{T}, k \overset{\text{ind}}{\sim} \text{N}(\mu_{ij}, \tau_{ij}^{-1}), \quad \mu_{ij} = U_{z_i z_j}, \tau_{ij} = T_{z_i z_j}, \quad 1 \le i < j \le n, \end{aligned} \tag{3.16}$$

where G_0 is a normal-gamma distribution whose hyperparameters are the same as in (3.12). While the model in (3.16) introduces spatial dependence to encourage locally contiguous clustering, it still allows any customer a chance to sit with any other customer, so that globally discontiguous clustering can be captured.

4 Bayesian Inference

MCMC is used to draw samples from the posterior distributions of the model parameters. In this section, we present the sampling scheme, the posterior inference of cluster configurations, and metrics to evaluate the estimation performance and clustering accuracy.

4.1 The MCMC Sampling Schemes

Our goal is to sample from the posterior distributions of the unknown parameters k, $\boldsymbol{z} = (z_1, \ldots, z_n) \in \{1, \ldots, k\}^n$, $\boldsymbol{U} = [U_{rs}] \in (-\infty, +\infty)^{k \times k}$ and $\boldsymbol{T} = [T_{rs}] \in (0, +\infty)^{k \times k}$. While methods such as reversible jump Markov chain Monte Carlo or even allocation samplers can be used for inference, they frequently suffer from poor mixing and slow

convergence. We extend Miller and Harrison (2018)'s approach to exploit the Pólya urn scheme for MRFC-MFM. Bayesian inference is performed using an efficient collapsed Gibbs sampler that analytically marginalizes out k. The sampler for MFM is presented in Algorithm 1 in the supplemental material, and the sampler for MRFC-MFM is presented in Algorithm 2 in the supplemental material. The only difference between the two algorithms lies in the posterior probability of an observation assigned to an existing cluster. Both algorithms efficiently cycle through the full conditional distributions of z_i given $\boldsymbol{z}_{-i}$, $\boldsymbol{U}$, and $\boldsymbol{T}$ for $i = 1, 2, \ldots, n$.

For the hyperparameters in both the simulation studies and the real data analysis, we use $\alpha = 1$, $\beta = 1$, $k_0 = 2$ and $\gamma = 1$. For μ_0, $\max_{i,j} \mathcal{S}_{ij}$ is assigned to diagonal terms and $\min_{i,j} \mathcal{S}_{ij}$ is assigned to off-diagonal terms in order to make it more informative. These choices for μ_0 ensure that the functions within a cluster are closer to each other than those in different clusters. We arbitrarily initialized the algorithms with nine clusters, and randomly allocated the cluster configurations. Various other choices were tested and we did not find any evidence of sensitivity to the initialization.

4.2 Post MCMC Inference

Dahl's method (Dahl, 2006) is a popular post-MCMC inference algorithm for the clustering configurations $\boldsymbol{z}$ and the estimated parameters. The inferences in Dahl's method are based on the membership matrices, $B^{(1)}, \ldots, B^{(M)}$, from the posterior samples. The membership matrix $B^{(t)}$ for the t-th post-burn-in MCMC iteration is defined as:

$$B^{(t)} = [B^{(t)}(i,j)]_{i,j\in\{1:n\}} = 1(z_i^{(t)} = z_j^{(t)})_{n\times n}, \quad t = 1, \ldots, M, \tag{4.1}$$

where $1()$ denotes the indicator function, i.e., $B^{(t)}(i,j) = 1$ indicates observations i and j are in the same cluster in the t-th posterior sample after burn-in iterations. Based on the membership matrices for the posterior samples, a Euclidean mean for membership matrices is calculated by:

$$\overline{B} = \frac{1}{M}\sum_{t=1}^{M} B^{(t)}.$$

The iteration with the least squared distance to $\overline{B}$ is obtained by

$$C_{LS} = \text{argmin}_{t\in(1:M)} \sum_{i=1}^{n}\sum_{j=1}^{n}\{B(i,j)^{(t)} - \overline{B}(i,j)\}^2. \tag{4.2}$$

The estimated parameters, together with the cluster assignments $\boldsymbol{z}$, are then extracted from the C_{LS}-th post burn-in iteration. An advantage of Dahl's method is that it utilizes the information in the empirical pairwise probability matrix $\overline{B}$.

A convergence diagnostic for the clustering algorithm is obtained using the Adjusted Rand index (ARI; Hubert and Arabie, 1985). As a chance-adjusted version of the Rand Index (RI; Rand, 1971), it measures the concordance between two clustering schemes. Taking values between 0 and 1, a large ARI value indicates high concordance. In particular, when two cluster configurations are identical, ARI takes the value 1.

4.3 Selection of λ

It is critical in our MRFC-MFM-fCluster algorithm to choose an appropriate value for λ, which controls the extent of spatial smoothness. The deviance information criterion (DIC; Spiegelhalter et al., 2002), the Bayesian equivalent of the Akaike information criterion (AIC; Akaike, 1973), has been one of the most commonly used model selection criteria under the Bayesian framework. However, the AIC does not exert enough penalization for clustering problems, which often leads to over-clustering results. Because of this, we suggest adopting a modified DIC (mDIC), in which the amount of the penalty in the traditional DIC is modified to match the Bayesian information criterion (BIC; Schwarz, 1978). The mDIC is calculated as

$$\text{mDIC} = \text{Dev}(\overline{\theta}) + \log(\frac{n \times (n+1)}{2}) p_D, \tag{4.3}$$

where

$$\text{Dev}(\theta) = -2 \log \prod_{1 \le i < j \le n : z_i = r, z_j = s} \frac{1}{\sqrt{2\pi T_{rs}^{-1}}} \exp\left\{ -\frac{T_{rs}(\mathcal{S}_{ij} - U_{rs})^2}{2} \right\},$$

$\theta = \{1 \le i < j \le n : z_i = r, z_j = s, U_{rs}, T_{rs}\}$, and

$$p_D = \overline{D(\theta)} - D(\overline{\theta}),$$

with $\overline{\theta}$ being the estimated parameters based on Dahl's method. The model with the smaller value of mDIC is preferred.

5 Simulation

In this section, we detail the simulation settings, the evaluation metrics, and the comparison performance results.

5.1 Simulation Setting and Evaluation Metrics

The spatial adjacency structure of the 51 states is used in the data simulation. We consider in total three partition settings with respective true number of clusters 3, 5, and 4. The first partition setting shown in Figure 3 contains a cluster which consists of two disjoint parts in the east and west. It is designed to mimic a fairly common economic pattern where geographically distant regions share similar ID patterns, and geographical proximity is not the only factor for determining homogeneity in ID. The second setting is the five-cluster partition shown in Figure 4. The third and final setting has four clusters, and the spatial constraint on clusters is "weaker" than under the other two designs, so that these clusters are composed of many spatially discontiguous states and regions, as shown in Figure 5.

Following Salem and Mount (1974), we generate $10,000$ simulated observations for each state from a Gamma distribution to mimic the long-tailed pattern commonly observed in econometrics data. In addition, to simulate the minor variations between IDs

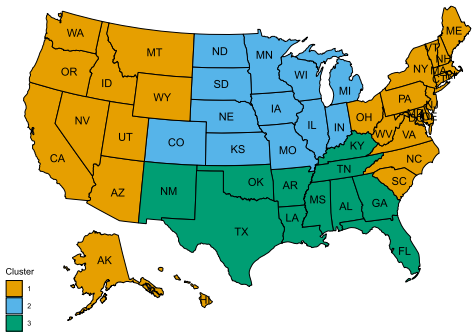

Figure 3: Illustration of the first partition setting with three true clusters, where the first cluster consists of two disjoint components.

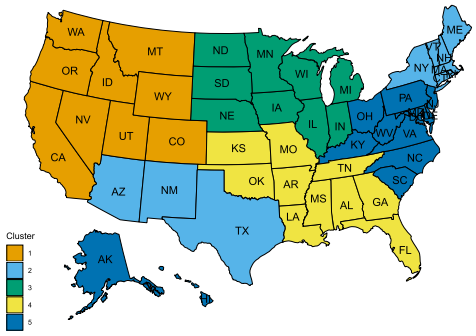

Figure 4: Illustration of the second partition setting with five clusters, where clusters 2 and 5 both have disjoint components.

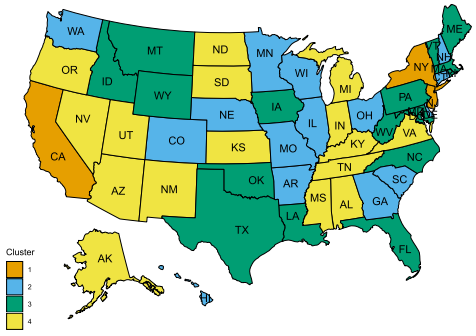

Figure 5: Illustration of the third partition setting with four clusters, where all the clusters are composed of disjoint states and regions.

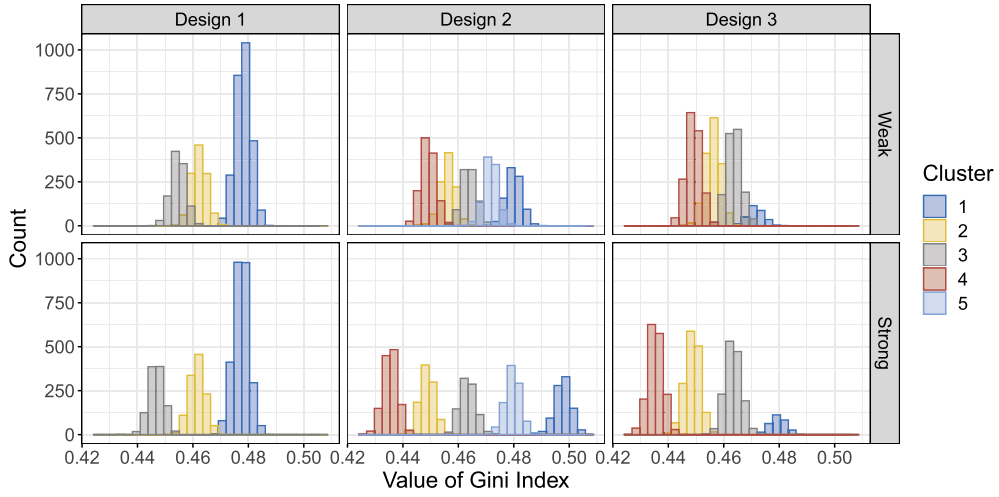

Figure 6: Histograms of Gini indices calculated from the simulated state-wise income data (5,100 in each panel from 100 replicates) for weak and strong signals under the three true partition settings.

of states within the same cluster, we add a noise term with probability 0.05 to each observation that also follows the Gamma distribution. We assume each cluster has its own set of distribution parameters shared by all the states within it. The true values of the parameters are set so that the Lorenz curves computed from the simulated data are comparable to those computed from real data (see Table 1). We consider two different parameter settings with small and large differences in income distributions between clusters, corresponding to weak and strong signal designs, respectively. For a total of 100 replicates, we show the Gini indices for different clusters of both weak and strong signal designs in Figure 6, which clearly exhibits major and minor overlapping among clusters, respectively.

The estimated number of clusters and ARI are used to evaluate the final clustering performance. ARI is calculated using the final clustering result obtained by Dahl's method for each replicate, and we calculate an average ARI over all replicates in each setting. The computation of ARI is carried out with the R package **mclust** (Scrucca et al., 2016). In each replicate of the simulation, the outcome is a clustering of the 51 states into several clusters. If the number of clusters for a replicate equals the true number of clusters (3 in the first design, 5 in the second, and 4 in the third), this replicate is counted towards one time that the number of clusters is correctly inferred. We report the total count of replicates with a correctly inferred number of clusters out of 100 replicates.

Design	Signal	Cluster	Design
Three Clusters			
	Weak	1	$\Gamma(1.15, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		2	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		3	$\Gamma(1.25, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
	Strong	1	$\Gamma(1.10, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		2	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		3	$\Gamma(1.30, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
Five Clusters			
	Weak	1	$\Gamma(1.10, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		2	$\Gamma(1.15, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		3	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		4	$\Gamma(1.25, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		5	$\Gamma(1.30, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
	Strong	1	$\Gamma(1.00, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		2	$\Gamma(1.10, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		3	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		4	$\Gamma(1.30, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		5	$\Gamma(1.40, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
Four Clusters			
	Weak	1	$\Gamma(1.15, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		2	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		3	$\Gamma(1.25, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
		4	$\Gamma(1.30, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.3, 50000)$
	Strong	1	$\Gamma(1.10, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		2	$\Gamma(1.20, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		3	$\Gamma(1.30, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$
		4	$\Gamma(1.40, 50000) + \text{Bin}(0.05) \cdot \Gamma(0.5, 50000)$

Table 1: Simulation designs with weak and strong signals. The symbol Γ denotes the Gamma distribution, and "Bin" denotes the binomial distribution.

5.2 Simulation Results

We first examine the inference results for the number of clusters, as well as the accuracy of clustering results from both MFM-fCluster and MRFC-MFM-fCluster. Each parameter setting listed in Table 1 is run with 100 replicates. For MRFC-MFM-fCluster, $\lambda \in \{0.5, 1, 1.5, 2, 2.5, 3\}$ are considered, and the best λ value is selected using mDIC within each replicate. The graph distance (GD; Bhattacharyya and Bickel, 2014) is used as the distance measure to construct the neighborhood graph used in the Markov random field model. Different upper limits of distance for two states to be considered as "neighbors" are used for the three designs. For the first three-cluster partition, the upper limit is set to 3. For the second five-cluster partition, however, due to the relatively small true cluster sizes, an upper limit of 1, i.e., only immediate neighbors, is adopted.

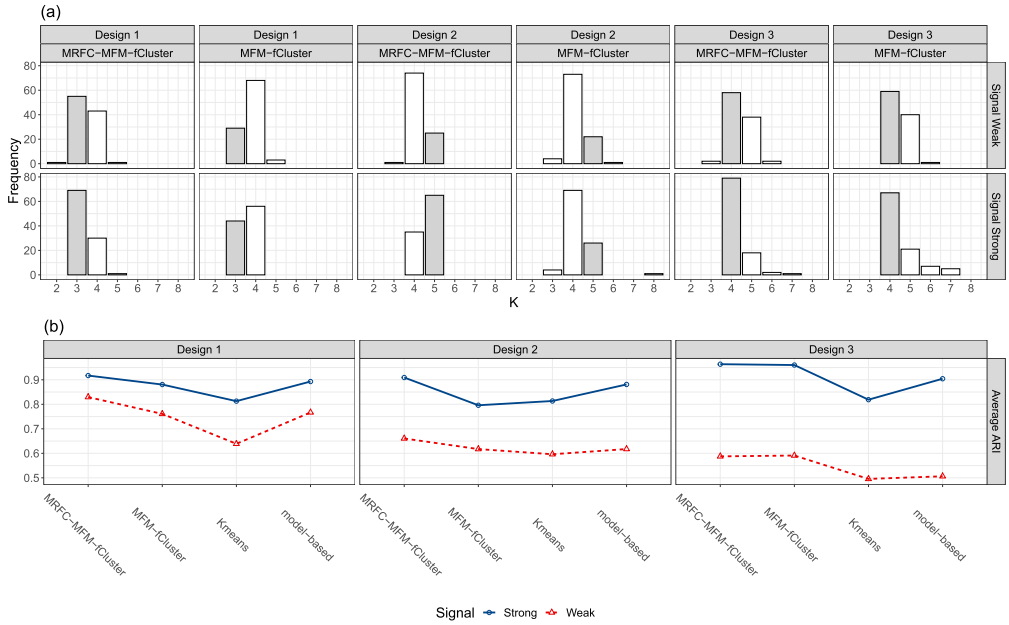

Figure 7: (a) Histogram of the number of clusters inferred by MRFC-MFM-fCluster and MFM-fCluster under different designs and signal strength settings. The grey bars correspond to the correct number of clusters. (b) Plot of the average ARI for all four methods under different designs and signal strength settings.

Similarly, for the third design, as the clusters have more disjoint components, and a state is more likely to have neighbors that belong to a different cluster from its own, we also use an upper limit of 1.

In addition to MFM-fCluster, we also consider two other competing methods. In the first competing method, we treat the SRVFs derived from Lorenz curves as vectors, and use K-means to cluster them. The second competing method is the model-based clustering for sparsely sampled functional data proposed by James and Sugar (2003), which is available in the R package **funcy**, and can be performed with the function `funcit()` using the option `method=''fitfclust''`. The clustering recovery performance of all three methods is measured using the ARI. For our proposed method, we present the average of the ARIs corresponding to the λ value selected by mDIC in each replicate. As neither K-means nor model-based clustering can estimate the number of clusters but instead require it to be provided, to make a fair comparison, for these methods we use the same number of clusters inferred for each replicate by the selected optimal λ.

Performances are visualized in Figure 7, and the average optimal λ's selected by mDIC are presented in Table 2. In Figure 7(a), it can be seen that under design 1, MFM-fCluster exhibits severe over-clustering, which produces four final clusters for more than 60 replicates under the weak signal setting, and more than 50 replicates in

	Design 1	Design 2	Design 3
Signal Weak	1.535	1.715	1.630
Signal Strong	1.680	1.630	1.610

Table 2: Average λ selected by mDIC for 100 simulation replicates for each combination of signal strength and true cluster design.

the strong signal setting. In contrast, even under the weak signal setting, MRFC-MFM-fCluster is able to correctly infer the true number of clusters for more than 50 replicates, and a notable number of 69 for the strong signal. Under design 2, as cluster sizes are relatively small, it is rather difficult for both MRFC-MFM-fCluster and MFM-fCluster to infer the number of clusters under the weak signal setting, as can be seen from the top middle two plots. With the strong signal, however, MRFC-MFM-fCluster is able to correctly identify the true number of clusters for more than half of the simulation replicates, while the performance of MFM-fCluster remains poor. Under design 3, as the true clusters are "messy" in the sense that there are no clear spatially contiguous states that belong to the same cluster, the performance of MRFC-MFM-fCluster is similar to that of MFM-fCluster in the weak signal setting. With the strong signal, however, MFM-fCluster again overclusters, producing for 67 replicates the correct K, while this number for MRFC-MFM-fCluster is 79. In Figure 7(b), our proposed method has the highest average ARI over 100 replicates for all six combinations of signal and partition design. The model-based functional clustering has the second best performance in designs 1 and 2, and the third best performance in design 3, while MFM-fCluster has the second best performance in design 3, and the third best in designs 1 and 2. In all cases, K-means performs the worst.

In addition, computation times for all methods are benchmarked using R package **microbenchmark** (Mersmann, 2019) on a desktop computer running Windows 10 Enterprise, with i7-8700K CPU@3.70GHz using single-core mode. A total of 20 replicates are performed to compute the average running time for each method. As expected, K-means takes the least time of 1.62 seconds due to its simple iterative algorithm. Unlike K-means which can only provide clusters without making statistical inference for cluster memberships and sizes, our proposed method utilizes conjugate forms for efficient Bayesian inference that provides not only estimates of clusters but also their uncertainty measures at only a slightly higher computation cost. Indeed, it takes on average 20.79 seconds for one simulated dataset with 500 MCMC iterations, as in our empirical studies 500 iterations are sufficient for the chain to converge and stabilize. The model-based approach, however, takes more than three minutes to finish. Due to the time-consuming nature of the model-based approach, the actual simulation studies are conducted on a 16-core desktop computer using parallel computation. The code is submitted for review and will be made publicly available at GitHub after the acceptance of the manuscript.

Finally, in the peer review process, an anonymous reviewer suggested the possibility of using more than a univariate measure of similarity. We have modified the code to take two similarity matrices, and run another 100 simulations under Design 1 to make a comparison of performance. Although the average ARIs remain over 0.9, the inference for K is rather poor. More details are included in Section 5 of the supplemental material.

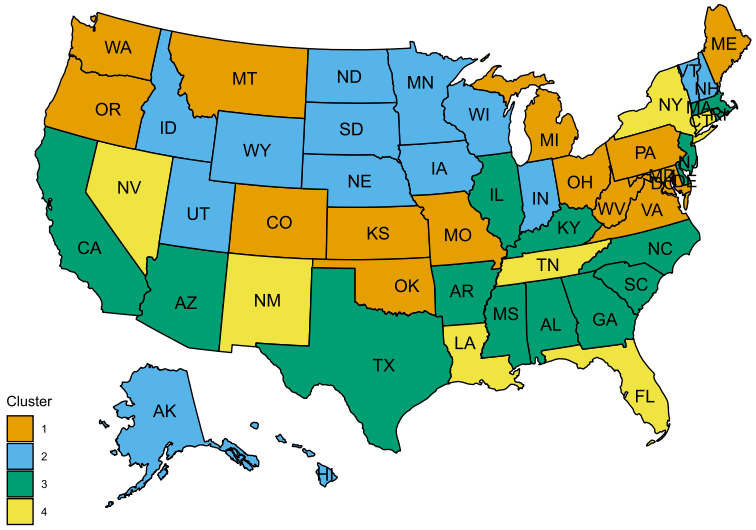

Figure 8: Illustration of the four clusters identified by the proposed method for the 51 states.

6 Analysis of PUMS Data

In this section, we apply the proposed MRFC-MFM-fCluster to a study of US household incomes in 2017. As with the simulation studies, the Lorenz curves for all states are obtained for the purpose of functional clustering. Based on (3.1) and (3.2), we get the inner product matrix $\boldsymbol{S}$. The spatial smoothing parameter λ is considered within the range of $\{0, 0.2, 0.4, \ldots, 3\}$, with $\lambda = 0$ corresponding to MFM-fCluster. The upper limit for considering states to be "neighbors" is considered within the range $\{1, 2, 3\}$. The mDIC is used to determine the optimal combination of these two parameters. From the sensitivity analyses presented in the supplemental material, the value of γ has only a marginal impact on the clustering performance. Therefore, it is set to be the same as in the simulation studies described in Section 4.1. In addition, consistent with the simulations, we choose $\alpha = 1$ and $\beta = 1$.

The final model with the smallest mDIC value has $\lambda = 0.8$ and upper limit 2 for defining neighbors. The final cluster configuration is visualized in Figure 8. There are, respectively, 14, 14, 15 and 8 states in clusters 1, 2, 3 and 4. Cluster 4 is the highest in terms of income inequality, and has an average Gini coefficient of 0.491. Cluster 2, with an average Gini of 0.435, exhibits the most equal income distribution among the four. Clusters 1 and 3 have average Gini values of 0.458 and 0.477.

A significant advantage of our proposed method is that it allows for globally discontinuous clusters. As illustrated in Figure 8, New Mexico and Tennessee are clustered together. Their 2017 Gini values are 0.4851 and 0.4858, respectively. Based on the 2010 American Community Survey from the U.S. Census Bureau (`https://factfinder.census.gov/`), their Gini coefficients have been historically very close. In addition,

there are several government policies that could be modified to affect certain clusters. For the states in Clusters 3 and 4, increasing the minimum wage and expanding the earned income tax credit are two strategies for improving the equality of ID. The median household incomes in the majority of Cluster 1 states are fairly low. Reduced income taxation will assist them in increasing their overall household income, at the cost of a little compromise in ID equality. Furthermore, an increase in government spending directly benefits the states in Cluster 1 by increasing household income. The states in Cluster 2 have the most balanced IDs and a mid-level median household income. Most of their government policies can be maintained to ensure sustainable economic growth. Another finding is that states with big urban regions tend to have less balanced IDs, which is consistent with the findings in Glassman and Branch (2017). According to Janikas et al. (2005) and Rey (2018), who evaluated income data from the Bureau of Economic Analysis, states with high (low) levels of internal inequality tend to be located near others with high (low) levels of internal inequality. This assertion corroborates the regional homogeneity tendencies for IDs across states identified by our method. Taking Clusters 1 and 2 as examples, these two clusters include a large number of neighboring states with low levels of internal inequality.

The posterior estimate of $\boldsymbol{U}$ in (3.16) is

$$\widehat{U} = \begin{pmatrix} 4.885 & 4.186 & 4.293 & 3.692 \\ 4.186 & 4.700 & 3.710 & 3.341 \\ 4.293 & 3.710 & 4.821 & 4.042 \\ 3.692 & 3.341 & 4.042 & 4.524 \end{pmatrix}. \tag{6.1}$$

It is noticeable that the diagonal entries of $\boldsymbol{U}$ are larger than the off-diagonal entries, which suggests the within-cluster similarity is much higher than between-cluster similarities. Cluster 1 has the least similarity to Cluster 4 based on (6.1), which is consistent with the results presented in Figure 9.

Finally, to make sure the cluster configuration presented here is not a random occurrence but reflects the true pattern demonstrated by the data, we ran 100 separate MCMC chains with different random seeds and initial values, and obtained 100 final clustering schemes. The RI between each scheme and the present clustering scheme in Figure 8 is calculated, and they average to 0.899, indicating high concordance of the conclusions regardless of random seeds. As suggested by a reviewer, we also use the sequentially-allocated latent structure optimization (SALSO) algorithm implemented in the R package **salso** (Dahl, 2020) to check for uncertainties in the presented clustering result. The details are included in Section 6 of the supplemental material.

7 Discussion

In this paper, we proposed both MFM-fCluster and MRFC-MFM-fCluster to capture spatial homogeneity of ID using the functional inner product of Lorenz curves. Parameter tuning is achieved using a modified version of DIC, the popular Bayesian model selection criterion. Extensive simulation studies demonstrate that MRFC-MFM outperforms the traditional MFM model in the pursuit of spatial homogeneity. It also outperforms K-means and model-based methods across a range of designs, and the comparison

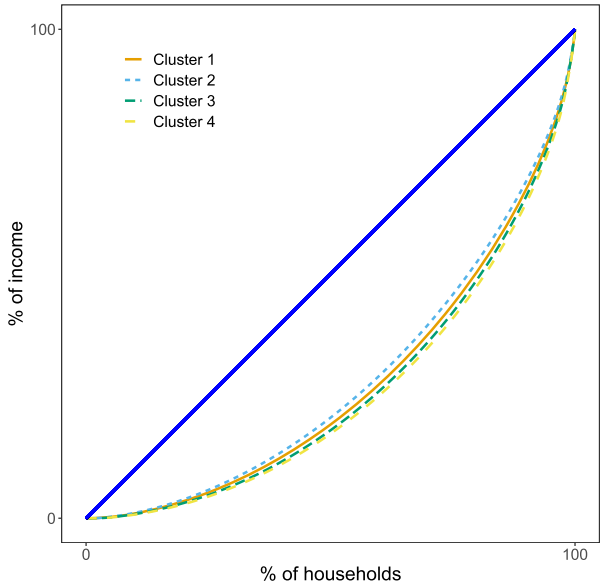

Figure 9: Average Lorenz curves for states in the four identified clusters.

of performance is relatively robust with respect to choices of the spatial smoothing parameters. A case study using the PUMS data reveals a number of important findings for IDs across the 51 states in the US. The results shown in Section 6 indicate that states that are near each other tend to be clustered together by our method, which is reasonable as they often share similar demographic information as well as median income, tax rate and urbanization. The results provide valuable insights to both residents and governors: residents could gain a better understanding of their state's conditions, and hence vote with their feet accordingly; governors, equipped with more *objective* and principled analysis of IDs, could make better data-informed policy design decisions.

A few topics beyond the scope of this paper are worth further investigation. In this paper, Fisher's Z-transformation of the inner product matrix is used. Modeling the original inner product matrix is an interesting alternative in future work. Independence is assumed between elements of the inner product matrix $\boldsymbol{S}$ for modeling and computational simplicity and convenience, and extending SBM to incorporate such edge dependence similar to Yuan and Qu (2018) is an interesting but nontrivial problem for future research. In addition, tuning of λ is criterion-based. Treating it as an unknown parameter and proposing a prior in a hierarchical model for it may improve the efficiency. Besides the geographical information, other auxiliary covariates, such as demographic information, could also be taken into account for clustering in our future work. While our clustering methods are based on a similarity matrix or dissimilarity matrix, the proposed MRFC-MFM clustering prior model can be adapted to other hierarchical model settings, including the case with multiple similarity matrices as responses (Paul et al., 2016; Lei et al., 2020) Extending our prior on functional data model with basis coeffi-

cients (Suarez et al., 2016) is also another interesting direction for future work. Finally, nonstationarity is an important consideration of spatial dependence. Considering a nonstationary cost function in our clustering process has the potential of broadening the applications of our proposed methods, and is worth further investigation.

Supplementary Material

Supplementary to "Bayesian Spatial Homogeneity Pursuit of Functional Data: an Application to the U.S. Income Distribution" (DOI: 10.1214/22-BA1320SUPP; .pdf).

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle." In Petrov, B. N. and Csaki, F. (eds.), *Second International Symposium on Information Theory*, 267–281. Akadémiai Kiado. MR0483125. 592

Aldous, D. J. (1985). "Exchangeability and Related Topics." In *École d'Été de Probabilités de Saint-Flour XIII—1983*, 1–198. Springer. MR0883646. doi: https://doi.org/10.1007/BFb0099421. 587

Antoniak, C. E. (1974). "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems." *The Annals of Statistics*, 2(6): 1152–1174. MR0365969. 586

Bartels, C. P. and Van Metelen, H. (1975). *Alternative Probability Density Functions of Income: A Comparison of the Lognormal-, Gamma-and Weibull-distribution with Dutch Data*. Vrije Universiteit, Economische Faculteit. 580

Bhattacharyya, S. and Bickel, P. J. (2014). "Community Detection in Networks Using Graph Distance." *arXiv preprint arXiv:1401.3915*. 595

Blackwell, D. and MacQueen, J. B. (1973). "Ferguson Distributions via Pólya Urn Schemes." *The Annals of Statistics*, 1(2): 353–355. MR0362614. 587

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions*. R package version 2.2.9. URL https://CRAN.R-project.org/package=pracma. 585

Cowell, F. A. and Victoria-Feser, M.-P. (2008). "Modelling Lorenz Curves: Robust and Semi-parametric Issues." In *Modeling Income Distributions and Lorenz Curves*, 241–253. Springer. 580

Dahl, D. B. (2006). "Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model." In Kim-Anh Do, M. V., Peter Müller (ed.), *Bayesian Inference for Gene Expression and Proteomics*, volume 4, 201–218. Cambridge University Press. MR2706330. 591

Dahl, D. B. (2020). *salso: Sequentially-Allocated Latent Structure Optimization*. R package version 0.1.16. URL https://CRAN.R-project.org/package=salso. 599

Gastwirth, J. L. (1972). "The Estimation of the Lorenz Curve and Gini Index." *The Review of Economics and Statistics*, 54(3): 306–316. MR0314429. 583

Geman, S. and Geman, D. (1984). "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6): 721–741. 588

Geng, J., Bhattacharya, A., and Pati, D. (2019). "Probabilistic Community Detection with Unknown Number of Communities." *Journal of the American Statistical Association*, 114(526): 893–905. MR3963189. doi: https://doi.org/10.1080/01621459.2018.1458618. 588

Gibrat, R. (1931). *Les inégalités économiques*. Recueil Sirey. 580

Gini, C. (1997). "Concentration and Dependency Ratios." *Rivista di Politica Economica*, 87: 769–792. 580

Giraldo, R., Delicado, P., and Mateu, J. (2012). "Hierarchical Clustering of Spatially Correlated Functional Data." *Statistica Neerlandica*, 66(4): 403–421. MR2983302. doi: https://doi.org/10.1111/j.1467-9574.2012.00522.x. 581

Glassman, B. and Branch, P. S. (2017). "Income Inequality Among Regions and Metropolitan Statistical Areas: 2005 to 2015." Technical report, U.S. Census Bureau. 599

Heaton, M. J., Christensen, W. F., and Terres, M. A. (2017). "Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences." *Technometrics*, 59(1): 93–101. MR3604192. doi: https://doi.org/10.1080/00401706.2015.1102763. 581

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). "Stochastic Blockmodels: First Steps." *Social Networks*, 5(2): 109–137. MR0718088. doi: https://doi.org/10.1016/0378-8733(83)90021-7. 586

Hoover, E. M. (1936). "The Measurement of Industrial Localization." *The Review of Economic Statistics*, 162–171. 580

Hu, G., Geng, J., Xue, Y., and Sang, H. (2022). "Supplementary to "Bayesian Spatial Homogeneity Pursuit of Functional Data: an Application to the U.S. Income Distribution"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1320SUPP. 590

Hubert, L. and Arabie, P. (1985). "Comparing Partitions." *Journal of Classification*, 2(1): 193–218. 591

Jacques, J. and Preda, C. (2014). "Functional Data Clustering: A Survey." *Advances in Data Analysis and Classification*, 8(3): 231–255. MR3253859. doi: https://doi.org/10.1007/s11634-013-0158-y. 581

James, G. M. and Sugar, C. A. (2003). "Clustering for Sparsely Sampled Functional Data." *Journal of the American Statistical Association*, 98(462): 397–408. MR1995716. doi: https://doi.org/10.1198/016214503000189. 596

Janikas, M. V., Rey, S. J., et al. (2005). "Spatial clustering, inequality and income convergence." *Région et Développement*, 21(2): 45–64. 599

Jiang, H. and Serban, N. (2012). "Clustering Random Curves Under Spatial Interdependence with Application to Service Accessibility." *Technometrics*, 54(2): 108–119. MR2929427. doi: https://doi.org/10.1080/00401706.2012.657106. 581

Kim, H.-M., Mallick, B. K., and Holmes, C. C. (2005). "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes." *Journal of the American Statistical Association*, 100(470): 653–668. MR2160567. doi: https://doi.org/10.1198/016214504000002014. 581

Knorr-Held, L. and Raßer, G. (2000). "Bayesian Detection of Clusters and Discontinuities in Disease Maps." *Biometrics*, 56(1): 13–21. 581

Lee, J., Gangnon, R. E., and Zhu, J. (2017). "Cluster Detection of Spatial Regression Coefficients." *Statistics in Medicine*, 36(7): 1118–1133. MR3621013. doi: https://doi.org/10.1002/sim.7172. 581

Lei, J., Chen, K., and Lynch, B. (2020). "Consistent community detection in multi-layer network data." *Biometrika*, 107(1): 61–73. MR4064140. doi: https://doi.org/10.1093/biomet/asz068. 600

Li, F. and Sang, H. (2019). "Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets." *Journal of the American Statistical Association*, 114(527): 1050–1062. MR4011757. doi: https://doi.org/10.1080/01621459.2018.1529595. 581

Lorenz, M. O. (1905). "Methods of Measuring the Concentration of Wealth." *Publications of the American Statistical Association*, 9(70): 209–219. 580

Mankiw, N. G. (2014). *Principles of Economics*. Cengage Learning. 580

McDonald, J. B. (1984). "Some Generalized Functions for the Size Distribution of Income." *Econometrica: Journal of the Econometric Society*, 52(3): 647–663. 580

McDonald, J. B. and Xu, Y. J. (1995). "A Generalization of the Beta Distribution with Applications." *Journal of Econometrics*, 66(1): 133–152. 580

Mersmann, O. (2019). *microbenchmark: Accurate Timing Functions*. R package version 1.4-7. URL https://CRAN.R-project.org/package=microbenchmark. 597

Miller, J. W. and Harrison, M. T. (2018). "Mixture Models with a Prior on the Number of Components." *Journal of the American Statistical Association*, 113(521): 340–356. MR3803469. doi: https://doi.org/10.1080/01621459.2016.1255636. 581, 587, 588, 590, 591

Neal, R. M. (2000). "Markov Chain Sampling Methods for Dirichlet Process Mixture Models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: https://doi.org/10.2307/1390653. 587

Orbanz, P. and Buhmann, J. M. (2008). "Nonparametric Bayesian Image Segmentation." *International Journal of Computer Vision*, 77(1-3): 25–45. 588

O'sullivan, A. and Sheffrin, S. M. (2007). *Prentice Hall Economics: Principles in Action*. Pearson/Prentice Hall. 579

Pareto, V. (1964). *Cours d'économie Politique*, volume 1. Librairie Droz. 580

Paul, S., Chen, Y., et al. (2016). "Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel." *Electronic Journal of Statistics*, 10(2): 3807–3870. MR3579677. doi: https://doi.org/10.1214/16-EJS1211. 600

Pitman, J. (1995). "Exchangeable and Partially Exchangeable Random Partitions." *Probability Theory and Related Fields*, 102(2): 145–158. MR1337249. doi: https://doi.org/10.1007/BF01213386. 587

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods." *Journal of the American Statistical Association*, 66(336): 846–850. 591

Rey, S. J. (2018). "Bells in space: The spatial dynamics of US interpersonal and inter-regional income inequality." *International Regional Science Review*, 41(2): 152–182. 599

Romano, E., Verde, R., and Cozza, V. (2011). "Clustering Spatial Functional Data: A Method Based on a Nonparametric Variogram Estimation." In Ingrassia, S., Rocci, R., and Vichi, M. (eds.), *New Perspectives in Statistical Modeling and Data Analysis*, 339–346. Springer. MR3051225. doi: https://doi.org/10.1007/978-3-642-11363-5_38. 581

Ryu, H. K. and Slottje, D. J. (1996). "Two Flexible Functional Form Approaches for Approximating the Lorenz Curve." *Journal of Econometrics*, 72(1-2): 251–274. 580

Salem, A. B. and Mount, T. D. (1974). "A Convenient Descriptive Model of Income Distribution: the Gamma Density." *Econometrica: Journal of the Econometric Society*, 42(6): 1115–1127. 592

Schwarz, G. (1978). "Estimating the Dimension of a Model." *The Annals of Statistics*, 6(2): 461–464. MR0468014. 592

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). "mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal*, 8(1): 289–317. doi: https://doi.org/10.32614/RJ-2016-021. 594

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society: Series B (Methodological)*, 64(4): 583–639. MR1979380. doi: https://doi.org/10.1111/1467-9868.00353. 592

Srivastava, A., Klassen, E., Joshi, S. H., and Jermyn, I. H. (2010). "Shape Analysis of Elastic Curves in Euclidean Spaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(7): 1415–1428. 585

Srivastava, A. and Klassen, E. P. (2016). *Functional and Shape Data Analysis*. Springer. MR3821566. 582

Suarez, A. J., Ghosal, S., et al. (2016). "Bayesian Clustering of Functional Data using Local Features." *Bayesian Analysis*, 11(1): 71–98. MR3447092. doi: https://doi.org/10.1214/14-BA925. 581, 601

Tucker, J. D. (2019). *fdasrvf: Elastic Functional Data Analysis*. R package version 1.9.2. URL https://CRAN.R-project.org/package=fdasrvf. 585

Tucker, J. D., Wu, W., and Srivastava, A. (2013). "Generative Models for Functional Data Using Phase and Amplitude Separation." *Computational Statistics & Data Analysis*, 61: 50–66. MR3063000. doi: https://doi.org/10.1016/j.csda.2012.12.001. 585

Winkler, G. (2012). *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*, volume 27. Springer Science & Business Media. MR1950762. doi: https://doi.org/10.1007/978-3-642-55760-6. 589

Yuan, Y. and Qu, A. (2018). "Community Detection with Dependent Connectivity." *arXiv preprint arXiv:1812.06406*. MR4319255. doi: https://doi.org/10.1214/20-aos2042. 600

Zhang, B., Sang, H., Luo, Z., and Huang, H. (2022). "Bayesian clustering of spatial functional data with application to a human mobility study during COVID-19." *The Annals of Applied Statistics*. Forthcoming. 581

Zhang, Z., Pati, D., and Srivastava, A. (2015). "Bayesian Clustering of Shapes of Curves." *Journal of Statistical Planning and Inference*, 166: 171–186. MR3390142. doi: https://doi.org/10.1016/j.jspi.2015.04.007. 585

Acknowledgments

The authors would like to thank the editor in chief, the editor, the associate editor, and two reviewers for their valuable comments, which helped improve the presentation of this paper. In addition, the authors thank Dr. Fred Huffer for his help in writing.