RIETI Discussion Paper Series 17-E-015

A Functional Linear Regression Model in the Space of Probability Density Functions

ARATA Yoshiyuki
RIETI

The Research Institute of Economy, Trade and Industry
http://www.rieti.go.jp/en/

RIETI Discussion Paper Series 17-E-015
March 2017

A Functional Linear Regression Model in the Space of Probability Density Functions

ARATA Yoshiyuki*†

Research Institute of Economy, Trade and Industry

Abstract

In this paper, we propose a functional linear regression model in the space of probability density functions. We treat a cross-sectional distribution of individual earnings as an infinite dimensional random variable. By an isometric transformation of density functions, the constrained nature of density functions is explicitly taken into account. Then, we introduce a regression model where the income distribution is a dependent variable. Asymptotic results for the significance test statistics of the coefficients are obtained. Applying this method to Japanese data, we figure out a functional relationship of the income distribution with economic growth. It is found that the change in income distribution associated with economic growth is characterized by a disproportional increase in the lower income class, reduction of the middle income earners, and irresponsiveness of the higher income earners. Since the information that the income distribution offers is preserved as a density function, this method enables us to obtain implications ignored by the usual statistical ones.

Keywords**:** Functional data analysis, Bayes space, Income distribution, Inequality

JEL Classification Numbers**:** C14; C13; D31

*Address: 1-3-1 Kasumigaseki Chiyoda-ku, Tokyo, 100-8901 Japan. E-mail:*arata-yoshiyuki@rieti.go.jp*.

†This research was conducted as a part of the Project 'Sustainable Growth and Macroeconomic Policy' undertaken at the Research Institute of Economy, Trade and Industry (RIETI). This paper utilizes the micro data of the "Basic Survey on Wage Structure" provided by the Ministry of Health, Labour and Welfare (MHLW) and the "Family Income and Expenditure Survey" provided by the Statistics Bureau, Ministry of Internal Affairs and Communications (MIC). I would like to thank Yuichi Ikeda, Irena Vodenska, Hiroshi Iyetomi, Takashi Unayama, Kazufumi Yamana, and Hiroshi Yoshikawa, as well as many colleagues, especially Shota Araki, Keisuke Kondo, Willem Thorbecke, Masayuki Morikawa, and Makoto Yano for helpful and constructive suggestions. I am grateful to participants in seminars at Niigata University, University of Tokyo and RIETI for useful comments.

1 Introduction

How does the shape of the distribution of income change when an economy falls into recession? Does economic growth contribute to reduce inequality? These questions have been one of the oldest subjects of economic inquiry, which are relevant for economists as well as policy makers. In fact, since at least the seminal paper by Kuznets (1955), there have been a huge volume of studies examining the relationship between the income distribution and economic growth. (e.g., Ahluwalia (1976), Alesina and Rodrik (1994), Barro (2000), and Galor and Moav (2004), just to name a few). The empirical methodology used in this line of research is regression-based models whose dependent variable is an inequality index (e.g., the Gini coefficient or the top 1% income share) or parameters characterizing a parametric distribution, though the choice varies from author to author. However, the use of a particular index may yield misleading results because it captures only one of the aspects of the income distribution and, therefore, the implications of the analysis may depend on the choice of the index. Inequality indexes do not necessarily display the same behavior and the interdependence among them is not clear, making the interpretation difficult and ambiguous.[1] Likewise, the parametric approach of assuming a functional form has similar problems because the choice of a parametric model is equivalent to imposing restrictions on the behavior of the income distribution. In other words, the effects of macroeconomic variables on the income distribution estimated in this approach are, at least partially, the consequence of the assumed functional form.

To overcome these deficiencies, we propose a functional linear regression model whose dependent variable is the income distribution. By treating the income distribution as an infinite-dimensional random variable, we figure out its functional relationship with economic growth. It enables us to extract the influence of economic growth on the income distribution without assuming a parametric distribution. This approach is advantageous because the information that the distribution has is not reduced to a small number of indexes or parameters but fully used in estimation.

The idea of dealing with probability density functions as random variables is not new in the existing literature. For example, Kneip and Utikal (2001) apply functional data analysis (FDA, see Section 2) to income and age distributions, and decompose them into principle components.

[1] For example, Acemoglu and Robinson (2015) shows that the top 1% share of income behaves quite differently from other measures of inequality in the cases of the South Africa and Sweden. They write, "the share of national income going to the top 0.1 percent or top 1 percent can give a distorted view of what is actually happening to inequality more broadly" (p.16).

Park and Qian (2012) propose a linear regression model where both dependent and explanatory variables are density functions. Chang et al. (2016) also develop functional principal component analysis (FPCA) and an unit roots test to analyze the nonstationarity in the time series of density functions. These studies assume that the space of probability density functions, denoted by F, is a proper subset of the space of square integrable functions, denoted by L^2, and treat a density function as an element of the L^2 space. However, recent studies, e.g., Delicado (2011), show that this approach is problematic. This is because elements in F have additional constraints, i.e., $\int f(x)dx = 1$ and $\forall x,\ f(x) \geq 0$ for $f \in F$. The behavior of density functions is constrained in F rather than in the L^2 space, which is ignored in this approach. It means that we seek the true value in the much larger space L^2 and, therefore, the accuracy of the estimation is substantially reduced. In fact, using artificial data, Delicado (2011) shows that this approach leads to results qualitatively different from the true model. In this paper, we take a different approach in that the space F is transformed based on the results given by Egozcue et al. (2006) and van den Boogaart et al. (2010, 2014). By this transformation, the two constraints are incorporated and the Hilbert structure is introduced into F. Based on this structure, we introduce a functional linear regression model and develop a statistical method to deal with a random variable in F.

We then apply this method to Japanese data and figure out the statistically significant relationship between the income distribution and economic growth, showing how the shape of the income distribution changes according to economic growth. In particular, since we can explicitly reveal the behavior of the density function, we can find the underlying behavior of the income distribution behind the variation in inequality indexes. Indeed, our analysis of Japanese income data shows that an increase in the Gini coefficient associated with economic growth is caused by the variation of the left half of the income distribution corresponding to the low and middle-income earners. Interestingly, the right half of the probability distribution is stable with respect to economic growth. A similar pattern can be observed when we decompose our data by industry. This striking feature of the income distribution, which has not been addressed in the literature, can be revealed only by our method. Put differently, our method and findings can be seen as a new test for theoretical models explaining inequality and economic growth because the true model must approximate not only the behavior of a few indexes but also that of the income distribution. In this sense, our method gives a hint as to what characteristics future works should aim to explain.

The remainder of this paper is organized as follows. Section 2 discusses the structure of the space of density functions and introduces the Bayes space. Section 3 introduces a functional linear regression model in the Bayes space and gives our asymptotic results. Section 4 conducts simulation studies. Section 5 provides results of our implementation. Section 6 concludes. Mathematical concepts and a proof are collected in the Appendix.

2 Bayes Space

Over the last decades, a new statistical methodology called FDA has established itself as an emerging area of statistics. It offers new effective tools to deal with samples of random functions, developing functional versions for a wide range of standard statistical models.[2] Previous studies such as Kneip and Utikal (2001), Park and Qian (2012), and Chang et al. (2016) can be seen as attempts to apply FDA to probability density functions. In these studies, density functions are treated as elements of the L^2 space and the statistical methods developed for L^2 functions are applied to density functions. However, this approach has serious problems. While functions are assumed to be L^2 functions in FDA, density functions f have two additional constraints: $\int f(x)dx = 1$ and $f(x) \geq 0$, for all x. The direct application of FDA to probability density functions ignores this constrained nature.[3] There is no guarantee that the estimation of a density function becomes a density function; an estimated density function may take negative values, i.e., $\hat{f}(x) < 0$ for some x. In fact, using artificial data, Delicado (2011) compares the prediction by FPCA used in Kneip and Utikal (2001) and a true model from which the artificial data are drawn in order to examine the accuracy of FPCA. They show that the prediction by FPCA is not only quantitatively but also qualitatively different from the true model. Thus, the two constraints must be incorporated in estimation to overcome these difficulties.

In the finite-dimensional case, the statistical method known as compositional data analysis has

[2] For review, see Ramsay and Silverman (2005), Ferraty and Vieu (2006), and Horváth and Kokoszka (2012).

[3] In the previous studies above, a centered density function is defined as deviation from the mean: $w_t \equiv f_t - Ef_t$. Then, w_t is treated as a random variable taking value in the space H_w:

$$H_w \equiv \left\{ w \Big| \int_K w(x)dx = 0, \int_K w^2(x)dx < \infty \right\}$$

where K is a compact support. However, due to the nonnegativeness of density functions f_t, only random functions w_t in H_w such that $f_t(x) = w_t(x) + \mathbb{E}f_t(x) \geq 0$ for all x are allowed. Thus, the space in which the random function w_t takes values is much smaller than H_w.

been developed to deal with these constraints, where the Aitchison geometry is introduced into the space of compositional data.[4] Since the seminal paper by Egozcue et al. (2006), subsequent papers (e.g., van den Boogaart et al. (2010, 2014)) generalize this geometry to the space of probability density functions called the Bayes space, incorporating the constraints of density functions. Based on their results, Hron et al. (2016) reformulate FPCA for density functions. In line with these works, we consider a functional linear regression model in the Bayes space in this paper. Before introducing the regression model, we review the Bayes space and its properties used in later sections.

Let λ be the Lebesgue measure on a measurable space $(I,\ \mathfrak{B}(I))$, where $I \equiv [a,b] \subset \mathbb{R}$ and $\mathfrak{B}(I)$ is σ-algebra of I. We define the Bayes space B as the set of measures μ equivalent to λ, i.e., $\lambda(A) = 0 \Leftrightarrow \mu(A) = 0,\ \forall A \in \mathfrak{B}(I)$. By the Radon-Nikodým theorem, an element of B is represented by its density $f_\mu = d\mu/d\lambda$. We introduce an equivalence relation called B-equivalence denoted by $=_B$: for $f_\mu, f_\nu \in B$, $f_\mu =_B f_\nu$ iff $\exists c > 0$ such that $f_\mu(x) = c \cdot f_\nu(x)$, a.e. Here, we use the convention $c \cdot (+\infty) = +\infty$. Then, we define the perturbation and powering, denoted by $\oplus$ and $\odot$, respectively, as follows:

$$(f \oplus g)(x) =_B f(x)g(x), \quad (\alpha \odot f)(x) =_B (f(x))^\alpha, \text{ a.e.}$$

As we will discuss later, these operations play a role of the addition of two elements and the multiplication by $\alpha \in \mathbb{R}$, respectively.

Next, we define a subset of B:

Definition 1 (B^2 space of measures)

$$B^2 \equiv \big\{ f \in B \big| \int_I |\ln f(x)|^2 dx < +\infty \big\}.$$

An inner product in B^2 is defined as follows:

Definition 2 (Inner product in B^2)

$$\langle f, g \rangle_{B^2} = \Big[\frac{1}{2\eta} \int_I \int_I \ln \frac{f(x)}{f(y)} \ln \frac{g(x)}{g(y)} dx dy \Big]$$

[4] For compositional data analysis, see, e.g., Aitchison (1986).

where $\eta = b - a$.

The next definition is a transformation of density functions.

Definition 3 (Centered log-ratio (clr) transformation) *For* $f \in B^2$,

$$\mathrm{clr}(f)(x) \equiv f_c(x) \equiv \ln f(x) - \frac{1}{\eta}\int_I \ln f(z)dz.$$

With these setups, Egozcue et al. (2006) and van den Boogaart et al. (2014) show that B^2 is a separable Hilbert space with $\oplus$, $\odot$ and the inner product $\langle \cdot, \cdot \rangle_{B^2}$. The important point is that B^2 is isometric to the Hilbert space $L_0^2 \equiv \{f \in L^2 : \int_I f(x)dx = 0\}$ by the clr transformation:

$$\langle f, g \rangle_{B^2} = \langle f_c, g_c \rangle_2 \equiv \int_I f_c(z) g_c(z) dz \tag{1}$$

In other words, elements in B^2 are transformed into a more tractable space L_0^2 with the structure preserved. Note that the clr transformation preserves the linearity:

$$\mathrm{clr}((\alpha \odot f) \oplus g) = \alpha \mathrm{clr}(f) + \mathrm{clr}(g)$$

Thus, the perturbation and powering defined in the B^2 space correspond to the usual addition and multiplication in the L_0^2 space, respectively. Moreover, if $\{\psi_i\}_{i=0,1,\ldots}$ with $\psi_0(x) = const.$ is an orthonormal basis in L^2, then $\{\varphi_i \equiv \exp(\psi_i)\}_{i=1,2,\ldots}$ is an orthonormal basis in B^2. Namely, a density function f whose logarithm is expressed by $\log f = \sum_{k=0}^{\infty} \alpha_k \psi_k$ with $\sum_{k=0}^{\infty} |\alpha_k|^2 < +\infty$, has the following representation:[5]

$$f =_B \oplus_{k=1}^{\infty} \alpha_k \odot \varphi_k, \;\; \alpha_k = \langle f, \varphi_k \rangle_{B^2}$$

Note that $\int_I \psi_i(x)dx = 0$ for $i \geq 1$ because ψ_0 is a constant function. Since the clr transformation is an one-to-one mapping, clr has its inverse, i.e., $\mathrm{clr}^{-1}(h) =_B \exp(h)$ for $h \in L_0^2$. Using these

[5] $\oplus_{j=1}^{k} z_j$ is the summation of z_j, that is,

$$\oplus_{j=1}^{k} z_j = \underbrace{z_1 \oplus z_2 \oplus \ldots \oplus z_k}_{k}$$

structure of B^2, we build a functional linear regression model in this space.

Remark. From equation (1), the induced distance $d_{B^2}(f,g)$ in B^2 can be explicitly written as follows:

$$d_{B^2}(f,g) \equiv \sqrt{\langle f \ominus g \rangle_{B^2}} = \sqrt{\langle f_c - g_c \rangle_2} = \left[\frac{1}{2\eta}\int_I\int_I\left(\ln\frac{f(x)}{f(y)} - \ln\frac{g(x)}{g(y)}\right)^2 dxdy\right]^{1/2}. \tag{2}$$

This distance has an important connection with parametric estimation. As a family of parametric distributions, consider the exponential family (denoted by Exp_I). Delicado (2011) shows that the distance for two elements in Exp_I is equivalent to the Euclidean distance between parameter values:[6] if $f, g \in Exp_I$,

$$d_{B^2}(f,g) = \|\boldsymbol{\theta}_f - \boldsymbol{\theta}_g\|_p,$$

where $\|\cdot\|_p$ is the usual Euclidean norm and $\boldsymbol{\theta}_f$ and $\boldsymbol{\theta}_g$ are parameters of f and g, respectively. Thus, the distance $d_{B^2}(\cdot,\cdot)$ can be seen as a natural generalization of $\|\cdot\|_p$ defined in the parametric space.

3 A Functional linear model in B^2

Based on the structure discussed above, we introduce a linear regression model in the B^2 space. In what follows, we consider the following functional linear regression model with a functional response f in B^2:

$$f = \beta_0 \oplus \oplus_{j=1}^k (x_j \odot \beta_j) \oplus \epsilon, \tag{4}$$

where we assume that $\boldsymbol{x} \in \mathbb{R}^k$, $\beta_j, \epsilon \in B^2$, $E[\epsilon] = \lambda$, and $\{\epsilon\}$ and $\{\boldsymbol{x}\}$ are two i.i.d. sequences and independent of each other. The full rank of $\Sigma_X = E[\boldsymbol{x}^T\boldsymbol{x}]$ are assumed. Note that the Lebesgue

[6] Density functions of the exponential family Exp_I are given by the following form:

$$f(x;\boldsymbol{\gamma}) = c(\boldsymbol{\gamma})g(x)\exp(\boldsymbol{\gamma}^T\mathbf{T}(\mathbf{x})), \quad \boldsymbol{\gamma}, \mathbf{T} \in \mathbb{R}^{\mathbf{p}}. \tag{3}$$

where $\boldsymbol{\gamma}$ is a parameter and $c(\boldsymbol{\gamma})$ is a normalization constant. Many of the commonly used distributions such as the normal, log-normal, gamma, and beta are included in Exp_I. Delicado (2011) shows that there exists a linear transformation matrix $\boldsymbol{M}$ such that $\boldsymbol{\theta} = \boldsymbol{\gamma}\boldsymbol{M}^{1/2}$ and $d_{B^2}(f,g) = \|\boldsymbol{\theta}_f - \boldsymbol{\theta}_g\|_p$.

measure λ is the neutral element in B^2, i.e., $g \oplus \lambda = g$ for $g \in B^2$. The meaning of equation (4) is essentially the same as the usual regression model; the dependent variable f is regressed on the explanatory variables $\boldsymbol{x}$ and the constant. For the convenience, set $y \equiv \mathrm{clr}(f)$, $e \equiv \mathrm{clr}(\epsilon)$, and $b_j \equiv \mathrm{clr}(\beta_j)$.

An important point in our estimation is the isometry from B^2 to L_0^2. Namely, an estimate $\hat{\boldsymbol{\beta}}$ minimizing the distance $d_{B^2}(f, \beta_0 \oplus \oplus_{j=1}^k (x_j \odot \beta_j))$ is equivalent to the least square estimate minimizing $d_2(y_i, b_0 + \sum_{j=1}^k x_{ij} b_j)$, where d_2 is the usual L^2-distance. Because of this property, we can deal with the regression model (4) in B^2 as if it were defined in L_0^2, which is easier to handle. Thus, in what follows, we consider a separable Hilbert space L_0^2, instead of B^2. By the clr transformation, equation (4) can be written in matrix form:

$$\mathbf{y}(t) = \mathbf{X}\boldsymbol{b}(t) + \mathbf{e}(t) = \mathbf{X_1}\boldsymbol{b_1}(t) + \mathbf{X_2}\boldsymbol{b_2}(t) + \mathbf{e}(t)$$

where $\mathbf{y} = \{y_i\}_{i=1,\ldots,N}$ and N is sample size. For later purpose, $\mathbf{X}$ is decomposed into two part, $\mathbf{X_1}$ and $\mathbf{X_2}$, and $\mathbf{X_2}$ has L columns. Note that the least square estimate $\hat{\boldsymbol{b}} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \in L_0^2$ because the estimate is a linear combination of elements of L_0^2. Since the clr transformation is invertible, coefficients of interest in B^2 are recovered by clr^{-1}.

In the remainder of this section, we present our method to test the significance of the coefficients. We begin with the definitions of the covariance operator $C_e : L_0^2 \to L_0^2$ and the covariant function $c : [a,b]^2 \to \mathbb{R}$ as follows:

$$C_e \equiv E[\langle e_i, \cdot \rangle e_i], \quad c(t,s) \equiv E[e_i(t)e_i(s)].$$

It should be noted that $C_e(z)$ can be written as

$$C_e(z)(t) = \int c(t,s)z(s)ds.$$

The operator C_e is determined by c. The Mercer lemma (see, e.g., p.24 in Bosq (2000)) shows that if $c(t,s)$ is continuous over $[a,b]^2$, then $c(t,s)$ can be expressed in the following form:[7]

$$c(t,s) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(t) \varphi_n(s), \tag{5}$$

where $\{\lambda_n\}$ is a decreasing sequence of positive values and $\{\varphi_n\}$ is a sequence of continuous functions such that

$$\int \varphi_n(s) \varphi_m(s) ds = \delta_{n,m}, \;\; n, m \in \mathbb{N}.$$

Note that $\int \varphi_n(s) ds = 0$ for positive λ_n.[8] Thus ,the $\{\varphi_n\}$ can be extended to form an orthonormal basis in L_0^2 by adding an orthonormal basis in the subspace complement to the subspace spanned by the original $\{\varphi_n\}$.

With these setups, we have the following proposition:

Proposition 4 *Set* $\Lambda \equiv \sum_{i=1}^{N} \|y_i(t) - \mathbf{x}_{1,i}\hat{\mathbf{b}}_1\|_2^2 - \sum_{i=1}^{N} \|y_i(t) - \mathbf{x}_i \hat{\mathbf{b}}\|_2^2$. *Under the null hypothesis that* $\mathbf{b}_2 = 0$, $\Lambda \xrightarrow{d} \sum_{n=1}^{\infty} \lambda_n \chi_n^2(L)$, *where* $\chi_n^2(L)$ *is an i.i.d. chi-squared random variable with* L *degrees of freedom.*

Proof. See the Appendix. ■

Using this proposition, we can test the significance of our coefficients. We also use the bootstrap method to numerically calculate the distribution of Λ under the null hypothesis.[9] We carry out the so-called residual resampling method here. The first step is to calculate the coefficient $\hat{\mathbf{b}}_1$ under the null hypothesis: the regression of y on $\mathbf{x}_1$. We obtain the estimates $\hat{y} = \mathbf{x}_1 \hat{\mathbf{b}}_1$ and the residuals e_0 from this fit. Next, data are simulated N times by

$$y^* = \mathbf{x}_1 \hat{\mathbf{b}}_1 + e^*,$$

[7]For simplicity, the continuity of $c(t,s)$ is assumed here. Since the discontinuity of probability density functions is not our focus, this assumption is innocent.

[8]Now assume conversely, that is, that $\exists n$, $\lambda_n > 0$ and $\int \varphi_n(s) ds \neq 0$. Since $\int c(t,s) ds = \int E[e_i(t) e_i(s)] ds = 0$, the Mercer lemma implies that $0 = \sum_{m=1}^{\infty} a_m \varphi_m$, where $a_m \equiv \lambda_m \int \varphi_m(s) ds$. Since $a_n \neq 0$, we have

$$\varphi_n = \frac{\sum_{m=1, m \neq n}^{\infty} a_m \varphi_m}{a_n}.$$

This means that $\{v_n\}$ are not orthogonal, leading to contradiction.

[9]Although a full understanding of the bootstrap application to functional models has not been reached yet, this subject has been discussed in the statistical literature. For example, see Cuevas et al. (2006).

where the N errors are sampled from the residuals e_0. Finally, we calculate Λ in Proposition 4 from bootstrap samples $(y_1^*, \mathbf{x}_1), (y_2^*, \mathbf{x}_2), ..., (y_N^*, \mathbf{x}_N)$. From these replicates of Λ, the one-sided p-value can be obtained.

4 Simulations

In this section, we explore the properties of our functional linear regression model developed above through a simulation study using the exponential family Exp_I.

4.1 Functional linear regression in the exponential family

In this subsection, we perform the regression analysis of simulated densities and compare the results with those of the usual FDA. The usual FDA means the direct application of functional linear regression analysis in L^2 to density functions, that is, density functions are assumed to be in L^2 and a functional linear model in L^2 is considered (the least squares estimate in L^2 is used). To this end, we use the exponential family Exp_I. As noted in Section 2, the Bayes space has an important relation with the exponential family.[10] Moreover, the density function (3) in Exp_I can be written as follows:

$$f(x; \boldsymbol{\gamma}) =_B g(x) \oplus \oplus_{j=1}^{k} \Big[\gamma_j \odot \exp(T_j(x))\Big]$$

Based on this structure, we generate pseudo samples following a linear model. That is, by viewing $\boldsymbol{\gamma}$ and $\mathbf{T}$ as random variables and coefficients, respectively, we generate n pseudo samples following a model $f = \beta_0 \oplus \oplus_{j=1}^{k} (x_j \odot \beta_j) \oplus \epsilon$, where $x_j = \gamma_j$ and $\beta_j = \exp(T_j)$. $\{e_i\}_{i=1,...,n}$ are i.i.d. random variables generated from an infinite-dimensional Gaussian distribution.[11] By regressing the density functions on the random parameters, we estimate the coefficients by each method.

Here, we consider the following two cases:

[10] For further information, see, e.g., van den Boogaart et al. (2010).

[11] An infinite-dimensional Gaussian random variable Z admits the following expansion:

$$Z \stackrel{d}{=} \sum_{n=1}^{\infty} \sqrt{\lambda_n} N_n \varphi_n, \tag{6}$$

where N_n is an independent standard Gaussian random variable. See the proof in the Appendix. Here, we fix $\sqrt{\lambda_n}$ to be $4/3 * 10^{-2}$ and use trigonometric series without the constant function as φ_n, and the sum is truncated up to its first 20 terms.

- Set 1: Lognormal distribution, $f_i^{Lognormal}(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp(-\frac{\ln^2(x)}{2\sigma_i^2})$. For simplicity, we set $s_i = 1/\sigma_i^2$. We note that the lognormal distribution corresponds to a 1-parametric exponential family with $\gamma_1 = s$, and $T_1(x) = -\ln^2(x)$. We first generate r_i from an uniform distribution over $[-1, 4]$ and put $s_i = \exp(r_i)$.
- Set 2: Gamma distribution, $f_{i,j}^{Gamma}(x) = \frac{\theta_j^{\kappa_i} x^{\kappa_i - 1}\exp(-x\theta_j)}{\Gamma(\kappa_i)}$. The gamma distribution corresponds to a 2-parametric exponential family with $\gamma_1 = \kappa$, $\gamma_2 = \theta$, $T_1(x) = \ln(x)$ and $T_2(x) = -x$. We let the explanatory variables κ_i and θ_j drawn from uniform distributions over the domain $[1.7, 4.5]$ and $[2.2, 5.0]$, respectively.

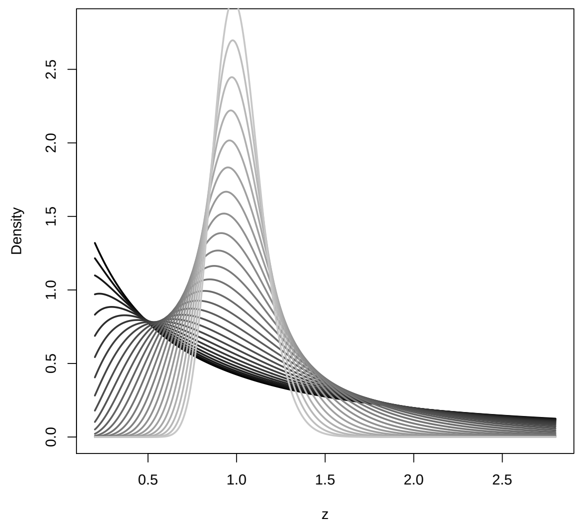
Adding infinite-dimensional Gaussian random variables to these densities, simulated densities to be used as the dependent variable f in the models are obtained.

Figure 1 shows the results on Set 1. As shown in Figure 1 (b), estimated densities by our method reproduce the original densities shown in Figure 1(a). On the other hand, FDA method reports very different results, as shown in Figure 1(c). Moreover, estimated densities in Figure 1(c) exhibit negative values for some x. This is caused by the fact that the sample space considered in FDA is the L^2 space and much larger than the Bayes space, even though the behavior of the density function is restricted within the space of density functions by definition. Namely, functions unnecessary for estimation are included in the sample space considered in FDA. Because of this, FDA is not guaranteed to yield proper density funcions and shows poor accuracy.

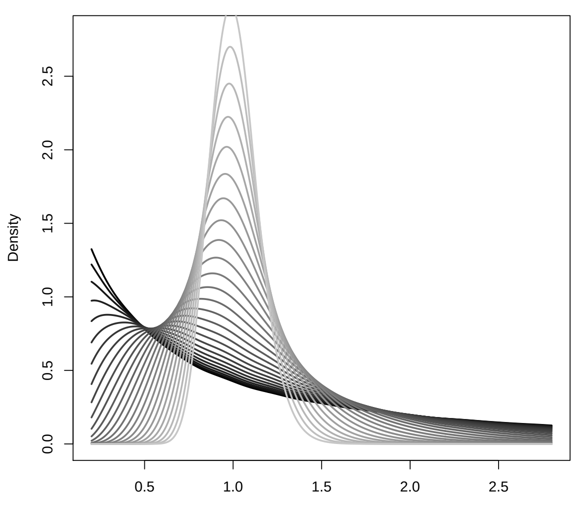
Figure 2 reports the results on Set 2, and the implications are essentially same as in Figure 1. As expected, while estimated densities by our method exactly reproduce the original ones, densities estimated by FDA show a very different picture and exhibit negative values. Especially when the income distribution is discussed as in the next section, the values of the density function are close to 0 for lower and higher incomes, and therefore, this problem is inevitable. In contrast, this problem can be avoided in our method because density functions are dealt within the space of the density functions.

4.2 Finite sample properties

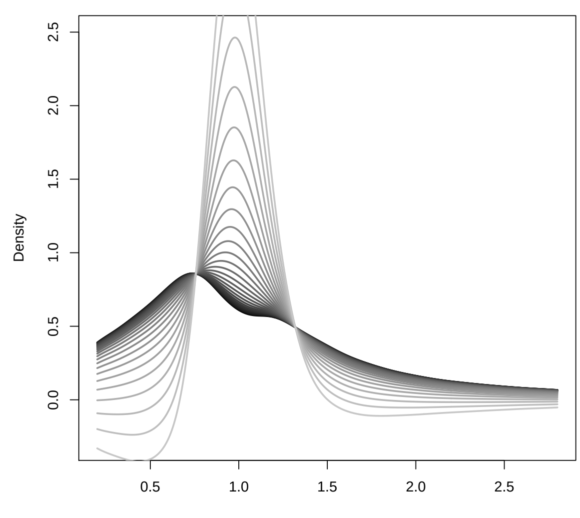
In this subsection, we examine the finite sample properties of Proposition 4. Here, we use the lognormal distribution as a true model: We generate n pseudo samples following $f = \beta_0 \oplus$



(a) Original densities.

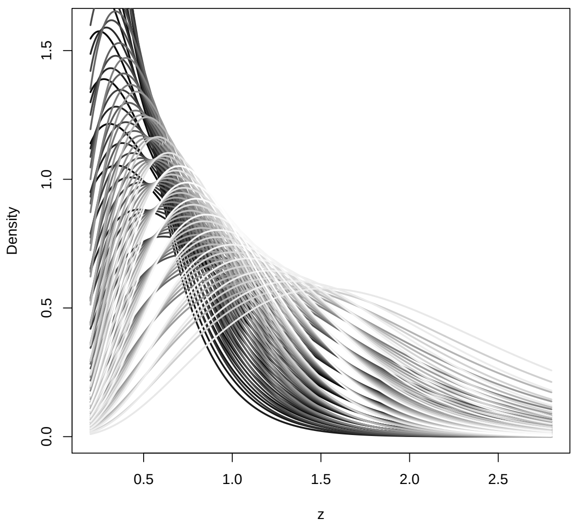


(b) Estimated densities by our method.

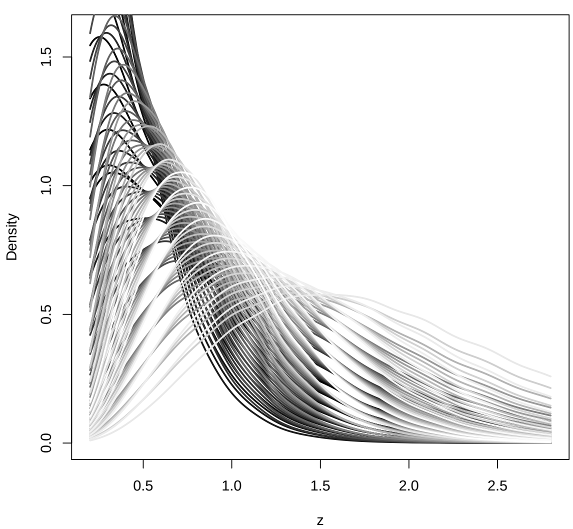


(c) Estimated densities by FDA.

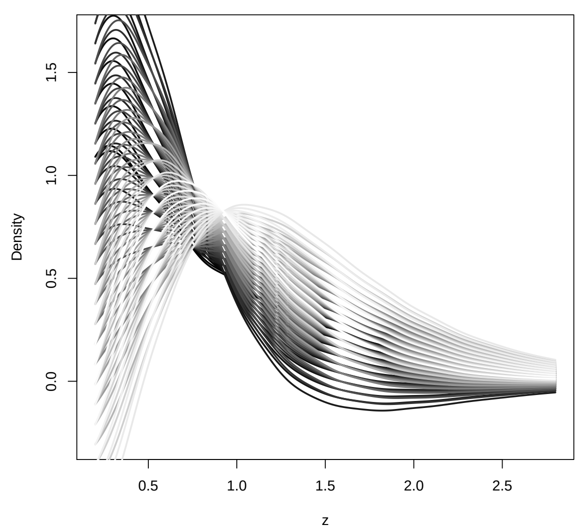
Figure 1: Lognormal distribution. In panel (a), densities of the lognormal distribution $f_i(x) = \frac{1}{\sqrt{2\pi}\sigma x}\exp(-\frac{\ln^2(x)}{2\sigma_i^2})$ with $1/\sigma_i^2 = \exp(-1.2 + 0.2i)$ for $i = 1, ..., 26$ are shown. Estimated densities in (b) and (c) means $\hat{f}$ based on the estimated coefficients by each method.



(a) Original densities.



(b) Estimated densities by our method.



(c) Estimated densities by FDA.

Figure 2: Gamma distribution. In panel (a), densities of the gamma distribution $f_{i,j}^{Gamma}(x) = \frac{\theta_j^{\kappa_i} x^{\kappa_i - 1} \exp(-x\theta_j)}{\Gamma(\kappa_i)}$ with $\kappa_i = 1.5 + 0.3i$, $\theta_j = 2 + 0.3j$, $i, j = 1, ..., 10$ are shown.

$(x_1 \odot \beta_1) \oplus \epsilon$, where $x_1 = 1/\sigma^2$ and $\beta_j = \exp(-\ln^2(x))$. We set $x_{1,i} = \exp(r_i)$, where r_i is drawn from an uniform distribution over $[0, 2]$. $\{e_i\}_{i=1,\ldots,n}$ are i.i.d. random variables generated from an infinite-dimensional Gaussian distribution.[12] Under these settings, we work on the null hypothesis H_0^1: $\beta_1 = \lambda$, which is equivalent to $b_1 = 0$ if the regression model is expressed in the L_0^2 space, i.e., $y = b_0 + b_1 x + e$. This procedure is repeated 1000 times and we count the number of rejection. Thus, the percentage of rejections under H_0^1 gives empirical power of our test.

For empirical size of our test, we generate n pseudo samples with $\beta_1 = \lambda$. x and ϵ are the same as before and the same procedure is repeated. The null hypothesis under these settings is denoted by H_0^2. The percentage of rejections gives empirical size of our test.

Both results are summarized in Table 1. We can see that almost all empirical sizes are close to the nominal values and approaches to them as N increases. It also suggests that even small sample sizes, e.g., $N = 25, 50$, show good power for our test in our settings.

Table 1: Empirical power and size. α is nominal size. All figures are in percentage of rejections.

	$n = 25$			$n = 50$			$n = 100$			$n = 500$		
α	1	5	10	1	5	10	1	5	10	1	5	10
H_0^1:	47.6	74.8	86.1	96.5	99.2	99.7	100	100	100	100	100	100
H_0^2:	0.0	2.8	8.1	.3	2.9	8.9	.7	4.4	8.8	.8	4.5	10.4

5 Implementation

We apply our statistical method to find a functional relationship between the income distribution and economic growth. Here, we use GDP growth rate as an indicator of economic growth. Since the information that the income distribution is not reduced to a few variables but preserved as a density function, our analysis is expected to reveal the relationship in more detail than by the usual regression analysis. Indeed, as we will see in the following, there is a striking feature of the behavior of the income distribution associated with economic growth, which has not been addressed in the economic literature.

[12] Infinite-dimensional Gaussian random variables are generated in the same manner as in the previous section, except $\sqrt{\lambda_n} = 2/3$ in equation (6).

5.1 Data Set

The data set used in our analysis is micro data from the Basic Survey on Wage Structure (BSWS) in Japan. The BSWS is a business establishment survey conducted annually by the Japanese government. Although the survey started in 1948, the data used in our analysis cover the period 1989–2014 due to data accessibility. The survey is conducted as of June of each year and includes monthly income before tax of individual workers including/excluding overtime payment. Regarding the bonus payment, the total bonus payment between 1 January and 31 December in the previous year is included in this survey. In addition to income, this survey provides us detailed information on individual workers such as age, sex, type of workers, working days/hours as well as on the establishment's attributes such as industry. In our analysis, we focus on ordinary workers' incomes including overtime and bonus payment. In BSWS, ordinary workers are defined as workers "to whom general scheduled working hours are applied," which exclude part-time workers whose "scheduled working hours a day or a week are less than those of general workers." Samples are drawn from almost all regions and industries except agriculture. A unique feature of this survey is its size: the number of samples is approximately 1.3 million workers (0.9 million of ordinary workers) per year. This is useful to our analysis because we have to estimate density functions from the micro data before carrying out our method. The huge size of data makes the estimation error negligible.

Since monthly income in June and annual bonus payment are recorded, income used in our analysis is defined as $monthly\ income + (1/12) * bonus\ payment$.[13] Relative income denoted by z is defined as income divided by average income, i.e., $z = \frac{income}{avg.\ income}$, which enables us to ignore price fluctuations and to focus on the change of the shape of the income distribution.[14] Fig 3 is a plot of estimated density functions of relative income z for 1989–2014, where the domain of z is restricted to $0.2 \leq z \leq 2.8$.[15] Here, kernel density estimation with the Gaussian kernel and Silverman's rule of thumb for the choice of the bandwidth is employed. Since density functions are treated as random variables taking values in the space of probability density functions, each density function shown in Figure 3 is a random sample f_i in equation (4).

[13] Exclusion of bonus payment makes no substantial change.

[14] Relative income is also used in the existing literature, e.g., Kneip and Utikal (2001).

[15] This excludes the top 1% and the bottom 2% of ordinary workers. This is done because the accuracy of the kernel density estimation is reduced by the scares of samples in the regions.

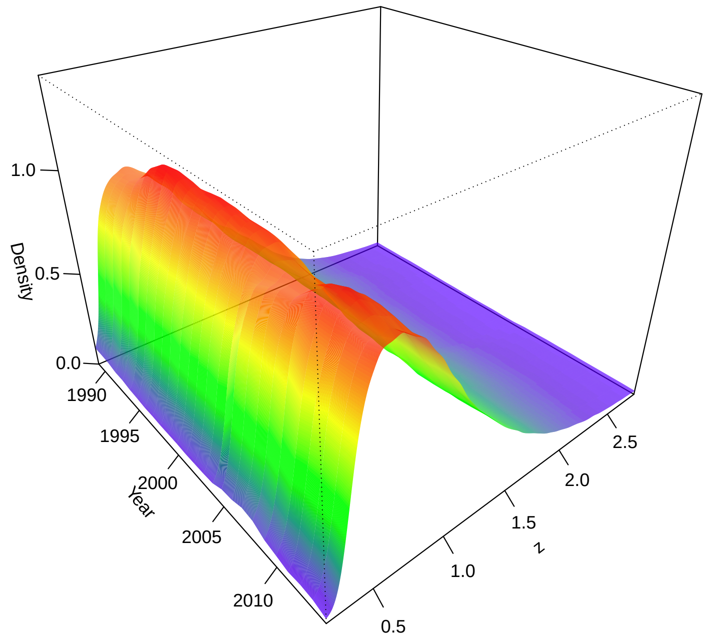


Figure 3: Density functions in 1989–2014.

5.2 Aggregate income distribution

First, treating the income distributions of all ordinary workers as a dependent variable, we estimate the effect of economic growth on the income distribution. Our regression model is as follows:

$$f = \beta_0 \oplus (x_1 \odot \beta_1) \oplus (x_2 \odot \beta_2) \oplus \epsilon, \tag{7}$$

where x_1 is the annual real GDP growth rate. x_2 is a dummy variable; 0 for the years 1989-2004, 1 for others, which is included to control the redesign of the survey in 2005.[16] Taking into account a time-lag in wage adjustments and the fact that this survey is conducted as of June of each year, we choose the GDP growth rate in the previous fiscal year.[17] Summary statistics of the GDP growth rates are given in Table 2.

We then estimate the coefficient $\hat{b}_1 = \text{clr}(\hat{\beta}_1)$, which is plotted in Figure 4(a). Using this estimate, we calculate $\bar{y} = \hat{b}_0 + \bar{x}_1 \cdot \hat{b}_1$ and $\hat{y}_\pm = \bar{y} \pm 2\sigma_1 \cdot \hat{b}_1$, where $\bar{x}_1$ and σ_1 are the mean

[16] For the purpose of getting more detailed information on non-regular workers, some appellations and questions asked in the survey were changed in 2005. Although the definition of ordinary workers remains the same, several studies have pointed out that this change causes a discontinuity between 2004 and 2005. See e.g., Lise et al. (2014).

[17] The time series of GDP are obtained from OECD data.

Table 2: Summary statistics of the GDP growth rate (%).

♯ of Obs.	Mean	Std. Dev.	Min	Max
26	1.54	2.54	-5.53	7.15

and the standard deviation of x_1, respectively. $\bar{y}$ and $\hat{y}_\pm$ are shown in Figure 4(b). The inverse transformation clr^{-1} of $\bar{y}$ and $\hat{y}_\pm$, i.e., probability density functions, are shown in Figure 4(c).

Next, we carry out the test for the significance of β_1 presented in Proposition 4. Our test statistics Λ is calculated to be .0447. We also calculate values of Λ under the null hypothesis: $\Lambda_{Chi,95} = .0271$, $\Lambda_{Chi,97} = .0322$, and $\Lambda_{Chi,99} = .0435$, where $\Lambda_{Chi,p}$ is the pth percentile value of Λ under the null hypothesis. Although the number of samples is small, this result suggests that the coefficient β_1 is statistically significant with p-value less than .01. In addition, the bootstrap result is presented in Figure 5(a). . From the histogram, we calculate the percentile values of Λ: $\Lambda_{B,95} = .0315$, $\Lambda_{B,97} = .0372$, and $\Lambda_{B,99} = .0541$. This shows that the coefficient is statistically significant with p-value $< .03$.

Finally, in order to find the relationship between the Gini coefficient and the GDP growth rate, we calculate the Gini coefficient from the estimate $\hat{y}_a = \hat{b}_0 + (\bar{x} + a \cdot \sigma_1) \cdot \hat{b}_1$, where a varies from -3 to 3. They are plotted in Figure 5, showing a positive relationship.[18] For economic interpretation of these results, see Section 5.4.

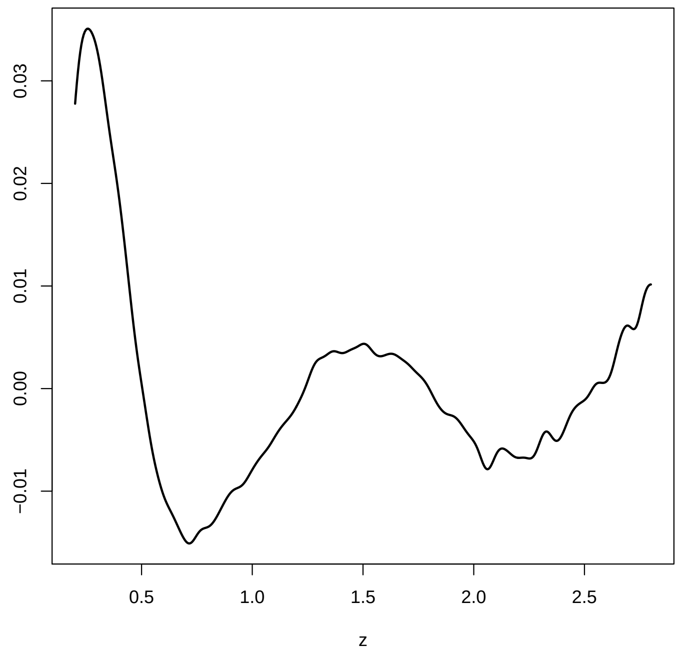
5.3 Income distributions by industry

Next, taking into account the differences across industries, we decompose ordinary workers by industry and obtain income distributions of 8 industries each year.[19] As an illustration, density functions in manufacturing and service industries are shown in Figure 6.

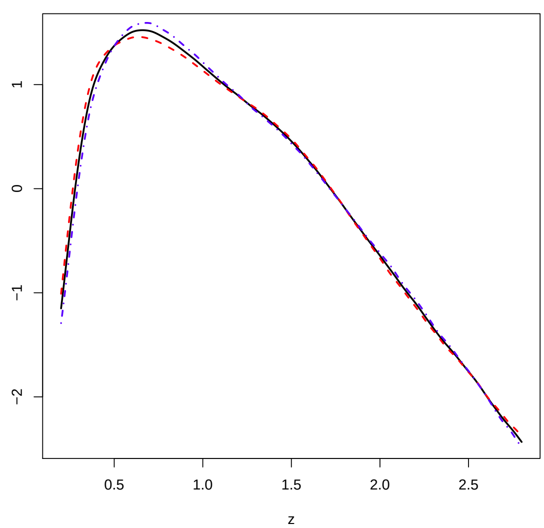
We assume that there is unobserved heterogeneity across industries, which is constant over time,

[18] Although the relationship in Figure 5(b) appears to be a linear function at a first glance, it is, strictly speaking, a convex function. While it inherits the linearity defined in the Bayes space, there is no reason to believe that the property yields a linear relationship between aggregated indexes and explanatory variables.

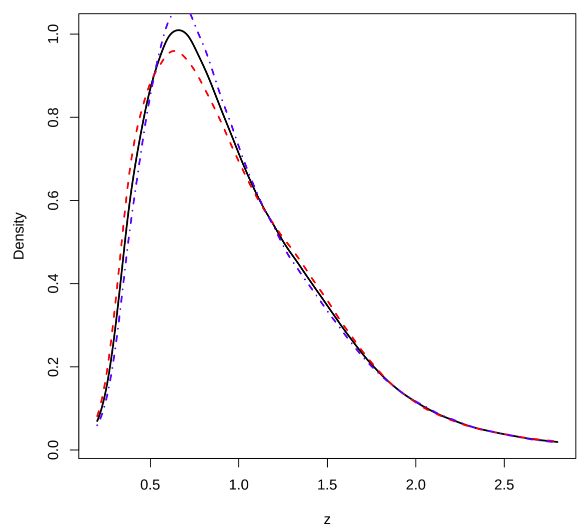
[19] Along with the Japan Standard Industrial Classification (the 10th revised edition, JSIC), the following 8 sectors are considered: construction, manufacturing, wholesaler & retailer, finance, real estate, transportation & information, utility, and service. The revision of JSIC was intermittently carried out (the 11th rev. in March 2002, the 12th rev. in November 2007, and the 13th rev. in October 2013. The 11th rev. JSIC has been employed in BSWS since 2005). The 11th and 12th revisions involve the changes of the definition of classification, increasing the number of major divisions. There is no straightforward way to connect the data before/after the revision in an consistent manner, we focus on the period 1989-2004, in which there was no change regarding major divisions.



(a) Coefficient $\hat{b}_1$.

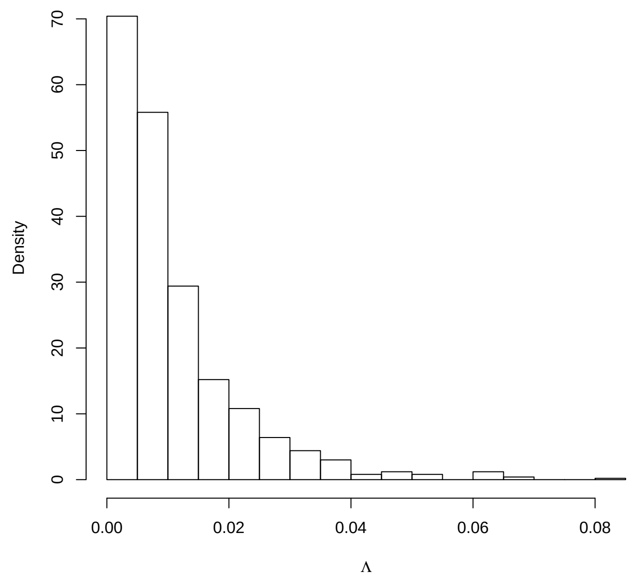


(b) $\bar{y}$ and $\hat{y}_{\pm}$.

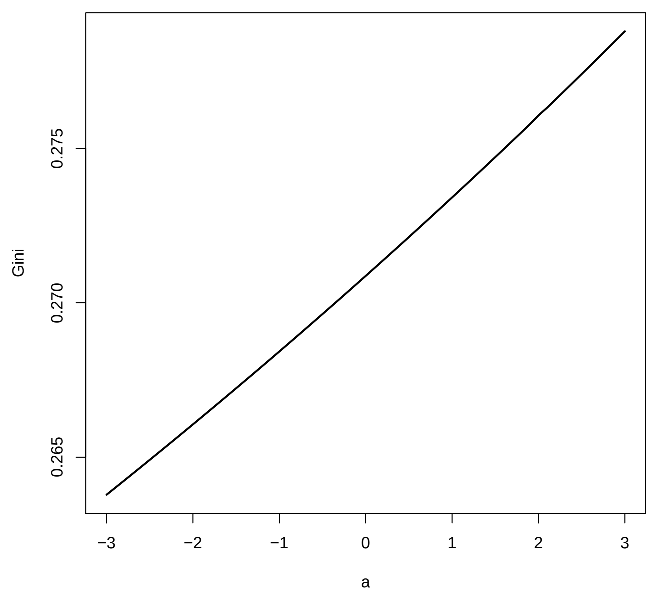


(c) $\mathrm{clr}^{-1}(\bar{y})$ and $\mathrm{clr}^{-1}(\hat{y}_{\pm})$.

Figure 4: Estimation results. In panel (b) and (c), the solid line is $\bar{y}$ and the dashed (dash-dot) line is $\hat{y}_+(\hat{y}_-)$.

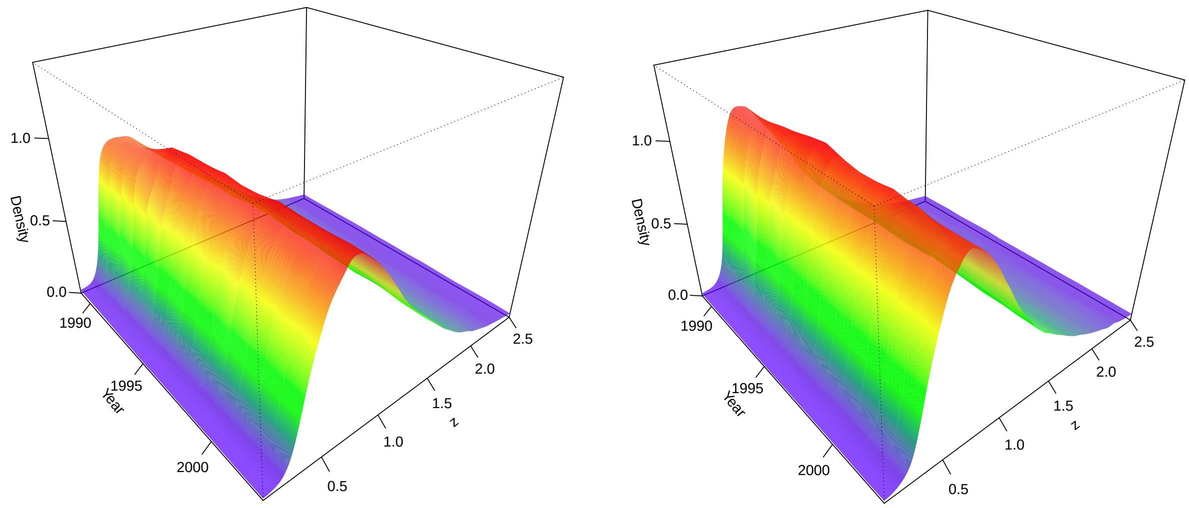


(a) Histogram of Λ.



(b) Gini coefficient.

Figure 5: Estimation results. In panel(a), the number of bootstrap samples is 1000. In panel(b), the relationship between the GDP growth and the Gini coefficient is shown.



Manufacturing. Service.

Figure 6: Density functions in 1989–2003.

but the response of the density function to growth in the industry is the same. Namely, including industry and year dummy variables, our regression model becomes as follows:

$$f = \beta_0 \oplus x_1 \odot \beta_1 \oplus \oplus_{j=1}^{7}(x_j^{Ind} \odot \beta_j^{Ind}) \oplus \oplus_{j=1}^{14}(x_j^{Year} \odot \beta_j^{Year}) \oplus \epsilon, \tag{8}$$

where x_j^{Ind} and x_j^{Year} are industry and year dummy variables, respectively. Here, construction in the year 1989 is the reference state. x_1 is the GDP growth rate for the industry and β_1 is common to all the industries. While x_j^{Year} is included in the model to control for changes of economic environment such as natural disasters, an economy-wide shock may also be captured by β_j^{Year}.[20] However, if the relationship between the income distribution and growth rates shown in the previous section is robust and not caused by some omitted variables such as policy changes, a similar relationship must be found due to variation of growth rates across industries. In fact, this is found as we will see in the following.

As in the previous section, we estimate the coefficient of the GDP growth rate, $\hat{b}_1$, which is shown in Figure 7(a). Using this estimate, we calculate $\bar{y} = \hat{b}_0 + \bar{x}_1 \cdot \hat{b}_1 + \hat{b}_1^{Ind}$ (i.e., manufacturing

[20]Here, we use the nominal GDP growth rate as x_1 by considering that the effect of inflation on the income distribution, if any, is removed of by including the year dummy. The data are obtained from Annual Report on National Accounts.

Table 3: Summary statistics of the annual GDP growth rate by industry(%). The total number of observations is 120.

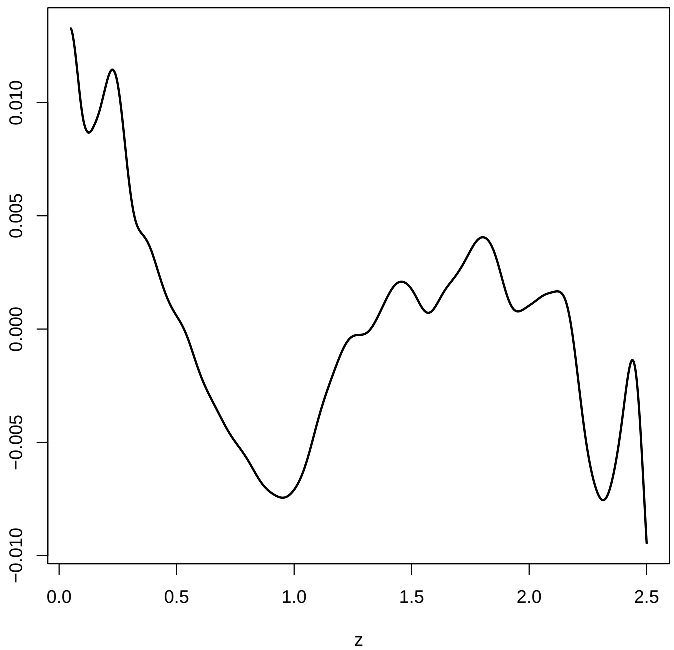
	Mean	Std. Dev.	Min	Max
Construction	1.1	6.78	-7.3	14.4
Manufacturing	.52	4.64	-6.6	7.9
Wholesaler & Retailer	2.56	4.77	-4.0	13.0
Finance	2.77	5.32	-7.7	15.7
Real estate	3.49	2.35	-0.1	7.5
Transportation & Information	2.53	3.36	-2.1	9.0
Utility	1.22	2.62	-3.5	6.6
Service	3.99	2.82	-0.2	8.7

in 1989) and $\hat{y}_{\pm} = \bar{y} \pm 2\sigma_1 \cdot \hat{b}_1$, which are shown in Figure 7(b). $\mathrm{clr}^{-1}(\bar{y})$ and $\mathrm{clr}^{-1}(\hat{y}_{\pm})$, i.e., probability density functions, are shown in Figure 7(c). The coefficient $\hat{b}_1$ shown in Figure 7(a) has characteristics similar to the one given in the previous section.

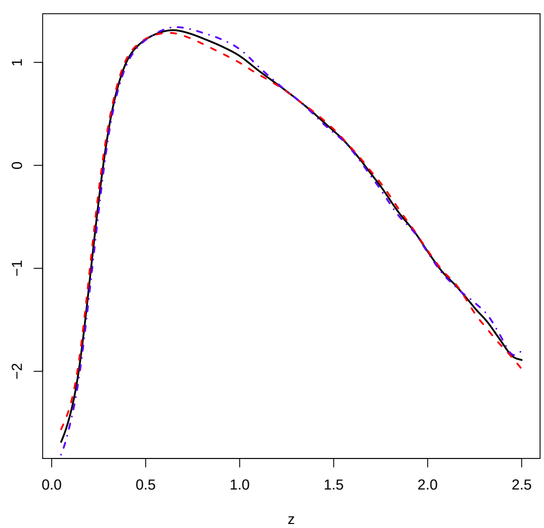
The test statistics Λ in Proposition 4 is .0586. Percentile values of Λ under the null hypothesis are as follows: $\Lambda_{Chi,95} = .0464$, $\Lambda_{Chi,97} = .0540$, and $\Lambda_{Chi,99} = .0706$, where the definition of $\Lambda_{Chi,p}$ is the same as before. The bootstrap result is also presented in Figure 8(a), which suggests that $\Lambda_{B,95} = .0449$, $\Lambda_{B,97} = .0506$, and $\Lambda_{B,99} = .0645$. Both tests show that the coefficient is statistically significant with p-value $< .03$. As before, the Gini coefficient is calculated from $\hat{y}_a = \hat{b}_0 + (\bar{x} + a \cdot \sigma_1) \cdot \hat{b}_1 + \hat{b}_1^{Ind}$. They are plotted in Figure 8(b), showing a positive relationship. The estimation result is essentially the same as the one in the previous section.

5.4 Discussion

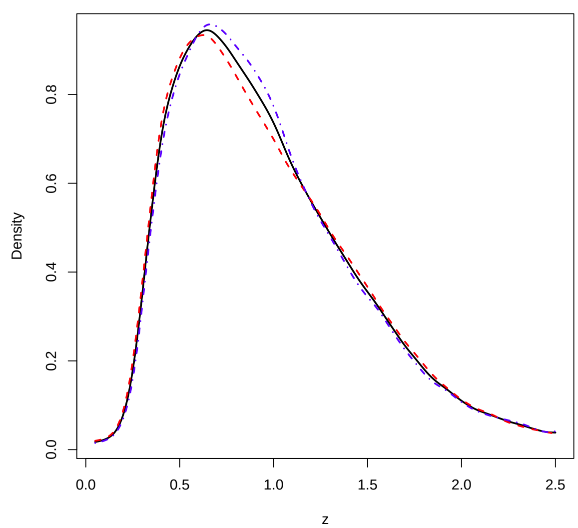
Figures 5(b) and 8(b) describe the positive relationship between the Gini coefficient and the GDP growth rate. This implies that the economic growth does not contribute to eliminating inequality measured by the Gini coefficient, at least in the sample period that we study. If we focus only on finding this relationship, we do not need to use our method; the usual time series analysis is sufficient. However, the question we aim to address in this paper is what happens behind this positive linkage. Is an increase in the Gini coefficient caused by an increase in poverty or by the emergence of a group of rich workers? Even if we are facing an increase in the Gini coefficient, economic implications as well as policy measures needed depend on these questions. The answer is explicitly given by Figures 4(a) and 7(a).



(a) Coefficient $\hat{b}_1$.

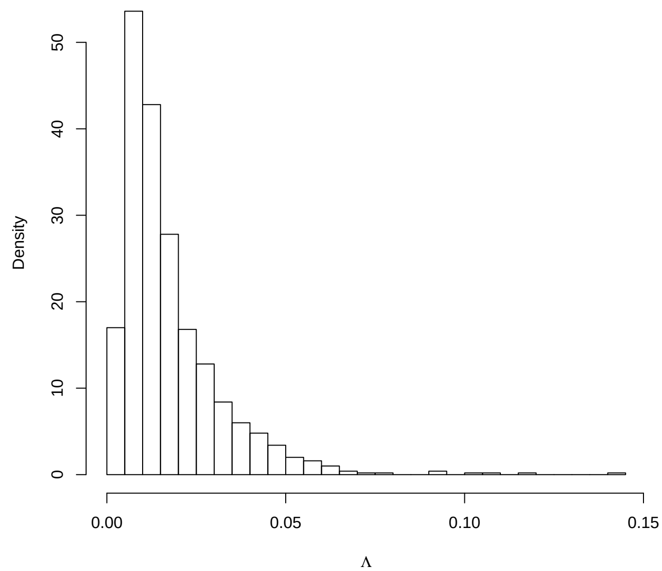


(b) $\bar{y}$ and $\hat{y}_\pm$.

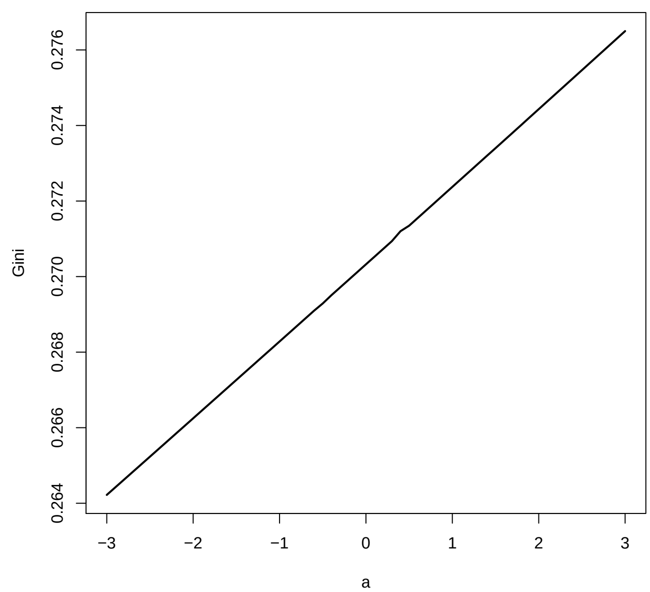


(c) $\mathrm{clr}^{-1}(\bar{y})$ and $\mathrm{clr}^{-1}(\hat{y}_\pm)$.

Figure 7: Estimation results. In panel (b) and (c), the solid line is $\bar{y}$ and the dashed (dash-dot) line is $\hat{y}_+(\hat{y}_-)$.



(a) Histogram of Λ.



(b) Gini coefficient

Figure 8: Estimation results. In panel(a), the number of bootstrap samples is 1000. In panel(b), the relationship between the GDP growth and the Gini coefficient is shown.

As shown in these Figures, the coefficient function $\hat{b}_1$ has distinctive features. $\hat{b}_1$ takes large values in the region of lower z ($z < .5$) and sharply drops to its minimum value around $z = 1$ or less. This means that an increase in the GDP growth rate is associated with a relative increase in lower income workers and decrease in the middle class. On the other hand, in the region of higher values of z (around $z = 1.5$ or more), it stays around 0. This means that there is no substantial change in this higher region.[21] Thus, most of the changes generating the variation of inequality measured by the Gini coefficient take place in the lower and middle regions.

So far, we have not discussed the causality: Is the change of the income distribution caused by economic growth or does the change of the income distribution lead to economic growth? In the literature, both causalities have been considered, though the results have been mixed (see reference in the Introduction). What we have done in this paper is to find a functional relationship and, therefore, we cannot give the definitive answer to the causality. However, it should be noted that the results of our analysis are used as a new test because the true model must approximate not only the behavior of a few indexes but also that of the income distribution. In the existing literature, several theoretical models explaining inequality and economic growth have been proposed, e.g., Galor and Tsiddon (1997), Galor and Moav (2000), Barlevy and Tsiddon (2006). For example, Barlevy and Tsiddon (2006) focus on technological changes and workers' optimal responses, showing that income inequality representing the disparity of skills is related to economic growth.[22] They also examine whether their model is consistent with empirical data by using inequality indexes such as top income shares and the Gini coefficient. However, the model has further implications: it also predicts the change of the income distribution behind that of inequality. According to their model, the change of income is monotone, i.e., the rich grow richer and the poor grow poorer when inequality rises, which implies that the income distribution becomes flatter when inequality is high. Namely, the functional coefficient b_1 takes a U or V-shaped form in this model. However, this conclusion is inconsistent, at least, with our results. As shown in Figures 4(a) and 7(a), the estimated coefficient $\hat{b}_1$ is quite different from a simple U or V-shaped form, especially in the region of large z. This suggests that the mechanism driving the change of inequality is different from what is envisaged in

[21] In addition, Figure 4(a) shows that there seems to be the second *hollow* around $z = 2$ rather than positive values.

[22] They argue that, in times of recessions, more able workers have more incentives to spend their time to master the new technology, and the disparity of skills among workers would grow at a faster rate. Economic growth matters because the return of their investments(i.e., their time) depends on it.

their model. In this way, our analysis can be used to further analyze the model prediction and to explore whether theoretical models are actually supported by empirical data.

6 Conclusions

In this paper, density functions have been consistently dealt with in the Bayes space rather than in the L^2 space as in FDA. Density functions are transformed by the clr transformation and the constrained nature of density functions are properly incorporated in estimation. Since this space has favorable properties such as linearity, we introduce a functional linear regression model in this space based on this structure. Since functions other than probability density functions are excluded from our consideration in the first place, our regression analysis necessarily yields density functions and shows good accuracy. Indeed, this point has been confirmed by simulation studies based on the exponential family.

We then apply this statistical method to cross-sectional distributions of individual earnings and the GDP growth rates to examine the relationship between inequality and economic growth. As the long-term controversy suggests, the relationship is integral to macroeconomic theory as well as policy-making, and in fact, previous studies have developed many indexes of inequality and analyzed the interdependence between the indexes and macroeconomic variables. However, since using the indexes is equivalent to dimension reduction, it is inevitable that a considerable part of information that an original income distribution has is lost. Rather than seeking appropriate indexes, we have treated an income distribution as an infinite-dimensional random variable in this paper. From this point of view, the indexes are regarded as only one of the aspects of the random variable. Our analysis can reveal a detailed relationship represented by a functional coefficient and shows that the changes of the income distribution take place in the lower and middle income regions. A disproportional increase in the lower class, a relative reduction of the middle class, and irresponsiveness of the higher class to economic growth are a striking feature of the behavior of the income distribution. These findings shed a new light on issues about inequality and economic growth. Moreover, the functional coefficient obtained can also be used as a new test in that the criteria are a probability density function itself rather than a few indexes.

In economics, there are many distributions other than the income distribution that has received increasing attention recently, e.g., the wealth distribution, the firm size distribution, the distribution of growth rates of firms, and the distribution of stock returns. The increasing availability of big data enables us to directly observe the probability density functions. However, as in the case of the income distribution, most of the analysis in this field still relies on a few indexes. Like our analysis, statistical methods that directly analyze density functions can fully utilize the power of big data and have great potential for discovering new economic dynamics ignored in previous studies.

A Proof of Proposition 4

Proof. Our proof is a generalization of the standard approach in the finite-dimensional case (see, e.g., Seber and Lee (2003)) and closely related to the one in Reimherr and Nicolae (2014) for FDA. Following Reimherr and Nicolae (2014), we show that the test statistics Λ converges to the weighted sum of i.i.d. χ^2 random variables. The key is the fact that the L_0^2 space is a separable Hilbert space.

Set the two least squares estimators:

$$\hat{\mathbf{b}} \equiv (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \quad \hat{\mathbf{b}}_1 \equiv (\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T\mathbf{y}.$$

Our test statistics Λ is the reduction in the sum of squared residuals by including $\mathbf{X}_2$. Straight calculations yield that Λ can be written as follows:

$$\Lambda = \int_I \mathbf{e}(t)^T[\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T]\mathbf{e}(t)dt$$

Since the L_0^2 space is a separable Hilbert space, we can use the central limit theorem in the Hilbert space:

$$N^{-1/2}\mathbf{X}^T\mathbf{e}(t) \xrightarrow{d} \mathbf{Z}(t)$$

where $\xrightarrow{d}$ represents the convergence in distribution and $\mathbf{Z}(t)$ is a Gaussian random variable.[23]

[23] A H-random variable X is called Gaussian if the characteristic function φ_X can be written as

$$\varphi_X(x) \equiv E[\exp i\langle x, X\rangle] = \exp(i\langle x, EX\rangle - \frac{1}{2}\langle C_X(x), x\rangle), \; x \in H,$$

Since a Gaussian random variable is determined by the expectation (= 0 in our case) and the covariance operator, $\mathbf{Z}$ can be written as $\mathbf{Z}(t) = \Sigma_X^{1/2}\mathbf{N}(t)$, where $\mathbf{N} \equiv \{N_l\}_{l=1,\dots,k}$ is a vector of i.i.d. Gaussian random variables in L_0^2; that is, $N_l \sim \mathcal{N}(0, C_e)$. Therefore, by the linear transformation property of Gaussian random variables,

$$\Lambda \xrightarrow{d} \int_I \mathbf{N}(t)^T \mathbf{A}\mathbf{N}(t)dt \stackrel{d}{=} \int_I \sum_{k=1}^{L} N_k(t)^2 dt.$$

where $\mathbf{A}$ is a projection matrix with rank L.[24] By the Karhunen-Loéve expansion, (see, e.g., Theorem 1.5 in Bosq (2000)) we have

$$N_l \stackrel{d}{=} \sum_{n=1}^{\infty} \sqrt{\lambda_n} N_n \varphi_n,$$

where the sequence (λ_n, φ_n) is defined in the Mercer lemma, i.e., the eigenfunctions and eigenvalues of C_e, and N_n is an independent standard Gaussian random variable. Therefore, $\Lambda \xrightarrow{d} \sum_{n=1}^{\infty} \lambda_n \chi_n^2(L)$. ■

B Family Income and Expenditure Survey (FIES)

For another implementation of our method, we examine the relationship between the wealth distribution and macroeconomic variables, which has received increasing attention recently, e.g., Piketty and Goldhammer (2014). Here, the wealth distribution are obtained from Family Income and Expenditure Survey in the period 2002-2015. As a dependent variable, the change of the wealth distribution, i.e., $f_{t+1} \ominus f_t$ is taken, where the duration of time intervals is 3 months. Explanatory variables are the difference between the interest rate and growth rate, $r_t - g_t$, and stock returns, $return_t$. Namely, we consider the following model:

$$f_{t+1} \ominus f_t = \beta_0 \oplus \beta_1 \odot (r_t - g_t) \oplus \beta_2 \odot return_t \oplus \epsilon, \tag{9}$$

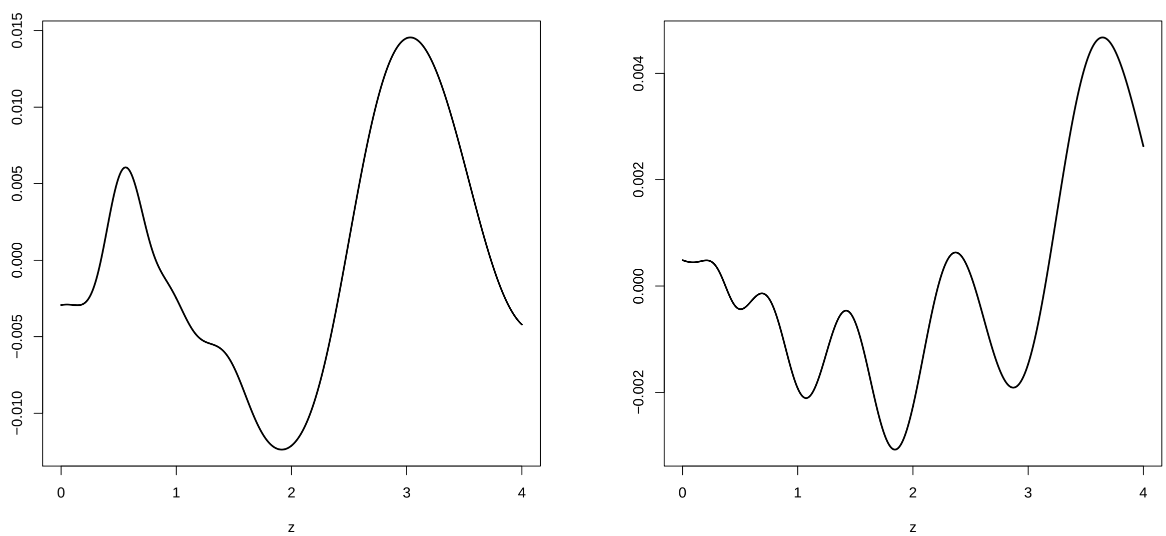
where EX is the expectation and C_X is the covariance operator.

[24]

$$\mathbf{A} \equiv \mathbf{I} - \Sigma_X^{1/2} \begin{pmatrix} \Sigma_{X,11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \Sigma_X^{1/2}$$

r_t is long-term interest rates, i.e., government bonds maturing in ten years. g_t is the GDP growth rate which is seasonally adjusted and measured in percentage change from previous quarter. $return_t$ is percentage change of share prices. All these data are obtained from OECD data. More detailed information is available from the author upon request.

As before, we calculate the functional coefficients, which are shown in Figure 9, and carry out our asymptotic test. For the coefficients b_1 and b_2, p-values are .0458 and .102, respectively.



Coefficient $\hat{b}_1$.

Coefficient $\hat{b}_2$.

Figure 9: Estimations.

Although the p-value is strictly speaking larger than 10%, the shape of $\hat{b}_2$ shown in the right panel of Figure 9 is informative. In contrast to the distribution of labor income discussed in Section 5, the function $\hat{b}_2$ takes large values in the region of high z, meaning that the right tail of the wealth distribution is much affected by stock returns. This is consistent with our intuition because the change of individual wealth, i.e., income considered here includes capital income. In other words, this clearly indicates unequally distribution of benefits from an increase in the stock price to households.

The interpretation of the coefficient $\hat{b}_1$ is much difficult. This shape presented in the left panel of Figure 9 suggests that an increase in $r_t - g_t$ promotes inequality, but is different from a U-shaped form. It might indicate the existence of another economic mechanism generating inequality different from what is envisaged in Piketty and Goldhammer (2014).

References

Acemoglu, D., Robinson, J. A., 2015. The rise and decline of general laws of capitalism. The Journal of Economic Perspectives 29 (1), 3–28.

Ahluwalia, M. S., 1976. Income Distribution and Development: Some Stylized Facts. The American Economic Review 66 (2), 128–135.

Aitchison, J., 1986. The statistical analysis of compositional data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd.

Alesina, A., Rodrik, D., 1994. Distributive politics and economic growth. The Quarterly Journal of Economics 109 (2), 465–490.

Barlevy, G., Tsiddon, D., 2006. Earnings inequality and the business cycle. European Economic Review 50 (1), 55–89.

Barro, R. J., 2000. Inequality and Growth in a Panel of Countries. Journal of Economic Growth 5 (1), 5–32.

Bosq, D., 2000. Linear Processes in Function Spaces: Theory and Applications. Vol. 149. Springer Science & Business Media.

Chang, Y., Kim, C. S., Park, J. Y., 2016. Nonstationarity in time series of state densities. Journal of Econometrics 192 (1), 152–167.

Cuevas, A., Febrero, M., Fraiman, R., 2006. On the use of the bootstrap for estimating functions with functional data. Computational Statistics & Data Analysis 51 (2), 1063–1074.

Delicado, P., 2011. Dimensionality reduction when data are density functions. Computational Statistics & Data Analysis 55 (1), 401–420.

Egozcue, J., Díaz-Barrero, J., Pawlowsky-Glahn, V., 2006. Hilbert Space of Probability Density Functions Based on Aitchison Geometry. Acta Mathematica Sinica 22 (4), 1175–1182.

Ferraty, F., Vieu, P., 2006. Nonparametric functional data analysis: theory and practice. Springer Science & Business Media.

Galor, O., Moav, O., 2000. Ability-biased technological transition, wage inequality, and economic growth. quarterly Journal of Economics, 469–497.

Galor, O., Moav, O., 2004. From Physical to Human Capital Accumulation: Inequality and the Process of Development. The Review of Economic Studies 71 (4), 1001–1026.

Galor, O., Tsiddon, D., 1997. Technological progress, mobility, and economic growth. The American Economic Review 87 (3), 363–382.

Horváth, L., Kokoszka, P., 2012. Inference for Functional Data with Applications. Vol. 200. Springer Science & Business Media.

Hron, K., Menafoglio, A., Templ, M., Hrůzová, K., Filzmoser, P., 2016. Simplicial principal component analysis for density functions in Bayes spaces. Computational Statistics & Data Analysis 94, 330–350.

Kneip, A., Utikal, K. J., 2001. Inference for density families using functional principal component analysis. Journal of the American Statistical Association 96 (454), 519–532.

Kuznets, S., 1955. Economic Growth and Income Inequality. The American Economic Review 45 (1), 1–28.

Lise, J., Sudo, N., Suzuki, M., Yamada, K., Yamada, T., 2014. Wage, income and consumption inequality in Japan, 1981–2008: From boom to lost decades. Review of Economic Dynamics 17 (4), 582–612.

Park, J. Y., Qian, J., 2012. Functional regression of continuous state distributions. Journal of Econometrics 167 (2), 397–412.

Piketty, T., Goldhammer, A., 2014. Capital in the Twenty-First Century. Harvard University Press.

Ramsay, J., Silverman, B., 2005. Functional Data Analysis, 2nd Edition. Springer Science & Business Media.

Reimherr, M., Nicolae, D., 2014. A functional data analysis approach for genetic association studies. The Annals of Applied Statistics 8 (1), 406–429.

Seber, G. A., Lee, A. J., 2003. Linear regression analysis, 2nd Edition. John Wiley & Sons.

van den Boogaart, K., Egozcue, J., Pawlowsky-Glahn, V., 2010. Bayes linear spaces. Statistics and Operations Research Transactions 34 (2), 201–222.

van den Boogaart, K., Egozcue, J., Pawlowsky-Glahn, V., 2014. Bayes Hilbert Spaces. Australian & New Zealand Journal of Statistics 56 (2), 171–194.