Department of Economics
Seminar Series

David Jacho-Chavez
Emory University

“Functional Principal Component Analysis of Density Families with Complex Survey Data on UK Prices”

Friday,
August 28, 2015
3:30 p.m.
212 Middlebush
On the Evolution of the United Kingdom Price Distributions

Ba M. Chu†  Kim P. Huynh‡  David T. Jacho-Chávez§  Oleksiy Kryvtsov¶

December 25, 2014

Abstract

We propose a functional principal components method that accounts for stratified random sample weighting to understand the evolution of distributions of monthly micro-level consumer prices for the United Kingdom (UK). We apply the method to publicly available monthly data on individual-good prices collected in retail stores by the UK Office for National Statistics for the construction of the UK Consumer Price Index from March 1996 to March 2014. Our novel method is amenable to visualization, and we find that dynamic evolution of the price distribution is highly correlated with economic indicators such as the unemployment rate and the inflation rate.

Keywords: Consumer Price Distributions, Nonparametric Methods, Stratified Random Sampling.

---

*We thank Stuart Berry, Songnian Chen, David Cook, Erwin Diewert, Allan Gregory, James MacKinnon, Gregor W. Smith, and Liangjun Su for helpful comments. We are grateful to the UK Office for National Statistics, especially, Derek Bird and Ainslie Restieaux for their assistance with this project; Rosie Foster and Rowan Kelsoe for their detailed responses to our queries regarding the data. Boyan Bejanov and Claudiu Motoc provided excellent scientific computation advice on using SQLite to handle large datasets and using npRmpi and snow parallel procedures on the EDITH high performance cluster. Tony Chernis, Geoff Halmo, and Anderson Nzabandora provided excellent research assistance. The views expressed in this paper are those of the authors. No responsibility for them should be attributed to the Bank of Canada. All remaining errors are the responsibility of the authors.

†Department of Economics, Carleton University, B-857 Loeb Building, 1125 Colonel By Drive, Ottawa, ON K1S 5B6, Canada. Phone: +1 (613) 520 2600. Fax: +1 (613) 520 3906. E-mail: ba.chu@carleton.ca

‡Bank of Canada, 234 Laurier Avenue West, Ottawa, Ontario K1A 0G9, Canada. Phone: +1 (613) 782 8698. Fax: +1 (613) 782-7764. E-mail: kim@huynh.tv.

§Department of Economics, Emory University, 1602 Fishburne Dr, Rich Bldg., 3rd Floor, Atlanta, GA 30322, USA. Phone: +1 (404) 727 6151. Fax: +1 (404) 727 4639. E-mail: djacho@emory.edu.

¶Bank of Canada, 234 Laurier Avenue West, Ottawa, Ontario K1A 0G9, Canada. Phone: +1 (613) 782 7818. Fax: +1 (613) 782 7764. E-mail: oleksiy.kryvtsov@gmail.com
1 Introduction

This paper investigates the evolution of the underlying distribution of prices used to construct the Consumer Price Index (CPI) for the United Kingdom as illustrated in Figure 1. In all countries, the CPI is the most commonly used measure of the average price that households pay for their consumption, and it is also widely used as a measure of the changes in the cost of living or inflation. The ubiquity of the CPI in the everyday lives of consumers and firms makes it one of the most watched, reported, and discussed statistics in a modern economy. National statistical agencies such as the Bureau of Labour Statistics (BLS) in the United States, the Office for National Statistics (ONS) in the United Kingdom, and Statistics Canada devote considerable resources to the efficient collection and dissemination of these data. Central banks all over the world rely on the CPI to formulate economic policies, especially countries with an explicit inflation target such as the United States (US), United Kingdom (UK), and Canada.\(^1\)

The measurement of the CPI has considerable economic implications, and it is therefore often debated and extensively studied. As an example, in the US the Boskin commission of 1996 suggested methodological changes that resulted in 1.1 percentage points per annum prior to 1996, see Boskin et al. (1997).\(^2\) The US Congressional Budget Office states that this upward bias contributes to about 148 billion US dollars to the deficit in 2006 and 691 billion USD to the national debt. In relative terms, this bias would be the fourth largest federal program, after social security, health care and defense! One of the key recommendation of the Boskin commission was to suggest creating representative price index that allows for flexibility and dynamism in the economy. Therefore, creating this superlative CPI is of first-order importance for the government.

In practice, different measures of the CPI are reported to account for the complex relative dynamics of its components.\(^3\) This relative dynamics manifests itself along two dimensions: (1) prices do not move in a synchronous manner across goods and services; for example, food and energy prices tend to be more volatile; and (2) relative expenditures for goods and services are also time-varying, as households shift their consumption in response to price changes, discounts, stockouts, changes in quality, or over the business cycle.\(^4\)

Understanding these complex dynamics of relative prices and expenditure weights requires information about micro-level prices and weights underlying the CPI. A limited availability of such data, however, hinders the advancement of tools and methods that could be used by

\(^1\)http://www.centralbanknews.info/p/inflation-targets.html provides a complete list of countries.
\(^2\)Greenlees (2006) provides an official response to the Boskin Commission recommendations.
\(^3\)Diewert (2012) provides a detailed discussion of the various CPI measures in the UK.
economists who study micro price dynamics. This paper utilizes monthly data on individual-good prices collected in retail stores by the ONS for the construction of the UK CPI from February 1996 to March 2014. These are the first-ever publicly available monthly data on individual-good prices collected by a national statistical agency.

The distributional analysis of prices in this data set poses a challenge. Similar to many other data sets collected by national statistical agencies, such as the ONS, the available price data is collected using stratified sampling techniques. In this situation, the incorporation of the probability distribution induced by the complex survey design into the statistical analysis is very important, because if this effect is not taken into account, subsequent statistical inference can be seriously flawed, see, e.g., Kish and Frankel (1974). The problem becomes more acute when the objective is to recover and characterize the dynamics of the underlying super-populations as in this research. Consequently we adapt a new method based on the functional principal component analysis (FPCA) of densities (first proposed by Kneip and Utikal, 2001) to study the distributional dynamics of prices, that utilizes the Sample Weighted Kernel Density (SWKD) estimator proposed by Buskirk (1998, 1999) and Bellhouse and Stafford (1999) instead. This reworking is necessary because Kneip and Utikal’s (2001) results are only valid when the data are independent realizations drawn from a common probability distribution in each time period, and do not directly apply to data from stratified samples. The asymptotic properties of this adaptation are developed here, adding to the emerging literature studying jointly complex survey sampling and functional data analysis, see, e.g., Cardot et al. (2010), Cardot and Josserand (2011), and Cardot et al. (2013).

The main contributions of this paper are threefold: First, we use this method to characterize the dynamics of price distributions by decomposing changes in the distribution into a time-invariant component, or the basis functions, and a time-varying component, or dynamic strength. Second, using this decomposition, we capture factors driving the changes in the distribution of prices in a parsimonious way and give them macroeconomic interpretation. Our analysis reveals that the dynamic strength components for price levels are highly correlated with the unemployment rate while those for their changes are highly correlated with the inflation rate. Finally, the results from our time-invariant basis functions indicate that higher order moments related to location-scale models are insufficient to characterize price distributions.

Apart from our empirical findings, these tools can be utilized by researchers that study the dynamics of price distributions using an ever-growing number of large panel data sets, similar

---

5 Only a few privileged external researchers have limited access to these data sets. Data access usually requires a lengthy application process and is limited to economists within the country. Conditions of access often preclude the use of powerful computing resources and specialized software. Release of results is subject to review and may involve substantial delays.

6 The price quote data are available via the ONS website: http://www.ons.gov.uk/ons/datasets-and-tables/index.html.
to the one we use in this paper. Furthermore, it is straightforward to adapt our method to other micro data sets that are used to study the dynamics of household income, credit and asset holdings, household demographic, educational and employment characteristics, firm investment, firm and bank balance sheets.

The rest of the paper is organized as follows: Section 2 provides a description of the UK price data used in the paper. Section 3 discusses the functional principal components analysis with complex survey design methodology, while section 4 discusses the results of this dynamic distributional analysis. Finally, section 5 concludes.

2 UK Consumer Price Index Data

To construct the CPI, the ONS surveys the prices for goods and services that are included in the household final monetary consumption expenditure component of the UK National Accounts. The survey includes prices for more than 1,100 individual goods or services a month, collected across more than 14,000 retail stores across the UK. The survey excludes housing portion of consumer prices, such as mortgage interest payments, house depreciation, insurance and other house purchase fees. Also expenditures for purposes other than final consumption are excluded, e.g., those for capital and financial transactions, direct taxes, cash gifts, etc.

The goods and services in the CPI are classified into 71 classes, according to the international (European) classification of household expenditure, Classification of Individual Consumption by Purpose (COICOP). A CPI “class” represents a basic group category, such as “Meat”, “Liquid Fuels” or “New Cars.” Each item in a given class is assigned an item weight, that reflects its relative importance in households’ consumption expenditures. Changes in expenditure weights over time reflect changes in expenditure composition of households’ consumption baskets.

Figure 1 shows that the CPI price level in the UK displays significant trend and cyclical variations across time in the UK. The price level grew by an average of 2.1% per year between 1996 and 2013, and price inflation fell by 2 to 3 percentage points in annualized terms during each of the two recessions in 2000’s. While these micro-level time variations may differ from each other, they are also interconnected. For example, prices of Jaguar automobiles in London are likely to move similarly to Honda prices in Manchester; prices of clothing and shoes go up

---

7Bryan and Cecchetti (1999), Baharad and Eden (2004), Klenow and Kryvtsov (2008), and Midrigan (2011) are examples of research using related data.

8Detailed description of the data underlying the CPI, the statistical methodology used, collection and validation of prices, calculation of weights, can be found in the “Consumer Price Indices Technical Manual” (2012). The price quote data are available via ONS website: http://www.ons.gov.uk/ons/datasets-and-tables/index.html.

9These prices are used to construct the Retail Prices Index (RPI).

10CPI inflation is used by the Government’s for its inflation target starting in December 2003. Before then, the index was published in the UK as the Harmonised Index of Consumer Prices (HICP).

11Weights are calculated based on the Household Final Monetary Consumption Expenditure (HFMCE) and ONS Living Cost and Food Survey (LCF).
and down in tandem. As we will show below, much of information on such complex micro-level price variations is contained in the dynamics of the distribution of prices.

2.1 Data Description

Prices are collected across 13 geographical regions, e.g., London, Wales, East Midlands. There are four levels of sampling for local price collection: locations, outlets within location, items within section and product varieties. For each geographical region locations and outlets based on a probability-proportional-to-size systematic sampling with a size measures based on the number of employees in the retail sector (locations) and the net retail floor space (outlets). The data set contains around 150 locations with an average of more than 90 outlets per location.

Representative items are selected based on a number of factors, including expenditure size and product diversity, variability of price movements, availability for purchase throughout the year (except for certain goods that are seasonal). There are currently over 510 items in the basket. Examples of representative items include: onions, men’s suit, single bed. Finally, for each item-outlet-location, individual products and varieties are chosen by price collectors based on their shelf size and regular stock replenishment.

Most prices are collected monthly via a stratified sampling. Exceptions include some services in household and leisure groups, and seasonal items. The sample period includes 217 months ($T$), from March 1996 till March 2014. The total number of observations is over 24 million ($n = \sum_{t=1}^{217} n_t$), or about 110,000 per month. These samples contains realizations and stratum weights for observations stratified by shop type, by region and by region and shop type. Two types of shop are identified for the stratum weights: multiples and independents. Retailers with fewer than 10 outlets are classified as independents, while retailers with 10 or more outlets are classified as multiples. Similarly, the UK is divided into 12 distinctive regions, namely London, South East, South West, Eastern, East Midlands, West Midlands, Yorkshire and the Humber, North West, North East, Scotland, Wales and Northern Ireland. Details regarding the collection and validation of prices, weights calculation and sampling methodologies can be found in ONS (2014). This study only utilizes the samples stratified by region and shop type yielding a total of 24 strata ($L$).

In addition to posted prices, the data set also contains information about some good characteristics during price collection. Prices for goods that are on special offer (available to all consumers) or on temporary sale comprise 6.4% of observation in the data set (4.5% weighted). Forced substitutions happen 8.0% (5.5% weighted) of time, with about a quarter (three quarters) of them corresponding to substitutions for items that are non-comparable (comparable) to previously priced items. An item can be temporarily out-of-stock (2.2% or 1.5% weighted) or permanently missing (0.5% or 0.3% weighted). Finally, a small subset of goods has distinct seasonal patterns and is treated separately: they include some items of clothing, gardening
products, holidays and air fares. For those seasonal goods for which prices are not available, such as clothing, gardening and food, prices are imputed based on prices observed at the end of the previous season or based on prices observed for “in-season” goods in the same item category; in addition, weights are adjusted in accordance with availability of such goods throughout the year.

To study the dynamics of distributions of prices and price changes, we construct the following variables: Let \( t \) denote a month, \( k \) denote an item-stratum bin and \( l \) denotes an individual product, so that \( p_{t,k,l} \) represents the natural logarithm of the price of a product or service \( l \) (in nominal British pounds) in month \( t \), for item-stratum bin \( k \); while \( p_{t,k,l} = p_{t,k,l} - p_{t-1,k,l} \) is the natural logarithm of the price change for that individual product in item-stratum bin \( k \) from month \( t - 1 \) to \( t \). Figure 2 displays estimates of the population mean, median and inter-quartile range of \( p_{t,k,l} \) for each of the 217 months. These descriptive statistics were constructed by estimating their corresponding population equations using the Horvitz-Thompson estimator and then solving them using the \textit{survey} R package by Lumley (2004). Figure 2 provides a visual summary of the data. The monthly distribution of log-prices has a dramatic shift in 01/1999. The latter can be attributed to a gradual change in location selection between 1995-1998 resulting in a complete random sample of locations from January 1999 onwards (see, i.e., ONS, 2014, Section 3.2.1). The quantiles fluctuate but without a noticeable trend until 2004/01 after which the 3rd quartile and median decreases until about 2006/1. After that the entire distribution (especially the median) shifts to higher prices from 2006 onwards. Overall, the interquartile range of the prices is about 0.40 to 5 pounds sterling.

Figure 2 provides two important insights. First, the dynamics of prices underlying the CPI and inflation rates reported in Figure 1, are complex. They stem from the lack of synchronization of prices across different goods and services in the basket. Second, fitting a simply location-scale models will fail to capture the real dynamics shown by these estimated population summary statistics, since the distribution clearly does not have the same functional form.

Considerable variations of prices at store and product levels suggest the observed dynamics of distributions of \( p_{t,k,l} \) or \( p_{t,k,l}^* \) could be driven by a subset of stores or goods with particularly volatile prices, creating numerical difficulties when recovering the overall movement of the super-population distribution of prices. Furthermore, in a given month some prices do not change from the previous month. We proceed by standardizing price levels and price changes as suggested in Klenow and Kryvtsov (2008). These are defined as \( x_{t,k,l} = (p_{t,k,l} - \overline{p}_{k,l}) / \overline{\sigma}_{p_{k,l}} \), and \( x_{t,k,l}^* = (p_{t,k,l}^* - \overline{p}_{k,l}) / \overline{\sigma}_{p_{k,l}}^* \), where \( \overline{p}_{k,l}, \overline{\sigma}_{p_{k,l}} \) are across-time estimated means of \( p_{t,k,l} \) for item-stratum \( k \) and individual product \( l \), and \( \overline{\sigma}_{p_{k,l}}, \overline{\sigma}_{p_{k,l}}^* \) are their respective estimates of their standard deviations. Hereafter, we refer to \( X \) and \( X^* \) as the ‘log(Price)’ and ‘\( \Delta \log(Price) \)’ respectively. Their underlying super-population distributions in each time period are the focus of our analysis hereafter.
3 Methodology

We start this section by first describing the general framework, notation and estimation strategy of the FPCA of density families estimated using complex survey data. The theoretical underpinnings derived here are more general than is needed for our empirical application, and therefore it can potentially accommodate more elaborate complex survey designs.

3.1 Nonparametric Kernel Density Estimation under Stratified Random Sampling

At each time period $t$, let $N_{t,1}, \ldots, N_{t,L}$ denote the sizes of $L$ sub-populations, $U_{N_{t,1}}, \ldots, U_{N_{t,L}}$, respectively at time $t$ such that the super-population size, $N_t = \sum_{k=1}^{L} N_{t,k}$. We assume that, for every $k \in \{1, \ldots, L\}$, the elements of each $U_{N_{t,k}}$ are realizations of the i.i.d. random variables, $X_{t,k,1}, \ldots, X_{t,k,N_{t,k}}$ (these would either represent log(Price) or $\Delta$log(Price) in our empirical application), distributed according to a density, $f_{t,N_{t,k}}$, also, $X_{t,k,i}$ and $X_{t,\ell,j}$, where $i = 1, \ldots, N_{t,k}$ and $j = 1, \ldots, N_{t,\ell}$, are independent for all $k \neq \ell$; and $X_{t,k,i}$ and $X_{s,\ell,j}$, where $i = 1, \ldots, N_{t,k}$ and $j = 1, \ldots, N_{s,\ell}$, are independent for all $t \neq s$. Define the stratum [sub-population] weight by $W_{t,k} = \frac{N_{t,k}}{N_t}$ the super-population density, $f_{N_t}$, is then the mixture of the sub-population densities, $f_{t,N_{t,1}}, \ldots, f_{t,N_{t,L}}$:

$$f_{N_t}(x) = \sum_{k=1}^{L} W_{t,k} f_{t,N_{t,k}}(x).$$

In what follows, we shall represent this super-population model by $M$.

For survey data generated by stratified random sampling, from each stratum, $U_{N_{t,k}}$ for $k = 1, \ldots, L$, it is assumed that one randomly samples without replacement a set of $n_{t,k}$ observations - denoted by $S_{t,k}$; then $S_t = \bigcup_{k=1}^{L} S_{t,k}$ is a stratified random sample of size $n_t = \sum_{k=1}^{L} n_{t,k}$ from the super-population, $U_t = \bigcup_{k=1}^{L} U_{N_{t,k}}$. We hereafter denote this sampling design by $D$, which can be defined by the design-based random variables, $I_{t,k,i} = 1$ if unit $i$ of the stratum $k$ belongs to $S_{t,k}$ and zero otherwise, for each $k = 1, \ldots, L$ and $k = 1, \ldots, N_{t,k}$. According to Cochran (1977, p. 29), $E[I_{t,k,i}] = \frac{n_{t,k}}{N_{t,k}}$ and $E[I_{t,k,i} I_{t,k,j}] = \frac{n_{t,k}(n_{t,k}-1)}{N_{t,k} (N_{t,k}-1)}$ for every $i \neq j$.

In view of Buskirk (1998, 1999) and Bellhouse and Stafford (1999), and after defining the stratum sampling weight by $w_{t,k} = \frac{N_{t,k}}{n_{t,k}}$, one can define the sample weighted kernel density
(SWKD) estimator as

\[
\hat{f}_{S_t,h}(x) = \frac{1}{N_t h} \sum_{k=1}^{L} w_{t,k} \sum_{i \in S_{t,k}} K \left( \frac{x - X_{t,k,i}}{h} \right)
\]

\[
= \frac{1}{N_t h} \sum_{k=1}^{L} w_{t,k} \sum_{i=1}^{N_{t,k}} I_{t,k,i} K \left( \frac{x - X_{t,k,i}}{h} \right),
\]

(1)

where \( K(\cdot) \) represents a standard kernel weighting function, and \( h \) denotes a scalar bandwidth that goes to zero as \( N_t \) becomes large. Buskirk (1999) demonstrates that \( \hat{f}_{S_t,h}(x) \) estimates the super-population density \( f_{N_t}(x) \) and its asymptotic equivalence with the infeasible estimator obtained by estimating the finite population quantity

\[
\hat{f}_{N_t,h}(x) = \sum_{k=1}^{L} W_{t,k} \hat{f}_{t,N_{t,k}}(x),
\]

where \( \hat{f}_{t,N_{t,k}}(x) \) are simple stratum kernel densities, i.e. Rosenblatt’s (1956) kernel estimator using all the data on sub-population \( k \). In what follows, we will use these two versions, and will assume without loss of generality that \( w_{t,k} = O(1) \) for every \( t = 1, \ldots, T \) and \( k = 1, \ldots, L \).

### 3.2 Functional Principal Component Analysis

Following the well-known Karhunen-Loève decomposition, Kneip and Utikal (2001) propose to represent each density, \( f_t \), at time \( t \) in terms of the principal functional components, \( g_1, \ldots, g_J \):

\[
f_{N_t} = f_\mu + \sum_{j=1}^{J} \theta_{t,j} g_j,
\]

(2)

where \( f_\mu = \frac{1}{T} \sum_{t=1}^{T} f_{N_t} \). In particular, let \( \mathcal{M} \) denote a variance-covariance matrix with the elements \( M_{ts} = \langle f_{N_t} - f_\mu, f_{N_s} - f_\mu \rangle \), where \( \langle f, g \rangle = \int_{\mathcal{X}} f(x) g(x) w(x) dx \), \( w(x) > 0 \) is some continuous, uniformly bounded weighting function \( \forall x \in \mathcal{X} \), and \( \mathcal{X} \) represent the compact support of \( f(x) \) and \( g(x) \). Kneip and Utikal (2001) demonstrate that \( \theta_{t,j} \) and \( g_j \) can be identified as:

\[
\theta_{t,r} = \frac{1}{2} \frac{\lambda_r^{1/2} p_r}{\sum_{t=1}^{T} \theta_{t,r}^2},
\]

(3)

\[
g_r = \frac{\sum_{t=1}^{T} \theta_{t,r} f_{N_t}}{\sum_{t=1}^{T} \theta_{t,r}^2},
\]

(4)

where \( p_1 = (p_{1,1}, \ldots, p_{T,1})^T; \ldots; p_r = (p_{1,r}, \ldots, p_{T,r})^T; \ldots \) are the eigenvectors of \( \mathcal{M} \); and \( \lambda_1 \geq \lambda_2 \geq \ldots \) are the eigenvalues of \( \mathcal{M} \).
Under an assumption of independent observations from $T$ populations, Kneip and Utikal’s (2001) approach to estimating the elements of (2) is based on first estimating each element of $\mathcal{M}$ by plugging in Rosenblatt’s (1956) kernel estimator, i.e. $\hat{\mathcal{M}}$. Estimators of $\theta_{t,r}$ and $g_r$ are then constructed from (3)-(4). Kneip and Utikal (2001) show the consistency of the resulting $\hat{\theta}_{t,r}$ and pointwise asymptotic normality of $\hat{g}_r$. Followup work by Huynh et al. (2011) extend the work to allow for categorical and continuous data.

3.3 Estimation with Stratified Random Sampling

However, when confronted with complex survey data, the SWKD estimator defined by (1) is better suited to estimate the super-population density, $f_{N_t}$, in each time period, as to account for the stratified random sampling nature of the data. Therefore, a natural estimator of $M$ is $\hat{M}_{ts} = \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{M}_{ts}$, where

$$\hat{M}_{ts} = \frac{1}{N_t h^2} \sum_{k=1}^{L} \frac{N_t^2}{n_{t,k}} \sum_{i \in S_{t,k}} K^2 \left( \frac{x - X_{t,k,i}}{h} \right) w(x) dx = \frac{1}{N_t h^2} \sum_{k=1}^{L} \frac{N_t^2}{n_{t,k}} \int K^2(u) du \left( \int w(v) f_{X_{t,k}}(v) dv + O(h) \right).$$

Therefore, by removing these terms, we can define the elements of the matrix, $\hat{\mathcal{M}}$, of the efficient estimators as

$$\hat{M}_{ts} = \hat{M}_{ts} - \frac{1}{T} \sum_{t=1}^{T} (\hat{M}_{tt} + \hat{M}_{ta}) + \frac{1}{T^2} \sum_{t=1}^{T} \sum_{s=1}^{T} \hat{M}_{ts},$$

where

$$\hat{M}_{tt}^* = \frac{1}{N_t} \sum_{k=1}^{L} \frac{N_t^2}{n_{t,k}} \sum_{i \neq j} I_{t,k,i} I_{t,k,j} (U_{t,k,i}, U_{t,k,j}) + \frac{1}{N_t} \sum_{k \neq \ell} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} I_{t,k,i} I_{t,\ell,j} (U_{t,k,i}, U_{t,\ell,j}),$$

$$\hat{M}_{ts}^* = \frac{1}{N_t N_s} \sum_{k=1}^{L} \sum_{\ell=1}^{L} \frac{N_t N_s}{n_{t,k} n_{s,\ell}} \sum_{i=1}^{N_t} \sum_{j=1}^{N_s} I_{t,k,i} I_{s,\ell,j} (U_{t,k,i}, U_{s,\ell,j}),$$

where $U_{t,k,i}(x) = \frac{1}{h} K \left( \frac{x - X_{t,k,i}}{h} \right)$. After calculating the eigenvalues $\hat{\lambda}_r$, and eigenvectors $\hat{p}_r$ of $\hat{\mathcal{M}}$ for $r = 1, 2, \ldots$, one can estimate (3) and (4) by the analogy principle, i.e.

$$\hat{\theta}_{t,r} = \hat{\lambda}_r^{1/2} \hat{p}_{t,r},$$

$$\hat{g}_r = \frac{\sum_{t=1}^{T} \hat{\theta}_{t,r} \hat{f}_{S_{t,h}}}{\sum_{t=1}^{T} \hat{\theta}_{t,r}^2},$$

(5)

(6)
where \( \hat{f}_{S_t,h} \) is defined in (1). One could use an estimator based on a different bandwidth say \( b (> h) \), e.g. \( \hat{f}_{S_t,b} \) as suggested by Kneip and Utikal (2001), but our proofs and empirical analysis are not affected by this choice.

On the other hand, by the same argument as Kneip and Utikal (2001), one can show that the estimator based on the super-population kernel density, \( \hat{f}_{N_t,h} \), for \( \mathcal{M} \), i.e. \( \hat{H} \), is given by

\[
\hat{H}_{ts} = \hat{H}_{ts}^* - \frac{1}{T} \sum_{i=1}^{T} (\hat{H}_{ti}^* + \hat{H}_{ts}^*) + \frac{1}{T^2} \sum_{i=1}^{T} \sum_{s=1}^{T} \hat{H}_{ts}^*,
\]

where

\[
\hat{H}_{ts}^* = (1 - \delta_{ts}) \frac{1}{N_t N_s} \sum_{k=1}^{L} \sum_{\ell=1}^{L} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{s,\ell}} \langle U_{t,k,i}, U_{s,\ell,j} \rangle \\
+ \delta_{ts} \left\{ \frac{1}{N_t^2} \sum_{k=1}^{L} \sum_{i \neq j} \langle U_{t,k,i}, U_{t,k,j} \rangle + \frac{1}{N_s^2} \sum_{k \neq \ell} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{s,\ell}} \langle U_{t,k,i}, U_{s,\ell,j} \rangle \right\},
\]

where \( \delta_{ts} = 1 \) if \( t = s \) and 0 otherwise.

### 3.4 Large Sample Properties

Here we adapt various results in Buskirk and Lohr (2005) to derive various finite sample and asymptotic properties of the resulting estimators of the eigenvalues and eigenvectors of \( \mathcal{M} \), and estimators defined in (5)-(6) under various modes of inference used with complex survey data. It should be noticed that these new results do not trivially follow from arguments in Kneip and Utikal (2001), because of the stratified sampling involved. In particular, we utilize the three frameworks common for asymptotic inference in this literature, namely: model-based (M), design-based (D), and a combination of the design-based and model-based (C).

Without loss of generality, we assume that, at each point in time, the number of observations, \( n_{t,k} \), sampled from each stratum \( k \) is proportional to the sub-population size \( N_{t,k} \). Our asymptotic analysis requires \( N_{t,\text{min}} = \min(N_{t,1}, \ldots, N_{t,L}) \) for every \( t = 1, \ldots, T \) to go to infinity such that \( W_{t,k} = O(1) \) and \( \max_{1 \leq k \leq L} w_{t,k} = O(1) \), which indeed implies that the number of strata \( L \) is fixed and that all strata are represented in the sample. Similarly, in order to make the derivation of the asymptotics more tractable, we also fix \( T \) and \( J \). That is, our analysis assumes that the number of strata are bounded but sample sizes within the strata increase. However, it is worth noting that, since data are independent across both strata and time, and sampling is in fact carried out independently outside each stratum, it is possible to extend our analysis to the case where \( T \) varies with \( N_{t,\text{min}}^\text{\text{}} \) at the cost of more involved algebraic manipulations. This is left for future research.
We now state some important assumptions used throughout the rest of this section:

**Assumption 1** The sub-population densities, \( f_{t,N_t,k} \) for \( t = 1, \ldots, T \) and \( k = 1, \ldots, L \) have partial derivatives up to order \( k^* + 1 \), satisfying a Lipschitz condition uniformly on \( \mathcal{X} \).

**Assumption 2** For any \( T, N_t \), and any fixed \( r \in \{1, \ldots, J\} \), there exist constants, \( 0 < C_{3,r} < \infty \) and \( 0 < C_{4,r} \leq C_{5,r} < \infty \) such that \( \min_{s=1,\ldots,T,s\neq r} |\lambda_r - \lambda_s| \geq C_{3,r}T \) and \( C_{4,r}T \leq \lambda_r \leq C_{5,r}T \).

**Assumption 3** Some conditions on the weight functions are given below.

1. The kernel weight \( K(u) \) used in the estimation procedure is a bounded probability density function with support, \( \{u : |u| \leq 1\} \), such that \( K(u) = K(-u) \), \( \int K^2(u)du < \infty \), and \( \int u^jK(u)du = 0 \) for all \( j = 1, \ldots, k^* \).

2. The weight function \( w(\cdot) \) in the inner product \( \langle \cdot, \cdot \rangle \) is a bounded continuous non-negative function supported on an open convex set of the real line.

**Assumption 4** Let \( N^{\min} = \min(N_1^{\min}, \ldots, N_T^{\min}) \); the bandwidth \( h \) verifies \( \sqrt{N^{\min}}h^{k^*+1} = o(1) \) and \( \sqrt{N^{\min}}h \to \infty \). In addition, \( \frac{N^{\max}}{N^{\min}} \to c < \infty \), where \( N^{\max} = \max(N_1^{\max}, \ldots, N_L^{\max}) \) with \( N_k^{\max} = \max(N_{1,k}, \ldots, N_{T,k}) \).

Assumption 1 is a stronger version of Assumption A1 in Kneip and Utikal (2001) in that it requires all sub-population densities and their partial derivatives to satisfy an uniform Lipschitz condition. Assumption 2 is the same as A2 in Kneip and Utikal (2001). Similarly, Assumption 3 is standard in the kernel smoothing literature, while Assumption 4 imposes a restriction on how fast \( h \) goes to zero as the number of subjects becomes large in each stratum and overall through time. The latter is stronger than the usual condition for pointwise consistency in the random sampling model; namely \( n_t h \to \infty \), and it is needed here because pointwise consistency is required for every possible finite population.

Because each \( \hat{M}_{ts} \) can be linearly represented in terms of the estimators \( \hat{M}_{ts}^* \) for \( M_{ts}^* = \langle f_{N_t}, f_{N_s} \rangle \), our proofs require the careful study the asymptotic properties of

\[
\hat{M}_{ts}^* - M_{ts}^* = \langle \hat{f}_{S_t,h} - f_{N_t}, f_{N_s} \rangle + \langle f_{N_t}, \hat{f}_{S_t,h} - f_{N_s} \rangle + \langle \hat{f}_{S_t,h} - f_{N_t}, \hat{f}_{S_s,h} - f_{N_s} \rangle - \frac{\delta_{ts}}{N_t^2} \sum_{k=1}^L w_{t,k}^2 \sum_{i=1}^{N_{t,k}} I_{t,k,i} \|U_{t,k,i}\|^2_w, \quad (8)
\]

and

\[
\hat{M}_{ts}^* - \hat{H}_{ts}^* = \langle \hat{f}_{S_t,h} - \hat{f}_{N_t,h}, \hat{f}_{N_s,h} \rangle + \langle \hat{f}_{N_t,h}, \hat{f}_{S_s,h} - f_{N_s,h} \rangle + \tau_{ts}, \quad (9)
\]

where \( \tau_{ts} = \langle \hat{f}_{S_t,h} - f_{N_t,h}, \hat{f}_{S_s,h} - f_{N_s,h} \rangle - \frac{\delta_{ts}}{N_t^2} \sum_{k=1}^L w_{t,k}^2 \sum_{i=1}^{N_{t,k}} I_{t,k,i} \|U_{t,k,i}\|^2_w - \frac{\delta_{ts}}{N_t^2} \sum_{k=1}^L \sum_{i=1}^{N_{t,k}} \|U_{t,k,i}\|^2_w \), and \( \| \cdot \|_w \) is the norm induced by the above-defined inner product.
Some algebraic calculations yield

\[\langle \hat{f}_{S_t,h} - \hat{f}_{N_t,h}, \hat{f}_{N_s,h} \rangle = \sum_{k=1}^{L} W_k S_{N_t,k}^{(1)},\]

\[\tau_{ts} = (1 - \delta_{ts}) \sum_{\ell=1}^{L} \left( \sum_{k=1}^{L} W_{t,k} W_{s,\ell} S_{N_t,k, N_s,\ell}^{(1)} - W_{s,\ell} S_{N_s,\ell}^{(1)} \right)\]

\[+ \delta_{ts} \left( \sum_{k=1}^{L} W_{t,k}^2 (S_{N_t,k}^{(2a)} - S_{N_t,k}^{(2b)}) + \sum_{k=1}^{L} \left( \sum_{\ell=1}^{L} W_{t,k} W_{t,\ell} S_{N_t,k, N_t,\ell}^{(1a)} - W_{t,k} S_{N_t,k}^{(3)} \right) \right),\]

where

\[S_{N_t,k}^{(1)} = \frac{1}{n_{t,k}} \sum_{i=1}^{N_{t,k}} \left( I_{t,k,i} - \frac{n_{t,k}}{N_{t,k}} \right) \langle U_{t,k,i}, \hat{f}_{N_s,h} \rangle,\]

(10)

\[S_{N_s,\ell}^{(1)} = \frac{1}{n_{s,\ell}} \sum_{i=1}^{N_{s,\ell}} \left( I_{s,\ell,i} - \frac{n_{s,\ell}}{N_{s,\ell}} \right) \langle U_{s,\ell,i}, \hat{f}_{N_t,h} \rangle,\]

(11)

\[S_{N_t,k, N_s,\ell}^{(1)} = \frac{1}{n_{t,k} n_{s,\ell}} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{s,\ell}} I_{t,k,i} \left( I_{s,\ell,j} - \frac{n_{s,\ell}}{N_{s,\ell}} \right) \langle U_{t,k,i}, U_{s,\ell,j} \rangle,\]

(12)

\[S_{N_t,k}^{(2a)} = \frac{1}{n_{t,k}^2} \sum_{i\neq j}^{N_{t,k}} \left( I_{t,k,i} - \frac{n_{t,k}}{N_{t,k}} \right) \langle U_{t,k,i}, U_{t,k,j} \rangle\]

(13)

\[S_{N_t,k}^{(2b)} = \frac{1}{n_{t,k}^2} \sum_{i\neq j}^{N_{t,k}} \left( I_{t,k,j} - \frac{n_{t,k}}{N_{t,k}} \right) \langle U_{t,k,i}, U_{t,k,j} \rangle,\]

(14)

\[S_{N_t,k, N_s,\ell}^{(1a)} = \frac{1}{n_{t,k} n_{s,\ell}} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{s,\ell}} I_{t,k,i} \left( I_{s,\ell,j} - \frac{n_{s,\ell}}{N_{s,\ell}} \right) \langle U_{t,k,i}, U_{t,\ell,j} \rangle,\]

(15)

\[S_{N_t,k}^{(3)} = \frac{1}{n_{t,k} n_{t,k}} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{t,k}} \left( I_{t,k,j} - \frac{n_{t,k}}{N_{t,k}} \right) \langle U_{t,k,i}, U_{t,\ell,j} \rangle.\]

(16)

The asymptotic behaviour of (10)–(16) is given in the proposition below (which effectively adapts Proposition 1 in Kneip and Utikal, 2001, p. 524). Our method of proof requires that for any random element, say \( V_n \), we define \( V_n = O_C(a_n) \) and \( V_n = o_C(b_n) \) if and only if \( \limsup_{n \to \infty} B \to \infty P_D \left( P_M \left( \frac{V_n}{a_n} > B \right) \right) = 0 \), where \( P_D \) is the probability induced by the stratified sampling design \( D \) and \( P_M \) is the probability induced by the super-population model \( M \), and \( \limsup_{n \to \infty} P_D \left( P_M \left( \frac{V_n}{a_n} > \epsilon \right) \right) = 0 \) for every \( \epsilon > 0 \) respectively. Similarly, for a sequence of random variables \( \{Z_m\} \), we say that \( "Z_m \to_d N(a_m, b_m)" \) if \( (Z_m - a_m) / b_m^{1/2} \) converges in distribution to a standard normal distribution, i.e. \( N(0, 1) \). We also use the standard convention
of defining $E_D[\cdot]$, and $V_D[\cdot]$ the unconditional expectation and variance under the stratified random sampling design, and as $E_M[\cdot]$, and $V_M[\cdot]$ the same but under the super-population model generating the stratified finite population.

**Proposition 1** Supposing that Assumptions 1, 3 and 4 hold. Then, as $N^{\text{min}} = \min(N_1^{\text{min}}, \ldots, N_T^{\text{min}}) \to \infty$,

$$
\widehat{M}_t^s - M_t^s = O_C \left( h^{N^{1/2}} + \frac{1}{N^{1/2}} \right) + o_C \left( \frac{1}{(N^{\text{min}})^{1/2}} + \frac{1}{h (N^{\text{min}})^{5/4}} \right).
$$

In addition, under the stratified sampling design $D$, $S_{N_t,k}^{(1)} \overset{d}{\to} N(0, \var_D(S_{N_t,k}^{(1)}))$, $S_{N_s,\ell}^{(1)} \overset{d}{\to} N(0, \var_D(S_{N_s,\ell}^{(1)}))$, $S_{N_t,k}^{(2a)} \overset{d}{\to} N(0, \var_D(S_{N_t,k}^{(2a)}))$, $S_{N_s,\ell}^{(2)} \overset{d}{\to} N(0, \var_D(S_{N_s,\ell}^{(2)}))$, and $S_{N_t,k}^{(3)} \overset{d}{\to} N(0, \var_D(S_{N_t,k}^{(3)}))$, where

$$
\var_D(S_{N_t,k}^{(1)}) = O_M(1/hN_t,k), \quad \var_D(S_{N_s,\ell}^{(1)}) = O_M(1/hN_s,\ell), \quad \var_D(S_{N_t,k}^{(2a)}) = O_M(1/hN_t,k), \quad \var_D(S_{N_s,\ell}^{(2)}) = O_M(1/hN_s,\ell), \quad \var_D(S_{N_t,k}^{(3)}) = O_M(1/hN_t,k).
$$

These variance terms are defined in the proof of this proposition in Appendix B in the supplemental material. The most important difference with the original result in Kneip and Utikal (2001) is that different rates are needed for consistency under the design-based, model-based or combined inference. This is because in design-based inference, one must have consistency under every possible finite sub-population and sample, while in model-based inference, Assumption 1 guarantees that misbehaving sub-populations will not occur.

The difference in rates of convergence also carries over to the following theorem establishing the design-based consistency of the estimated eigenvalues $\widehat{\lambda}_r$, and eigenvectors $\widehat{p}_r$ of $\widehat{M}$ for $r = 1, 2, \ldots$.

**Theorem 1** Let $h = (N^{\text{min}})^{-\theta}$ with $\frac{1}{2(k+1)} \leq \theta < \frac{1}{2}$. Suppose that Assumptions 1-3 hold. Then, for any fixed $r \in \{1, \ldots, J\}$,

$$
E_D^{1/2}[|\widehat{\lambda}_r - \lambda_r|^2] = O_M \left( \frac{T^{1/2}}{(N^{\text{min}})^{1/4}} \right),
$$

$$
E_D[\|p_r - \widehat{p}_r - S_r q_r\|_2] = O_M \left( \frac{1}{(N^{\text{min}})^{1/2}} \right),
$$

where $S_r = \sum_{s=1}^T \frac{1}{N_s - \lambda_r} p_s p_s^\top$ and the elements of the vector $q_r$ are given by

$$
q_{t,r} = \sum_{s=1}^T \frac{1}{N_s - \lambda_r} \left( \langle \widehat{f}_{S_t,h} - f_t, f_N - f_r \rangle - \langle \widehat{f}_{S,t} - f_t, f_N - f_r \rangle \right)
+ \langle \widehat{f}_{S_t,h} - f_N, f_N - f_r \rangle - \langle \widehat{f}_{S,t} - f_t, f_N - f_r \rangle.
$$
where \( \hat{f}_{S,h} = \frac{1}{T} \sum_{t=1}^{T} \hat{f}_{S,t,h} \);

\[
E_D[\hat{\theta}_{t,r} - \theta_{t,r}] = O_M \left( \frac{1}{(N_{\min})^{1/4}} \right);
\]  

(20)

and

\[
\sum_{r=J+1}^{T} (\hat{\lambda}_{r} - E_D[\hat{\lambda}_{r}]) = \sum_{t=1}^{T} \left( 1 - \frac{1}{T} - \sum_{r=1}^{J} p_{t,r}^2 \right) (\hat{M}_{t}^* - E_D[\hat{M}_{t}^*])
\]

\[
- \sum_{t \neq s} \left( \frac{1}{T} + \sum_{r=1}^{J} p_{t,r} p_{s,r} \right) (\hat{M}_{ts}^* - E_D[\hat{M}_{ts}^*]),
\]  

(21)

where \((\hat{M}_{t}^* - E_D[\hat{M}_{t}^*]) = \sum_{k=1}^{L} W_{t,k}^2 S_{N_{t,k},N_{t,k}}(4) + \sum_{k \neq \ell} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{t,\ell}} (I_{t,k,i} I_{t,\ell,j} - \frac{n_{t,k}(n_{t,k} - 1)}{N_{t,k}(N_{t,k} - 1)} \langle U_{t,k,i}, U_{t,\ell,j} \rangle) \langle U_{t,k,i}, U_{t,\ell,j} \rangle, \) and \(S_{N_{t,k},N_{t,k}}(4) = \frac{1}{n_{t,k} n_{t,k}} \sum_{i=1}^{N_{t,k}} \sum_{j=1}^{N_{t,k}} (I_{t,k,i} I_{t,\ell,j} - \frac{n_{t,k}(n_{t,k} - 1)}{N_{t,k}(N_{t,k} - 1)} \langle U_{t,k,i}, U_{t,\ell,j} \rangle) \langle U_{t,k,i}, U_{t,\ell,j} \rangle; \) moreover, under the stratified sampling distribution \(D,\)

\[
S_{N_{t,k},N_{t,k}}(4) \quad \Rightarrow \quad N(0, \text{var}_D(S_{N_{t,k},N_{t,k}}(4))),
\]  

(22)

\[
S_{N_{t,k},N_{t,\ell}}(2) \quad \Rightarrow \quad N(0, \text{var}_D(S_{N_{t,k},N_{t,\ell}}(2))),
\]  

(23)

\[
S_{N_{t,\ell},N_{t,\ell}}(2) \quad \Rightarrow \quad N(0, \text{var}_D(S_{N_{t,\ell},N_{t,\ell}}(2))),
\]  

(24)

where \(\text{var}_D(S_{N_{t,k},N_{t,k}}(4)) = O_M \left( \frac{1 + h}{N_{t,k}} + \frac{1}{N_{t,k}} + \frac{h}{N_{t,k}} \right), \) \(\text{var}_D(S_{N_{t,k},N_{t,\ell}}(2)) = O_M \left( \frac{1}{N_{t,k} N_{t,\ell}} + \frac{1}{N_{t,k}} + \frac{1}{N_{t,\ell}} \right), \) and \(\text{var}_D(S_{N_{t,\ell},N_{t,\ell}}(2)) = O_M \left( \frac{1}{N_{t,\ell} N_{t,\ell}} + \frac{1}{N_{t,\ell}} + \frac{1}{N_{t,\ell}} \right).\)

Finally, the next theorem provides the pointwise design-based asymptotic normality for the estimator \(\hat{g}_r\) for every \(r \in \{1, \ldots, J\},\) using the SWKD-based estimator of the basis function \(g_r\) defined in (4). Recall that the covariance matrix of the super-population kernel density estimator in (3.1) is \(\{\hat{H}_{ts}\}_{t,s=1}^{T}\) with \(\hat{H}_{ts} = \langle \hat{f}_{N,h} - \hat{f}_{N,h}, \hat{f}_{N,h} - \hat{f}_{N,h} \rangle - \frac{h}{N_{t,k}} \sum_{k=1}^{L} \sum_{i=1}^{N_{t,k}} \|U_{t,k,i}\|_2^2;\) i.e. (7), where \(\hat{f}_{N,h} = T^{-1} \sum_{t=1}^{T} \hat{f}_{N,t,h};\) and let \(\tilde{p}_{t,r} = (\tilde{p}_{t,1}, \ldots, \tilde{p}_{t,J})^\top\) for \(r = 1, \ldots, J\) and \(\tilde{\lambda}_r \geq \tilde{\lambda}_2 \geq \cdots \geq \tilde{\lambda}_J\) represent the eigenvectors and eigenvalues of \(\{\hat{H}_{ts}\}_{t,s=1}^{T}\) respectively. Define \(\hat{\theta}_{t,r} = \frac{\tilde{p}_{t,r}}{\sum_{t=1}^{T} \tilde{p}_{t,r}}, \)

\[
\hat{f}_{S,h}(x) = (\hat{f}_{S1,h}(x), \ldots, \hat{f}_{ST,h}(x))^\top, \quad \text{and} \quad S_{N_{t,k}}^{(5)}(x) = \frac{1}{n_{t,k}} \sum_{i=1}^{N_{t,k}} (I_{t,k,i} - \frac{n_{t,k}}{N_{t,k}}) U_{t,k,i}(x).
\]

**Theorem 2** Suppose that Assumptions 1-4 hold. Then, under the stratified sampling distribu-
tion $D$, 

$$
\frac{\tilde{g}_r(x) - \tilde{g}_r(x)}{\sqrt{\sum_{t=1}^T \sum_{k=1}^L W_{t,k}^2 E_D[(S_{N_t,k}^{(5)}(x))^2]}} \overset{d}{\to} N(0, 1).
$$

The asymptotic normality established in the last theorem is not surprising as 4 demonstrates that $\tilde{g}_r(\cdot)$ can be written as a linear combination of estimated densities that themselves have pointwise asymptotic normal distributions.

## 4 Results

We use the estimators described in the previous section to analyze the time evolution of the distribution of the standardized natural logarithm series of prices and their changes excluding series as previously described. We utilized the R packages npRmpi (cross-validation) and snow (numerical integration). All computations performed on EDITH, the Bank of Canada High Performance Cluster, which consists of 664 cores on 36 nodes.\(^{12}\)

Figure 3 utilizes Hyndman et al.’s (1996) stacked density plots and Hyndman’s (1996) highest density regions to display the resulting SWKD estimators for both variables each month. Month-specific bandwidths were chosen by Duin’s (1976) likelihood cross-validation using a second-order gaussian kernel throughout, i.e. $\hat{h}_{CV,t}$. This bandwidth selection mechanism is justified here because under the assumption that $n_{t,k}/N_{t,k} = n_t/N_t$ for all $k = 1, 2, \ldots$, the optimal bandwidth choice, $h$, to calculate \((1)\) has the same order of magnitude as the optimal bandwidth in the random sampling case, see Buskirk and Lohr (2005, p. 181).

We observe that the distributions of price levels (top panels) and price changes (bottom panels) exhibit complex dynamics. First, from month to month, distributions shift their entire mass as well as different density regions, captured by shaded areas on the right-hand side plots. Second, the relative sizes of density regions change across time, reflecting fluctuations in price dispersion. Third, changes in relative density regions are asymmetric from month to month, resulting in a seesaw shape pattern of the distribution (top left panel). Fourth, the densities at the peak and at the tails vary extensively (bottom left panel). Finally, fluctuations in density regions stand out during recessions. For example, the peak of price-level distributions is markedly higher during mid-2000’s, and the 50% density is smaller for price changes during both recessions.\(^ {13}\)

We proceeded to operationalize the estimators described in Section 3.3 by setting the weighting function $w(\cdot) = 1$ over the observed support of the entire sample (as permitted in Assumption

\[^{12}\] All jobs utilized about 250 cores; bandwidth cross-validation takes about 24 hours while numerical integration takes about an hour.

\[^{13}\] Using a specialized sub-sample of U.S. CPI micro data, Berger and Vavra (2014) finds that price changes become substantially more disperse during recessions.
3). The bandwidths, \( h \), in the right-hand side of (6) were calculated for each month and set equal to \( \tilde{h}^{5/4}_{CV,t} \). A total of \( T \times (T - 1)/2 = 23,653 \) univariate numerical integrals were calculated using an adaptive quadrature algorithm in Piessens et al. (1983). Results are shown in Figure 4.

### 4.1 Dynamic Scree Plots and Strength Components

The left-hand side plots in Figure 4 display the resulting dynamic scree plots or the contribution of the respective eigenvalue to the total eigenvalue. Visual inspection of the dynamic scree plot for the log(Price) reveals that the first four components accounts for about 12\% for log(Price) after which the dynamic scree plot is relatively flat. However, for the \( \Delta \)log(Price) the first four components accounts for 8\% of the variation.

For illustrative purposes, we also provide the estimated first four dynamic strength components, i.e. \( \theta_{t,r} \) defined in (5), on the right-hand side of Figure 4. All are normalized with respect to their starting values, so they portray deviations from the initial condition or \( \theta_{1996-03,r} \), for \( r = 1, \ldots, 4 \). The dynamics are different for price levels and changes; the first four dynamic strength components deviations for log(Price) are persistent and displays some non-stationarity while \( \Delta \)log(Price) are less persistent and stationary.

The first dynamic strength component for the log(Price) oscillates around zero until about 2007 and then starts to increase for the rest of the period with a first-order autocorrelation, AR(1), parameter of 0.98. This movement is a precursor to the rise in the unemployment during the Great Recession in the UK (shaded in yellow). The second component AR(1) is 0.95 and is positive for most of the sample period. By contrast, the third and fourth components have AR(1) parameters of 0.85 and 0.76. For \( \Delta \)log(Price) all dynamic strength components are stationary but persistent series with AR(1) parameters of 0.50, 0.56, 0.31, and 0.14, respectively.

By design, the dynamic strength components are statistical concepts, not economics; however, if the most important components correlate with macroeconomic variables, they can be given macroeconomic interpretation. Figure 5 displays the non-normalized first three estimated strength components for log(Price) and \( \Delta \)log(Price) along with the UK monthly unemployment rate and the inflation, respectively. For the log(Price), the sample correlation with the first three dynamics strength components and the unemployment rate are 0.69, -0.35, and 0.29, respectively. These highly persistent dynamic strength components for the log(Price) could be used to forecast the unemployment rate. For \( \Delta \)log(Price), the correlation of the first three dynamics strength components and the inflation rate is -0.19, 0.48, and 0.20, respectively. In this case, the correlation and persistence parameters are weaker so the forecasting ability is much lower.
4.2 Basis Functions and their Deformations

The dynamic strength components summarizes the dynamics of the distributions through time, but they do not speak directly on the parts of the distributions that are evolving. To understand this phenomena, we plot the first four bootstrap bias-corrected estimated basis functions, $\hat{g}_r$, $r = 1, \ldots, 4$, see, i.e. (6), along with their contributions to the overall distributional deformations, $\hat{\theta}_{t,j}g_r$, for three relevant time periods in Figures 6 and 7 for log(Price) and $\Delta$log(Price) respectively. Based on the point-wise asymptotic normality result in Theorem 2, a 95% bootstrap point-wise confidence interval based on 999 replications are also shown as gray areas in these plots. To mimic the original stratified sampling design in each month, the following bootstrap procedure was implemented:

1. For each month, $t = 1, \ldots, 217$, 199 bootstrap samples within strata are taken and bootstrap replication weights constructed as suggested in Canty and Davison (1999, Section 3.4, pp. 383-384).

2. Based on the 199 bootstrap samples and replication weights generated in the first step, bootstrap analogues of the original estimator of (4) are then constructed using the same bandwidths and kernel functions used for the original sample.

4.2.1 log(Price) Deformations

The first component basis function for log(Price) is estimated precisely in Figures 6. These same graphs also plot the corresponding deformation for three important periods of the British economy in the past 18 years; the beginning of two recessions in August 2009 and December 2011 as well as the beginning of a normal period of economic activity in April 2001. Interestingly, the shape of the first deformation at the onset of the recessions indicate that during periods of economic stress, the underlying baseline distribution of prices shifts towards higher prices (right-tail of price levels becomes fatter) and if prices change, they do it towards cheaper values (left-tail of price levels becomes fatter). The opposite seems to happen during periods of normal economic activity.

For the second basis function, recessions are associated with the shift of prices from the middle of the distribution toward the tails. Such shift is described as a selection effect by Golosov and Lucas (2007). The third basis function is imprecisely estimated and is almost equivalent to the deformations. The fourth basis function has a similar story to the second basis function but in this case the middle of the distribution is gaining at the expense of the tails.
4.2.2 $\Delta \log(\text{Price})$ Deformations

Note that only the second basis function for $\Delta \log(\text{Price})$ is estimated somewhat precisely in Figure 7 (top right plot). In the recessions, left and right tails of $\Delta \log(\text{Price})$ are gaining at the expense of the middle mass of distribution. In addition, the right tail has a larger weight relative to the left tail. This finding is consistent with the predictions from models generated by Midrigan (2011). The rest of the basis functions and deformations are virtually zero, so the changes over time are not statistically significant. This result underscores that the dynamics of distribution for $\Delta \log(\text{Price})$ manifests itself mostly via time-varying thickness of the tails of the distributions.

5 Concluding Remarks

We adapt the functional principal component analysis of Kneip and Utikal (2001) to accommodate complex survey designs to study the dynamics of price distributions over time. We apply the method to publicly available monthly data on individual-good prices collected in retail stores by the ONS for the construction of the U.K. CPI from March 1996 to March 2014. Evidence from this large price panels obtained with the help of this method can provide information about the dynamic evolution of the distribution of prices. We find that these dynamics are highly correlated with business cycle indicators such as the UK unemployment and inflation rate.

Further, understanding the dynamic evolution of price distributions can be used to understand the micro-founded principles in macroeconomic models used to understand the design and implementation of monetary policy. Furthermore, FPCA method can be applied to other micro data sets that are used to study the dynamics of household income, credit and asset holdings, household demographic, educational and employment characteristics, firm investment, firm and bank balance sheets.

6 Supplemental Material

In the supplemental appendix, we provide detailed mathematical proofs of the main theoretical results of the paper.

References


Figure 1: UK Monthly CPI & Inflation

Top graph plots the monthly CPI (source: http://www.ons.gov.uk), and the bottom graph plots the implied monthly inflation in percentages. The base year is 2005.

The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions. The red solid line represents zero.
Figure 2: Summary Statistics for the Logarithm of the Raw Price Data

Per-month Horvitz-Thompson estimates of the mean (black line), 1st quartile (gray line), median (blue line), and 3rd quartile (red line) of the logarithm of raw prices calculated using the relevant functions in the survey R package by Lumley (2004).

The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions.

\(a\) Per-month Horvitz-Thompson estimates of the mean (black line), 1st quartile (gray line), median (blue line), and 3rd quartile (red line) of the logarithm of raw prices calculated using the relevant functions in the survey R package by Lumley (2004).

\(b\) The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions.
Figure 3: Estimated Distributions of log(Price) & ∆ log(Price)

Left graphs utilize Hyndman et al.’s (1996) stacked density plots device, while the right-hand-side graphs plot Hyndman’s (1996) highest density regions per month with 50% (dark gray), 95% (gray), 99% (light gray) probability coverage, and the mode (●).

Bandwidths for each month were chosen by Duin’s (1976) likelihood cross-validation as readily available in the np R package by Hayfield and Racine (2007). The algorithm was randomly started 10 different times as to avoid finding local optima. A second-order gaussian kernel is used throughout.

The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions.
Figure 4: Estimated Dynamic Scree Plots & Strength Components: log(Price) & Δ log(Price)

The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions. The red solid line represents zero, while the blue solid line draws 1/T. The estimated dynamic scree plots are normalized to add up to one.

Each plotted estimated strength components are normalized to start at zero on 03/1996.
The light yellow areas highlights the UK’s late 2000s (04/2008-09/2009), and the double-dip (10/2011-06/2012) recessions.

Top plot: Left hand side axis is the scale for the non-normalized estimated strength components for log(Price) (The red solid line represents its zero), while the right hand side axis is the scale for the UK monthly unemployment rate measured in percentages.

Bottom plot: Left hand side axis is the scale for the non-normalized estimated strength components for \( \Delta \log(\text{Price}) \) (The red solid line represents its zero), while the right hand side axis is the scale for the UK monthly inflation as measured by the CPI measured in percentages.
Each plot displays bootstrap bias-corrected estimates of the first four basis functions, $\hat{g}_j(\cdot)$, for $j = 1, \ldots, 4$ (black solid line), as well as three deformations corresponding to the beginnings of two recessions, i.e. 08/2009 (dotted line) and 11/2011 (dot-dashed line), as well as the beginning a period of normal economic activity, i.e. 04/2001. The red solid line represents zero.

The light gray areas represent 95% point-wise bootstrap confidence intervals based on 999 replications as explained in the main text.

Canty and Davison’s (1999) bootstrap was implemented each month as readily available in the survey R package by Lumley (2004).
Figure 7: Estimated First Four Basis Functions & their Deformations for $\Delta \log(\text{Price})$

Each plot displays bootstrap bias-corrected estimates of the first four basis functions, $\hat{g}_j(\cdot)$, for $j = 1, \ldots, 4$ (black solid line), as well as three deformations corresponding to the beginnings of two recessions, i.e. 08/2009 (dotted line) and 11/2011 (dot-dashed line), as well as the beginning a period of normal economic activity, i.e. 04/2001. The red solid line represents zero.

The light gray areas represent 95% point-wise bootstrap confidence intervals based on 999 replications as explained in the main text.

Canty and Davison’s (1999) bootstrap was implemented each month as readily available in the survey R package by Lumley (2004).