International Journal of Forecasting 35 (2019) 1304–1317

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting of density functions with an application to cross-sectional and intraday returns

Piotr Kokoszka [a,*], Hong Miao [b], Alexander Petersen [c], Han Lin Shang [d]

[a] Department of Statistics, Colorado State University, United States
[b] Department of Finance and Real Estate, Colorado State University, United States
[c] Department of Statistics and Applied Probability, University of California, Santa Barbara, United States
[d] Research School of Finance, Actuarial Studies and Statistics, Australian National University, Australia

ARTICLE INFO

Keywords:
Compositional data analysis
Constrained functional time series
Density function forecasting
Log quantile density transformation

ABSTRACT

This paper is concerned with the forecasting of probability density functions. Density functions are nonnegative and have a constrained integral, and thus do not constitute a vector space. The implementation of established functional time series forecasting methods for such nonlinear data is therefore problematic. Two new methods are developed and compared to two existing methods. The comparison is based on the densities derived from cross-sectional and intraday returns. For such data, one of our new approaches is shown to dominate the existing methods, while the other is comparable to one of the existing approaches.

1. Introduction

There are many problems in which the prediction of future probability densities functions is useful; for example, the prediction of densities of income, fertility, mortality, and densities of several types of returns on financial assets. Motivated by the need to forecast cross-sectional and intraday returns, we propose two new methods of predicting densities, and compare them to two existing methods.

The importance of density forecasting in finance has long been recognized, and is stated aptly by Crnkovic and Drachman (1997, p. 47) as follows: "At the heart of market risk measurement is the forecast of the probability density functions (PDFs) of the relevant market variables ... a forecast of a PDF is the central input into any decision model for asset allocation and/or hedging ... therefore, the quality of risk management will be considered synonymous with the quality of PDF forecasts". Also, as was pointed out by Ross (2017), forecasting the density of financial returns allows us to recover the pricing kernel, the market risk premium and the probability of a catastrophe, and to construct model-free tests of the efficient market hypothesis. The recovery theorem of Ross (2017) enables one to determine the market's forecast of returns. In financial risk management, there is more interest in the density of the future returns, as better density forecasts can produce better risk forecasts in terms of the value-at-risk (VaR) or conditional-value-at-risk (CVaR) forecast. Lee, Xi, and Zhang (2014) proposed a multiplicative decomposition of the financial returns in order to improve the density forecasts of financial returns and show that the risk forecasts produced from the density forecast using the decomposition and maximum entropy are superior to the approaches used more broadly, especially in extreme tail events of a large loss. This paper enriches the set of tools that can be used for forecasting the densities that are relevant in financial applications. Examples of the most commonly studied densities are shown in Fig. 1.

While density functions, or alternative characterizations of their underlying distributions, may be thought of as elements of a Hilbert space, they do not constitute a *linear* subspace. If the densities are treated as elements of L^2, the predicted curves will be elements of L^2, but will not

[*] Correspondence to: Department of Statistics, Colorado State University, Fort Collins, CO 80523, United States.
E-mail address: Piotr.Kokoszka@colostate.edu (P. Kokoszka).

https://doi.org/10.1016/j.ijforecast.2019.05.007

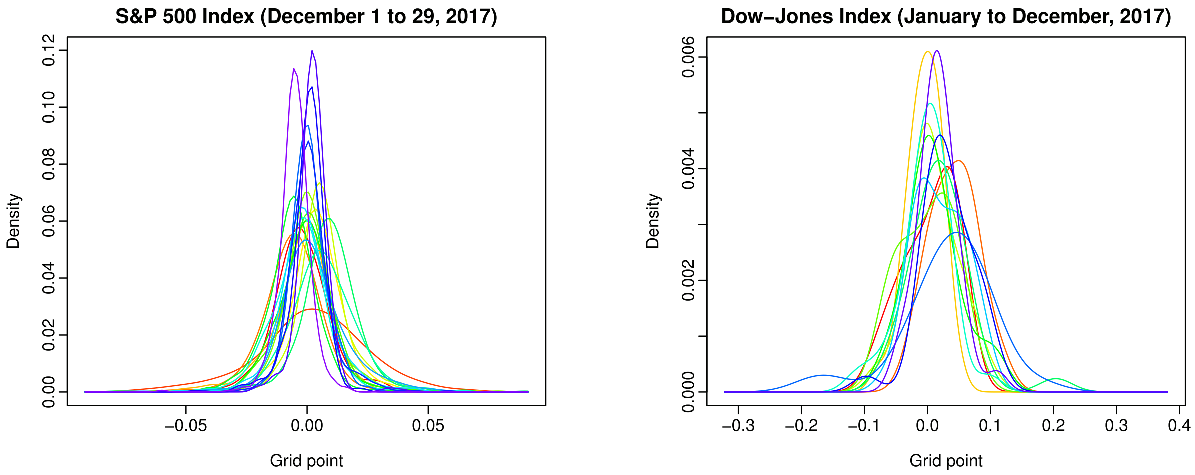

Fig. 1. Densities of cross-sectional monthly returns.

necessarily be densities. This is why the usual methods of prediction, which are based on Hilbert space formalism and go back to the work of Kolmogorov and Wiener, are not applicable directly. A natural way of dealing with such constraints is to peel them away by means of an invertible transformation that maps densities onto a linear space. Invertibility is key, as it allows one to perform analyses in the transformed space, then visualize and interpret the results in terms of densities by means of the inverse map. We propose two approaches that fall into this general paradigm: one based on compositional data analysis, the other on a suitable transformation of densities studied by Petersen and Müller (2016). Recent prediction methods for unconstrained curves are explained by Kokoszka and Reimherr (2017, Chapter 8), see also Hyndman and Ullah (2007) and Shang and Hyndman (2011). The forecasting of functional time series with constraints that differ from those considered here has been considered by Canale and Vantini (2016), for example.

Probability density functions have been studied from many angles within the broad framework of functional data analysis. Jones and Rice (1992) use kernel density estimation for the nonparametric estimation of density functions, and display a large functional dataset based on functional principal component (FPC) analysis. Kneip and Utikal (2001) use kernel density estimation to obtain annual income densities, and study the temporal evolution of income density functions in the United Kingdom from 1968 to 1988. Nerini and Ghattas (2007) consider regression trees where the responses are density functions, and use functional principal component analysis to interpret the main mode of variation in each terminal node. van der Linde (2008) proposes a Bayesian functional principal component analysis and applies it to a simulated dataset consisting of nonparametric density estimates. Delicado (2011) considers a compositional data analysis (CoDa) to analyze density functions and implement dimension-reduction techniques on the constrained compositional data space. Srivastava, Klassen, Joshi, and Jermyn (2011) consider a time-warping function in registration, where square root transformations of densities reside in the Hilbert space.

The paper is organized as follows. Section 2 describes the approaches that we study, with a focus on the two new approaches (Sections 2.1 and 2.2). Section 3 is dedicated to the comparison of the methods, with its last subsection summarizing the findings of this research.

2. Forecasting approaches

This section describes the forecasting approaches that we consider in this paper. The approaches based on compositional data analysis (Section 2.1) and the log quantile transformation (Section 2.2) are new. The approach introduced in Section 2.3 was proposed by Horta and Ziegelmann (2018) for forecasting the densities of intraday returns. The three approaches above observe the densities $f_t, t = 1, 2, \ldots, n$, and we want to forecast future densities $f_{n+h}, h \geq 1$. The parametric approach presented in Section 2.4, based on fitting skewed t-distributions, was proposed by Wang (2012) in the context of predicting monthly cross-sectional returns. It uses past returns rather than past densities as inputs.

2.1. Compositional data analysis

Density functions, which are nonnegative functions that integrate to one, share some features with compositional data (Aitchison, 1986; Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015). Compositional data are defined as a random vector of K nonnegative components, $D = [d_1, \ldots, d_K]$, the sum of which is a specified constant, typically set equal to 1 (portion), 100 (percentage) or 10^6 (parts per million). Thus, sample space of compositional data is the simplex

$$S^K = \left\{ D = (d_1, \ldots, d_K)^\top, \quad d_k > 0, \quad \sum_{k=1}^{K} d_k = c \right\},$$

where c is a fixed constant, and $^\top$ denotes the transpose. The simplex sample space is a $K - 1$ dimensional subset of the Euclidean space R^K.

Compositional data arise in many scientific fields, such as geology (geochemical elements), economics (income/

expenditure distribution), medicine (body composition), food industry (food composition), chemistry (chemical composition), agriculture (nutrient balance bionomics), environmental sciences (soil contamination), ecology (abundance of different species), and demography (life-table death counts). For example, Scealy, de Caritat, Grunsky, Tsagris, and Welsh (2015) study the concentration of chemical elements in sediment and rock samples, while (Scealy & Welsch, 2014) analyze households' total weekly expenditure on food and housing costs. In the field of demography, Boucher, Canudas-Romo, Oeppen, and Vaupel (2017) treat the age-specific life-table death count (d_k) as compositional data and use a principal component approach to forecast life-table death counts.

The method that we propose consists of the following steps.

1. Compute the geometric mean function

$$\alpha_n(u) = \exp\left\{\frac{1}{n}\sum_{t=1}^{n} \ln[f_t(u)]\right\} \tag{1}$$

   and set

$$s_t(u) = \frac{f_t(u)/\alpha_n(u)}{\int f_t(u)/\alpha_n(u)du}. \tag{2}$$

   The geometric mean standardizes the ranges, so that no range dominates the weighting. The standardization in Eq. (2) ensures that $\int s_t(u)du = 1$, a condition that is imposed commonly in compositional data analysis.

2. Apply the centered log-ratio transformation given by

$$\beta_t(u) = \ln\left(\frac{s_t(u)}{g_t}\right), \tag{3}$$

   where g_t is the geometric mean given by

$$g_t = \exp\left\{\int \ln[s_t(u)]du\right\}.$$

   The log-ratio transformation in Eq. (3) removes the constraints on f_t.

3. Apply FPC analysis to the transformed data $\{\beta_1(u), \dots, \beta_n(u)\}$, i.e., compute the Karhunen–Loéve expansions

$$\beta_t(u) = \sum_{\ell=1}^{n} \hat{\beta}_{t,\ell}\hat{\phi}_\ell(u) \approx \sum_{\ell=1}^{L} \hat{\beta}_{t,\ell}\hat{\phi}_\ell(u), \tag{4}$$

   where $\hat{\phi}_1(u), \dots, \hat{\phi}_L(u)$ are the first L estimated FPCs, and $\hat{\beta}_{t,1}, \dots, \hat{\beta}_{t,L}$ are their scores.

   We determine L by the explained variance criterion:

$$L = \text{argmin}_{L:L\geq 1}\left\{\sum_{\ell=1}^{L}\hat{\lambda}_\ell \Big/ \sum_{\ell=1}^{n}\hat{\lambda}_\ell \geq \delta\right\},$$

   where $\hat{\lambda}_\ell$ is the sample variance of the $\hat{\beta}_{t,\ell}$, $t = 1, 2, \dots, n$. In our implementation, we use $\delta = 85\%$ (see e.g. Horváth & Kokoszka, 2012, p. 41).

4. Forecast the L FPC scores. This can be done in many ways. We use an exponential smoothing forecasting method. Alternatively, a univariate ARIMA method or a multivariate time series forecasting method (Aue, Norinho, & Hörmann, 2015) can be used.

   Obtain the h-step-ahead forecast $\hat{\beta}_{n+h,\ell}$ of the ℓth principal component score. We utilize an automatic algorithm developed by Hyndman and Khandakar (2008) to determine the optimal exponential smoothing model based on the corrected Akaike information criterion (AIC) of Hurvich and Tsai (1993). Conditioning on the estimated principal components and the observed data, the forecast of $\beta_{n+h}(u)$ is given by

$$\hat{\beta}_{n+h|n}(u) = \sum_{\ell=1}^{L} \hat{\beta}_{n+h|n,\ell}\hat{\phi}_\ell(u). \tag{5}$$

5. Transform back to the compositional data; i.e., take the inverse centered log-ratio transformation given by

$$\hat{s}_{n+h|n}(u) = \frac{\exp[\hat{\beta}_{n+h|n}(u)]}{\int \exp[\hat{\beta}_{n+h|n}(u)]du},$$

   where $\hat{\beta}_{n+h|n}(u)$ denotes the forecasts in Eq. (5).

6. Finally, we add back the geometric means, to obtain the forecasts of the density function

$$\hat{f}_{n+h|n}(u) = \frac{\hat{s}_{n+h|n}(u)\alpha_n(u)}{\int \hat{s}_{n+h|n}(u)\alpha_n(u)du},$$

   where $\alpha_n(u)$ is the geometric mean function given in Eq. (1).

The functions β_t in Eq. (3) are given by

$$\beta_t(u) = \ln f_t(u) - \frac{1}{n}\sum_{t=1}^{n}\ln[f_t(u)] - \ln\int \frac{f_t(u)}{\alpha_n(u)}du.$$

The middle term on the right-hand side does not depend on t, and so has no impact on predictions. The third term is a normalizing constant that depends on t, and could potentially either improve the predictions or make them worse. Following the advice of a referee, we explored how a modification without this standardization would perform. This is equivalent to working with $\beta_t(u) = \ln(f_t(u))$, and using $\hat{f}_{n+h|n}(u) = \exp\hat{\beta}_{n+h|n}(u)$ instead of steps 5 and 6. We will refer to this approach as CoDa (no standardization), and to the approach described in steps 1–6 as CoDa (standardization).

A related approach can be found in a recent paper by Hron, Menafoglio, Templ, Hrüzová, and Filzmoser (2016), where the compositional data approach of Delicado (2011) was extended to define a version of functional principal component analysis on samples of densities. A related theoretical development is presented by Egozcue, Diaz-Barrero, and Pawlowsky-Glahn (2006). The method of Hron et al. (2016) cannot be applied to the densities we study, at least not without some fairly substantial modifications. Their setup is to begin with the collection of densities on an interval $[a, b]$, for which the log density is square integrable. This key requirement means that the density must be positive on $[a, b]$ for their methodology to work. The support of return densities (i.e., the collection of grid points for which the estimated density is strictly positive) is different for each month or day, even within the same data set, and therefore the method of Hron et al. (2016) cannot be applied directly without first adjusting the estimates to be strictly positive on a common interval.

2.2. Log quantile density transformation

Our approach is based on the ideas of Petersen and Müller (2016). The general paradigm is as follows. We transform the densities f_t, $t \le n$, to the space L^2 using a suitable transformation ψ, and set $Y_t = \psi(f_t)$. There are several prediction methods that are known to work well for functions in the "unrestricted" Hilbert space L^2. Applying one of these methods, we obtain the predicted curves $\widehat{Y}_{n+h}$, $1 \le h \le H$. We then apply the inverse transformation and obtain the predicted densities $\hat{f}_{t+h} = \psi^{-1}(\widehat{Y}_{t+h})$. The key difficulty in this approach is that we must ensure that $\psi^{-1}(\widehat{Y}_{t+h})$ is a density. In the context of functional principal components analysis, Petersen and Müller (2016) considered two specific transformations: the log hazard transformation and the log quantile density (LQD) transformation. We explain in Appendix B why the log hazard transformation is not suitable for our purpose, and focus on the LQD transformation. The main theoretical novelty of our transformation approach relative to that of Petersen and Müller (2016) is to define a modified LQD transformation that allows for densities with different supports to be transformed to functions on a common domain and analyzed jointly. Petersen and Müller (2016) studied the theoretical properties of FPCA after LQD transformation. The method has since been applied successfully in the context of functional regression models, where densities appear as predictors (Petersen, Chen, & Müller, 2019), as well as in a distributional regression framework in an engineering application (Chen, Bao, Li, & Spencer Jr, 2019). This paper is the first to demonstrate the utility of this approach for forecasting distributional time series. In particular, we are the first to provide algorithms for computing the modified LQD transformation and its inverse.

The density is only one of many characterizations of a distribution. Each representation has its own interpretation and properties, and we will demonstrate how these can be used to obtain a functional representation of the distribution that is free of nonlinear constraints. Let F_t and Q_t be the cumulative distribution function (cdf) and quantile functions corresponding to f_t, respectively, i.e.

$$F_t(x) = \int_{-\infty}^{x} f_t(u)du, \quad -\infty < x < \infty,$$

$$Q_t(s) = F_t^{-1}(s) = \inf\{y \in (0,1) : F_t(y) \ge s\}, \quad s \in (0,1).$$

Each of the three functions, f_t, F_t and Q_t, uniquely characterizes the same distribution, and can be used as a functional data object. However, they are all subject to nonlinear constraints that make the application of typically linear functional data methods inadequate, namely that $f_t \ge 0$, $\int_{-\infty}^{\infty} f_t(x)dx = 1$, $0 \le F_t(x) \le 1$, $F_t' \ge 0$ and $Q_t' \ge 0$, where $'$ denotes a first-order derivative. Of these three, Q_t is the least constrained, and it is easy to see that the so-called quantile density (Jones, 1992; Parzen, 1979; Tukey, 1965) $q_t = Q_t'$ is only constrained to be nonnegative. Thus, all constraints can be removed completely by computing the log quantile density (LQD)

$$Y_t(s) = \log(q_t(s)) = -\log(f_t \circ Q_t(s)), \quad s \in [0,1].$$

The last equality follows because

$$q_t(s) := Q_t'(s) = (F_t^{-1})'(s) = \frac{1}{F_t' \circ F_t^{-1}(s)} = \frac{1}{f_t \circ Q_t(s)}. \tag{6}$$

However, the transformation from f_t to its LQD is not invertible, since for any constant c, the quantile density of $f_t(\cdot - c)$, the density shifted f_t by c, is also q_t. Petersen and Müller (2016) assumed that the densities had a common known support $\mathcal{T}$, and developed a modification of the inverse transformation so that densities with support $\mathcal{T}$ could be mapped back and forth.

In our application, the densities do not always have the same support, so we are forced to deal with the non-invertibility issue in a different way. Although the supports differ in our data set, most have most of their mass concentrated around zero, and have bounded support with endpoints

$$\begin{aligned} r_t^+ &= \inf\{r > 0 : f_t(r) = 0\}, \\ r_t^- &= \sup\{r < 0 : f_t(r) = 0\}. \end{aligned} \tag{7}$$

Thus, in what follows, we assume that $0 \in \mathrm{supp}(f_t)$, which is the case for all densities that we consider. This will facilitate a modified definition of the log quantile density transformation that is indeed invertible, and thus suitable for our purposes. While the assumption $0 \in \mathrm{supp}(f_t)$ is natural for the data sets that we consider, this may be adapted to any sample of density functions for which the intersection of their supports contains a known point.

Definition A.1 in Appendix A defines a class $\mathcal{D}$ of densities for which the steps described in the remainder of this section are justified. This section provides an algorithmic description. We first present two algorithms that show how the forward (density to L^2) and backward (L^2 to density) transformations are executed. In fact, our transformations act between the space of densities and the product space $L^2[0,1] \times (0,1)$, as is stated in the following definition.

Definition 2.1. *For a density $f \in \mathcal{D}$, with $\mathcal{D}$ as in Definition A.1, let F and Q be the cdf and corresponding quantile function, and define*

$$Y(s) := -\log\{f(Q(s))\}, \quad s \in [0,1]. \tag{8}$$

The modified log quantile density transformation is the map

$$\psi : \mathcal{D} \to L^2[0,1] \times (0,1), \quad \text{given by} \quad \psi(f) = (Y, F(0)). \tag{9}$$

Note that this definition includes the possibility that $Y(0)$ and/or $Y(1)$ is/are infinite. Proposition A.1 ensures that $Y \in L^2[0,1]$ and that the transformation ψ is invertible. As was mentioned above, this definition can be generalized slightly under the assumption that there is a known point x_0 which lies in the support of all densities f_t. In this case, one would set $\psi(f) = (Y, F(x_0))$. An even more flexible alternative would be to choose a fixed percentile level $s_0 \in (0,1)$ and define $\psi(f) = (Y, Q(s_0))$, which would not require any overlap of the supports of the f_t. However, for the sake of simplicity and to conform with the motivating data examples, we consider the specific form in Eq. (9).

Let $f \in \mathcal{D}$ be a density whose values are known on a discrete grid $r^- = x_0 < x_1 < \cdots < x_L = r^+$, where $x_j = 0$ for some j. In what follows, c will represent the value $F(0)$, s_l will represent grid points on $[0,1]$, and $Y(s_l)$ is given by Eq. (8).

Algorithm 1 Forward transformation

```
Input: Data pairs $(x_l, f(x_l))$, $l = 0, \ldots, L$
Output: Data pairs $(s_l, Y(s_l))$, $l = 0, \ldots, L$ and $c$
For $l = 1, \ldots, L$:
    1. Compute $F(x_l) = \int_{x_0}^{x_l} f(x)dx$, $l = 1, \ldots, L$ by numerical integration of the pairs $(x_l, f(x_l))$.
    2. Define $s_l = F(x_l)$ so that $Q(s_l) = x_l$.
    3. Compute $Y(s_l) = -\ln(f(x_l))$.
    4. Find $j$ such that $x_j = 0$ and set $c = F(x_j) = F(0)$.
```

For the inverse transformation, one begins with a function g (corresponding to the output Y of Algorithm 1) that is continuous on $(0,1)$. Available are discrete observations $(s_l, g(s_l))$, where $0 = s_0 < s_1 < \cdots < s_L = 1$ is a grid. We also have a value $c \in (0,1)$ that specifies the value of the cdf at zero for the target density f. We assume that $s_j = c$ for some $j \neq 0$ and L. Thus, g is the predicted function, and c corresponds to $\widehat{F}_{n+h|n}(0)$.

Algorithm 2 Backward transformation

```
Input: Data pairs $(s_l, g(s_l))$, $l = 0, \ldots, L$ and $c$
Output: Data pairs $(x_l, f(x_l))$, $l = 0, \ldots, L$
For $l = 1, \ldots, L$:
    1. Find $j$ such that $s_j = c$.
    2. Compute $Q(s_l) = \int_{s_j}^{s_l} \exp\{g(s)\}ds$, $l = 1, \ldots, L$ by numerical integration of the pairs $(s_l, \exp\{g(s_l)\})$.
    3. Define $x_l = Q(s_l)$ so that $F(x_l) = s_l$.
    4. Compute $f(x_l) = \exp\{-g(s_l)\}$
```

Multiple LQD functions and/or densities can be computed using the above algorithms, and it is often desirable for these to be computed on a common grid. As an optional post-processing step to either algorithm, one can use linear interpolation to impute onto a common, equi-spaced grid of values.

With the above two algorithms, the full prediction procedure can be summarized in the following steps.

1. Transform the densities f_t into (Y_t, c_t), as in Algorithm 1, $t = 1, 2, \ldots, n$.
2. Compute the predictions $(\widehat{Y}_{n+h}, \hat{c}_{n+h})$, $h \geq 1$.
3. Transform each pair $(\widehat{Y}_{n+h}, \hat{c}_{n+h})$ into the predicted density $\hat{f}_{n+h|n}$, as in Algorithm 2.

We predict $Y_{n+h} \in \mathcal{D}$ by first obtaining a mean estimate $\hat{\mu}_Y(s)$ and covariance estimate $\widehat{C}_n$, from which we obtain eigenfunction estimates $\hat{v}_j$ and predicted FPC scores $\hat{\xi}_{n+h|n}$. The predicted value of the future LQD is then

$$\widehat{Y}_{n+h|n}(s) = \hat{\mu}_Y(s) + \sum_{j=1}^{p} \hat{\xi}_{n+h,j}\hat{v}_j(s),$$

which would at least be continuous on $(0,1)$, even if it diverges at the boundary. We use a scalar time series model to predict $c_{n+h} = F_{n+h}(0)$ using the time series of values $c_t = F_t(0)$, leading to a value $\hat{c}_{n+h|n}$. The prediction of the future density f_{n+h} is then given by

$$\hat{f}_{n+h|n} = \psi^{-1}(\widehat{Y}_{n+h|n}, \hat{c}_{n+h|n}),$$

where ψ^{-1} is the inverse transformation described in Algorithm 2.

2.3. Dynamic functional principal component regression

This method is based on the dynamic functional principal components (DFPCs) introduced by Bathia, Yao, and Ziegelmann (2010), which should not be confused with the spectral domain DFPCs developed by Hörmann, Kidziński, and Hallin (2015). The method is implemented exactly as described by Horta and Ziegelmann (2018), using their R software. We refer to their paper for its justification. It proceeds in the following steps.

1. Compute the kernel

$$\widehat{K}(u,v) = \frac{1}{(n-p)^2}\sum_{t,s=1}^{n-p}\sum_{k=1}^{p}\left[f_t(u) - \bar{f}(u)\right] \times \left[f_s(v) - \bar{f}(v)\right]\left\langle f_{t+k} - \bar{f}, f_{s+k} - \bar{f}\right\rangle$$

and its $\hat{d}$ orthonormal eigenfunctions $\hat{\psi}_1(\cdot), \ldots, \hat{\psi}_{\hat{d}}(\cdot)$, corresponding to nonzero eigenvalues.

2. Approximate the densities f_t, $t = 1, 2, \ldots, n$, by

$$\hat{f}_t(u) = \bar{f}(u) + \sum_{j=1}^{\hat{d}} \hat{\eta}_{tj}\hat{\psi}_j(u),$$

$$\hat{\eta}_{tj} = \int_{\mathcal{I}} \left[f_t(u) - \bar{f}(u)\right]\hat{\psi}_j(u)du.$$

3. Fit a vector autoregressive (VAR) model with the order selected by the AIC to the vectors

$$\hat{\boldsymbol{\eta}}_t = \left[\hat{\eta}_{t1}, \ldots, \hat{\eta}_{t\hat{d}}\right]^\top,$$

and compute the predicted vectors $\hat{\eta}_{n+h}$.

4. Compute the predictions

$$\hat{f}_{n+h|n}(u) = \bar{f}(u) + \sum_{j=1}^{\hat{d}} \hat{\eta}_{n+h,j}\hat{\psi}_j(u).$$

5. Since the predicted functions in the previous step do not have to be nonnegative and do not have to

integrate to one, set $f^{+}_{n+h|n}(u) = \max\left[0, \widehat{f}_{n+h}(u)\right]$ and

$$\hat{f}_{n+h|n}(u) = \frac{f^{+}_{n+h|n}(u)}{\int f^{+}_{n+h|n}(u)du}.$$

2.4. Skewed t distribution

As was noted above, this method has been used only in the context of the prediction of densities of cross-sectional returns. We follow exactly the approach of Wang (2012, Chapter 5), who also kindly provided the R code. Earlier work on modeling the cross-section of returns includes that by Lillo and Mantegna (2000) and Cont (2001). In particular, Cont (2001) emphasized that a parametric model must have at least four parameters in order to successfully reproduce the empirical properties: a location parameter, a scale parameter, a parameter that captures the decay in the tails, and an asymmetry parameter that allows the left and right tails to have different behaviors.

The skewed t distribution has been studied extensively. For example, Jones and Faddy (2003) proposed a tractable form and the associated likelihood inference. Azzalini and Capitanio (2003) constructed a skewed t distribution from a skewed normal distribution, while Fernández and Steel (1998) presented a general method for transforming any symmetric and unimodal distribution to a skewed distribution. Although these two methods are equivalent, we follow Fernández and Steel's (1998) method for constructing a skewed t distribution and fit it to the cross-sectional and intraday returns. The density function of this skewed t distribution is

$$f(u|\mu,\sigma,\nu,\lambda) = \begin{cases} \sqrt{(1-\lambda)(1+\lambda)}\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\sigma^2}\Gamma(\frac{\nu}{2})}\left(1+\frac{(u-\mu)^2(1+\lambda)}{\lambda^2\nu(1-\lambda)}\right)^{-\frac{\nu+1}{2}} \\ \quad \text{if } u \le \lambda; \\ \sqrt{(1-\lambda)(1+\lambda)}\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\sigma^2}\Gamma(\frac{\nu}{2})}\left(1+\frac{(u-\mu)^2(1-\lambda)}{\lambda^2\nu(1+\lambda)}\right)^{-\frac{\nu+1}{2}} \\ \quad \text{if } u \ge \lambda. \end{cases} \tag{10}$$

where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ is the scale parameter, $-1 < \lambda < 1$ is the skewness parameter, and $\nu > 0$ is degrees of freedom.

We estimate the four parameters by maximizing the logarithm of the log likelihood

$$\ln(L) = nC(\nu,\lambda) - \frac{\nu+1}{2}\sum_{u_i>\mu}\ln\left(1+\frac{(u_i-\mu)^2(1-\lambda)}{\lambda^2\nu(1+\lambda)}\right) \tag{11}$$

$$- \frac{\nu+1}{2}\sum_{u_i\le\mu}\ln\left(1+\frac{(u_i-\mu)^2(1+\lambda)}{\lambda^2\nu(1-\lambda)}\right),$$

where

$$C(\nu,\lambda) = \ln\frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi\sigma^2}\Gamma(\frac{\nu}{2})} + \frac{1}{2}\ln\frac{(1-\lambda)(1+\lambda)}{\nu\pi\lambda^2}.$$

The prediction method is summarized as follows.

1. For each $t = 1, 2, \ldots, n$, maximize the log likelihood in Eq. (11) to obtain the MLEs $\hat{\mu}_t, \hat{\sigma}_t, \hat{\nu}_t, \hat{\lambda}_t$.
2. Transform the estimated parameters to obtain an approximately stationary vector-valued time series
$$\boldsymbol{\theta}_t = [\theta_{t1}, \theta_{t2}, \theta_{t3}, \theta_{t4}]^\top, \quad t = 2, \ldots, n,$$
where
$$\theta_{t1} = \hat{\mu}_t, \quad \theta_{t2} = \ln\left(\frac{\hat{\sigma}^2_t}{\hat{\sigma}^2_{t-1}}\right),$$
$$\theta_{t3} = \ln\left(\frac{\zeta_t}{\zeta_{t-1}}\right), \quad \theta_{t4} = \hat{\lambda}_t,$$
and where ζ_t is the cdf of the standard t distribution with $\hat{\nu}_t$ degrees of freedom evaluated at 2; $\zeta_t = F_{\hat{\nu}_t}(2)$. (The introduction of ζ_t smooths out the jumps in the values of $\hat{\nu}_t$.)
3. Fit a VAR model to the $\boldsymbol{\theta}_t$, $t \le n$, with the order selected via the AIC. Compute the forecasts $\hat{\boldsymbol{\theta}}_{n+h}$, $h \ge 1$.
4. Using the components of $\hat{\boldsymbol{\theta}}_{n+h}$, compute the predicted parameters $\hat{\mu}_{n+h|n}, \hat{\sigma}_{n+h|n}, \hat{\nu}_{n+h|n}, \hat{\lambda}_{n+h|n}$ and the predicted densities
$$\hat{f}_{n+h|n}(u) = f\left(u|\hat{\mu}_{n+h|n}, \hat{\sigma}_{n+h|n}, \hat{\nu}_{n+h|n}, \hat{\lambda}_{n+h|n}\right).$$

3. Comparison of the four prediction methods

3.1. Data sets

We consider four financial data sets, two consisting of cross-sectional returns and two of five-minute intraday returns. Both are amongst the most extensively studied forms of financial data, with hundreds of contributions to date; see Harvey, Liu, and Zhu (2016) for a recent review of cross-sectional returns and Renault (2017) for intraday returns.

3.1.1. Dow Jones cross-sectional returns

The Dow Jones Industrial Average (DJIA) is a stock market index that shows how 30 large publicly-owned companies based in the United States have traded during a standard NYSE trading session. We consider monthly cross-sectional returns from April 2004 to December 2017. The data were obtained from the CRSP database (Center for Research in Security Prices). We thus have a sample of log-price observations, denoted by $(q_{j,t}, j = 1, \ldots, 30)$, for each month $t = 1, \ldots, n-1$. We define the jth return in month t as

$$r^*_{j,t} := q_{j,t+1} - q_{j,t}, \qquad j = 1, 2, \ldots, 30,$$

where $r^*_{j,t}$ is the log return for the jth company at the close of month t.

We can then estimate the density f_t for each month t from these data, as is explained in Section 3.2. Fig. 2 presents three density functions for three representative months that have mostly negative, mostly positive and close to zero returns, respectively.

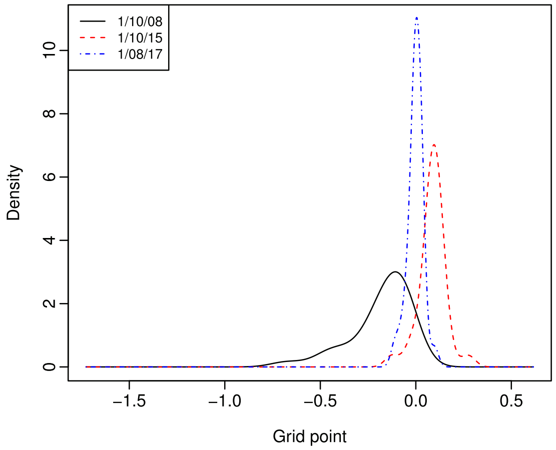

Fig. 2. Kernel density estimates with a Gaussian kernel and bandwidth selected by Silverman's rule of thumb for the monthly Dow-Jones index from April 2004 to December 2017. The solid black line shows a density function of one representative month that has mostly negative returns; the red dashed line shows a density function of one representative month that has mostly positive returns; the blue dash-dotted line shows a density function of one representative month that has returns close to zero.

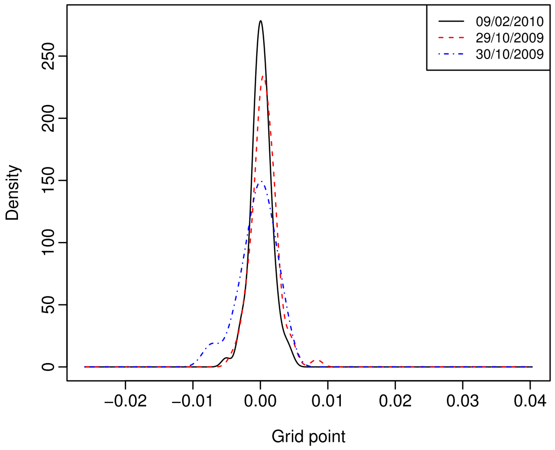

Fig. 3. Three density estimates, obtained in the same way as the estimates in Fig. 2, for the Bovespa intraday return. These densities are leptokurtic and are generally centered around zero.

3.1.2. S&P 500 cross-sectional returns

The Standard & Poor's 500 (SPX) index is an American stock market index based on the market capitalizations of 500 large companies with common stocks listed on the NYSE or NASDAQ. We consider monthly cross-sectional returns from April 2004 to December 2017. These data were obtained from the CRSP database.

3.1.3. Bovespa intraday returns

These data, which cover 305 trading days from September 1, 2009, to November 6, 2010, were made available by Capse Investimentos. This is exactly the data set that was used by Horta and Ziegelmann (2018). The tick-by-tick series is sampled at 5-minute intervals, such that we have a sample $(p_{i,t}, i = 1, \dots, m_t + 1)$ for each day $t = 1, \dots, n$, where $m_t + 1$ is the number of log-price observations within day t. We define the ith 5-minute return on day t as

$$r_{i,t} := p_{i+1,t} - p_{i,t}, \qquad i = 1, \dots, m_t. \tag{12}$$

For each day t, the 5-minute returns in Eq. (12) are distributed according to some density f_t, which we estimate as is explained in Section 3.2. Examples of such densities are given in Fig. 3.

3.1.4. XLK intraday returns

These intraday returns are constructed in the same way as the Bovespa returns but the underlying asset is XLK, the Technology Select Sector SPDR Fund. The time period is exactly the same as for the Bovespa data.

3.2. Density estimation

The true densities are not observable. We work with densities which are outputs of kernel estimators:

$$f_t(u) = \frac{1}{n_t h_t} \sum_{i=1}^{n_t} K\left(\frac{u - r_{i,t}}{h_t}\right), \qquad t = 1, \dots, n,$$

where $K(x)$ is a kernel and h_t is a bandwidth. We consider two commonly-used kernels, Gaussian and Epanechnikov, given by

$$K(u) = \left(\sqrt{2\pi}\right)^{-1} \exp^{-u^2/2}, \qquad \text{(Gaussian kernel)}$$

$$K(u) = \frac{3}{4}\left(1 - u^2\right), \qquad |x| \leq 1. \qquad \text{(Epanechnikov kernel)}$$

We select the bandwidths using Silverman's rule of thumb (ROT), which leads to

$$\hat{h}_t = 1.06 \times \hat{\sigma}_t \times n_t^{-1/5}, \qquad \text{(Gaussian kernel)}$$

$$\hat{h}_t = 2.34 \times \hat{\sigma}_t \times n_t^{-1/5}, \qquad \text{(Epanechnikov kernel)}$$

where $\hat{\sigma}_t$ denotes the sample standard deviation of the returns $r_{i,t}$, $i = 1, \dots, n_t$. We also consider the direct plug-in (DPI) method of Sheather and Jones (1991) for bandwidth selection.

We compare the predicted future density $\hat{f}_{n+h|n}$ with its estimate f_{n+h} obtained using the above methods. Section 3.3 explains the measures of the discrepancy that we use.

3.3. Measures of forecast accuracy

We measure the difference between the forecast density and the estimated future density by considering the discrete version of the Kullback–Leibler divergence (KLD; see Kullback & Leibler, 1951), the square root of the Jensen–Shannon divergence (JSD; see Shannon, 1948), and the mean L_p-norms with $p = 1, 2, \infty$. The KLD is designed to measure the loss of information when we choose an approximation. For two probability density functions, denoted by $g_t(u)$ and $\hat{g}_t(u)$, the discrete version of the KLD is given by

$$\begin{aligned}\mathrm{KLD} &= D_{\mathrm{KL}}\left(g_t \parallel \hat{g}_t\right) + D_{\mathrm{KL}}\left(\hat{g}_t \parallel g_t\right)\\ &= \sum_{i=1}^{N} g_t(u_i)\cdot\left[\ln g_t(u_i) - \ln \hat{g}_t(u_i)\right]\\ &\quad + \sum_{i=1}^{N} \hat{g}_t(u_i)\cdot\left[\ln \hat{g}_t(u_i) - \ln g_t(u_i)\right],\end{aligned}$$

which is symmetric and nonnegative.

An alternative is given by the JSD, defined as

$$\mathrm{JSD} = \frac{1}{2}D_{\mathrm{KL}}\left(g_t \parallel \delta_t\right) + \frac{1}{2}D_{\mathrm{KL}}\left(\hat{g}_t \parallel \delta_t\right), \tag{13}$$

where δ_t measures a common quantity between g_t and $\hat{g}_t$. We consider the simple mean and the geometric mean, given by $\delta_t = \frac{1}{2}\left(g_t + \hat{g}_t\right)$ or $\delta_t = \sqrt{g_t\hat{g}_t}$. The JSD is locally proportional to the Fisher information metric, and is similar to the Hellinger metric, in the sense that it induces the same affine connection on a statistical manifold, and is equal to half the so-called Jeffreys divergence. We make the JSD a metric between any two probability densities by taking its square root (see e.g. Fuglede & Topsøe, 2004).

Let $\hat{\epsilon}_t(x)$ be the difference between $g_t(x)$ and $\hat{g}_t(x)$; then, the L_1-norm, L_2-norm and L_∞-norm are

$$\begin{aligned}|\hat{\epsilon}|_1 &= \frac{1}{n_1}\sum_{t=1}^{n_1}\sum_{i=1}^{N}|\hat{\epsilon}_t(x_i)|,\\ |\hat{\epsilon}|_2 &= \frac{1}{n_1}\sum_{t=1}^{n_1}\sqrt{\sum_{i=1}^{N}[\hat{\epsilon}_t(x_i)]^2},\\ |\hat{\epsilon}|_\infty &= \frac{1}{n_1}\sum_{t=1}^{n_1}\max_i|\hat{\epsilon}_t(x_i)|,\end{aligned}$$

where n_1 denotes the forecasting period.

3.4. Tables of forecast accuracy measures

We use the expanding window approach, which is fairly standard for comparisons of this type (see e.g. Zivot & Wang, 2006, Chapter 9). Parameter estimates and their forecasts are computed over an expanding window, with the start point being at the beginning of the sample and the end point starting at the end of the first sample, which is of a reasonable length, and moving forward one step at a time until it reaches the end of the sample. This gives one fewer forecasts than end points.

For the Dow Jones and S&P 500 data, which cover the same period of time, we proceed as follows. We use the first 110 density estimates, from April 2004 to May 2013, to produce a one-step-ahead forecast of the June 2013 density, which we treat as an unknown future density. Next, using an expanding-window approach, we re-estimate the parameters in the density forecasting methods using the first 111 estimated densities, from April 2004 to June 2013, and compute the one-step-ahead forecast of the density with monthly index 112. We iterate this process by increasing the sample size by one month at a time until we reach the end of the data period in December, 2017. This process produces 55 one-step-ahead forecasts, which we compare to the actual density estimates. Note that we do not observe true densities in practice, but can construct density estimates from the available data.

Table 1 compares the density estimation accuracies of the four methods for the Dow-Jones cross-sectional returns. Of the two kernel functions, it is generally advantageous to use the Epanechnikov kernel function because of the superiority of its estimation accuracy over that of the Gaussian kernel function. Of the two bandwidth selection methods, it is better to use the rule-of-thumb because of its superior estimation accuracy and high computational speed relative to the plug-in method. The CoDa method performs the best as measured by the Kullback–Leibler divergence and the Jensen–Shannon divergence with the geometric mean, while the log quantile density transformation method performs the best as measured by the Jensen–Shannon divergence with the simple mean, L_1-norm, L_2-norm and L_∞-norm. The difference in accuracy between the CoDa methods with and without standardization is small, though the CoDa method without standardization has a slight edge.

Table 2 compares the density estimation accuracies of the four methods for the S&P 500 cross-sectional returns. Of the two kernel functions, it is better to use the Epanechnikov kernel function because of its superior estimation accuracy relative to the Gaussian kernel function. Of the two bandwidth selection methods, it is better to use the rule-of-thumb because of its superior estimation accuracy and fast computational speed relative to the plug-in method. As measured by the Jenson-Shannon divergence with a simple mean, the log quantile density transformation method generally performs the best, while the CoDa method performs the best as measured by the remaining criteria. The difference in accuracy between the CoDa methods with and without standardization is small.

Regarding the Bovespa and XLK data, we start with the first 203 densities from September 1, 2009, to June 30, 2010. We then increase the sample size by one day at a time until we reach the end of the data period on November 26, 2010. This process produces 102 one-step-ahead forecasts, which we compare to the actual density estimates.

Table 3 compares the density estimation accuracies of the four methods for the Bovespa intraday returns. Of the two kernel functions, it is generally better to use the Epanechnikov kernel function because of its superior estimation accuracy relative to the Gaussian kernel function. Of the two bandwidth selection methods, it is better to use the rule-of-thumb because of its superior estimation accuracy and fast computational speed relative to the plug-in method. The log quantile density transformation method generally performs the best as measured by the Jensen–Shannon divergence with a simple mean, while the CoDa method generally performs the best as measured by the Kullback–Leibler divergence. The difference in accuracy between the CoDa methods with and without standardization is small.

Table 4 compares the density estimation accuracies of the four methods for the XLK intraday returns. Of the

Table 1
Forecast accuracies of the four methods, Dow-Jones cross-sectional returns.

Parameter			JSD		norm		
selection	Density forecast method	KLD	Simple	Geometric	L_1	L_2	L_∞
Gaussian kernel							
ROT	Horta-Ziegelmann	1.3070	3.5986	9.4038	1039.36	46.96	3.87
	LQDT	1.0421	**3.0129**	6.9443	948.77	**45.83**	3.65
	CoDa (standardization)	0.6658	3.2359	5.1780	953.42	46.67	3.66
	CoDa (no standardization)	**0.6510**	3.1785	**5.0572**	**943.62**	46.19	**3.62**
	Skewed-*t*	1.3590	5.2532	10.4784	1324.97	64.40	5.19
DPI	Horta-Ziegelman	1.4751	4.0955	10.6577	1144.32	53.76	4.84
	LQDT	1.2569	**3.4896**	6.7053	**1043.12**	**52.36**	**4.40**
	CoDa (standardization)	0.8225	3.8350	6.4700	1055.21	54.02	4.46
	CoDa (no standardization)	**0.8088**	3.7884	**6.3581**	1049.78	53.74	4.43
	Skewed-*t*	1.5841	5.6072	12.0493	1369.45	67.97	5.73
Epanechnikov kernel							
ROT	Horta-Ziegelmann	1.4725	2.2652	8.3987	756.94	29.23	1.79
	LQDT	1.1833	**2.0165**	7.0648	**720.00**	**29.02**	**1.77**
	CoDa (standardization)	**0.6998**	2.3528	5.5992	740.74	30.51	1.91
	CoDa (no standardization)	0.7004	2.3397	**5.5767**	740.22	30.47	1.90
	Skewed-*t*	1.5595	5.4505	11.5386	1423.94	65.65	5.38
DPI	Horta-Ziegelman	1.7475	2.5834	9.8658	835.16	33.79	2.27
	LQDT	1.1197	**2.1370**	6.7053	**757.23**	**32.64**	**2.17**
	CoDa (standardization)	0.7495	2.7820	6.0593	844.10	37.33	2.57
	CoDa (no standardization)	**0.7476**	2.7679	**6.0312**	845.94	37.41	2.59
	Skewed-*t*	1.6668	5.2563	11.7497	1367.93	63.80	5.15

Note: The smallest error for each choice of kernel function, bandwidth selection method and evaluation criterion is highlighted in bold.

Table 2
Forecast accuracies of the four methods, S&P 500 cross-sectional returns.

Parameter			JSD		norm		
selection	Density forecast method	KLD	Simple	Geometric	L_1	L_2	L_∞
Gaussian kernel							
ROT	Horta-Ziegelmann	0.5315	1.9986	3.1032	222.62	17.00	2.33
	LQDT	0.4252	1.8165	2.5232	213.10	17.22	2.31
	CoDa (standardization)	**0.3156**	**1.7994**	**2.3023**	**208.71**	**16.61**	**2.19**
	CoDa (no standardization)	0.3233	1.8465	2.3550	211.29	16.81	2.21
	Skewed-*t*	0.5560	3.0961	3.6383	286.04	23.54	3.26
DPI	Horta-Ziegelman	0.6525	2.3241	3.8281	242.33	19.30	2.82
	LQDT	0.5292	**2.0994**	3.1572	**230.07**	19.33	2.75
	CoDa (standardization)	0.4152	2.2679	3.1047	232.78	19.22	2.66
	CoDa (no standardization)	**0.4149**	2.2681	**3.1010**	232.52	**19.18**	**2.65**
	Skewed-*t*	0.6365	3.2910	4.0951	290.11	24.18	3.40
Epanechnikov kernel							
ROT	Horta-Ziegelmann	0.6080	1.7090	3.1196	197.00	13.10	1.50
	LQDT	0.4065	**1.5867**	2.2334	190.29	13.26	1.48
	CoDa (standardization)	0.2939	1.5933	2.1443	188.15	**13.09**	**1.44**
	CoDa (no standardization)	**0.2925**	1.5916	**2.1278**	**188.14**	13.10	1.45
	Skewed-*t*	0.6496	3.5922	4.1716	323.77	25.88	3.74
DPI	Horta-Ziegelman	0.5920	1.8191	3.0113	209.05	15.53	1.97
	LQDT	0.4423	**1.6529**	2.4081	199.37	15.52	1.93
	CoDa (standardization)	0.3289	1.7698	2.4310	199.87	15.35	1.86
	CoDa (no standardization)	**0.3250**	1.7493	**2.3986**	**198.84**	**15.27**	**1.85**
	Skewed-*t*	0.5921	2.9930	3.5924	281.16	23.04	3.17

two kernel functions, it is better to use the Epanechnikov kernel function because of its superior estimation accuracy relative to the Gaussian kernel function. Of the two bandwidth selection methods, it is better to use the rule of thumb because of its superior estimation accuracy and high computational speed relative to the plug-in method. When the kernel function is the Epanechnikov kernel, the log quantile density transformation method generally performs the best. However, when the kernel function is the Gaussian kernel, the skewed-*t* distribution produces the smallest Kullback–Leibler divergence and Jensen–Shannon divergence with the geometric mean, while the log quantile density transformation method produces the smallest errors as measured by the Jensen–Shannon divergence with the simple mean and almost all of the distance metrics.

Table 3
Forecast accuracies of the four methods, Bovespa intraday returns.

Parameter			JSD		norm		
selection	Density forecast method	KLD	Simple	Geometric	L_1	L_2	L_∞
Gaussian kernel							
ROT	Horta-Ziegelmann	0.4009	1.9098	6.1713	16993.19	**977.07**	116.99
	LQDT	0.4258	**1.6634**	6.0687	**16313.87**	979.59	**116.70**
	CoDa (standardization)	**0.2271**	1.7360	**3.70**	16351.17	988.20	117.39
	CoDa (no standardization)	0.2278	1.7448	3.7038	16391.76	989.64	117.47
	Skewed-*t*	0.2750	1.9909	3.9774	19261.90	1186.31	147.91
DPI	Horta-Ziegelmann	0.4888	**1.9968**	7.5246	18896.40	**1164.28**	156.96
	LQDT	0.5947	2.0552	8.6969	**18405.96**	1170.16	**155.07**
	CoDa (standardization)	0.3279	2.2459	5.5030	18790.69	1208.02	159.40
	CoDa (no standardization)	**0.3257**	2.2425	5.45	18820.87	1212.29	160.58
	Skewed-*t*	0.3656	2.2396	**5.3992**	20360.42	1293.39	171.70
Epanechnikov kernel							
ROT	Horta-Ziegelmann	0.6486	1.7364	8.7790	15791.96	**759.33**	66.98
	LQDT	0.4695	**1.5042**	**6.3631**	**15224.69**	762.86	**66.59**
	CoDa (standardization)	0.3979	2.1160	6.8356	15961.10	802.75	68.89
	CoDa (no standardization)	**0.3940**	2.1265	6.7361	16003.43	804.04	68.63
	Skewed-*t*	0.5136	4.4523	6.4589	34469.29	2070.45	252.45
DPI	Horta-Ziegelmann	0.7665	2.4774	12.4275	**17035.77**	**904.35**	93.21
	LQDT	0.6004	**1.6917**	8.2761	17203.12	909.11	**93.02**
	CoDa (standardization)	0.4968	2.4988	8.5900	17754.95	971.08	97.27
	CoDa (no standardization)	0.4949	2.5266	8.5308	17866.92	976.65	97.73
	Skewed-*t*	**0.4744**	3.0147	**6.0645**	26348.37	1610.32	200.54

Table 4
Forecast accuracies of the four methods, XLK intraday returns.

Parameter			JSD		norm		
selection	Density forecast method	KLD	Simple	Geometric	L_1	L_2	L_∞
Gaussian kernel							
ROT	Horta-Ziegelman	0.2831	1.5095	4.2809	11257.47	**680.30**	83.59
	LQDT	0.3831	**1.3411**	5.2559	**10891.16**	682.67	**83.18**
	CoDa (standardization)	0.3231	2.6076	4.9518	14689.67	877.64	107.51
	CoDa (no standardization)	0.3579	2.8919	5.2173	15053.57	907.20	113.52
	Skewed-*t*	**0.2666**	1.7418	**3.8736**	13701.89	882.55	115.80
DPI	Horta-Ziegelman	0.3571	1.7629	5.4100	12348.01	777.02	113.04
	LQDT	0.4790	**1.5513**	6.6771	**11821.43**	**771.23**	**103.77**
	CoDa (standardization)	0.4489	3.5761	6.7166	17335.82	1094.71	161.35
	CoDa (no standardization)	0.4541	3.4848	6.5883	16403.03	1036.22	154.35
	Skewed-*t*	**0.3239**	1.8855	**4.7631**	14099.58	920.55	125.05
Epanechnikov kernel							
ROT	Horta-Ziegelman	0.5824	1.5024	7.6044	10933.73	550.49	51.10
	LQDT	**0.3909**	**1.2324**	**5.1780**	**10195.88**	**536.08**	**49.40**
	CoDa (standardization)	1.0400	7.6509	14.2193	24178.29	1406.33	222.95
	CoDa (no standardization)	1.0427	7.6440	14.2826	24174.29	1408.00	199.82
	Skewed-*t*	0.4823	3.9036	5.9900	24524.30	1540.05	195.18
DPI	Horta-Ziegelman	0.5614	1.5377	7.4848	11373.07	611.13	63.45
	LQDT	**0.4337**	**1.2728**	5.7667	**10706.14**	**593.15**	**60.24**
	CoDa (standardization)	1.0997	8.1558	15.1706	25875.30	1647.52	318.18
	CoDa (no standardization)	1.0829	7.7589	15.5553	25432.32	1520.35	222.01
	Skewed-*t*	0.4483	2.8957	**5.5946**	19887.16	1256.74	161.24

3.5. Model confidence set

The model confidence set procedure proposed by Hansen, Lunde, and Nason (2011) consists of a sequence of tests that permit the construction of a set of "superior" models for which the null hypothesis of equal predictive ability (EPA) is not rejected at a specified confidence level. The EPA test statistic can be evaluated for any arbitrary loss function. The model confidence set (MCS) procedure is a sequential testing procedure that eliminates the worst model at each step until the hypothesis of equal predictive ability is accepted for all of the models that belong to a set of superior models. The selection of the worst model is determined by an elimination rule that is consistent with the test statistic. This paper uses the $T_{\max}$ test statistic, which is a default statistic in the R package MCS and performed well in the study by Shang and Haberman (2018). With the bandwidth being selected by the rule-of-thumb and for both kernel functions, we use the MCS procedure to select a set of superior models for the four data sets. For each of the five error criteria, Table 5 indicates the superior set of models with daggers, †.

Table 5 can be summarized as follows. With four data sets, two different kernel functions and six forecast error criteria, there are 48 cases. The numbers of cases in which each specific method is superior are as follows: (1) LQDT: 24; (2) CoDa (no standardization): 14; (3) CoDa (standardization): 13; (4) dynamic FPCR: 5; (5) skewed-*t*: 2. Differentiating between the two types of densities, we see that the LQDT and CoDa methods are both recommended for the monthly cross-sectional returns. For the daily intraday returns, the LQDT is recommended.

3.6. Conclusions

We have compared the performances of five methods using four raw data sets, from which eight data sets consisting of densities have been constructed. These eight data sets cannot lead to a complete picture, but they already show a fairly clear pattern. With the exception of a focus on the densities of returns in finance, the data sets have not been selected using any specific criteria, and we have not experimented with any other data sets, so the results presented in Section 3.4 are not a selection of “favorable" results. We used six forecast quality criteria, meaning that each of the five methods is compared in $8 \times 6 = 48$ scenarios.

The predictions are almost always better when we use ROT rather than DPI, so our summary of the results in Section 3.4 focuses on the criteria computed for the bandwidth selected by ROT. The LQDT method is the best in 25 of these scenarios, while one of the two CoDa methods is the best in 17 scenarios. The CoDa (no standardization) works a little better. The performance of the Horta-Ziegelmann method is practically the same as the CoDa methods; it is either the best or the second best in 50% of cases. For some data, the skewed-*t* method performs well according to the KLD criterion. In many cases, the differences between the measures of forecast accuracy are small. However, a pattern that has emerged suggests that the LQDT method, which has been developed both theoretically and numerically in this paper, should be used as the first choice for the purpose of the prediction of cross-sectional and intraday returns. This is confirmed further by the MCS analysis in Section 3.5, where we noted that the LQDT and CoDa methods are both recommended for the monthly cross-sectional returns.

The rankings of the methods may be different for other time series of densities, e.g. those arising in population or medical research. It is hoped that the algorithmic descriptions and background presented in this paper will allow researchers in fields other than finance to assess the usefulness of these forecast methods.

Acknowledgments

The authors acknowledge the insightful comments and suggestions made by an Associate Editor and two reviewers; they led to a much-improved manuscript. Han Lin Shang would like to thank for valuable comments seminar participants at the Mailman School of Public Health, Columbia University and the School of Mathematics and Statistics, University of New South Wales. Piotr Kokoszka and Alexander Petersen were partially supported by United States National Science Foundation grants DMS-1737795 and DMS-1811888, respectively.

Appendix A. Theoretical justification of the LQDT method of Section 2.2

This section defines the class of densities for which the transformations defined in Section 2.2 are well defined, and shows that the inverse transformation produces a density.

Definition A.1. *Let $\mathcal{D}$ be the space of densities with support (r^-, r^+), $-\infty < r^- < 0 < r^+ < \infty$, such that f is continuous on $[r^-, r^+]$ for all $f \in \mathcal{D}$, and either $f(r) > 0$ on $[r^-, r^+]$ or there exist constants $c_j, \gamma_j > 0$ and $r_j \in (r^-, r^+)$ such that*

$$c_0(r-r^-)^{\gamma_0} \le f(r) \le c_1(r-r^-)^{\gamma_1}, \quad r^- < r < r_0,$$
$$c_2(r^+-r)^{\gamma_2} \le f(r) \le c_3(r^+-r)^{\gamma_3}, \quad r_1 < r < r^+.$$

Proposition A.1. *Suppose that $f \in \mathcal{D}$. Then, Y defined by Eq. (8) is in $L^2[0,1]$. Furthermore, let $c \in (0,1)$ and let g be a continuous function on $(0,1)$. Then $Q_{g,c}(s) := \int_c^s e^{g(u)}du$ is continuous and increasing on $(0,1)$, it has a well-defined inverse $F_{g,c} = Q_{g,c}^{-1}$, and the inverse log quantile density transformation given by*

$$f_{g,c}(r) := \psi^{-1}(g,c)(r) \qquad \text{(A.1)}$$
$$= \begin{cases} \exp\left\{-g \circ F_{g,c}(r)\right\}, \\ \quad -\infty \le Q_{g,c}(0) < r < Q_{g,c}(1) \le \infty, \\ 0, \quad \textit{otherwise}, \end{cases}$$

is well-defined and satisfies $F'_{g,c}(r) = f_{g,c}(r) \ge 0$, $F_{g,c}(0) = c$, and $\int_{-\infty}^{\infty} f_{g,c}(r)dr = 1$.

Proof. Clearly, $Y \in L^2[0,1]$ if f is strictly positive in its support $[r^-, r^+]$. Otherwise, for small $\epsilon > 0$,

$$\int_0^{\epsilon} Y^2(s)ds = \int_{r^-}^{Q(\epsilon)} \log^2\{f(r)\}f(r)dr \quad (r = Q(s))$$
$$\le c_1 \int_{r^-}^{Q(\epsilon)} \left[\log(c_0) + \gamma_0 \log(r-r^-)\right]^2 \times (r-r^-)^{\gamma_1}dr \quad \left(w = -\log(r-r^-)\right)$$
$$= c_1 \int_{-\log(Q(\epsilon)-r^-)}^{\infty} [\log(c_0) - \gamma_0 w]^2 \times e^{-(\gamma_1+1)w}dw < \infty.$$

A similar argument shows that $\int_{1-\epsilon}^{1} Y^2(s)ds < \infty$.

Next, take g, c, $Q_{g,c}$, $F_{g,c}$, and $f_{g,c}$ as in the statement of the proposition, so that $f_{g,c} \ge 0$ is clear. As $Q_{g,c}(c) = 0$, we also have $F_{g,c}(0) = c$. The limits

$$Q_{g,c}(0) := \lim_{s\to 0} Q_{g,c}(s), \quad Q_{g,c}(1) := \lim_{s \to 1} Q_{g,c}(s)$$

are well-defined, since $Q_{g,c}$ is increasing, where we may have $Q_{g,c}(0) = -\infty$, $Q_{g,c}(1) = \infty$, or both. Taking the

Table 5
MCS procedure using the $T_{\max}$ test applied to the error criteria in the validation set.

Data	Kernel	Density forecast method	KLD	JSD Simple	 Geometric	norm L_1	 L_2	 L_∞
DJI	Gaussian	Horta-Ziegelman						
		LQDT		†		†	†	
		CoDa (standardization)						
		CoDa (no standardization)	†		†	†		†
		Skewed-*t*						
	Epanechnikov	Horta-Ziegelman						
		LQDT		†		†	†	†
		CoDa (standardization)						
		CoDa (no standardization)	†		†			
		Skewed-*t*						
S&P 500	Gaussian	Horta-Ziegelman						
		LQDT						
		CoDa (standardization)	†	†	†	†	†	†
		CoDa (no standardization)						
		Skewed-*t*						
	Epanechnikov	Horta-Ziegelman					†	
		LQDT		†				
		CoDa (standardization)		†		†	†	†
		CoDa (no standardization)	†	†	†	†	†	†
		Skewed-*t*						
Bovespa	Gaussian	Horta-Ziegelman					†	†
		LQDT		†		†		†
		CoDa (standardization)	†		†			†
		CoDa (no standardization)						†
		Skewed-*t*						
	Epanechnikov	Horta-Ziegelman					†	
		LQDT		†	†	†		†
		CoDa (standardization)						
		CoDa (no standardization)	†					
		Skewed-*t*						
XLK	Gaussian	Horta-Ziegelman					†	
		LQDT		†		†		†
		CoDa (standardization)						
		CoDa (no standardization)						
		Skewed-*t*	†		†			
	Epanechnikov	Horta-Ziegelman						
		LQDT	†	†	†	†	†	†
		CoDa (standardization)						
		CoDa (no standardization)						
		Skewed-*t*						

Note: The model(s) that are selected to belong to the superior set of models in each case are indicated by †.

change of variables $s = F_{g,c}(u)$, we have $u = Q_{g,c}(s)$ and $du = e^{g(s)}ds$, so that for $Q_{g,c}(0) \le r \le Q_{g,c}(1)$,

$$\int_{-\infty}^{r} f_{g,c}(u)du = \int_{Q_{g,c}(0)}^{r} \exp\left\{-g \circ F_{g,c}(u)\right\} du$$
$$= \int_{0}^{F_{g,c}(r)} e^{-g(s)}e^{g(s)}ds = F_{g,c}(r),$$

i.e., $F'_{g,c}(r) = f_{g,c}(r)$. Plugging in $r = Q_{g,c}(1)$ proves that $f_{g,c}$ is indeed a density with support $(Q_{g,c}(0), Q_{g,c}(1))$.

Lastly, we verify that this is truly the inverse. For a density $f \in \mathcal{D}$ with cdf F, quantile function Q, and $q = Q'$, set $(g, c) = \psi(f)$, so $Q(c) = 0$ and $g(s) = -\log(f \circ Q(s))$ is continuous on (0, 1). Then, as $q(\cdot) = 1/(f \circ Q(\cdot))$ by Eq. (6),

$$Q_{g,c}(s) = \int_{c}^{s} e^{g(u)}du = \int_{c}^{s} \exp\{-\log(f \circ Q(u))\}\, du$$
$$= \int_{c}^{s} q(u)du = Q(u).$$

As a result, $F_{g,c} = F$ and, for $r \in (r^-, r^+) = (Q(0), Q(1))$,

$$f_{g,c}(r) = \exp\left\{-g \circ F_{g,c}(r)\right\}$$
$$= \exp\{\log(f \circ Q \circ F(r))\} = f(r),$$

i.e. $\psi^{-1} \circ \psi(f) = f$.

In the other direction, if g is continuous on (0, 1) and $c \in (0, 1)$, set $f_{g,c} = \psi^{-1}(g, c)$ to be the inverse-mapped density with cdf $F_{g,c}$ and quantile function $Q_{g,c} = F^{-1}_{g,c}$. The log quantile density transformation of $f_{g,c}$ is $(Y, F_{g,c}(0)) =$

$\psi(f_{g,c})$, where

$$\begin{aligned} Y(s) &= -\log(f_{g,c} \circ Q_{g,c}(s)) \\ &= -\log\left(\exp\left\{-g \circ F_{g,c} \circ Q_{g,c}(s)\right\}\right) = g(s) \end{aligned}$$

and $F_{g,c}(0) = c$, i.e. $\psi \circ \psi^{-1}(g, c) = (g, c)$. ■

Appendix B. Log hazard transformation

This section shows that the log hazard transformation in Eq. (B.1) maps densities into the space L^2 under weak assumptions, but the inverse transformation does not necessarily transform the predicted L^2 curves into densities. Define

$$Y_t(r) = \psi(f_t)(r) = \begin{cases} \log\left\{\dfrac{f_t(r)}{1 - F_t(r)}\right\}, & \text{if } f_t(r) > 0, \\ 0, & \text{if } f_t(r) = 0. \end{cases} \tag{B.1}$$

The following conditions are sufficient to ensure $Y_t \in L^2(-\infty, \infty)$, and are not too strong. Recall that we assume that f_t is continuous and that Eq. (7) holds. Suppose that

$$\begin{aligned} f_t(r) \geq c_0(r - r_t^-)^{\gamma_0}, &\quad r_t^- < r < r_0, \\ c_1(r_t^+ - r)^{\gamma_1} \leq f_t(r) \leq c_2(r_t^+ - r)^{\gamma_2}, &\quad r_1 < r < r_t^+, \end{aligned}$$

for some $c_j, \gamma_j > 0$ and $r_j \in (r_t^-, r_t^+)$. Then, for small $r - r_t^-$,

$$\begin{aligned} |Y_t(r)| &= \log\left(\frac{1 - F_t(r)}{f_t(r)}\right) \leq -\log(f_t(r)) \\ &\leq \log(c_0^{-1}) - \gamma_0 \log(r - r_t^-). \end{aligned}$$

Since $\int_0^\epsilon \log^2(s)ds < \infty$, for small ϵ,

$$\begin{aligned} \int_{r_t^-}^{r_t^- + \epsilon} Y_t^2(s)ds &\leq \int_{r_t^-}^{r_t^- + \epsilon} \left[\log(c_0) + \gamma_0 \log(s - r_t^-)\right]^2 ds \\ &< \infty. \end{aligned}$$

Similarly, for $r_t^+ - r$ small, $1 - F_t(r) \geq c_1(\gamma_1 + 1)^{-1}(r_t^+ - r)^{\gamma_1 + 1} = c_1'(r_t^+ - r)^{\gamma_1'}$, so that

$$\begin{aligned} |Y_t(r)| &= \log\left(\frac{f_t(r)}{1 - F_t(r)}\right) \\ &\leq \log(c_2/c_1') + (\gamma_2 - \gamma_1')\log(r_t^+ - r), \end{aligned}$$

and $\int_{r_t^+ - \epsilon}^{r_t^+} Y_t^2(s)ds < \infty$. Thus, $Y_t \in L^2(-\infty, \infty)$.

Suppose that g is a continuous function with supp(g) $= [a, b]$. In our prediction context, $g = \widehat{Y}_{n+h}$. Set

$$\Lambda_g(x) = \int_a^x e^{g(y)}dy, \quad x \in (a, b)$$

and

$$\psi^{-1}(g)(x) = \exp\left\{g(x) - \Lambda_g(x)\right\},$$

and denote $f(x) = \psi^{-1}(g)(x)$. Then, $f \geq 0$ and

$$\int_a^b f(x)dx = \int_0^{\Lambda_g(b)} e^{-w}dw = 1 - e^{-\Lambda_g(b)}.$$

Thus, for f to be a density, we would need $\Lambda_g(b) = \infty$, which could be understood as a limit as $x \uparrow b$; i.e., we would need

$$\lim_{x \uparrow b} \int_a^x e^{g(y)}dy = \infty. \tag{B.2}$$

However, Eq. (B.2) is not necessarily true in general; for example, if $\lim_{x \uparrow b} g(y) = g(b) < \infty$. Eq. (B.2) holds if $g(y) = \log\left\{\frac{h(y)}{1 - H(y)}\right\}$ for some density h.

An additional issue is that Y_t necessarily diverges at r_t^- and r_t^+, which are random and differ for each t, making Y_t an ill-behaved stochastic process.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. London: Chapman & Hall.

Aue, A., Norinho, D. D., & Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, *110*, 378–392.

Azzalini, A., & Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew *t*-distribution. *Journal of the Royal Statistical Society. Series B.*, *65*(2), 367–389.

Bathia, N., Yao, Q., & Ziegelmann, F. (2010). Identifying the finite dimensionality of curve time series. *The Annals of Statistics*, *38*, 3353–3386.

Boucher, M.-P. B., Canudas-Romo, V., Oeppen, J., & Vaupel, J. W. (2017). Coherent forecasts of mortality with compositional data analysis. *Demographic Research*, *37*, 527–566.

Canale, A., & Vantini, S. (2016). Constrained functional time series: Applications to the Italian gas market. *International Journal of Forecasting*, *32*, 1340–1351.

Chen, Z., Bao, Y., Li, H., & Spencer Jr, B. F. (2019). LQD-RKHS-Based distribution-to-distribution regression methodology for restoring the probability distributions of missing SHM data. *Mechanical Systems and Signal Processing*, *121*, 655–674.

Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, *1*, 223–236.

Crnkovic, C., & Drachman, J. (1997). Quality control. In S. Grayling (Ed.), *VAR: Understanding and applying value-at-risk*. Risk Books.

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

Egozcue, J. J., Diaz-Barrero, J. L., & Pawlowsky-Glahn, V. (2006). Hilbert Space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica*, *22*, 1175–1182.

Fernández, C., & Steel, M. F. J. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association: Theory and Methods*, *93*(441), 359–371.

Fuglede, B., & Topsøe, F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In *Proceedings of International Symposium on Information Theory*.

Hansen, P. R., Lunde, A., & Nason, J. M. (2011). The model confidence set. *Econometrica*, *79*(2), 453–497.

Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the cross-section of expected returns. *Review of Financial Studies*, *29*, 5–68.

Hörmann, S., Kidziński, L., & Hallin, M. (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society. Series B.*, *77*, 319–348.

Horta, E., & Ziegelmann, F. (2018). Dynamics of financial returns densities: a functional approach applied to the bovespa intraday index. *International Journal of Forecasting*, *34*, 75–88.

Horváth, L., & Kokoszka, P. (2012). *Inference for functional data with applications*. New York: Springer.

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.

Hurvich, C. M., & Tsai, C.-L. (1993). A corrected akaike information criterion for vector autoregressive model selection. *Journal of Time Series Analysis*, *14*(3), 271–279.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *27*(3).

Hyndman, R. J., & Ullah, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Computational Statistics & Data Analysis*, *51*(10), 4942–4956.

Jones, M. C. (1992). Estimating densities, quantiles, quantile densities and density quantiles. *Annals of the Institute of Statistical Mathematics*, *44*, 721–727.

Jones, M. C., & Faddy, M. J. (2003). A skew extension of the *t*-distribution, with applications. *Journal of the Royal Statistical Society. Series B.*, *65*(1), 159–174.

Jones, M. C., & Rice, J. A. (1992). Displaying the important features of a large collection of similar curves. *The American Statistician*, *46*, 140–145.

Kneip, A., & Utikal, K. (2001). Inference for density families using functional principapl component analysis. *Journal of the American Statistical Association*, *96*(454), 519–542.

Kokoszka, P., & Reimherr, M. (2017). Introduction to functional data analysis. Boca Raton: CRC Press.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lee, T.-H., Xi, Z., & Zhang, R. (2014). Density and risk forecast of financial returns using decomposition and maximum entropy. In *Working paper*. University of California at Riverside.

Lillo, F., & Mantegna, R. N. (2000). Statistical properties of statistical ensembles of stock returns. *International Journal of Theoretical and Applied Finance*, *3*(3), 405–408.

Nerini, D., & Ghattas, B. (2007). Classifying densities using functional regression trees: Applications in oceanology. *Computational Statistics & Data Analysis*, *51*(10), 4984–4993.

Parzen, E. (1979). Nonparametric statistical modeling. *Journal of the American Statistical Association*, *74*, 105–121.

Pawlowsky-Glahn, V., Egozcue, J., & Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester: Wiley.

Petersen, A., Chen, C.-J., & Müller, H.-G. (2019). Quantifying and visualizing intraregional connectivity in resting-state functional magnetic resonance imaging with correlation densities. *Brain Connectivity*, *9*(1), 37–47.

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

Renault, T. (2017). Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking & Finance*, *84*, 25–40.

Ross, S. A. (2017). The recovery theorem. *Journal of Finance*, *70*, 615–648.

Scealy, J. L., de Caritat, P., Grunsky, E. C., Tsagris, M. T., & Welsh, A. H. (2015). Robust principal component analysis for power transformed compositional data. *Journal of the American Statistical Association*, *110*(509), 136–148.

Scealy, J. L., & Welsch, A. H. (2014). Colours and cocktails: Compositional data analysis 2013 lancaster lecture. *Australian and New Zealand Journal of Statistics*, *56*, 145–169.

Shang, H. L., & Haberman, S. (2018). Model confidence sets and forecast combination: An application to age-specific mortality. *Genus: Journal of Population Sciences*, *74*, 19.

Shang, H. L., & Hyndman, R. J. (2011). Nonparametric time series forecasting with dynamic updating. *Mathematics and Computers in Simulation*, *81*(7), 1310–1324.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Labs Technical Journal*, *27*(3), 379–423.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B.*, *53*, 683–690.

Srivastava, A., Klassen, E., Joshi, S. H., & Jermyn, I. H. (2011). Shape analysis of elastic curves in Euclidean spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(7), 1415–1428.

Tukey, J. W. (1965). Which part of the sample contains the most information? *Proceedings of the National Academy of Sciences of the United States of America*, *53*, 127–134.

van der Linde, A. (2008). Variational Bayesian functional PCA. *Computational Statistics & Data Analysis*, *53*(2), 517–533.

Wang, J. (2012). A state space model approach to functional time series and time series driven by differential equations. Rutgers University.

Zivot, E., & Wang, J. (2006). Rolling analysis of time series. In *Modeling financial time series with S-PLUS*. New York: Springer.

Piotr Kokoszka is a Professor of Statistics at Colorado State University. He has published over 130 papers on various aspects of statistics and applied probability. He published two monographs of functional data analysis. He is a fellow of the Institute of Mathematical Statistics. He serves on editorial boards of several journals, including Journal of Multivariate Analysis, Journal of Time Series Analysis and Scandinavian Journal of Statistics.

Hong Miao is an Associate Professor of Finance at Colorado State University. He published over 30 articles, mostly in empirical finance. He is a U.S. Bank Research Fellow.

Alexander Petersen is an Assistant Professor of Statistics and Applied Probability at the University of California Santa Barbara. He has published several influential papers in the Annals of Statistics, focusing on functional data analysis on manifolds.

Hanlin Shang is an Associate Professor of Statistics at Australian National University. He published over 30 papers, focusing functional data analysis with applications to demographic research. He serves on the edito- rial boards of the Journal of Computational and Graphical Statistics and the Australian and New Zealand Journal of Statistics.