Modeling and forecasting daily electricity loads via functional clustering and curve linear regression

Haeran Cho* Yannig Goude† Xavier Brossat† Qiwei Yao‡,§

December 13, 2013

Abstract

We propose a methodology for modeling and forecasting daily electricity load. Two main ingredients of our approach are (i) clustering pairs of successive daily load curves into homogeneous sub-groups, and (ii) modeling the dependence between the successive curves within each of such sub-groups via curve linear regression. For the former task, we adopt the k-centers functional clustering (k-CFC) from Chiou and Li (2007) which simultaneously accounts for the dissimilarities between clusters in terms of both the mean and the covariance functions. Besides, we propose some model selection criteria which are applicable to the functional data together with the k-CFC, to identify the number of clusters as well as the optimal clustering of the data. For the latter part of the methodology, the curve linear regression technique from Cho et al. (2013a) plays a significant part, which reduces the curve regression problem to a finite number of scalar linear regression problems via singular value decomposition in a Hilbert space. The combined methodology is applied to a range of simulated datasets as well as to the French electricity load data collected between 1996 and 2009, where various model selection methods are investigated and the SVD-based curve linear regression methods are compared to other competitors.

1 Introduction

Electricity load forecast is an essential entry in optimizing the power system scheduling as electricity cannot be stored or discharged without incurring extra costs. Therefore it is of great importance for electricity providers to model and forecast electricity loads

*School of Mathematics, University of Bristol, UK.
†Électricité de France, France.
‡Department of Statistics, London School of Economics, UK.
§Guanghua School of Management, Peking University, China

accurately over short-term (from one hour to one month ahead) and long-term (from one month to five years ahead) horizons, and various models have been proposed for the purpose.

The French energy company Électricité de France (EDF) manages a large panel of production units across Europe, which includes water dams, nuclear plants, wind turbines, coal and gas plants. Based on the vast knowledge on French electricity consumption patterns accumulated over 20 years, EDF has developed a forecasting model which consists of complex regression models based on past loads, temperature and calendar events, coupled with classical time series models such as the seasonal ARIMA (SARIMA) (Bruhns et al. 2005). This operational model performs very well, attaining about 1% mean absolute percentage error in forecasting the electricity consumption in France over one day horizon. Due to its complexity, however, the model may have a limited capacity in adapting to constantly changing electricity consumption environments, which may be attributed to the opening of new markets, technological innovations, social and economic changes, to name a few.

Cho et al. (2013a) and Cho et al. (2013b) recognized the strategic importance of developing an adaptive forecasting model and proposed to model the dependence across consecutive daily loads via curve linear regression. By regarding each daily load as a curve, the proposed model takes advantage of the continuity of the load curves in statistical modeling, while embedding some nonstationary features, such as daily patterns (see Figure 1), into a stationary framework in a functional space. The key ingredient of the proposed curve linear regression technique is the dimension reduction based on the singular value decomposition (SVD) in a Hilbert space, which effectively reduces the problem to several ordinary (i.e. scalar) linear regression problems. Compared to the EDF operational model, this approach does not incorporate much of the data-specific knowledge, while maintaining competitive prediction accuracy when applied to the French electricity consumption data.

In fitting the curve linear regression model, it is implicitly assumed that the dependence structure between the regressor (e.g. electricity load of the current day) and the response (electricity load of the next day) curves, such as their profiles and covariance function, does not change across observations. However, this assumption does not appear reasonable in electricity load modeling due to meteorological and economic factors, as demonstrated below with the electricity load data observed between 1996 and 2009 in France.

Firstly, as shown in Figure 1, there exist systematic discrepancies in the profiles and the variability of daily load curves observed on different days of a week and in different

months. While successive daily loads on Mondays–Tuesdays in June behave similarly, they are distinctively different from those observed on Saturdays–Sundays in June, and also from those observed on Mondays–Tuesdays in January. Those profile discrepancies are reflected predominantly in the locations and magnitudes of daily peaks. Typically in France, daily peaks occur at noon in summer and in the evening in winter due to economic cycle as well as the usage of electrical heating and lighting. Hence, the profiles of successive daily curves vary over different days within a week, and also over different months within a year.

In addition, to gain an insight into the possible seasonal variation present in the covariance between successive daily loads (denoted by $X_i(\cdot)$ and $X_{i+1}(\cdot)$), we examine the projections of daily load curves onto the first left singular function, which is obtained from the SVD of the sample covariance operator ($\text{cov}\{X_i(\cdot), X_{i+1}(\cdot)\}$) from the pooled dataset, see Figure 2. Note that each $X_i(\cdot)$ has been de-meaned with the mean curve obtained by averaging out all the daily curves observed on the same day of a week. If the profiles of the pairs of curves as well as their covariance undergo seasonal changes, we expect such seasonality to be reflected in the above projections over the span of one year. Indeed, the boxplots in Figure 2 indicate that in cold climate, the relationship between two consecutive daily loads is more volatile than that in warmer climates, and that the covariance structure is time-varying over a year's period.

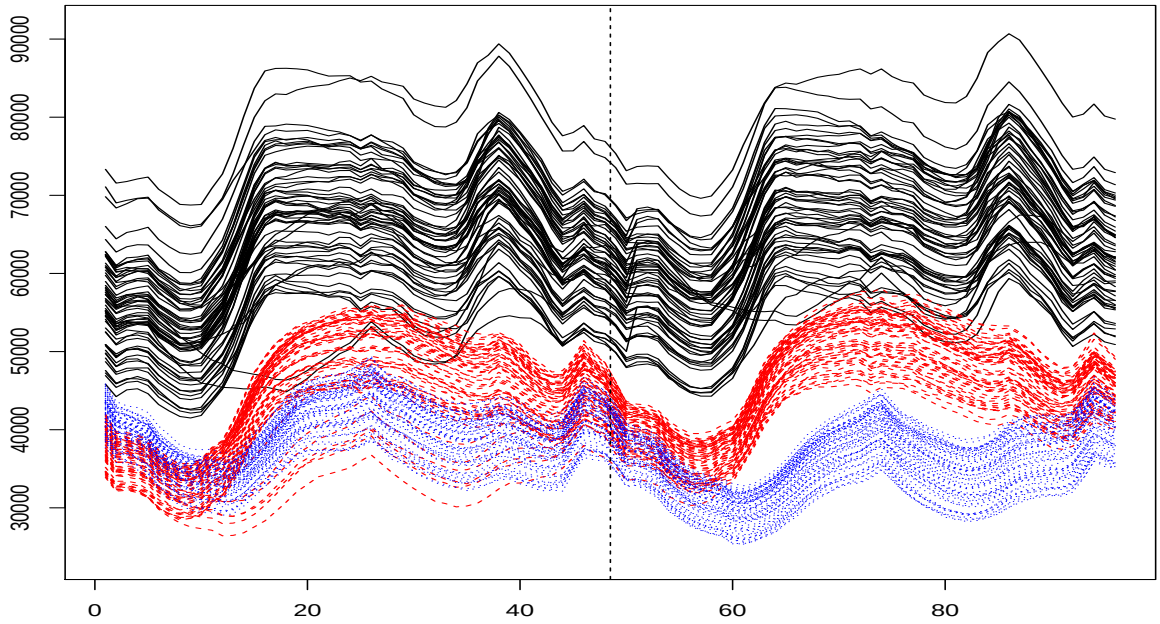

Figure 1: French electricity loads (in MW) on Mondays–Tuesdays in January (solid), Mondays–Tuesdays in June (broken) and Saturdays–Sundays in June (dotted) between 1996 and 2009.

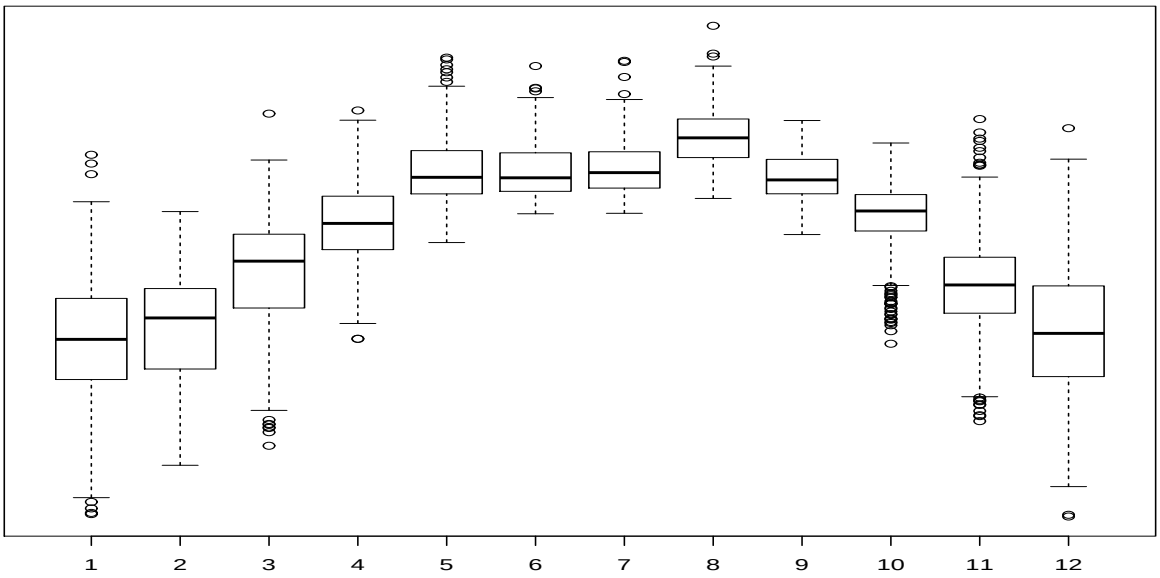

Figure 2: Boxplots of the successive daily load projections with respect to different months.

From the above observations, it is evident that prior to modeling the electricity load data via curve linear regression, an essential step is to partition the pairs of regressor and response curves into sub-groups with homogeneous dependence structure. The importance of clustering daily electricity loads was emphasized in Cho et al. (2013b), where it was shown that when the daily loads were well-partitioned, the proposed curve linear regression method could directly be applied to the data without any trend or seasonality modeling (as in Cho et al. (2013a) with the two-stage hybrid method), and still achieve even superior forecasting performance. Also, it was further argued that a data-driven clustering method might be necessary to handle the ever-changing nature of the electricity load data.

Hence in this paper, we address the problem of functional data clustering with the view of modeling and forecasting in the following step using a curve linear regression model. We note that this combined approach may be applicable to a wide range of functional data in general besides the application to electricity load modeling.

The dynamic nature of real-life data has been noted in the functional data analysis literature, and several clustering methods have been proposed and applied to various datasets. For example in Antoniadis et al. (2011), wavelet-based dissimilarity measures were introduced to detect patterns and clusters in functional data, and applied to cluster daily electricity loads into sub-groups of homogeneous profiles. Chiou (2012) adopted a subspace projected functional clustering method termed k-centers functional clustering (k-CFC) from Chiou and Li (2007), to identify distinct daily traffic flow patterns and predict future traffic flow based on the observed flow in transportation management.

We refer to Jacques and Preda (2013) for an overview of functional data clustering.
In particular, the k-CFC takes into account both the mean function and the modes of variation differentials between any two clusters, and thus is distinguished from many other clustering methods where the clustering criteria account for centrality features only, usually the mean function. Regarding the problem of clustering the pairs of regressor and response curves, say $\{X_i(\cdot), Y_i(\cdot)\}$, as that of clustering the *joined* curves $Z_i(\cdot) \equiv (X_i, Y_i)(\cdot)$, we show that the k-CFC is applicable to the problem of partitioning the pairs of observations into sub-groups of homogeneous linear dependence structure between $X_i(\cdot)$ and $Y_i(\cdot)$.
Many clustering methods, including the k-CFC, require a pre-determined number of clusters as an input, which is closely linked to selecting the "optimal" clustering of the data. Numerous suggestions have been made in the literature, such as objective function-based approaches (Caliński and Harabasz (1974), Hartigan (1975), Rousseeuw (1987), Krzanowski and Lai (1988), Tibshirani et al. (2001), James and Sugar (2003)), and pairwise testing-based approaches (Li and Chiou (2011)). However, difficulties arise as different model selection methods lead to different optimal clustering of the same data. We discuss a range of cluster number selection criteria, some of which are extensions of existing methods for multivariate data clustering to the functional data framework, and present a comparative study on various simulated datasets which is a separate contribution of this work.
The rest of the paper is organized as follows. In Section 2, we provide an overview of the curve linear regression technique introduced in Cho et al. (2013a). Then in Section 3, we describe the k-centers functional clustering algorithm and discuss its applicability to our framework, as well as presenting various model selection methods for choosing the final clustering. A comprehensive simulation study is conducted in Section 4, where the combined methodology of functional clustering and curve linear regression, is assessed alongside other competitors. In Section 5, we study the performance of the methodology on the French electricity load data in its modified form, and Section 6 closes the paper with some concluding remarks.

2 Curve linear regression via dimension reduction

As noted in Introduction, every day at noon, EDF forecasts the half-hourly consumption of electricity for the next 24 hours. We can regard such half-hourly loads for the next 24 hours from the noon on the i-th day as a curve response ($\equiv Y_i(\cdot)$), and the loads for the 24 hours up to the noon on the i-th day as a curve regressor ($\equiv X_i(\cdot)$). Then

the following curve linear regression model can be adopted to model the dependence between the successive daily loads:

$$Y_i(u) = \mu_Y(u) + \int_{\mathcal{I}_2} \{X_i(v) - \mu_X(v)\}\beta(u,v)dv + \varepsilon_i(u) \quad \text{for} \quad u \in \mathcal{I}_1, \tag{1}$$

where $\mu_Y(u) = \mathbb{E}\{Y_i(u)\}$, $\mu_X(v) = \mathbb{E}\{X_i(v)\}$ and $\mathcal{I}_1$ and $\mathcal{I}_2$ denote the supports of $Y_i(\cdot)$ and $X_i(\cdot)$, respectively. The linear operator β is a regression coefficient function defined on $\mathcal{I}_1 \times \mathcal{I}_2$, and $\varepsilon_i(\cdot)$ is noise and assumed to satisfy $\mathbb{E}\{\varepsilon_i(u)\} = 0$ for all $u \in \mathcal{I}_1$.

Functional linear regression models have been discussed in various settings as listed in Chapter 12 of Ramsay and Silverman (2005), including the case where both the response and the regressor variables are functions as in (1). Dimension reduction based on Karhunen-Loève decomposition, or functional principal component (FPC) analysis, has played a key role in the functional data literature. The conventional approach to the problem in (1) based on the dimension reduction, is to expand $Y_i(\cdot)$ and $X_i(\cdot)$ via Karhunen-Loève decomposition, then to fit simple linear regression models between the terms from such expansions. We note that this approach is equivalent to the dimension reduction based on principal component analysis in multivariate analysis, and refer to it by "FPC" in the subsequent sections. For further references on functional linear models, see e.g. Ramsay and Dalzell (1991), Chiou et al. (2004), Yao et al. (2005) and Hall and Horowitz (2007).

Similarly, an empirical Bayes approach proposed in Zhou et al. (2011) also builds a predictive model between the principal component scores of $Y_i(\cdot)$ and $X_i(\cdot)$, which is based on the expectation of the posterior distribution when normality is imposed on the prior distribution of the scores. We refer to this method by "Bayes" for further discussion in Section 4.

Since the principal components do not necessarily represent the directions in which $X_i(\cdot)$ and $Y_i(\cdot)$ are most correlated, Cho et al. (2013a) introduced an alternative where the singular value decomposition (SVD) in a Hilbert space was adopted to single out the directions upon which the projections of $Y_i(\cdot)$ were most correlated with $X_i(\cdot)$. While this approach is closely related to the canonical correlation analysis, its focus is on regressing $Y_i(\cdot)$ on $X_i(\cdot)$ and therefore does not treat $Y_i(\cdot)$ and $X_i(\cdot)$ on an equal footing, which is different from, and much simpler than, the latter. In what follows, we lay out the details of the SVD-based curve linear regression method in a generic setting. Let $\{Y_i(\cdot), X_i(\cdot)\},\ i = 1, \ldots, n$, be a random sample where $Y_i(\cdot) \in \mathcal{L}_2(\mathcal{I}_1)$, $X_i(\cdot) \in \mathcal{L}_2(\mathcal{I}_2)$, and let $\mathcal{I}_1$ and $\mathcal{I}_2$ be two compact subsets of $\mathbb{R}$ (which are allowed to be different). We denote by $\mathcal{L}_2(\mathcal{I})$ the Hilbert space consisting of all the square integrable curves

defined on the set $\mathcal{I}$, which is equipped with the inner product $\langle f, g\rangle = \int_{\mathcal{I}} f(u)g(u)du$ for any $f, g \in \mathcal{L}_2(\mathcal{I})$. For now, it is assumed that $\mathbb{E}\{Y_i(u)\} = 0$ for all $u \in \mathcal{I}_1$ and $\mathbb{E}\{X_i(v)\} = 0$ for all $v \in \mathcal{I}_2$. The covariance operator between $Y_i(\cdot)$ and $X_i(\cdot)$ is denoted by $\Sigma(u, v) = \text{cov}\{Y_i(u),\, X_i(v)\}$. Under the assumption

$$\int_{\mathcal{I}_1} \mathbb{E}\{Y_i(u)^2\}du + \int_{\mathcal{I}_2} \mathbb{E}\{X_i(v)^2\}dv < \infty, \tag{2}$$

Σ defines the following two bounded operators between $\mathcal{L}_2(\mathcal{I}_1)$ and $\mathcal{L}_2(\mathcal{I}_2)$,

$$f_1(u) \to \int_{\mathcal{I}_1} \Sigma(u, v)f_1(u)du \in \mathcal{L}_2(\mathcal{I}_2) \;\text{ and }\; f_2(v) \to \int_{\mathcal{I}_2} \Sigma(u, v)f_2(v)dv \in \mathcal{L}_2(\mathcal{I}_1)$$

for any $f_l(\cdot) \in \mathcal{L}_2(\mathcal{I}_l),\ l = 1, 2$.
Performing the SVD on Σ, there exists a triple sequence $\{\lambda_j, \varphi_j(\cdot), \psi_j(\cdot)\},\ j = 1, 2, \ldots$ which satisfies

$$\Sigma(u, v) = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\, \varphi_j(u)\, \psi_j(v), \tag{3}$$

where $\{\varphi_j(\cdot)\}$ is an orthonormal basis of $\mathcal{L}_2(\mathcal{I}_1)$, $\{\psi_j(\cdot)\}$ is that of $\mathcal{L}_2(\mathcal{I}_2)$, and the squared singular values $\{\lambda_j\}$ are ordered in a decreasing manner as $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. See Smithies (1937) for further discussion on the SVD in a Hilbert space.
Since $\{\varphi_j(\cdot)\}$ and $\{\psi_j(\cdot)\}$ are the orthonormal basis of $\mathcal{L}_2(\mathcal{I}_1)$ and $\mathcal{L}_2(\mathcal{I}_2)$, we may write

$$Y_i(u) = \sum_{j=1}^{\infty} \xi_{ij}\varphi_j(u), \qquad X_i(v) = \sum_{k=1}^{\infty} \eta_{ik}\psi_k(v), \tag{4}$$

where ξ_{ij} and η_{ik} are random variables defined as $\xi_{ij} = \langle Y_i, \varphi_j\rangle$ and $\eta_{ik} = \langle X_i, \psi_k\rangle$. From (3), it is straightforward to derive that

$$\text{cov}(\xi_{ij},\, \eta_{ik}) = \mathbb{E}(\xi_{ij}\eta_{ik}) = \begin{cases} \sqrt{\lambda_j} & \text{for } \; j = k, \\ 0 & \text{for } \; j \neq k. \end{cases} \tag{5}$$

The dimensionality of the functional data has been defined in various contexts, e.g. see Hall and Vial (2006) and Bathia et al. (2010). A correlation dimension between the two curves $Y_i(\cdot)$ and $X_i(\cdot)$ was defined in Cho et al. (2013a) with the eigenvalues λ_j.
Definition 1. If $\lambda_r > 0$ and $\lambda_{r+1} = 0$, the correlation between $Y_i(\cdot)$ and $X_i(\cdot)$ is r-dimensional.
When the correlation between $Y_i(\cdot)$ and $X_i(\cdot)$ is r-dimensional, it follows from (5) that

$\text{cov}\{\xi_{ij},\, X_i(v)\} = 0$ for all $j > r$ and $v \in \mathcal{I}_2$, from which we can conclude that the curve linear regression model (1) has an equivalent representation by r (scalar) linear regression models. It was summarized in Theorem 1 of Cho et al. (2013a) as follows: under the assumptions

(A1) the regression coefficient operator β is in the Hilbert space $\mathcal{L}_2(\mathcal{I}_1 \times \mathcal{I}_2)$, and

(A2) $\varepsilon_i(\cdot)$ are i.i.d. with $\mathbb{E}\{\varepsilon_i(u)\} = 0$ and $\mathbb{E}\{X_i(v)\varepsilon_j(u)\} = 0$ for any $u \in \mathcal{I}_1,\ v \in \mathcal{I}_2$ and $i, j \geq 1$,

the curve regression model (1) may be represented equivalently by

$$\begin{array}{ll} \xi_{ij} = \sum_{k=1}^{\infty} \beta_{jk}\eta_{ik} + \varepsilon_{ij} & \text{for } \ j = 1, \ldots, r, \\ \xi_{ij} = \varepsilon_{ij} & \text{for } \ j = r+1, r+2, \ldots, \end{array} \tag{6}$$

where $\varepsilon_{ij} = \int_{\mathcal{I}_1} \varphi_j(u)\varepsilon_i(u)du$, and $\beta_{jk} = \int_{\mathcal{I}_1 \times \mathcal{I}_2} \varphi_j(u)\psi_k(v)\beta(u,v)dudv$.

2.1 Estimation and forecasting

Given the observed pairs of curves $\{Y_i(\cdot),\, X_i(\cdot)\},\ i = 1, \cdots, n$, let

$$\widehat{\Sigma}(u,v) = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(u) - \bar{Y}(u)\}\{X_i(v) - \bar{X}(v)\},$$

where $\bar{Y}(u) = n^{-1}\sum_i Y_i(u)$ and $\bar{X}(v) = n^{-1}\sum_i X_i(v)$. Performing the SVD on $\widehat{\Sigma}(u,v)$, we obtain the estimators for $\{\lambda_j, \varphi_j(\cdot), \psi_j(\cdot)\}$ as $\{\widehat{\lambda}_j, \widehat{\varphi}_j(\cdot), \widehat{\psi}_j(\cdot)\},\ j = 1, 2, \ldots$. The SVD of $\widehat{\Sigma}(u,v)$ can be transformed into the eigenanalysis of a non-negative definite matrix, see Section 2.2.2 of Bathia et al. (2010).

Adopting the results from Bathia et al. (2010) on the consistency of $\widehat{\lambda}_j$, Cho et al. (2013a) proposed the use of a ratio-based estimator

$$\widehat{r} = \arg\max_{1 \leq j \leq d} \widehat{\lambda}_j / \widehat{\lambda}_{j+1} \tag{7}$$

to determine the correlation dimension, where d is a pre-specified upper bound on r. In Cho et al. (2013b), it was noted that this estimator should be used with caution as different components of the SVD can have different degrees of "strength" in the sense that, there may exist some $k < r$ for which non-zero $\lambda_j \neq 0,\ j > k$ are considerably smaller than $\lambda_j,\ j \leq k$. Further discussion on this point in the framework of factor

analysis can be found in Lam and Yao (2012). Heuristically, we may estimate r as

$$\widehat{r} = \max\{1 \leq j \leq d : \widehat{\lambda}_j / \widehat{\lambda}_{j+1} > M\}, \tag{8}$$

for sufficiently chosen M to avoid neglecting such smaller non-zero eigenvalues.
Taking into account the fact that $\text{var}(\eta_{ik}) \to 0$ as $k \to \infty$ (see (2) and (4)), we may include only the first Q terms $\eta_{ik},\ k = 1, \ldots, Q$ in the r multiple linear regression models, and obtain the ordinary least squares (OLS) estimator of the finite number of linear coefficients. Note that, while the OLS estimator of β_{jk} is unbiased (assuming $\beta_{jk} = 0$ for all $k > Q$), its variance tends to increase with Q in finite sample performance. That is, if Q is selected too large, we may end up with a model which fits the data too closely but performs poorly in prediction. We use $Q = 15$ in the subsequent simulation study and real data analysis which works reasonably well.
Once the $\widehat{r}$ scalar linear regression models between $\widehat{\xi}_{ij} = \langle Y_i - \bar{Y}, \widehat{\varphi}_j \rangle$ and $\widehat{\eta}_{ik} = \langle X_i - \bar{X}, \widehat{\psi}_k \rangle$, $k = 1, \ldots, Q$, are fitted as

$$\widehat{\xi}_{ij} = \sum_{k=1}^{Q} \widehat{\beta}_{jk} \widehat{\eta}_{ik}, \quad j = 1, \ldots, \widehat{r},$$

the next step is to predict $Y(\cdot)$ for a given regressor curve $X(\cdot)$. The predictor $Y(u)$ takes the following form $\widehat{Y}(u) = \bar{Y}(u) + \sum_{j=1}^{\widehat{r}} \widehat{\xi}_j \widehat{\varphi}_j(u)$, where $\widehat{\xi}_j$ are predicted as $\widehat{\xi}_j = \sum_{k=1}^{Q} \widehat{\beta}_{jk} \widehat{\eta}_k$ with $\widehat{\eta}_k = \langle X - \bar{X}, \widehat{\psi}_k \rangle$.

3 Functional data clustering

The curve linear regression framework discussed in Section 2 assumes that the pairs of curves have an identical dependence structure as in (1) across the observations, sharing the common mean functions and covariance operator (and thus regression coefficient operator). However in practice, this assumption is often found too strong as demonstrated in Introduction. Therefore, an essential step prior to building forecasting models is to partition the pairs of curves into homogeneous sub-groups, each of which consists of $\{X_i(\cdot), Y_i(\cdot)\}$ that can be modeled as sharing an identical dependence structure. In other words, denoting such sub-groups by

$$\mathcal{C}_k = \left\{ \{X_i^{(k)}(\cdot), Y_i^{(k)}(\cdot)\},\ i = 1, \ldots, n_k \right\}\ k = 1, \ldots, K^*$$

with n_k denoting the number of curves belonging to $\mathcal{C}_k$, there exist $\mu_X^{(k)}(\cdot) \in \mathcal{L}_2(\mathcal{I}_2)$, $\mu_Y^{(k)}(\cdot) \in \mathcal{L}_2(\mathcal{I}_1)$, and bounded operators $\Sigma_X^{(k)}$, $\Sigma_Y^{(k)}$ and $\Sigma^{(k)}$ for each k, such that any $\{X_i^{(k)}(\cdot), Y_i^{(k)}(\cdot)\} \in \mathcal{C}_k$ satisfy

- $\mathbb{E}\{X_i^{(k)}(v)\} = \mu_X^{(k)}(v)$, $\mathbb{E}\{Y_i^{(k)}(u)\} = \mu_Y^{(k)}(u)$ and
- $\mathrm{cov}\{X_i^{(k)}(v), X_i^{(k)}(v')\} = \Sigma_X^{(k)}(v, v')$, $\mathrm{cov}\{Y_i^{(k)}(u), Y_i^{(k)}(u')\} = \Sigma_Y^{(k)}(u, u')$ and $\mathrm{cov}\{Y_i^{(k)}(u), X_i^{(k)}(v)\} = \Sigma^{(k)}(u, v)$

for all $u, u' \in \mathcal{I}_1$, $v, v' \in \mathcal{I}_2$. Further, there exists a regression coefficient operator $\beta^{(k)}$ for each k which satisfies

$$Y_i^{(k)}(u) = \mu_Y^{(k)}(u) + \int_{\mathcal{I}_2} \{X_i^{(k)}(v) - \mu_X^{(k)}(v)\}\beta^{(k)}(u, v)dv + \varepsilon_i^{(k)}(u) \tag{9}$$

for all $i = 1, \ldots, n_k$, where $\varepsilon_i^{(k)}(\cdot)$ is i.i.d. noise with $\mathbb{E}\{\varepsilon_i^{(k)}(u)\} = 0$ and $\mathbb{E}\{X_i^{(k)}(v)\varepsilon_{i'}^{(k)}(u)\} = 0$ for all $u \in \mathcal{I}_1$, $v \in \mathcal{I}_2$ and $i, i' = 1, \ldots, n_k$.

Let $Z_i^{(k)}(\cdot)$ be the curve joining the regressor and the response curves, i.e. $Z_i^{(k)}(\cdot) \equiv (X_i^{(k)}, Y_i^{(k)})(\cdot)$, and $Z_i^{(k)}(\cdot) \in \mathcal{L}_2(\mathcal{I})$ where $\mathcal{I} = \mathcal{I}_2 \cup \mathcal{I}_1$. Such composite curves have been employed in multivariate functional principal component analysis (PCA), to examine the simultaneous variation of more than one function. We refer to Chapter 8.5 of Ramsay and Silverman (2005) for further details, such as the definition of an inner product for a composite curve.

Under our definition of $\mathcal{C}_k$, the pairs of curves in each $\mathcal{C}_k$ form $Z_i^{(k)}(\cdot)$, $i = 1, \ldots, n_k$ which share the identical mean curve $\mu^{(k)}(\cdot) \equiv \mathbb{E}\{Z_i^{(k)}(\cdot)\} = (\mu_X^{(k)}, \mu_Y^{(k)})(\cdot)$. Also, there exists a covariance operator satisfying $\Gamma^{(k)}(t, t') \equiv \mathrm{cov}\{Z_i^{(k)}(t), Z_i^{(k)}(t')\}$, $t, t' \in \mathcal{I}$ for all $i = 1, \ldots, n_k$ since

$$\Gamma^{(k)}(t, t') = \begin{cases} \Sigma_X^{(k)}(t, t') & \text{when } \ t \in \mathcal{I}_2, t' \in \mathcal{I}_2, \\ \Sigma_Y^{(k)}(t, t') & \text{when } \ t \in \mathcal{I}_1, t' \in \mathcal{I}_1, \\ \Sigma^{(k)}(t, t') & \text{when } \ t \in \mathcal{I}_1, t' \in \mathcal{I}_2 \ \text{ or } \ t \in \mathcal{I}_2, t' \in \mathcal{I}_1. \end{cases}$$

Hence, the task of clustering the pairs of curves $\{X_i(\cdot), Y_i(\cdot)\}$, $i = 1, \cdots, n$ is equivalently accomplished by clustering the joined curves $Z_i(\cdot)$ into the sub-groups of homogeneous mean and covariance functions. The k-centers functional clustering (k-CFC) introduced in Chiou and Li (2007) is designed specifically to achieve this goal, and the details of the procedure are discussed in the following section.

3.1 k-centers functional clustering

Chiou and Li (2007) proposed the k-CFC to account for between-cluster inhomogeneities in both the mean functions and the modes of variation. The k-CFC is similar to the k-means algorithm in that it iteratively re-classifies the observations based on the L_2-distance between each observed curve and the cluster centers. However in the k-CFC, cluster centers are assigned individually for each observation as its projections onto the functional principal component (PC) spaces corresponding to different clusters, and thus both the mean and the covariance function differentials are simultaneously taken into consideration. As with many clustering techniques, the k-CFC requires the total number of clusters K^* as an input, which we assume to be known throughout this section.

Adopting the functional framework from Chiou and Li (2007), let random curve $Z_i(\cdot)$ be independently sampled from a mixture of K^* stochastic processes in $\mathcal{L}_2(\mathcal{I})$, where each sub-process $Z^{(k)}(\cdot)$ is associated with a cluster $\mathcal{C}_k$. Also let $C_i \equiv C(Z_i) \in \{1, \ldots, K^*\}$ denote the random variable representing the cluster membership of $Z_i(\cdot)$. For each sub-process $Z^{(k)}(\cdot)$, there exist $\mu^{(k)}(\cdot)$ and $\Gamma^{(k)}$ as defined previously, such that $\mathbb{E}\{Z_i(t)|C_i = k\} = \mu^{(k)}(t)$ and $\text{cov}\{Z_i(t), Z_i(t')|C_i = k\} = \Gamma^{(k)}(t, t')$. Since for all $Z_i(\cdot)$,

$$\int_{\mathcal{I}} \mathbb{E}\{Z_i(t)\}^2 dt = \int_{\mathcal{I}_1} \mathbb{E}\{Y_i(u)\}^2 du + \int_{\mathcal{I}_2} \mathbb{E}\{X_i^{(k)}(v)\}^2 dv < \infty \tag{10}$$

under (2), the following Karhunen-Loève expansion is valid for $Z^{(k)}(\cdot)$:

$$Z^{(k)}(t) = \mu^{(k)}(t) + \sum_{j=1}^{\infty} \zeta_j^{(k)} \rho_j^{(k)}(t) \quad \text{for all} \quad t \in \mathcal{I}, \tag{11}$$

where $\zeta_j^{(k)} \equiv \zeta_j^{(k)}(Z^{(k)}) = \langle Z^{(k)} - \mu^{(k)}, \rho_j^{(k)} \rangle$. Also, $\{\rho_j^{(k)}(\cdot)\}_j$ are eigenfunctions associated with $\Gamma^{(k)}$, with the corresponding eigenvalues $\{\sigma_j^{(k)}\}_j$ such that $\langle \Gamma(t, \cdot), \rho_j^{(k)} \rangle = \sigma_j^{(k)} \rho_j^{(k)}(t)$. Further, we assume that the eigenvalues are in non-increasing order.

We denote an operator projecting $Z_i(\cdot)$ onto the functional PC space associated with $\mathcal{C}_k$ by $\mathcal{P}_k(\cdot)$, i.e.,

$$\mathcal{P}_k(Z_i)(t) = \mu^{(k)}(t) + \sum_{j=1}^{\infty} \zeta_j^{(k)}(Z_i) \rho_j^{(k)}(t)$$

with $\zeta_j^{(k)}(Z_i) = \langle Z_i - \mu^{(k)}, \rho_j^{(k)} \rangle$. If $C(Z_i) = k^*$, then $\mathcal{P}_{k^*}(Z_i)$ is identical to the Karhunen-Loève expansion of $Z_i(\cdot)$ and thus the L_2-norm $\|Z_i - \mathcal{P}_{k^*}(Z_i)\|_2 = 0$, while there exists

a discrepancy between $Z_i(\cdot)$ and $\mathcal{P}_k(Z_i)$ for all other $k \neq k^*$. Motivated by this, Chiou and Li (2007) proposed to assign the K^* cluster centers to each $Z_i(\cdot)$ as its projections $\mathcal{P}_k(Z_i)(\cdot)$ associated with the K^* clusters, and to determine its cluster membership based on the L_2-distances between $Z_i(\cdot)$ and $\mathcal{P}_k(Z_i)(\cdot)$. In other words, the cluster membership $\widehat{C}(Z_i)$ is assigned as

$$\widehat{C}(Z_i) = \arg \min_{k \in \{1,\ldots,K^*\}} \|Z_i - \widehat{\mathcal{P}}_k(Z_i)\|_2, \tag{12}$$

where $\widehat{\mathcal{P}}_k(\cdot)$ is the projection operator estimated from the observations belonging to the current $\mathcal{C}_k$.

To estimate these operators, the observations must have already been clustered into $\mathcal{C}_k$, $k = 1, \ldots, K^*$. As this is unattainable in practice, the k-CFC takes an iterative updating approach similar to that of the k-means algorithm. Namely, given an initial clustering, $\mu^{(k)}(\cdot)$ and $\{\rho_j^{(k)}(\cdot)\}$ are estimated from the observations currently belonging to $\mathcal{C}_k$ for each k, and thus $\widehat{\mathcal{P}}_k(\cdot)$ is estimated. Then based on the estimated cluster centers, $\widehat{\mathcal{P}}_k(Z_i)$, each $Z_i(\cdot)$ is re-classified according to (12), and consequently the clusters and the corresponding projection operators are updated. This re-classification is repeated until a certain termination condition is met, e.g. when there are no more curves to be re-classified.

Chiou and Li (2007) suggested to initially partition the observations by applying the k-means algorithm to the first few functional PC scores obtained from the pooled dataset. In some applications, there may be some exogenous information which can readily be used to produce an initial clustering. For example, in electricity load data analysis, calendar variables can serve as criteria for initial clustering. We further discuss on this point in Section 5.

Since the expansion in (11) involves infinitely many terms, we identify a finite number d_k for each $\mathcal{C}_k$ such that only the first d_k leading eigenfunctions are employed in the projection operator. A similar problem of estimating a curve dimensionality is addressed in Section 2.1. However, we note that our focus here is to find d_k which leads to a well-performing truncated projection operator

$$\widetilde{\mathcal{P}}_k(Z)(t) = \mu^{(k)}(t) + \sum_{j=1}^{d_k} \zeta_j^{(k)}(Z)\rho_j^{(k)}(t)$$

in the sense that $\|\widehat{\mathcal{P}}_k(Z) - \mathcal{P}_k(Z)\|_2$ is small, without seeking to identify the underlying dimensionality. To this end, we adopt one of the suggestions made in Chiou and Li

(2007) and select d_k based on the cumulative percentage of total variance:

$$d_k = \min\left\{ q \geq 1 \,:\, \frac{\sum_{j=1}^{q} \sigma_j^{(k)}}{\sum_{j=1}^{\infty} \sigma_j^{(k)} \mathbb{I}(\sigma_j^{(k)} > 0)} > \tau \right\}, \tag{13}$$

where τ is a pre-specified value within $(0, 1)$. In the simulation study reported later, $\tau = 0.8$ is used which works reasonably well.

For each $\mathcal{C}_k$, we estimate $\mu^{(k)}(\cdot)$ and $\Gamma^{(k)}$ at each iteration as $\widehat{\mu}^{(k)}(t) = n_k^{-1} \sum_{i=1}^{n_k} Z_i^{(k)}(t)$ and $\widehat{\Gamma}^{(k)}(t, t') = n_k^{-1} \sum_{i=1}^{n_k} \{Z_i^{(k)}(t) - \widehat{\mu}^{(k)}(t)\}\{Z_i^{(k)}(t') - \widehat{\mu}^{(k)}(t')\}$ for all $t, t' \in \mathcal{I}$. Then performing eigenanalysis on $\widehat{\Gamma}^{(k)}$, we estimate the eigenvalues and eigenfunctions $\{\widehat{\sigma}_j^{(k)}, \widehat{\rho}_j^{(k)}(\cdot)\}_j$ as well as d_k, and consequently the cluster centers for each observed curve $Z_i(\cdot)$ is estimated as

$$\widehat{\mathcal{P}}_k(Z_i)(t) = \widehat{\mu}^{(k)}(t) + \sum_{j=1}^{d_k} \widehat{\zeta}_j^{(k)}(Z_i) \widehat{\rho}_j^{(k)}(t), \quad \text{where } \widehat{\zeta}_j^{(k)}(Z_i) = \langle Z_i - \widehat{\mu}^{(k)}, \widehat{\rho}_j^{(k)} \rangle.$$

In the re-classification step, if the current membership of $Z_i(\cdot)$ is k^*, the mean curve and the covariance operator of $\mathcal{C}_{k^*}$ are obtained as leave-one-out estimators, and these estimates are used in the corresponding cluster center $\widehat{\mathcal{P}}_{k^*}(Z_i)(\cdot)$.

In summary, the k-CFC algorithm takes the following steps to iteratively partition the curve observations into homogeneous sub-groups.

k-CFC algorithm

Step 0: Initial clustering. From the pooled data, obtain $\bar{Z}(t) = n^{-1} \sum_{i=1}^{n} Z_i(t)$ and $\widehat{\Gamma}(t, t') = n^{-1} \sum_{i=1}^{n} (Z_i(t) - \bar{Z}(t))(Z_i(t') - \bar{Z}(t'))$. Then perform eigenanalysis on $\widehat{\Gamma}$ to obtain eigenfunctions $\widehat{\rho}_j(\cdot)$ and the corresponding eigenvalues $\widehat{\sigma}_j$ for $j = 1, 2, \ldots$. For a d chosen similarly as d_k in (13) with $\widehat{\sigma}_j$ replacing $\widehat{\sigma}_j^{(k)}$, apply the k-means clustering to the first d functional PC scores $\widehat{\zeta}_j(Z_i) = \langle Z_i - \bar{Z}, \widehat{\rho}_j \rangle$ and produce initial clusters $\mathcal{C}_k,\ k = 1, \ldots, K^*$.

Step 1 Update the projection operators $\widehat{\mathcal{P}}_k$ based on the current clustering. Then for each observation $Z_i(\cdot)$, obtain its projections onto the K^* functional PC spaces as $\widehat{\mathcal{P}}_k(Z_i)(\cdot),\ k = 1, \ldots, K^*$ and re-classify the curve according to $\widehat{C}(Z_i) = \arg\min_{1 \leq k \leq K^*} \|Z_i - \widehat{\mathcal{P}}_k(Z_i)\|_2$. Based on this re-classification step, update the clusters $\mathcal{C}_k,\ k = 1, \ldots, K^*$.

Step 2 Repeat Step 1 until $\mathcal{C}_k$ no longer change.

Once equipped with $\mathcal{C}_k,\ k = 1, \ldots, K^*$ returned by the k-CFC, as a new observation

$Z(\cdot)$ is made, it is classified to one of the K^* clusters based on the membership criterion in (12), i.e. $\widehat{C}(Z) = \arg\min_{k=1,\ldots,K^*} \|Z - \widehat{\mathcal{P}}_k(Z)\|_2$.
There still remains a non-trivial question on the choice of the total number of clusters K^*. In Section 3.3, we list a number of methods from the relevant multivariate clustering literature, and propose their extensions in the context of k-CFC, some of which are motivated by the theoretical properties of k-CFC discussed in the next section.

3.2 Properties of the k-CFC

Pollard (1981) showed that in multivariate data classification, the centers of clusters returned by the k-means algorithm were consistent. The two essential tools are used in the proof: the fact that the optimal sample cluster centers are contained in a compact ball and the strong law of large numbers. However, due to the infinite-dimensional nature of the functional data, it still remains as a challenging task to extend the above arguments to the case of k-CFC algorithm. Instead, Chiou and Li (2007) provide with some conditions under which the functional PC spaces associated with any two *true* clusters $\mathcal{C}_k^*$ and $\mathcal{C}_l^*$ are distinguished by the L_2-distance measure (12).
Let $c_i = C^*(Z_i)$ indicate the true cluster membership of $Z_i(\cdot)$ in the sense that $Z_i(\cdot)$ is generated from the sub-process corresponding to the cluster $\mathcal{C}_{c_i}^*$, and let $|\mathcal{C}_k^*| = n_k^*$ for all $k = 1, \ldots, K^*$. We impose the following condition on the number of observations in each true cluster to study the asymptotic properties of the estimated cluster centers.

(C1) There exists $\delta \in (0, 1)$ such that $n^{-\delta} \cdot n_k^* \to \infty$.

Each sub-process corresponding to $\mathcal{C}_k^*$ is equipped with the mean function $\mu^{*(k)}(\cdot)$ and the pair of eigenvalues and eigenfunctions $\{\sigma_j^{*(k)}, \rho_j^{*(k)}(\cdot)\}$, $j = 1, 2, \ldots$. The projection operator $\mathcal{P}_k^*(\cdot)$ associated with $\mathcal{C}_k$ is defined accordingly with $\mu^{*(k)}(\cdot)$ and $\{\rho_j^{*(k)}(\cdot)\}_{j=1}^{\infty}$, and so does its truncated operator $\widetilde{\mathcal{P}}_k^*$ with $\mu^{*(k)}(\cdot)$ and $\{\rho_j^{*(k)}(\cdot)\}_{j=1}^{d_k}$.
The consistency of $\widehat{\rho}_j^{*(k)}(\cdot)$ estimated within each $\mathcal{C}_k^*$ can be shown by adapting the proof of Theorem 1 in Bathia et al. (2010), in the sense that

$$\sup_{t \in \mathcal{I}} |\widehat{\rho}_j^{*(k)}(t) - \rho_j^{*(k)}(t)| = O_p(n^{-\delta/2})$$

for a compact $\mathcal{I}$, provided that $\mathbb{E}\{\int_{\mathcal{I}} Z_i(t)^2 dt\}^2 < \infty$. Similarly, supposing that $\sup_{t \in \mathcal{I}} |\widehat{\mu}^{*(k)}(t) - \mu^{*(k)}(t)| = O_p(n^{-\delta/2})$ as done in Assumption 1 of Chiou and Li (2007), we can show the following for estimated projection operator $\widehat{\mathcal{P}}_k^*(Z_i)(t) = \widehat{\mu}^{*(k)}(t) + \sum_{j=1}^{d_k} \widehat{\zeta}^{*(k)}(Z_i)\widehat{\rho}_j^{*(k)}(t)$:

$$\|\widehat{\mathcal{P}}_k^*(Z_i) - \widetilde{\mathcal{P}}_k^*(Z_i)\|_2 = O_p(n^{-\delta/2}), \tag{14}$$

see Section A in Appendix for the proof.
Using the above result, Lemma 1 of Chiou and Li (2007) shows that for all $l \neq c_i$,

$$\begin{aligned} & \|\widehat{\mathcal{P}}_c^*(Z_i) - \widehat{\mathcal{P}}_l^*(Z_i)\|_2^2 = \|\widetilde{\mathcal{P}}_c^*(Z_i) - \widetilde{\mathcal{P}}_l^*(Z_i)\|_2^2 + O_p(n^{-\delta}) \\ = \;& \|\mu^{*(c)} - \mu^{*(l)}\|^2 + \sum_{j=1}^{d_c} |\zeta_j^{*(c)}(\widetilde{\mathcal{P}}_c^*(Z_i))|^2 - \sum_{j=1}^{d_l} |\zeta_j^{*(l)}(\widetilde{\mathcal{P}}_c^*(Z_i))|^2 \\ + \;& 2\sum_{j=1}^{d_c} \zeta_j^{*(c)}(\widetilde{\mathcal{P}}_c^*(Z_i))\langle \mu^{*(c)} - \mu^{*(l)}, \rho_j^{*(k)}\rangle + \sum_{j=1}^{d_l} |\langle R_c(Z_i), \rho_j^{*(l)}\rangle|^2 + O_p(n^{-\delta}) \quad (15) \end{aligned}$$

(the subscript i is dropped from c_i for notational brevity), where $R_c(Z_i) = \sum_{j=d_c+1}^{\infty} \zeta_j^{*(c)}(Z_i)\rho_j^{*(c)}$, i.e. the residual resulting from the truncation of Karhunen-Loève expansion. Since $\sigma_j^{*(c)} \to 0$ as $j \to \infty$ under (10), we may assume that $d_l \sum_{j=d_c+1}^{\infty} \sigma_j^{*(c)} \to 0$ as $d_c = d_c(n) \to \infty$, which leads to the term $\sum_{j=1}^{d_l} |\langle R_c(Z_i), \rho_j^{*(l)}\rangle|^2 \to 0$ in probability.
Defining $\mathcal{M}^{*(k)}$ as the functional space spanned by $\{\rho_j^{*(k)}(\cdot),\ j = 1, \ldots, d_k\}$ for all k, the following are referred to as *non-identifiability conditions* in Chiou and Li (2007):

(C2) $\mu^{*(k)}(t) = \mu^{*(l)}(t)$ for all $t \in \mathcal{I}$, or $\mu^{*(k)}(\cdot), \mu^{*(l)}(\cdot) \in \mathcal{M}^{*(l)}$.

(C3) $\mathcal{M}^{*(k)} \subseteq \mathcal{M}^{*(l)}$.

When both (C2) and (C3) are met with $k = c_i$ and some $l \neq c_i$, the right-hand side of (15) converges to zero as $n \to \infty$. This implies that the two clusters $\mathcal{C}_c^*$ and $\mathcal{C}_l^*$ are indistinguishable from each other in terms of the corresponding cluster centers for $Z_i(\cdot)$. On the other hand, when either (C2) or (C3) is not met, $\|\widehat{\mathcal{P}}_{c_i}^*(Z_i) - \widehat{\mathcal{P}}_l^*(Z_i)\|_2^2$ is bounded away from zero. Hence, it is easily seen that the true membership of $Z_i(\cdot)$ can be identified by the L_2-distance criterion in (12), as long as every pair of the true clusters are identifiable in the sense that either of (C2) or (C3) is *not* met for any $k \neq l$. As the true clusters are unknown, let alone the true total number of clusters K^*, the estimated mean functions and eigenfunctions have smaller convergence rates than those assumed above, and the consistency of the estimated cluster centers also suffer in the re-classification procedure. However, we may apply these findings in assessing the clusters finally returned by the k-CFC with some K as an input for the number of clusters, and thus derive a model selection criterion. Indeed some of the cluster number selection methods described in the next section actively make use of the theoretical properties discussed above.

3.3 Identifying the optimal number of clusters

Most clustering procedures including the k-CFC algorithm, require the knowledge of the total number of clusters as an input, which is often closely related to the quality of the clustering outcome yet is largely unavailable in practice. This problem of determining the number of clusters is widely regarded as one of the most difficult challenges, as remarked by Gordon (1999) (Chapter 6) and the references therein.

A commonly adopted approach is to apply a classification technique with a range of K as an input and then assess the resulting clusters to estimate the optimal number of clusters. As noted by Tibshirani et al. (2001), clustering assessment methods may be categorized into two, as global and local approaches. The former builds an object function evaluating the clusters which is optimized as a function of K. Examples of such global methods are provided in Section 3.3.1. The forward functional testing method proposed in Li and Chiou (2011) is a local method in the sense that for any pairs of clusters, it tests whether to merge them or not according to the identifiability conditions, see Section 3.3.2 for more details.

3.3.1 Global approaches

Many global approaches employ some objective functions for evaluating the relative quality of clustering, and ultimately selecting an appropriate number of clusters. They are often constructed with certain within-cluster (WC) and between-clusters (BS) dissimilarity measures, noting that when the observations are clustered "optimally", we expect the former to be small and the latter to be large.

The silhouette statistic (Rousseeuw 1987) is an example of such objective functions originally applied to multivariate data clustering. Its definition can be extended to be applicable in the functional clustering framework as follows. Given the clusters returned with an input cluster number K, for an observation $Z_i(\cdot)$ with $c_i = C(Z_i)$, let $a_i(K)$ denote the average distance between $Z_i(\cdot)$ and all the other curve members in $\mathcal{C}_{c_i}$ (WC dissimilarity). Also, denoting by l_i^o the index of the "nearest" cluster for $Z_i(\cdot)$ in the sense that $l_i^o = \arg\min_{1\le k\ne c_i\le K} \sum_{j=1}^{n_k} \|Z_i - Z_j^{(k)}\|_2^2$, let $b_i(K)$ be the average distance between $Z_i(\cdot)$ and the curve members in $\mathcal{C}_{l_i^o}$ (BC dissimilarity). Then the silhouette statistic is defined as

$$s_i^o(K) = \frac{b_i(K) - a_i(K)}{\max\{a_i(K), b_i(K)\}},$$

and the optimal number of clusters $\widehat{K}$ is chosen as where $S^o(K) = \sum_{i=1}^n s_i^o(K)$ is maximized.

$s_i^o(\cdot)$ may be modified to exploit some properties of the estimated projection operators

discussed in Section 3.2. We define l_i to denote another index of the nearest cluster for $Z_i(\cdot)$, according to the distance between $Z_i(\cdot)$ and its projections onto the $K-1$ clusters $\mathcal{C}_k$, $k \neq c_i$, i.e. $l_i = \arg\min_{1\leq k\neq c_i\leq K} \|Z_i - \widehat{\mathcal{P}}_k(Z_i)\|_2^2$. Then the modified silhouette statistic is of the form

$$s_i(K) = \frac{\|\widehat{\mathcal{P}}_{c_i}(Z_i) - \widehat{\mathcal{P}}_{l_i}(Z_i)\|_2^2}{\|Z_i - \widehat{\mathcal{P}}_{c_i}(Z_i)\|_2^2}.$$

With a slight abuse of notation, let $s_i(K^*)$ denote the silhouette statistic obtained under the true clustering, i.e., not only $K = K^*$ but also $c_i = C^*(Z_i)$ for all i. Also assume that each cluster is identifiable in the sense that either of (C2) or (C3) is not met. Then, the denominator of $s_i(K^*)$ is close to 0 while its numerator is bounded away from 0 (as a result of the arguments below (see (14) and the arguments below (15)), which implies that $S(K^*) = \sum_{i=1}^n s_i(K^*)$ satisfies $n^{-\delta} \cdot S(K^*) \to \infty$.

On the other hand, when $K < K^*$, the denominator of $s_i(K)$ is expected to be bounded away from 0 for some clusters, as the corresponding $\widehat{\mathcal{P}}_{c_i}(\cdot)$ does not well-approximate the true Karhunen-Loève expansion of $Z_i(\cdot)$. Also when $K > K^*$, it is likely that some of the true clusters are split into two, which results in $s_i(K)$ with their numerators of the order $n^{-\delta}$ for those $Z_i(\cdot)$ belonging to the split clusters. In summary, assuming that the size of each true cluster is sufficiently large, i.e., $\delta = 1$ and thus $n_k^* \asymp n$ for all $k = 1, \ldots, K^*$, we expect that $S(K)$ is maximized at $K = K^*$.

Other dissimilarity measures that are frequently adopted in the literature are WC and BC sum of squares (SS). However, some adjustments are necessary to the notion of WCSS commonly used in the multivariate clustering literature, since in the k-CFC, cluster centers are determined for each $Z_i(\cdot)$ separately, in place of the cluster means conventionally adopted in these measures.

Recall that the k-means algorithm sets out to minimize the following WCSS

$$\sum_{k=1}^{K}\sum_{i=1}^{n_k} \|Z_i^{(k)} - \widehat{\mu}^{(k)}\|_2^2,$$

where $Z_i^{(k)}(\cdot)$ denotes the i-th observation in $\mathcal{C}_k$ and $\widehat{\mu}^{(k)}(\cdot)$ denotes the cluster center of $\mathcal{C}_k$. We propose a new definition of WCSS in the framework of k-CFC as

$$W(K) = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \|Z_i^{(k)} - \widehat{\mathcal{P}}_k(Z_i^{(k)})\|_2^2 = \sum_{i=1}^{n} \|Z_i - \widehat{\mathcal{P}}_{c_i}(Z_i)\|_2^2.$$

Similarly, we use the following as the BCSS

$$B(K) = \sum_{i=1}^{n} \|\widehat{\mathcal{P}}_{c_i}(Z_i) - \widehat{\mu}(Z_i)\|_2^2,$$

where $\widehat{\mu}(Z_i)(\cdot)$ is the pooled cluster center for $Z_i(\cdot)$, which may take the forms

(i) $\widehat{\mu}(Z_i)(t) = n^{-1} \sum_{k=1}^{K} n_k \widehat{\mathcal{P}}_k(Z_i)(t)$, or

(ii) $\widehat{\mu}(Z_i)(t) = \widehat{\mathcal{P}}(Z_i)(t)$ with $\widehat{\mathcal{P}}(\cdot)$ denoting the projection operator to the functional PC space of the pooled data.

Besides, we compute the degrees of freedom of $W(K)$ in the spirit of the effective degrees of freedom (see e.g. (3.60) of Hastie et al. (2001)) as

$$df(K) = K + \sum_{k=1}^{K} (n_k - 1) \frac{\sum_{j=1}^{d_k} \widehat{\sigma}_j^{(k)}}{\sum_k \widehat{\sigma}_j^{(k)} \mathbb{I}(\widehat{\sigma}_j^{(k)} > 0)},$$

where the second term accounts for the projection onto the functional PC space in $\widehat{\mathcal{P}}_{k_i}(Z_i)(\cdot)$.
Using these definitions, we propose to extend some cluster number selection criteria suggested in the multivariate clustering literature, to functional clustering setting.

Caliński and Harabasz (1974): The variance ratio criterion

$$CH(K) = \frac{B(K)/(K-1)}{W(K)/(n - df(K))},$$

analogous to F-statistic in univariate data analysis, finds the best partitioning as where $CH(K)$ is maximized for $K \geq 2$. We denote the criteria obtained with two different choices of $\widehat{\mu}(Z_i)(\cdot)$ as in the above (i) and (ii), by $CH_1(\cdot)$ and $CH_2(\cdot)$, respectively.

Hartigan (1975): Defined using WCSS only as

$$H(K) = \frac{1}{n - K - 1} \left\{ \frac{W(K) - W(K+1)}{W(K+1)} \right\},$$

it is feasible to obtain $H(K)$ for all $K \geq 1$. Instead of deriving a test criterion from a F-distribution, Hartigan (1975) suggested to increase the number of clusters as long as $H(K) > 10$, i.e. $\widehat{K}$ is selected as the smallest K for which $H(K) \leq 10$.

Krzanowski and Lai (1988): Denoting the dimensionality of the data by p,

$$KL(K) = \left| \frac{(K-1)^{2/p} W(K-1) - K^{2/p} W(K)}{K^{2/p} W(K) - (K+1)^{2/p} W(K+1)} \right|$$

returns the number of cluster as where it is maximized for $K \geq 2$. When applying $KL(K)$ for functional clustering, p may be set as the size of the discrete grids over which the curves are observed.

Occasionally, the k-CFC returns different clustering configurations for the same given number of clusters, due to e.g. cluster merging or the different choices of initial clusters. To cope with such situations, we may regard $S^o(\cdot)$, $S(\cdot)$, $CH_1(\cdot)$ and $CH_2(\cdot)$, not as a function of K but as that of cluster configuration, and identify the optimal clustering of the data as where each function is maximized. Note that this adjustment is not applicable to $H(\cdot)$ and $KL(\cdot)$.

3.3.2 Forward functional testing

Motivated by the identifiability conditions (C2)–(C3), Li and Chiou (2011) proposed a multiple hypothesis testing scheme named forward functional testing (FFT). It evaluates the clustering by testing whether there exists any pair of clusters which are not identifiable (i.e., both (C2) and (C3) are met). If so, such two clusters are merged into one whereas, if there exists no such pair of clusters, we can sufficiently believe that the clustering is valid.
More specifically, the following null hypotheses are tested

$$\mathrm{H}_{01}(k,l) : \mu^{(k)}(\cdot) = \mu^{(l)}(\cdot) \quad \text{and} \quad \mathrm{H}_{02}(k,l) : \mathcal{M}^{(k)} = \mathcal{M}^{(l)},$$

for all pairs of $\mathcal{C}_k$ and $\mathcal{C}_l$, $k \neq l$. When the sub-processes corresponding to the two clusters are identifiable, the discrepancy is reflected in the test statistics

$$\begin{aligned} \mathcal{D}_1(k,l) &= \|\widehat{\mu}^{(k)} - \widehat{\mu}^{(l)}\|_2^2, \quad \text{and} \\ \mathcal{D}_2(k,l) &= \sum_{j=1}^{d} \omega_j^{(k)} \|\widehat{\rho}_j^{(k)} - \widehat{\mathcal{P}}_l(\widehat{\rho}_j^{(k)})\|_2^2 + \omega_j^{(l)} \|\widehat{\rho}_j^{(l)} - \widehat{\mathcal{P}}_k(\widehat{\rho}_j^{(l)})\|_2^2, \end{aligned}$$

where $d = \min(d_k, d_l)$ and $\omega_j^{(k)}$ is defined as $\omega_j^{(k)} = \widehat{\sigma}_j^{(k)} / \sum_{j'} \widehat{\sigma}_{j'}^{(k)} \mathbf{I}(\widehat{\sigma}_{j'}^{(k)} > 0)$.
By repeatedly generating bootstrap samples according to the resampling scheme given in Efron and Tibshirani (1994), we compare $\mathcal{D}_1(k,l)$ and $\mathcal{D}_2(k,l)$ against the test statistics that are similarly obtained from the bootstrap samples, and thus produce p-values.

Starting from the initial cluster number $K = K_0 \geq 2$, the FFT procedure is applied to the clusters returned by the k-CFC with K as an input number of clusters. We update K by one, as long as either $\mathrm{H}_{01}(k, l)$ or $\mathrm{H}_{02}(k, l)$ is rejected for all pairs of (k, l) according to the chosen multiple testing method. Otherwise, the procedure stops and we set the optimal number of clusters as $\widehat{K} = K - 1$. For further details of the FFT procedure, including bootstrap sample generation, see Sections 3.1–3.2 of Li and Chiou (2011).

Due to the bootstrap sampling, the FFT procedure is computationally intensive compared to the global approaches listed in Section 3.3.1. Besides, when the input K is set greater than K^*, the k-CFC often splits (approximately) homogeneous clusters into two. Then, the bootstrap samples from such split clusters resemble the data too closely, such that the FFT still rejects the null hypotheses and causes false alarms. Also, the k-CFC occasionally merges two or more clusters into one for a large input of K. However, sometimes the FFT procedure may be quitted before reaching such a large K, and does not benefit from clustering merging which may lead to inferior clustering configuration as observed in the simulation study.

Remark 1. We note that some of the model selection methods described above are not applicable to identify the true cluster number when $K^* = 1$. However, when the input cluster number K is greater than K^*, it has been empirically observed that the k-CFC merges multiple clusters into one during the re-classification steps. This feature was investigated on simulated datasets in Li and Chiou (2011), where they showed that when $K > K^*$, merging occurs with high frequencies for the input values of K closer to K^*. Since the FFT is stopped when merging occurs, the true $K^* = 1$ was identified over 93% of simulated datasets generated in the paper.

4 Simulation study

In this section, we conduct a simulation study with two aims. First, we investigate the performance of various model selection techniques described in Section 3.3 in combination with the k-CFC, whether they are able to identify the true clusters (and the total number) and predict correct cluster memberships on the test set. Then we proceed to a comparative study between the SVD-based curve linear regression technique and its competitors (recall FPC and Bayes described in Section 2), by fitting different curve linear regression models fitted within each cluster and examining their predictive performance. In this manner, we attempt to fully appreciate the joint methodology combining clustering and forecasting steps.

We adopt a range of models to simulate curve data. The models (M1)–(M4) were first employed in Li and Chiou (2011), although the size of each cluster and observation grids are set differently here. The curve observations are generated as

$$Z_i^{(k)}(t_l) = \mu^{(k)}(t_l) + \sum_{j=1}^{d_k} \zeta_{ij}^{(k)} \rho_j^{(k)}(t_l) + \varepsilon_i^{(k)}(t_l), \quad i = 1, \ldots, n_k;\ k = 1, \ldots, K^*, \qquad (16)$$

where each curve is observed on equispaced grids $t_l = (l-1)/(p-1)$ for $l = 1, \ldots, p = 60$. The random effects $\zeta_{ij}^{(k)}$ are generated independently from $\mathcal{N}(0, \sigma_j^{(k)})$. For the specifics of the choice of $\mu^{(k)}(\cdot)$, $\rho_j^{(k)}(\cdot)$ and $\sigma_j^{(k)}$ we refer to Section 4.1 of Li and Chiou (2011). Unlike in the original models, we include the measurement error $\varepsilon_i^{(k)}(t_l)$ which is generated at each t_l as an i.i.d. random variable following $\mathcal{N}(0, \sigma_{d_k}^{(k)}/200)$.
For each model, we generate both the training set (denoted by $\{Z_i^{(k)}(\cdot),\ i = 1, \ldots, n_k\},\ k = 1, \ldots, K$) and the test set ($\{\widetilde{Z}_i^{(k)}(\cdot),\ i = 1, \ldots, \widetilde{n}_k\},\ k = 1, \ldots, K$). For the training set, n_k is independently generated from $\mathcal{U}(300, 400)$ in (M1)–(M3), $\mathcal{U}(150, 250)$ in (M4) and $\mathcal{U}(100, 200)$ in (M5), while for the test set, $\widetilde{n}_k$ is independently generated from $\mathcal{U}(30, 60)$ in (M1)–(M3) and $\mathcal{U}(20, 50)$ in (M4)–(M5). Finally, when we refer to the pooled dataset, we adopt the notations $\{Z_i(\cdot),\ i = 1, \ldots, n\}$ for the training set and $\{\widetilde{Z}_i(\cdot),\ i = 1, \ldots, \widetilde{n}\}$ for the test set.

(M1) $K^* = 3$ **with a shared mean function but different eigenspaces.**

(M2) $K^* = 3$ **with a shared eigenspace but different mean functions.**

(M3) $K^* = 3$ **with different mean functions and eigenspace.**

(M4) $K^* = 6$ **with different mean functions and eigenspace:** This model is included to see whether the model selection criteria are effective for identifying the true cluster number when it is larger than that considered in (M1)–(M3).

(M5) $K^* = 4$ **with three classes generated from the electricity load data:** As observed in Introduction, we expect that pairs of daily electricity loads observed on the same consecutive days of a week in the same month, behave similarly in terms of their profiles and covariance structures. Therefore we define four classes for $Z_i^{(k)}(\cdot) = (X_i^{(k)}, Y_i^{(k)})(\cdot)$ as in Table 1, and each $Z_i^{(k)}(\cdot)$ is observed over $p = 96$ equispaced grids (every half an hour). Within each class, $Z_i^{(k)}(\cdot)$ is generated as in (16), where $\mu^{(k)}(\cdot)$, $\rho_j^{(k)}(\cdot)$, $\sigma_j^{(k)}$ and d_k are estimated using the observations from the French electricity load dataset (collected between 1996 and 2009) corresponding to each class, and the measurement error is generated at each t_l as an i.i.d. random variable following $\mathcal{N}(0, \widehat{\sigma}_{d_k}^{(k)}/200)$.

Table 1: Four classes in (M5).

k	1		2		3		4	
	day	month	day	month	day	month	day	month
$X_i^{(k)}$	Tuesday	July	Friday	June	Saturday	September	Friday	April
$Y_i^{(k)}$	Wednesday	July	Saturday	June	Sunday	September	Saturday	April

Sample datasets generated from (M1)–(M5) are plotted in Figures 3–7.

4.1 Performance of the k-CFC and model selection methods

We study the performance of the k-CFC algorithm along with the model selection techniques discussed in Section 3.3, such as the objective function-based methods ($S^o(\cdot)$, $S(\cdot)$, $CH(\cdot)$, $H(\cdot)$ and $KL(\cdot)$), and the FFT procedure. Besides the estimated number of clusters ($\widehat{K}$), we adopt the adjusted Rand index to assess the quality of cluster configurations returned by different methods. The adjusted Rand index is a measure of agreement between two clusterings, such that a value close to 1 indicates higher agreement between the two, see Hubert and Arabie (1985) for further details. We obtain the adjusted Rand index between the estimated and the true cluster memberships on the training set ($\widehat{\theta}$), and that between the predicted and the true cluster memberships on the test set ($\widetilde{\theta}$). We also report the quality of the clustering resulted from the following approaches as a reference:

- **oracle:** apply the k-CFC with K^* known,
- **initial:** apply Step 0 of the k-CFC with K^* known, and
- k**-means** : apply the k-means algorithm by regarding the curves observed over size p grids as vectors of length p, with K^* known.

Note that the three approaches are infeasible in practice, since they all assume K^* is known. The k-CFC is applied to the simulated datasets from (M1)–(M5) with $K = 2$ as an initial input, and it is increased by one until $K = 10$ is reached ($K = 12$ in the case of (M4)). None of the model selection methods returned the maximum allowed K as the final number of clusters in the simulation study below.

In view of the curve linear regression modeling conducted in the next step, we regard each curve as consisting of the regressor and the response curves, each part being observed over equal length of grids ($p/2$), and we denote the regressor and the response curves by $\{X_i(\cdot), Y_i(\cdot)\}_{i=1}^n$ for the training set and $\{\widetilde{X}_i(\cdot), \widetilde{Y}_i(\cdot)\}_{i=1}^{\widetilde{n}}$ for the test set. This

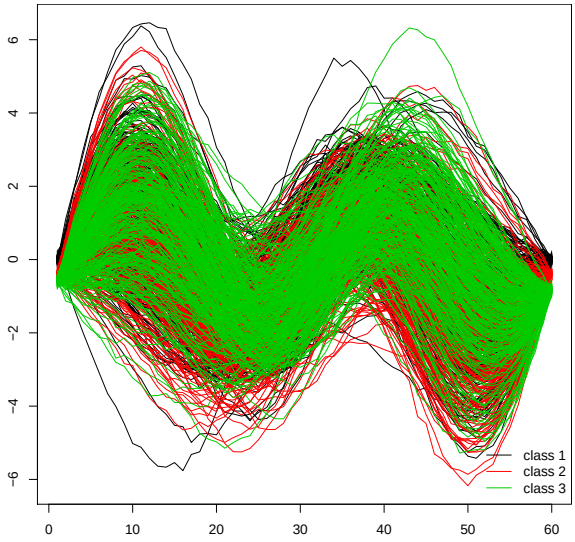

Figure 3: $Z_i(\cdot),\ i = 1, \ldots, n$ generated from (M1) where $K^* = 3$.

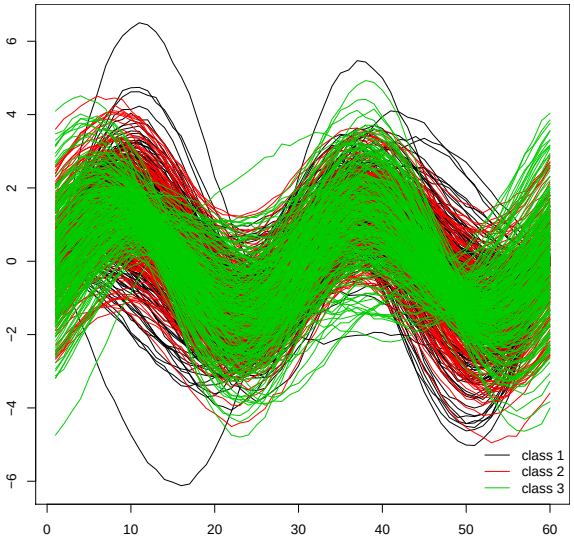

Figure 4: $Z_i(\cdot),\ i = 1, \ldots, n$ generated from (M2) where $K^* = 3$.

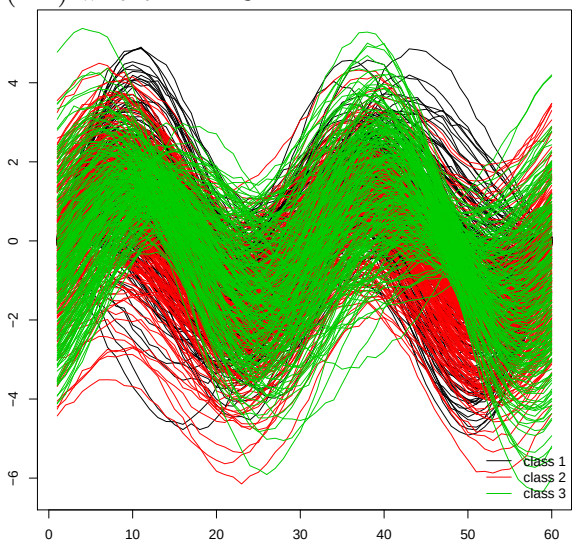

Figure 5: $Z_i(\cdot),\ i = 1, \ldots, n$ generated from (M3) where $K^* = 3$.

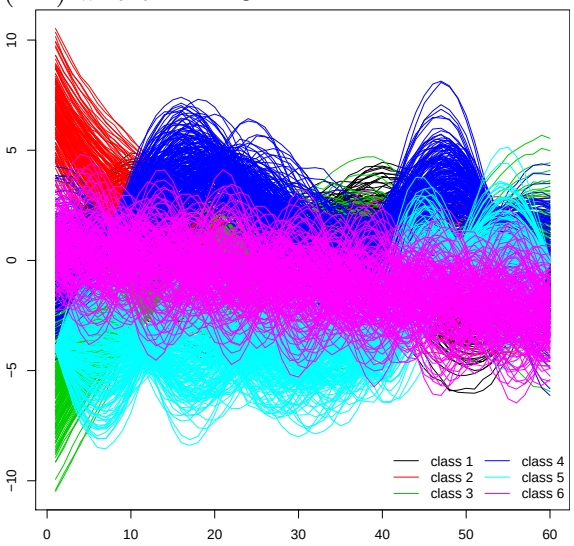

Figure 6: $Z_i(\cdot),\ i = 1, \ldots, n$ generated from (M4) where $K^* = 6$.

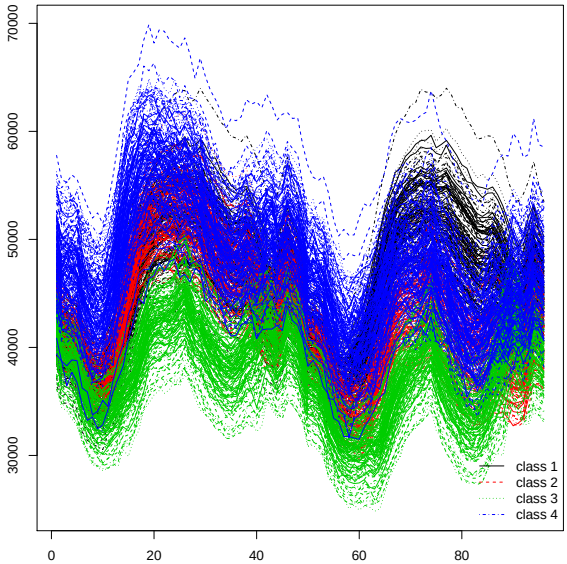

Figure 7: $Z_i(\cdot),\ i = 1, \ldots, n$ generated from (M5) where $K^* = 4$.

notion comes naturally in the setting of (M5), where each curve joins the electricity loads observed on two consecutive days.

Then, when predicting the cluster memberships of the test set, we assume $\widetilde{Z}_i(\cdot)$ is observed partially such that only $\widetilde{X}_i(\cdot)$ is available, and its membership is predicted as follows:

$$\widehat{C}(\widetilde{Z}_i) = \arg\min_{k=1,\ldots,K} \|\widetilde{X}_i - \widehat{\mathcal{P}}_k^X(\widetilde{X}_i)\|_2^2, \tag{17}$$

where each $\widehat{\mathcal{P}}_k^X(\cdot)$ is estimated from the training set using $X_i^{(k)}(\cdot)$ belonging to $\widehat{\mathcal{C}}_k$ in the same manner as $\widehat{\mathcal{P}}_k(\cdot)$. The results reported in Tables 2–6 are based on the 100 replications generated from each model.

Overall, it is apparent that the k-CFC can identify the true clustering for the models considered in this study, when combined with a suitable model selection technique, regardless of whether the dissimilarities between the clusters originates from either the mean function or the modes of variation or both. Some of the model selection criteria perform as well as the oracle approach, and occasionally even better due to the cluster merging. Although equipped with the prior information on true K^*, the initial or the k-means procedures often return poor clusterings as they do not account for the entire dependence structure differentials. Gordon and Henderson (1977) noted that the k-means clustering with the sum of squares criterion tends to find clusters of hyperspherical shapes, such that long, straggling sets of points (e.g., when there exist high correlations within each observed vector) are broken up into several different groups, which accounts for poor clusterings returned by the two methods.

Since the simulated datasets used in the study are generated from the models in the form of the truncated Karhunen-Loève expansion in (16), the k-CFC performs well when combined with $S(\cdot)$, an objective function which is designed specifically to take into account the properties of the projection operators employed by the classification criterion. It outperforms other model selection methods in terms of both the total number of clusters and the adjust Rand index for (M1)–(M4). On the other hand, $S(\cdot)$ tends to under-estimate the number of clusters for (M5), which implies that the within-cluster and between-clusters dissimilarity measures adopted by $S(\cdot)$ are not suitable for detecting differences among the clusters in (M5).

While the FFT performs well for (M1)–(M3) and (M5), correctly identifying K^* over at least 70% of the simulated datasets, it tends to over-estimate the number of clusters for (M4) by additionally splitting a true cluster. This may be due to the bootstrap resampling scheme which, as mentioned in Section 3.3.2, generates bootstrap samples

that resemble the data too closely and thus leads to false alarms.
$H(\cdot)$ prefers clusterings with $K > K^*$ and large adjusted Rand indices, presumably by including an additional cluster resulted from splitting a true cluster. While the FFT achieves better accuracy than $H(\cdot)$ in terms of the number of total clusters, $\widehat{\theta}$ and $\widetilde{\theta}$ returned by the two methods are comparable. We note that global approaches are allowed to compare the clusterings returned by the k-CFC with a wider range of K as an input compared to the FFT (due to its stopping rule), and thus benefit from cluster merging occurring during the re-classification step.
$CH_1(\cdot)$ and $CH_2(\cdot)$ often under-estimate the number of clusters, as the reduction in WCSS due to adding another cluster does not outweigh the reduction in BCSS according to the scalings adopted by these criteria. Exceptions occur for (M4) (in the case of $CH_1(\cdot)$) and (M5), where the two methods prove to be effective in identifying the true clusters, and this is more striking for the latter model where other well-performing criteria fail to do so. Finally, since $S^o(\cdot)$ and $KL(\cdot)$ are originally proposed for multivariate data clustering (the latter even involves the dimensionality p in the criterion), their performance is not outstanding on most of the simulated examples.
We conclude that depending on the underlying structure of the data, different model selection methods are particularly more suited.

4.2 Performance of curve linear regression models

To assess the performance of curve linear regression methods when combined with the k-CFC, we apply the three methods described in Section 2 (denoted by SVD, FPC and Bayes respectively), to the curves clustered as in the previous section. Using each technique, we estimate the curve linear regression models within the clusters resulted from applying various model selection methods on the training set. Then on the test set, according to the predicted memberships of $\widetilde{Z}_i$ (see (17)), we forecast the corresponding response $\widetilde{Y}_i(\cdot)$ as $\widehat{\widetilde{Y}}_i(\cdot)$ using the curve linear regression model estimated from the cluster $\widehat{C}(\widetilde{Z}_i)$. The forecasting performance of SVD, FPC and Bayes is evaluated using the root-mean-square error (RMSE) $\sqrt{\widetilde{n}^{-1}\sum_{i=1}^{\widetilde{n}} \|\widetilde{Y}_i - \widehat{\widetilde{Y}}_i\|_2^2}$ with $\widetilde{n} = \sum_{k=1}^{K} \widetilde{n}_k$. In Tables 2–6, we report the prediction errors of different curve linear regression methods combined with different model selection criteria for clustering. We also present the prediction results obtained with the knowledge of the true memberships of $\widetilde{Z}_i$, which is reported under the heading "oracle*".
Naturally, good membership forecasts lead to better forecasts of $\widetilde{Y}_i$, and the advantage is more pronounced with the SVD method than the other two. It is also this SVD-based

Table 2: Summary of simulation results for (M1) where $K^* = 3$. The largest $\widehat{\theta}$, $\widetilde{\theta}$ and the smallest RMSE, among those returned by different model selection methods and forecasting models, are marked by $*$.

	clustering							RMSE					
	$\widehat{K}$							SVD		FPC		Bayes	
	1	2	3	4	5	$\widehat{\theta}$	$\widetilde{\theta}$	mean	sd	mean	sd	mean	sd
S^o	0	95	**5**	0	0	0.586	0.585	0.094	0.02	0.518	0.121	1.53	0.144
S	0	0	**100**	0	0	1*	1*	0.074*	0.001	0.421	0.122	1.523	0.127
CH_1	0	94	**6**	0	0	0.596	0.594	0.092	0.006	0.516	0.123	1.53	0.143
CH_2	0	95	**5**	0	0	0.586	0.585	0.094	0.018	0.52	0.119	1.53	0.144
H	0	0	**69**	30	1	0.953	0.958	0.074*	0.001	0.424	0.12	1.524	0.125
KL	0	37	**62**	1	0	0.823	0.826	0.087	0.036	0.467	0.13	1.521	0.144
FFT	0	0	**92**	8	0	0.85	0.84	0.109	0.072	0.447	0.134	1.529	0.126
initial	–	–	–	–	–	0.021	0.423	0.282	0.021	0.694	0.088	1.737	0.185
k-means	–	–	–	–	–	0.017	0.413	0.279	0.019	0.7	0.076	1.723	0.165
oracle	–	–	–	–	–	1	1	0.074	0.001	0.421	0.122	1.523	0.127
oracle*	–	–	–	–	–	–	–	0.074	0.001	0.421	0.122	1.523	0.127

Table 3: Summary of simulation results for (M2) where $K^* = 3$.

	clustering									RMSE					
	$\widehat{K}$									SVD		FPC		Bayes	
	1	2	3	4	5	6	7	$\widehat{\theta}$	$\widetilde{\theta}$	mean	sd	mean	sd	mean	sd
S^o	0	86	**12**	2	0	0	0	0.396	0.337	0.419	0.248	0.79	0.133	1.565	0.193
S	0	0	**100**	0	0	0	0	0.999*	0.993*	0.056*	0.006	0.466	0.157	1.538	0.119
CH_1	0	65	**34**	0	1	0	0	0.665	0.583	0.326	0.305	0.723	0.205	1.569	0.196
CH_2	0	97	**3**	0	0	0	0	0.412	0.339	0.44	0.264	0.793	0.139	1.553	0.209
H	0	0	**93**	7	0	0	0	0.988	0.983	0.056*	0.006	0.467	0.156	1.539	0.119
KL	0	22	**63**	5	3	5	2	0.856	0.834	0.129	0.135	0.558	0.19	1.534	0.162
FFT	0	8	**77**	10	3	2	0	0.771	0.729	0.219	0.259	0.602	0.215	1.562	0.128
initial	–	–	–	–	–	–	–	0.029	0.014	0.647	0.071	0.805	0.115	1.552	0.232
k-means	–	–	–	–	–	–	–	0.026	0.022	0.632	0.07	0.769	0.083	1.501	0.17
oracle	–	–	–	–	–	–	–	0.999	0.993	0.056	0.006	0.466	0.157	1.538	0.119
oracle*	–	–	–	–	–	–	–	–	–	0.054	0.001	0.47	0.158	1.542	0.115

linear regression method which consistently achieves the smallest forecasting RMSE in combination with well-performing model selection methods, and in such case, the RMSE is as small as the error of the oracle* predictor. As noted in Cho et al. (2013a), good performance of the SVD-based method is attributed to the fact that SVD singles out the directions upon which the projections of $Y_i(\cdot)$ are most correlated with $X_i(\cdot)$. These arguments do not apply to the principal components, and it is reflected in relatively poor performance of the FPC and the Bayes methods.

Occasionally, over-estimating the number of clusters obtains good forecasting results. As mentioned previously, a clustering with $K > K^*$ returned by the k-CFC is likely to have an additional cluster resulted from further partitioning a true homogeneous cluster, which lead to over-fitted forecasting models. This is observable especially with the clustering results returned by $H(\cdot)$, which is consistent with the observations made on its clustering performance in the previous section.

Table 4: Summary of simulation results for (M3) where $K^* = 3$.

	clustering							RMSE					
	$\widehat{K}$							SVD		FPC		Bayes	
	1	2	3	4	5	$\widehat{\theta}$	$\widetilde{\theta}$	mean	sd	mean	sd	mean	sd
S^o	0	86	**12**	1	1	0.53	0.412	0.516	0.273	1.033	0.173	1.911	0.231
S	0	0	**100**	0	0	1*	0.996*	0.064*	0.004	0.573	0.18	1.943	0.152
CH_1	0	86	**14**	0	0	0.558	0.44	0.511	0.323	1.026	0.202	1.907	0.235
CH_2	0	99	**1**	0	0	0.502	0.377	0.547	0.297	1.041	0.173	1.907	0.242
H	3	0	**71**	25	1	0.929	0.933	0.083	0.111	0.582	0.194	1.942	0.156
KL	0	14	**85**	0	1	0.93	0.912	0.119	0.159	0.633	0.242	1.927	0.169
FFT	0	6	**84**	9	1	0.845	0.805	0.213	0.243	0.723	0.254	1.935	0.19
initial	–	–	–	–	–	0.073	0.089	0.757	0.108	1.015	0.102	1.941	0.207
k-means	–	–	–	–	–	0.056	0.13	0.719	0.1	0.995	0.102	1.914	0.177
oracle	–	–	–	–	–	1	0.996	0.064	0.004	0.573	0.18	1.943	0.152
oracle*	–	–	–	–	–	–	–	0.062	0.001	0.573	0.18	1.943	0.152

Table 5: Summary of simulation results for (M4) where $K^* = 6$.

	clustering											RMSE					
	$\widehat{K}$											SVD		FPC		Bayes	
	1	2	3	4	5	6	7	8	9	$\widehat{\theta}$	$\widetilde{\theta}$	mean	sd	mean	sd	mean	sd
S^o	0	0	0	0	33	**67**	0	0	0	0.943	0.939	0.202	0.258	0.574	0.155	1.654	0.121
S	0	0	0	0	0	**93**	7	0	0	0.994*	0.991*	0.182	0.242	0.529	0.131	1.631	0.124
CH_1	0	0	0	0	5	**92**	3	0	0	0.99	0.986	0.189	0.251	0.537	0.137	1.635	0.126
CH_2	0	24	8	0	41	**27**	0	0	0	0.705	0.644	0.549	0.568	0.894	0.45	2.09	0.702
H	0	0	0	2	2	**17**	46	26	7	0.939	0.937	0.105*	0.165	0.506	0.13	1.633	0.121
KL	0	3	7	18	4	**29**	37	2	0	0.835	0.827	0.268	0.34	0.65	0.27	1.794	0.334
FFT	0	0	0	0	8	**56**	33	3	0	0.909	0.911	0.189	0.235	0.567	0.154	1.649	0.123
initial	–	–	–	–	–	–	–	–	–	0.727	0.801	0.445	0.268	0.764	0.108	1.791	0.142
k-means	–	–	–	–	–	–	–	–	–	0.608	0.78	0.674	0.468	0.93	0.313	1.923	0.307
oracle	–	–	–	–	–	–	–	–	–	0.986	0.98	0.201	0.251	0.542	0.14	1.646	0.127
oracle*	–	–	–	–	–	–	–	–	–	–	–	0.053	0.001	0.468	0.096	1.635	0.114

Table 6: Summary of simulation results for (M5) where $K^* = 4$.

	clustering							RMSE							
	$\widehat{K}$							SVD		FPC		Bayes	
	1	2	3	4	5	6	7	$\widehat{\theta}$	$\widetilde{\theta}$	mean	sd	mean	sd	mean	sd
S^o	0	0	100	**0**	0	0	0	0.713	0.682	941.28	175.78	1827.54	499.11	1279.18	847.96
S	0	97	3	**0**	0	0	0	0.351	0.297	2552.76	401.60	3145.63	347.39	3217.01	497.91
CH_1	0	0	2	**98**	0	0	0	0.996	0.973	579.36	164.01	1456.93	320.73	1057.21	665.15
CH_2	0	0	1	**99**	0	0	0	0.998*	0.975*	577.36*	161.96	1451.24	315.66	1055.89	665.49
H	0	0	0	**35**	51	12	2	0.933	0.914	597.20	163.31	1443.15	315.93	1048.94	637.81
KL	0	32	3	**60**	5	0	0	0.776	0.743	1222.41	936.47	2002.39	864.94	1808.85	1169.36
FFT	0	0	0	**71**	29	0	0	0.962	0.95	593.83	159.03	1461.84	325.61	1078.02	670.59
initial	–	–	–	–	–	–	–	0.827	0.875	848.32	203.40	1954.53	401.30	1151.56	494.41
k-means	–	–	–	–	–	–	–	0.784	0.787	861.14	387.26	1880.41	573.51	1279.27	788.46
oracle	–	–	–	–	–	–	–	0.999	0.977	573.95	160.14	1451.43	315.54	1052.82	66.49
oracle*	–	–	–	–	–	–	–	–	–	399.95	26.16	1365.56	333.52	947.83	711.47

5 Clustering and forecasting of daily electricity loads

We apply the combined methodology of functional data clustering and curve linear regression modeling, to the problem of forecasting daily electricity loads. To this end, we use the French electricity load dataset collected between 1 January 1996 and 31 December 2008 as the training set ($\{Z_i(\cdot)\}_{i=1}^n,\ n = 4749$), and that collected between 1 January 2009 and 31 December 2009 as the test set ($\{\widetilde{Z}_i(\cdot)\}_{i=1}^n,\ n = 365$).
As EDF produces one day-ahead forecast of electricity loads at noon, we set the observations made over the half-hour grids from the noon of day $(i-1)$ to the noon of day i as the regressor curve $X_i(\cdot)$ (total 48 grids). Accordingly, we define the corresponding response curve $Y_i(\cdot) = X_{i+1}(\cdot)$ and the joined curve $Z_i(\cdot) = (X_i, Y_i)(\cdot)$, and the test set is defined similarly as $\widetilde{Z}_i(\cdot) = (\widetilde{X}_i, \widetilde{Y}_i)(\cdot)$.
To recap, the typical modeling and forecasting procedure is as follows.

Step 1: Clustering and modeling.

Step 1.1: Choose an initial cluster number K_0 and let $K \leftarrow K_0$.

Step 1.2: Apply the k-CFC to partition $Z_i(\cdot),\ i = 1, \ldots, n$ into K homogeneous subgroups and let $K \leftarrow K + 1$.

Step 1.3: Repeat Step 1.2 until K reaches the pre-specified maximum cluster number K_1, and apply model selection methods to identify $\widehat{K}$ and the final clustering $\widehat{\mathcal{C}}_k,\ k = 1, \ldots, \widehat{K}$.

Step 1.4: Within each $\widehat{\mathcal{C}}_k$, fit a curve linear regression model

$$Y_i^{(k)}(u) = \mu_Y^{(k)}(u) + \int_{v \in \mathcal{I}_2} \beta^{(k)}(u, v)\{X_i^{(k)}(v) - \mu_X^{(k)}(v)\} + \varepsilon_i^{(k)}(u)$$

by estimating $\widehat{\mu}_X^{(k)}(\cdot)$, $\widehat{\mu}_Y^{(k)}(\cdot)$ and $\widehat{\beta}^{(k)}$.

Step 2: Forecasting.

Step 2.1: Classify each $\widetilde{Z}_i(\cdot)$ to one of the $\widehat{K}$ clusters as described in (17), i.e. its membership $\widehat{C}(\widetilde{Z}_i)$ is determined as

$$\widehat{C}(\widetilde{Z}_i) = \arg \min_{1 \le k \le \widehat{K}} \|\widetilde{X}_i - \widehat{\mathcal{P}}_k^X(\widetilde{X}_i)\|_2^2.$$

Step 2.2: Forecast each $\widetilde{Y}_i(\cdot)$ as $\widehat{\widetilde{Y}}_i(u) = \widehat{\mu}_Y^{(c)}(u) + \int_{v \in \mathcal{I}_2} \widehat{\beta}^{(c)}(u, v)\{\widetilde{X}_i(v) - \widehat{\mu}_X^{(c)}(v)\}$ with $c = \widehat{C}(\widetilde{Z}_i)$.

The difficulty in performing the Step 1.4 of the above procedure on the current training set is that, as noted in Introduction, the mean and the modes of variation of $Z_i(\cdot)$ depend heavily on the calendar variables such as the corresponding day of a week and the month of a year. Therefore, it is conceivable that there are a lot more classes in this dataset compared to those models considered (where $K = 3, \ldots, 6$) in Section 4. Indeed, classifying the training set with the above two calendar variables as a classification rule, we obtained 92 classes in total. (Note that in applying the classification rule, the size of each class was required to be greater than or equal to 10, and the classes with fewer than 10 members were re-classified to the classes with more than 10 members according to the L_2-distance in (12)).

We have observed that the model selection criteria in Section 3.3.1 often failed to return a sufficiently large K as the number of clusters. That is, when combined with the k-CFC as in the above Step 1, they often attained the highest value at $K = 2$ (minimum allowable number of clusters) then gradually decreased as K increased. The FFT also prefers the clustering with small K, and tends not to reject the null hypotheses for K greater than 10.

Instead, we have devised a new local approach to accommodate the prior knowledge on the electricity load patterns, as well as effectively handling the large size of the data. Commencing with the $K = 92$ classes obtained as above, we merge two classes $\mathcal{C}_{k_0}$ and $\mathcal{C}_{l_0}$ which are *nearest* to each other in the sense that

$$(k_0, l_0) = \arg \min_{1 \le k \ne l \le K} \sum_{i=1}^{n_k} \|Z_i^{(k)} - \widehat{\mathcal{P}}_l(Z_i^{(k)})\|_2^2 + \sum_{i=1}^{n_l} \|Z_i^{(l)} - \widehat{\mathcal{P}}_k(Z_i^{(l)})\|_2^2. \tag{18}$$

With the reduced number of classes, we repeatedly perform the merging until $K = 2$. At each iteration, the curve linear regression models are fitted within each class using the three methods SVD, FPC and Bayes (as in the above Step 1.4), each test set observation is classified to one of the K clusters (Step 2.1), and the forecast of $\widetilde{Y}_i$ is produced (Step 2.2).

(a) When fitting the curve linear regression models, it has been noted in Cho et al. (2013a) and Cho et al. (2013b) that including the temperature observations $T_i(\cdot)$ (made over the same grids as $X_i(\cdot)$) and forecasts $T_i^F(\cdot)$ (made over the same grids as $Y_i(\cdot)$) in the curve regressor, often bring in superior forecasts. Therefore below we report the RMSE obtained on the fitted ($\{\widehat{Y}_i\}_{i=1}^n$) and predicted ($\{\widehat{\widetilde{Y}}_i\}_{i=1}^{\widetilde{n}}$) curves from the curve linear regression models with the joined curve $(X_i, T_i, T_i^F)(\cdot)$ as regressor. Note that the fitted curves $\{\widehat{Y}_i\}_{i=1}^n$ are obtained using leave-one-out estimators of $\mu_Y^{(k)}(\cdot)$, $\mu_X^{(k)}(\cdot)$ and $\beta^{(k)}$ in the curve linear regression models.

(b) It has been pointed out during our application study that, while $Z_i(\cdot)$ joining two consecutive daily loads are identifiable on any day of a week, it is not the case when performing the membership prediction. For example, while daily electricity load curves observed on Tuesday (i.e. from Monday midday to Tuesday midday), Wednesday, Thursday and Friday of the same week behave very close to each other, the load curves on Tuesday–Thursday and that on Friday behave markedly different due to the economic cycle. It implies that the membership prediction criterion in Step 2.2 is likely to regard $\widetilde{Z}_i(\cdot)$ on Tuesday-Thursday and that on Friday as belonging to the same class, and this misclassification leads to poor one-day ahead load forecasting. To remedy this, when classifying the test data, we use the seven-day load curves $X_i^o(\cdot) = (X_{i-6}, \ldots, X_i)(\cdot)$ and $\widetilde{X}_i^o(\cdot) = (\widetilde{X}_{i-6}, \ldots, \widetilde{X}_i)(\cdot)$ in place of $X_i(\cdot)$ and $\widetilde{X}_i(\cdot)$, respectively.

The results of merging are summarized in Figure 8, which shows the minimum distance between the clusters, the fitted RMSE of $\{\widehat{Y}_i(\cdot)\}_{i=1}^n$ and the forecast RMSE of $\{\widehat{\widetilde{Y}}_i(\cdot)\}_{i=1}^{\widetilde{n}}$ from SVD and FPC. The results from Bayes are omitted from the figure for better presentation, as the corresponding RMSE (in the range of 6000–10000 MW) is far greater than the RMSE produced by the other two methods. For a reference, some model selection criteria computed on the clustering at each iteration as well as the results from performing the FFT, are presented in Figure 9.

Overall, the SVD performs far superior to the FPC, consistently with the results from our simulation study in Section 4.2. As expected, the minimum distance between any two clusters increases as K decreases. On the other hand, the fitted and the forecast RMSE remain almost constant as K decreases from $K = 92$ to around $K = 30$, which implies that not much is lost in terms of between-class dissimilarities by locally merging the classes defined by the calendar variables up to a certain K.

By studying which of the two classes are merged at each iteration, we gain an insight into the patterns of successive daily loads; full list of the merged classes and the corresponding calendar variables is provided in the table of Appendix B, which also reports the iteration-by-iteration fitted and the forecast RMSE. From the table, it can be concluded that the merging occurs between the classes associated with warmer months (April–October) first then moves on to those associated with colder months (November–December, January–March). Also, the classes corresponding to Tuesday–Thursday within the same month tend to get merged together, then those corresponding to nearby months (e.g. June and July) are merged. Overall, these patterns agree with the opinions of the experts at EDF.

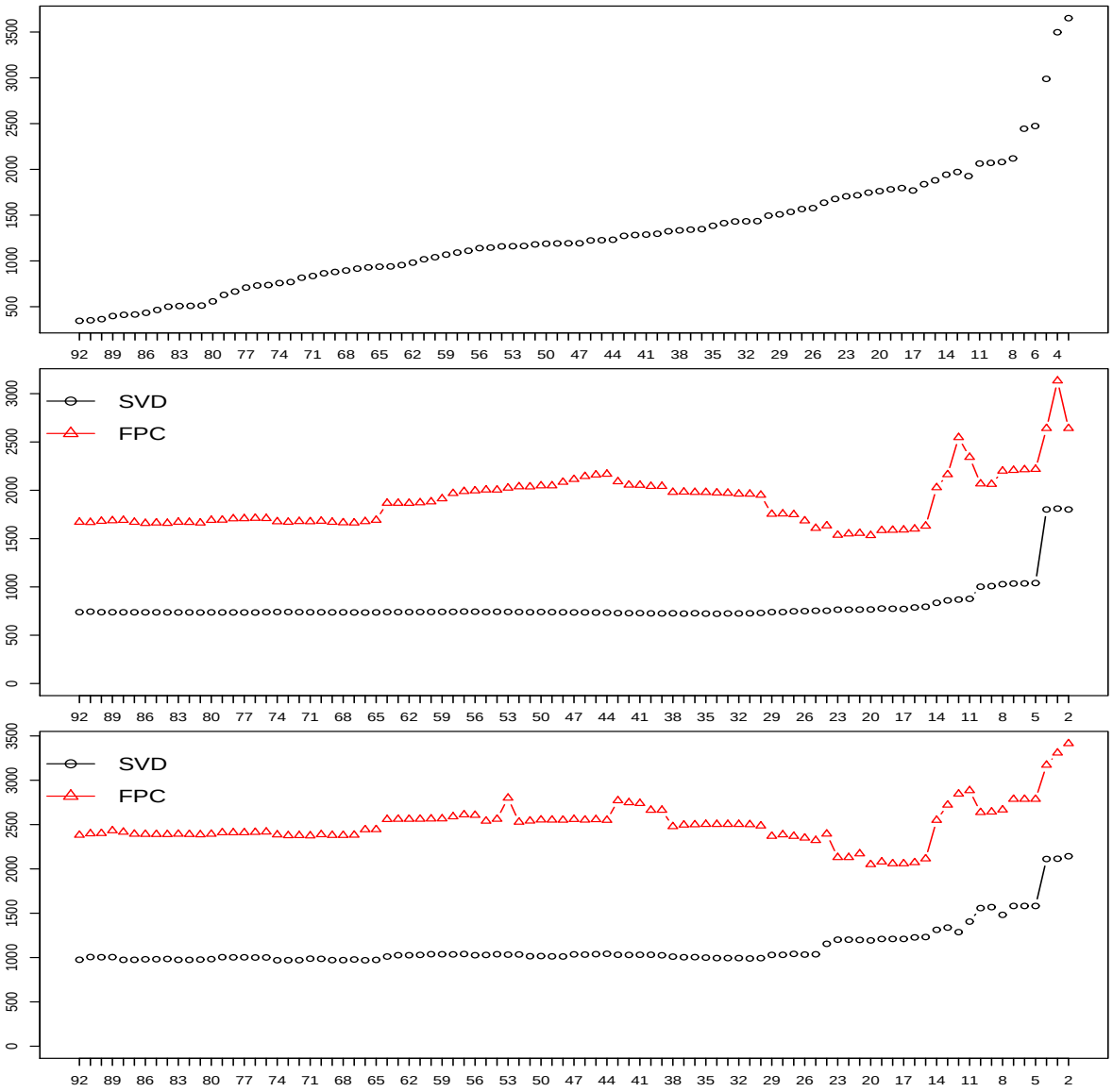

Figure 8: Top: distance between $\mathcal{C}_{k_0}$ and $\mathcal{C}_{l_0}$ at each iteration; middle: fitted RMSE (in MW) of $\{\widehat{Y}_i(\cdot)\}_{i=1}^n$ from SVD (empty circle) and FPC (empty triangle) on the training set; bottom: forecast RMSE (in MW) of $\{\widetilde{\widehat{Y}}_i(\cdot)\}_{i=1}^{\widetilde{n}}$ on the test set.

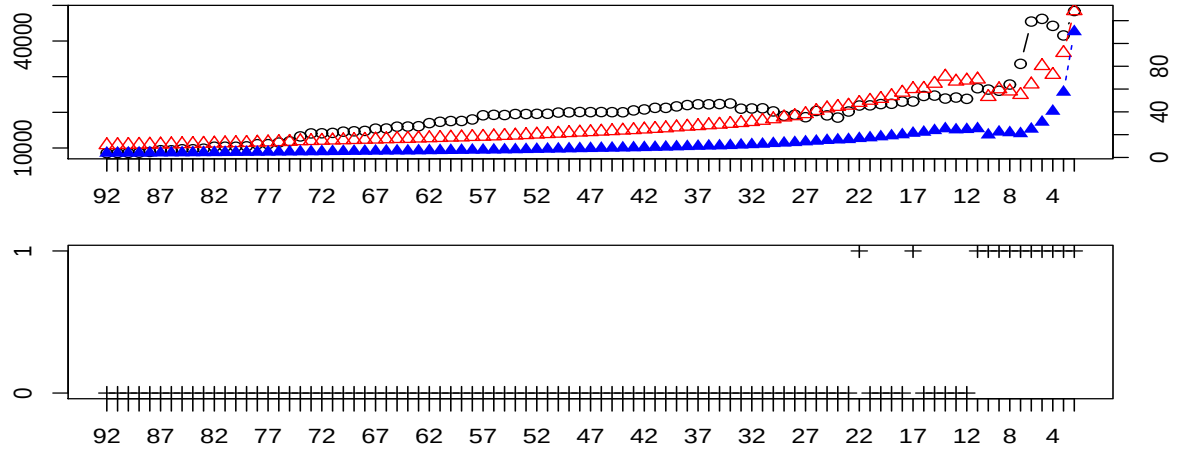

Figure 9: Top:$S(K)$ (empty circle, left y-axis), $CH_1(K)$ (empty triangle, right y-axis) and $CH_2(K)$ (filled triangle, right y-axis) over $K = 92, \ldots, 2$; bottom: The results of performing FFT over $K = 92, \ldots, 2$ where 0 indicates not rejecting the null hypotheses (and thus not requiring an additional cluster) and 1 indicates rejecting at least one of the null hypotheses at $\alpha = 0.05$

6 Conclusions

In this paper, we have studied the combined methodology of functional data clustering and curve linear regression for modeling and forecasting daily electricity loads. Functional data clustering is achieved by the k-CFC, which partitions the data into sub-groups of homogeneous mean and covariance structures such that the curve linear regression technique is applicable to each of the sub-groups separately. We have also proposed some extensions of clustering selection criteria and studied their performance on simulated datasets in conjunction with the k-CFC.

The French electricity load dataset, according to the classification rule based on the calendar variables, consists of 92 classes. Applying the methodology to the dataset in its modified form, it is clear that with the reduced number of classes, the curve linear regression method still returns forecasts which are as good as those from the 92 classes. Also we can collect information on the patterns of daily electricity loads from studying the order of the occurring of merging, which coincides with the observations made from extensively studying the French electricity load patterns at EDF.

While it still remains as a task to develop a fully adaptive clustering procedure for the electricity load curves which does not rely on any prior information, it would be interesting to see how well the current methodology works when applied to the domestic and industrial usage separately. We defer these questions to future work.

A Lemma 1

Lemma 1. The estimated projection operator $\widehat{\mathcal{P}}_k^*$ satisfies

$$\|\widehat{\mathcal{P}}_k^*(Z_i) - \widetilde{\mathcal{P}}_k^*(Z_i)\|_2 = O_p\{(n_k^*)^{-1/2}\},$$

for all $k = 1, \ldots, K^*$.

Proof. Note that

$$\begin{aligned}
\widehat{\zeta}_j^{*(k)}(Z_i) &= \langle Z_i - \widehat{\mu}^{*(k)}, \widehat{\rho}_j^{*(k)} \rangle = \langle Z_i - (\mu^{*(k)} + O_p\{(n_k^*)^{-1/2}\}), \rho_j^{*(k)} + O_p\{(n_k^*)^{-1/2}\} \rangle \\
&= \langle Z_i - \mu^{*(k)}, \rho_j^{*(k)} \rangle + O_p\{(n_k^*)^{-1/2}\} = \zeta_j^{*(k)}(Z_i) + O_p\{(n_k^*)^{-1/2}\}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\widehat{\mathcal{P}}_k^*(Z_i)(t) &= \widehat{\mu}^{*(k)}(t) + \sum_{j=1}^{d_k} \widehat{\zeta}_j^{*(k)}(Z_i)\widehat{\rho}_j^{*(k)}(t) \\
&= \mu^{*(k)}(t) + O_p\{(n_k^*)^{-1/2}\} + \sum_{j=1}^{d_k} (\zeta_j^{*(k)}(Z_i) + O_p\{(n_k^*)^{-1/2}\})(\rho_j^{*(k)}(t) + O_p\{(n_k^*)^{-1/2}\}) \\
&= \mu^{*(k)}(t) + \sum_{j=1}^{d_k} \zeta^{*(k)}(Z_i)\rho_j^{*(k)}(t) + O_p\{(n_k^*)^{-1/2}\} = \widetilde{\mathcal{P}}_k^*(Z_i)(t) + O_p\{(n_k^*)^{-1/2}\}.
\end{aligned}$$

B Table summarizing the results from merging

List of the two classes merged at each iteration and the associated calendar variables along with the iteration-by-iteration fitted and the forecast RMSE. Three leading days and months associated with $X_i(\cdot)$ are presented. NAs correspond to classes which consist of invalidated observations or those from bank holidays.

	RMSE (MW)		Class k_0						Class l_0					
K	fitted	forecast	month			day			month			day		
92	739	975	Jul	–	–	Wed	–	–	Jul	–	–	Thu	–	–
91	744	1006	Jun	–	–	Wed	–	–	Jun	–	–	Thu	–	–
90	738	1004	Aug	–	–	Thu	–	–	Aug	–	–	Wed	–	–
89	738	1005	Jul	–	–	Tue	–	–	Jul	–	–	Thu	Wed	–
88	737	974	Aug	–	–	Thu	Wed	–	Aug	–	–	Tue	–	–
87	737	974	Jun	–	–	Tue	–	–	Jun	–	–	Thu	Wed	–
86	737	980	Jun	–	–	Sat	–	–	Jul	Jun	–	Sat	–	–
85	737	981	Sep	–	–	Wed	–	–	Sep	–	–	Thu	–	–
84	737	984	May	–	–	Fri	–	–	Jun	–	–	Fri	–	–
83	736	974	Sep	–	–	Tue	–	–	Sep	–	–	Wed	Thu	–
82	736	974	May	–	–	Wed	–	–	May	–	–	Thu	–	–
81	735	977	Jun	–	–	Thu	Tue	Wed	Jul	–	–	Tue	Thu	Wed
80	736	982	May	–	–	Tue	–	–	May	–	–	Thu	Wed	–

79	736	1004	Jun	May	–	Fri	–	–	Jul	–	–	Fri	–	–
78	736	1002	Jun	–	–	Mon	–	–	Jul	–	–	Mon	–	–
77	735	1003	May	–	–	Sat	–	–	Jun	Jul	–	Sat	–	–
76	735	1000	Jul	–	–	Sun	–	–	Aug	–	–	Sun	–	–
75	739	1001	May	–	–	Tue	Thu	Wed	Jul	Jun	–	Thu	Tue	Wed
74	740	970	Jul	Jun	–	Mon	–	–	Aug	–	–	Mon	–	–
73	740	970	Oct	–	–	Wed	–	–	Oct	–	–	Thu	–	–
72	739	971	Oct	–	–	Tue	–	–	Oct	–	–	Thu	Wed	–
71	738	987	Jun	Jul	May	Sat	–	–	Aug	Jul	–	Sat	–	–
70	738	985	Apr	–	–	Tue	–	–	Apr	–	–	Wed	–	–
69	737	971	Apr	–	–	Mon	–	–	NA	–	–	NA	–	–
68	737	971	Apr	–	–	Tue	Wed	–	Apr	–	–	Thu	–	–
67	736	978	Nov	–	–	Wed	–	–	Nov	–	–	Thu	–	–
66	735	970	Oct	–	–	Sat	–	–	Sep	–	–	Sat	–	–
65	736	973	Jul	Jun	May	Thu	Tue	Wed	Aug	–	–	Tue	Thu	Wed
64	739	1011	May	–	–	Sun	–	–	Jun	–	–	Sun	–	–
63	739	1026	May	–	–	Mon	–	–	NA	–	–	NA	–	–
62	739	1026	Sep	–	–	Fri	–	–	Oct	–	–	Fri	–	–
61	740	1030	Sep	–	–	Sun	–	–	Oct	–	–	Sun	–	–
60	741	1038	Jun	Jul	May	Fri	–	–	Aug	–	–	Fri	–	–
59	742	1037	May	–	–	Mon	–	–	Jul	Jun	Aug	Mon	–	–
58	742	1036	Sep	–	–	Tue	Wed	Thu	Oct	–	–	Thu	Wed	Tue
57	745	1040	Nov	–	–	Sun	–	–	Dec	–	–	Sun	–	–
56	744	1026	Nov	–	–	Tue	–	–	Nov	–	–	Thu	Wed	–
55	741	1029	Feb	–	–	Sat	–	–	Mar	–	–	Sat	–	–
54	743	1037	Jan	–	–	Wed	–	–	Jan	–	–	Thu	–	–
53	741	1033	Jan	–	–	Tue	–	–	Jan	–	–	Wed	Thu	–
52	740	1035	Feb	–	–	Thu	–	–	Feb	–	–	Tue	–	–
51	738	1015	Sep	–	–	Mon	–	–	Oct	–	–	Mon	–	–
50	740	1017	Feb	–	–	Tue	Thu	–	Feb	–	–	Wed	–	–
49	739	1013	Dec	Nov	–	Sat	–	–	Nov	–	–	Sat	–	–
48	737	1012	Dec	–	–	Tue	–	–	Nov	–	–	Tue	Thu	Wed
47	736	1036	Dec	–	–	Wed	–	–	Nov	Dec	–	Tue	Thu	Wed
46	736	1034	Dec	–	–	Fri	–	–	Nov	–	–	Fri	–	–
45	734	1038	Nov	Dec	–	Tue	Wed	Thu	Dec	–	–	Thu	–	–
44	734	1042	Jan	–	–	Tue	Wed	Thu	Nov	Dec	–	Tue	Thu	Wed
43	729	1031	Jan	–	–	Fri	–	–	Nov	Dec	–	Fri	–	–
42	728	1030	Feb	–	–	Fri	–	–	Mar	–	–	Fri	–	–
41	729	1031	Jan	–	–	Mon	–	–	Nov	–	–	Mon	–	–
40	726	1031	Mar	–	–	Wed	–	–	Mar	–	–	Thu	–	–
39	725	1026	Jan	–	–	Sat	–	–	Nov	Dec	–	Sat	–	–
38	727	1010	Feb	–	–	Sun	–	–	Mar	–	–	Sun	–	–
37	723	1004	Apr	–	–	Sun	–	–	Jun	May	–	Sun	–	–
36	727	1005	Feb	–	–	Tue	Thu	Wed	Mar	–	–	Thu	Wed	–
35	723	1000	Feb	–	–	Mon	–	–	Mar	–	–	Mon	–	–
34	722	994	NA	–	–	NA	–	–	Jun	Aug	Jul	Fri	–	–
33	725	994	NA	–	–	NA	–	–	Jun	Apr	May	Sun	–	–
32	725	994	Feb	Mar	–	Thu	Wed	Tue	Mar	–	–	Tue	–	–
31	726	990	Jun	Jul	Aug	Sat	–	–	Sep	Oct	–	Sat	–	–
30	730	993	Jul	Jun	Aug	Tue	Thu	Wed	Sep	Oct	–	Thu	Tue	Wed
29	739	1030	Jan	Nov	–	Mon	–	–	Dec	–	–	Mon	–	–
28	739	1031	Jun	Apr	May	Sun	–	–	Sep	Oct	–	Sun	–	–
27	747	1041	Sep	Oct	–	Fri	–	–	Jun	Aug	Jul	Fri	–	–
26	749	1034	Sep	Oct	–	Mon	–	–	Jul	Jun	Aug	Mon	–	–
25	753	1037	Apr	–	–	Sat	–	–	Sep	Jun	Oct	Sat	–	–

24	754	1155	Jan	Nov	Dec	Tue	Thu	Wed	Mar	Feb	–	Thu	Tue	Wed
23	764	1202	NA	–	–	NA	–	–	Jul	Aug	–	Sun	–	–
22	764	1202	Jan	Nov	Dec	Sat	–	–	Mar	Feb	–	Sat	–	–
21	765	1199	Jan	–	–	Sun	–	–	Nov	Dec	–	Sun	–	–
20	766	1193	Sep	Oct	Jul	Mon	–	–	Apr	–	–	Mon	–	–
19	777	1210	Jan	Nov	Dec	Fri	–	–	Mar	Feb	–	Fri	–	–
18	774	1210	Sep	Oct	Jun	Sun	–	–	NA	–	–	NA	–	–
17	772	1210	Jul	Aug	–	Sun	–	–	Sep	Oct	Jun	Sun	–	–
16	787	1227	Apr	–	–	Fri	–	–	Jun	Sep	Oct	Fri	–	–
15	794	1231	Apr	–	–	Thu	Tue	Wed	Sep	Oct	Jul	Thu	Tue	Wed
14	835	1313	Mar	Feb	Jan	Sat	–	–	Sep	Jun	Oct	Sat	–	–
13	860	1338	Sep	Oct	Jun	Sun	–	–	Mar	Feb	–	Sun	–	–
12	869	1287	Sep	Mar	Oct	Sun	–	–	Jan	Nov	Dec	Sun	–	–
11	877	1406	Mar	Feb	Jan	Thu	Tue	Wed	Sep	Oct	Jul	Thu	Tue	Wed
10	1002	1558	Jan	Nov	Dec	Mon	–	–	Mar	Feb	–	Mon	–	–
9	1008	1569	Mar	Feb	Jan	Fri	–	–	Jun	Sep	Oct	Fri	–	–
8	1028	1482	Mar	Jan	Feb	Mon	–	–	Sep	Oct	Jul	Mon	–	–
7	1035	1582	Mar	Jun	Sep	Fri	–	–	NA	–	–	NA	–	–
6	1036	1582	NA	–	–	NA	–	–	Mar	Sep	Jun	Sat	–	–
5	1040	1582	Mar	Sep	Oct	Thu	Tue	Wed	Mar	Jun	Sep	Fri	–	–
4	1801	2112	Mar	Sep	Oct	Thu	Fri	Tue	Mar	Sep	Jun	Sat	–	–
3	1811	2115	Sep	Mar	Oct	Mon	–	–	Mar	Sep	Oct	Thu	Sat	Fri
2	1842	2143	–	–	–	–	–	–	–	–	–	–	–	–

References

Antoniadis, A., Brosat, X., Cugliari, J., and Poggi, J. (2011), "Clustering functional data using wavelets," *Arxiv preprint arXiv:1101.4744*.

Bathia, N., Yao, Q., and Ziegelmann, F. (2010), "Identifying the finite dimensionality of curve time series," *Annals of Statistics*, 38, 3352–3386.

Bruhns, A., Deurveilher, G., and Roy, J. S. (2005), "A non-linear regression model for mid-term load forecasting and improvements in seasonality," in *The 15th Power Systems Computation Conference, Liege, Belgium*.

Caliński, T. and Harabasz, J. (1974), "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, 3, 1–27.

Chiou, J.-M. (2012), "Dynamical functional prediction and classification, with application to traffic flow prediction," *Annals of Applied Statistics*, 6, 1588–1614.

Chiou, J. M. and Li, P. L. (2007), "Functional clustering and identifying substructures of longitudinal data," *Journal of the Royal Statistical Society: Series B*, 69, 679–699.

Chiou, J. M., Müller, H. G., and Wang, J. L. (2004), "Functional response models," *Statistica Sinica*, 14, 659–677.

Cho, H., Goude, Y., Brossat, X., and Yao, Q. (2013a), "Modelling and forecasting daily electricity load curves: a hybrid approach," *Journal of the American Statistical Association*, 108, 7–21.

— (2013b), "Modelling and forecasting daily electricity load via curve linear regression," *In submission*.

Efron, B. and Tibshirani, R. J. (1994), *An Introduction to the Bootstrap*, Chapman and Hall/CRC.

Gordon, A. D. (1999), *Classification, (Chapman & Hall/CRC Monographs on Statistics & Applied Probability)*, Chapman and Hall/CRC.

Gordon, A. D. and Henderson, J. T. (1977), "An algorithm for Euclidean sum of squares classification," *Biometrics*, 355–362.

Hall, P. and Horowitz, J. L. (2007), "Methodology and convergence rates for functional linear regression," *Annals of Statistics*, 35, 70–91.

Hall, P. and Vial, C. (2006), "Assessing the finite dimensionality of functional data," *Journal of the Royal Statistical Society: Series B*, 68, 689–705.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer New York.

Hubert, L. and Arabie, P. (1985), "Comparing partitions," *Journal of classification*, 2, 193–218.

Jacques, J. and Preda, C. (2013), "Functional data clustering: a survey," *Inria Research Report No. 8198*.

James, G. M. and Sugar, C. A. (2003), "Clustering for sparsely sampled functional data," *Journal of the American Statistical Association*, 98, 397–408.

Krzanowski, W. J. and Lai, Y. (1988), "A criterion for determining the number of groups in a data set using sum-of-squares clustering," *Biometrics*, 44, 23–34.

Lam, C. and Yao, Q. (2012), "Factor modelling for high-dimensional time series: inference for the number of factors," *Annals of Statistics*, 40, 694–726.

Li, P. and Chiou, J. (2011), "Identifying cluster number for subspace projected functional data clustering," *Computational Statistics & Data Analysis*, 55, 2090–2103.

Pollard, D. (1981), "Strong Consistency of k-Means Clustering," *Annals of Statistics*, 9, 135–140.

Ramsay, J. O. and Dalzell, C. J. (1991), "Some tools for functional data analysis (with discussions)," *Journal of the Royal Statistical Society: Series B*, 53, 539–572.

Ramsay, J. O. and Silverman, B. W. (2005), *Functional Data Analysis*, Springer, New York.

Rousseeuw, P. J. (1987), "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, 20, 53–65.

Smithies, F. (1937), "The eigenvalues and singular values of integral equations." in *Proceedings of the London Mathematical Society*, pp. 255–279.

Tibshirani, R., Walther, G., and Hastie, T. (2001), "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B*, 63, 411–423.

Yao, F., Müller, H. G., and Wang, J. L. (2005), "Functional data analysis for sparse longitudinal data," *Journal of the American Statistical Association*, 100, 577–590.

Zhou, R. R., Serban, N., and Gebraeel, N. (2011), "Degradation modeling applied to residual lifetime prediction using functional data analysis," *Annals of Applied Statistics*, 5, 1586–1610.