Electronic Journal of Statistics
Vol. 13 (2019) 2093–2120
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1573

Regularised forecasting via smooth-rough partitioning of the regression coefficients

Hyeyoung Maeng* **and Piotr Fryzlewicz**

Department of Statistics, London School of Economics, UK
e-mail: h.maeng@lse.ac.uk*;* p.fryzlewicz@lse.ac.uk

Abstract: We introduce a way of modelling temporal dependence in random functions $X(t)$ in the framework of linear regression. Based on discretised curves $\{X_i(t_0), X_i(t_1), \dots, X_i(t_T)\}$, the final point $X_i(t_T)$ is predicted from $\{X_i(t_0), X_i(t_1), \dots, X_i(t_{T-1})\}$. The proposed model flexibly reflects the relative importance of predictors by partitioning the regression parameters into a smooth and a rough regime. Specifically, unconstrained (rough) regression parameters are used for observations located close to $X_i(t_T)$, while the set of regression coefficients for the predictors positioned far from $X_i(t_T)$ are assumed to be sampled from a smooth function. This both regularises the prediction problem and reflects the 'decaying memory' structure of the time series. The point at which the change in smoothness occurs is estimated from the data via a technique akin to change-point detection. The joint estimation procedure for the smoothness change-point and the regression parameters is presented, and the asymptotic behaviour of the estimated change-point is analysed. The usefulness of the new model is demonstrated through simulations and four real data examples, involving country fertility data, pollution data, stock volatility series and sunspot number data. Our methodology is implemented in the R package `srp`, available from CRAN.

Keywords and phrases: Change-point detection, prediction, penalised spline, functional linear regression.

Received October 2018.

1. Introduction

Over the last few decades, functional data analysis (FDA) has been growing in importance and enjoying increased attention. Functional objects arise in many contexts and the applications in the literature include prediction of daily curves of particulate matter in the air (Aue et al., 2015), testing stationarity of intraday price curves of a financial asset (Horváth et al., 2014), modelling the dynamics of fertility rate (Chen et al., 2017), studying the effect of air pollution on mortality rate across cities (Kong et al., 2016), prediction of the protein content of meat from spectral curves (Zhu et al., 2014), investigation of a bike sharing system by predicting bike pick-up counts (Han et al., 2017), choosing predictive days from daily egg-laying counts for fruit flies (Ji and Müller, 2017) and predicting

[*] Corresponding author.

sucrose content of orange juice from its near-infrared spectrum (Ferraty et al., 2010).

In this paper, we consider random functions $X_i \in L^2[a,b]$ where $i = 1,\ldots,n$ and $[a,b]$ is a compact subset of $\mathbb{R}$. If the functions are used as a predictor for explaining a scalar response variable Y, this simply describes the standard functional linear regression which has been widely studied in the literature. The reader is referred to Reiss et al. (2017) for a review of numerous approaches to scalar-on-function regression. On the other hand, if the random functions X_i are believed to possess temporal dependence and are analysed by separating the domain they live on into shorter units, we call such a data structure functional time series. Functional time series analysis has been an active field of research in recent years. The best-known model in this area is the first-order functional autoregressive model proposed by Bosq (2000). Other recent contributions include testing for stationarity (Horváth et al., 2014), testing for mean functions in a two-sample problem (Horváth et al., 2013), testing for error correlation (Gabrys et al., 2010) and prediction (Antoniadis et al., 2006; Aue et al., 2015).

In practice, functional data are often observed on a grid, rather than continuously. The observation of i.i.d. square-integrable random functions $X_i(t) \in L^2[a,b]$ on an equispaced grid $\{t_0, t_1, \ldots, t_T\}$ gives the discretised curves $\{X_i(t_0), X_i(t_1), \ldots, X_i(t_T)\}$ for $i = 1,\ldots,n$ where $t_0 = a$ and $t_T = b$. Based on these design points, our objective in this work is to predict the final point $X_i(t_T)$ from the past observations $\{X_i(t_0),\ldots,X_i(t_{T-1})\}$. This is an important applied problem in a variety of fields, including public health, earth sciences, finance and environment, as our data examples of Section 4 illustrate. Arguably the simplest statistical framework for expressing the dependence of $X_i(t_T)$ on $\{X_i(t_0),\ldots,X_i(t_{T-1})\}$ is linearity, and with this in mind, this work focuses on the following model:

$$X_i(t_T) = \mu + \sum_{j=1}^{T} \alpha_j X_i(t_{T-j}) + \varepsilon_i, \quad i = 1,\ldots,n. \tag{1}$$

We now discuss its specifics. In our asymptotic considerations, we work with a fixed T; however, in practice, T can be large (e.g. two of the datasets in Section 4 have T roughly of the order of n), which inevitably brings us into a high-dimensional setting and the set of parameters α_j cannot be estimated well by classical approaches. In addition, we often experience a high degree of collinearity between the predictors. As a way of regularising the problem, our proposal in this work is to split the set of parameters $\{\alpha_1,\ldots,\alpha_T\}$ into two, $\{\alpha_1,\ldots,\alpha_q\}$ and $\{\alpha_{q+1},\ldots,\alpha_T\}$, and assume that the second set is discretised from a smooth curve $\beta(t)$, which gives

$$X_i(t_T) = \mu + \sum_{j=1}^{q} \alpha_j X_i(t_{T-j}) + \sum_{j=q+1}^{T} \beta(t_{T-j}) X_i(t_{T-j}) + \varepsilon_i, \quad i = 1,\ldots,n, \tag{2}$$

where the final point $X_i(t_T)$ is a scalar response variable, $\{X_i(t_{T-j}),_{j=1,\ldots,T}\} \in \mathbb{R}^T$ represents scalar predictors and ε_i denotes the error term with $E(\varepsilon_i|X_i(t_{T-j})$

$_{,j=1,\ldots,T}) = 0$ and unknown variance σ^2. Since all the dependent and independent variables are obtained from random functions, we assume them to be random. The unknown parameter set contains a constant $\mu \in \mathbb{R}$, real and scalar $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T \in \mathbb{R}^q$, real and functional $\beta \in L^2[t_0, t_{T-q-1}]$ and a change-point index parameter q. Throughout the paper, we will be referring to (2) as the Smooth-Rough Partition (SRP) model. The SRP model assumes that the change-point index q is unknown, and we estimate it from the data via a change-point detection technique. This is possible because we will be assuming that the coefficients α_j are rougher than the coefficients $\beta(t_{T-j})$, i.e. exhibit more variation.

We now motivate the smooth-rough partitioning idea in more detail. The partitioning of the regression coefficients into two classes of smoothness captures the difference in the relative importance of the observations in predicting the final point $X_i(t_T)$. Constraining the β's to be smooth reflects the relatively lower importance of the more remote observations, whose influence on $X_i(t_T)$ is 'bundled together' by the smoothness restriction in β. By contrast, the unconstrained parameters $\boldsymbol{\alpha}$ are not connected to each other in any (functional) way, so are able to capture any arbitrary linear influence of the near observations on $X_i(t_T)$. The smoothness assumptions on $(\boldsymbol{\alpha}, \beta)$ will be specified in Section 5.

The smooth-rough partitioning results in regression estimation that is interpretable in the sense that it automatically separates the effects that can be seen as "long-term" (these are the ones corresponding to the smooth portion of the parameter vector) from those that can be seen as "instantaneous" (these are the ones that correspond to the rough portion of the parameter vector). In other words, the SRP framework can be seen as a "two-scale" approach to linear prediction, where the two scales are defined by both the smoothness and the extent of the regression parameter vector (i.e. the long, smooth portion and the short, rough portion). As will be demonstrated in Section 4, this two-scale framework is useful in various real-world datasets e.g. fertility rate data, high-frequency stock volatility series, Mexico city pollution data and sunspot number data. Each of them appears to display both long-term and instantaneous temporal dependences, which (as we illustrate) are well captured by the SRP model. For example, it is reasonable to believe that in the context of the pollution data, the level of pollution in a particular curve at a particular time depends both on the overall shape and level of the curve up until the current time (which could be seen as the long-term effect) and the levels immediately preceding the current time in question (which can be seen as the instantaneous effect). We can attach similar interpretations to the other datasets studied in the paper.

Additionally, the SRP framework can also be useful in the modelling and forecasting of univariate time series, especially those that are believed to be well modelled as AR (autoregressive) processes with large orders, in which case the SRP smoothing device would be able to offer both regularisation and (hopefully) interpretability, especially if the time series is believed to possess long-range dependence (which will typically be the case if an AR model with a large order is used in the first place). Our sunspot example in Section 4 illustrates this.

Model (2) covers two special cases: 1) in the case of $q = T$, i.e. if we ignore the constrained part, then it has the form of multiple linear regression $X_i(t_T) = \mu + \sum_{j=1}^{T} \alpha_j X_i(t_{T-j}) + \varepsilon_i$ and 2) when $q = 0$, i.e. without the unconstrained part, if the summation is replaced by integration with a large enough T, then it becomes scalar on function regression with $X_i(t_T) = \mu + \int_{t_0}^{t_{T-1}} \beta(t) X_i(t) dt + \varepsilon_i$. Unlike the former, completely unconstrained case, the regularisation in model (2) operates in a way that reduces the model's degrees of freedom. In the examples of Section 4, we empirically show that the full model (2) exhibits better prediction performance than these two extreme cases. This further justifies our efforts in proposing a methodology for detecting the change-point index q automatically from the data.

Other ways of regularising the functional linear regression coefficient have been proposed in the literature. In particular, some researchers have used ideas from variable selection to obtain $\beta(t) = 0$ for the non-informative subintervals and $\beta(t) \neq 0$ for the informative ones. James et al. (2009) employ the LASSO and Dantzig selector with the aim of improving the interpretability of $\beta(t)$ while Zhou et al. (2013) use the Dantzig selector and SCAD approaches. Lin et al. (2015) propose a functional version of SCAD by combining the SCAD method and smoothing splines to obtain a smooth and sparse estimator for the functional coefficient. By contrast, we do not regularise by finding null subregions of $\beta(t)$ but by imposing different smoothness constraints over different sections of the parameter curve. The 'null subregion' and our approach are compared and contrasted in Sections 3 and 4.

In addition, our approach is different from the functional linear regression model with points of impact (Kneip et al., 2016) in the sense that our unrestricted coefficients are grouped into a single region that is the nearest to the time-location of the response variable, which (in contrast to Kneip et al. (2016)) allows us not to have to remove the observations adjoining the points of impact in estimating their locations, which would be impossible in our time series context. Other methods related to Kneip et al. (2016) but less so to our work have also been proposed: McKeague and Sen (2010) aim to estimate a single point of impact with the motivation from gene expression data and Ferraty et al. (2010) fit a nonparametric model after finding several predictive design points. The performance of our technique is compared to that of Ferraty et al. (2010) in Sections 3 and 4.

Change-point detection ideas have been proposed in other functional regressions contexts before. Hall and Hooker (2016) find the truncation point θ under the truncated functional linear model $Y_i = \mu + \int_0^{\theta} \beta(t) X_i(t) dt + \varepsilon_i$. Goia and Vieu (2015) use two functions $\beta_1(t)$ and $\beta_2(t)$ by dividing the interval into two with one discontinuity point. They suggest the partitioned functional single index model, $Y_i = \mu + g_1\big(\int_{[0,\lambda]} \beta_1(t) X_i(t) dt\big) + g_2\big(\int_{(\lambda,1]} \beta_2(t) X_i(t) dt\big) + \varepsilon_i$, where g_1 and g_2 are smooth functions to be estimated and the breakpoint λ identifies a discontinuity in the functional regression coefficient. Neither of these methods use their concept of change-point detection to differentiate between two classes of smoothness.

If q in model (2) were known, the skeleton of our model would be similar to that of partial functional linear regression with both scalar and functional covariates, recently studied by e.g. Kong et al. (2016), Zhou et al. (2016), Zhou and Chen (2012), Shin and Lee (2012), Shin (2009), Aneiros-Pérez and Vieu (2008) and Goia (2012). Apart from assuming q to be unknown, (2) is different in that it operates in a time series context.

The remainder of the article is organised as follows. Section 2 describes the model and the parameter estimation procedure. The supporting simulation studies are outlined in Section 3, with further real-data illustrations in Section 4 regarding country fertility data, Mexico city pollution data, stock volatility series and sunspot number data. The relevant theoretical results are presented in Section 5 and we end with additional discussion in Section 6. The SRP methodology is implemented in the R package `srp` and the proofs of our main theoretical results are in Appendix A.

2. Model and its estimation

We work with the discretised curves $\{X_i(t_0), \ldots, X_i(t_T)\}_{i=1,\ldots,n}$ observed from each function $X_i(t)$ on the equispaced $T+1$ discrete points including both endpoints. Since the regression coefficients vary by q, we rewrite model (2) as

$$X_i(t_T) = \mu^q + \sum_{j=1}^{q} \alpha_j^q X_i(t_{T-j}) + \sum_{j=q+1}^{T} \beta^q(t_{T-j}) X_i(t_{T-j}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where $1 \leq q \leq T$. The point t_{T-q} is where a sudden smoothness change occurs in the sequence of the regression coefficients, with the coefficients α_j^q being unconstrained in terms of their smoothness and the coefficients $\beta^q(t_{T-j})$ assumed to be a sampled version of a smooth function. The change-point location in (3) is the same for all i's. Our expectation is that q is substantially smaller than T and the optimal q is chosen by examining a number of q's over a subset of $\{1, \ldots, T\}$, which we specify in Section 2.1. The reason why T is assumed to be fixed is that if we were to allow $T \to \infty$, then t_T would asymptotically approach t_{T-1} and we could simply predict $X(t_T)$ by $X(t_{T-1})$.

The set of unknown parameters in (3) can be categorised into two types: 1) change-point t_{T-q} and 2) regression coefficients $(\mu^q, \boldsymbol{\alpha}^q, \beta^q)$. Our interest includes the estimation of the underlying smooth function $\beta(t)$. Broadly speaking, two possible ways exist: 1) estimate $(\hat{\beta}^q(t_0), \ldots, \hat{\beta}^q(t_{T-q-1}))$ and then use interpolation to obtain the functional form of $\hat{\beta}^q(t)$ or 2) obtain the interpolant $\{X(t),\ t \in [t_0, t_{T-q-1}]\}$ and then estimate the function $\hat{\beta}^q(t)$ through basis expansion. In this paper, we use the latter approach as it is more popular and the former approach needs a particular penalty to make it feasible if T is close to or exceeding n. Examples of the former can be found in Cardot et al. (2007) and Crambes et al. (2009).

The interpolant $\{X_i(t),\ t \in [t_0, t_{T-q-1}]\}$ is obtained from the discrete observations $(X_i(t_0), \ldots, X_i(t_{T-q-1}))$ using natural cubic splines with knots at

$(t_0, \dots, t_{T-q-1})$. As stated in Crambes et al. (2009), the essential property of natural splines is that for any vector, the unique natural spline interpolant exists and it can be expressed as a B-spline expansion with dimension equal to 'number of knots + 2' (in our case $T - q + 2$) as follows:

$$X_i(t) = \sum_{h=1}^{T-q+2} d_{ih} B_h(t), \quad t \in [t_0, t_{T-q-1}], \tag{4}$$

where $B_h(t)$ is a set of basis functions for the normalised B-splines $\{B_h\}_{h=1,\dots,T-q+2}$.

Dimension reduction is necessary for the estimation of $\beta(t)$. The required regularisation is usually achieved by a basis expansion, which enables a finite number of basis functions to approximate the infinite-dimensional function. Numerous approaches are available, such as via the Fourier series, functional principal components (PC), splines or wavelets. The reader is referred to Ramsay and Silverman (2005) for more details. In what follows, we use B-splines. Cardot et al. (2003) argue that spline estimators should be preferred to the functional PC approach when $X(t)$ is rough and the functional coefficient is smooth, which is the case we are interested in. Moreover, a spline estimator is not directly affected by the estimation of the eigenstructure of the covariance operator of $X(t)$.

Let $\mathcal{S}$ be the space of splines defined on $[t_0, t_{T-q-1}]$ with degree s and $k-1$ equispaced interior knots where $L = k + s$ denotes the dimension of $\mathcal{S}$. Then one can derive a set of basis functions from the normalised B-splines $\{B_l\}_{l=1,\dots,L}$ to approximate $\beta^q(t)$ as

$$\beta^q(t) \approx \sum_{l=1}^{L} b_l B_l(t), \quad t \in [t_0, t_{T-q-1}], \tag{5}$$

where b_l represents the corresponding coefficient. For each t_{T-q}, the set of the regression parameters simplifies to $\delta^q = (\mu^q, \boldsymbol{\alpha}^q, b_1^q, \dots, b_L^q)^T \in \mathbb{R}^{1+q+L}$ where $\boldsymbol{\alpha}^q = (\alpha_1^q, \dots, \alpha_q^q)^T$. The choice of L is considered in Section 2.2.

2.1. Joint estimation procedure for parameters

We suggest a one-stage estimation procedure for the change-point and the regression parameters. Since the parameter q represents the number of scalar parameters, under fixed L, q itself determines the dimension of the model. Thus, using the well-known criterion of Schwarz (1978), we estimate q by minimising

$$SIC(q) = n \cdot \log M(q) + (q + L + 1) \cdot \log n, \tag{6}$$

where

$$M(q) = \frac{1}{n} \sum_{i=1}^{n} \left\{ X_i(t_T) - \hat{\mu}^q - \sum_{j=1}^{q} \hat{\alpha}_j^q X_i(t_{T-j}) - \sum_{j=q+1}^{T} \hat{\beta}^q(t_{T-j}) X_i(t_{T-j}) \right\}^2, \tag{7}$$

and $(\hat{\mu}^q, \hat{\alpha}_j^q,\ \hat{\beta}^q(t_{T-j}))$ are repeatedly estimated for each q by minimising the following sum of squared errors with appropriate penalisations,

$$\begin{aligned}(\hat{\boldsymbol{\alpha}}^q, \hat{\beta}^q(t)) = &\underset{\boldsymbol{\alpha}^q, \beta^q(t)}{\text{argmin}} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \tilde{X}_i(t_T) - \sum_{j=1}^{q} \alpha_j^q \tilde{X}_i(t_{T-j}) - \int_{t_0}^{t_{T-q-1}} \beta^q(t) \tilde{X}_i(t) dt \right\}^2 \right. \\ &\left. + \lambda_1 \delta_0^{q\,T} \delta_0^q + \lambda_2 \int_{t_0}^{t_{T-q-1}} \left\{ \beta^{q(m)}(t) \right\}^2 dt \right], \qquad (8) \\ \hat{\mu}^q = &\bar{X}(t_T) - \sum_{j=1}^{q} \hat{\alpha}_j^q \bar{X}(t_{T-j}) - \sum_{j=q+1}^{T} \hat{\beta}^q(t_{T-j}) \bar{X}(t_{T-j}),\end{aligned}$$

where $\delta_0^q = (\boldsymbol{\alpha}^q, b_1^q, \ldots, b_L^q)^T \in \mathbb{R}^{q+L}$, $\tilde{X}_i(t_{T-j})$ and $\tilde{X}_i(t)$ are demeaned predictors, $\bar{X}(t_{T-j}) = \frac{1}{n}\sum_{i=1}^{n} X_i(t_{T-j})$ and $\beta^{q(m)}(t)$ is the m^{th} derivative of $\beta^q(t)$ with the positive integer m satisfying $m < s$ where s denotes the degree of space $\mathcal{S}$.

The penalty terms in (8) contain two tuning parameters: λ_1 controls a ridge-type penalty and λ_2 governs the smoothness of the estimated $\hat{\beta}^q(t)$. We do not explicitly specify assumptions for the magnitudes of λ_1 and λ_2, but instead, as in Hall and Hooker (2016), our theoretical conditions (in Section 5) are phrased in terms of the appropriate convergence rates (Assumptions 3 and 4). In practice, only the initial values of λ_1 and λ_2 need to be specified by the user and the optimal values are selected automatically via a cross-validation-type criterion described in Section 2.2. If q were known, our task would be to estimate the regression parameters $(\mu^q, \boldsymbol{\alpha}^q, \beta^q)$. However, we assume that q is not known and estimate the parameters $(q, \mu^q, \boldsymbol{\alpha}^q, \beta^q)$ jointly. We preserve the original time scale of $\beta^q(t)$ instead of rescaling it to $[0, 1]$ so that we can place $\hat{\boldsymbol{\alpha}}^{\hat{q}}$ and $\hat{\beta}^{\hat{q}}(t)$ on the same time scale.

Let $q_0, \boldsymbol{\alpha}_0, \beta_0$ denote the true values of the parameters $q, \boldsymbol{\alpha}, \beta$, respectively. Typically, as a function of q, $M(q)$ decreases sharply as $q \uparrow q_0$, and becomes relatively flat (as $n \to \infty$) for $q \geq q_0$. For $q > q_0$, the smooth function $\beta_0(t)$ is estimated by the scalar estimators $(\hat{\alpha}_q, \ldots, \hat{\alpha}_{q_0+1})$ on the interval $[t_{T-q}, t_{T-q_0-1}]$. As the smoothness of $(\hat{\alpha}_q, \ldots, \hat{\alpha}_{q_0+1})$ is unrestricted, the fit is typically good, which causes the flat shape of $M(q)$. Conversely, when $q < q_0$, some of the unrestricted parameters, $(\alpha_{0,q_0}, \ldots, \alpha_{0,q+1})$, are estimated by the smooth $(\hat{\beta}^q(t_{T-q_0}), \ldots, \hat{\beta}^q(t_{T-q-1}))$, which typically causes $M(q)$ to be away from its minimum for $q < q_0$. The SIC penalty "lifts" the flat part of $M(q)$ and enables us to estimate the q parameter close to its true value. This is shown theoretically in Section 5 and numerically in Sections 3 and 4.

When finding the optimal q in (6), although q can in principle be large enough up to $q = T$, we recommend examining $1 \leq q \leq \bar{q}$, where $\bar{q}$ is substantially smaller than T. In the examples considered in Sections 3 and 4, we take $\bar{q} = \min(\lceil T \times 0.1 \rceil, 30)$. Based on our empirical experience, when q is large, there is the possibility that the optimisation of the two tuning parameters, λ_1 and λ_2 in (8), becomes unstable in that it becomes highly dependent on the selection of their initial values. In addition, examining the entire range $1 \leq q \leq T$ can make the algorithm unnecessarily slow especially when both T and n are large. In practice, even if we do not restrict q to be small (which introduces the stability and speed issues referred to above), the minimiser $\hat{q}$ of $SIC(q)$ in (6), if computed successfully despite the potential stability issues, is typically obtained to be substantially smaller than T.

2.2. Selection of the tuning parameters

To select the tuning parameters, we use the `magic` function from the R package `mgcv` (Wood (2006)). The `mgcv` includes various regression models such as GAM or the generalised ridge regression. The `magic` function is useful in that it is able to optimise over more than one penalty parameters (λ_1 and λ_2 in our case) by minimising GCV based on Newton's method. The results also give the estimators $(\hat{\boldsymbol{\alpha}}^{\hat{q}}, \hat{\beta}^{\hat{q}}(t))$ in (8) under each selected $\hat{q}$.

Regarding the dimension of β^q, we typically set L to be large but substantially smaller than $T-q$. As mentioned in Ruppert (2002), the number of basis function tends not to play an important role in functional linear regression with a roughness penalty, if we choose it to be large enough to prevent undersmoothing. Following the rule of thumb from Ruppert (2002), we use $L = 35$ in Sections 3 and 4, except in cases in which $T < 40$, when we use $L = 9$.

3. Simulations

In this section, we evaluate the finite-sample performance of our approach. We expect the performance of our method to vary depending on the size of change between $\beta_0(t)$ and $\boldsymbol{\alpha}_0^T$ and on the degree of fluctuations in the $\boldsymbol{\alpha}_0^T$ coefficients relative to the smoothness of $\beta_0(t)$.

Based on the model (3), we consider the following four parametric cases – Case 1: $\mu_0 = 0.0180$, $\boldsymbol{\alpha}_0 = (0.4, 0.2, 0.1)^T$, Case 2: $\mu_0 = -0.0836$, $\boldsymbol{\alpha}_0 = (0.6, -0.5, 0.4)^T$, Case 3: $\mu_0 = -0.0239$, $\boldsymbol{\alpha}_0 = (0.4, 0.2, 0.1)^T$ and Case 4: $\mu_0 = -0.0742$, $\boldsymbol{\alpha}_0 = (0.4, -0.2, 0.1)^T$, to investigate how the performance of change-point detection is affected by the degree of changes in the regression parameters. The true change-point index parameter is $q_0 = 3$ for all cases as shown in Figure 1 and $\beta_0(t)$ is available from the R package `srp`. In the data generating process based on the model (3), we use the Gaussian noise ε_i with the signal-to-noise ratio, defined as $\sigma^2_{\mathbf{X}}/\sigma^2$, equal to 4 where σ^2 is the error variance. In Cases 1 and 3, $\boldsymbol{\alpha}_0$ shows less fluctuation than in Cases 2 and 4. The size $|\alpha_{0,3} - \beta_0(t_{T-4})|$ of the change-point is approximately 0.4 in Case 2 and approximately 0.1 in the remaining three cases. Case 3 is similar to Case 1 except that its $\beta_0(t) = b_0 + b_1 t$ is linear. We simulate $n = 300$ independent copies of each process, in which the length of the sample is $T + 1 = 360$ (see formula (3)).

In each of 100 Monte Carlo runs, we split $n = 300$ observations into training and test sets of sizes $n_1 = 150$ and $n_2 = 150$, respectively. The training sample is used to obtain $\hat{q}$ and $(\hat{\boldsymbol{\alpha}}, \hat{\beta})$ by minimising (6) and (8), respectively. The

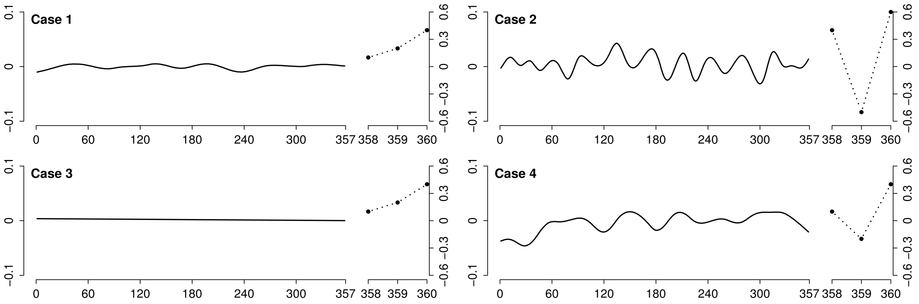

FIG 1. *True regression parameters of Cases 1-4 with different scale for each $\beta_0(t)$ (solid line) and $\boldsymbol{\alpha}_0^T$ (dots).*

accuracy of the regression parameter estimators can be evaluated by comparing $(\hat{\boldsymbol{\alpha}}^q, \hat{\beta}^q(t))$ and $(\boldsymbol{\alpha}_0, \beta_0(t))$; however, if the change-point is incorrectly estimated, i.e. $\hat{q} \neq q_0$, the length of the vector $\hat{\boldsymbol{\alpha}}^q$ is not matched with that of $\boldsymbol{\alpha}_0$ and neither is $\hat{\beta}^q(t)$. To circumvent this, we discretise $\hat{\beta}^q(t)$ and $\beta_0(t)$ and define $\hat{\boldsymbol{\gamma}}_{\hat{q}}$ and $\boldsymbol{\gamma}_0$ of dimension $T \times 1$ as $\hat{\boldsymbol{\gamma}}_{\hat{q}} = \left(\hat{\alpha}_1^{\hat{q}}, \ldots, \hat{\alpha}_{\hat{q}}^{\hat{q}}, \hat{\beta}^{\hat{q}}(t_0), \ldots, \hat{\beta}^{\hat{q}}(t_{T-\hat{q}-1})\right)^T$ and $\boldsymbol{\gamma}_0 = \left(\alpha_{0,1}, \ldots, \alpha_{0,q_0}, \beta_0(t_0), \ldots, \beta_0(t_{T-q_0-1})\right)^T$, which enables us to use the following sum-of-squared-errors (SSE) criterion:

$$\text{SSE} = \left[\hat{\boldsymbol{\gamma}}_{\hat{q}} - \boldsymbol{\gamma}_0\right]^T \left[\hat{\boldsymbol{\gamma}}_{\hat{q}} - \boldsymbol{\gamma}_0\right]. \tag{9}$$

The prediction performance is examined in the test sample by computing the mean-square prediction error (MSPE),

$$\text{MSPE} = \frac{1}{n_2} \sum_{i=1}^{n_2} \{X_i(t_T) - \hat{X}_i(t_T)\}^2, \tag{10}$$

where $\hat{X}_i(t_T)$ is the prediction using the estimated parameters $(\hat{q}, \hat{\mu}^{\hat{q}}, \hat{\boldsymbol{\alpha}}^{\hat{q}}, \hat{\beta}^{\hat{q}}(t))$.

3.1. Competing methods

We compare the performance of our approach to the following existing methodologies: multiple linear regression (**MLR**), ridge regression (**RIDGE**), functional linear regression with penalised B-splines (**FLR**, Cardot et al. (2003)), interpretable functional linear regression (**FLiRTI**, James et al. (2009)), most-predictive design points approach (**MPDP**, Ferraty et al. (2010)) and functional nonparametric regression (**NP**, Ferraty and Vieu (2002)). We also compare our proposal (**SRP**$_{\mathcal{C}}$) with its simplified version named **SRP**$_{\mathcal{L}}$, which follows the form of **SRP**$_{\mathcal{C}}$ except that $\beta_0(t)$ is estimated as a linear function. The corresponding objective functions for the parametric methods are as follows:

$$\textbf{MLR}: \hat{\boldsymbol{\alpha}}^{\hat{q}_1} = \underset{\boldsymbol{\alpha}^{\hat{q}_1}}{\text{argmin}} \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \tilde{X}_i(t_T) - \sum_{j=1}^{\hat{q}_1} \alpha_j^{\hat{q}_1} \tilde{X}_i(t_{T-j}) \right\}^2,$$

$$\textbf{FLR}: \hat{\beta}(t) = \underset{\beta(t)}{\operatorname{argmin}} \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \tilde{X}_i(t_T) - \int_{t_0}^{t_{T-1}} \beta(t) \tilde{X}_i(t) dt \right\}^2 + \lambda \int_{t_0}^{t_{T-1}} \left\{ \beta^{(m)}(t) \right\}^2 dt,$$

$$\textbf{SRP}_{\mathcal{L}}: (\hat{\boldsymbol{\alpha}}^{\hat{q}_2}, \hat{b}_0, \hat{b}_1) = \underset{\boldsymbol{\alpha}^{\hat{q}_2}, b_0, b_1}{\operatorname{argmin}} \frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ \tilde{X}_i(t_T) - \sum_{j=1}^{\hat{q}_2} \alpha_j^{\hat{q}_2} \tilde{X}_i(t_{T-j}) - \int_{t_0}^{t_{T-\hat{q}_2-1}} (b_0 + b_1 t) \tilde{X}_i(t) dt \right\}^2.$$

The objective function of our method ($\mathrm{SRP}_{\mathcal{C}}$) is in (8) and we determine $\hat{q}_1$ and $\hat{q}_2$ for MLR and $\mathrm{SRP}_{\mathcal{L}}$ by minimising $SIC(q)$ in (6) with appropriate $M(q)$ for each. In the implementation of FLR, we use cubic smoothing splines ($s = 3$) with the dimension $L = 35$ for both $\beta(t)$ and $X_i(t)$ where the derivative order of $\beta(t)$ is $m = 2$ and λ is selected by minimising GCV. Ridge parameter is also optimised by minimising GCV. For the implementation of other methods, we follow the suggestions of each paper for selecting the tuning parameters and the R code is available on the web (FLiRTI: http://www-bcf.usc.edu/~gareth/research/Research.html, MPDP and NP: http://www.math.univ-toulouse.fr/~ferraty/).

3.2. Simulation results

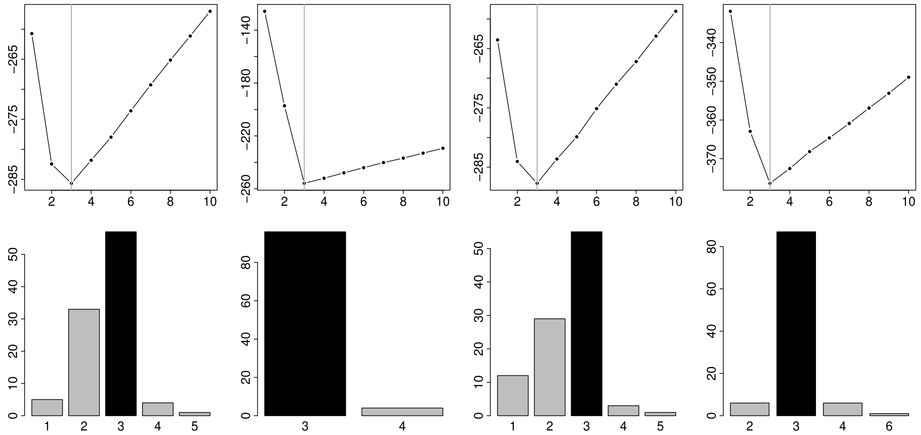

Fig 2. *(1^{st} row) Mean of $\{SIC(q)\}_{1 \le q \le 10}$ defined in formula (6) over 100 simulation runs for Cases 1-4 (1^{st}-4^{th} column); (2^{nd} row) Barplots of the 100 $\hat{q}$ estimated by minimising $\{SIC(q)\}_{1 \le q \le 30}$ where the black bars indicate the true change-point index parameter $q_0 = 3$.*

The top row of Figure 2 shows that the mean of 100 $SIC(q)$ is minimised at true $q_0 = 3$ for all cases. Case 2 shows a more rapid decrease than the other cases when $q \uparrow q_0$ due to the larger size of change at the change-point. Similarly, in the bottom row, we see that the mode of $\hat{q}$ is $q_0 = 3$ in all cases. Since Cases 1 and 3 have a relatively smooth $\boldsymbol{\alpha}$, $\hat{q} = 1, 2 (< q_0)$ are selected more frequently than in Cases 2 and 4, which have relatively more fluctuating $\boldsymbol{\alpha}$'s. Figure 3 provides numerical evidence of the increased closeness of $\hat{q}$ to q_0 in Case 4 as the sample size n increases.

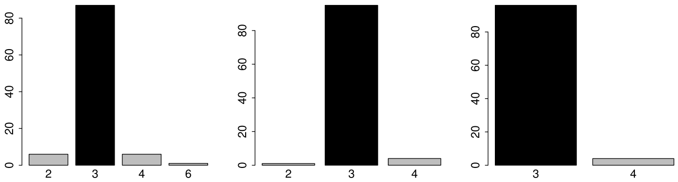

FIG 3. *Barplots of the* 100 $\hat{q}$ *estimated by minimising* $SIC(q)$ *in formula* (6) *with increasing* $n = 300, 600, 1200$ *under Case 4. The black bars indicate the true change-point index parameter* $q_0 = 3$.

As is apparent from Table 1, FLR and RIDGE perform systematically worse than the others. Our proposal, $\text{SRP}_{\mathcal{C}}$, outperforms the others in Cases 1, 2 and 4 and the difference is the most striking in Cases 2 and 4, in which a sudden smoothness change occurs. In Case 3, in which there is no clear smoothness change around the change-point and the true $\beta(t)$ is linear, $\text{SRP}_{\mathcal{L}}$ turns out to be the best-performing method.

TABLE 1
The mean(sd) of SSE($\times 10^2$*) defined in formula* (9) *over* 100 *simulation runs for the parametric methods in all cases. Bold: methods with the lowest mean of SSE.*

Case	MLR	FLR	FLiRTI	$\text{SRP}_{\mathcal{L}}$	$\text{SRP}_{\mathcal{C}}$	RIDGE
1	1.39(0.73)	5.32(1.33)	1.11(0.44)	1.43(0.69)	**1.00**(0.86)	19.90(2.05)
2	10.24(2.59)	75.08(1.76)	31.25(9.24)	9.09(1.03)	**2.06**(0.76)	72.80(2.87)
3	0.79(0.55)	5.28(1.30)	0.78(0.39)	**0.64**(0.56)	1.08(1.20)	19.12(1.91)
4	11.31(1.38)	21.37(1.63)	9.72(2.32)	6.96(0.79)	**1.03**(0.44)	24.30(1.23)

Examining Figure 4, while the misestimation in $\text{SRP}_{\mathcal{C}}$ is mainly located around the true change-point, in FLiRTI and FLR it is scattered over the whole interval. In addition, the graph offers visual confirmation of the superior performance of $\text{SRP}_{\mathcal{C}}$ in Cases 1, 2 and 4. In particular, in Cases 2 and 4, FLR ignores the sudden fluctuation in $\boldsymbol{\alpha}$ by estimating it as a smooth function. Unlike FLR and FLiRTI, $\text{SRP}_{\mathcal{C}}$ shows its advantages not only when scale changes are present (Cases 1 and 3) but also when a sudden smoothness change occurs at the change-point (Cases 2 and 4).

Table 2 contains two more columns than Table 1 as the mean-square prediction error can also be obtained for the nonparametric methods, MPDP and NP, which do not involve the estimation of $(\hat{\boldsymbol{\alpha}}, \hat{\beta}(t))$. In all cases considered, FLR, MPDP, NP and RIDGE show worse prediction performance than the other methods. $\text{SRP}_{\mathcal{C}}$ performs better than FLiRTI for all cases (but more noticeably so in Cases 2 and 4). $\text{SRP}_{\mathcal{C}}$ is superior to $\text{SRP}_{\mathcal{L}}$ in all cases except Case 3

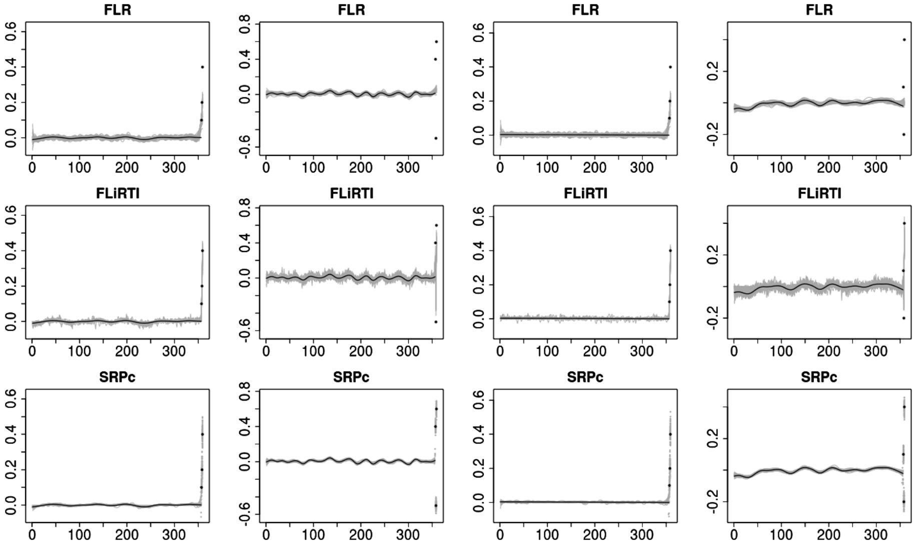

Fig 4. *True (black) and* 100 *estimated (grey) regression parameters for Cases 1-4*($1^{st} - 4^{th}$ *column) with three methods, FLR*(1^{st} *row), FLiRTI*(2^{nd} *row) and SRP*$_{\mathcal{C}}$(3^{rd} *row). The corresponding numerical summaries of these results are in Table 1.*

Table 2

The mean(sd) of MSPE($\times 10^2$) *defined in formula* (10) *over* 100 *simulation runs for all methods in all cases. Bold: methods with the lowest mean of MSPE.*

Case	MLR	FLR	FLiRTI	SRP$_{\mathcal{L}}$	SRP$_{\mathcal{C}}$	MPDP	NP	RIDGE
1	21.83	23.39	20.48	22.12	**18.95**	26.22	79.04	43.52
	(2.7)	(3.2)	(2.7)	(2.8)	(3.2)	(5.1)	(9.9)	(7.0)
2	53.97	83.38	51.71	50.81	**27.55**	69.20	102.21	94.76
	(7.0)	(9.9)	(9.3)	(7.1)	(4.4)	(21.4)	(11.5)	(11.3)
3	17.26	22.01	17.86	**15.61**	16.80	21.41	74.82	41.35
	(2.1)	(3.0)	(2.5)	(2.1)	(3.3)	(3.8)	(9.7)	(6.8)
4	30.48	28.17	22.17	22.05	**10.88**	39.18	43.54	35.68
	(4.2)	(4.2)	(4.1)	(2.8)	(1.6)	(15.7)	(5.6)	(4.3)

which is expected since Case 3 includes a linear $\beta_0(t)$. However, SRP$_{\mathcal{C}}$ is not far behind SRP$_{\mathcal{L}}$ in Case 3 as the smoothness of $\hat{\beta}(t)$ is flexibly controlled by the automatically chosen penalty.

4. Data applications

In this section, our methodology is applied to country fertility data, Mexico city pollution data, stock volatility series and sunspot number data. The data can be obtained from the Human Fertility Database (https://www.humanfertility.org/), the R package `aire.zmvm`, the Wharton Research Data Services (https://wrds-web.wharton.upenn.edu/wrds/) and the Base R datasets available from CRAN, respectively.

4.1. Country fertility rate data

Forecasting future fertility rates has a great impact on governments in planning children's service and education. We use fertility rates at age 20, recorded for 36 years from 1974 to 2009 for 31 countries around the world. As shown in Figure 5, the fertility rates at age 20 show an overall decreasing trend in all countries and although it is not illustrated in this paper, similar patterns are observed at ages 21–26, while fertility rates at ages 30–39 have obvious increasing trends in recent years from 1990 onwards, which reflects the phenomenon of more women deferring childbirth to a later age.

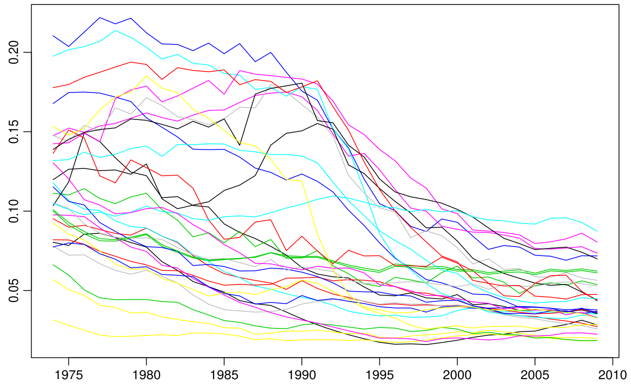

Fig 5. *The fertility rates at age 20 from 1974 to 2009 for 31 countries.*

The final observation recorded in 2009 is predicted from the past observations from 1974 to 2008. To compare the prediction power of the new model with competitors, we split the whole dataset into a training sample of size $n_1 = 26$ and a test set of size $n_2 = 5$ randomly 100 times and compute the mean, median and standard deviation of the 100 mean-square prediction errors defined in (10). In the training set, the B-spline expansion with dimension $L = 9$ is used for $\mathrm{SRP}_{\mathcal{C}}$, $\mathrm{SRP}_{\mathcal{L}}$ and FLR. As found in Table 3, MLR, $\mathrm{SRP}_{\mathcal{C}}$ and $\mathrm{SRP}_{\mathcal{L}}$ lead to similar performance in prediction, which is better than that of the nonparametric methods (MPDP, NP), the full functional model (FLR), the full scalar setting (RIDGE) and FLiRTI.

Table 3

The mean, median and standard deviation of 100 MSPE's ($\times 10^6$) defined in formula (10) for all methods described in Section 3.1, for the case study in Section 4.1. Bold: methods with the three lowest MSPE's.

	MLR	FLR	FLiRTI	$\mathrm{SRP}_{\mathcal{L}}$	$\mathrm{SRP}_{\mathcal{C}}$	MPDP	NP	RIDGE
mean	**3.36**	12.60	5.99	**3.45**	**3.73**	5.38	139.55	7.73
median	**2.98**	9.15	3.95	**3.12**	**3.28**	3.65	118.48	4.97
sd	2.00	10.70	6.13	1.94	2.32	5.33	114.58	7.39

As shown in Figure 6, $\hat{q}_{\cdot} = 1, 2$ are the most frequently selected as the optimal size of scalar variables for MLR, $\mathrm{SRP}_{\mathcal{L}}$ and $\mathrm{SRP}_{\mathcal{C}}$. Although MLR and

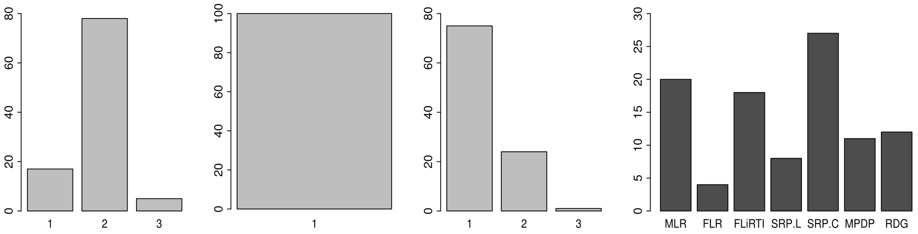

Fig 6. *Barplots of the* 100 $\hat{q}_1$ *for MLR (first), 100* $\hat{q}_2$ *for SRP*$_{\mathcal{L}}$ *(second) and 100* $\hat{q}$ *for SRP*$_{\mathcal{C}}$ *(third) estimated by minimising* $\{SIC(q), 1 \leq q \leq 4\}$ *in formula* (6) *and the frequency barplot of the best-performing method (with the lowest MSPE) out of the 100 samples (fourth) for the case study in Section 4.1.*

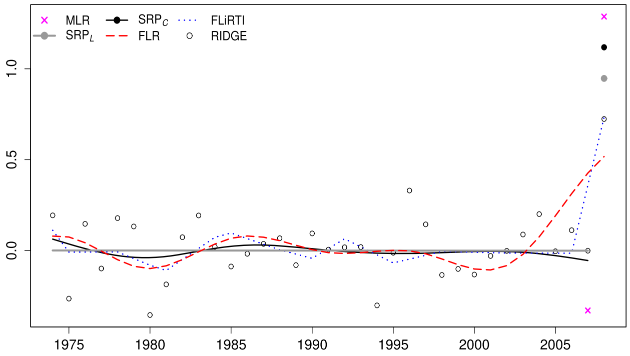

Fig 7. *A randomly selected estimated regression coefficients of the six parametric methods (MLR, SRP*$_{\mathcal{L}}$*, SRP*$_{\mathcal{C}}$*, FLR, FLiRTI, RIDGE) for predicting fertility rates at age 20 in 2009 from the past observations (1974-2008).*

SRP$_{\mathcal{L}}$ seem to be slightly better than SRP$_{\mathcal{C}}$ in prediction in Table 3, Figure 6 shows that SRP$_{\mathcal{C}}$ is the most frequently selected as the best-performing method in terms of MSPE from 100 samples. In Figure 7, the functional estimators $\hat{\beta}(t)$ for FLR and FLiRTI and the discrete ones for RIDGE live in the whole interval $t \in [t_0, t_{T-1}]$ while SRP$_{\mathcal{C}}$, MLR and SRP$_{\mathcal{L}}$ assign the corresponding subintervals for $\hat{\boldsymbol{\alpha}}$ with the optimally chosen $\hat{q} = 1$, $\hat{q}_1 = 2$ and $\hat{q}_2 = 1$ (respectively). The estimated curves for FLR and FLiRTI and the estimated coefficients for RIDGE appear to be relatively oscillatory over the entire interval under a fixed smoothness while the smoothness of the SRP estimators varies as dictated by their design. Interestingly, all parametric methods give a large size of the regression coefficient at year 2008, which contrasts with the coefficients for years 1974–2007 which are close to zero. In a time series context, this indicates that the fertility rate in 2008 is more influential for predicting the fertility rate in 2009 than the older observations are.

4.2. Nitrogen oxides in Mexico City

We use the daily curves of hourly average nitrogen oxides level in Mexico City, recorded at the Pedregal station in 2016. As shown in Figure 8, daily curves contain 24 observations each and have similar patterns including two peaks around hours 9 and 21. The final observation recorded at hour 24 is predicted from the past observations indexed 1 to 23. We split the whole dataset into a training sample of size $n_1 = 161$ and a test set of size $n_2 = 86$ randomly 100 times and compute the mean, median and standard deviation of the 100 mean-square prediction errors defined in (10). In the training set, the B-spline expansion with dimension $L = 9$ is used for SRP$_\mathcal{C}$, SRP$_\mathcal{L}$ and FLR. As found in Table 4 and Figure 9, SRP$_\mathcal{C}$ gives the best prediction among all methods and is also the most frequently selected as the best-performing one from the 100 samples in terms of MSPE. As shown in Figure 9, $\hat{q} = 3$ is the most frequently selected as the optimal size of scalar variables for SRP$_\mathcal{C}$ while $\hat{q}_{.} = 2$ is so for MLR and SRP$_\mathcal{L}$.

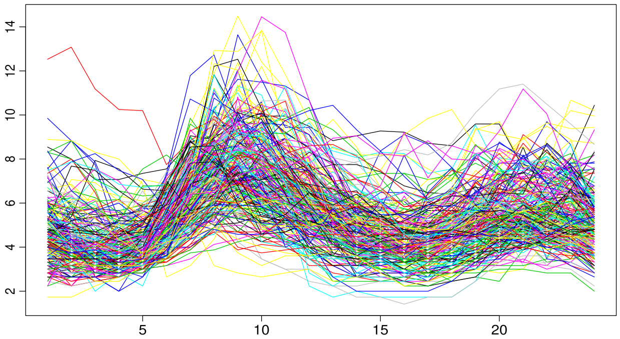

FIG 8. *The daily curves of hourly average nitrogen oxides (parts per billion) at the Pedregal station in Mexico City in 2016.*

In Figure 10, it is interesting to observe that the smooth portion of the SRP parameter vector appears to be non-trivially different from zero, which, together with the fact that the SRP model outperforms its competitors in the forecasting exercise reported above, provides evidence for the existence and impact of the long-term temporal dependence in this dataset. It is also apparent that all the methods attempt to fit a particularly large-size regression coefficient at hour 23. The SRP$_\mathcal{C}$ curve detects a change at hour 20, where it experiences a seemingly non-trivial drop. It would be difficult for us to conclude that this drop is merely caused by a boundary effect as the RIDGE solution (in which there are no boundary effects to speak of) also experiences a dip at that point. In the same manner, the sudden increase observed in the FLR curve at hour 23 does not appear to be a mere boundary effect, but it also reflects this method's own effort to fit the influential predictor under its own smoothness constraints. The results in Table 4 show that it is useful to apply two different regularisations, as done in SRP$_\mathcal{C}$, depending on the perceived importance of predictors, rather than estimating the regression coefficients under an unvarying regularisation, as done in RIDGE.

Table 4
The mean, median and standard deviation of 100 MSPE's ($\times 10^2$) defined in formula (10) for all methods described in Section 3.1, for the case study in Section 4.2. Bold: methods with the three lowest MSPE's.

	MLR	FLR	FLiRTI	SRP$_{\mathcal{L}}$	SRP$_{\mathcal{C}}$	MPDP	NP	RIDGE
mean	75.50	86.44	**73.88**	75.41	**72.35**	74.92	126.09	**74.42**
median	75.38	85.16	**74.04**	75.13	**71.84**	74.23	126.99	**73.41**
sd	12.92	14.03	12.96	14.10	13.18	13.13	26.63	12.94

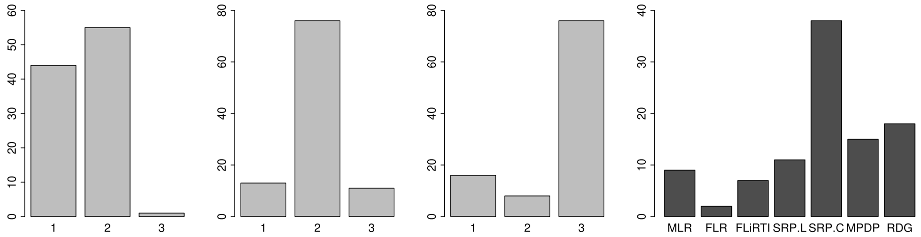

Fig 9. *Barplots of the 100 $\hat{q}_1$ for MLR (first), 100 $\hat{q}_2$ for SRP$_{\mathcal{L}}$ (second) and 100 $\hat{q}$ for SRP$_{\mathcal{C}}$ (third) estimated by minimising $\{SIC(q), 1 \leq q \leq 3\}$ in formula (6) and the frequency barplot of the best-performing method (with the lowest MSPE) out of the 100 samples (fourth) for the case study in Section 4.2.*

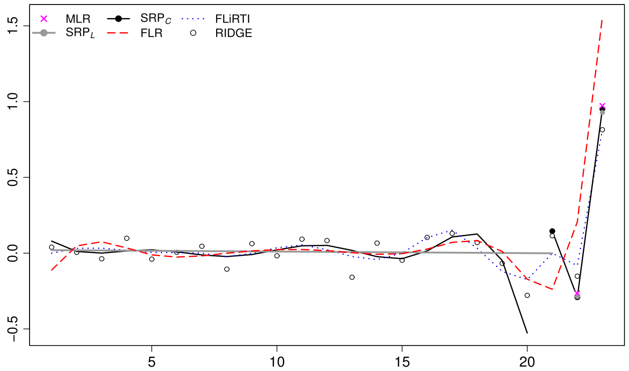

Fig 10. *A randomly selected estimated regression coefficients of the six parametric methods (MLR, SRP$_{\mathcal{L}}$, SRP$_{\mathcal{C}}$, FLR, FLiRTI, RIDGE) for predicting the average of nitrogen oxides level at hour 24.*

4.3. High frequency volatility series

In financial data analysis, modelling high-frequency volatility has attracted much attention in recent years. Especially, in the functional framework, nonparametric methods have been extensively studied (Bandi and Phillips, 2003; Reno, 2008; Kristensen, 2010). Müller et al. (2011) emphasise the random nature of volatility functions under the assumption that the repeated realisations of the volatility trajectories come from a suitable functional volatility process. Our interest is also in the random nature of functional observations rather than in modelling potential dependencies between curves, therefore, as in Müller et al. (2011), we view the daily curves as i.i.d. random functions. We aim to predict the latest point of the curves from the past observations.

Specifically, our methodology is applied to the prediction of the Disney stock volatility where the raw observations contain $n = 248$ trading days available from January 2, 2013 to December 30, 2013 and each curve has 395 grid points of closing prices recorded every 1 minute. The volatility trajectories are obtained from the return series in the same way as in Müller et al. (2011), however we retain the roughness of volatility trajectories by using natural cubic splines as in (4) rather than smoothing them. This is important as volatility is not observable but typically estimated to be oscillatory, thus an extra smoothing step can possibly cause the loss of important information as stated in Kneip et al. (2016).

We split the dataset into a training and a test set of size $n_1 = n_2 = 124$ randomly 100 times and in the training set, the B-spline expansion with dimension $L = 35$ is used for SRP$_\mathcal{C}$, SRP$_\mathcal{L}$ and FLR. Figure 11 shows that $\hat{q}_1 = 3$ is the most frequently chosen for MLR while $\hat{q}_2 = 1$ and $\hat{q} = 1$ are the most frequently selected for SRP$_\mathcal{L}$ and SRP$_\mathcal{C}$, respectively.

Similar to the previous examples in Sections 4.1 and 4.2, Figure 12 shows that all the parametric methods reflect the 'fading memory' of the time series by assigning a large-size regression coefficient for observations located close to the closing volatilities, which contrasts with the coefficients for intervals positioned far from the closing volatility. As found in Table 5 and Figure 11, SRP$_\mathcal{C}$ leads to the best prediction among all methods and is also the most frequently selected as the best-performing one in terms of MSPE from 100 samples.

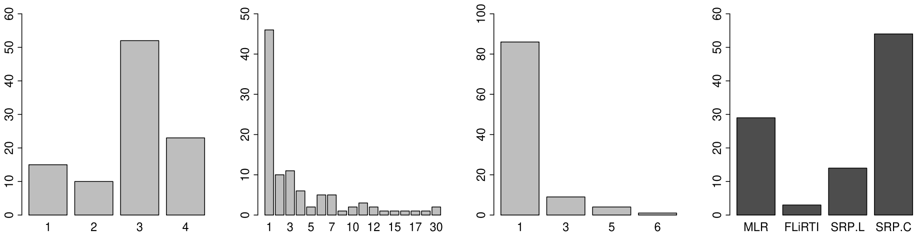

Fig 11. *Barplots of the* 100 $\hat{q}_1$ *for MLR (first), 100* $\hat{q}_2$ *for SRP*$_\mathcal{L}$ *(second) and 100* $\hat{q}$ *for SRP*$_\mathcal{C}$ *(third) estimated by minimising* $\{SIC(q), 1 \leq q \leq 30\}$ *in formula* (6) *and the frequency barplot of the best-performing method (with the lowest MSPE) out of the 100 samples (fourth) for the case study in Section 4.3.*

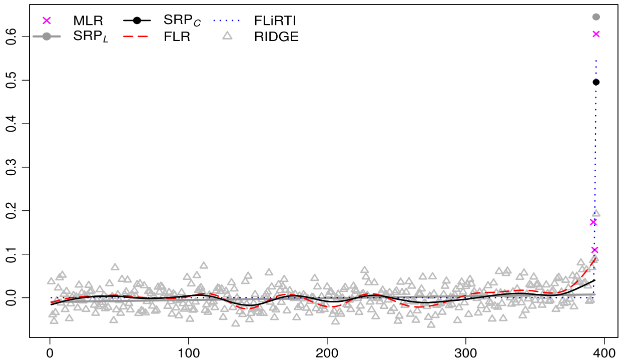

Fig 12. *A randomly selected estimated regression coefficients of the six parametric methods (MLR, SRP$_\mathcal{L}$, SRP$_\mathcal{C}$, FLR, FLiRTI, RIDGE) for predicting closing volatility of the Disney stock data from January to December in 2013.*

Table 5

The mean, median and standard deviation of 100 MSPE's defined in formula (10) for all methods described in Section 3.1, for the case study from Section 4.3. Bold: methods with the three lowest MSPE's.

	MLR	FLR	FLiRTI	SRP$_\mathcal{L}$	SRP$_\mathcal{C}$	MPDP	NP	RIDGE
mean	**2.88**	4.10	3.13	**2.96**	**2.78**	3.02	6.29	4.34
median	**2.80**	4.05	3.08	2.91	**2.72**	**2.77**	6.18	4.29
sd	0.56	0.58	0.68	0.56	0.51	1.52	0.71	0.48

4.4. Monthly numbers of sunspots

In this section, we demonstrate the usefulness the SRP framework in univariate time series modelling, as an alternative to the AR model, which is often used in time series forecasting. The SRP model is similar to the AR model in that they both specify the fading memory structure of the time series under linear dependence of the output variable on its own previous values. In practice, the AR(p) model is usually fitted with a small p for simplicity, interpretability and better forecasting performance, however it may fail in the presence of longer-range effects. In this case, the SRP model can also be used for the forecasting of a univariate time series, where it becomes an autoregressive (AR) model with a large order (e.g. AR(T) in (2) with a fixed T) under the smooth-rough regularisation. The middle plot of Figure 13 shows that the monthly sunspot number series may need large-order autoregression (even up to or exceeding order 100), in which case it may be advantageous to use the SRP model over plain AR modelling.

The sunspot number data contains 3177 observations available from 1749 to 2013 and we perform a square root transformation to the raw data. We split the whole dataset into a training sample of size $n_1 = 2223$ and a test set of size $n_2 = 954$ and create the data matrix for each set via a moving window with a prespecified number $T + 1 = 151$ of discrete points in one curve (150 for covariates and 1 for the response variable), i.e. from the

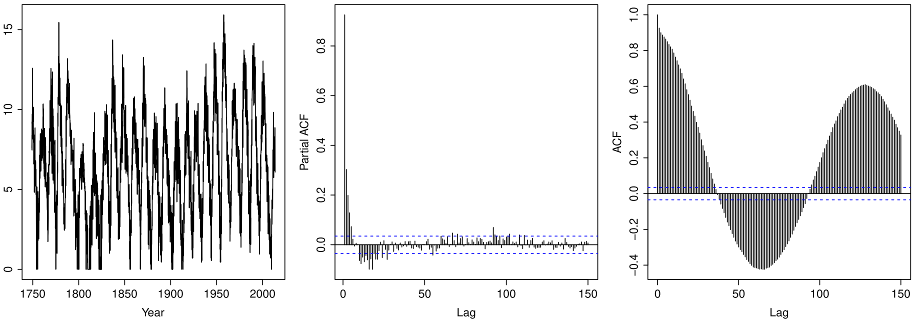

FIG 13. *Square-rooted monthly numbers of sunspots from 1749 to 2013 (left), its partial autocorrelation function with maximum lag=150 (middle) and the autocorrelation function with maximum lag=150 (right).*

univariate time series $(x_1, x_2, \ldots, x_{n_1})$ in the training sample, we create 2073 curves, $X_1(t) = (x_1, x_2, \ldots, x_{151})$, $X_2(t) = (x_2, x_3, \ldots, x_{152})$, …, $X_{n_1-151+1} = (x_{n_1-150}, x_{n_1-149}, \ldots, x_{n_1})$. In the same way, we create 804 curves for the test sample. In each curve, we use the last points as the response variable and the covariates are the remaining 150 observations. Due to the temporal dependence in the entire dataset, we do not randomly repeat the construction of the training and test sets.

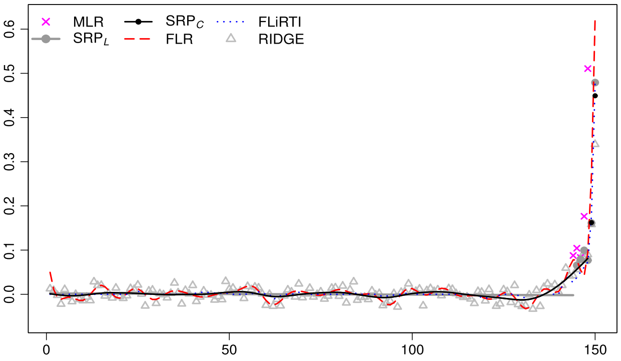

FIG 14. *Estimated regression coefficients of the six parametric methods (MLR, SRP$_{\mathcal{L}}$, SRP$_{\mathcal{C}}$, FLR, FLiRTI, RIDGE) for predicting the sunspot number of next month from past 150 months of sunspot number.*

From the training set, with $L = 35$, the optimal change-point index parameter for MLR, SRP$_{\mathcal{L}}$ and SRP$_{\mathcal{C}}$ are chosen as $\hat{q}_1 = 5$, $\hat{q}_2 = 6$, $\hat{q} = 2$ (respectively) from $\{q : 1 \leq q \leq 15\}$ as shown in Figure 14. As the optimal size $\hat{q}_1 = 5$ for MLR is obtained by minimising the SIC criterion, the estimated regression coefficients are very close to that of the AR(5) model and the significance of the first five lags is already revealed in the partial autocorrelation function in Figure 13. In Figure 14, the FLR and RIDGE estimators appear to be relatively oscillatory over the entire interval, while the estimators for FLiRTI and $\mathrm{SRP}_{\mathcal{C}}$ are relatively smoother. We also obtain the OLS (ordinary least squares) estimator which is slightly more fluctuating than RIDGE, but is not included in Figure 14. As is apparent from Table 6, our approach shows an improvement in prediction compared to the other methods. From this example, $\mathrm{SRP}_{\mathcal{C}}$ appears to be a useful substitution for a classical AR(p) model with a small p, especially when the memory of a time series is relatively long.

Table 6

MSPE ($\times 10^2$) defined in formula (10) for all parametric methods described in Section 3.1 and OLS, for the case study from Section 4.4. Bold: methods with the three lowest MSPE's.

	MLR	FLR	FLiRTI	$\mathrm{SRP}_{\mathcal{L}}$	$\mathrm{SRP}_{\mathcal{C}}$	RIDGE	OLS
MSPE	11.67	12.09	12.59	**11.09**	**10.72**	11.17	**11.11**

5. Theoretical results

In this section, we assume that the SRP model in (3) is correct and explore the asymptotic behaviour of $\hat{q}$, the estimator of the change-point index q_0. There is a one-to-one correspondence between q and t_{T-q}, so we will be interchangeably considering $\hat{q}$ and $t_{T-\hat{q}}$. We denote the true values of scalars $\boldsymbol{\alpha}$ and function β by $(\boldsymbol{\alpha}_0, \beta_0)$ and assume the following conditions.

Assumption 1. *$\beta_0(t)$ is continuous on $t \in [t_0, t_{T-q_0-1}]$ and $\boldsymbol{\alpha}_0$ is composed of the finite number of scalars $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \ldots, \alpha_{0,q_0})^T$ on $t \in [t_{T-q_0}, t_{T-1}]$.*

Assumption 2. *The true change-point $t_{T-q_0} \in (t_0, t_{T-1}]$ is where the change of smoothness occurs in the sequence of true regression parameters. When $q_0 > 1$, taking q_1 such that $1 \leq q_1 < q_0$, for any $q \in [q_1, q_0)$, there exist $\delta_1, \delta_2, \delta_3 > 0$ such that (a) $\inf_{1 \leq j \leq q} |\alpha_{0,j} - \hat{\alpha}_j^q| > \delta_1$, (b) $\inf_{q_0 < j \leq T} |\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j})| > \delta_2$ and (c) $\inf_{q < j \leq q_0} |\alpha_{0,j} - \hat{\beta}^q(t_{T-j})| > \delta_3$.*

As mentioned in the discussion of the shape of the function $M(q)$ in Section 2.1, Assumption 2 quantifies the non-convergences occurring when $q < q_0$. The next two assumptions list the converging components of $M(q)$ when $q \geq q_0$. Our Assumptions 3 and 4 are similar to the assumptions made on estimated regression coefficients in Hall and Hooker (2016).

Assumption 3. *Taking q_2 such that $1 \leq q_0 < q_2 < T$,*

$$(a) \quad \sup_{q_0 \leq q \leq q_2} \frac{1}{n} \sum_{i=1}^{n} \Big[\sum_{1 \leq j \leq q_0} (\alpha_{0,j} - \hat{\alpha}_j^q) \tilde{X}_i(t_{T-j}) \Big]^2 = O_p(n^{-1}),$$

$$(b) \quad \sup_{q_0 \leq q \leq q_2} \frac{1}{n} \sum_{i=1}^{n} \Big[\sum_{q < j \leq T} (\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j})) \tilde{X}_i(t_{T-j}) \Big]^2 = O_p(n^{-1}),$$

$$(c) \quad \sup_{q_0 < q \leq q_2} \frac{1}{n} \sum_{i=1}^{n} \Big[\sum_{q_0 < j \leq q} (\beta_0(t_{T-j}) - \hat{\alpha}_j^q) \tilde{X}_i(t_{T-j}) \Big]^2 = O_p(n^{-1}).$$

Assumption 4. *When q_2 is as in Assumption 3,*

$$(a) \quad \sup_{q_0 \le q \le q_2} \left| \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon}) \sum_{1 \le j \le q_0} (\alpha_{0,j} - \hat{\alpha}_j^q) \tilde{X}_i(t_{T-j}) \right| = O_p(n^{-1}),$$

$$(b) \quad \sup_{q_0 \le q \le q_2} \left| \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon}) \sum_{q < j \le T} (\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j})) \tilde{X}_i(t_{T-j}) \right| = O_p(n^{-1}),$$

$$(c) \quad \sup_{q_0 < q \le q_2} \left| \frac{1}{n} \sum_{i=1}^{n} (\varepsilon_i - \bar{\varepsilon}) \sum_{q_0 < j \le q} (\beta_0(t_{T-j}) - \hat{\alpha}_j^q) \tilde{X}_i(t_{T-j}) \right| = O_p(n^{-1}).$$

Assumption 5. *The independent and identically distributed errors ε_i are independent of the predictors. We further assume $E(\boldsymbol{X}^T\boldsymbol{X}) + E(\varepsilon^2) < \infty$ with $E(\varepsilon) = 0$, where $\boldsymbol{X}_{n\times T} = (X(t_0), X(t_1), \ldots, X(t_{T-1}))$.*

Assumption 6. *Writing the singular value decomposition of the covariance matrix of $\boldsymbol{X}$ as $K_{(k_1,k_2)} = cov(X(t_{k_1}), X(t_{k_2})) = \sum_{j=1}^{T} v_j \boldsymbol{\psi}_j \boldsymbol{\psi}_j^T$ where $v_1 \ge v_2 \cdots > 0$ are eigenvalues, and $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ are the corresponding eigenvectors, we assume that the eigenvalues decay sufficiently fast so that the condition $\sum_{j=1}^{T} v_j^{1/2} \|\boldsymbol{\psi}_j\|_\infty < \infty$ holds.*

We are now ready to state our main result.

Theorem 5.1. *If $\hat{q}$ is any value of q which minimises (6) on the interval $[q_1, q_2]$ when q_1 and q_2 are chosen to satisfy $1 \le q_1 < q_0 < q_2 < T$, then under the Assumptions 1–6, we have $P(\hat{q} = q_0) \to 1$ as $n \to \infty$.*

Technical proof of Theorem 5.1 is available in Appendix A. We end this section with further brief justification of our assumptions by comparing them to similar assumptions made in some related recent works.

The B-spline expansion employed in this article can be replaced with other bases, for instance the set of eigenfunctions of the covariance operator of $X(t)$. Cai and Hall (2006) investigate this case and derive the parametric rates with this methodology. Hall and Hooker (2016) mention that the methods used by Cai and Hall (2006) can give the rate of convergence of $\beta^q(t)$ in $(n^{-1/2}, n^0)$ for Assumption 3-(b) under appropriate smoothness conditions for $\beta(t)$, $X(t)$ and the covariance function measured by the spacing of the eigenvalues in a fully functional setting (that is, when $q = 0$ in our case). Similarly, Crambes et al. (2009) derive the rate of convergence for the general spline classes which is comparable to that of Cai and Hall (2006), under the usual smoothness assumptions on $\beta(t)$ and $X(t)$ defined by the continuity of its derivatives. The methods used in Crambes et al. (2009) can give the rate in Assumption 3-(b) under appropriate smoothness conditions for $\beta(t)$ and $X(t)$ in a full functional setting. Since our model contains scalar covariates and has the ridge type penalty in (8), we postulate the same or slightly slower rates, which are also supported by our numerical experience.

6. Discussion

The SRP model represents a compromise between a completely unregularised and a completely regularised linear model in that it keeps all the effects as non-zero but partitions them into two classes of regularity. This makes it a useful alternative to sparsity-based approaches as retaining the smooth non-zero regression parameter can be beneficial for prediction, as this paper demonstrates.

The SRP approach can in principle be applied in any context in which potential regressors have been pre-ordered in terms of their importance as is the case in the time series setting studied in this paper.

Appendix A: Technical proofs

The proof of Theorem 5.1 in Section 5 is presented. The preparatory lemma is developed first and the main part of the proof is presented in Section A.1.

Lemma A.1. *Let $1 \leq q_1 < q_0$ as in Assumption 2. If Assumptions 1, 2, 5 and 6 hold, then uniformly in $q \in [q_1, q_0)$,*

$$(a) \quad \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})\left\{\sum_{j=1}^{q}(\alpha_{0,j} - \hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\} = O_p(n^{-1/2}|q|),$$

$$(b) \quad \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})\left\{\sum_{j=q_0+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}$$
$$= O_p(n^{-1/2}|T - q_0|),$$

$$(c) \quad \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})\left\{\sum_{j=q+1}^{q_0}(\alpha_{0,j} - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\} = O_p(n^{-1/2}|q_0 - q|).$$

Our Lemma A.1 is similar to the Lemma in the recent work of Hall and Hooker (2016) who study the consistency of truncation point in functional linear regression with one functional predictor. The proof of Lemma A.1 can be simply obtained by following the methods used in Hall and Hooker (2016) and by having a discrete version of it, i.e. replacing a curve with a vector, under our assumptions.

A.1. Proof of Theorem 5.1

Let q_1 and q_2 as in Assumptions 2 and 3, respectively. Since $X_i(t_T) = \mu + \sum_{j=1}^{q_0}\alpha_{0,j}\{X_i(t_{T-j}) - EX(t_{T-j})\} + \sum_{j=q_0+1}^{T}\beta_0(t_{T-j})\{X_i(t_{T-j}) - EX(t_{T-j})\} + \varepsilon_i$, we have $X_i(t_T) - \bar{X}(t_T) = \sum_{j=1}^{q_0}\alpha_{0,j}\tilde{X}_i(t_{T-j}) + \sum_{j=q_0+1}^{T}\beta_0(t_{T-j})\tilde{X}_i(t_{T-j}) + (\varepsilon_i - \bar{\varepsilon})$, thus $M(q)$ defined in (7) is expanded as

$$M(q) = \frac{1}{n}\sum_{i=1}^{n}\left[X_i(t_T) - \hat{\mu} - \sum_{j=1}^{q}\hat{\alpha}_j^q X_i(t_{T-j}) - \sum_{j=q+1}^{T}\hat{\beta}^q(t_{T-j})X_i(t_{T-j})\right]^2$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\left[X_i(t_T) - \bar{X}(t_T) - \sum_{j=1}^{q}\hat{\alpha}_j^q \tilde{X}_i(t_{T-j}) - \sum_{j=q+1}^{T}\hat{\beta}^q(t_{T-j})\tilde{X}_i(t_{T-j})\right]^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{j=1}^{q_0}\alpha_{0,j}\tilde{X}_i(t_{T-j}) + \sum_{j=q_0+1}^{T}\beta_0(t_{T-j})\tilde{X}_i(t_{T-j}) - \sum_{j=1}^{q}\hat{\alpha}_j^q \tilde{X}_i(t_{T-j})\right. \\
&\qquad \left. - \sum_{j=q+1}^{T}\hat{\beta}^q(t_{T-j})\tilde{X}_i(t_{T-j}) + (\varepsilon_i - \bar{\varepsilon})\right]^2,
\end{aligned}
$$

where $q \in [q_1, q_2]$. $M(q)$ has a different form for three cases: 1) $q > q_0$, 2) $q < q_0$ and 3) $q = q_0$. Firstly, if $q > q_0$, for $q \in (q_0, q_2]$, we have

$$
\begin{aligned}
M(q) = \frac{1}{n}\sum_{i=1}^{n}&\left[\sum_{j=1}^{q_0}(\alpha_{0,j} - \hat{\alpha}_j^q)\tilde{X}_i(t_{T-j}) + \sum_{j=q+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right. \\
&\left. + \sum_{j=q_0+1}^{q}(\beta_0(t_{T-j}) - \hat{\alpha}_j^q)\tilde{X}_i(t_{T-j}) + (\varepsilon_i - \bar{\varepsilon})\right]^2. \qquad (11)
\end{aligned}
$$

If $q < q_0$, for $q \in [q_1, q_0)$,

$$
\begin{aligned}
M(q) = \frac{1}{n}\sum_{i=1}^{n}&\left[\sum_{j=1}^{q}(\alpha_{0,j} - \hat{\alpha}_j^q)\tilde{X}_i(t_{T-j}) + \sum_{j=q_0+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right. \\
&\left. + \sum_{j=q+1}^{q_0}(\alpha_{0,j} - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j}) + (\varepsilon_i - \bar{\varepsilon})\right]^2. \qquad (12)
\end{aligned}
$$

Lastly, when $q = q_0$,

$$
\begin{aligned}
M(q) = \frac{1}{n}\sum_{i=1}^{n}&\left[\sum_{j=1}^{q_0}(\alpha_{0,j} - \hat{\alpha}_j^{q_0})\tilde{X}_i(t_{T-j}) + \sum_{j=q_0+1}^{T}(\beta_0(t_{T-j})\right. \\
&\left. - \hat{\beta}^{q_0}(t_{T-j}))\tilde{X}_i(t_{T-j}) + (\varepsilon_i - \bar{\varepsilon})\right]^2. \qquad (13)
\end{aligned}
$$

A.1.1. Convergence rates of $M(q)$

Now we explore the behaviour of $M(q)$. For the first case, 1) $q > q_0$, under Assumptions 3 and 4, (11) simplifies to

$$
\begin{aligned}
M(q) =&\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{q_0}(\alpha_{0,j} - \hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=q+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}^2
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=q_0+1}^{q}(\beta_0(t_{T-j})-\hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}^2\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=1}^{q_0}(\alpha_{0,j}-\hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=q+1}^{T}(\beta_0(t_{T-j})-\hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=q_0+1}^{q}(\beta_0(t_{T-j})-\hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}+\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2\\
=&O_p(1/n)+\mathbf{V}, \qquad (14)
\end{aligned}
$$

uniformly in $q\in(q_0,q_2]$, where $\mathbf{V}$ refers to the error term which does not depend on q. In the second case, 2) $q<q_0$, using Lemma A.1, (12) simplifies to

$$
\begin{aligned}
M(q)=&\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{q}(\alpha_{0,j}-\hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}^2\\
&+\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=q_0+1}^{T}(\beta_0(t_{T-j})-\hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}^2\\
&+\frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=q+1}^{q_0}(\alpha_{0,j}-\hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}^2\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=1}^{q}(\alpha_{0,j}-\hat{\alpha}_j^q)\tilde{X}_i(t_{T-j})\right\}\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=q_0+1}^{T}(\beta_0(t_{T-j})-\hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}\\
&+\frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})\left\{\sum_{j=q+1}^{q_0}(\alpha_{0,j}-\hat{\beta}^q(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}+\frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i-\bar{\varepsilon})^2\\
=&M_1(q)+M_2(q)+M_3(q)+O_p(n^{-1/2}|q|)+O_p(n^{-1/2}|T-q_0|)\\
&+O_p(n^{-1/2}|q_0-q|)+\mathbf{V}, \qquad (15)
\end{aligned}
$$

uniformly in $q\in[q_1,q_0)$, where

$$M_1(q)=\sum_{1\le k_1,k_2\le q}\{\alpha_{0,k_1}-\hat{\alpha}_{k_1}^q\}\{\alpha_{0,k_2}-\hat{\alpha}_{k_2}^q\}\hat{K}_{(k_1,k_2)}, \qquad (16)$$

$$M_2(q)=\sum_{q_0+1\le k_1,k_2\le T}\{\beta_0(t_{T-k_1})-\hat{\beta}^q(t_{T-k_1})\}\{\beta_0(t_{T-k_2})-\hat{\beta}^q(t_{T-k_2})\}\hat{K}_{(k_1,k_2)}, \qquad (17)$$

$$M_3(q)=\sum_{q+1\le k_1,k_2\le q_0}\{\alpha_{0,k_1}-\hat{\beta}^q(t_{T-k_1})\}\{\alpha_{0,k_2}-\hat{\beta}^q(t_{T-k_2})\}\hat{K}_{(k_1,k_2)}, \qquad (18)$$

and $\hat{K}_{(k_1,k_2)}$ is the empirical version of K defined in Assumption 6. Now we define

$$\kappa_3(q) = \sum_{q+1\le k_1,k_2\le q_0} \{\alpha_{0,k_1} - \hat{\beta}^q(t_{T-k_1})\}\{\alpha_{0,k_2} - \hat{\beta}^q(t_{T-k_2})\}K_{(k_1,k_2)},$$

to deal with $M_3(q)$. If we show that, for any bounded vector $\boldsymbol{z} = (z_0, ..., z_{T-1})^T$,

$$\sup_{u,v\in[0,T-1]} \left| \sum_{k_1=u}^{v} \sum_{k_2=u}^{v} z_{k_1} z_{k_2} \left\{ \hat{K}_{(k_1,k_2)} - K_{(k_1,k_2)} \right\} \right| \to 0 \quad \text{in probability}, \tag{19}$$

then we can argue that $\sup_{q\in[q_1,q_0)} \left|M_3(q) - \kappa_3(q)\right| \to 0$ in probability by taking a vector $\boldsymbol{z}$ with its elements $z_j = (\alpha_{0,j} - \hat{\beta}^q(t_{T-j}))$ if $q+1 \le j \le q_0$ and $z_j = 0$ otherwise. We can simply derive (19) under Assumption 5 and the appropriate inequalities as in Hall and Hooker (2016). Similarly, $\kappa_1(q)$ and $\kappa_2(q)$ can be defined for $M_1(q)$ and $M_2(q)$, respectively and following from Assumption 2, $\kappa_1(q)$, $\kappa_2(q)$ and $\kappa_3(q)$ are strictly positive whenever $q < q_0$.

Lastly, when $q = q_0$, under Assumptions 3 and 4, (13) can be simplified as

$$\begin{aligned} M(q) =& \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=1}^{q_0}(\alpha_{0,j} - \hat{\alpha}_j^{q_0})\tilde{X}_i(t_{T-j})\right\}^2 \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left\{\sum_{j=q_0+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^{q_0}(t_{T-j}))\tilde{X}_i(t_{T-j})\right\}^2 \\ &+ \frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})\left\{\sum_{j=1}^{q_0}(\alpha_{0,j} - \hat{\alpha}_j^{q_0})\tilde{X}_i(t_{T-j})\right\} \\ &+ \frac{2}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})\left\{\sum_{j=q_0+1}^{T}(\beta_0(t_{T-j}) - \hat{\beta}^{q_0}(t_{T-j}))\tilde{X}_i(t_{T-j})\right\} \\ &+ \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 \\ =& O_p(1/n) + \mathbf{V}. \end{aligned} \tag{20}$$

A.1.2. Expansions of $SIC(q)$ based on $M(q)$

To prove Theorem 5.1, it suffices to show that $SIC(q) - SIC(q_0)$ is positive for both cases 1) $q > q_0$ and 2) $q < q_0$. If $q > q_0$, for $\epsilon > 0$,

$$\begin{aligned} SIC(q) - SIC(q_0) =& n \cdot \log\left(\frac{M(q)}{M(q_0)}\right) + (q - q_0)\cdot \log n \\ =& n \cdot \log\left(1 - \frac{M(q_0) - M(q)}{M(q_0)}\right) + (q - q_0)\cdot \log n \end{aligned}$$

$$\geq - n(1+\epsilon)\left(\frac{M(q_0)-M(q)}{M(q_0)}\right) + (q-q_0)\cdot \log n.$$

Since $M(q_0) - M(q) = O_p(1/n)$ for $q > q_0$ by (14) and (20), $SIC(q) - SIC(q_0)$ is guaranteed to be positive as $n \to \infty$.

Conversely, if $q < q_0$,

$$\begin{aligned} SIC(q) - SIC(q_0) =& n \cdot \log\left(\frac{M(q)}{M(q_0)}\right) + (q-q_0)\cdot \log n \\ \geq& n \cdot \log\left(\frac{M(q)}{M(q_0)}\right) - q_0 \cdot \log n. \end{aligned}$$

Since it can be simply shown that $\frac{M(q)}{M(q_0)} > 1 + \frac{1}{n}$ for $q > q_0$ from (15) and (20), $SIC(q) - SIC(q_0)$ is guaranteed to be positive as $n \to \infty$. Hence, we simply deduce that $P(\hat{q} = q_0) \to 1$ as $n \to \infty$.

Acknowledgements

We thank the associate editor and the referee for their comments and suggestions which led to an improved version of this paper. Piotr Fryzlewicz's work was supported by the Engineering and Physical Sciences Research Council grant No. EP/L014246/1.

References

Aneiros-Pérez, G. and Vieu, P. (2008). Nonparametric time series prediction: A semi-functional partial linear modeling. *Journal of Multivariate Analysis*, 99:834–857. MR2405094

Antoniadis, A., Paparoditis, E., and Sapatinas, T. (2006). A functional wavelet–kernel approach for time series prediction. *Journal of the Royal Statistical Society, Series B*, 68:837–857. MR2301297

Aue, A., Norinho, D. D., and Hörmann, S. (2015). On the prediction of stationary functional time series. *Journal of the American Statistical Association*, 110:378–392. MR3338510

Bandi, F. M. and Phillips, P. C. (2003). Fully nonparametric estimation of scalar diffusion models. *Econometrica*, 71:241–283. MR1956859

Bosq, D. (2000). *Linear Processes in Function Spaces*. New York: Springer-Verlag. MR1783138

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *The Annals of Statistics*, 34:2159–2179. MR2291496

Cardot, H., Crambes, C., Kneip, A., and Sarda, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis*, 51:4832–4848. MR2364543

Cardot, H., Ferraty, F., and Sarda, P. (2003). Spline estimators for the functional linear model. *Statistica Sinica*, 13:571–591. MR1997162

Chen, K., Delicado, P., and Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *Journal of the Royal Statistical Society, Series B*, 79:177–196. MR3597969

Crambes, C., Kneip, A., and Sarda, P. (2009). Smoothing splines estimators for functional linear regression. *The Annals of Statistics*, 37:35–72. MR2488344

Ferraty, F., Hall, P., and Vieu, P. (2010). Most-predictive design points for functional data predictors. *Biometrika*, 97:807–824. MR2746153

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics*, 17:545–564. MR1952697

Gabrys, R., Horváth, L., and Kokoszka, P. (2010). Tests for error correlation in the functional linear model. *Journal of the American Statistical Association*, 105:1113–1125. MR2752607

Goia, A. (2012). A functional linear model for time series prediction with exogenous variables. *Statistics and Probability Letters*, 82:1005–1011. MR2910049

Goia, A. and Vieu, P. (2015). A partitioned single functional index model. *Computational Statistics*, 30:673–692. MR3404355

Hall, P. and Hooker, G. (2016). Truncated linear models for functional data. *Journal of the Royal Statistical Society, Series B*, 78:637–653. MR3506796

Han, K., Müller, H.-G., and Park, B. U. (2017). Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions. *Preprint*. MR3706793

Horváth, L., Kokoszka, P., and Reeder, R. (2013). Estimation of the mean of functional time series and a two-sample problem. *Journal of the Royal Statistical Society, Series B*, 75:103–122. MR3008273

Horváth, L., Kokoszka, P., and Rice, G. (2014). Testing stationarity of functional time series. *Journal of Econometrics*, 179:66–82. MR3153649

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37:2083–2108. MR2543686

Ji, H. and Müller, H.-G. (2017). Optimal designs for longitudinal and functional data. *Journal of the Royal Statistical Society, Series B*, 79:859–876. MR3641411

Kneip, A., Poß, D., Sarda, P., et al. (2016). Functional linear regression with points of impact. *The Annals of Statistics*, 44:1–30. MR3449760

Kong, D., Xue, K., Yao, F., and Zhang, H. H. (2016). Partially functional linear regression in high dimensions. *Biometrika*, 103:147–159. MR3465827

Kristensen, D. (2010). Nonparametric filtering of the realized spot volatility: A kernel-based approach. *Econometric Theory*, 26:60–93. MR2587103

Lin, Z., Cao, J., Wang, L., and Wang, H. (2015). A smooth and locally sparse estimator for functional linear regression via functional scad penalty. *Preprint*. MR3640188

McKeague, I. W. and Sen, B. (2010). Fractals with point impact in functional linear regression. *The Annals of Statistics*, 38:2559–2586. MR2676898

Müller, H.-G., Sen, R., and Stadtmüller, U. (2011). Functional data analysis for volatility. *Journal of Econometrics*, 165:233–245. MR2846647

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

York: Springer-Verlag. MR2168993

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, 85:228–249. MR3686566

Reno, R. (2008). Nonparametric estimation of the diffusion coefficient of stochastic volatility models. *Econometric Theory*, 24:1174–1206. MR2440739

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757. MR1944261

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464. MR0468014

Shin, H. (2009). Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139:3405–3418. MR2549090

Shin, H. and Lee, M. H. (2012). On prediction rate in partial functional linear regression. *Journal of Multivariate Analysis*, 103:93–106. MR2823711

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. CRC press. MR2206355

Zhou, J. and Chen, M. (2012). Spline estimators for semi-functional linear model. *Statistics and Probability Letters*, 82:505–513. MR2887465

Zhou, J., Chen, Z., and Peng, Q. (2016). Polynomial spline estimation for partial functional linear regression models. *Computational Statistics*, 31:1107–1129. MR3528648

Zhou, J., Wang, N.-Y., and Wang, N. (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica*, 23:25–50. MR3076157

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in reproducing kernel hilbert spaces. *Journal of the Royal Statistical Society, Series B*, 76:581–603. MR3210729