SLANTS: Sequential Adaptive Nonlinear Modeling of Time Series

Qiuyi Han, Jie Ding, *Student Member, IEEE,* Edoardo M. Airoldi, and Vahid Tarokh, *Fellow, IEEE*

***Abstract*—We propose a method for adaptive nonlinear sequential modeling of time series data. Data is modeled as a nonlinear function of past values corrupted by noise, and the underlying non-linear function is assumed to be approximately expandable in a spline basis. We cast the modeling of data as finding a good fit representation in the linear span of multi-dimensional spline basis, and use a variant of l_1-penalty regularization in order to reduce the dimensionality of representation. Using adaptive filtering techniques, we design our online algorithm to automatically tune the underlying parameters based on the minimization of the regularized sequential prediction error. We demonstrate the generality and flexibility of the proposed approach on both synthetic and real-world datasets. Moreover, we analytically investigate the performance of our algorithm by obtaining both bounds on prediction errors and consistency in variable selection.**

***Index Terms*—Adaptive Filtering, Data prediction, Group LASSO, Nonlinearity, Sequential Modeling, SLANTS, Spline, Time Series.**

I. Introduction

SEQUENTIALLY observed multi-dimensional time series are emerging in various applications. In most of these applications, modeling nonlinear functional inter-dependency between present and past data is crucial for both representation and prediction. This is a challenging problem, especially when fast online implementation, adaptivity to new data generating processes, and ability to handle high dimensions need to be simultaneously taken into account in nonlinear modeling. For example, environmental science combines high dimensional weather signals for real time prediction [1]. In epidemics, huge amount of online search data is used to form fast prediction of influenza epidemics [2]. In finance, algorithmic traders demand adaptive models to accommodate a fast changing stock market. In robot autonomy, there is the challenge of learning the high dimensional movement systems [3]. These tasks usually take high dimensional input signals which may contain a large number of irrelevant signals. In all these applications, methods to remove redundant signals and learn the nonlinear model with low computational complexity are well sought after. This motivates our work in this paper, where we propose an approach to sequential nonlinear adaptive modeling of potentially high dimensional time series.

Inference of nonlinear models has been a notoriously difficult problem, especially for large dimensional data [3]–[5]. In low dimensional settings, there have been remarkable parametric and nonparametric nonlinear time series models that have been applied successfully to data from various domains. Examples include threshold models [6], generalized autoregressive conditional hetero-scedasticity models [7], multivariate adaptive regression splines (MARS) [4], generalized additive models [8], functional coefficient regression models [9], etc. However, some of these methods may suffer from prohibitive computational complexity. Variable selection using some of these approaches is yet another challenge as they may not guarantee the selection of significant predictors (variables that contribute to the true data generating process) given limited data size. In contrast, there exist high dimensional nonlinear time series models that are mostly inspired by high dimensional statistical methods. There are typically two kinds of approaches. In one approach, a small subset of significant variables is first selected and then nonlinear time series models are applied to selected variables. For example, independence screening techniques such as [10]–[12] or the MARS may be used to do variable selection. In another approach, dimension reduction method such as least absolute shrinkage and selection operator (LASSO) [13] are directly applied to nonlinear modeling. Sparse additive models have been developed in recent works of Ravikumar et al. [14] and Huang et al. [5]. In the work of Bazerque et al. [15], splines additive models together with group-sparsity penalty was proposed and applied to spectrum cartography. These offline approaches seem promising and may benefit from additional reductions in computational complexity.

In this work, inspired by the second approach, we develop a new method referred to as Sequential Learning Algorithm for Nonlinear Time Series (SLANTS). A challenging problem in sequential inference is that the data generating process varies with time, which is common in many practical applications [1]–[3]. We propose a method that can help address sequential inference of potentially time-varying models. Moreover, the proposed method provides computational benefits as we avoid repeating batch estimation upon sequential arrival of data. Specifically, we use the spline basis to dynamically approximate the nonlinear functions. The algorithm can efficiently give unequal weights to data points by design, as typical in adaptive filtering. We also develop an online version of group LASSO for dimensionality reduction (i.e. simultaneous

This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers W911NF-14-1-0508 and N66001-15-C-4028, by National science Foundation (NSF) grant numbers IIS-1149662 and IIS-1409177, and by Office of Naval Research (ONR) grant numbers N00014-14-1-0485 and N00014-17-1-2131.

Q. Han and E. Airoldi are with the Department of Statistics, J. Ding and V. Tarokh are with the School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes three demonstrating videos and two real datasets used in the paper. Contact hqychr@gmail.com for further questions about this work.

estimation and variable selection). To this end, the group LASSO regularization is re-formulated into a recursive estimation problem that produces an estimator close to the maximum likelihood estimator from batch data. We theoretically analyze the performance of SLANTS. Under reasonable assumptions, we also provide an estimation error bound, and a backward stepwise procedure that guarantees consistency in variable selection.

The outline of this paper is given next. In Section II, we formulate the problem mathematically and present our inference algorithm. In Section III, we present our theoretical results regarding prediction error and model consistency. In Section IV, we provide numerical results using both synthetic and real data examples. The results demonstrate excellent performance of our method. We make our conclusions in Section V.

II. Sequential modeling of nonlinear time series

In this section, we first present our mathematical model and cast our problem as l_1-regularized linear regression . We then propose an expectationmaximization (EM) type algorithm to sequentially estimate the underlying coefficients. Finally we disclose methods for tuning the underlying parameters. Combining our proposed EM estimation method with automatic parameter tuning, we tailor our algorithm to sequential time series applications.

A. Formulation of SLANTS

Consider a multi-dimensional time series given by

$$\boldsymbol{X}_t = [X_{1,t}, \ldots, X_{D,t}]^{\mathrm{T}} \in \mathbb{R}^D, t = 1, 2, \ldots$$

Our main objective in this paper is to predict the value of $\boldsymbol{X}_T$ at time T given the past observations $\boldsymbol{X}_{T-1}, \ldots, \boldsymbol{X}_1$. For simplicity, we present our results for the prediction of scalar random variable $X_{1,T+1}$. We start with the general formulation

$$X_{1,T} = f(\boldsymbol{X}_{T-1}, \ldots, \boldsymbol{X}_{T-L}) + \varepsilon_T, \tag{1}$$

where $f(\cdot, \cdots, \cdot)$ is smooth (or at least piece-wise smooth), ε_t are independent and identically distributed (i.i.d.) zero mean random variables and the lag order L is a finite but unknown nonnegative integer.

We rewrite the model in (1) as

$$X_{1,T} = f(X_{1,T-1}, \ldots, X_{1,T-L}, \ldots, X_{D,T-1}, \ldots, X_{D,T-L}) + \varepsilon_T.$$

With a slight abuse of notation, we rewrite the above model (1) as

$$Y_T = f(X_{1,T}, \ldots, X_{\tilde{D},T}) + \varepsilon_T, \tag{2}$$

with observations $Y_T = X_{1,T}$ and $[X_{1,T}, \ldots, X_{\tilde{D},T}] = [X_{1,T-1}, \ldots, X_{1,T-L}, \ldots, X_{D,T-1}, \ldots, X_{D,T-L}]$, where $\tilde{D} = DL$. To estimate $f(\cdot, \cdots, \cdot)$, we consider the following least squares formulation

$$\min_f \sum_{t=1}^{T} w_{T,t}(Y_t - f(X_{1,t}, \ldots, X_{\tilde{D},t}))^2 \tag{3}$$

where $\{w_{T,t} \in [0,1]\}$ are weights used to emphasize varying influences of the past data. The weights may also be used to accommodate different variance levels across dimensions. The appropriate choice of $\{w_{T,t} \in [0,1]\}$ will be later discussed in Section II-C.

In order to estimate the nonlinear function $f(\cdot, \cdots, \cdot)$, we further assume a nonlinear additive model, i.e.

$$f(X_{1,t}, \ldots, X_{\tilde{D},t}) = \mu + \sum_{i=1}^{\tilde{D}} f_i(X_i), \quad E\{f_i(X_i)\} = 0, \tag{4}$$

where f_i are scalar functions, and expectation is with respect to the stationary distribution of X_i. The second condition is required for identifiability. To estimate f_i, we use B-splines (extensions of polynomial regression techniques [16]). In our presentation, we consider the additive model mainly for brevity. Our methods can be extended to models where there exist interactions among $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{\tilde{D}}$ using multidimensional splines in a straight-forward manner.

We assume that there are v spline basis of degree ℓ for each f_i. Incorporating the B-spline basis into regression, we write

$$\begin{aligned} f_i(x) &= \sum_{j=1}^{v} c_{i,j} b_{i,j}(x), \\ b_{i,j}(x) &= B(x \mid s_{i,1}, \ldots, s_{i,v-\ell+1}) \end{aligned} \tag{5}$$

where $s_{i,1}, \ldots, s_{i,v-\ell+1}$ are the knots and $c_{i,j}$ are the coefficients associated with the B-spline basis. Replacing these into (3), the problem of interest is now the minimization of

$$\hat{e}_T = \sum_{t=1}^{T} w_{T,t} \left\{ Y_t - \mu - \sum_{i=1}^{\tilde{D}} \sum_{j=1}^{v} c_{i,j} b_{i,j}(X_{i,t}) \right\}^2 \tag{6}$$

over $c_{i,j}$, $i = 1, \ldots, \tilde{D}$, $j = 1, \ldots, v$, under the constraint

$$\sum_{t=1}^{T} \sum_{j=1}^{v} c_{i,j} b_{i,j}(x_i) = 0, \text{ for } i = 1, \ldots, L. \tag{7}$$

which is the sample analog of the constraint in (4). Equivalently, we obtain an unconstrained optimization problem by centering the basis functions. Let $b_{i,j}(x_{i,t})$ be replaced by $b_{i,j}(x_{i,t}) - T^{-1} \sum_{t=1}^{T} b_{i,j}(x_{i,t})$. By proper rearrangement, (6) can be rewritten into a linear regression form

$$\hat{e}_T = \sum_{t=1}^{T} w_{T,t}(Y_t - \boldsymbol{z}_t^{\mathrm{T}} \boldsymbol{\beta}_T)^2 \tag{8}$$

where $\boldsymbol{\beta}_T$ is a $(1 + \tilde{D}v) \times 1$ column vector to be estimated and $\boldsymbol{z}_t$ is $(1 + \tilde{D}v) \times 1$ column vector $\boldsymbol{z}_t = [1, b_{1,1}(x_{1,t}), \ldots, b_{1,v}(x_{1,t}), \ldots, b_{\tilde{D},1}(x_{\tilde{D},t}), \ldots, b_{\tilde{D},v}(x_{\tilde{D},t})]$. Let Z_T be the design matrix of stacking the row vectors $\boldsymbol{z}_t^{\mathrm{T}}, t = 1, \ldots, T$. Note that we have used $\boldsymbol{\beta}_T$ instead of a fixed $\boldsymbol{\beta}$ to emphasize that $\boldsymbol{\beta}_T$ may vary with time. We have used bold style for vectors to distinguish them from matrices. Let W_T be the diagonal matrix whose elements are $w_{T,t}, t = 1, \ldots, T$. Then the optimal β_T in (8) can be

recognized as the MLE of the following linear Gaussian model

$$\boldsymbol{Y}_T = Z_T\boldsymbol{\beta}_T + \boldsymbol{\varepsilon} \tag{9}$$

where $\boldsymbol{\varepsilon} \in \mathcal{N}(0, W_T^{-1})$. Here, we have used $\mathcal{N}(\boldsymbol{\mu}, V)$ to denote Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix V.

To obtain a sharp model from large L, we further assume that the expansion of $f(\cdot, \cdots, \cdot)$ is sparse, i.e., only a few additive components f_i are active. Selecting a sparse model is critical as models of over large dimensions lead to inflated variance, thus compromising the predictive power. To this end, we give independent Laplace priors for each sub-vector of $\boldsymbol{\beta}_T$ corresponding to each f_i. Our objective now reduces to obtaining the maximum a posteriori estimator (MAP)

$$\begin{aligned}&\log p(\boldsymbol{Y}_T \mid \boldsymbol{\beta}_T, Z_T) - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 \\ &= -\frac{1}{2}\sum_{t=1}^{T} w_{T,t}(Y_t - \boldsymbol{z}_t^{\mathrm{T}}\boldsymbol{\beta}_T)^2 - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 + c\end{aligned} \tag{10}$$

where c is a constant that depends only on W_T. The above prior corresponds to the so called group LASSO [17]. The bold $\boldsymbol{\beta}_{T,i}$ is to emphasize that it is not a scalar element of $\boldsymbol{\beta}_T$ but a sub-vector of it. It will be interesting to consider adaptive group LASSO [18], i.e., to use $\lambda_{T,i}$ instead of a unified λ_T and this is currently being investigated. We refer to [5] for a study of adaptive group LASSO for batch estimation.

B. Implementation of SLANTS

In order to solve the optimization problem given by (10), we build on an EM-based solution originally proposed for wavelet image restoration [19]. This was further applied to online adaptive filtering for sparse linear models [20] and nonlinear models approximated by Volterra series [21]. The basic idea is to decompose the optimization (10) into two parts that are easier to solve and iterate between them. One part involves linear updates, and the other involves group LASSO in the form of orthogonal covariance which leads to closed-form solution.

For now, we assume that the knot sequence $t_{i,1}, \ldots, t_{i,v}$ for each i and v is fixed. Suppose that all the tuning parameters are well-defined. We introduce an auxiliary variable τ_T that we refer to as the *innovation parameter*. This helps us to decompose the problem so that underlying coefficients can be iteratively updated. It also allows the sufficient statistics to be rapidly updated in a sequential manner. The model in (9) now can be rewritten as

$$\boldsymbol{Y}_T = Z_T\boldsymbol{\theta}_T + W_T^{-\frac{1}{2}}\boldsymbol{\varepsilon}_1,\ \boldsymbol{\theta}_T = \boldsymbol{\beta}_T + \tau_T\boldsymbol{\varepsilon}_2,$$

where

$$\boldsymbol{\varepsilon}_1 \in \mathcal{N}(0, I - \tau_T^2 W_T^{\frac{1}{2}} Z_T Z_T^{\mathrm{T}} W_T^{\frac{1}{2}}), \quad \boldsymbol{\varepsilon}_2 \in \mathcal{N}(0, I) \tag{11}$$

We treat $\boldsymbol{\theta}_T$ as the missing data, so that an EM algorithm can be derived. By basic calculations similar to that of [19], we obtain the kth step of EM algorithm

E step:

$$Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_T^{(k)}) = -\frac{1}{2\tau_T^2}\|\boldsymbol{\beta} - \boldsymbol{r}^{(k)}\|_2^2 - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_i\|_2 \tag{12}$$

where

$$\boldsymbol{r}^{(k)} = (I - \tau_T^2 A_T)\hat{\boldsymbol{\beta}}_T^{(k)} + \tau_T^2 B_T, \tag{13}$$

$$A_T = Z_T^{\mathrm{T}} W_T Z_T,\ \boldsymbol{B}_T = Z_T^{\mathrm{T}} W_T \boldsymbol{Y}_T. \tag{14}$$

The derivation of Equation (12) is included in the appendix.

M step: $\hat{\boldsymbol{\beta}}_T^{(k+1)}$ is the maximum of $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_T^{(k)})$ given by

$$\hat{\boldsymbol{\beta}}_{T,i}^{(k+1)} = \left[1 - \frac{\lambda_T \tau_T^2}{\|\boldsymbol{r}_i^{(k)}\|_2}\right]_+ \boldsymbol{r}_i^{(k)}, \quad i = 1, \ldots, \tilde{D}. \tag{15}$$

Suppose that we have obtained the estimator $\hat{\boldsymbol{\beta}}_T$ at time step T. Consider the arrival of the $(T+1)$th point $(y_{T+1}, \boldsymbol{z}_{T+1})$, respectively corresponding to the response and covariates of time step $T+1$. We first compute $\boldsymbol{r}_{T+1}^{(0)}$, the initial value of $\boldsymbol{r}$ to be input the EM at time step $T+1$:

$$\boldsymbol{r}_{T+1}^{(0)} = (I - \tau_T^2 A_{T+1})\hat{\boldsymbol{\beta}}_T + \tau_T^2 \boldsymbol{B}_{T+1}, \tag{16}$$

where

$$\begin{aligned}A_{T+1} &= (1-\gamma_{T+1})A_T + \gamma_{T+1}\boldsymbol{z}_{T+1}\boldsymbol{z}_{T+1}^{\mathrm{T}}, \\ \boldsymbol{B}_{T+1} &= (1-\gamma_{T+1})\boldsymbol{B}_T + \gamma_{T+1} y_{T+1}\boldsymbol{z}_{T+1}.\end{aligned} \tag{17}$$

Then we run the above EM for $K > 0$ iterations to obtain an updated $\hat{\boldsymbol{\beta}}_{T+1}$.

Remark 1: In the above equation, $\{\gamma_t\}$ is a nonnegative sequence which we refer to as the step sizes. We shall elaborate on its relation with $\{W_t\}$ in Subsection II-C.

SLANTS can be efficiently implemented. The recursive computation of A_T (resp. B_T) reduces the complexity from $O(\tilde{D}^3)$ to $O(\tilde{D}^2)$ (resp. from $O(\tilde{D}^2)$ to $O(\tilde{D})$). Moreover, straightforward computations indicate that the complexity of SLANTS at each time t is $O(\tilde{D}^2)$, which does not depend on T. Coordinate descent [22] is perhaps the most widely used algorithm for batch LASSO. Adapting coordinate descent to sequential setting has the same complexity for updating sufficient statistics. But straightforward use of batch LASSO has complexity $O(\tilde{D}^2 T)$.

Theorem 1: At each iteration, the mapping from $\hat{\boldsymbol{\beta}}_T^{(k)}$ to $\hat{\boldsymbol{\beta}}_T^{(k+1)}$ is a contraction mapping for any τ_T, whenever the absolute values of all eigenvalues of $I - \tau_T^2 A_{T+1}$ are less than one. In addition, there exists a unique global maximum point of (10) denoted by $\hat{\boldsymbol{\beta}}_T$, and the error $\|\hat{\boldsymbol{\beta}}_T^{(k+1)} - \hat{\boldsymbol{\beta}}_T\|_2$ decays exponentially in k.

Remark 2: The theorem states that EM can converge exponentially fast to the MAP of (10). From its assumption, it can be directly calculated that (10) as a function of $\boldsymbol{\beta}_T$ is strictly concave. We note that the assumption is not mild, so the application of Theorem 1 is limited. But the proposed algorithm does converge exponentially fast in our various synthetic and real data experiments. The proof of Theorem 1 is given in the appendix.

C. The choice of tuning parameters: from a prequential perspective

To evaluate the predictive power of an inferential model estimated from all the currently available data, ideally we would apply it to independent and identically generated datasets. However, it is not realistic to apply this cross-validation idea to real-world time series data, since real data is not permutable and has a "once in a lifetime" nature. As an alternative, we adopt a prequential perspective [23] that the goodness of a sequential predictive model shall be assessed by its forecasting ability.

Specifically, we evaluate the model in terms of the one-step prediction errors upon each newly arrived data point and subsequently tune the necessary control parameters, including regularization parameter λ_t and innovation parameter τ_t (see details below). Automatic tuning of the control parameters are almost a necessity in many real-world applications in which any theoretical guidance (e.g., our Theorem 2) may be insufficient or unrealistic. Throughout our algorithmic design, we have adhered to the prequential principle and implemented the following strategies.

The choice of $w_{T,t}$: In view of equation (17), $w_{T,t}$ is determined by $w_{1,1} = \gamma_1$, and

$$w_{t,t} = \gamma_t,\ w_{t,j} = w_{t-1,j}(1-\gamma_t), \quad j = 1, \ldots, t-1,$$

for $t > 1$.

It includes two special cases that have been commonly used in the literature. The first case is $\gamma_t = 1/t$. It is easy to verify that $w_{T,t} = 1/T, t = 1, \ldots, T$ for any T. This leads to the usual least squares. The second case is $\gamma_t = c$ where c is a positive constant. It gives $w_{T,t} = c(1-c)^{T-t}, t = 1, \ldots, T$. From (3), the estimator of f remains unchanged by rescaling $w_{T,t}$ by $1/c$, i.e. $w_{T,t} = (1-c)^{T-t}$ which is a series of powers of $1-c$. The value $1-c$ has been called the "forgetting factor" in the signal processing literature and used to achieve adaptive filtering [20].

The choice of τ_T: Because the optimization problem

$$\log p(\boldsymbol{Y}_T \mid \boldsymbol{\beta}_T) - \lambda_T \sum_{i=1}^{L} \|\boldsymbol{\beta}_{T,i}\|_2 \tag{18}$$

is convex, as long as τ_T is proper, the EM algorithm converges to the optimum regardless of what τ_T is. But τ_T affects the speed of convergence of EM as $\lambda_T \tau_T^2$ determines how fast β_T shrinks. Intuitively the larger τ_T is, the faster is the convergence. Therefore we prefer τ_T to be large and proper. A necessary condition for τ_T to be proper is to ensure that the covariance matrix of $\boldsymbol{\varepsilon}_1$ in

$$\boldsymbol{\varepsilon}_1 \in \mathcal{N}(0, I - \tau_T^2 W^{\frac{1}{2}} Z_T Z_T^{\mathrm{T}} W^{\frac{1}{2}}), \quad \boldsymbol{\varepsilon}_2 \in \mathcal{N}(0, I) \tag{19}$$

is positive definite. Therefore, there is an upper bound $\bar{\tau}_T$ for τ_T, and $\bar{\tau}_T$ converges to a positive constant $\bar{\tau}$ under some mild assumptions (e.g. the stochastic process X_t is stationary). Extensive experiments have shown that $\bar{\tau}_T/2$ produces satisfying results in terms of model fitting. However, it is not computationally efficient to calculate $\bar{\tau}_T$ at each T in SLANTS. Nevertheless without computing $\bar{\tau}_T$, we can determine if $\tau_T < \bar{\tau}_T$ by checking the EM convergence. If τ_T exceeds $\bar{\tau}_T$, the EM would diverge and coefficients go to infinity exponentially fast. This can be proved via a similar argument to that of proof of Theorem 1. This motivates a lazy update of τ_T with shrinkage only if EM starts to diverge.

The choice of λ_T: On the choice of regularization parameter λ_T, different methods have been proposed in the literature. The common way is to estimate the batch data for a range of different λ_T's, and select the one with minimum cross-validation error. To reduce the underlying massive computation required for such an approach, in the context of Bayesian LASSO [24], [25] proposed an sequential Monte Carlo (SMC) based strategy to efficiently implement cross-validation. The main proposal is to treat the posterior distributions educed by an ordered sequence of λ_T as $\pi_t, t = 0, 1, \ldots$, the target distributions in SMC, and thus avoid the massive computation of applying Markov chain Monte Carlo (MCMC) for each λ independently. Another method is to estimate the hyper-parameter λ_T via empirical Bayes method [24]. In our context, however, it is not clear whether the Bayesian setting with MCMC strategy can be efficient, as the dimension Lv can be very large. An effective implementation technique is to run three channels of our sequential modeling, corresponding to $\lambda_T^- = \lambda_T/\delta, \lambda_T, \lambda_T^+ = \lambda_T * \delta$, where $\delta > 1$ is a small step size. The one with minimum average prediction error over the latest window of data was chosen as the new λ_T. For example, if λ_T^- gives better performance, let the three channels be $\lambda_T^-/\delta, \lambda_T^-, \lambda_T^- * \delta$. If there is an underlying optimal λ^* which does not depend on T, we would like our channels to converge to the optimal λ^* by gradually shrinking the stepsize δ. Specifically in case that the forgetting factor $\gamma_t = 1/t$, we let $\delta_T = 1 + \frac{1}{T}(\delta - 1)$ so that the step size $\delta_T \to 1$ at the same speed as weight of new data.

The choice of knots: The main difficulty in applying spline approximation is in determining the number of the knots to use and where they should be placed. Jupp [26] has shown that the data can be fit better with splines if the knots are free variables. de Boor suggests the spacing between knots is decreased in proportion to the curvature (second derivative) of the data. It has been shown that for a wide class of stationary process, the number of knots should be of the order of $O(T^{\zeta})$ for available sample size T and some positive constant ζ to achieve a satisfying rate of convergence of the estimated nonlinear function to the underlying truth (if it exists) [27]. Nevertheless, under some assumptions, we will show in Theorem 2 that the prediction error can be upper bounded by an arbitrarily small number (which depends on the specified number of knots). It is therefore possible to identify the correct nonzero additive components in the sequential setting. On the other hand, using a fixed number of knots is computationally desirable because sharp selection of significant spline basis/support in a potentially varying environment is computationally intensive. It has been observed in our synthetic data experiments that the variable selection results are not very sensitive to the number of knots as long as this number is moderately large (e.g. around $v = 10$).

III. Theoretical results

Consider the harmonic step size $\gamma_t = 1/t$. For now assume that the sequential update at each time t produces $\hat{\boldsymbol{\beta}}_t$ that is the same as the penalized least squares estimator given batch data. We are interested in two questions. First, how to extend the current algorithm in order to take into account an ever-increasing number of dimensions? Second, is it possible to select the "correct" nonzero components as sample size increases?

The first question is important in practice as any prescribed finite number of dimensions/time series may not contain the data-generating process, and it is natural to consider more candidates whenever more samples are obtained. It is directly related to the widely studied high-dimensional regression for batch data. In the second question, we are not only interested in optimizing the prediction error but also to obtain a consistent selection of the true nonzero components. Moreover, in order to maintain low complexity of the algorithm, we aim to achieve the above goal by using a fixed number of spline basis. We thus consider the following setup. Recall the predictive model (1) and its alternative form (2). We assume that L is fixed while D is increasing with sample size T at certain rate.

Following the setup of [28], we suppose that each X_d takes values from a compact interval $[a, b]$. Let $[a, b]$ be partitioned into J equal-sized intervals $\{I_j\}_{j=1}^J$, and let $\mathfrak{F}$ denote the space of polynomial splines of degree $\ell \geq 1$ consisting of functions $g(\cdot)$ satisfying 1) the restriction of $g(\cdot)$ to each interval is a polynomial of degree ℓ, and 2) $g(\cdot) \in C^{\ell-1}[a, b]$ ($\ell - 1$ times continuously differentiable). Typically, splines are called linear, quadratic or cubic splines accordingly as $\ell = 1, 2$, or 3. There exists a normalized B-spline basis $\{b_j\}_{j=1}^v$ for $\mathfrak{F}$, where $v = J + \ell$, and any $f_i(x) \in \mathfrak{F}$ can be written in the form of (5). Let $k \leq \ell$ be a nonnegative integer, $\beta \in (0, 1]$ that $p = k+\beta > 0.5$, and $M > 0$. Suppose each considered (non)linear function f has kth derivative, $f^{(k)}$, and satisfies the Holder condition with exponent β: $|f^{(k)}(x) - f^{(k)}(x')| < M|x - x'|^\beta$ for $x, x' \in [a, b]$. Define the norm $\|f\|_2 = \sqrt{\int_a^b f(x)^2 dx}$. Let $f^* \in \mathfrak{F}$ be the best L_2 spline approximation of f. Standard results on splines imply that $\|f_d - f_d^*\|_\infty = O(v^{-p})$ for each d. The spline approximation is usually an estimation under a mis-specified model class (unless the data-generating function is low-degree polynomials), and large v narrows the distance to the true model. We will show that for large enough v, it is possible to achieve the aforementioned two goals. To make the problem concrete, we need the following assumptions on the data-generating procedure.

Assumption 1: The number of additive components is finite and will be included into the candidate set in finite time steps. In other words, there exists a "significant" variable set $S_0 = \{i_1, \ldots, i_{D_0}\}$ such that 1) $f_d(x) \neq 0$ for each $d \in S_0$, 2) $f_d(x) \equiv 0$ for $d \notin S_0$, and 3) both D_0 and i_{D_0} are finite integers that do not depend on sample size T.

We propose two steps for a practitioner targeting two goals given below.

Step 1. (unbiasedness) This step aims to discover the significant variable set with probability close to one as more data is collected. The approach is to minimize the objective function in (10), and it can be efficiently implemented using the proposed sequential algorithm in Section II-B with negligible error (Theorem 1). In the case of equal weights $w_{T,t} = 1/T$, it can be rewritten as

$$\|Y_T - Z_T\boldsymbol{\beta}_T\|_2^2 + \tilde{\lambda}_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 \tag{20}$$

where $\tilde{\lambda}_T = 2T\lambda_T$. Due to Assumption 1, the significant variable set S_0 is included in the candidate set $\{1, \ldots, \tilde{D}\}$ for sufficiently large T. Our selected variables are those whose group coefficients are nonzero, i.e. $S_1 = \{d : 1 \leq d \leq \tilde{D}, \hat{\boldsymbol{\beta}}_{T,d} \neq \mathbf{0}\}$. We are going to prove that all the significant variables will be selected by minimizing (20) with appropriately chosen $\tilde{\lambda}_T$, i.e., $S_0 \subseteq S_1$.

Step 2. (minimal variance) The second step is optional and it is applied only when a practitioner's goal is to avoid selecting any redundant variables outside S_0. Suppose that we obtain a candidate set of $\tilde{D}$ variables S_1 (satisfying $S_0 \subseteq S_1$ from the previous step). Since a thorough search over all subsets of variables is computationally demanding, we use a backward stepwise procedure. We start with the set of selected variables S_1, delete one variable at a time by minimizing the MSE of a spline model with $v_T = T^\zeta$ number of equally spaced knots. We note that v_T in the optional Step 2 can be different from the v in SLANTS. Specifically, suppose that at step k ($k = 1, 2, \ldots$), the survived candidate models are indexed by $\mathcal{S}^{(k)}$. We solve the least-squares problem for each $\bar{d} \in \mathcal{S}^{(k)}$

$$\hat{e}_{\bar{d}}^{(k)} = \min_{\mu, c_{d,j}} \sum_{t=1}^{T} \left(Y_t - \mu - \sum_{d \in \mathfrak{S}} \sum_{j=1}^{v_T} c_{d,j} b_{d,j}(X_{d,t}) \right)^2 \tag{21}$$

where $\mathfrak{S} = \mathcal{S}^{(k-1)} - \{\bar{d}\}$, and select $\bar{d} = \bar{d}_k^*$ that minimize the $\hat{e}_{\bar{d}}^{(k)}$ with minimum denoted by $\hat{e}^{(k)}$. Here $A - B$ denotes the set of elements that are in a set A but not in a set B. We let $\mathcal{S}^{(k)} = \mathcal{S}^{(k-1)} - \{\bar{d}_k^*\}$. By default, we let $\mathcal{S}^{(0)} = S_1$ and use $\hat{e}^{(0)}$ to denote the minimum of (21) with $\mathfrak{S} = S_1$. If $\hat{e}^{(k-1)} - \hat{e}^{(k)} < (v_T \log T)/T$, i.e., the gain of goodness of fit is less than the incremented Bayesian information criterion (BIC) penalty [29], then we stop the procedure and output $S_2 = \mathcal{S}^{(k-1)}$; otherwise we proceed to the $(k+1)$th iteration. We prove that the finally selected subset S_2 satisfies $\lim_{T\to\infty} \mathrm{pr}(S_2 = S_0) = 1$.

Before we proceed to the theoretical result, we introduce some necessary assumptions and their interpretations.

Assumption 2: There is a positive constant c_0 such that $\min_{d \in S_0} \|f_d\|_2 \geq c_0$.

Assumption 3: The noises ε_t are sub-Gaussian distributed, i.e., $E(e^{w\varepsilon_t}) \leq e^{w^2\sigma^2/2}$ for a constant $\sigma > 0$ and any $w \in \mathbb{R}$.

Assumption 4: Suppose that S_1 is a finite subset of $\{1, \ldots, \tilde{D}\}$. In addition, the "design matrix" Z_{S_1} satisfies $Z_{S_1}^{\mathrm{T}} Z_{S_1}/T \geq \kappa$ for a positive constant κ that depend only on v (the number of splines).

We use $o_p(1)$ and $O_p(1)$ to denote a sequence of random variables that converges in probability to zero, and that is stochastically bounded, respectively. We use $O(1)$ to denote a bounded deterministic sequence.

Theorem 2: Suppose that Assumptions 1-4 hold. Then for any given v it holds that

$$\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 \leq 8c_2 v^{-2p}/\kappa + O_p(T^{-1}\log\tilde{D}) + O_p(T^{-1}) + O(T^{-2}\tilde{\lambda}^2) \quad (22)$$

for some positive constant c_2. If we further assume that $\log\tilde{D} = o(T)$, $\tilde{\lambda} = o(T)$, then there exists a constant $c_1 > 0$ such that for all $v > c_1 c_0^{-1/p}\max\{1, c_0^{-\frac{1}{p(2p+1)}}\}$, $\lim_{T\to\infty}\mathrm{pr}(S_0 \subseteq S_1) = 1$.

Remark 3: Theorem 2 gives an error bound between the estimated spline coefficients with the oracle, where the first term is dominating. As a result, if v is sufficiently large, then it is guaranteed that S_0 will be selected with probability close to one. We note that the constant c_1 depends only on the true nonlinear function and the selected spline basis function. In proving Theorem 2, Assumption 2-3 serve as standard conditions to ensure that a significant variable is distinguishable, and that any tail probability could be well bounded. Assumption 4 is needed to guarantee that if the estimated coefficients $\hat{\beta}$ produces low prediction errors, then it is also close to the true (oracle) coefficients. This assumption is usually guaranteed by requiring $\tilde{\lambda} > c\sqrt{T\log D}$. See for example [5], [30].

To prove the consistency in step 2, we also need the following assumption (which further requires that the joint process is strictly stationary and strongly mixing).

Assumption 5: $\sup_x\{E(|Y_t|^r|\boldsymbol{X}_t = x)\} < \infty$ for some $r > 2$.

The α-mixing coefficient is defined as $\alpha_S(j) = \sup\{P(E_y \cap E_x) - P(E_y)P(E_x) : E_y \in \sigma(\{(Y_{\tilde{t}}, X_{d,\tilde{t}}, d \in S) : \tilde{t} \leq n\}), E_x \in \sigma(\{(Y_{\tilde{t}}, X_{d,\tilde{t}}, d \in S) : \tilde{t} \geq n+j\})\}$, where $\sigma(\cdot)$ denotes the σ-field generated by the random variables inside the parenthesis.

Assumption 6: The process $\{(X_{d,t}, d \in S_1)\}$ is strictly stationary, and the joint process $\{(Y_t, X_{d,t}, d \in S_1)\}$ is α-mixing with coefficient

$$\alpha_{S_1}(j) \leq \min\{O(j^{-2.5\zeta/(1-\zeta)}), O(j^{-2r/(r-2)})\},$$

where ζ has been defined in Step 2.

Theorem 3: Suppose that Assumptions 1-6 hold, then the S_2 produced by the above step 2 satisfies $\lim_{T\to\infty}\mathrm{pr}(S_2 = S_0) = 1$.

IV. Numerical results

In this section, we present experimental results to demonstrate the theoretical results and the advantages of SLANTS on both synthetic and real-world datasets. The synthetic experiments include cases where the data-generating model is fixed over time, is varying over time, or involves large dimensionality.

A. Synthetic data experiment: modeling nonlinear relation in stationary environment

The purpose of this experiment is to show the performance of SLANTS in stationary environment where the data-generating model is fixed over time. We generated synthetic data using the following nonlinear model

$$X_{1,t} = \epsilon_{1,t}$$
$$X_{2,t} = 0.5X_{1,t-1}^2 - 0.8X_{1,t-7} + 0.2\epsilon_{2,t},\ t = 1,\ldots,500$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are i.i.d. standard Gaussian. The goal is to model/forecast the series $X_{2,t}$. We choose $L = 8$, and place $v = 10$ quadratic splines in each dimension. The knots are equally spaced between the 0.01 and 0.99 quantiles of observed data. The initial L values of $X_{2,t}$ are set to zeros. We choose the step size $\gamma_t = 1/t$ to ensure convergence.

Simulation results are summarized in Fig. 1. The left-top plot shows the convergence of all the $2\times 8\times 10 = 160$ spline coefficients. The right-top plot shows how the eight nonlinear components $f_d, d = 1,\ldots,8$ evolve, where the number 1-8 indicate each additive component (splines). The values of each function are centralized to zero for identifiability. The remaining two plots show the optimal choice of control parameters λ_t and τ_t that have been automatically tuned over time. In the experiment, the active components f_1 and f_7 are correctly selected and well estimated. It is remarkable that the convergence is mostly achieved after only a few incoming points (less than the number of coefficients 160).

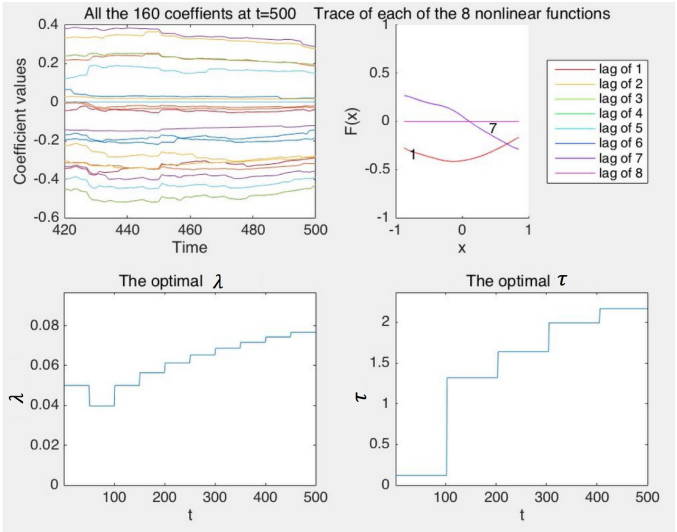

Fig. 1. Four subplots show the estimated coefficients of splines, nonlinear functions, and trace plots of automatically-tuned regularization parameter λ_t and innovation parameter τ_t. A demo video is available in the supplement.

B. Synthetic data experiment: modeling nonlinear relation in adaptive environment

The purpose of this experiment is to show the performance of SLANTS in terms of prediction and nonlinearity identification when the underlying date generating model varies over time.

We have generated a synthetic data using the following nonlinear model where there is a change at time $t = 500$,

$$X_{1,t} = \epsilon_{1,t}$$
$$X_{2,t} = 0.5X_{1,t-1}^2 - 0.8X_{1,t-7} + 0.2\epsilon_{2,t},\ t = 1,\ldots,500$$
$$X_{1,t} = u_{1,t}$$
$$X_{2,t} = -2X_{1,t-1}^2 + \exp(X_{1,t-7}) + 0.2\epsilon_{2,t},\ t = 501,\ldots,1000$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are i.i.d. standard Gaussian. $u_{1,t}$ are i.i.d. uniform on $[-1, 1]$. The goal is to model the series $X_{2,t}$. Compared with the previous experiment, the only difference is that the forgetting factor is set to $\gamma = 0.99$ in order to track potential changes in the underlying true model. Fig. 2 shows that SLANTS successfully tracked a change after the change point $t = 500$. The top plot in Fig. 2 shows the inference results right before the change. It successfully recovers the quadratic pattern of lag 1 and linear effect of lag 7. The bottom plot in Fig. 2 shows the inference results at $t = 1000$. It successfully finds the exponential curve of lag 7 and reversed sign of the quadratic curve of lag 1. From the bottom left subplot we can see how the autotuning regularization parameter decreases since the change point $t = 500$.

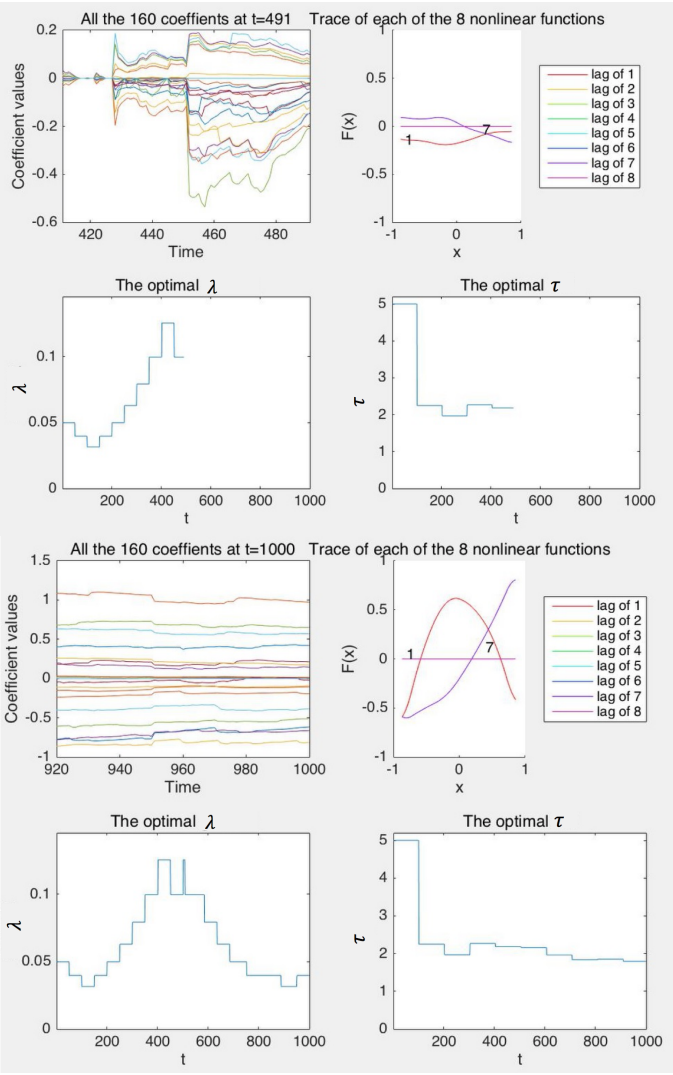

Fig. 2. Two plots stacked vertically, each consisting of four subplots that show the estimated coefficients of splines, nonlinear functions, and trace plots of automatically-tuned regularization parameter λ_t and innovation parameter τ_t at time $t = 491$ and $t = 1000$ respectively. A demo video is available in the supplement.

C. Synthetic data experiment: causal discovery for multi-dimensional time series

The purpose of this experiment is to show the performance of SLANTS in identifying nonlinear functional relation (thus Granger-type of causality) among multi-dimensional time series. We have generated a 9-dimensional time series using the following nonlinear network model,

$$
\begin{aligned}
X_{1,t} &= \epsilon_{1,t} \\
X_{2,t} &= 0.6X_{3,t-1} + \epsilon_{2,t} \\
X_{3,t} &= 0.3X_{4,t-2}^2 + \epsilon_{3,t} \\
X_{4,t} &= 0.7X_{5,t-1} - 0.2X_{5,t-2} + \epsilon_{4,t} \\
X_{5,t} &= -0.2X_{2,t-1}^2 + \epsilon_{5,t} \\
X_{6,t} &= 0.5X_{6,t-2} + 1 + \epsilon_{6,t} \\
X_{7,t} &= 2\exp(-X_{7,t-2}^2) + \epsilon_{7,t} \\
X_{8,t} &= 6X_{7,t-1} - 5X_{9,t-2} + \epsilon_{8,t} \\
X_{9,t} &= -X_{6,t-1} + 0.9X_{7,t-2} + \epsilon_{9,t}
\end{aligned}
$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are i.i.d. standard Gaussian. The initial L values are set to zero. The goal is to model each dimension and draw sequential causality graph based on the estimation. We choose $L = 2$, $v = 10$ and $\gamma_t = 1/t$. For illustration purpose, we only show the estimation for $X_{9,t}$. The left-top plot shows the 9 dimensional raw data that are sequentially obtained. The right-top plot shows the convergence of the $DLv = 9 \times 2 \times 10 = 180$ coefficients in modeling $X_{9,t}$. The right-bottom plot shows how the nonlinear components $f : X_{6,t-1} \mapsto X_{9,t}$ and $f : X_{7,t-2} \mapsto X_{9,t}$ evolve. Similar as before, the values of each function are centralized to zero for identifiability. The left-bottom plot shows the causality graph, which is the digraph with black directed edges and edge labels indicating functional relations. For example, in modeling $X_{9,t}$, if the function component corresponding to $X_{6,t-1}$ is nonzero, then we draw a directed edge from 6 to 9 with label 1; if the function components corresponding to both $X_{6,t-1}$ and $X_{6,t-2}$ are nonzero, then we draw a directed edge from 6 to 9 with label 12. The true causality graph (determined by the above data generating process) is drawn as well, in red thick edges. From the simulation, the discovered causality graph quickly gets close to the truth.

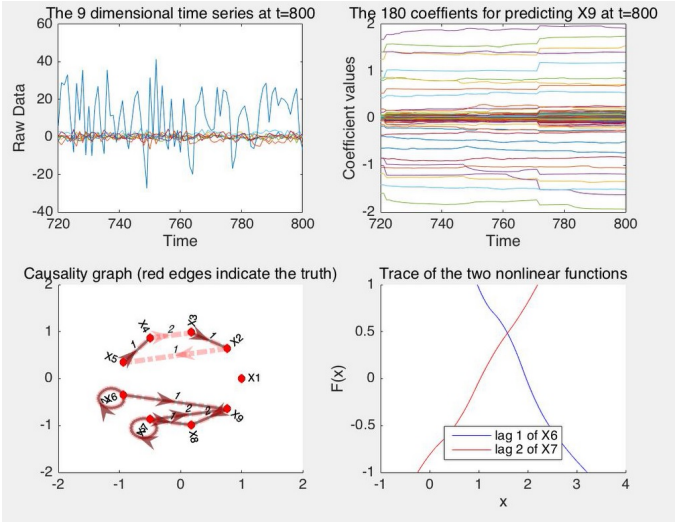

Fig. 3. Four subplots show the time series data, convergence of the coefficients, causality graph, and trace plot of the nonlinear functions. A demo video is available in the supplement.

D. Synthetic data experiment: computational cost

The purpose of this experiment is to show that SLANTS is computationally efficient by comparing it with standard batch group LASSO algorithm. We use the same data generating process in the first synthetic data experiment, and let the size of data be $T = 100, 200, \ldots, 1000$.

We compare SLANTS with the standard R package "grplasso" [31] and "gglasso" [32] which implement widely used group LASSO algorithms. The package "gglasso" implements the efficient active-set algorithm proposed in [33]. For the two packages, at each time t, solution paths on a fixed grid of 100 penalties are calculated. To provide fair comparisons, we run SLANTS in two ways. The first is the proposed algorithm with adaptive tuned penalties. In the table, it is denoted as SLANTS(a). The second is SLANTS without adaptive tuning but also run on a fixed grid of 100 equivalent penalties as in "grplasso" and "gglasso", denoted as SLANTS(b). In computing solution paths, we adopted the techniques suggested in [33]. The results are shown in Table I.

Table I shows the time in seconds for SLANTS(a), SLANTS(b), gglasso, and grplasso to run through a dataset sequentially with different size T. Each run is repeated 30 times and the standard error of running time is shown in parenthesis. From Table I, the computational cost of SLANTS grows linearly with T while gglasso and grplasso grow much faster. Moreover, the prediction error is very similar for SLANTS(b), gglasso and grplasso on the grid of penalties. This is understandable as they calculate the solution to the same optimization problem. SLANTS(a) approaches the optimal prediction error as the penalty parameter is stabilized. But SLANTS(a) is faster than SLANTS(b) as it only calculates solutions to three penalties at each time. In summary, both SLANTS(a) and SLANTS(b) are computationally faster than existing batch algorithms with comparable prediction performance.

The computational cost of SLANTS is slightly larger than that of grplasso when $T < 100$. This is because SLANTS is written purely in R, while the core part of gglasso and grplasso is implemented in Fortran (which is usually a magnitude faster than R). However, the growth of computational cost of SLANTS is much slower than that of grplasso, and thus SLANTS is faster for large T.

TABLE I
THE TABLE SHOWS THE COMPUTATIONAL COST IN SECONDS WITH STANDARD ERROR IN PARENTHESIS FOR SLANTS(A), SLANTS(B), GGLASSO, AND GRPLASSO, WITH INCREASING T.

T	SLANTS(a)	SLANTS(b)	gglasso	grplasso
100	4.8(0.1)	35.5(2.2)	32.6(2.5)	9.9(3.6)
200	11.1(0.3)	82.5(3.8)	110.8(7.9)	98.3(9.3)
300	15.4(0.7)	131.4(5.6)	204.3(9.2)	238.6(16.7)
400	21.4(0.7)	180.3(7.2)	296.2(10.7)	392.3(21.4)
500	26.0(0.9)	228.8(9.0)	386.8(12.1)	563.2(26.3)
600	31.3(1.1)	277.0(10.8)	477.5(13.4)	753.3(30.6)
700	37.1(1.2)	324.8(12.7)	569.4(15.0)	961.3(34.6)
800	42.1(1.4)	372.3(14.5)	663.0(19.1)	1189.0(38.5)
900	46.3(1.6)	419.4(16.3)	758.6(20.4)	1435.7(43.3)
1000	53.3(1.8)	466.3(18.1)	856.5(21.3)	1702.5(46.8)

E. Real data experiment: Boston weather data from 1980 to 1986

In this experiment, we study the daily Boston weather data from 1980 Jan to 1986 Dec. with $T = 2557$ points in total. The data is a six-dimensional time series, with each dimension corresponding respectively to temperature (K), relative humidity (%), east-west wind (m/s), north-south wind (m/s), sea level pressure (Pa), and precipitation (mm/day). In other words, the raw data is in the form of $X_{d,t}, d = 1, \ldots, 6, t = 1, \ldots, T$. We plot the raw data corresponding to year 1980 (i.e. $X_{d,t}, d = 1, \ldots, 6, t = 1, \ldots, 366$) in Fig. 4.

We compare the predictive performance of SLANTS with that of a linear model. For brevity, suppose that we are going to predict the east-west wind. We chose the autoregressive model of order 3 (denoted by AR(3)) as the representative linear model. The order was chosen by applying *Bridge criterion* [34] to the batch data of T observations. We started processing the data from $t_0 = 10$, and for each $t = t_0+1, \ldots, T$ the one-step ahead prediction error $\hat{e}_t$ was made by applying AR(3) and SLANTS to the currently available $t-1$ observations. The cumulated average prediction error at time step t is computed to be $\sum_{i=t_0+1}^{t} \hat{e}_i/(t-t_0)$, where $\hat{e}_i$ is the squared difference between the true observation and our prediction at time step i. The results are shown in Fig. 5(a). At the last time step, the significant (nonzero) functional components are the third, fourth, and sixth dimension, corresponding to EW wind, NS wind, precipitation, have been plotted in Fig. 5 (b), (c), (d), respectively. From the plot, the marginal effect of $X_{4,t}$ on $X_{3,t+1}$ is clearly nonlinear. It seems that the correlation is low for $X_{4,t} < 0$ and high for $X_{4,t} > 0$. In fact, if we let $\mathfrak{T} = \{t : X_{4,t} > 0\}$, the correlation of $\{X_{4,t} : t \in \mathfrak{T}\}$ with $\{X_{3,t+1} : t \in \mathfrak{T}\}$ is 0.25 (with p value 1.4×10^{-8}) while $\{X_{4,t} : t \notin \mathfrak{T}\}$ with $\{X_{3,t+1} : t \notin \mathfrak{T}\}$ is -0.05 (with p value 0.24)

F. Real data experiment: the weekly unemployment data from 1996 to 2015

In this experiment, we study the US weekly unemployment initial claims from Jan 1996 to Dec 2015. The data is a one-dimensional time series with $T = 1043$ points in total. we plot the raw data in Fig. 6.

Though the data exhibits strong cyclic pattern, it may be difficult to perform cycle-trend decomposition in a sequential setting. We explore the power of SLANTS to do lag selection to compensate the lack of such tools.

We compare three models. The first model, AR(5), is linear autoregression with lag order 5. The lag order was chosen by applying Bridge criterion [34] to the batch data. The second and third are SLANTS(1) with linear spline and SLANTS(2) with quadratic splines. SLANTS(1) have 1 spline per dimension, which is exactly LASSO with auto-tuned penalty parameter in SLANTS. SLANTS(2) have 8 splines per dimension. We allow SLANTS to select from a maximum lag of 55, which is roughly the size of annual cycle of 52 weeks.

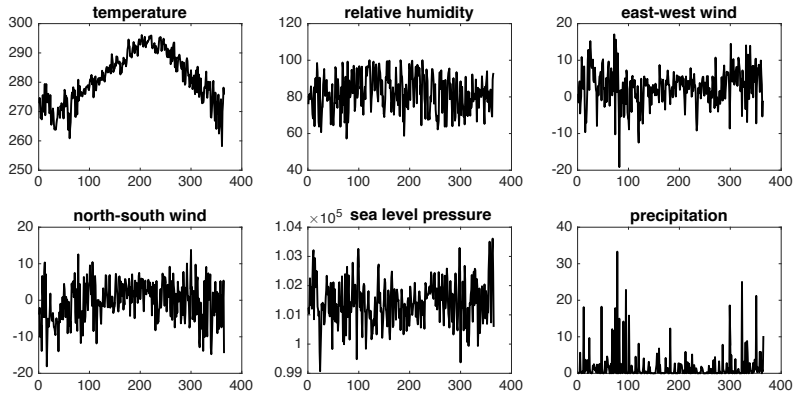

Fig. 4. A graph showing the raw data of (a) temperature (K), (b) relative humidity (%), (c) east-west wind (m/s), (d) north-south wind (m/s), (e) sea level pressure (Pa), and (f) precipitation (mm/day).

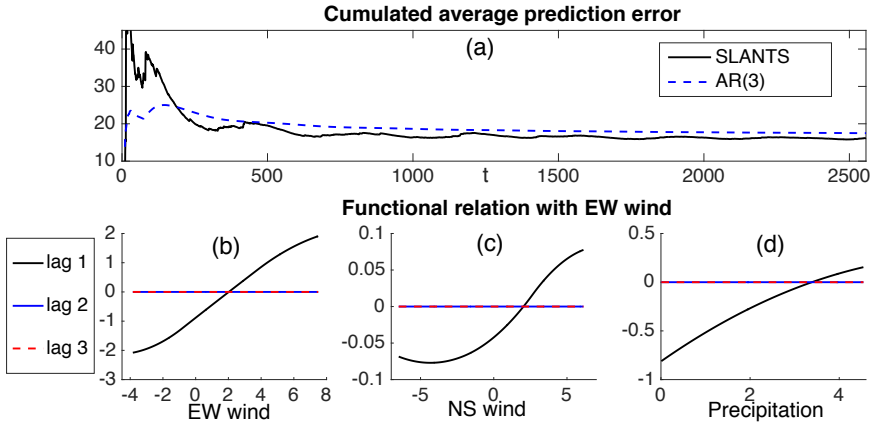

Fig. 5. A graph showing (a) the cumulated average one-step ahead prediction error of east-west wind (m/s) produced by two approaches, and east-west wind decomposed into nonlinear functions of lagged values of (b) east-west wind, (c) north-south wind (m/s), and (c) precipitation (mm/day). The functions were output from SLANTS at the last time step $t = T$.

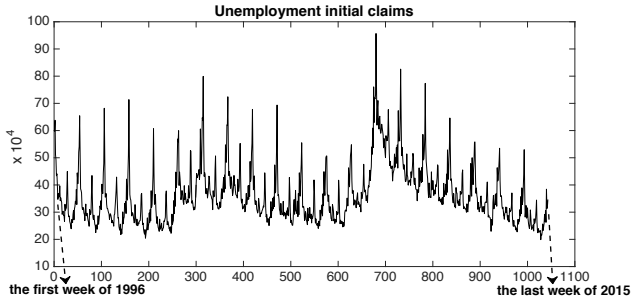

Fig. 6. A graph showing the raw data of the number of unemployment initial claims.

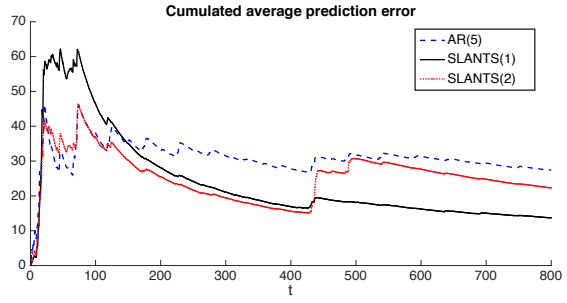

Fig. 7. A graph showing the cumulated average one-step ahead prediction error at each time step produced by three approaches: linear autoregressive model, SLANTS with linear splines, and SLANTS with quadratic splines.

Fig. 7 shows the cumulative average one-step ahead prediction error at each time step by the above three approaches. Here we plot the fits to the last 800 data points due to the unstable estimates of AR and SLANTS at the beginning. The results show that SLANTS is more flexible and reliable than linear autoregressive model in practical applications. Both SLANTS(1) and SLANTS(2) selected lag 1,2,52,54 as significant predictors. It is interesting to observe that SLANTS(2) is preferred to SLANTS(1) before time step 436 (around the time when the 2008 financial crisis happened) while the simpler model SLANTS(1) is preferred after that time step. The fitted quadratic splines from SLANTS(2) are almost linear, which means the data has little nonlinearity. So SLANTS(1) performs best overall.

V. Concluding Remarks

To address several challenges in time series prediction that arises from environmental science, economics, and finance, we proposed a new method to model nonlinear and high dimensional time series data in a sequential and adaptive manner. The performance of our method was demonstrated by both

synthetic and real data experiments. We also provided rigorous theoretical analysis of the rate of convergence, estimation error, and consistency in variable selection of our method.

Future work may include modeling and joint prediction of $\boldsymbol{X}_{1,T}, \ldots, \boldsymbol{X}_{D,T}$. Currently, the prediction is separated into D individual problems. The performance may be further enhanced by considering potential correlations of innovations in each series. Adaptive placement of knots is another direction for future work. The knot sequence should adequately cover the range of data. In this paper, we assumed that the range of data is known. In some practical applications, however, the range may vary over time. In such case, it would be helpful to add a rejuvenation step that routinely updates the empirical domain of the data (and thus the knot placement).

APPENDIX

We prove Theorems 1-3 in the appendix. For any real-valued column vector $x = [x_1, \ldots, x_m]$, we let $\|x\|_2 = (\sum_{i=1}^m x_i^2)^{1/2}$, $\|x\|_A = x^{\mathrm{T}} A x$ denote respectively the ℓ_2 norm and matrix norm (with respect to A, a positive semidefinite matrix).

PROOF OF THEOREM 1

At time T and iteration k, we define the functions $h(\cdot)$ and $g(\cdot)$ that respectively map $\hat{\boldsymbol{\beta}}_T^{(k)}$ to $\boldsymbol{r}_T^{(k)}$ and from $\boldsymbol{r}_T^{(k)}$ to $\hat{\boldsymbol{\beta}}_T^{(k+1)}$, namely $\hat{\boldsymbol{\beta}}_T^{(k)} \overset{h}{\mapsto} \boldsymbol{r}_T^{(k)}$, $\boldsymbol{r}_T^{(k)} \overset{g}{\mapsto} \hat{\boldsymbol{\beta}}_T^{(k+1)}$. Suppose that the largest eigenvalue of $I - \tau^2 A_{T+1}$ in absolute value is ξ ($\xi < 1$). We shall prove that

$$\|g(h(\boldsymbol{\chi}_1)) - g(h(\boldsymbol{\chi}_2))\|_2 \le \xi \|\boldsymbol{\chi}_1 - \boldsymbol{\chi}_2\|_2. \tag{23}$$

It suffices to prove that $\|h(\boldsymbol{\alpha}_1) - h(\boldsymbol{\alpha}_2)\|_2 \le \xi \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2$ and $\|g(\boldsymbol{\chi}_1) - g(\boldsymbol{\chi}_2)\|_2 \le \|\boldsymbol{\chi}_1 - \boldsymbol{\chi}_2\|_2$ for any vectors $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\chi}_1, \boldsymbol{\chi}_2$. The first inequality follows directly from the definition of $\boldsymbol{r}^{(k)}$ in the E step, and $h(\boldsymbol{\alpha}_1) - h(\boldsymbol{\alpha}_2) = (I - \tau^2 A_T)(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)$. To prove the second inequality, we prove

$$\|g(\boldsymbol{\chi}_{1,i}) - g(\boldsymbol{\chi}_{2,i})\|_2 \le \|\boldsymbol{\chi}_{1,i} - \boldsymbol{\chi}_{2,i}\|_2, \tag{24}$$

where $\boldsymbol{\chi}_{k,i}$ ($i = 1, \ldots, L$) are subvectors (groups) of corresponding to $\hat{\boldsymbol{\beta}}_{T,i}^{(k)}$ for either $k = 1$ or $k = 2$. For brevity we define $\tilde{\tau} = \lambda_T \tau_T^2$. We prove (24) by considering three possible cases: 1) $\|\boldsymbol{\chi}_{1,i}\|_2, \|\boldsymbol{\chi}_{2,i}\|_2 \ge \tilde{\tau}$; 2) one of $\|\boldsymbol{\chi}_{1,i}\|_2$ and $\|\boldsymbol{\chi}_{2,i}\|_2$ is less than $\tilde{\tau}$ while the other is no less than $\tilde{\tau}$; 3) $\|\boldsymbol{\chi}_{1,i}\|_2, \|\boldsymbol{\chi}_{2,i}\|_2 < \tilde{\tau}$. For case 1), $g(\boldsymbol{\chi}_{1,i}) = g(\boldsymbol{\chi}_{2,i}) = \mathbf{0}$ and (24) trivially holds. For case 2), assume without loss of generality that $\|\boldsymbol{\chi}_{2,i}\|_2 < \tilde{\tau}$. Then

$$\begin{aligned}\|g(\boldsymbol{\chi}_{1,i}) - g(\boldsymbol{\chi}_{2,i})\|_2 &= \|g(\boldsymbol{\chi}_{1,i})\|_2 = \|\boldsymbol{\chi}_{1,i}\|_2 - \tilde{\tau} \\ &\le \|\boldsymbol{\chi}_{1,i}\|_2 - \|\boldsymbol{\chi}_{2,i}\|_2 \le \|\boldsymbol{\chi}_{1,i} - \boldsymbol{\chi}_{2,i}\|_2.\end{aligned}$$

For case 3), we note that $g(\boldsymbol{\chi}_{k,i})$ is in the same direction of $\boldsymbol{\chi}_{k,i}$ for $k = 1, 2$. We define the angle between $\boldsymbol{\chi}_{1,i}$ and $\boldsymbol{\chi}_{2,i}$ to be θ, and let $a = \|\boldsymbol{\chi}_{1,i}\|$, $b = \|\boldsymbol{\chi}_{2,i}\|$. By the Law of Cosines, to prove $\|g(\boldsymbol{\chi}_1) - g(\boldsymbol{\chi}_2)\|_2^2 \le \|\boldsymbol{\chi}_1 - \boldsymbol{\chi}_2\|_2^2$ it suffices to prove that

$$\begin{aligned}&(a - \tilde{\tau})^2 + (b - \tilde{\tau})^2 - 2(a - \tilde{\tau})(b - \tilde{\tau})\cos(\theta) \\ &\le a^2 + b^2 - 2ab\cos(\theta).\end{aligned} \tag{25}$$

By elementary calculations, Inequality (25) is equivalent to $2\{1 - \cos(\theta)\}\{(a + b)\tilde{\tau} - \tilde{\tau}^2)\} \ge 0$, which is straightforward.

Finally, Inequality (23) and Banach Fixed Point Theorem imply that there exists a *unique* fixed point $\hat{\boldsymbol{\beta}}_T$ and,

$$\|\hat{\boldsymbol{\beta}}_T^{(k)} - \hat{\boldsymbol{\beta}}_T\|_2 \le \frac{\xi^k}{1 - \xi} \|\hat{\boldsymbol{\beta}}_T^{(1)} - \hat{\boldsymbol{\beta}}_T^{(0)}\|_2$$

which decays exponentially in k for any given initial value $\hat{\boldsymbol{\beta}}_T^{(0)}$.

Moreover, the fixed point $\hat{\boldsymbol{\beta}}_T$ is MAP, because each EM iteration increases the value in (10) *implicitly* by increasing the value in $Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_T^{(k)})$ (see the justification of EM algorithm [35], [36]).

PROOF OF THEOREM 2

The proof follows standard techniques in high-dimensional regression settings [5], [30]. We only sketch the proof below. For brevity, $\hat{\boldsymbol{\beta}}_T$ and $\hat{\boldsymbol{\beta}}_{T,d}$ are denoted as $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_d$, respectively.

Let $\tilde{S}_1 = S_0 \cup S_1$ be the set union of truly nonzero set of coefficients and the selected nonzero coefficients. By the definition of $\tilde{S}_1$, we have

$$\begin{aligned}&\|Y - Z_{\tilde{S}_1} \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 + \tilde{\lambda} \sum_{d \in \tilde{S}_1} \|\hat{\boldsymbol{\beta}}_d\|_2 \\ &\le \|Y - Z_{\tilde{S}_1} \boldsymbol{\beta}_{\tilde{S}_1}\|_2^2 + \tilde{\lambda} \sum_{d \in \tilde{S}_1} \|\boldsymbol{\beta}_d\|_2.\end{aligned} \tag{26}$$

Define $\boldsymbol{\rho} = Y - Z\boldsymbol{\beta}$, and $\boldsymbol{\psi} = Z_{\tilde{S}_1}(\hat{\boldsymbol{\beta}}_{\tilde{S}_1} - \boldsymbol{\beta}_{\tilde{S}_1})$. We obtain

$$\begin{aligned}\|\boldsymbol{\psi}\|_2^2 &\le 2\boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\rho} + \tilde{\lambda} \sum_{d \in \tilde{S}_1} (\|\boldsymbol{\beta}_d\|_2 - \|\hat{\boldsymbol{\beta}}_d\|_2) \\ &\le 2\boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\rho} + \tilde{\lambda} \sum_{d \in S_0} (\|\boldsymbol{\beta}_d\|_2 - \|\hat{\boldsymbol{\beta}}_d\|_2) \\ &\le 2\boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\rho} + \tilde{\lambda}\sqrt{|S_0|}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2 \\ &\le 2\boldsymbol{\psi}^{\mathrm{T}} \boldsymbol{\rho} + \tilde{\lambda}\sqrt{|S_1|}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2 \\ &\le 2\|\boldsymbol{\psi}\|_2 \|\boldsymbol{\rho}\|_2 + \tilde{\lambda}\sqrt{|S_1|}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2\end{aligned}$$

where the first inequality is rewritten from (26), the second and fourth follow from $S_0 \subseteq \tilde{S}_1$, the third and fifth follow from Cauchy inequality. From the above equality and $2\|\boldsymbol{\psi}\|_2\|\boldsymbol{\rho}\|_2 \le \|\boldsymbol{\psi}\|_2^2/2 + 2\|\boldsymbol{\rho}\|_2^2$, we obtain

$$\|\boldsymbol{\psi}\|_2^2 \le 4\|\boldsymbol{\rho}\|_2^2 + 2\tilde{\lambda}\sqrt{|S_1|}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2. \tag{27}$$

On the other hand, Assumption 4 gives $\|\boldsymbol{\psi}\|_2^2 \ge \kappa T \|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2$. Therefore,

$$\begin{aligned}\kappa T \|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 &\le 4\|\boldsymbol{\rho}\|_2^2 + 2\tilde{\lambda}\sqrt{|S_1|}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2 \\ &\le 4\|\boldsymbol{\rho}\|_2^2 + \frac{2\tilde{\lambda}^2|S_1|}{\kappa T} + \frac{\kappa T}{2}\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2\end{aligned}$$

which implies that

$$\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 \le 8\|\boldsymbol{\rho}\|_2^2/(\kappa T) + 4\tilde{\lambda}^2|S_1|/(\kappa T)^2. \tag{28}$$

In order to bound $\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2$, it remains to bound $\|\boldsymbol{\rho}\|_2$. Since ρ_t can be written as

$$\varepsilon_t + \sum_{d \in \tilde{S}_1} \{f_d(X_{d,t}) - f_d^*(X_{d,t})\} + (\mu - \bar{Y}),$$

where $(\mu - \bar{Y}) = O_p(T^{-1})$ and $\|f_d - f_d^*\|_\infty = O(v^{-p} + v^{1/2}T^{-1/2})$ [5, Lemma 1], we obtain $\|\boldsymbol{\rho}\|_2^2 \le 2\|\boldsymbol{\varepsilon}\|_{P_X}^2 + c_2 T v^{-2p} + O_p(1)$ for sufficiently large T, where c_2 is a constant that does not depend on v, and P_X is the projection matrix of $Z_{\tilde{S}_1}$. On the other side,

$$\|\varepsilon\|_{P_X}^2 \le \|Z_{\tilde{S}_1}^{\mathrm{T}}\boldsymbol{\varepsilon}\|_2^2/(\kappa T).$$

Therefore,

$$\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 \le 8c_2 v^{-2p}/\kappa + O(T^{-2}\|Z_{\tilde{S}_1}^{\mathrm{T}}\boldsymbol{\varepsilon}\|_2^2) + O_p(T^{-1}) + O(T^{-2}\tilde{\lambda}^2).$$

To finish the proof of (22), it remains to prove that $\|Z_{\tilde{S}_1}^{\mathrm{T}}\boldsymbol{\varepsilon}\|_2^2 = O_p(T\log\tilde{D})$. Note that the elements of ε are not i.i.d. conditioning on $Z_{\tilde{S}_1}$ due to time series dependency, which is different from the usual regression setting. However, for any of the $|S_1|v$ column of $Z_{\tilde{S}_1}$, say $\boldsymbol{z}_{d,j}$, the inner product $\boldsymbol{z}_{d,j}^{\mathrm{T}}\boldsymbol{\varepsilon} = \sum_{t=1}^T z_{d,j,t}\varepsilon_t$ is the sum of a martingale difference sequence (MDS) with sub-exponential condition. Applying the Bernstein-type bound for a MDS, we obtain for all $w > 0$ that

$$\mathrm{pr}\left(\left|\sum_{t=1}^T z_{d,j,t}\varepsilon_t\right| > w\right) \le 2\exp\left\{-w^2/(2\sum_{t=1}^T \eta_t)\right\}, \text{ where}$$

$$\eta_t \triangleq var z_{d,j,t}\varepsilon_t \le z_{d,j,t}^2\sigma^2 \le \sup_{x\in[a,b]}\{b_{d,j}(x)\}^2\sigma^2.$$

Thus, $\sum_{t=1}^T z_{d,j,t}\varepsilon_t$ is a sub-Gaussian random variable for each d, j. By applying similar techniques used in the maximal inequality for Gaussian random variables [37],

$$\max_{d\in\tilde{S}_1, 1\le j\le v} E(T^{-1/2}\boldsymbol{z}_{d,j}^{\mathrm{T}}\boldsymbol{\varepsilon}) \le O(T^{-1/2}(\log\tilde{D})^{1/2}).$$

Therefore,

$$\|Z_{\tilde{S}_1}^{\mathrm{T}}\boldsymbol{\varepsilon}\|_2^2 \le |S_1|vT \max_{d\in\tilde{S}_1, 1\le j\le v}\{E(T^{-1/2}\boldsymbol{z}_{d,j}^{\mathrm{T}}\boldsymbol{\varepsilon})\}^2 \le O_p(T\log\tilde{D}).$$

To prove $\lim_{T\to\infty}\mathrm{pr}(S_0 \subseteq S_1) = 1$, we define the event E_0 as "There exists $d \in S_0$ such that $\hat{\boldsymbol{\beta}}_d = 0$ and $\boldsymbol{\beta}_d \neq 0$". Under event E_0, let d satisfy the above requirement. Since $\|f_d - f_d^*\|_\infty = O(v^{-p} + v^{1/2}T^{-1/2})$, there exists a constant $c_1^{'}$ such that for all $v \ge c_1^{'}c_0^{-1/p}$ and sufficiently large T, $\|f_d^*\|_2 \ge c_0/2$. By a result from [38], $\|\boldsymbol{\beta}_d\|_2^2/v \ge c_2^{'}\|f_d^*\|_2^2$ holds for some constant $c_2^{'}$. Then, under E_0 it follows that $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_2^2 \ge \|\boldsymbol{\beta}_d\|_2^2 \ge c_2^{'}vc_0^2/4 \ge 16c_2v^{-2p}/\kappa$ for all $v \ge c_1^{''}c_0^{-2/(2p+1)}$, where $c_1^{''}$ is some positive constant. This contradicts the bound given in (22) for large T.

Proof of Theorem 3

Recall that the backward selection procedure produces a nested sequence of subsets $S_2 = \mathcal{S}^{(K)} \subseteq \cdots \subseteq \mathcal{S}^{(1)} \subseteq \mathcal{S}^{(0)} = S_1$ with corresponding MSE $\hat{e}^{(k)}$ $(k = 0, \ldots, K)$, where $0 \le K \le |S_1| - |S_2|$. In addition, $\mathcal{S}^{(k)} = \mathcal{S}^{(k-1)} - \{\bar{d}_k^*\}$ for some $\bar{d}_k^* \in \mathcal{S}^{(k-1)}$. It suffices to prove that as T goes to infinity, with probability going to one i) $S_0 \subseteq \mathcal{S}^{(k)}$ for each $k = 0, \ldots, K$, and ii) $|S_2| = |S_0|$.

Following a similar proof by [27, Proof of Theorem 1], it can be proved that for any k, conditioned on $S_0 \subseteq \mathcal{S}^{(k-1)}$, we have $\hat{e}^{(k-1)} - \hat{e}^{(k)} = O_p(v_T/T)$ if $S_0 \subseteq \mathcal{S}^{(k-1)}$, and $\hat{e}^{(k-1)} - \hat{e}^{(k)} = c + o_p(1)$ for some constant $c > 0$ if $S_0 \not\subseteq \mathcal{S}^{(k-1)}$. Note that the penalty increment $(v_T\log T)/T$ is larger than $O_p(v_T/T)$ and smaller than $c + o_p(1)$ for large T. By successive application of this fact finitely many times, we can prove that $S_0 \subseteq \mathcal{S}^{(k)}$ for each $k = 0, \ldots, K$, and that $|S_2| = |S_0|$ with probability close to one.

Derivation of Equation (12) in SLANTS

We need to compute

$$Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_T^{(k)}) = E_{\boldsymbol{\theta}_T|(\hat{\boldsymbol{\beta}}_T^{(k)}, \boldsymbol{Y}_T)} \log p(\boldsymbol{Y}_T, \boldsymbol{\theta}_T \mid \boldsymbol{\beta}_T) - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_i\|_2$$

up to a constant (which does not depend on $\boldsymbol{\beta}$). The complete log-likelihood is

$$\log p(\boldsymbol{Y}_T, \boldsymbol{\theta}_T \mid \boldsymbol{\beta}) = C_0 - \frac{\|\boldsymbol{\theta}_T - \boldsymbol{\beta}\|^2}{2\tau_T^2} = C_1 - \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\theta}_T}{2\tau_T^2},$$

where C_1 and C_2 are constants that do not involve $\boldsymbol{\beta}$. So it remains to calculate $E_{\boldsymbol{\theta}_T|(\hat{\boldsymbol{\beta}}_T^{(k)}, \boldsymbol{Y}_T)}\boldsymbol{\theta}_T$. Note that $\boldsymbol{Y}_T|\boldsymbol{\theta}_T \sim N(Z_T\boldsymbol{\theta}_T, W_T^{-1} - \tau_T^2 Z_T Z_T^{\mathrm{T}})$, $\boldsymbol{\theta}_T|\hat{\boldsymbol{\beta}}_T^{(k)} \sim N(\hat{\boldsymbol{\beta}}_T^{(k)}, \tau_T^2 I)$. Thus, $\boldsymbol{\theta}_T \mid (\hat{\boldsymbol{\beta}}_T^{(k)}, \boldsymbol{Y}_T)$ is Gaussian with mean

$$E_{\boldsymbol{\theta}_T|(\hat{\boldsymbol{\beta}}_T^{(k)}, \boldsymbol{Y}_T)}\boldsymbol{\theta}_T = \boldsymbol{r}^{(k)}.$$

It follows that

$$Q(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}_T^{(k)}) = -\frac{1}{2\tau_T^2}\|\boldsymbol{\beta} - \boldsymbol{r}^{(k)}\|_2^2 - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_i\|_2.$$

Acknowledgement

The authors thank Dr. Lu Shen for suggesting the Boston weather data. The authors also thank associate editor Dr. Morten Mørup and three anonymous reviewers for their reviewing the paper and providing insightful comments.

References

[1] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Adv. Neural. Inf. Process. Syst.*, 2015, pp. 802–810.

[2] S. Yang, M. Santillana, and S. Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 112, no. 47, pp. 14 473–14 478, 2015.

[3] S. Vijayakumar, A. D'souza, and S. Schaal, "Incremental online learning in high dimensions," *Neural computation*, vol. 17, no. 12, pp. 2602–2634, 2005.

[4] J. H. Friedman, "Multivariate adaptive regression splines," *Ann. Stat.*, pp. 1–67, 1991.

[5] J. Huang, J. L. Horowitz, and F. Wei, "Variable selection in nonparametric additive models," *Ann. Stat.*, vol. 38, no. 4, p. 2282, 2010.

[6] H. Tong, *Threshold models in non-linear time series analysis*. Springer Science & Business Media, 2012, vol. 21.

[7] C. Gouriéroux, *ARCH models and financial applications*. Springer Science & Business Media, 2012.

[8] T. J. Hastie and R. J. Tibshirani, *Generalized additive models*. CRC Press, 1990, vol. 43.

[9] Z. Cai, J. Fan, and Q. Yao, "Functional-coefficient regression models for nonlinear time series," *J. Amer. Statist. Assoc.*, vol. 95, no. 451, pp. 941–956, 2000.

[10] K. Zhang and A. Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *UAI*. AUAI Press, 2009, pp. 647–655.

[11] K. Zhang, J. Peters, D. Janzing, and B. Schölkopf, "Kernel-based conditional independence test and application in causal discovery," *arXiv preprint arXiv:1202.3775*, 2012.

[12] J. Fan, Y. Feng, and R. Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *J. Amer. Statist. Assoc.*, 2012.

[13] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *J. R. Stat. Soc. Ser. B*, pp. 267–288, 1996.

[14] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, "Sparse additive models," *J. Roy. Statist. Soc. Ser. B*, vol. 71, no. 5, pp. 1009–1030, 2009.

[15] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Group-lasso on splines for spectrum cartography," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4648–4663, 2011.

[16] G. Wahba, *Spline models for observational data*. Siam, 1990, vol. 59.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[18] H. Zou, "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[19] M. A. Figueiredo and R. D. Nowak, "An em algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, 2003.

[20] B. Babadi, N. Kalouptsidis, and V. Tarokh, "Sparls: The sparse rls algorithm," *IEEE Trans. Signal Process.*, vol. 58, no. 8, pp. 4013–4025, 2010.

[21] G. Mileounis, B. Babadi, N. Kalouptsidis, and V. Tarokh, "An adaptive greedy algorithm with application to nonlinear communications," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 2998–3007, 2010.

[22] H. T. Friedman, J. and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Stat. Softw.*, vol. 33, no. 1, pp. 1–22, 2008.

[23] A. P. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *J. Roy. Statist. Soc. Ser. A*, pp. 278–292, 1984.

[24] T. Park and G. Casella, "The bayesian lasso," *J. Amer. Statist. Assoc.*, vol. 103, no. 482, pp. 681–686, 2008.

[25] L. Bornn, A. Doucet, and R. Gottardo, "An efficient computational approach for prior sensitivity analysis and cross-validation," *Can J. Stat.*, vol. 38, no. 1, pp. 47–64, 2010.

[26] D. L. Jupp, "Approximation to data by splines with free knots," *SIAM. J. Numer. Anal.*, vol. 15, no. 2, pp. 328–343, 1978.

[27] J. Z. Huang and L. Yang, "Identification of non-linear additive autoregressive models," *J. Roy. Statist. Soc. Ser. B*, vol. 66, no. 2, pp. 463–477, 2004.

[28] C. J. Stone, "Additive regression and other nonparametric models," *Ann. Stat.*, pp. 689–705, 1985.

[29] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.

[30] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the LASSO and generalizations*. CRC Press, 2015.

[31] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.

[32] Y. Yang and H. Zou, "A fast unified algorithm for solving group-lasso penalize learning problems," *Statistics and Computing*, vol. 25, no. 6, pp. 1129–1141, 2015.

[33] V. Roth and B. Fischer, "The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *ICML*. ACM, 2008, pp. 848–855.

[34] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: a new criterion for autoregression," *IEEE Trans. Inf. Theory*, 2017.

[35] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, pp. 1–38, 1977.

[36] C. J. Wu, "On the convergence properties of the em algorithm," *Ann. Stat.*, pp. 95–103, 1983.

[37] A. W. Van Der Vaart and J. A. Wellner, *Weak Convergence*. Springer, 1996.

[38] C. D. Boor, *A practical guide to splines*. Springer-Verlag New York, 1978, vol. 27.

Qiuyi Han (hqychr@gmail.com) received the B.S. in Mathematics and Physics from Tsinghua University in 2012. She is a Ph.D. student in the department of Statistics at Harvard University. Her research interests are statistical network analysis, high dimensional statistics and machine learning.

Jie Ding (jieding@fas.harvard.edu) received Bachelor of Science degree from Tsinghua University in May 2012, majoring in Mathematics and Electrical Engineering. He received Master of Arts degree in Statistics in May 2016, and Ph.D. degree in Engineering Sciences in March 2017, both from Harvard University. His research areas are statistical inference, machine learning, signal processing, and combinatorics. His recent goal is to establish a reliable, efficient, and widely applicable time series prediction system.

Edoardo M. Airoldi (airoldi@fas.harvard.edu) is an Associate Professor of Statistics at Harvard University, where he has been directing the Harvard Laboratory for Applied Statistical Methodology & Data Science since 2009. He holds a Ph.D. in Computer Science and an M.Sc. in Statistics from Carnegie Mellon University, and a B.Sc. in Mathematical Statistics and Economics from Bocconi University. His current research focuses on statistical theory and methods for designing and analyzing experiments on large networks, and on modeling and inferential issues that arise in analyses that leverage network data. His work has appeared in journals across statistics, computer science and engineering, including Annals of Statistics, Journal of the American Statistical Association, Journal of Machine Learning Research, Proceedings of the National Academy of Sciences, and Nature. He is the recipient of several research awards including an Alfred Sloan Research Fellowship and a Shutzer Fellowship from the Radcliffe Institute of Advanced Studies. He delivered an IMS Medallion Lecture at JSM 2017 in Baltimore.

Vahid Tarokh (vahid@seas.harvard.edu) received the Ph.D. in electrical engineering from the University of Waterloo, Ontario, Canada in 1995. He then worked at AT&T Labs-Research and AT&T wireless services until August 2000 as Member, Principal Member of Technical Staff, and finally as the Head of the Department of Wireless Communications and Signal Processing. In Sept 2000, he joined MIT as an Associate Professor where he worked until June 2002. In June 2002, he joined Harvard University as a Professor of Electrical Engineering. He was named Perkins Professor and Vinton Hayes Senior Research Fellow of Electrical Engineering in 2005. His current research areas are in statistical signal processing and data analysis.