MODELING NONLINEARITY IN MULTI-DIMENSIONAL DEPENDENT DATA

Qiuyi Han, Jie Ding, Edoardo Airoldi, and Vahid Tarokh

Harvard University, Cambridge, MA, 02138 USA

ABSTRACT

Given massive data that may be time dependent and multi-dimensional, how to efficiently explore the underlying functional relationships across different dimensions and time lags? In this work, we propose a methodology to sequentially and adaptively model nonlinear multivariate time series data. Data at each time step and dimension is modeled as a nonlinear function of past values corrupted by noise, and the underlying nonlinear 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 bounds of the prediction errors.

***Index Terms*—** Adaptive Filtering, Group LASSO, Nonlinear Models, Spline Regression, Time Series.

1. INTRODUCTION

Sequentially observed vector time series are emerging in various applications. In most these applications modeling nonlinear functional inter-dependency between present and past data is crucial for both representation and prediction. This is a challenging problem given that often in various applications fast online implementation, adaptivity and ability to handle high dimensions are basic requirements for 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 vector time series.

Relation to prior work: 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.

Contribution: 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

This work is supported by Defense Advanced Research Projects Agency (DARPA) grant numbers W911NF-14-1-0508 and N66001-15-C-4028.

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 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.

2. 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 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 vector time series applications.

2.1. Formulation of SLANTS

Consider a multi-dimensional vector 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.

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, μ is a constant, 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 [14]). In our presentation, for brevity, we consider the additive model mainly but note that 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 straightforward manner.

Incorporating the B-spline basis into regression, we write

$$f_i(x) = \sum_{j=1}^{v} c_{i,j} b_{i,j}(x),\ b_{i,j}(x) = B(x \mid s_{i,1}, \ldots, s_{i,v}) \tag{5}$$

where $s_{i,1}, \ldots, s_{i,v}$ are the knots and $c_{i,j}$ are the coefficients associated with the B-spline basis. Here, we have assumed that there are v spline basis of degree k for each f_i. 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 $b_{i,j}(x_{i,t}) - \frac{1}{T}\sum_{t=1}^{T} b_{i,j}(x_{i,t})$. 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)

$$\log p(\boldsymbol{Y}_T \mid \boldsymbol{\beta}_T, Z_T) - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 \tag{10}$$

$$= -\frac{1}{2} \sum_{t=1}^{T} w_{T,t} (Y_t - \boldsymbol{z}_t^{\mathrm{T}} \boldsymbol{\beta}_T)^2 - \lambda_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 + c$$

where c is a constant that depends only on W_T. The above prior corresponds to the so called group LASSO. 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 [15], 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.

2.2. 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 [16]. This was further applied to online adaptive filtering for sparse linear models [17] and nonlinear models approximated by Volterra series [18, 19]. 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 expectationmaximization (EM) algorithm can be derived. By basic calculations similar to that of [16], we obtain the kth step/iteration 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

$$\boldsymbol{r}^{(k)} = (I - \tau_T^2 A_T) \hat{\boldsymbol{\beta}}_T^{(k)} + \tau_T^2 B_T,$$
$$A_T = Z_T^{\mathrm{T}} W_T Z_T, \ \boldsymbol{B}_T = Z_T^{\mathrm{T}} W_T \boldsymbol{Y}_T.$$

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

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

$$\hat{\boldsymbol{\beta}}_{T,i}^{(k+1)} = \left[1 - \frac{\lambda_T \tau_T^2}{\|\boldsymbol{r}_i^{(k)}\|_2}\right]_+ \boldsymbol{r}_i^{(k)}, \quad i = 1, \ldots, \tilde{D}. \tag{13}$$

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 $T+1$:

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

where

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

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

Remark 1 *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 [20] 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 $O(\tilde{D}^2 T)$.*

Remark 2 *Here, we provide a short 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

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

In the following Theorem, we show that EM can converge exponentially fast to the MAP of (10).

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.*

3. THEORETICAL RESULTS

Consider the harmonic step size $\gamma_t = 1/t$. We assume that L is fixed while D is increasing with sample size T at certain rate. Following the setup of [21], 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.*

Suppose that a practitioner aims to discover the significant variable set with probability close to one as more data is collected. Our approach is to minimize the objective function in (10), and it can be efficiently implemented using the proposed sequential algorithm in Section 2.2 with negligible error (Theorem 1). In the case of equal weights $w_{T,t} = 1/T$, it can be rewritten as

$$\|Y_T - Z_T\boldsymbol{\beta}_T\|_2^2 + \tilde{\lambda}_T \sum_{i=1}^{\tilde{D}} \|\boldsymbol{\beta}_{T,i}\|_2 \tag{14}$$

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 (14) with appropriately chosen $\tilde{\lambda}_T$, i.e., $S_0 \subseteq S_1$.

Assumption 2 *There is a positive constant c_0 such that $\min_{d \in S_0} \|f_d\|_2 \geq c_0$. 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 3 *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).*

Theorem 2 *Suppose that Assumptions 1-3 hold. Then for any given v it holds that*

$$\|\boldsymbol{\beta}_{\tilde{S}_1} - \hat{\boldsymbol{\beta}}_{\tilde{S}_1}\|_2^2 \leq 8c_2 v^{-2p}/\kappa + O_p(T^{-1}\log\tilde{D}) + O_p(T^{-1}) + O(T^{-2}\tilde{\lambda}^2) \tag{15}$$

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 $v > c_1 c_0^{-1/p} \max\{1, c_0^{-\frac{1}{p(2p+1)}}\}$, $\lim_{T\to\infty} \mathrm{pr}(S_0 \subseteq S_1) = 1$.

The proofs will appear in a journal version of the work.

4. REFERENCES

[1] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[2] Shihao Yang, Mauricio Santillana, and SC Kou, "Accurate estimation of influenza epidemics using google search data via argo," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14473–14478, 2015.

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

[4] Jerome H Friedman, "Multivariate adaptive regression splines," *The annals of statistics*, pp. 1–67, 1991.

[5] Jian Huang, Joel L Horowitz, and Fengrong Wei, "Variable selection in nonparametric additive models," *Annals of statistics*, vol. 38, no. 4, pp. 2282, 2010.

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

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

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

[9] Zongwu Cai, Jianqing Fan, and Qiwei Yao, "Functional-coefficient regression models for nonlinear time series," *Journal of the American Statistical Association*, vol. 95, no. 451, pp. 941–956, 2000.

[10] Kun Zhang and Aapo Hyvärinen, "On the identifiability of the post-nonlinear causal model," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 647–655.

[11] Kun Zhang, Jonas Peters, and Dominik Janzing, "Kernel-based conditional independence test and application in causal discovery," in *In Uncertainty in Artificial Intelligence*. Citeseer, 2011.

[12] Jianqing Fan, Yang Feng, and Rui Song, "Nonparametric independence screening in sparse ultra-high-dimensional additive models," *Journal of the American Statistical Association*, 2012.

[13] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[14] Grace Wahba, *Spline models for observational data*, vol. 59, Siam, 1990.

[15] Hui Zou, "The adaptive lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[16] Mário AT Figueiredo and Robert D Nowak, "An em algorithm for wavelet-based image restoration," *Image Processing, IEEE Transactions on*, vol. 12, no. 8, pp. 906–916, 2003.

[17] Behtash Babadi, Nicholas Kalouptsidis, and Vahid Tarokh, "Sparls: The sparse rls algorithm," *Signal Processing, IEEE Transactions on*, vol. 58, no. 8, pp. 4013–4025, 2010.

[18] Gerasimos Mileounis, Behtash Babadi, Nicholas Kalouptsidis, and Vahid Tarokh, "An adaptive greedy algorithm with application to nonlinear communications," *IEEE Transactions on Signal Processing*, vol. 58, no. 6, pp. 2998–3007, 2010.

[19] Gerasimos Mileounis, Behtash Babadi, Nicholas Kalouptsidis, and Vahid Tarokh, "An adaptive greedy algorithm with application to sparse narma identification.," in *ICASSP*, 2010, pp. 3810–3813.

[20] Hastie T. Friedman, J. and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2008.

[21] Charles J Stone, "Additive regression and other nonparametric models," *The annals of Statistics*, pp. 689–705, 1985.