Iterative Stepwise Selection and Threshold for Learning Causes in Time Series

Jianxin Yin JIANXINYIN@MATH.PKU.EDU.CN
Shaopeng Wang WANGSHOP@GMAIL.COM
Wanlu Deng SHIRLEYPKU@GMAIL.COM
Ya Hu TERESAHU@PKU.EDU.CN
Zhi Geng ZGENG@MATH.PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing 100871, China

Abstract

When we explore the causal relationship among time series variables, we first remove the potential seasonal term then we deal with the problem in the feature selection framework. For a time series with seasonal term, we use several sequences of $sin(t)$ and $cos(t)$ functions with different frequencies to design a 'pseudo' design matrix, and the seasonal term is removed by getting the regression residual of the original series on this 'pseudo' design matrix. An iterative stepwise subset selection and threshold method are then applied. The cut-value for the threshold is selected by an EBIC criterion. Some simulations are performed to assess our method. In the PROMO task of the Potluck challenge, we apply our method and obtain a specificity of above 77% while keep the sensitivity of around 89% on the PROMO task.

Keywords: functional data, iterative threshold, linear model, seasonal term, stepwise subset selection, structural equation model, time series.

1. Introduction

In the Potluck challenge, we try to select the causal processes from the 1000 promotions for each product sales series separately. For a time series $Y(t)$ with seasonal terms, we can decompose it into three parts:

$$Y(t) = S(t) + T(t) + N(t) \tag{1}$$

where $S(t)$ denotes the seasonal term, $T(t)$ denotes the trend term which may be influenced by the other processes and $N(t)$ is the noise part. There exist many methods to model the seasonal term $S(t)$ in the literature(see Brockwell and Davis (1991), Box et al. (1994)). Here we treat $S(t)$ as a continuous periodic function and approximate it by a series of periodic functions bases in the *sin* and *cos* functions. We use a linear structure equation model(SEM) to model the other processes' causal influence on the target process. That is,

$$T(t) = \beta_0 + \beta_1 * x_1(t) + \cdots + \beta_p * x_p(t) \tag{2}$$

where $x_i(t)$, $i = 1, \cdots, p$ stands for the other processes which may also be influenced by the seasonal factor. So we also apply the same model of seasonal term on each $x_i(t)$ process. Equation (2) is then treated as a simple linear regression model and a stepwise selection(Weisberg (1985)) procedure is applied to screening out the influential independent variables. This stepwise selection is applied iteratively to eliminate the possible "boundary" variables(see the next section). To get a sparse model, we further use a kind of threshold on the regression coefficients to select the significant subset. This procedure is also applied iteratively to get a converged subset. And the hyper-parameter of the cut-point is selected by an extended BIC criterion. Some simulation study shows that our method can get the consistent result under different kinds of $S(t)$ and $N(t)$ process. This paper is organized as following. In Section 2, the preprocessing for the seasonal term is described, in Section 3 the stepwise selection procedure is mentioned and the iterative threshold method with its hyper-parameter selection method is introduced. Section 4 is the numerical study. Finally the Section 5 gives some discussion on our method.

2. Preprocessing: Filtering the seasonal term

For a given time series, we use a series of continuous periodic functions to filter out the seasonal term. Suppose that the period length is T while in our problem, $T = 365$. Then we generate the periodical sequences $\sin(2\pi t/k)$, $\cos(2\pi t/k)$, for $t = 1, \cdots, 1095$ and $k = T$, $T/2$, $T/3$, $T/4$, $T/5$. It is obvious that the period for each sequence is k. Denote a design matrix $Z = [z_1, z_2, \cdots, z_{1095}]^\top$, where

$$z_t = \Big(1 \ \sin(2\pi t/T) \ \cos(2\pi t/T) \cdots \sin(10\pi t/T) \ \cos(10\pi t/T)\Big)^\top$$

$S = \big(S(1), \cdots, S(1095)\big)^\top$ is estimated as $\hat{S} = Z(Z^\top Z)^{-1}Z^\top Y$. Then we remove the seasonal term expressed in the regression value on this design matrix to get the residual as the input of our next analysis. $Y^* = Y - \hat{S}$, $x_i^* = x_i - Z(Z^\top Z)^{-1}Z^\top x_i$ for $i = 1, \cdots, 1000$. Here Y is the realization of 1095 days of certain product sales in our problem and x_i is the realization of 1095 days for some promotion method. To simplify the notation, we still use the Y for Y^* and x_i for x_i^* respectively. The k is selected up to $T/5$ is determined by experience. We assume that the continuous periodic seasonal function can be approximated well enough by its Fourier expansion up to the fifth order.

3. Feature Selection

Since we have reduce our time series problem into a simple linear feature selection problem after removing the seasonal term, we can omit the subindex t and write our model

$$y = X_p\beta_p + \varepsilon \tag{3}$$

where p is the dimension of the original feature space.

3.1 Iterative stepwise subset selection

We use the stepwise selection(SW) to select the influential $x_i(t)$s for each $Y(t)$ through relation (1) and (2). When the significant level for entering(*penter*) is different from the one

for removing($premove$), there exists certain situation that some features on the "boundary" (here we mean that the significant level is between the $penter$ and $premove$) can be dropped in the next round of stepwise selection on the remained feature set. For example, if x_2 is only significant for the response when x_1 is in the model; suppose x_1 enters first, then x_2 can enter, but later x_1 is removed and x_2 is not removed (if it is on the "boundary"). For the next round of SW selection, x_2 will be removed. So for the purpose of sparsity, we use the $stepwisefit$ procedure in Matlab iteratively to select the feature set with $penter = 0.05$ and $premove = 0.1$.

3.2 Iterative threshold selection

Before the following analysis, each column of X_p is standardized to have zero mean and unit variance. Suppose that the true model has a dimension d and denoted as X_d, then the above relation (3) can be represented as

$$y = X_d\beta_d^* + \varepsilon \tag{4}$$

where $\beta_d^* = \{\beta_j : \beta_j \neq 0, 1 \leq j \leq p\}$ with a dimension $d < p$. Initialize $\mathcal{M}^{(0)}$ as the output of the iterative SW selection. $\mathcal{M}^{(i)}$ is obtained in an iteratively manner:

$$\mathcal{M}^{(i)} = \left\{1 \leq j \leq \| \mathcal{M}^{(i-1)} \|: |\hat{\beta}_j^{(i-1)}| \geq \alpha * \max_{1\leq k\leq \|\mathcal{M}^{(i-1)}\|}(|\hat{\beta}_k^{(i-1)}|)\right\} \tag{5}$$

where $\hat{\beta}^{(i-1)}$ is the least square estimate for regression coefficient vector of y on $X_{\mathcal{M}^{(i-1)}}$ and $\| \cdot \|$ denotes the cardinality of a set. Intuitively, we drop those features whose absolute values of regression coefficients are smaller than $\alpha * 100\%$ of the current largest one (in absolute value).

Denote the true feature set as $\mathcal{M}_T = \{1 \leq j \leq p : \beta_j \neq 0\}$. And use the note $|\beta|_{min} = \min_{1\leq j\leq p} |\beta_j|$. In order to justify our selection procedure, we need the following two assumptions on the underlying model.

- *Assumption1* There exists a constant number $c_0 > 0$, such that $|\beta^*|_{min}/|\beta^*|_{max} \geq c_0$
- *Assumption2* For any sub-model of the true model $\mathcal{M}^s \subset \mathcal{M}_T$, $|\beta^s|_{min}/|\beta^s|_{max} \geq c_0$, where c_0 is the same constant in assumption 1.

Remark. Assumption 1 says that the ratio of the two extremes of the true coefficients is significantly apart from 0. Assumption 2 want to regulate the behavior of the load on every feature subset. It's not the possible weakest requirement.

Under the above assumptions, we have the following results.

Theorem 1 *Suppose assumptions 1 and 2 are true, under our model setup (3)-(5), then with probability tending to one (as $n \longrightarrow \infty$) that there exists a constant α such that*
(I) If $\mathcal{M}^{(i)} \supsetneq \mathcal{M}_T$, then $\| \mathcal{M}^{(i+1)} \| < \| \mathcal{M}^{(i)} \|$.
(II) If $\mathcal{M}^{(i)} \subseteq \mathcal{M}_T$, then $\| \mathcal{M}^{(i+1)} \| = \| \mathcal{M}^{(i)} \|$.

Proof The α can be chosen as a positive number that $0 < \alpha < c_0$, where $c_0 = |\beta^*|_{min}/|\beta^*|_{max}$. For case (I), since the estimate $\hat{\beta}^{(i)}$ is an unbiased consistent estimator for β_p, then with

probability tending to one $|\hat{\beta}^{(i)}|_{max} \approx |\beta^*|_{max}$ and $|\hat{\beta}^{(i)}|_{min} \approx 0$. Then $|\hat{\beta}^{(i)}|_{min}/|\hat{\beta}^{(i)}|_{max} < \alpha$, so at least we can drop one variable. Similarly, for case (II), with assumption 2, we can see that the same α is also appropriate here for $|\hat{\beta}^{(i)}|_{min}/|\hat{\beta}^{(i)}|_{max} > \alpha$. ■

From the above theorem, one can see that from an over-fit model including the true features, threshold on the regression coefficients can remove the unrelated features and the iteration can repeat this process until it converges to the true model or its subset. And it will not continue to delete variables as long as it is covered by the true subset. But what is the case when we begin from a subset that $\mathcal{M}^{(i)} \not\supseteq \mathcal{M}_T$? There is no assertion can be made here, but from the experience we have, the iteration converges in finite steps under this case.

3.3 Extended-BIC criterion for model selection

Ordinary BIC is likely inconsistent when $p > \sqrt{n}$ (Chen and Chen (2008)). We used the extended-BIC(EBIC) criterion (Chen and Chen (2008)) to select the hyper-parameter α. From the simulation study in the next section, we can see that the EBIC outperforms the ordinary BIC and prediction MSE criterion in measure of SN. The extended-BIC is defined as:

$$EBIC(\mathcal{M}) = log(\hat{\sigma}^2_{(\mathcal{M})}) + n^{-1} \parallel \mathcal{M} \parallel \times(\log n + 2\log p)$$

Then we search the minimum value for this $EBIC(\mathcal{M}_\alpha)$ index by α in an interval. Denote $\hat{\mathcal{M}} = argmin_{\alpha\in[a,b]} EBIC(\mathcal{M}_\alpha)$. In practice, we choose $[a, b] = [0.1, 0.3]$.

4. Numerical Studies

In the simulation study, a linear additive model of (1) is considered. We consider three types of $S(t)$, extra lag-effect in the relationship between $T(t)$ and $x_i(t)$s rather than (2), and an ARMA noise with an acceptable signal-to-noise ratio for $N(t)$.The simulation result shows that our approaches have a good robust performance although we never take into account the lagged effect in $T(t)$ and ARMA in $N(t)$. We use the specificity(shorted as SP) and sensitivity(shorted as SN) to evaluate. To write them out explicitly,

$$SP = \frac{\#\Big\{j : \hat{\beta}_j \neq 0 \ \&\ \beta_j \neq 0\Big\}}{\#\Big\{j : \beta_j \neq 0\Big\}} \qquad SN = \frac{\#\Big\{j : \hat{\beta}_j \neq 0 \ \&\ \beta_j \neq 0\Big\}}{\#\Big\{j : \hat{\beta}_j \neq 0\Big\}}.$$

4.1 Simulation method

We simulate the similar model configurations compared to our PROMO task under model (1). For the $S(t)$, we select three types of periodic continuous function to represent it.

Type(1) $S(t; n, m, \phi_1, \phi_2, T) = \sum_{i=1}^{n} \sin\big(2\pi it/T + \phi_1\big) + \sum_{j=0}^{m} \cos\big(2\pi jt/T + \phi_2\big)$.
Type(2) $S(t; a_1, a_2, a_3, T) = t(T-t)\big[(t-a_1)(t-a_2)(t-a_3) - 200\big]$.
Type(3) Twice moving average of a random $N(0, \sigma^2)$ series with smoothing window width h which is sampled from [50, 80] uniformly.

We generate a total of 50 $S(t)$s for each seasonal type on the domain $t \in [0, 365]$ and then extend them to span in $[0, 1095]$ periodically. The parameters in the three types are randomly assigned except T which is set to 365.

For $T(t)$, firstly we generate 1000 binary series $x_i(t)$ for $t = 1, \cdots, 1095$, $i = 1, \cdots, 1000$ which are potential causes of each $Y(t)$ series through the relation (1) and (2). We generate 500 seasonal $x_i(t)$s(with period 365) and 500 non-seasonal $x_i(t)$s. For each point in $x_i(t)$, a binary number is generated from a binomial distribution $B(p)$ where p is sampled uniformly on $[0.2, 0.8]$. The seasonal one is the triplication of the function defined in $[0, 365]$. Then the coefficients β in (2) are defined as following: the number of non-zero β_j is uniform sampled from $[1, 50]$ while the non-zero values are sampled uniformly from interval $[0.4, 1]$ and assigned a negative sign with probability 0.2. Besides, we suppose covariates $x_i(t)$s may have lagged effect on the target $T(t)$, which often exists in the real world. Inspired by the power decay lagged effect model in (Box et al. (1975)), we use the backward operator $\omega * \sum_{lag}(\frac{1}{2}B)^{(lag-1)}$ on $x_i(t)$s, where ω represents the lagged effect coefficients randomly taking a smaller absolute value than the corresponding major effect but a possible opposite sign. The lag length lag is randomly selected in $\{0, 1, 2, 3, 4, 5\}$. We add the lagged effect to the $T(t)$.

We model the noise term $N(t)$ as an ARMA model whose parameter (p, d) is randomly assigned. Finally,we get the simulated target series $Y(t)$ from the summation of the above three series multiplied with appropriate scale parameters, such that the signal-to-noise rate is around 4 . Then we apply an iterative SW procedure followed by an iterative threshold procedure proposed above to select the significant subset. Different kinds of criterion for α are compared in their accuracy measures.

4.2 Simulation results

Table 1 tells us that the model selected by EBIC is comparably good in SP while is overwhelmingly better than BIC and MSE criterion in SN.

Seasonal Function	**criterion**	**SP**	**SN**
Type (1)	BIC	0.98	0.76
	EBIC	0.96	0.99
	MSE	0.99	0.59
Type (2)	BIC	0.96	0.57
	EBIC	0.92	0.94
	MSE	0.97	0.46
Type (3)	BIC	0.96	0.57
	EBIC	0.93	0.93
	MSE	0.96	0.47

Table 1: Comparison of three criterions.

From Table 2 we can see that the iterative SW process is necessary for there are more than 30% of the case that there are 'boundary' variables in our selected model. Table 3 also

Seasonal Function	1	2	3	$\geqslant 4$
Type (1)	0.7	0.3	0	0
Type (2)	0.58	0.38	0.04	0
Type (3)	0.62	0.36	0.02	0

Table 2: Iteration number distribution for iterative SW process.

Seasonal Function	**criterion**	1	2	3	4	5	6	7	$\geqslant 8$
Type (1)	BIC	0	0.14	0.28	0.22	0.06	0.14	0.04	0.12
	EBIC	0	0.24	0.38	0.24	0.12	0	0	0.02
	MSE	0.38	0.26	0.14	0.08	0.04	0.06	0.02	0.02
Type (2)	BIC	0	0.06	0.26	0.28	0.12	0.08	0.02	0.18
	EBIC	0	0	0.44	0.34	0.16	0.04	0.02	0
	MSE	0.44	0.08	0.24	0.12	0.04	0.02	0.02	0.04
Type (3)	BIC	0	0.04	0.24	0.26	0.16	0.18	0.06	0.06
	EBIC	0	0.04	0.28	0.38	0.1	0.14	0.02	0.04
	MSE	0.42	0.12	0.12	0.16	0.02	0.1	0.04	0.02

Table 3: Iteration number distribution for the iterative Threshold process.

supports that the iteration for Threshold process is necessary although they finally converge with large probability.

4.3 Results on the PROMO task

We apply our algorithm to the PROMO task and get a specificity of above 77% and sensitivity around 89%.

5. Discussion

One may doubt that whether only the iterative threshold process can do the variable selection job well enough. When feature space is of high dimension, from our experience, only the iterative threshold without stepwise selection can lead to terrible results. The $stepwisefit$ procedure in Matlab is very efficient in computation, and the computation for one time of iterative threshold can be negligible.

References

Box, G.E.P., and G.C. Tiao, Intervention analysis with applications to economic and environmental problems, *J. Amer. Statist. Assoc.*, 70, 70-79, 1975.

Box, G.E.P., Jenkins, G.M. and Reinsel, G.C., *Time Series Analysis: Forecasting and Control, 3rd Edition*. Pearson Education Asia Ltd., 1994.

P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods (Second Edition)*. Springer-Verlag New York, Inc., 1991.

Chen, J. and Chen Z. , Extended Bayesian information criterion for model selection with large model spaces, *Biometrika*, 95, 759-771, 2008.

S. Weisberg, *Applied Linear Regression (Second Edition)*. John Wiley & Sons, Inc., 1985.