Statistica Sinica **33** (2023), 1025-1045
doi:https://doi.org/10.5705/ss.202021.0078

A STRATIFIED PENALIZATION METHOD FOR SEMIPARAMETRIC VARIABLE LABELING OF MULTI-OUTPUT TIME-VARYING COEFFICIENT MODELS

Ting Zhang[1], Weiliang Wang[2] and Yu Shao[2]

[1] *University of Georgia and* [2] *Boston University*

Abstract: In a time-varying coefficient model, the regression coefficient is allowed to change over time as a nonparametric function to capture the time-varying feature. Owing to its popularity in time series applications, where the assumption of independence typically does not hold, it is desirable to allow dependent and nonstationary observations. We consider the problem of semiparametric variable labeling and estimation for multi-output time-varying coefficient models in a time series setting, where a variable can be labeled as time-varying, time-constant, or irrelevant, in a nested structure. We first show that the natural approach of imposing separate penalties on the local linear estimator and its derivative do not work as intended for semiparametric labeling, owing to the lack of connection between the coefficient and the derivative estimators in the popular local linear method. We then propose a stratified fix that borrows information from the coefficient estimator and combines it with the derivative into the same stratum that simultaneously achieves successful labeling and estimation. We establish the theoretical properties of the proposed method, including its estimation and labeling consistency, for a general class of nonstationary processes. Numerical examples, including a Monte Carlo simulation study and a real-data application, are presented to illustrate the proposed method.

Key words and phrases: Kernel smoothing, local linear estimation, nonstationary time series, time-varying coefficient model, variable selection.

1. Introduction

Linear regression models have been recognized as powerful and popular statistical tools for studying the relationship between a response variable and a set of explanatory variables. However, for applications to time series data, numerous empirical examples have suggested that the regression coefficient does not necessarily stay as a constant, and can change over time with other aspects of the data, making the observed time series nonstationary. For example, Fan and Zhang (1999) studied the relationship between the number of daily hospital admissions and the level of multiple pollutants in Hong Kong, finding a time-varying

Corresponding author: Ting Zhang, Department of Statistics, University of Georgia, Athens, GA 30602, USA. E-mail: tingzhang@uga.edu.

relationship. Gao and Hawthorne (2006) regressed the global temperature series on the Southern Oscillation Index (SOI), and argued that at least the intercept term should be treated as time varying in a nonparametric fashion, owing to a lack of knowledge about the change. Zhang and Wu (2015) considered the problem of modeling U.S. treasury yields, and found statistical evidence for a time-varying linear drift for the yield curve rates with six-month maturity. Such applications motivated the time-varying coefficient model, in which the regression coefficient is no longer assumed to be a constant, but is modeled as a nonparametric function of time to capture the time-varying feature.

The time-varying coefficient model is related to the varying coefficient model, which has been studied extensively; see, for example, Fan and Zhang (1999), Zhang, Lee and Song (2002), Xia, Zhang and Tong (2004), Ahmad, Leelahanon and Li (2005), Fan and Huang (2005), Li and Liang (2008), Wang, Li and Huang (2008), Wang and Xia (2009), Tang et al. (2012), Xue and Qu (2012), Cheng, Honda and Zhang (2016), and an excellent review by Fan and Zhang (2008). However, in a varying coefficient model, the observations are typically assumed to be independent samples, and the distribution from which the index variable is sampled is often assumed to have a continuous density function that is bounded away from zero and infinity on its support. This prevents allowing the deterministic time as the index variable, and as a result, different treatments are often needed for the time-varying coefficient model. In particular, by using lagged values as potential explanatory variables, the time-varying coefficient model can include the influential time-varying autoregressive model (Rao (1970); Dahlhaus, Neumann and Sachs (1999); Moulines, Priouret and Roueff (2005); Van Bellegem and Dahlhaus (2006)) as a special case. However, the latter model cannot be covered by the varying coefficient model with a random index. In addition, when there is only an intercept term in the model, the time-varying coefficient model reduces to the mean nonstationary model of Johnstone and Silverman (1997), Wu and Zhao (2007), Zhang and Wu (2011), and Zhang (2016), which has been widely used in nonparametric trend estimation and testing problems.

For time-varying coefficient models with nonstationary time series observations, Zhou and Wu (2010) constructed simultaneous confidence bands for the coefficient functions, and Zhang and Wu (2012) considered an integrated squared test that can be more suitable for detecting smooth and dense changes. In addition to estimating the coefficient functions and testing the hypotheses associated with them, an important problem is to label or partition the variables into time-varying, time-constant, and irrelevant categories. This tricategory labeling task in a nested structure has been studied mainly using a two-step approach, where one focuses separately on a bicategory labeling task at each step. For example, Li and Liang (2008) assumed prior knowledge on the partition between time-varying and time-constant components, and applied a penalized quasi-likelihood method to label the time-constant variables in the parametric part, and a separate generalized likelihood ratio test to label the time-varying variables in the nonparametric part; see also Li, Chen and Lin (2009). Zhang and Wu (2012) first use an information criterion to label zero and nonzero variables, and then further apply an integrated squared nonparametric test among the labeled nonzero variables to label time-constant variables. Zhang (2015) use a penalized local linear method to first label irrelevant variables, and then apply an information criterion on the remaining variables to further label the time-varying ones.

The main focus of this study is to consider a penalized local linear method that can simultaneously achieve successful tricategory labeling and semiparametric estimation in a single step. Unlike the basis expansion approach, which can borrow results directly from the well-developed penalized least squares, extending the popular local linear method (Fan and Gijbels (1996)) to the penalized setting can be nontrivial. As a result, even in the important work of Li and Liang (2008), penalized methods are only used for variable selection in the parametric component, and variable selection in the nonparametric component is still handled using the generalized likelihood ratio test. Wang and Xia (2009) first considered a penalized kernel estimator by vectorizing the local constant estimator on a set of discrete time points, so that one can place a penalty directly on the norm of that vector to obtain sparse solutions. Zhang (2015) proposed a local linear shrinkage method that can handle the additional derivative estimator from the more sophisticated local linear method. This method is able to work with nonparametric kernel estimators in their original function form without having to vectorize on a discrete set. However, in the aforementioned studies, penalized kernel estimation is used mainly to label irrelevant variables, and making it work for labeling time-varying and time-constant variables can be nontrivial. In Section 2, we demonstrate that the natural approach of penalizing the derivative estimator from the local linear method may not work as intended, because a zero derivative estimator will not guarantee a constant coefficient estimator, owing to the lack of connection between the two in the local linear estimation. To address this, in Section 3.1, we propose a new stratified penalization method that is able to automatically yield nonparametric coefficient estimators for time-varying variables, constant estimators for time-constant variables, and zero estimators for irrelevant variables. As a result, the method achieves the task of tricategory labeling and semiparametric estimation simultaneously in a computationally efficient manner. Note that we consider the multi-output setting in which a variable is labeled as time-constant or irrelevant if its coefficient function is uniformly constant or zero for all outputs. We establish the theoretical properties of the proposed method, including its estimation and labeling consistency, in Section 3.2 for a general class of nonstationary processes. Numerical experiments, including a Monte Carlo simulation study and a real-data analysis, are provided in Section 4 to illustrate the proposed method and examine its finite-sample performance. Section 5 concludes the paper.

2. Direct Penalization on the Derivative: A Natural Approach and its Issue

Consider the time-varying coefficient model

$$y_{i,n} = \boldsymbol{x}_{i,n}^\top \boldsymbol{\beta}(t_{i,n}) + e_{i,n}, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\boldsymbol{\beta}: \ [0,1] \to \mathbb{R}^p$ is the coefficient function, $t_{i,n} = i/n$ represents the time, and $(e_{i,n})$ is a sequence of random noise. The coefficient function in model (2.1) can be estimated using the local constant estimator (Wang and Xia (2009)), which at each time point $t \in [0,1]$ can be obtained by

$$\breve{\boldsymbol{\beta}}(t) = \underset{\boldsymbol{\eta}(t) \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \{y_{i,n} - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}(t)\}^2 K\left(\frac{t_{i,n} - t}{b_n}\right), \tag{2.2}$$

where $K(\cdot)$ is a kernel function and b_n is the bandwidth. For the kth component of $\boldsymbol{\eta}(\cdot)$, let $|\eta_k|_{[0,1]} = \{\int_0^1 |\eta_k(t)|^2 dt\}^{1/2}$ denote its norm. Then, the penalized local constant estimator minimizes

$$\int_0^1 \sum_{i=1}^n \{y_{i,n} - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}(t)\}^2 K\left(\frac{t_{i,n} - t}{b_n}\right) dt + \sum_{k=1}^p f_{\lambda_k}(|\eta_k|_{[0,1]}), \tag{2.3}$$

where $f_{\lambda_k}(\cdot)$ is the penalty function for the kth variable, with the tuning parameter λ_k that controls the degree of penalization. Unlike (2.2), the penalized estimator in (2.3) can achieve parameter estimation and variable selection simultaneously in a computationally efficient manner; see, for example, Zhang (2015). However, for the time-varying coefficient model, it is often the case that one is interested in distinguishing not only relevant and irrelevant variables, but also time-varying and time-constant components; see, for example, Cai, Fan and Yao (2000), Fan and Zhang (2000), Li and Liang (2008), Zhang and Wu (2012), and Zhang (2015), among others. This makes it a tricategory labeling problem, in which each variable is labeled as time-varying, time-constant, or irrelevant if the associated coefficient function is a time-varying nonparametric function, a uniform constant, or zero, respectively.

To achieve the aforementioned tricategory labeling, a natural approach is to consider the local linear method (Fan and Gijbels (1996)), which estimates not only the coefficient function, but also its derivative:

$$\{\tilde{\boldsymbol{\beta}}(t), \tilde{\boldsymbol{\beta}}'(t)\} = \operatorname*{argmin}_{\boldsymbol{\eta}(t), \boldsymbol{\eta}'(t) \in \mathbb{R}^p} \sum_{i=1}^{n} \{y_{i,n} - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}(t) - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}'(t)(t_{i,n} - t)\}^2 K\left(\frac{t_{i,n} - t}{b_n}\right).$$

Let

$$\Omega_n(\{\boldsymbol{\eta}(t), \boldsymbol{\eta}'(t)\}_{t\in[0,1]}) = \int_0^1 \sum_{i=1}^{n} \{y_{i,n} - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}(t) - \boldsymbol{x}_{i,n}^\top \boldsymbol{\eta}'(t)(t_{i,n} - t)\}^2 K\left(\frac{t_{i,n} - t}{b_n}\right) dt.$$

Then, similarly to (2.3), we can consider its penalized version that minimizes

$$\Omega_n(\{\boldsymbol{\eta}(t), \boldsymbol{\eta}'(t)\}_{t\in[0,1]}) + \sum_{k=1}^{p} f_{\lambda_k}(|\eta_k|_{[0,1]}) + \sum_{k=1}^{p} f_{\tau_k}(|\eta'_k|_{[0,1]}), \tag{2.4}$$

where the first penalty penalizes the coefficient estimator and shrinks it to zero for irrelevant variables, whereas the second penalty penalizes the derivative estimator and shrinks it to zero for time-constant variables; see, for example, Gao (2019) and Chan, Gao and Palma (2022). The penalized local linear estimator in (2.4) is intuitively straightforward in the sense that it exploits the derivative estimator from the local linear method and takes advantage of the mathematical connection between a function being constant and its derivative being zero. However, in the following, we show that directly imposing a penalty on the derivative, as in (2.4), may not work as intended, because a zero derivative estimator does not necessarily guarantee that the associated coefficient estimator will be a constant. This is mainly because the mathematical connection between a function and its derivative does not carry over to the local linear estimation scheme. Specifically, for the true coefficient function, it is mathematically guaranteed that a zero derivative will lead to a constant function. However, when performing the local linear estimation, the derivative term is regarded as the coefficient of an additional explanatory variable, namely, $\boldsymbol{x}_{i,n}(t_{i,n} - t)$. As a result, such a coefficient being zero does not necessarily guarantee that the coefficient of $\boldsymbol{x}_{i,n}$ will be the same for different time points, because they are simply treated as the coefficients for different variables $\boldsymbol{x}_{i,n}(t_{i,n} - t)$ and $\boldsymbol{x}_{i,n}$. As a result, additional manual flattening is often needed when implementing such a method. This typically leads to an altered optimization problem that is different from (2.4) in a nonnegligible manner, making its theoretical property difficult to understand.

To provide further insight into the aforementioned issue, we consider the simple setting when $p = 1$ with a time-constant coefficient, that is, when

$$y_{i,n} = x_{i,n}\beta + e_{i,n}, \quad i = 1, \dots, n, \tag{2.5}$$

and apply the penalized estimator in (2.4) to determine whether it will automatically reduce to a constant in this case. For simplicity, we assume that $\beta \neq 0$, and focus on the semi-oracle estimator $\check{\beta}(t)$, $t \in [0, 1]$, that minimizes

$$\Omega_n(\{\eta(t), \eta'(t)\}_{t\in[0,1]}) + f_\tau(|\eta'_k|_{[0,1]}). \tag{2.6}$$

This differs from (2.4) by dropping the first penalty term on the regression coefficient, because it is known to be nonzero, hence the term semi-oracle.

Theorem 1. *Assume that $f_\tau(x) = \tau|x|$ takes the LASSO penalty. Then, for any given data $(x_{i,n}, y_{i,n})$, for $i = 1, \dots, n$, with nondegenerate local designs for each $t \in [0, 1]$, the semi-oracle estimator $\check{\beta}(t)$, for $t \in [0, 1]$, that minimizes* (2.6) *among all square integrable continuous functions is equivalent to the local constant estimator in* (2.2) *for a sufficiently large choice of τ.*

Because the regression coefficient in (2.5) is assumed to be a nonzero constant, the oracle choice of the tuning parameters in (2.4) should be $\lambda = 0$ and $\tau = +\infty$. That is, no penalty should be put on the regression coefficient to reduce the bias caused by the penalized estimation of a nonzero coefficient function, whereas a sufficient penalization should be put on the derivative to force the coefficient function estimator to become a constant. However, from Theorem 1, the direct penalization on the derivative as in (2.4) may not work as intended, even with this oracle choice of the tuning parameters. In particular, when the penalty function satisfies the natural condition that $f_\tau(0) = 0$ (Fan and Li (2001)), using the oracle tuning $\lambda = 0$ in (2.4) makes it equivalent to (2.6), the solution of which, by Theorem 1, becomes the local constant estimator in (2.2) for a sufficiently large choice of τ with an oracle choice of $\tau = +\infty$. Therefore, imposing a direct penalization on the derivative, as in (2.4), typically does not yield a time-constant regression coefficient estimator, which can cause ambiguity when labeling time-varying and time-constant variables. The result in Theorem 1 can be generalized to penalty functions other than the LASSO, but the main purpose here is to show that directly penalizing the derivative may not work as intended, even when the popular and successful LASSO penalty is used. We propose a stratified fix that is able to automatically produce nonparametric coefficient estimators for time-varying variables, constant estimators for time-constant variables, and zero estimators for irrelevant variables, thus achieving the goals of tricategory labeling and semiparametric estimation at the same time.

3. Stratified Penalization: A Fix

3.1. Methodology

Here, we consider the multi-output time-varying coefficient model

$$\boldsymbol{y}_{i,n} = B(t_{i,n})^\top \boldsymbol{x}_{i,n} + \boldsymbol{e}_{i,n}, \quad i = 1, \ldots, n, \tag{3.1}$$

where $\boldsymbol{y}_{i,n} \in \mathbb{R}^d$ is the multi-output response vector, $\boldsymbol{x}_{i,n} \in \mathbb{R}^p$ is the set of explanatory variables, $B : [0,1] \to \mathbb{R}^{p\times d}$ is the coefficient function matrix, with its kth row being the coefficient function vector for the kth variable, and $(\boldsymbol{e}_{i,n})$ is a sequence of random vectors that form a triangular array of multivariate nonstationary processes that can depend on $(\boldsymbol{x}_{i,n})$ to accommodate heteroscedastic errors. Compared with the single-response setting (2.1), variable selection in the multi-output setting typically requires additional effort; see, for example, Turlach, Venables and Wright (2005), Rothman, Levina and Zhu (2010), Chen and Huang (2012), and Lee and Liu (2012), among others. The aforementioned studies consider variable selection for multi-output regression models in the traditional setting when the regression coefficient is assumed to be a constant. Here, we consider the time-varying setting (3.1), in which the regression coefficient can change over time as a nonparametric function. In this case, one is interested in labeling not only relevant and irrelevant variables, but also time-varying and time-constant variables. Here, a variable is said to be time constant or irrelevant if its coefficient function is uniformly a constant or zero, respectively, for all outputs. This cannot be achieved by performing variable selection separately on each of the response variables.

Let $\Theta(t) = \{\theta_{l,k,j}(t)\}_{l,k,j}$ be a three-way tensor function, with $\Theta_{0,\cdot,\cdot}(t) = \{\theta_{0,k,j}(t)\}_{k,j} = B(t)$, $\Theta_{1,\cdot,\cdot}(t) = \{\theta_{1,k,j}(t)\}_{k,j} = b_n B'(t)$, and its norm $|\Theta|_{[0,1]} = \{\sum_{l,k,j} \int_0^1 \theta_{l,k,j}(t)^2 dt\}^{1/2}$. Then, we can write the multi-output kernel criterion function as

$$\Upsilon_n(\{\Theta(t)\}_{t\in[0,1]}) =$$
$$\int_0^1 \sum_{i=1}^{n} \left| \boldsymbol{y}_{i,n} - \Theta_{0,\cdot,\cdot}(t)^\top \boldsymbol{x}_{i,n} - \Theta_{1,\cdot,\cdot}(t)^\top \boldsymbol{x}_{i,n} \left(\frac{t_{i,n}-t}{b_n}\right) \right|^2 K\left(\frac{t_{i,n}-t}{b_n}\right) dt.$$

To construct appropriate penalty structures that can achieve simultaneous tricategory labeling and semiparametric estimation, instead of imposing penalties directly on the coefficient part and the derivative part, as in (2.4), we propose decomposing the norm of Θ according to the different strata implied by the nested tricategory labeling structure. In particular, the irrelevant label stratum has a projection of zero, the time-constant label stratum has a projection of $\bar{\boldsymbol{\theta}}_{k,\cdot} = \int_0^1 \boldsymbol{\theta}_{0,k,\cdot}(t)dt$, where $\boldsymbol{\theta}_{0,k,\cdot}(t) = \{\theta_{0,k,1}(t), \ldots, \theta_{0,k,d}(t)\}^\top$, and the time-varying label stratum has a projection of $\boldsymbol{\theta}_{0,k,\cdot}(t) - \bar{\boldsymbol{\theta}}_{k,\cdot}$, together with $\boldsymbol{\theta}_{1,k,\cdot}(t) = \{\theta_{1,k,1}(t), \ldots, \theta_{1,k,d}(t)\}^\top$. This motivates us to consider the stratified penalized local linear (SPLL) estimator $\hat{\Theta}(t) = \{\hat{\theta}_{l,k,j}(t)\}_{l,k,j}$, for $t \in [0,1]$, that minimizes

$$\Upsilon_n(\{\Theta(t)\}_{t\in[0,1]}) + \sum_{k=1}^p f_{\lambda_{k,n}}(|\bar{\boldsymbol{\theta}}_{k,\cdot}|)$$
$$+ \sum_{k=1}^p g_{\tau_{k,n}}\left(\left[\int_0^1 \{|\boldsymbol{\theta}_{0,k,\cdot}(t) - \bar{\boldsymbol{\theta}}_{k,\cdot}|^2 + |\boldsymbol{\theta}_{1,k,\cdot}(t)|^2\}dt\right]^{1/2}\right),$$

where $f_{\lambda_{k,n}}(\cdot)$ and $g_{\tau_{k,n}}(\cdot)$ are penalty functions with nonnegative tuning parameters $\lambda_{k,n}$ and $\tau_{k,n}$, respectively. We establish its theoretical properties, including the estimation and labeling consistency, in Section 3.2 for a general class of nonstationary processes, and our results are directly applicable to vector time-varying autoregressive models. The two terms $|\bar{\boldsymbol{\theta}}_{k,\cdot}|$ and $\int_0^1 |\boldsymbol{\theta}_{0,k,\cdot}(t) - \bar{\boldsymbol{\theta}}_{k,\cdot}|^2 dt$ in the penalty have their own and separate goals. In particular, the term $\int_0^1 |\boldsymbol{\theta}_{0,k,\cdot}(t) - \bar{\boldsymbol{\theta}}_{k,\cdot}|^2 dt$ is mainly used to merge with the derivative $\int_0^1 |\boldsymbol{\theta}_{1,k,\cdot}(t)|^2 dt$ as a group structure. This ensures that, for a time-constant variable, both will be penalized to zero, making the coefficient function $\boldsymbol{\theta}_{0,k,\cdot}(t) \equiv \bar{\boldsymbol{\theta}}_{k,\cdot}$ be a constant and its derivative $\boldsymbol{\theta}_{1,k,\cdot}(t) \equiv 0$, thus solving the issue of the direct derivative penalization approach. The other additional term $|\bar{\boldsymbol{\theta}}_{k,\cdot}|$ ensures that once the regression coefficient function is shrunken to its constant projection $\bar{\boldsymbol{\theta}}_{k,\cdot}$, it can further penalize the coefficient to zero to correctly label irrelevant variables.

3.2. Theoretical properties

In (3.1), we allow nonstationary time series observations, which have been studied extensively in the literature; see, for example, Priestley (1965), Dahlhaus (1997), Cheng and Tong (1998), Mallat, Papanicolaou and Zhang (1998), Giurcanu and Spokoiny (2004), Ombao, von Sachs and Guo (2005), Zhou and Wu (2010), and Zhang (2013), and the references therein. Let $(\boldsymbol{\epsilon}_i)$ be a sequence of independent and identically distributed (i.i.d.) innovations, and denote its shift process by $\boldsymbol{\mathcal{F}}_i = (\ldots, \boldsymbol{\epsilon}_{i-1}, \boldsymbol{\epsilon}_i)$. Here, we follow Zhang (2015) and assume that

$$\max_{1\le i\le n} \|\boldsymbol{x}_{i,n} - \boldsymbol{G}(t_{i,n}, \boldsymbol{\mathcal{F}}_i)\| = O(n^{-1}), \quad \max_{1\le i\le n} \|\boldsymbol{e}_{i,n} - \boldsymbol{H}(t_{i,n}, \boldsymbol{\mathcal{F}}_i)\| = O(n^{-1}), \tag{3.2}$$

for some measurable functions $\boldsymbol{G}$ and $\boldsymbol{H}$, such that $(\boldsymbol{x}_{i,n})$ and $(\boldsymbol{e}_{i,n})$ are proper sequences of random vectors with $E(\boldsymbol{e}_{i,n} \mid \boldsymbol{x}_{i,n}) = \mathbf{0}$. Compared with the exact representation in Wu (2005), the approximate framework in (3.2) allows the popular time-varying autoregressive model (Rao (1970); Dahlhaus, Neumann and Sachs (1999); Moulines, Priouret and Roueff (2005); Van Bellegem and Dahlhaus (2006)) and covers a wide range of linear and nonlinear processes; see also the discussion in Zhang and Wu (2012). Let $\boldsymbol{\epsilon}_0^\star$ be identically distributed as $\boldsymbol{\epsilon}_0$, but independent of the sequence $(\boldsymbol{\epsilon}_i)$. We can define the coupled shift process $\boldsymbol{\mathcal{F}}_i^\star = (\ldots, \boldsymbol{\epsilon}_{-1}, \boldsymbol{\epsilon}_0^\star, \boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_i)$. Then, for any collection of processes $\{\boldsymbol{J}(t; \boldsymbol{\mathcal{F}}_i)\}_{i\in\mathbb{Z}}$ on $t \in [0,1]$, the functional dependence measure of Wu (2005) can be written as

$$\delta_{i,q}(\boldsymbol{J}) = \sup_{t\in[0,1]} \|\boldsymbol{J}(t; \boldsymbol{\mathcal{F}}_i) - \boldsymbol{J}(t; \boldsymbol{\mathcal{F}}_i^\star)\|_q, \quad \Delta_{0,q}(\boldsymbol{J}) = \sum_{i=0}^{\infty} \delta_{i,q}(\boldsymbol{J}),$$

where $\delta_{i,q}(\boldsymbol{J})$ measures the dependence of $\boldsymbol{J}(t; \boldsymbol{\mathcal{F}}_i)$ on the innovation $\boldsymbol{\epsilon}_0$, and its cumulative effect is measured by $\Delta_{0,q}(\boldsymbol{J})$. Let $\boldsymbol{L} = \boldsymbol{G} \times \boldsymbol{H}^\top$. We assume the following conditions.

(A1) The coefficient matrix function $B \in \mathcal{C}^3[0,1]$, the class of three times continuously differentiable matrix-valued functions on $[0,1]$.

(A2) The underlying process satisfies $\Delta_{0,4}(\boldsymbol{G}) + \Delta_{0,2}(\boldsymbol{L}) < \infty$.

(A3) There exists a constant $0 < C < \infty$ such that

$$\|\boldsymbol{G}(t_1, \boldsymbol{\mathcal{F}}_i) - \boldsymbol{G}(t_2, \boldsymbol{\mathcal{F}}_i)\| + \|\boldsymbol{L}(t_1, \boldsymbol{\mathcal{F}}_i) - \boldsymbol{L}(t_2, \boldsymbol{\mathcal{F}}_i)\| \le C|t_1 - t_2|$$

holds uniformly for all $t_1, t_2 \in [0,1]$.

(A4) The smallest eigenvalue of $E\{\boldsymbol{G}(t, \boldsymbol{\mathcal{F}}_i)\boldsymbol{G}(t, \boldsymbol{\mathcal{F}}_i)^\top\}$ is bounded away from zero on $[0,1]$.

Condition (A1) is a smoothness condition on the regression coefficient matrix function, which is a common assumption for nonparametric kernel estimation; see, for example, Fan and Gijbels (1996). Condition (A2) is a short-range dependence condition quantified by the functional dependence measure of Wu (2005); see also Zhou and Wu (2010) and Zhang and Wu (2012). Condition (A3) is a stochastic

Lipschitz continuity condition, under which the underlying process can be locally approximated by a stationary process within a small window; see, for example, the discussion in Zhang and Wu (2011). Condition (A4) is a regularity condition that prevents the design matrix from being singular in probability; see also Zhang (2015). Throughout this paper, we assume that the kernel function $K \in \mathcal{K}$, the collection of symmetric functions in $\mathcal{C}^1[-1,1]$, with $\int_{-1}^{1} K(v)dv = 1$. Examples include the Epanechnikov kernel $K(v) = 0.75\max(1-v^2, 0)$, the Bartlett kernel $K(v) = \max(1-|v|, 0)$, and the rectangle kernel $K(v) = 0.5\mathrm{I}(|v| \leq 1)$, with $\mathrm{I}(\cdot)$ being the indicator function, among many others. We also assume the following conditions on the penalty functions.

(P1) $f_\lambda(0) = 0$ and $g_\tau(0) = 0$.

(P2) $\lambda^{-1}\sup_{x\in\mathbb{R}} |f'_\lambda(x)| < \infty$ and $\tau^{-1}\sup_{x\in\mathbb{R}} |g'_\tau(x)| < \infty$.

(P3) $\lambda^{-1}\liminf_{x\to 0+} |f'_\lambda(x)| > 0$ and $\tau^{-1}\liminf_{x\to 0+} |g'_\tau(x)| > 0$.

Conditions (P1)–(P3) are natural requirements for good penalty functions (Fan and Li (2001)), and are satisfied by many popular choices, such as the LASSO penalty of Tibshirani (1996), the hard thresholding penalty of Antoniadis (1997), and the SCAD penalty of Fan and Li (2001). Let $\mathcal{D}_v$, $\mathcal{D}_c$, and $\mathcal{D}_0$ denote subsets of time-varying, time-constant, and irrelevant variables, respectively, and we further write $\mathcal{D}_v = \mathcal{D}_{v0} \cup \mathcal{D}_{v1}$, where $\mathcal{D}_{v0}$ is the set of time-varying variables with $|\bar{\boldsymbol{\theta}}_{k,\cdot}| = 0$ and $\mathcal{D}_{v1} = \mathcal{D}_v \setminus \mathcal{D}_{v0}$. Theorems 2 and 3 provide the estimation consistency and labeling consistency for the proposed SPLL estimator.

Theorem 2. *Assume* (A1)–(A4), (P1), (P2), $b_n \to 0$, *and* $nb_n \to \infty$. *If* $\{(nb_n)^{-1/2} + b_n^2\}(\max_{k\in\mathcal{D}_c\cup\mathcal{D}_{v1}} \lambda_{k,n} + \max_{k\in\mathcal{D}_v} \tau_{k,n}) = O(1)$, *then the norm*

$$|\hat{\Theta} - \Theta|_{[0,1]} = O_p\{(nb_n)^{-1/2} + b_n^2\}.$$

Theorem 3. *Assume* (A1)–(A4), (P1)–(P3), $b_n \to 0$, *and* $nb_n^2 \to \infty$. *If* $\{(nb_n)^{-1/2} + b_n^2\}(\max_{k\in\mathcal{D}_c\cup\mathcal{D}_{v1}} \lambda_{k,n} + \max_{k\in\mathcal{D}_v} \tau_{k,n}) = O(1)$, $\min_{k\in\mathcal{D}_0} \lambda_{k,n}/\{(nb_n)^{1/2} + nb_n^3\} \to \infty$, *and* $\min_{k\in\mathcal{D}_v^c} \tau_{k,n}/\{(nb_n)^{1/2} + nb_n^3\} \to \infty$, *then*

$$\mathrm{pr}\left\{\max_{k\in\mathcal{D}_0\cup\mathcal{D}_c} \sup_{t\in[0,1]} \left|\hat{\boldsymbol{\theta}}_{0,k,\cdot}(t) - \int_0^1 \hat{\boldsymbol{\theta}}_{0,k,\cdot}(s)ds\right| = 0 \text{ and } \max_{k\in\mathcal{D}_0\cup\mathcal{D}_c} \sup_{t\in[0,1]} |\hat{\boldsymbol{\theta}}_{1,k,\cdot}(t)| = 0\right\} \to 1,$$

and

$$\mathrm{pr}\left\{\max_{k\in\mathcal{D}_0} \sup_{t\in[0,1]} |\hat{\boldsymbol{\theta}}_{0,k,\cdot}(t)| = 0 \text{ and } \max_{k\in\mathcal{D}_0} \sup_{t\in[0,1]} |\hat{\boldsymbol{\theta}}_{1,k,\cdot}(t)| = 0\right\} \to 1.$$

By Theorem 3, for time-constant or irrelevant variables, the SPLL estimator proposed in Section 3.1 automatically produces a constant coefficient function with a zero derivative estimator. In addition, for irrelevant variables nested within $\mathcal{D}_0 \cup \mathcal{D}_c$, the coefficient function is regularized to zero uniformly over time. Therefore, it achieves the simultaneous tricategory variable labeling and semiparametric estimation, without having to decompose the problem into two bicategory labeling subproblems, as in Li and Liang (2008) and Zhang and Wu (2012). Note that in Li and Liang (2008), prior knowledge is assumed on the partition between the time-varying and time-constant variables, and therefore their labeling problem is supervised. This study concerns the unsupervised setting, where we do not assume we have this prior knowledge. In the following section, we describe the implementation of the proposed SPLL method, and examine its finite-sample performance using a Monte Carlo simulation study and a real-data analysis.

4. Implementation

4.1. Computational algorithm

Although penalized methods and their computational algorithms have been widely studied in the literature, existing results focus mainly on the traditional linear regression model with constant coefficients; see, for example, Tibshirani (1996), Knight and Fu (2000), Fan and Li (2001), Efron et al. (2004), Yuan and Lin (2006), and Zou and Li (2008), and the references therein. In addition, different penalty terms are usually put on the coefficients associated with the variables. In the current setting, the two penalty terms f and g can both be associated with the same variable, but for different purposes, where one regularizes a time-varying function into a constant, and the other shrinks a constant toward zero. Furthermore, the current setting requires appropriately combining localized least squares functions, as in traditional kernel regression methods, with suitably constructed global penalization terms to achieve successful semiparametric variable labeling and estimation. In the Supplementary Material, we describe an iterative algorithm that can be used to compute the SPLL estimator proposed in Section 3.1. Next, we discuss the choice of the tuning parameters.

4.2. Tuning parameter selection

Implementing the proposed SPLL method requires two sets of tuning parameters. The first ($\tau_{k,n}$) controls the degree of regularization from the time-varying stratum to the time-constant stratum, and the second ($\lambda_{k,n}$) controls the degree of regularization from the time-constant stratum to the irrelevant label stratum. Here, we adopt the idea of the adaptive LASSO (Zou (2006)), and set

$$\lambda_{k,n} = \lambda_n \cdot \left| \int_0^1 \tilde{\boldsymbol{\theta}}_{0,k,\cdot}(t)dt \right|^{-1},$$

$$\tau_{k,n} = \tau_n \cdot \left[\int_0^1 \left\{ \left| \tilde{\boldsymbol{\theta}}_{0,k,\cdot}(t) - \int_0^1 \tilde{\boldsymbol{\theta}}_{0,k,\cdot}(s)ds \right|^2 + |\tilde{\boldsymbol{\theta}}_{1,k,\cdot}(t)|^2 \right\} dt \right]^{-1/2}, \quad (4.1)$$

for some tuning parameters λ_n and τ_n that do not depend on k, where $\tilde{\Theta}(t) = \{\tilde{\theta}_{l,k,j}(t)\}_{l,k,j}$, for $t \in [0,1]$, can be taken as the unpenalized local linear estimator. Note that the norm $[\int_0^1 \{|\tilde{\boldsymbol{\theta}}_{0,k,\cdot}(t) - \int_0^1 \tilde{\boldsymbol{\theta}}_{0,k,\cdot}(s)ds|^2 + |\tilde{\boldsymbol{\theta}}_{1,k,\cdot}(t)|^2\}dt]^{1/2}$ is relatively small for time-constant variables and relatively large for time-varying variables. Thus, the choice in (4.1) can lead to adaptive tuning for different variables. When the bandwidth $b_n = cn^{-1/5}$, for some $0 < c < \infty$, has the asymptotic mean squared error optimal rate, following a similar discussion to that in Zhang (2015), one can simply use the asymptotic choice $\lambda_n = n^{1/5}$ and $\tau_n = n^{1/5}$. For any set $\mathcal{D}$, we use $|\mathcal{D}|$ to denote its cardinality. To provide a finite-sample data-driven choice of the pair (λ_n, τ_n), we consider a natural extension of the information criterion used in Zhang (2015) to the current setting, and minimize

$$\mathrm{EIC}(\lambda,\tau) = \log\{\Upsilon_n(\{\hat{\Theta}(t;\lambda,\tau)\}_{t\in[0,1]})\} + \frac{\log n}{nb_n} \cdot |\hat{\mathcal{D}}_0^c(\lambda,\tau)| + \frac{\log n}{nb_n} \cdot |\hat{\mathcal{D}}_v(\lambda,\tau)|,$$

which can also be viewed as a semiparametric extension of the traditional BIC. The simulation study in Section 4.3 shows that this data-driven tuning selector performs reasonably well for the current tricategory variable labeling and semiparametric estimation.

4.3. Simulation results

We conduct Monte Carlo simulations to examine the finite-sample performance of the proposed SPLL method. For this, let $\boldsymbol{\epsilon}_k = (\epsilon_{k,1}, \ldots, \epsilon_{k,p-1})^\top \in \mathbb{R}^{p-1}$, for $k \in \mathbb{Z}$, be independent innovation vectors with independent Rademacher components, and let $P_j(t)$ be the jth order Legendre polynomial. Let $\boldsymbol{M}^\diamond = (0.2^{|i-j|})_{1\le i,j\le p-1}$ and $\boldsymbol{P}(t) \in \mathbb{R}^{(p-1)\times(p-1)}$ be a diagonal matrix with the jth diagonal element $P_j(2t-1)/4$. Then, the vector $\boldsymbol{\xi}_k = \boldsymbol{M}^\diamond \boldsymbol{\epsilon}_k$ has dependent components, and we form $\boldsymbol{x}_{i,n} = \sum_{j=0}^\infty \boldsymbol{P}(i/n)^j \boldsymbol{\xi}_{i-j}$, which is a nonstationary process, owing to the coefficients being time varying. Let $\varepsilon_{k,l}$, for $k \in \mathbb{Z}$, $l \in \{1, \ldots, d\}$, be an array of independent standard normal random variables that is also independent of the process $(\boldsymbol{\epsilon}_k)$. We then form the nonstationary nonlinear process

$\boldsymbol{\zeta}_{i,n} = (\zeta_{i,1,n}, \ldots, \zeta_{i,d,n})^\top$, with $\zeta_{i,l,n} = \varepsilon_{i,l} + 2(i/n - 0.5)^2\{|\varepsilon_{i-1,l}| - (2/\pi)^{1/2}\} + \sum_{j=1}^{\infty} j^{-2}\varepsilon_{i-j,l}$. We consider the following multi-output time-varying coefficient model with heteroscedastic errors:

$$\boldsymbol{y}_{i,n} = \boldsymbol{\beta}_0(t_{i,n}) + \sum_{k=1}^{p-1} \boldsymbol{\beta}_k(t_{i,n})x_{i,k,n} + 0.5\sigma(x_{i,2,n}^2 + x_{i,3,n}^2)^{1/2}\boldsymbol{\zeta}_{i,n}, \quad i = 1, \ldots, n.$$

Let $n = 500$. We consider the following configurations on the variable labeling, where we use *a*-*b*-*c* to indicate a configuration with a time-varying variables, b nonzero time-constant variables, and c zero variables.

(2-2-16) Two time-varying variables: $\boldsymbol{\beta}_0(t) = \{3(2t-1)^2, 2(2t-1)^3\}^\top$ and $\boldsymbol{\beta}_2(t) = \{2\sin(2\pi t - 1), 2\cos(2\pi t + 1)\}^\top$; two time-constant variables: $\boldsymbol{\beta}_1(t) = (-1, \pi/3)^\top$ and $\boldsymbol{\beta}_3(t) = (1.5, -2^{1/2})^\top$; and 16 zero variables $\boldsymbol{\beta}_4(t) = \cdots = \boldsymbol{\beta}_{19}(t) = (0,0)^\top$.

(5-5-10) Five time-varying variables: $\boldsymbol{\beta}_0(t) = \{3(2t-1)^2, 2(2t-1)^3\}^\top$, $\boldsymbol{\beta}_2(t) = \{2(2t-1), 2\cos(2\pi t - 1)\}^\top$, $\boldsymbol{\beta}_4(t) = \{2\cos(2\pi t), 2(2t-1)^2\}^\top$, $\boldsymbol{\beta}_6(t) = \{\cos(\pi t) + 1, 2\sin\{\exp(\pi t - 2)\}\}^\top$, and $\boldsymbol{\beta}_8(t) = [\exp(-2t+1), 2\{\sin(-2\pi t + 1)\}^3]^\top$; five time-constant variables: $\boldsymbol{\beta}_1(t) = (2, -1.5)^\top$, $\boldsymbol{\beta}_3(t) = (1.5, -\pi/3)^\top$, $\boldsymbol{\beta}_5(t) = (\pi/2, \pi^{1/2})^\top$, $\boldsymbol{\beta}_7(t) = (-3^{1/2}, 2^{1/2})^\top$, and $\boldsymbol{\beta}_9(t) = (-e^{1/2}, 3/\pi)^\top$; and 10 zero variables $\boldsymbol{\beta}_{10}(t) = \cdots = \boldsymbol{\beta}_{19}(t) = (0,0)^\top$.

(2-8-10) Two time-varying variables: $\boldsymbol{\beta}_0(t) = \{3(2t-1)^2, 2(2t-1)^3\}^\top$ and $\boldsymbol{\beta}_2(t) = \{2\sin(2\pi t), 2\cos(2\pi t)\}^\top$; eight time-constant variables: $\boldsymbol{\beta}_1(t) = (1, 5^{1/2}/2)^\top$, $\boldsymbol{\beta}_3(t) = (\pi/2, -1.3)^\top$, $\boldsymbol{\beta}_4(t) = (e^{1/2}, -1.5)^\top$, $\boldsymbol{\beta}_5(t) = (1.5, 4/\pi)^\top$, $\boldsymbol{\beta}_6(t) = (1.2, 3^{1/2})^\top$, $\boldsymbol{\beta}_7(t) = (0.8, 7^{1/3})^\top$, $\boldsymbol{\beta}_8(t) = (-2^{1/2}, 1)^\top$, and $\boldsymbol{\beta}_9(t) = (-5/\pi, \pi/3)^\top$; and 10 zero variables $\boldsymbol{\beta}_{10}(t) = \cdots = \boldsymbol{\beta}_{19}(t) = (0,0)^\top$.

The three configurations above represent cases with a small number of nonzero variables, a balanced number of time-varying and time-constant variables, and an unbalanced number of time-varying and time-constant variables, respectively. For each configuration, we consider two noise levels $\sigma \in \{1, 2\}$, and apply the proposed SPLL method for the semiparametric variable labeling and estimation. We also make a comparison with the two-step procedure of Zhang (2015), denoted by Zhang15, and the direct derivative penalization method of Gao (2019), denoted by Gao19. Throughout our numerical experiments, we use the LASSO penalty function. The results with $\sigma = 2$ are reported in Table 1. The results with $\sigma = 1$ follow a similar pattern, and are provided in the Supplementary Material. Note that a case is considered to be under-labeled if at least one component of one variable is mislabeled from time varying to time constant, from time constant to zero, or from time varying to zero. On the other hand, a case is considered to be over-labeled if there is no under-labeling, and at least one variable is mislabeled from zero to time constant, from time constant to time varying, or from zero to time varying. We also report the labeling consistency ratio (LCR), defined as the proportion of correctly labeled variables, along with the mean squared error (MSE) of the associated semiparametric estimates.

We observe the following from our simulation results. First, the proposed SPLL method performs reasonably well, because for most of the cases considered, it produces the highest proportion of correctly labeled models, especially for the challenging cases when $\sigma = 2$. Its performance is also reasonably robust to different choices of the bandwidth. In addition, it typically leads to semiparametric estimates with a much smaller MSE than that of the two-step procedure of Zhang (2015). Note that even for the less challenging cases when $\sigma = 1$, as reported in Table 1 of the Supplementary Material, for which both the method of Zhang (2015) and the proposed method produce almost ideal results on variable labeling, the proposed method continues to yield semiparametric estimates with a much smaller MSE. This is mainly because the two-step procedure of Zhang (2015) does not fully use the labeling information when performing the nonparametric estimation, whereas the proposed method is able to interactively take advantage of the labeling information during the iteration. Furthermore, the proposed stratified method seems to improve on the direct derivative penalization method of Gao (2019). The reported LCR values show that the method of Gao (2019), although correctly labeling most of the variables, exhibits difficulty on a few variables, resulting in a very low proportion of correctly labeled models for certain configurations. Note that the proportion of correctly labeled models is a much more demanding metric than the LCR, because it does not allow even a single mislabeled variable. The less satisfactory performance of the method of Gao (2019) is mainly because it relies almost exclusively on derivative estimates to identify time-constant variables, and it is well known that, in local linear estimation, derivative estimates typically exhibit subpar quality compared with coefficient estimates; see, for example, Fan and Gijbels (1996). In addition, the underlying theory of such a direct derivative penalization method can be ambiguous and difficult to understand, as illustrated in Section 2. In contrast, the proposed stratified local linear method combines information from both coefficient estimates and their derivatives into the same stratum to label time-constant variables, and is theoretically guaranteed to yield consistent semiparametric labeling and estimation.

Table 1. Simulation results for $\sigma = 2$, based on 100 realizations for each configuration.

Model	b_n	Method	Under-label	Correct	Over-label	MSE	LCR
2-2-16	0.1	SPLL	0.00	0.99	0.01	0.0335	0.9995
		Zhang15	0.09	0.66	0.25	0.0765	0.9730
		Gao19	0.89	0.06	0.05	0.1931	0.9355
	0.2	SPLL	0.00	1.00	0.00	0.0283	1.0000
		Zhang15	0.10	0.81	0.09	0.0652	0.9810
		Gao19	0.00	0.26	0.74	0.0366	0.9085
	0.3	SPLL	0.00	1.00	0.00	0.0375	1.0000
		Zhang15	0.10	0.75	0.15	0.0702	0.9785
		Gao19	0.00	0.09	0.91	0.0414	0.8425
5-5-10	0.1	SPLL	0.04	0.94	0.02	0.0795	0.9970
		Zhang15	0.46	0.44	0.10	0.2510	0.9375
		Gao19	1.00	0.00	0.00	0.4982	0.7855
	0.2	SPLL	0.00	0.96	0.04	0.0610	0.9980
		Zhang15	0.12	0.79	0.09	0.1926	0.9635
		Gao19	0.61	0.14	0.25	0.1504	0.9385
	0.3	SPLL	0.00	0.98	0.02	0.0714	0.9990
		Zhang15	0.12	0.74	0.14	0.1857	0.9650
		Gao19	0.02	0.14	0.84	0.0848	0.8795
2-8-10	0.1	SPLL	0.00	0.98	0.02	0.0403	0.9990
		Zhang15	0.10	0.71	0.19	0.1682	0.9595
		Gao19	0.76	0.11	0.13	0.2111	0.9420
	0.2	SPLL	0.00	1.00	0.00	0.0352	1.0000
		Zhang15	0.12	0.81	0.07	0.1570	0.9650
		Gao19	0.00	0.41	0.59	0.0450	0.9405
	0.3	SPLL	0.00	1.00	0.00	0.0443	1.0000
		Zhang15	0.13	0.79	0.08	0.1593	0.9665
		Gao19	0.00	0.18	0.82	0.0501	0.9020

4.4. Data analysis

In this section, we apply the results to study the influence of El Niño-Southern Oscillation, characterized by the Southern Oscillation Index (SOI), on temperature anomalies, which is an important problem in climate science, and has been studied by Privalsky and Jensen (1995), Zheng and Basher (1999), Gao and Hawthorne (2006), McLean (2014), and Zhang (2015), among others. In this data analysis, we focus on determining what lags of the SOI should be used, and whether they should be treated as time-varying explanatory variables. For this, we consider the multi-output time-varying coefficient model

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0(t_{i,n}) + \sum_{j=1}^{25} \boldsymbol{\beta}_j(t_{i,n}) x_{i,j-13} + \boldsymbol{e}_i, \quad i = 1, \ldots, n, \tag{4.2}$$

where $\boldsymbol{y}_i = (y_{i,1}, y_{i,2})^\top$ are temperature anomalies from the northern and southern hemispheres, and x_k is the k-month-ahead SOI for $k \in \{-12, \ldots, 12\}$. For comparison with existing results, we use monthly data from 01/1936 to 12/2019, which can be downloaded from the Climatic Research Unit website: `https://crudata.uea.ac.uk/cru/data/temperature/`. In this case, the sample size $n = 984$, and time series plots are provided in Figure 1. We then apply the SPLL method proposed in Section 3 for the semiparametric variable labeling and estimation of (4.2), where the tuning parameters are selected by using the extended information criterion described in Section 4.2. For the bandwidth, we use a two-step selection procedure. We first use the asymptotic bandwidth $b_n^\circ = n^{-1/5}$ to obtain an initial labeling. Then, we apply the dependence-adjusted generalized cross-validation described in Section 4.3 of Zhang and Wu (2012) to the selected nonzero variables to obtain a data-driven bandwidth; see also Section 4.1 of Zhang (2015). Our analysis results show that, among the 25 lags considered, $x_{i,-2}$ and $x_{i,0}$ are labeled as time-constant variables, and all other lags are labeled as zero variables. In addition, the intercept is labeled as a time-varying variable, suggesting the following semiparametric multi-output model:

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0(t_{i,n}) + \boldsymbol{\beta}_{11} x_{i,-2} + \boldsymbol{\beta}_{13} x_{i,0} + \boldsymbol{e}_i, \quad i = 1, \ldots, n,$$

where the estimated time-varying coefficients $\hat{\boldsymbol{\beta}}_0(\cdot) = \{\hat{\beta}_{0,1}(\cdot), \hat{\beta}_{0,2}(\cdot)\}^\top$ are plotted in Figure 2 for the northern and southern hemispheres, and the estimated time-constant coefficients are given by $\hat{\boldsymbol{\beta}}_{11} = (-0.011, -0.015)^\top$ and $\hat{\boldsymbol{\beta}}_{13} = (-0.015, -0.014)^\top$.

Note that the method of Zhang (2015) leads to the multi-output model

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0(t_{i,n}) + \boldsymbol{\beta}_{10} x_{i,-3} + \boldsymbol{\beta}_{11} x_{i,-2} + \boldsymbol{\beta}_{12} x_{i,-1} + \boldsymbol{\beta}_{13} x_{i,0} + \boldsymbol{e}_i, \quad i = 1, \ldots, n.$$

Compared with the model selected by the proposed SPLL method, it shares the same insight that the effect of the SOI on temperature anomalies can be viewed as time-constant, whereas the intercept should be treated as time varying. This provides a data-driven approach to verify the semiparametric assumption commonly used in the climate science literature; see, for example, McLean (2014), who poses the possibility of a time-varying relationship, but does not explore statistical tools other than the simple linear regression to investigate further. On the other hand, the proposed SPLL method selects only $x_{i,-2}$ and $x_{i,0}$ as important lags, which differs from the aforementioned model selected by the method of Zhang (2015). The simulation results reported in Table 1 show that the method of Zhang (2015) can have a higher probability of producing over-labeled models

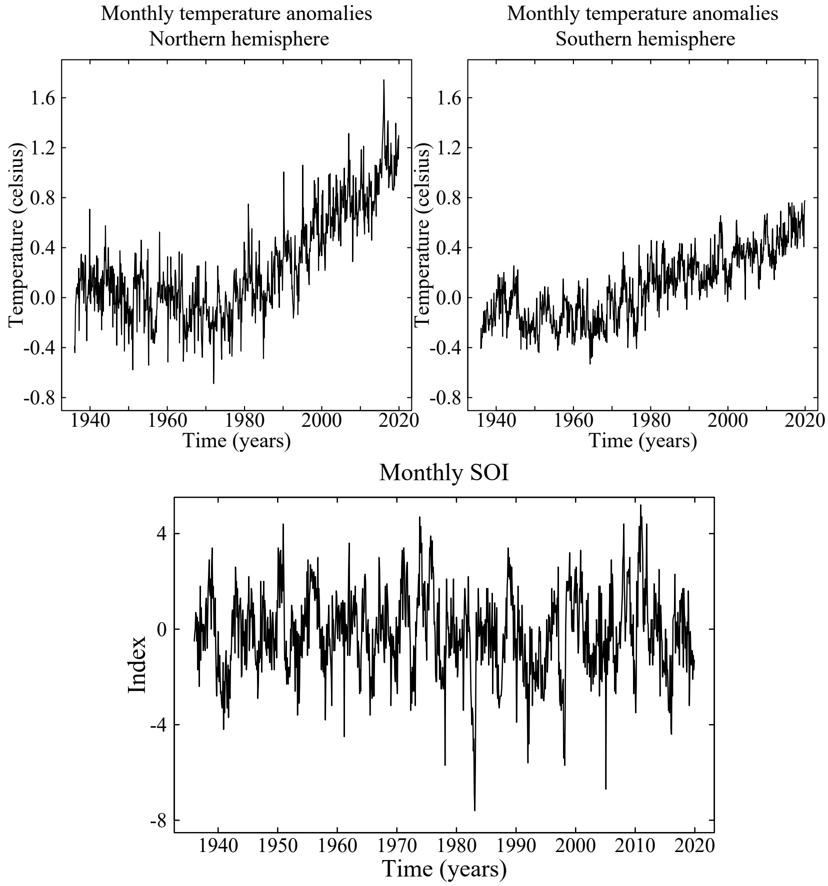

Figure 1. Time series plots for monthly temperature anomalies from the northern hemisphere (top left), monthly temperature anomalies from the southern hemisphere (top right), and the monthly SOI (bottom) during the period 01/1936–12/2019.

than the proposed SPLL method does, especially when there are a lot of zero coefficients, as in the 2-2-16 configuration. Therefore, we believe the model produced by the SPLL method is more reasonable. The climate science literature tends to find a lag between the SOI and temperature anomalies, mainly by relying on the correlation of annual averages calculated starting from different months. In contrast, the current analysis allows for the possibility that temperature anomalies may be associated with multiple lags of the SOI.

5. Conclusion

In addition to the current context, the proposed stratified penalization method sheds new light on the broader problem of how to incorporate penalization into

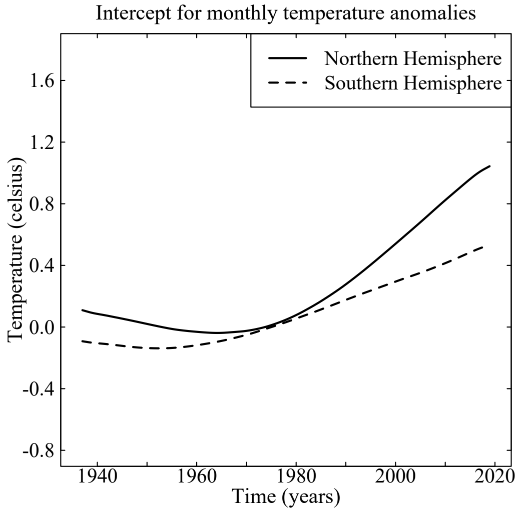

Figure 2. Estimated time-varying coefficients $\hat{\boldsymbol{\beta}}_0(\cdot) = \{\hat{\beta}_{0,1}(\cdot), \hat{\beta}_{0,2}(\cdot)\}^\top$ for the northern (solid) and southern (dashed) hemispheres.

kernel smoothing for labeling variables into more than two categories in a nested structure. Levina, Rothman and Zhu (2008) considered a nested LASSO method when different variables have a natural ordering for bicategory labeling, whereas in the current setting, each variable has its own nested structure for labeling. We expect that the proposed stratified penalization method can be generalized and will be useful in other problems that involve multi-category labeling with a nested structure, such as when the time-constant label is replaced by a more general parametric label. In addition, one may consider multiple parametric labels in a nested structure, such as polynomials with nested orders. However, detailed formulations of such problems and developing stratified penalization variants as their solutions are beyond the scope of this study, and are left to future work.

Supplementary Material

The online Supplementary Material contains a detailed description of the iterative algorithm for computing the proposed SPLL estimator, additional simulation results, and technical proofs of our results in Sections 2 and 3.

Acknowledgments

We thank the editor, associate editor, and two anonymous reviewers for their helpful comments and suggestions. We also thank Professor Ngai Hang Chan for mentioning the reference Chan, Gao and Palma (2022) that relates to the thesis of Gao (2019). The research was supported by the NSF CAREER Award DMS-1848035/DMS-2131821.

References

Ahmad, I., Leelahanon, S. and Li, Q. (2005). Efficient estimation of a semiparametric partially linear varying coefficient model. *The Annals of Statistics* **33**, 258–283.

Antoniadis, A. (1997). Wavelets in statistics: A review. *Journal of the Italian Statistical Society* **6**, 97–130.

Cai, Z., Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series. *Journal of the American Statistical Association* **95**, 941–956.

Chan, N. H., Gao, L. and Palma, W. (2022). Simultaneous variable selection and structural identification for time-varying coefficient models. *Journal of Time Series Analysis* **43**, 511–531

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association* **107**, 1533–1545.

Cheng, B. and Tong, H. (1998). K-stationarity and wavelets. *Journal of Statistical Planning and Inference* **68**, 129–144.

Cheng, M.-Y., Honda, T. and Zhang, J.-T. (2016). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *Journal of the American Statistical Association* **111**, 1209–1221.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The Annals of Statistics* **25**, 1–37.

Dahlhaus, R., Neumann, M. H. and Sachs, R. V. (1999). Nonlinear wavelet estimation of time-varying autoregressive processes. *Bernoulli* **5**, 873–906.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* **32**, 407–499.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications*. Chapman & Hall, London.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli* **11**, 1031–1057.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics* **27**, 1491–1518.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying-coefficient models. *Scandinavian Journal of Statistics* **27**, 715–731.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and its Interface* **1**, 179–195.

Gao, J. and Hawthorne, K. (2006). Semiparametric estimation and testing of the trend of temperature series. *The Econometrics Journal* **9**, 332–355.

Gao, L. (2019). *Simultaneous Variable Selection and Structural Identification for Time-Varying Coefficient Models*. Ph.D. Thesis. The Chinese University of Hong Kong, Hong Kong.

Giurcanu, M. and Spokoiny, V. (2004). Confidence estimation of the covariance function of stationary and locally stationary processes. *Statistics & Decisions* **22**, 283–300.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *Journal of the Royal Statistical Society. Series B (Methodological)* **59**, 319–351.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* **28**, 1356–1378.

Lee, W. and Liu, Y. (2012). Simultaneous multiple response regression and inverse covariance matrix estimation via penalized Gaussian maximum likelihood. *Journal of Multivariate Analysis* **111**, 241–255.

Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *The Annals of Applied Statistics* **2**, 245–263.

Li, D., Chen, J. and Lin, Z. (2009). Variable selection in partially time-varying coefficient models. *Journal of Nonparametric Statistics* **21**, 553–566.

Li, R. and Liang, H. (2008). Variable selection in semiparametric regression modeling. *The Annals of Statistics* **36**, 261–286.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *The Annals of Statistics* **26**, 1–47.

McLean, J. (2014). Late twentieth-century warming and variations in cloud cover. *Atmospheric and Climate Sciences* **4**, 727–742.

Moulines, E., Priouret, P. and Roueff, F. (2005). On recursive estimation for time varying autoregressive processes. *The Annals of Statistics* **33**, 2610–2654.

Ombao, H., von Sachs, R. and Guo, W. (2005). SLEX analysis of multivariate nonstationary time series. *Journal of the American Statistical Association* **100**, 519–531.

Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **27**, 204–237.

Privalsky, V. E. and Jensen, D. T. (1995). Assessment of the influence of ENSO on annual global air temperatures. *Dynamics of Atmospheres and Oceans* **22**, 161–178.

Rao, T. S. (1970). The fitting of non-stationary time-series models with time-dependent parameters. *Journal of the Royal Statistical Society: Series B (Methodological)* **32**, 312–322.

Rothman, A. J., Levina, E. and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 947–962.

Tang, Y., Wang, H. J., Zhu, Z. and Song, X. (2012). A unified variable selection approach for varying coefficient models. *Statistica Sinica* **22**, 601–628.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288.

Turlach, B. A., Venables, W. N. and Wright, S. J. (2005). Simultaneous variable selection. *Technometrics* **47**, 349–363.

Van Bellegem, S. and Dahlhaus, R. (2006). Semiparametric estimation by model selection for locally stationary processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 721–746.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association* **104**, 747–757.

Wang, L., Li, H. and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association* **103**, 1556–1569.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 14150–14154.

Wu, W. B. and Zhao, Z. (2007). Inference of trends in time series. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69**, 391–410.

Xia, Y., Zhang, W. and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika* **91**, 661–681.

Xue, L. and Qu, A. (2012). Variable selection in high-dimensional varying-coefficient models with global optimality. *Journal of Machine Learning Research* **13**, 1973–1998.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

Zhang, T. (2013). Clustering high-dimensional time series based on parallelism. *Journal of the American Statistical Association* **108**, 577–588.

Zhang, T. (2015). Semiparametric model building for regression models with time-varying parameters. *Journal of Econometrics* **187**, 189–200.

Zhang, T. (2016). Testing for jumps in the presence of smooth changes in trends of nonstationary time series. *Electronic Journal of Statistics* **10**, 706–735.

Zhang, T. and Wu, W. B. (2011). Testing parametric assumptions of trends of a nonstationary time series. *Biometrika* **98**, 599–614.

Zhang, T. and Wu, W. B. (2012). Inference of time-varying regression models. *The Annals of Statistics* **40**, 1376–1402.

Zhang, T. and Wu, W. B. (2015). Time-varying nonlinear regression models: Nonparametric estimation and model selection. *The Annals of Statistics* **43**, 741–768.

Zhang, W., Lee, S.-Y. and Song, X. (2002). Local polynomial fitting in semivarying coefficient model. *Journal of Multivariate Analysis* **82**, 166–188.

Zheng, X. and Basher, R. E. (1999). Structural time series models and trend detection in global and regional temperature series. *Journal of Climate* **12**, 2347–2358.

Zhou, Z. and Wu, W. B. (2010). Simultaneous inference of linear models with time varying coefficients. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 513–531.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics* **36**, 1509–1533.

Ting Zhang

Department of Statistics, University of Georgia, 310 Herty Drive, Athens, GA 30602, USA.

E-mail: tingzhang@uga.edu

Weiliang Wang

Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA.

E-mail: weiliang@bu.edu

Yu Shao

Department of Mathematics and Statistics, Boston University, 111 Cummington Mall, Boston, MA 02215, USA.

E-mail: yshao19@bu.edu

(Received February 2021; accepted August 2021)