Estimation and Variable Selection in Sequential Linear Models: SCAD-Penalized Method with Applications

Yiwen Yuan , Junfeng Shang * , Chao Gu

Posted Date: 18 March 2026

doi: 10.20944/preprints202603.1439.v1

Keywords: Lasso method; adaptive Lasso; SCAD penalty; sequential linear models; time series; high-dimensional modeling

 doi:10.20944/preprints202603.1439.v1

Article

Estimation and Variable Selection in Sequential Linear Models: SCAD-Penalized Method with Applications

Yiwen Yuan [1], Junfeng Shang [1,*] and Chao Gu [2]

[1] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403, USA
[2] Department of Mathematics and Data Analytics, The Citadel-The Military College of South Carolina, Charleston, SC 29409, USA
* Correspondence: jshang@bgsu.edu

Abstract

Sequential linear models can be adopted to describe the data where the response variable depends on lagged outcomes and fixed effects variables. In estimation and variable selection and for predicting the response variable with high accuracy, we propose the penalized method based on Smoothly Clipped Absolute Deviation Penalty (SCAD) in the sequential linear models. We conduct the simulations where the SCAD-penalized method is compared with other methods including the ordinary least squares (OLS), Lasso, and adaptive Lasso in the sequential linear models. The simulation results demonstrate that the SCAD-penalized method in the sequential linear models excels in estimation with better accuracy and precision and in variable selection with better prediction. We apply the proposed method to two real data sets for further illustrating the performance of the SCAD-penalized method in the sequential linear modeling.

Keywords: Lasso method; adaptive Lasso; SCAD penalty; sequential linear models; time series; high-dimensional modeling

1. Introduction

In the clinical field, biological studies, and econometrics, many data are dynamic and change over time. To model and analyze such dynamic relationships between variables in time series data, the Autoregressive Distributed Lag (ADL, see Pesaran and Shin (1999)) model is a powerful tool and has gained significant traction, for its ability to capture both short-term and long-term dependencies to facilitate the exploration of complex temporal dynamics (Keele and Kelly 2006). The ADL model incorporates both autoregressive terms (lags of the dependent variable) and distributed lag terms (lags of the independent variables), allowing researchers to examine how past values of the explanatory variables influence the current value of the dependent variable. This makes it particularly useful in contexts where past information plays a crucial role in shaping future outcomes.

For many situations, the time-varying dependent variable of interest not only depends on its previous state but also is correlated with time-varying predictors. For such situations, the sequential modeling approach can be adopted, see Shestopaloff et al. (2022). This sequential linear modeling approach can effectively model the dependent variable by its last state and predictors in sequential time points. Designed to handle sequential information, sequential linear models are more adapted for absorbing time-series process and predicting the dependent variable by time-varying lagged outcomes and predictors. For example, to cope with the sequential predictive modeling, Shestopaloff et al. (2022) involved sequentially predicting outcomes at intermediate time points, which are then used as predictors in the model for the subsequent time point, as interventions involving recovery times or chronic conditions entail outcome measures at various intermediate follow-up points.

An ordinary least squares (OLS) regression can also model dynamic data, and the past values of the lagged dependent response variable are incorporated to demonstrate the influence of dynamic

 doi:10.20944/preprints202603.1439.v1

variables on the prediction process. However, Keele and Kelly (2017) proposed that there exists a problem of correlation effects because the lagged dependent variable results in the coefficients of the independent variables being biased towards lower values. To deal with this problem in the sequential time-series data, Keele and Kelly (2017) introduced the lagged dependent variable in the ordinary least squares (OLS) regression to capture the prediction, where the lagged dependent variable coefficient indicates the timing effect based on the independent variable. Qiu et al. (2015) combined the estimated residuals with SCAD-penalty and Yule-Walker equations to determine the order of the autoregressive error and estimate the autoregressive parameters. Nicholson et al. (2020) proposed a new class of hierarchical lag structures (HLag) which allows the notion of lag selection into a convex regularizer and improves the forecasting performance.

Motivated by the ADL and above related work, we employ a sequential linear modeling approach and utilize sequential linear models to describe the dependent variable through incorporating the lagged dependent variable and independent variables varying with time. Like the ADL model, sequential linear modeling faces challenges in estimation due to overfitting and multicollinearity, particularly when the number of potential explanatory variables is large. To address these challenges, recent advances in penalized regression techniques, such as the Lasso (Tibshirani 1996) and its variants, have gained popularity for variable selection and parameter estimation in dynamic models such as the ADL. These methods impose penalties on the regression coefficients, allowing for sparse models that are both interpretable and predictive.

We propose using the SCAD-penalized method to estimate and select variables in the sequential linear modeling. The SCAD method, introduced by Fan and Li (2001), is known for its ability to handle large-scale variable selection problems while providing less bias in coefficient estimation compared to traditional Lasso. The SCAD-penalized method has become a preferred choice in high-dimensional regression models due to its superior ability to perform variable selection while preserving model accuracy. We aim to evaluate the performance of the sequential linear modeling estimated with the SCAD penalty and compare it with the Lasso and adaptive Lasso (Zou 2006; Zou and Hastie 2005), which have been widely applied for their effectiveness in variable selection and regularization. Our study will examine strengths and weaknesses of these penalized regression approaches by comparing their ability in estimation and variable selection. Specifically, we will assess whether the SCAD-penalized method outperforms Lasso and adaptive Lasso in terms of predictive accuracy and coefficient estimation. By doing so, we aim to contribute to the literature on penalized regression techniques and their applications to dynamic modeling, providing insights into the choice of regularization methods for dynamic and time series data.

As follows, Section 2 introduces the sequential linear models. Section 3 discusses the SCAD-penalized method along with the Lasso and adaptive Lasso for the sequential linear modeling. The simulations are presented in Section 4. Two applications are conducted in Section 5. Section 6 concludes and discusses.

2. Sequential Linear Models

We have a sample of data with repeated outcomes $y_{i,t}$ with $i = 1, ..., n$ and a sequence of timelines $t = 1, ..., T$.

At each time point t, we have a model that sequentially predicts the outcomes of $y_{i,1}, ..., y_{i,t}$ based on the predictors of $\boldsymbol{x_{i,t}} = (1, x_{i1,t}, ..., x_{ip,t})^T$ with a parameter/coefficient vector $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$, p is the number of predictors, and a parameter/coefficient vector $\boldsymbol{\gamma_t}$ with its length varying with t, $t = 1, \cdots, T$, denoted by $\boldsymbol{\gamma_t} = (\gamma_{1,t}, \cdots, \gamma_{t,t})^T$.

We can also have subsequent outcomes at any t time point denoted as $\boldsymbol{y_{i,t}} = (y_{i,1}, ..., y_{i,t})^T$. As the timeline t increases for one unit at each time, the outcomes of $\boldsymbol{y_{i,t}}$ can be predicted based on the covariates or predictors of $\boldsymbol{x_{i,t}}$ and the prior predicted outcomes of $\boldsymbol{y_{i,t-1}}$. The distribution of the predicted outcomes follows a normal distribution denoted as $\hat{y}_{i,t} \sim N\big(E(\hat{y}_{i,t}), V(\hat{y}_{i,t})\big)$. We can repeat the previous models above until the end timeline $t = T$.

 doi:10.20944/preprints202603.1439.v1

Before proceeding with the sequential modeling, we can visualize a picture of an example. Measures on patients can be collected at different time points, and these measures include outcomes such as a measure detecting an illness and predictors which are all time-varying such as many factors which affect this illness. We can easily assume that the outcomes at the current point can be predicted by the predictors at the current time point and the state of previous outcomes.

We can therefore have sequential linear models as follows.

At the first time point $t = 1$, the linear model is written as

$$y_{i,1} = \boldsymbol{x}_{\boldsymbol{i,1}}^T \boldsymbol{\beta} + \epsilon_{i,1},$$

where the error term is normally distributed as $\epsilon_{i,1} \sim N(0, \sigma^2)$ and σ^2 is the variance.

At the second time point $t = 2$, the subsequent linear model is written as

$$y_{i,2} = \boldsymbol{x}_{\boldsymbol{i,2}}^T \boldsymbol{\beta} + \boldsymbol{y}_{\boldsymbol{i,1}}^T \boldsymbol{\gamma_1} + \epsilon_{i,2},$$

where the error term is normally distributed as $\epsilon_{i,2} \sim N(0, \sigma^2)$, σ^2 is the variance, and the outcomes $\boldsymbol{y_{i,1}} = (y_{i,1})^T$.

At the third time point $t = 3$, the subsequent linear model is written as

$$y_{i,3} = \boldsymbol{x}_{\boldsymbol{i,3}}^T \boldsymbol{\beta} + \boldsymbol{y}_{\boldsymbol{i,2}}^T \boldsymbol{\gamma_2} + \epsilon_{i,3},$$

where the error term is normally distributed as $\epsilon_{i,3} \sim N(0, \sigma^2)$, σ^2 is the variance, and the outcomes $\boldsymbol{y_{i,2}} = (y_{i,1}, y_{i,2})^T$.

For any t time point, we can have the sequential linear models written as

$$y_{i,t} = \boldsymbol{x}_{\boldsymbol{i,t}}^T \boldsymbol{\beta} + \boldsymbol{y}_{\boldsymbol{i,t-1}}^T \boldsymbol{\gamma_{t-1}} + \epsilon_{i,t}, \tag{1}$$

where the error term is normally distributed as $\epsilon_{i,t} \sim N(0, \sigma^2)$, σ^2 is the variance, and the outcomes $\boldsymbol{y_{i,t-1}} = (y_{i,1}, \cdots, y_{i,t-1})^T$. Repeating the above subsequent models until the end timeline $t = T$.

We remark that the sequential linear models involve the lagged dependent variables over each time point, so the key assumptions need to be checked before fitting the sequential linear model and predicting, and they are described in what follows.

1. The relationship between the dependent variable and the independent variable is linear.
2. The error terms at each time point should be independent and follow normal distribution with constant variance.
3. The independent variables are not perfectly collinearity at each time point.
4. The independent variables are not correlated with the error term.
5. The incorporated fixed effect for each observation in the sequential linear model is allowed.

After model (2) is introduced, these key assumptions will be phrased in mathematical ways for the proofs. But for now, general ideas for the model assumptions can be easily grasped from the above description. For the sequential modeling as shown in Equation (1), we can see that this model is a combination of linear predictors and autoregressive time-series regressions. In the methodology section, we will propose penalized methods to conduct estimation. Such methods own a sparsity property, implying that adopting a penalty term can shrink some parameter estimates to zero. With the sparsity property, the employed methods for estimation can simultaneously select variables. Estimating parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma_{t-1}}$ in model (1) targets to select the predictors/features within the parameter vector $\boldsymbol{\beta}$ and the number of lags for autoregressive regression in $\boldsymbol{\gamma_{t-1}}$. Regard to the number of lags, for example, if the estimation turns out only $\gamma_{t-2,t}$ and $\gamma_{t-1,t}$ are significant (do not shrink to zero), we can conclude that the response variable at the current time point can be predicted by those for the last two points, which means the lag is 2 in the part of autoregressive regression. As time goes on, the response variables become increasingly uncorrelated, and the past lagged response variables will be shrunk to zero. Theoretically, it is interpretable that the response variable at the

 doi:10.20944/preprints202603.1439.v1

current time point only depends on some close points. So, the lag cannot be a large integer. Our purpose is to propose a well-functioning penalized method to select the predictors in $\boldsymbol{\beta}$ and lag in $\boldsymbol{\gamma_{t-1}}$.

Model (1) serves as the sequential modeling with linear models with time-varying predictors and autoregressive time-series model with the selected lag by the proposed method. To have a succinct form for it, we can have a $n * (p+1)$ matrix containing all the predictors denoted by

$$\boldsymbol{X_t} = \begin{bmatrix} 1 & x_{11,t} & x_{12,t} & ... & x_{1p,t} \\ 1 & x_{21,t} & x_{22,t} & ... & x_{2p,t} \\ 1 & x_{31,t} & x_{32,t} & ... & x_{3p,t} \\ \vdots & & \vdots & & \vdots \\ 1 & x_{n1,t} & x_{n2,t} & ... & x_{np,t} \end{bmatrix},$$

where again, n is the sample size and p is the number of predictors.

The response/outcome variables in the timeline t are represented by

$$\boldsymbol{Y_t} = \begin{bmatrix} y_{1,1} & y_{1,2} & ... & y_{1,t} \\ y_{2,1} & y_{2,2} & ... & y_{2,t} \\ y_{3,1} & y_{2,3} & ... & y_{3,t} \\ \vdots & & \vdots & \vdots \\ y_{n,1} & y_{n,2} & ... & y_{n,t} \end{bmatrix},$$

which is an $n * t$ vector matrix.

To represent model (1), the outcome at time point t, we can use the outcome $\boldsymbol{Y_{t-1}}$ at time point $t-1$ as the predictors, and all the predictors are combined into one matrix denoted by $\boldsymbol{Z_t} = \left[\boldsymbol{X_t}, \boldsymbol{Y_{t-1}}\right]$. This notation will be used in the next section when introducing the methods.

For any t time point and a succinct format, we can rewrite the sequential linear models in Equation (1) by

$$\boldsymbol{y_t} = \boldsymbol{X_t\beta} + \boldsymbol{Y_{t-1}\gamma_{t-1}} + \boldsymbol{\epsilon_t}, \tag{2}$$

where $\boldsymbol{X_t} = (\boldsymbol{X}_{1,t}, ..., \boldsymbol{X}_{p,t})^T$, which is the predictor matrix containing vectors at time t; $\boldsymbol{y_t} = (y_{1,t}, ..., y_{n,t})^T$, which is the response/outcome vector at time t; and $\boldsymbol{\epsilon_t}$ is the error term vector containing all error terms at t time point, i.e., $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \cdots, \epsilon_{n,t})^T$. We have i.i.d. and normally distributed $\epsilon_{1,t}, \cdots, \epsilon_{n,t}$ with variance σ^2. Again, we can repeat model (2) until the end timeline $t = T$. When $t = 1$, the initial time point, the model does not involve the term $\boldsymbol{Y_{t-1}\gamma_{t-1}}$, and $\boldsymbol{\gamma_{t-1}} = (\gamma_{1,t-1}, \cdots, \gamma_{t-1,t-1})^T$.

To adopt penalized methods in model (2), we can have model assumptions for weak or not autocorrelation and heteroscedasticity. Let s be an positive integer for labeling the lags. Without loss of generality and similar to those in Stock and Watson (2003) and Gu and Ratnasingam (2025), the assumptions are:

A1. $E(\boldsymbol{\epsilon_t}|\boldsymbol{y_{t-1}}, \boldsymbol{y_{t-2}}, \ldots, \boldsymbol{y_{t-s}}, \boldsymbol{X_t}) = 0$;

A2. $\boldsymbol{y_t}$ and $\boldsymbol{y_{t-s}}$ become independent as s gets large, i.e.,

$$\lim_{s\to\infty} \text{Corr}(\boldsymbol{y_t}, \boldsymbol{y_{t-s}}) = 0;$$

A3. The random variables $(\boldsymbol{y_t}, \boldsymbol{X}_{1,t}, \ldots, \boldsymbol{X}_{p,t})$ have a stationary distribution at each time t;

A4. There is no perfect multicollinearity;

A5. $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \epsilon_{2,t}, \cdots, \epsilon_{n,t})^T$ does not vary over time and represents an $n-$vector of independent identically distributed random variables, with $E(\epsilon_{1,t}) = 0$, $\text{Var}(\epsilon_{1,t}) = \sigma^2 < \infty$, and $E(\epsilon_{1,t}^4) < \infty$;

We note that A1 assumes that the error term does not involve information for other variables in the model and is not affected by the predictors and lags. A2 assumes an asymptotic uncorrelated

 doi:10.20944/preprints202603.1439.v1

relationship with a weakening of linear dependence over time rather than full independence, but does not imply full statistical independence unless additional distributional assumptions are imposed and needed. A3 assumes a stationary distribution, and this is a commonly used one for time series models to ensure the model is not going out of boundary. A4 and A5 guarantee the model and parameters are estimable.

For the sequential models in Equations (1) and (2), we hope to conduct the selection and estimation of variables and then we can have the most appropriate model to describe the data with varying time, predict the outcomes, and other inferences.

After we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma_{t-1}}$ with $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}_{\boldsymbol{t-1}}$ respectively, we can have the prediction of the dependent variable for the next time point as

$$\hat{\boldsymbol{y}}_{\boldsymbol{t}} = \boldsymbol{X_t}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{Y}}_{\boldsymbol{t-1}}\hat{\boldsymbol{\gamma}}_{\boldsymbol{t-1}}, \tag{3}$$

where $t > 1$. Equation (3) is the formulation for the prediction. Each time point's predicted output is a function of the predictors at the current point and predicted output from the previous step. There are various methods available for making comparisons used for variable selection, optimizing the parameters, and making the model prediction. The accuracy of the variance estimation is also measured to show a similar or better performance across the subsequent data sets and various continuous outcomes.

The predicted response variable $\boldsymbol{y_t}$, at any time point $t > 1$, incorporates the lagged response variable $\boldsymbol{Y_{t-1}}$ at the previous time points with the assumptions that the independent and identically distributed residuals follow normal distributions. The coefficients in front of the lagged term $\boldsymbol{Y_{T-1}}$ often decrease as the time lag increases (Brockwell and Davis 2002).

3. SCAD-Penalized Method for Estimation and Variable Selection

As sequential linear models (1) and (2) are introduced in the last section, we will conduct the parameter estimation and variable selection for the sequential modeling. To facilitate prediction with the most appropriate model, we adopt the penalized methods to improve the performance in estimation and variable selection and mitigate overfitting. One typical penalized method is the Lasso (Tibshirani 1996), which introduces a penalty term, adding advantages that estimators possess consistency and sparsity and making the model more interpretable and efficient in predicting and handling dependencies in the sequential modeling. Beginning with the Lasso method, this section will then introduce the adaptive Lasso and propose the SCAD-penalized method in the sequential linear models for estimation and variable selection.

3.1. SCAD-Penalized Method for Estimation and Variable Selection

In what follows, we present the methods that conduct estimation and variable selection adopted in this study. The basic idea is to penalize Equation (2) to obtain the estimators for the parameters. Let $\boldsymbol{Z_t}$ be the matrix containing all predictors $\boldsymbol{X_t}$ and previous outcomes $\boldsymbol{Y_{t-1}}$ denoted by $\boldsymbol{Z_t} = (\boldsymbol{X_t}, \boldsymbol{Y_{t-1}})$, and the elements of $\boldsymbol{Z_t}$ is $z_{i,t}$ for $i = 1, 2, \cdots, n$ at time point t. When we do regression in model (2), the design matrix is $\boldsymbol{Z_t}$. For the regression in this model, let $\boldsymbol{\alpha_t}$ be the parameter vector containing both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma_{t-1}}$ denoted by $\boldsymbol{\alpha_t} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}_{\boldsymbol{t-1}}^T)^T$. We note that the dimension of parameter vector $\boldsymbol{\beta}$ is $p+1$ and the dimension of parameter vector $\boldsymbol{\gamma_{t-1}}$ is $t-1$ for each time point t, so the length of $\boldsymbol{\alpha_t}$ is $(p+1)+(t-1)$ for each time point t. That is, the dimension of the parameter vector $\boldsymbol{\alpha_t}$ is $p+t$ for time point t. We denote each parameter element in vector $\boldsymbol{\alpha_t}$ by α_{jt} with $j = 1, \cdots, p+t$ for each time point t.

To estimate parameter vector $\boldsymbol{\alpha_t}$, we start with the Lasso method. According to model (2), we can estimate $\boldsymbol{\alpha_t}$ through minimizing the penalized squared errors, and the estimators $\hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{\boldsymbol{l}}$ are derived by

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{\boldsymbol{l}} = \arg\min_{\boldsymbol{\alpha_t}} ||\boldsymbol{y_t} - \boldsymbol{Z_t}\boldsymbol{\alpha_t}||^2 + \lambda \sum_j |\alpha_{jt}|, \tag{4}$$

 doi:10.20944/preprints202603.1439.v1

where again, we have the parameter element α_{jt} with $j = 1, \cdots, p+t$ for time point t, and the α_{jt} is the jth parameter element at time point t, and λ is the tuning parameter.

The Lasso method was initiated by Tibshirani (1996). Similar to the theory in it, it can be proved that the Lasso method of Equation (4) is a regularization method for parameter estimation and variable selection simultaneously and the estimators possess consistency and sparsity. Specifically, the penalty term $l_1 = \lambda \sum |\alpha_{jt}|$ can shrink the parameters of non-significant variables to zero, and the Lasso method therefore owns the property of sparsity in estimation, and relying on this property, variable selection can be implemented. Starting from the Lasso method, we can use different penalty terms in the sequential linear models as shown in Equation (2) for improvement on estimation and variable selection.

The sequential linear modeling involves modeling observed/generated data at each time point, and each model depends on the previous ones. The model often exhibits multi-collinearity for which predictors are highly correlated. The Lasso regression method in the sequential linear model helps handle multi-collinearity by automatically selecting relevant predictors and potentially excluding correlated ones. The Lasso regression method addresses temporal dependencies by considering the sequential order of observations. It can be used to select relevant lagged values or features that contribute to predicting future values. In addition, the Lasso regression may be applied to model time trends and automatically select the most influential time points.

When assigning different weights to different parameters in $\boldsymbol{\alpha_t}$, it is called adaptive Lasso as discussed in Zou (2006), essentially an l_1 penalization method. With certain regularity conditions, they derive the estimators of parameter vector $\boldsymbol{\alpha_t}$ as

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{a} = \arg\min_{\boldsymbol{\alpha_t}} ||\boldsymbol{y_t} - \boldsymbol{Z_t}\boldsymbol{\alpha_t}||^2 + \lambda \sum_j w_j |\alpha_{jt}|, \tag{5}$$

where again, we have the parameter element α_{jt} with $j = 1, \cdots, p+t$ for each time point t, and w_j is the weight for each parameter element α_{jt} (e.g., reciprocal of the parameter), and λ is the tuning parameter and needs to be selected.

The Lasso and adaptive Lasso in Equations (4) and (5) are employed for comparison in the performance of estimation and variable selection with the SCAD-penalized method to be introduced below.

We address that the penalized method with the SCAD penalty as in Fan and Li (2001) is applied to estimate the parameters in sequential linear models (1) and (2), and the estimation can be completed by using Equations (6) and (7) as below. With certain regularity conditions as in Fan and Li (2001), Zhao and Yu (2006), Kim et al. (2008), etc, we can derive the estimators of parameter vector $\boldsymbol{\alpha_t} = (\boldsymbol{\beta}^{\boldsymbol{T}}, \boldsymbol{\gamma}_{\boldsymbol{t-1}}^{\boldsymbol{T}})^{\boldsymbol{T}}$ by minimizing the penalized function shown in Equation (6), which is the sum of the squared error and a SCAD penalty term, and now we have

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{t}}^{s} = \arg\min_{\boldsymbol{\alpha_t}} ||\boldsymbol{y_t} - \boldsymbol{Z_t}\boldsymbol{\alpha_t}||^2 + \sum_j P_\lambda^s(|\alpha_{jt}|), \tag{6}$$

where $|\cdot|$ denotes the absolute value, $p_\lambda^s(|\cdot|)$ are the penalty functions depending on λ and are allowed not to be identical for all $j = 1, \cdots, p+t$, according to the formula for estimating the coefficients in the SCAD-penalized regression as in Fan and Li (2001) and Lim et al. (2008), the penalty function is explicitly written as

$$p_\lambda^s(|\alpha_{jt}|) = \begin{cases} \lambda|\alpha_{jt}| & \text{if} \quad |\alpha_{jt}| \le \lambda, \\ (2a\lambda|\alpha_{jt}| - |\alpha_{jt}|^2 - \lambda^2)/[2(a-1)] & \text{if} \quad \lambda < |\alpha_{jt}| \le a\lambda, \\ (a+1)\lambda^2/2 & \text{if} \quad |\alpha_{jt}| > a\lambda, \end{cases} \tag{7}$$

 doi:10.20944/preprints202603.1439.v1

where $a > 2$, and a is an additional tuning parameter to λ and controls the smoothness of the penalty function. λ is the tuning parameter that controls the strength of the penalty. Equation (7) forms a quadratic spline function with knots at λ and $a\lambda$. It is continuous, with the first derivative of

$$\lambda\left\{I(|\alpha_{jt}| \leq \lambda) + \frac{(a\lambda - |\alpha_{jt}|)_+}{(a-1)\lambda} I(|\alpha_{jt}| > \lambda)\right\}, \tag{8}$$

where $(\cdot)_+ = \max\{0, \cdot\}$. Expression (8) implies the SCAD penalty is continuously differentiable on $\mathbb{R}$ except for being singular at the origin, with the derivatives being all zeroes outside the interval $[-a\lambda, a\lambda]$.

We conduct the SCAD-penalized regression using Equation (6), and the SCAD-penalized estimators possess the oracle property at time t, which can be proved as in the references such as Kim et al. (2008). With certain regular conditions, the SCAD-penalized estimators are consistent in selecting the true model and are asymptotically normally distributed. Specifically, in the sequential linear model in (2), the design matrix is $\boldsymbol{Z_t}$, and we can partition it as $\boldsymbol{Z_t} = (\boldsymbol{Z_t}^{(1)}, \boldsymbol{Z_t}^{(2)})$, where $\boldsymbol{Z_t}^{(1)}$ represents the design matrix corresponding to nonzero regression coefficients and $\boldsymbol{Z_t}^{(2)}$ represents the design matrix corresponding to zero regression coefficients. We can also write $\boldsymbol{Z_t} = (\boldsymbol{Z}_t^1, \cdots, \boldsymbol{Z}_t^{p+t})$, i.e., $\boldsymbol{Z}_t^j$ represents the design vector corresponding to the jth parameter coefficient for the predictors or the lags of $\boldsymbol{y_t}$, $j = 1, \cdots, p+t$, at time point t.

For a clear notation, we have $\hat{\boldsymbol{\alpha}}_t^s = (\hat{\alpha}_{1t}^s, \ldots, \hat{\alpha}_{(p+t)t}^s)^\top$. Let $\mathcal{A}_t = \{j, t : \alpha_{jt} \neq 0\}$ be an index set of the coefficients for significant variables and suppose $\mathcal{A}_t^* = \{j, t : \hat{\alpha}_{jt}^s \neq 0\}$ is the index set of the significant SCAD estimators based on the historical sample at time t with the cardinality $|\mathcal{A}_t^*| = K^*$. We organize $\hat{\boldsymbol{\alpha}}_t^s$ as $(\hat{\boldsymbol{\alpha}}_t^{s(1)}, \hat{\boldsymbol{\alpha}}_t^{s(2)})$, accordingly. We also let $\boldsymbol{M_t^1} = (\boldsymbol{Z_t}^{(1)})^T \boldsymbol{Z_t}^{(1)}/n$.

As shown in Gu and Ratnasingam (2025), to ascertain the asymptotic properties and derive the limiting distribution of the estimators within the high-dimensional and sparse framework, we incorporate the following supplementary regularity conditions (Zhao and Yu (2006)), in conjunction with $A1 - A5$:

A6. There exists a positive constant C_1 at time t such that $\frac{1}{n}\boldsymbol{Z}_k^\top \boldsymbol{Z}_k \leq C_1$ for all $k = 1, \ldots, K$ and for all n;

A7. There exists a positive constant C_2 such that $\xi^T \boldsymbol{M_t^1} \xi \geq C_2$ for all $||\xi||_2^2 = 1$;

A8. $K^* = \mathcal{O}(n^{b_1})$ for some $0 < b_1 < 1$;

A9. There exist positive constants b_2 and C_3 such that $b_1 < b_2 \leq 1$ and

$$n^{(1-b_2)/2} \min_{k=1,\ldots,p^*,\ldots,p^*+S} |\alpha_t^{(1)}| \geq C_3.$$

We have these assumptions on the design matrix for regularity conditions to achieve oracle/sparsity estimators in the SCAD-penalized method. As in A6, we first assume the squared mean of each $\boldsymbol{Z}_t^j$ is bounded by a positive value, that is, there is a positive constant C_1 such that $(\boldsymbol{Z}_t^j)^T \boldsymbol{Z}_t^j/n \leq C_1$, for all $j = 1, \cdots, p+t$, at time t. This regularity condition is shown as A1 in Kim et al. (2008). When both the predictors and the lags of $\boldsymbol{y_t}$ are normalized, this condition may become trivial.

As in $A7$, for the oracle/sparsity property, we also need to necessitate a condition for maintaining eigenvalues of matrix $\boldsymbol{M_t^1}$ properly positive to ensure a proper inverse, that is, the design matrix satisfies the restricted eigenvalue condition to ensure that the predictors and lags of $\boldsymbol{y_t}$ are spread out enough to allow for correct variable selection even if the number of predictors and lags of $\boldsymbol{y_t}$ is much larger than the sample size. This restricted eigenvalue condition can be further described, i.e., there is a positive constant C_2 such that $\xi^T \boldsymbol{M_t^1} \xi \geq C_2$ for all $||\xi||_2^2 = 1$. This condition is shown as A2 in Kim et al. (2008). Combining other regularity conditions including A3 and A4 in Kim et al. (2008), as shown in A8 and A9, we can prove the SCAD-penalized estimators from Equation (6) owns the oracle/sparsity property, the assumption $A8$ addresses the divergence rate of the variable dimension p compared to the sample size n, ensuring the number of significant predictors K^* grows at a controlled polynomial

 doi:10.20944/preprints202603.1439.v1

rate relative to n and preventing overfitting while maintaining a sparsity structure in the model. The last assumption mandates a gap of size n^{b_2} between the decay rates of $\alpha_t^{(1)}$ and $n^{-\frac{1}{2}}$, ensuring that the estimation is not overwhelmed by the error terms. See more details in Gu and Ratnasingam (2025).

Again, with these regularity conditions, we can prove the SCAD-penalized estimators from Equation (6) owns the oracle/sparsity property, which means that asymptotically the SCAD-penalized estimators shrink coefficients to zero for non-significant predictors or lags of $\boldsymbol{y_t}$ while preserving coefficients for significant predictors or lags of $\boldsymbol{y_t}$ in the SCAD-penalized regression. If only there are significant variables in the regression, this condition is satisfied. When we conduct the SCAD-penalized regression, the condition is met because the design matrix $\boldsymbol{Z_t}^{(1)}$ is not empty, i.e., there are significant variables in the regression.

In the context of the assumptions $A1 - A9$, we propose that the SCAD penalized estimators satisfy the *oracle* property at time t, the proof is not presented here and similar to that in (Gu and Ratnasingam (2025)).

We can therefore have the consequences of consistency in variable selection and estimation respectively, and they are written as:

- Consistency in variable selection, $\lim_{n\to\infty} P(\mathcal{A}_t^* = \mathcal{A}_t) = 1$;
- $\sqrt{n}\mathbb{A}_t(n^{-1}\boldsymbol{Z}_t^{(1)\top}\boldsymbol{Z}_t^{(1)}/\sigma^2)^{1/2}(\hat{\boldsymbol{\alpha}}_t^{s(1)} - \boldsymbol{\alpha}_t^{(1)}) \xrightarrow{d} N(\boldsymbol{0}, H)$, where $\mathbb{A}_t$ is an arbitrary matrix with $\mathbb{A}_t\mathbb{A}_t^\top \to H$ and H is a $K^* \times K^*$ nonnegative symmetric matrix which includes the elements of M_t in the set $\mathcal{A}_t$.

Simply speaking, consistency in variable selection means that asymptotically the probability of selecting correct variables for the model goes to one; consistency in estimation means the estimators will converge to true parameters or coefficients with a normal distribution.

As the result of the sparsity property, the SCAD-penalized method selects the true model or correctly selects significant variables in the model, which yields a sparse set of solutions and approximately unbiased coefficients for significant predictors or lags of $\boldsymbol{y_t}$. For notation brevity, we let $\hat{\boldsymbol{\alpha}}_t^{s(1)} = \hat{\boldsymbol{\alpha}}_t^s$. So, the solution to the SCAD-penalized regression is expressed as

$$\hat{\alpha}_{jt}^s = \begin{cases} \operatorname{sgn}(\hat{\alpha}_{jt})(|\hat{\alpha}_{jt}| - \lambda)_+ & \text{if} \quad |\hat{\alpha}_{jt}| \le 2\lambda, \\ [(a-1)\hat{\alpha}_{jt} - \operatorname{sgn}(\hat{\alpha}_{jt})a\lambda]/(a-2) & \text{if} \quad 2\lambda < |\hat{\alpha}_{jt}| \le a\lambda, \\ \hat{\alpha}_{jt} & \text{if} \quad |\hat{\alpha}_{jt}| > a\lambda, \end{cases} \tag{9}$$

where $\operatorname{sgn}(\cdot)$ represents the sign function and $\hat{\alpha}_{jt}$ is an estimator for α_{jt}, for example, the ordinary least squares estimator can be used. The optimal pair (λ, a) can be theoretically obtained through a two-dimensional grid search using criteria such as cross-validation, which can be quite computationally expensive. Fan and Li (2001) recommended setting $a = 3.7$ as a good choice for diverse scenarios. They further showed that the choice of a does not significantly improve the performance of the whole process compared to λ.

In the context of the sequential linear models, we employ the penalized method with the SCAD penalty as shown in Equations (7), (8), and (9) to enhance variable selection and refine the estimation of coefficients. The sparsity property, combined with the smooth penalization of the SCAD penalty, proves beneficial for handling the sequential and correlated nature of time series data.

We remark that the SCAD penalty (Fan and Li 2001) is more effective in parameter estimation and variable selection especially in high-dimensional data by promoting sparsity, achieving asymptotic unbiasedness, and ensuring oracle property. With similar advantages as the Lasso, the SCAD-penalized method can remove variables with small coefficients. Since the SCAD penalty is non-concave and encourages small coefficients to shrink toward zero, it applies heavy penalization to small coefficients, while large coefficients remain nearly unbiased. As the sample size grows, the SCAD-penalized method can consistently identify the relative accurate non-zero coefficients. Such a property can

 doi:10.20944/preprints202603.1439.v1

appropriately select the lag order in the sequential linear modeling and therefore raise the predicted accuracy in sequential linear models (1) and (2).

As shown in the simulation results in the next section, the SCAD penalty is a better method for solving the problem of penalized methods in parameter estimation and variable selection. Instead of applying the same penalty to the coefficients, the SCAD applies the non-concave penalty that effectively shrinks the smaller coefficients toward zero, while the larger coefficients are preserved for more accurate estimation. This balance makes the SCAD-penalized method a more effective way in sparsity and variable selection. To ensure sparsity and consistency, the SCAD-penalized method can accurately identify significant variables as discussed in Fan and Li (2001). Again, the design matrix also satisfies the restricted eigenvalue condition to ensure that the predictors are spread out enough to allow for correct variable selection even if the number of predictors is much larger than the sample size. The SCAD-penalized method also achieves the oracle property under regularity conditions by controlling the number of shrinkage coefficients as shown in Fan and Li (2006).

The regularization parameter λ controls the shrinkage effects on the coefficients for the predictors $\boldsymbol{Z_t} = (\boldsymbol{X_t}, \boldsymbol{Y_{t-1}})$ which combines the independent variables $\boldsymbol{X_t}$ and the lagged dependent variable $\boldsymbol{Y_{t-1}}$. The coefficient $\boldsymbol{\alpha_t} = (\boldsymbol{\beta}^T, \gamma_{t-1}^T)^T$ combines the parameters in front of independent variables $\boldsymbol{X_t}$ and the lagged coefficients in front of $\boldsymbol{Y_{t-1}}$. The response variable is $\boldsymbol{y_t}$ and the lagged dependent response variables are denoted as $\boldsymbol{y_1, y_2, ..., y_{t-1}}$.

The sequential linear modeling often involves dependencies between observations at different time points, and the penalized method with the SCAD penalty helps mitigate multicollinearity issues by encouraging a parsimonious representation of the underlying patterns. This combination is particularly valuable when the model for each time point builds upon information from previous points. Sequential modeling accounts for dependencies between observations across different time points, utilizing lagged response variables to predict future outcomes. The SCAD-penalized method effectively selects significant variables while minimizing the risk of over-shrinking large coefficients. Xie and Huang (2009) employed the SCAD penalty to promote sparsity in the partially linear model while using polynomial splines to estimate the nonparametric function. Their work demonstrates that the SCAD penalty not only achieves consistent variable selection but also accurately identifies the correct model structure, enabling efficient parameter estimation under the true model.

As discussed previously, while the fundamental principles of the penalized methods remain consistent across simple linear regression plus time-series models, the differences lie in the handling of temporal dependencies, multi-collinearity, and the unique characteristics of time-ordered data. By integrating the penalized method with the SCAD penalty into the combination of linear regression and time series model, we aim to achieve accurate predictions, reduce overfitting, and enhance the interpretability of the model. This approach strikes a balance between variable selection and coefficient estimation, providing a robust framework for modeling time-dependent data.

3.2. Selection of Tuning Parameter

The process of selecting the regularization tuning parameter in the penalized methods is crucial. In the sequential linear model, where temporal dependencies are present, careful consideration is needed to balance the penalty on the coefficients and the preservation of important features. To adjust for such balance, the tuning parameter in the penalized methods should be appropriately selected.

As discussed in last section, the SCAD-penalized method estimates the parameters in the sequential linear model by shrinking the non-significant parameters to zero and reducing the bias of significant parameter estimates. Therefore, the penalized method with SCAD penalty can consistently select the significant parameters in the model. Fan and Li (2001) proposed that the SCAD method is applicable in both parametric and high-dimensional nonparametric models by using wavelets and splines. The convergence rate of the proper choice of the penalized likelihood estimators depends on the regularization parameter. So, an appropriately selected tuning parameter is needed for effectively selecting significant variables in the model and efficiently computing in estimation, which will ensure a productive and rewarding process of variable selection and accurate parameter estimation. Tuning

 doi:10.20944/preprints202603.1439.v1

the regularization parameter in the penalized methods is a critical step, particularly when applying it to sequential linear models with temporal dependencies.

To facilitate the selection of the tuning parameter, cross-validation is commonly used, for which traditional random data splitting may not be appropriate due to the sequential nature of the data. The sequential linear model also involves additional complexities, such as creating lagged variables with time effect coefficients, which requires a careful balance between penalizing coefficients and preserving important features. In this context, the SCAD-penalized Lasso proves effective in addressing these challenges and enhancing feature selection, which is also the reason why the performance of the SCAD-penalized outperforms the other adopted methods as demonstrated in the simulations of next section.

Again, the penalized methods with different penalty terms include the process of cross-validation to choose the best tuning parameter, which controls the weight of the shrinkage to parameter estimates. In the sequential linear model with the Lasso method, we have the L_1 penalty function as

$$p_\lambda(|\alpha_{jt}|) = \lambda|\alpha_{jt}|,$$

which leads to the thresholding rule of $\alpha_{jt} = \text{sgn}(\alpha_{jt})(|\alpha_{jt} - \lambda|)_+$, where $\text{sgn}(\cdot)$ represents the sign function and $(\cdot)_+ = \max\{0, \cdot\}$.

With the adaptive Lasso method, we have the penalty function as

$$p_\lambda(|\alpha_{jt}|) = \lambda \sum_j w_j|\alpha_{jt}|,$$

where w_j is the weight assigned to the coefficient α_{jt}. The adaptive Lasso's penalty on each coefficient is adjusted based on the corresponding Lasso estimate. This variable-specific shrinkage allows adaptive Lasso to focus more on variables with larger estimated coefficients.

The purpose of using adaptive weights is to down weight the importance of certain variables and upweight others, allowing for more effective variable selection. Variables with larger Lasso estimates receive smaller weights, while variables with smaller Lasso estimates receive larger weights. The optimization problem for the adaptive Lasso involves minimizing the sum of the squared differences between the observed responses and the predicted values, subject to the adaptive Lasso penalty.

Before fitting the model via various methods, we split the data into training and testing sets in a ratio of 8:2. The training set, consisting of 80% of the data, is used to estimate the parameters, identify significant variables, and tune regularization parameters. The testing set, comprising the remaining 20% of the data, evaluates the model's predictive performance. In the sequential linear model, we use the training set of data to fit the model via the SCAD, Lasso, and adaptive Lasso methods and the test set to make model predictions. One step before fitting the model is to tune the appropriate parameter to make the mean squared errors as small as possible.

Cross-validation is a powerful technique to ensure the robustness of the generalization performance and avoid overfitting in the sequential linear modeling. By splitting the data into the training and validation sets in K-fold cross-validation, the cross-validation helps the SCAD-penalized method to select the significant variables and prevent over-shrinkage, and it is also used in combination with the Lasso or the other penalized methods adopted in this study to assess its performance and select an optimal regularization tuning parameter.

The prediction error serves as a measure in cross-validation. With the training set fixed, the cross-validation is to estimate the expected prediction error. We use the validation data set to show the performance of the prediction model. Especially for the high-dimensional data set, K-fold cross-validation is to choose the available part to fit the model and use the left part of the data to test the performance. The rule of selecting the best tuning parameter λ is to minimize the estimated prediction error.

According to Equation (7), here are the details of the SCAD-penalized method selecting variables with the combined coefficients α_{jt} in the sequential linear models.

 doi:10.20944/preprints202603.1439.v1

1. When the $|\alpha_{jt}|$ is small, the SCAD penalty works as the Lasso method that will shrink parameters to exact zero.

2. When the $|\alpha_{jt}|$ is large, the SCAD penalty is the constant value that avoids the bias for large parameters.

3. When the $|\alpha_{jt}|$ is moderate, the SCAD penalty attaches different weights to parameters.

4. Simulations

4.1. Simulation Settings

To check the performance of the SCAD-penalized method in the sequential linear models in equation (2), we conduct a simulation study with the same sample size and different numbers of predictors, along with a set of continuous outcomes based on fifty-time points $t = 1, ..., 50$. We therefore have different simulated settings. Based on each of the simulation settings, we compare the SCAD-penalized method with the OLS, the Lasso, and the adaptive Lasso in traditional linear and sequential linear models. The optimal tuning parameters λ is selected in the Lasso, SCAD-penalized, and adaptive Lasso methods respectively.

Consider a dataset with repeated measures of outcomes $y_{i,t}$ for a sample $i = 1, ..., n$ through time point $t = 1, ..., T$. A model is trained to sequentially predict the outcome at each time point as the baseline data matrix $\boldsymbol{x_i} = (x_{i,1}, ..., x_{i,p})^T$ and a set of parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ for modeling the outcome at time t, where p is the number of predictors. The number of predictors is set as low ($p = 10$), medium ($p = 100$), and high ($p = 1000$) dimensions in the simulation study.

At the first time point $t = 1$, the model is $y_{i,1} = \boldsymbol{x_{i,1}\beta} + \epsilon_{i,1}$, where $\epsilon_{i,1} \sim N(0, \sigma^2)$.

Then for any time point t, considering the sequential linear model with time series as model (2), we have

$$\boldsymbol{y_t} = \boldsymbol{X_t\beta} + \boldsymbol{Y_{t-1}\gamma_{t-1}} + \boldsymbol{\epsilon_t}, \tag{10}$$

where $\boldsymbol{y_t} = (y_{1,t}, y_{2,t}, ..., y_{n,t})^T$, which is the response/outcome vector at time t, and $\boldsymbol{\epsilon_t}$ is the error term vector containing all error terms at t time point, i.e., $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \epsilon_{2,t}, \cdots, \epsilon_{n,t})^T$. Again, we can repeat model (10) to generate data until the end timeline $t = T$. When $t = 1$, the initial time point, the model does not involve the term $\boldsymbol{Y_{t-1}\gamma_{t-1}}$.

Consider the number of observations is $n = 100$. The three levels of predictor numbers are $p = 10$, $p = 100$, and $p = 1000$ respectively. We independently generate the simulated data matrix $\boldsymbol{X_t}$ from the multivariate normal distribution with mean $\boldsymbol{0_{p*1}}$ and standard deviation of $\boldsymbol{1_{p*p}}$ at each time point.

The values of the true associated regression coefficients β_j are set as $\beta_1 = \beta_3 = 2, \beta_5 = \beta_7 = \beta_9 = -3$ and the rest of β_j values are set to be 0. We set up the lag is 1, and the time effect coefficient of true γ is 0.6. Iteratively, we have $T = 50$. Splitting the data is by a ratio of 4:1 for the training and test sets.

The generated data are fitted by linear model (LM) and sequential linear model (SLM) respectively and all through the OLS, Lasso, adaptive Lasso, SCAD-penalized methods, and adaptive Lasso methods.

4.2. Simulation Results for Estimation and Variable Selection

The simulation results of the estimation of the coefficients and the mean prediction testing errors are the average values over 1000 replicates.

For the estimation, the average predicted error is used as a measure, and the tables and figures are demonstrated to show the estimation results. Table 1 features the average prediction errors for each method.

According to Hastie et al. (2020), relative risk (RR) is the accuracy metric for methods comparisons. We expect the score of relative risk (RR) to be as small as possible, which denotes a more accurate

 doi:10.20944/preprints202603.1439.v1

coefficient estimation result. The formula of the relative risk (RR) is shown below for parameter vector $\boldsymbol{\alpha}_t$. The RR($\hat{\boldsymbol{\alpha}}_t$) can be estimated by

$$\frac{(\hat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t)^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_t)}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_t},$$

where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\boldsymbol{Z}_t$, and again $\boldsymbol{Z}_t$ represents the matrix containing all predictors $\boldsymbol{X}_t$ and previous outcomes $\boldsymbol{Y}_{t-1}$ denoted by $\boldsymbol{Z}_t = (\boldsymbol{X}_t, \boldsymbol{Y}_{t-1})$. We can calculate RR for the last time point at $t = T$. The perfect score is 0 which means the estimator $\hat{\boldsymbol{\alpha}}_t$ is equal to $\boldsymbol{\alpha}_t$ and the score of RR of 1 means that the estimator $\hat{\boldsymbol{\alpha}}_t$ is zero. Table 2 features the RR values for each method.

Table 1. Mean absolute prediction errors (MAPE).

	MAPE					
Dimension	Low ($p = 10$)		Medium ($p = 100$)		High ($p = 1000$)	
	LM	SLM	LM	SLM	LM	SLM
OLS	0.0128	0.0349	NA	NA	NA	NA
Lasso	0.0129	0.0224	0.0132	0.0125	0.0995	0.0362
SCAD	0.0128	0.0218	0.0127	0.0104	0.0994	0.0009
adaptive Lasso	0.0247	0.0204	0.0127	0.0210	0.0999	0.1009

Table 2. Relative risks (RR).

	RR					
Dimension	Low ($p = 10$)		Medium ($p = 100$)		High ($p = 1000$)	
	LM	SLM	LM	SLM	LM	SLM
OLS	0.3549	0.0092	NA	NA	NA	NA
Lasso	0.3561	0.0066	0.4654	0.0155	0.3942	0.0201
SCAD	0.3561	0.0031	0.4734	0.0037	0.3851	0.0019
adaptive Lasso	0.3561	0.0038	0.4733	0.0029	0.3873	0.1359

Table 1 shows that with time series data, for all models and different dimensions, the average absolute prediction error of the SCAD-penalized method holds the smallest value among all methods.

For the results fitted by the sequential linear model, Table 1 illustrates that the SCAD-penalized method has a relatively small mean absolute prediction error, indicating a high prediction accuracy among those methods. For instance, in the medium dimension of $p = 100$ with time series, the estimated $\hat{\gamma}$ is 0.56. The average prediction errors for the Lasso, SCAD, and the adaptive Lasso are 0.0125, 0.0104, and 0.0210 respectively. Thus, in the medium dimension with time series, the SCAD-penalized method fitted by a sequential linear model is the best choice among the various methods. Similar to the case for the medium dimension, with $p = 1000$ in the high-dimension setting, the average absolute prediction error is 0.0994 which is the smallest of the SCAD-penalized method.

Table 2 also shows that fitted by the sequential linear model for the generated data with time series, the SCAD-penalized method generally reaches a relatively low mean RR values, confirming that the SCAD-penalized method performs best in estimating parameters among all the adopted methods here.

Additionally, we can observe that for the generated data with time series, the estimation results shown by the RR as in Table 2 and prediction results shown by the MAPE as in Table 2 imply sequential linear modeling functions better than a simple linear model. This consequence confirms advantages of the sequential linear modeling in models (2) and (3). When the fixed effects and the time series data are involved, the sequential linear modeling is a good choice to conduct modeling as the ADL for dynamic data as discussed in the Introduction section.

In the generated data, we also conduct simulations for variable selection. In our simulations, we set five non-zero true β_j values as the significant variables, which are the first, third, fifth, seventh, and ninth variables, while the rest are insignificant. The individual asymptotic confidence interval is calculated as the difference between the 97.5% and 2.5% quantiles of the estimates for each parameter.

 doi:10.20944/preprints202603.1439.v1

To examine 95% estimation accuracy of significant variables using various methods including the OLS, Lasso, SCAD-penalized, and adaptive Lasso, the asymptotic confidence intervals (ACI) for all significant variables and for the number of significant variables are summarized in Tables 3, 4, and 5 corresponding to three different dimensions. Again, the generated data and all the methods are fitted by a traditional linear model (LM) and the sequential linear model (SLM).

Table 3 features the asymptotic 95% confidence intervals for the number of significant variables and parameters in various methods in the settings with time series. Based on the simulated data, with the number of parameters of 10, we expect a 95% confidence interval of significant variables infinitely closing to the true setting, which is 5. To examine the 95% asymptotic confidence intervals for the five significant variables in the true settings using various methods including the OLS, Lasso, SCAD, and adaptive Lasso, we calculate the 95% confidence intervals of the estimated coefficients based on 1000 repetitions. In low-dimension settings in the simple linear model (LM), the best performance in variable selection is adaptive Lasso because the confidence intervals have the smallest ranges and are close to the true value of 5. Besides, the asymptotic 95% confidence intervals for parameters have the least variations and are the closest to the true parameters.

Further, the confidence intervals for the time lag effects (γ) are also included and the values show that the SCAD-penalized method possesses the narrowest 95% confidence intervals for the parameters and for the number of significant variables. With the lagged time effect in the sequential linear model (SLM), the asymptotic confidence interval results for each significant variable including the lagged time coefficient γ are summarized in Table 3. In low dimension, the true setting of significant variables is 6 in the sequential linear model (SLM), and the SCAD-penalized method has the smallest range (6 to 9.53) for number of significant variables. So, regarding variable selection, the SCAD-penalized method achieves the best selection result. Based on the simulated data, there is 95% asymptotic confidence to predict that the lagged time effect coefficient is between 0.59 and 0.66 which is the smallest range of confidence interval among all the methods.

Table 3. Individual 95% asymptotic confidence intervals (ACI) in low dimension $p = 10$.

	ACI (LM)			
True value	OLS	Lasso	SCAD	adaptive Lasso
Num of sig variables = 5	(5, 10)	(5, 10)	(5, 8.58)	(5, 5)
$\beta_1 = 2$	(1.86, 2.14)	(1.82, 2.11)	(1.86, 2.14)	(1.84, 2.12)
$\beta_3 = 2$	(1.82, 2.12)	(1.78, 2.09)	(1.82, 2.12)	(1.80, 2.10)
$\beta_5 = -3$	(-3.12, -2.86)	(-3.08, -2.83)	(-3.12, -2.86)	(-3.11, -2.85)
$\beta_7 = -3$	(-3.14, -2.85)	(-3.11, -2.80)	(-3.13, -2.85)	(-3.13, -2.83)
$\beta_9 = -3$	(-3.14, -2.87)	(-3.11, -2.83)	(-3.14, -2.87)	(-3.13, -2.85)
	ACI (SLM)			
True value	OLS	Lasso	SCAD	adaptive Lasso
Num of sig variables = 6	(18, 41.1)	(6, 10.53)	(6, 9.53)	(6, 13)
$\beta_1 = 2$	(1.41, 2.63)	(1.52, 2.23)	(1.67, 2.36)	(1.59, 2.41)
$\beta_3 = 2$	(1.46, 2.67)	(1.50, 2.16)	(1.68, 2.29)	(1.66, 2.28)
$\beta_5 = -3$	(-3.58, -2.36)	(-3.10, -2.51)	(-3.26, -2.68)	(-3.23, -2.65)
$\beta_7 = -3$	(-3.59, -2.46)	(-3.18, -2.56)	(-3.32, -2.68)	(-3.25, -2.65)
$\beta_9 = -3$	(-3.61, -2.23)	(-3.17, -2.53)	(-3.26, -2.69)	(-3.21, -2.66)
$\gamma = 0.6$	(0.54, 0.73)	(0.54, 0.65)	(0.59, 0.66)	(0.56, 0.65)

In the medium dimension settings, we compare the three penalized methods, the Lasso, SCAD-penalized, and adaptive Lasso. Based on the medium dimension data, we have 95% confidence to predict that the number of significant variables falls between 6 and 11 in the SCAD-penalized method in the SLM, having the smallest range based on the 1000 repetition times among all three methods in both fitted models. The summarized results are in Table 4. We have 95% confidence to predict that the lagged coefficient is between 0.6 and 0.68 via the SCAD-penalized method based on the medium dimension of $p = 100$.

 doi:10.20944/preprints202603.1439.v1

Table 4. Individual 95% asymptotic confidence intervals (ACI) in medium dimension $p = 100$.

	ACI (LM)		
True value	Lasso	SCAD	adaptive Lasso
Num of sig variables = 5	(5, 18.03)	(5, 11.03)	(5, 24)
$\beta_1 = 2$	(1.73, 2.03)	(1.86, 2.15)	(1.84, 2.13)
$\beta_3 = 2$	(1.72, 2.04)	(1.85, 2.16)	(1.83, 2.15)
$\beta_5 = -3$	(-3.05, -2.73)	(-3.16, -2.85)	(-3.15, -2.83)
$\beta_7 = -3$	(-3.04, -2.71)	(-3.15, -2.83)	(-3.15, -2.82)
$\beta_9 = -3$	(-3.04, -2.72)	(-3.16, -2.84)	(-3.15, -2.83)
	ACI (SLM)		
True value	Lasso	SCAD	adaptive Lasso
Num of sig variables = 6	(6, 15)	(6, 11)	(8, 23)
$\beta_1 = 2$	(1.29, 2.00)	(1.71, 2.29)	(1.32, 2.21)
$\beta_3 = 2$	(1.28, 2.00)	(1.69, 2.34)	(1.35, 2.21)
$\beta_5 = -3$	(-3.02, -2.27)	(-3.32, -2.69)	(-3.24, -2.42)
$\beta_7 = -3$	(-2.99, -2.28)	(-3.30, -2.69)	(-3.22, -2.42)
$\beta_9 = -3$	(-3.03, -2.29)	(-3.34, -2.70)	(-3.23, -2.48)
$\gamma = 0.6$	(0.53, 0.63)	(0.60, 0.68)	(0.52, 0.63)

Table 5 shows that in the high-dimension setting, the SCAD-penalized method receives the best individual asymptotic 95% confidence interval for the parameters and the number of parameters. Based on the high-dimension data of $p = 1000$ predictors with 1000 repetitions via the SCAD method, we are 95% confident to predict that the number of significant variables is between 5 and 12.53 for the LM and between 6 and 17 in the SLM, which are the smallest ranges among the three methods. It means that the SCAD-penalized method is very likely to select a correct number of significant variables compared with the other methods here. The asymptotic confidence intervals for each significant variable are also close to their true coefficient values in the SCAD-penalized method. In addition, based on the generated data via the SCAD method in a high-dimension setting, we have 95% confidence to predict that the lagged time effect coefficient γ falls between 0.59 and 0.68 which is narrower than the method of the Lasso. The method of the adaptive Lasso has a 95% asymptotic confidence interval which is out of the range of the true value of γ.

Table 5. Individual 95% asymptotic confidence intervals in high dimension $p = 1000$.

	ACI (LM)		
True value	Lasso	SCAD	adaptive Lasso
Num of sig variables = 5	(5, 21.63)	(5, 12.53)	(20.48, 143.63)
$\beta_1 = 2$	(1.67, 1.98)	(1.87, 2.16)	(1.85, 2.13)
$\beta_3 = 2$	(1.64, 1.96)	(1.84, 2.15)	(1.83, 2.13)
$\beta_5 = -3$	(-2.98, -2.67)	(-3.16, -2.85)	(-3.15, -2.83)
$\beta_7 = -3$	(-2.97, -2.69)	(-3.14, -2.88)	(-3.13, -2.84)
$\beta_9 = -3$	(-2.96, -2.60)	(-3.16, -2.82)	(-3.13, -2.80)
	ACI (SLM)		
True value	Lasso	SCAD	adaptive Lasso
Num of sig variables = 6	(11, 29)	(6, 17)	(30, 67.63)
$\beta_1 = 2$	(1.07, 1.92)	(1.74, 2.26)	(0.00, 2.01)
$\beta_3 = 2$	(1.17, 1.96)	(1.75, 2.37)	(0.00, 2.03)
$\beta_5 = 2$	(-2.97, -2.01)	(-3.28, -2.71)	(-3.01, -0.34)
$\beta_7 = -3$	(-2.94, -2.11)	(-2.70, -2.69)	(-2.85, -0.94)
$\beta_9 = -3$	(-2.80, -2.15)	(-3.24, -2.67)	(-2.83, -0.79)
$\gamma = 0.6$	(0.52, 0.63)	(0.59, 0.68)	(0.00, 0.53)

5. Applications

In last section, the simulation results demonstrate that the SCAD-penalized method has advantages over other penalized methods employed in the study. To further compare the effectiveness of

 doi:10.20944/preprints202603.1439.v1

various penalized methods, we apply them to two longitudinal study data, one named Heart.valve from joineR package in R studio Lim et al. (2008) and the other one named Austin House Price from Pierce (2021), to conduct estimation and variable selection.

5.1. Heart Valve Data

5.1.1. Data Description

The data applied here is about the aortic valve replacement surgery from the joineR package in R studio Lim et al. (2008) and longitudinal measuring the different heart function outcomes after the surgery under an observational study based on the effects of different heart valve outcomes. For each patient's identification number, the effective subset of data is the part where the patients' observational time point equals or is greater than 5 times. Among this sub-dataset, the observed time point equal to 5 is employed as the training set of data, and the observed time point greater than 5 is set as the testing set of data. With the other various preoperative predictors, the column named log.lvmi is the response variable. The longitudinal column named lvmi measured the ventricular mass index at the follow-up visit.

The data frame consists of 988 observations for each of 25 predictors including sex (gender of patient, 0=Male and 1=Female), age (age of the patient at the day of surgery, years), status (censoring indicator, 1 = died and 0 = lost at follow up), prenyha (preoperative New York Heart Association (NYHA), classification), size (size of the valve, millimeters), con.cabg (concomitant coronary artery bypass graft), creat (preoperative serum creatinine, µmol/mL), dm (preoperative diabetes), acei (preoperative use of ace inhibitor), emergenc (operative urgency, 0 = elective, 1 = urgent, and 3 = emergency).

5.1.2. Results for Estimation and Variable Selection

To evaluate the log of left ventricular mass index values based on the patients' factors at the 5 times follow-up visits, the boxplot in Figure 1 visually explains the distribution of the log.lvmi values and the skewness by displaying the quartiles and the average values at each time point. Figure 1 shows that the distribution of the log of the left ventricular mass index (standardized) at each follow-up visit after the surgery is symmetric, and only when the time point is 5, the response variable has a right skewness. In addition, the medians are approximately the same, indicating that the log of left ventricular mass index values in the sequential time points is highly correlated and the current response variable highly depends on those for previous time points.

From the medical aspect, the value of the lvmi indicates the association between the left ventricular mass index (LVMI), as discussed by Shamszad et al. (2012), which is an important measure of the likelihood of heart morbidity and mortality among adults diagnosed with hypertension. The left ventricular mass index has been assessed, as in Lim et al. (2008), as strongly correlated with the post-surgery heart recovery of patients.

There are two types of models applied here. One is the simple linear model, and the other is the sequential linear model proposed by Equation (3). The methods used for comparison in each of the two models are the OLS, Lasso, SCAD-penalized, and adaptive Lasso. Table 6 features the mean absolute predicted error corresponding to each method. Among all values, the mean absolute prediction error of the sequential linear model via the adaptive Lasso and SCAD-penalized method is 0.0001 which is the lowest value among all the methods. We therefore can see that the SCAD-penalized method functions best in prediction.

 doi:10.20944/preprints202603.1439.v1

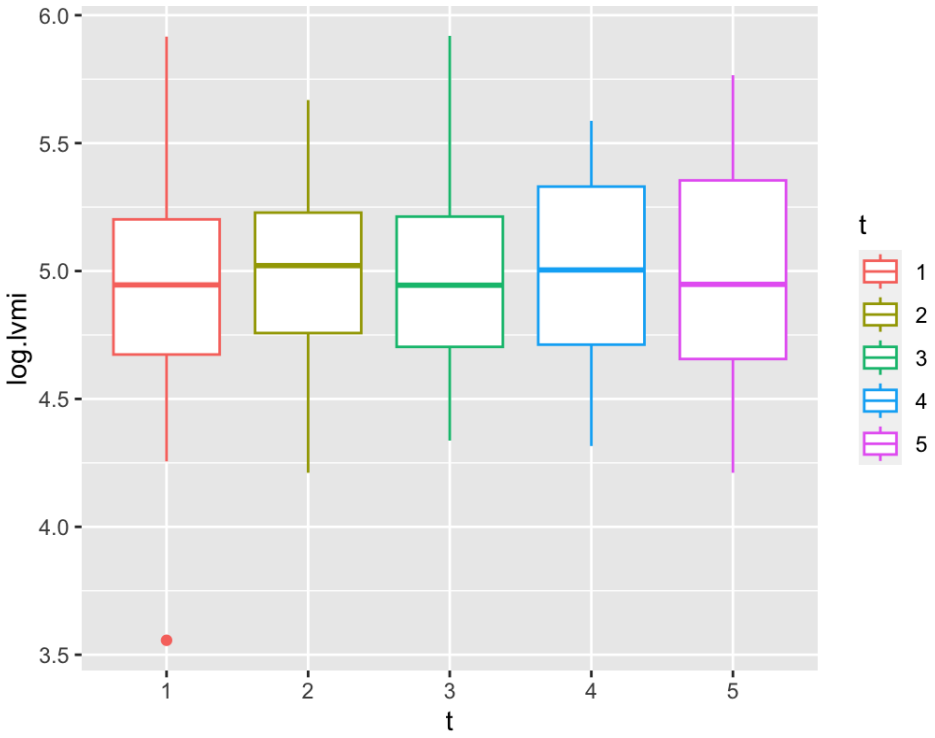

Figure 1. Boxplot of response variable (log.lvmi) for 5 time points.

Table 6. Mean absolute prediction errors (MAPE) in Heart.valve data.

	LM	SLM
OLS	0.169	0.037
Lasso	0.151	0.003
SCAD	0.129	0.001
adaptive Lasso	0.146	0.001

For the variable selection, the estimated coefficients for the selected variables via the SCAD-penalized method are summarized in Table 7 for the LM and in Table 8 for the SLM. As shown in Table 7, if the goal is only to check the predictors which have a significant effect on the response variable (the natural log transformation of left ventricular mass index), the selected significant predictors are clearly presented in the table, and the SCAD-penalized method in the simple linear model is a good choice. As shown in Table 8, the outputs via the SCAD-penalized method for the proposed sequential linear models show that the natural log transformation of left ventricular mass index as the response variable depends on the previous one time point, that is, the fourth lag variable. Since the lagged dependent variables are close and highly correlated, the SCAD-penalized method removes the unnecessary predictor variables in the model to validate the prediction accuracy, resulting the the fixed effects are not in the SLM and only one lagged response variable (y_4) is. Depending on the research goal, we could choose the SCAD-penalized method in different models.

To visually demonstrate the estimation and variable selection in this data set, Figure 2 features the coefficient estimates for significant variables in the linear model, and Figure 3 features the coefficients estimates for significant variables in the sequential linear model. The SCAD method in the LM detects ten significant variables with the estimated coefficients shown in Table 7, and the other estimated coefficients for insignificant variables are shrunken to be zero; while the SCAD method in the SLM detects only one significant variable because the other lagged dependent variables are highly correlated in the model. Through removing the unnecessary (insignificant) lagged dependent variables in the model, the SCAD method validates the prediction accuracy. The prediction performance is shown in Table 6, indicating that the SLM achieves a dramatically better behavior in prediction.

 doi:10.20944/preprints202603.1439.v1

Table 7. Coefficients estimates for significant variables via SCAD-penalized method in LM (Heart.valve data).

Selected variables	Estimated coefficients (LM)
sex	-0.299
age	0.007
status	-0.555
bsa	-0.541
lvh	-0.464
prenyha	-0.122
size	0.047
con.cabg	-0.181
creat	-0.005
acei	0.344
lv	-0.050
emergenc	0.338
hc	0.160

Table 8. Coefficients estimates for significant variables via SCAD-penalized method in SLM (Heart.valve data).

Selected variables	Coefficients estimates (SLM)
y_4	1.047

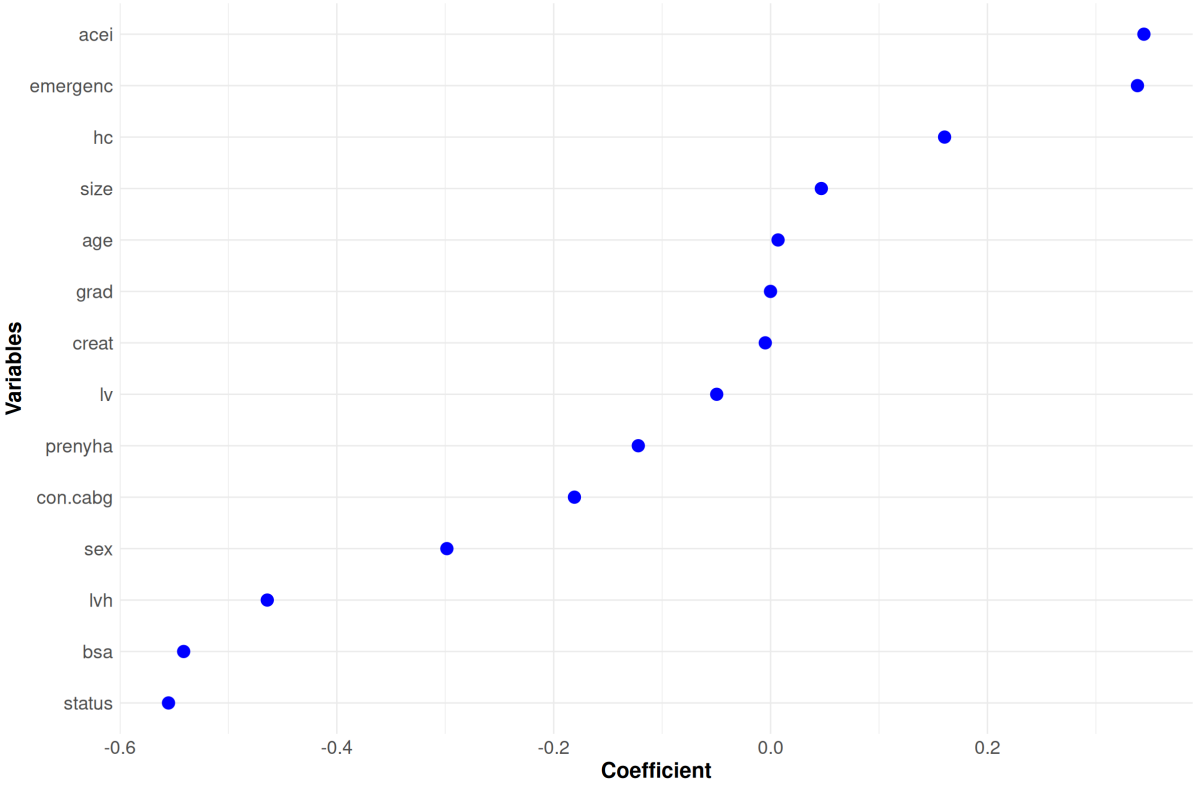

Figure 2. Coefficients estimates for significant variables in LM.

Last but not least, we check the regression assumptions for the SLM, and the output plots are shown in Figure 4. The residual plot shows that the error terms roughly follow a normal distribution with some skewness in both tails of Q-Q residuals plot. The error terms are spread equally along the ranges of the dots in the Scale-Location plot which shows an approximately constant variance and no clear pattern. The independent variables (significant predictors) are independent. Based on these plots, the assumptions of the SLM are complied.

 doi:10.20944/preprints202603.1439.v1

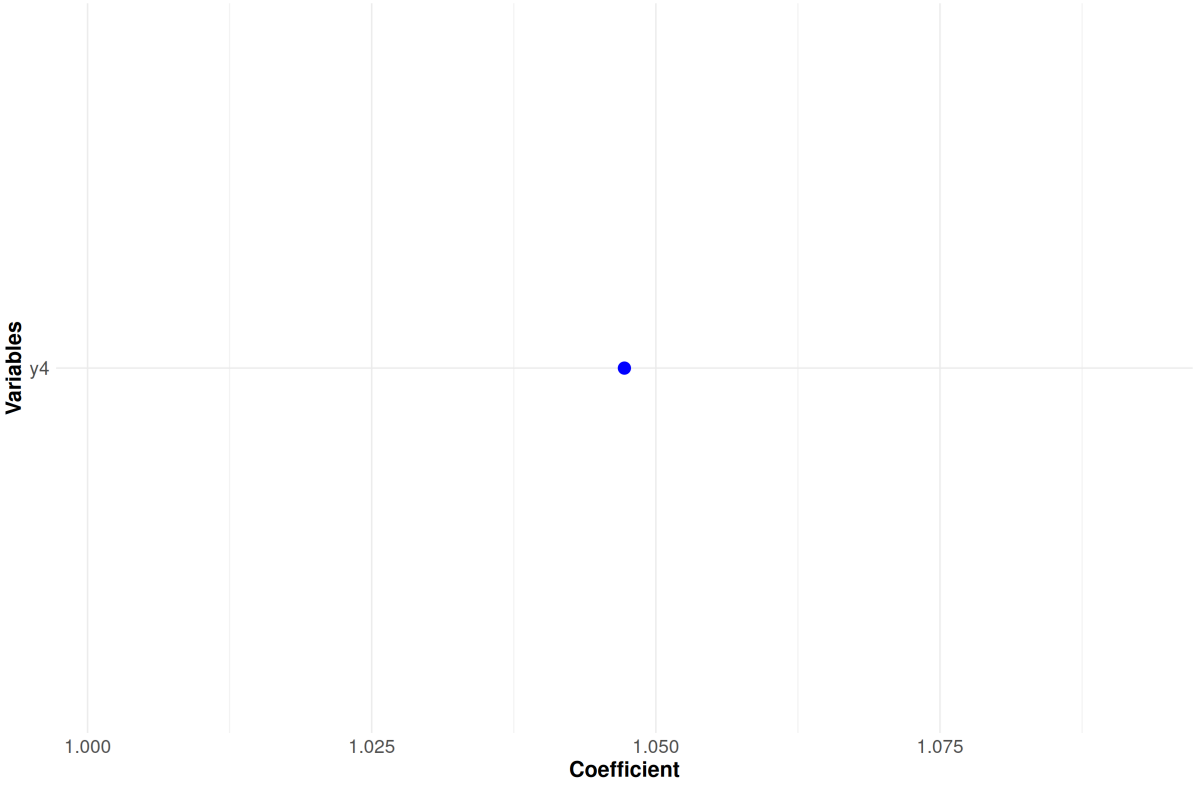

Figure 3. Coefficients estimates for significant variables in SLM.

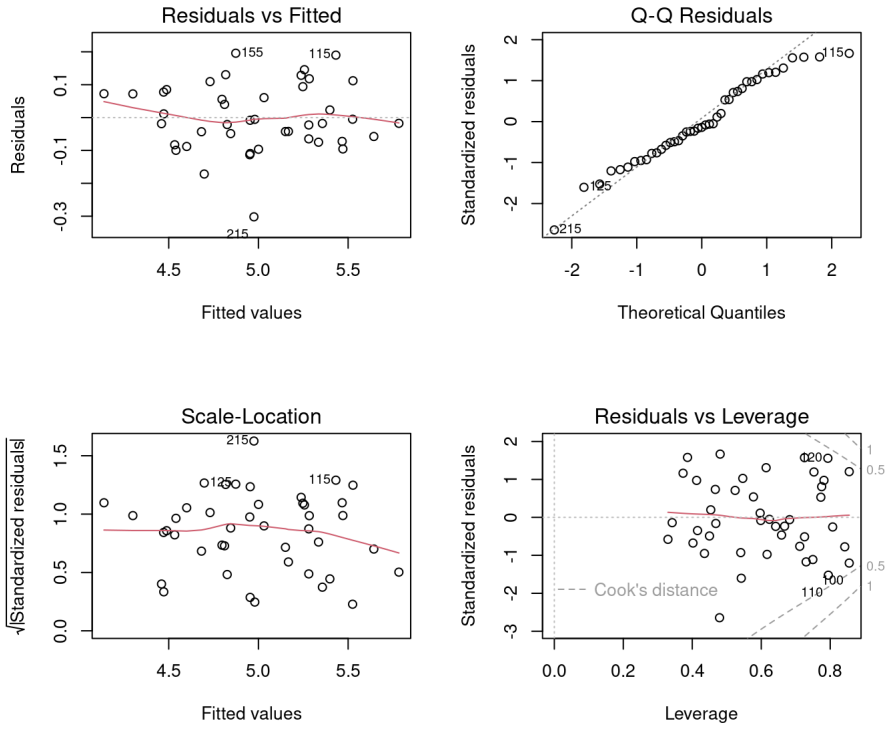

Figure 4. SLM diagnostics in Heart.valve data.

5.2. Austin House Price Data

5.2.1. Data Description

This data is collected from the website Kaggle (Pierce 2021) which was initially sourced from a project by author Eric Pierce three years ago. In 2021, the Austin Housing market stood out as one of the most dynamic markets, and these listings provided insight into the market's evolution over the

 doi:10.20944/preprints202603.1439.v1

last few years. The dataset includes various categories that affect the housing price. After cleaning the irrelevant predictors, there are 38 variables left in this application.

The data frame consists of 15171 observations for each of 38 predictors which include hasAssociation (indicates if there is a Homeowners Association associated with the listing), hasSpa (indicating if the home has a spa), numOfPhotos (the number of photos in the Zillow listing), avgSchoolRating (the average school rating of all school types, i.e., Middle, High, in the Zillow listing), and numOfBathrooms (the number of bathrooms in a property).

5.2.2. Results for Estimation and Variable Selection

The objective is to forecast the latest available price at the time of data collection for properties built each year. In terms of data cleaning, the initial step involves converting character variables into factors in RStudio. Then we eliminate any observations containing missing values. Next, we identify duplicate properties based on their year of construction, selecting those with a year built in 2000 or later as the applied dataset. We create a new data frame containing two variables: “yearbuilt" and “repeated numbers". This enables us to match the original dataset and select each segment of the data accurately. Then we split the data into the training and test sets. The training set consists of duplicated yearbuilt values that are less than or equal to 180, while the remaining data forms the test set.

After the manipulation of the data, we have 37 unique time points. The sequential linear model includes the 37th time point data and the previous 36 years’ property prices as predictor variables. Then we apply the methods of ordinary least squares (OLS), Lasso, SCAD-penalized, and adaptive Lasso in the sequential linear model to make a comparison among these methods.

To check the average properties’ latest price in the 37-year time points, we plot the line chart in Figure 5. The mean of the latest property’s prices fluctuates a lot during these years. It shows an increasing trend as time goes on with evident price changes over each time point. So, the house price should have time-series dependent variables and other fixed effects independent variables in the sequential linear modeling.

For the estimation and variable selection, the estimated coefficients for the selected variables via the SCAD-penalized method are summarized in Table 9 for the sequential linear model. The outputs via the SCAD-penalized method for the proposed sequential linear modeling show that the current house price as the response variable depends on the previous time point at $t = 25$ ($\boldsymbol{y_{25}}$), and some other independent variables/predictors such as hasAssociation, hasSpa, numOfPhotos, avgSchoolRating, and numOfBathrooms.

The map in Figure 6 shows the distribution of properties. Every point on the graph symbolizes a property and is distinguished by one of four colors: red, yellow, green, and blue. Red signifies the lowest level of recent property prices, yellow represents the second-lowest, green indicates a moderate price, and blue indicates the highest recent property prices. From this map, we can see which distribution has a relatively higher value of the properties.

As shown in Table 9, in addition to the selected variables such as the number of bathrooms, whether the property has a view, whether it includes a spa, and whether it is part of an association, and the other lagged dependent variables are all removed from the sequential linear models because of the high correlation among them. The selected variables significantly impact on the market price of properties in Austin, TX, in recent years. We note that the variables of the number of bathrooms and whether the properties have a spa positively affect property prices while being part of an association negatively impacts the estimated property prices in the Austin, TX market over the past three years.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 March 2026 doi:10.20944/preprints202603.1439.v1

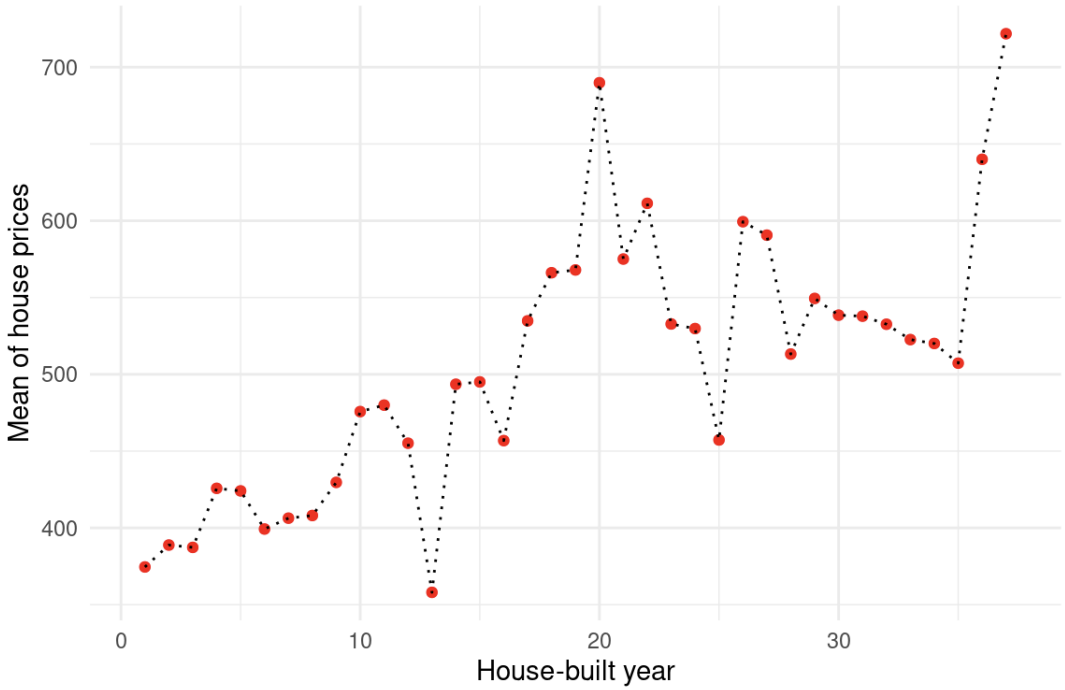

Figure 5. Mean of latest house prices built over time points.

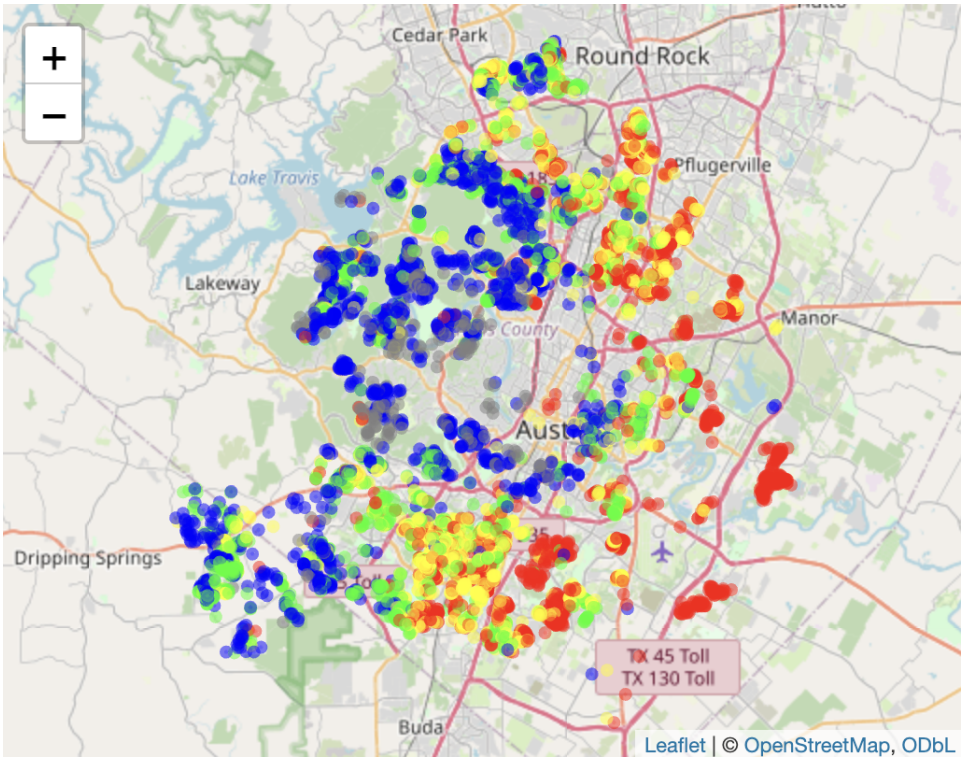

Figure 6. The Distribution of Properties' Leaflet Map

The mean absolute prediction errors for each method are summarized in Table 10, and again, the SCAD-penalized method owns the smallest mean of the prediction errors. Among all the penalized methods with different penalties, the SCAD-penalized method has the smallest mean of absolute prediction error 89.089 in the table.

 doi:10.20944/preprints202603.1439.v1

Table 9. Coefficients estimates of significant variables via SCAD-penalized method in SLM (Austin house price data).

	Estimated coefficients
hasAssociation	-354.962
hasSpa	78.606
numOfPhotos	0.168
avgSchoolRating	1.551
numOfBathrooms	323.626
y_{25}	0.013

Table 10. Comparison of Mean (Absolute) Prediction Errors in SLM (Austin, TX house price data).

	MAPE
Lasso	183.591
SCAD	89.089
Adaptive Lasso	116.206

For the significant variable selection, the coefficients estimated using the Lasso method in the sequential linear model is shown in Figure 7. In addition to the selected variables such as the number of bathrooms, whether the properties have a view, whether they include a spa, and whether they are part of an association, the model also includes several lagged dependent variables. These variables significantly impact on the market price of properties in Austin, TX, in recent years. Figure 7 apparently reflects the Lasso method shrinks the estimated coefficients of insignificant variables to zero, which vividly represents the sparsity property of the Lasso method.

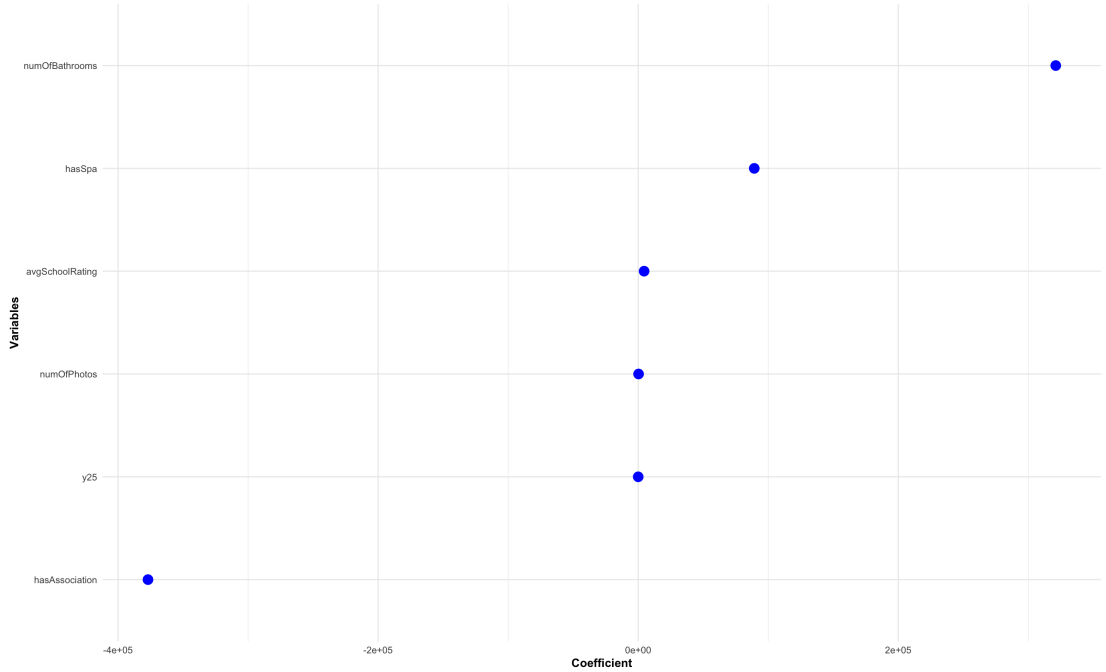

Figure 7. Coefficients estimates for significant variables via the Lasso method

From the coefficient plot via the adaptive Lasso method in Figure 8, the significant variables are the subset of the ones in the Lasso coefficient plot which does not select the lagged dependent variables in the model. Based on the Lasso coefficients, the adaptive Lasso applies different penalty weights to each coefficient, and the lagged dependent variable are all removed because of the high correlation among them. As the result of the adaptive Lasso method, the number of selected significant variables is smaller than that by the Lasso method, confirming a fact that the Lasso method, compared with the adaptive Lasso method, may overfit the model. With fewer significant variables in the model selected

 doi:10.20944/preprints202603.1439.v1

by the adaptive Lasso, the prediction outperforms the Lasso method as shown in Table 10 where the MAPE for the adaptive Lasso is much less than that of Lasso.

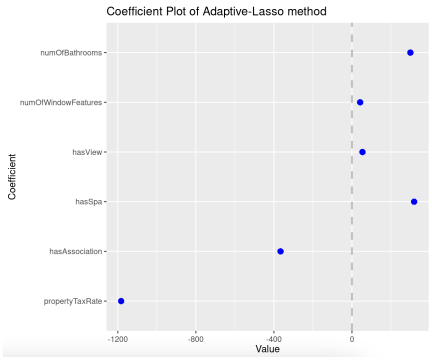

Figure 8. Coefficient estimates for significant variables via adaptive Lasso method

In Figure 9, the significant variables not only include the significant variables in the adaptive-lasso method but also include the lagged dependent properties price in the 25-time point. We notice that the number of bathrooms and whether the properties have a spa positively affect property prices while being part of an association negatively impacts the estimated property prices in the Austin, TX market over the past three years. We can therefore conclude the SCAD method can not only select the predictors in the SLM but also select the order of the lagged dependent variables. Because of this property, the SCAD method can effectively be applied in the sequential linear models.

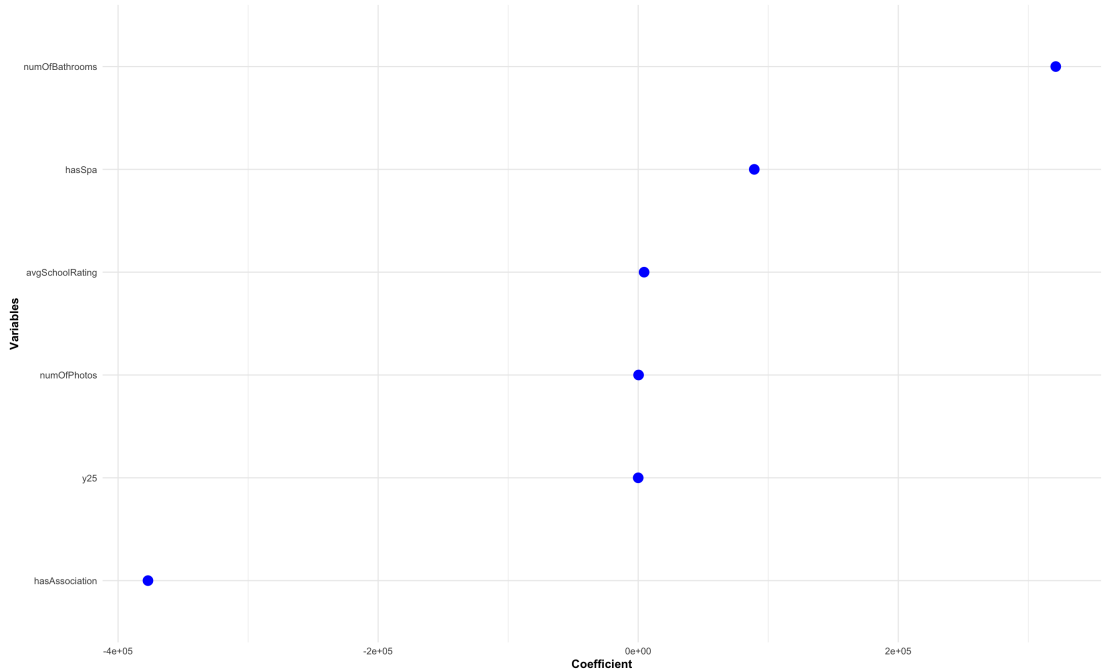

Figure 9. Coefficients estimates for significant variables via SCAD method

Among all the Lasso methods with different penalties, the SCAD method has the smallest mean of absolute prediction error 89.089 in Table 10.

Last but not least, we check the regression assumptions for the SLM, and the output plots are shown in Figure 10. The first residual plot shows a nonlinear pattern, which implies some nonlinear predictors may be needed in modeling the response variable. The Q-Q plot exhibits an approximate normal distribution and small skewness in both tails. In the Scale-Location plot, there exists a possibility that the variance is not homogeneous, so heteroscedasticity needs to be practically considered in the variance estimation. A few outliers can be observed from the leverage plot. Roughly the disparities

 doi:10.20944/preprints202603.1439.v1

from the assumptions for the SLM are not very far, the sequential linear models utilized here can reflect the trend for the prediction of the house prices, and the estimation and selected variables are very informative in the model interpretation.

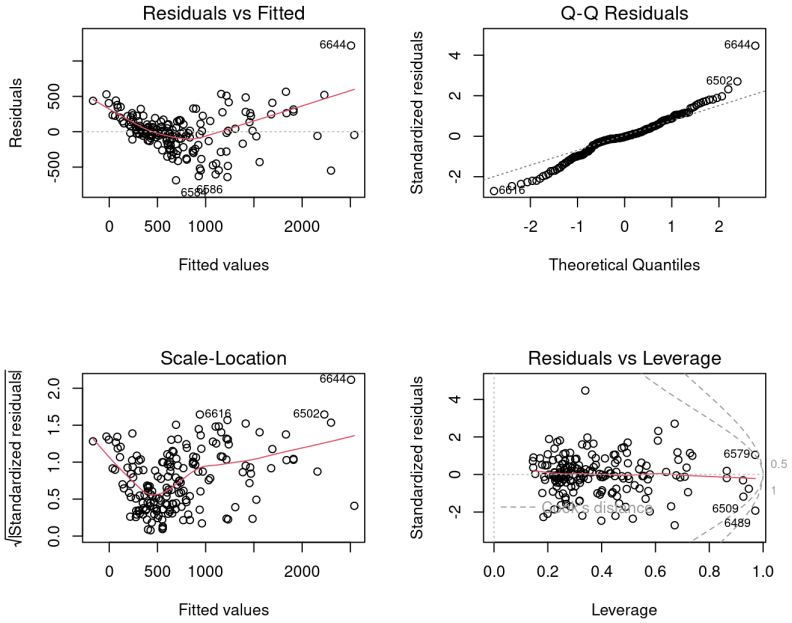

Figure 10. SLM diagnostics in Austin, TX house price data.

6. Concluding Remarks and Discussion

In sequential linear modeling, the measurements taken at each time point make great progress by longitudinally predicting the data results at intermediate time points. Such repeated sequential modeling can improve the model prediction accuracy because correlation measures are correctly described. With the focus on the sequential linear regression model, the lagged dependent variables are repeatedly included as the input variables in the next predicting models. We therefore explore sequential linear modeling and its performance in estimation and variable selection.

In the sequential linear models proposed in this study, we propose the SCAD-penalized method to improve the performance of estimation and variable selection. The asymptotic theory can validate the proposed method in that asymptotically the true model can be selected. The penalty term is expected to perform better in estimation and variable selection than the Lasso and adaptive Lasso methods. The Lasso method can do the variable selection by shrinking some coefficients to zero which is useful in medium or high-dimensional settings. The adaptive Lasso method applies the adaptive penalty weights to parameter estimates, improving over the Lasso method in estimation and variable selection. The SCAD-penalized method is designed to have different shrinkage to parameter estimates. When the parameter coefficients with the SCAD penalty are small, the penalty is similar to the Lasso and shrinks the corresponding coefficients to zero. When the coefficients are large, the SCAD penalty remains constant, reducing bias and likely indicating those variables are significant. When the parameter coefficients are moderate, the penalty will grow at a lower rate than the Lasso, reducing the shrinkage rate for parameter estimates.

We conduct simulations to explore the performance of the proposed SCAD-penalized method. With the purpose of comparison, we adopt other existing methods, including the Lasso and adaptive Lasso methods. Also, the ordinary least squares method is utilized in low-dimension settings. Across the datasets of low, medium, and high dimensions, we generate the base datasets of $p = 10, 100$, and 1000 with 1000 replications, and the error terms follow the standard normal distribution.

By incorporating lagged dependent variables into the model as predictors at each time point, we aim to compare the methods of ordinary least squares (OLS), Lasso, SCAD, and adaptive Lasso.

 doi:10.20944/preprints202603.1439.v1

Among these methods, the OLS is only applied in low-dimensional data settings and does not function well when the number of predictors p significantly exceeds the number of observations n.

The simulation results show that the SCAD-penalized method makes progress on the Lasso by reducing the bias in parameter estimation and by lowering the predicted errors in variable selection. The SCAD penalty improves parameter estimation accuracy and variation, resulting in better accurate predictions in medium or high-dimension settings. The simulation results are summarized through the outputs of mean absolute prediction errors (MAPE), relative risks (RR), and the asymptotic confidence interval (ACI) across various methods in different scenarios. The 95% asymptotic confidence intervals show that the SCAD-penalized method possesses the closest prediction results to those in the true settings, such as the number of significant variables and true parameter values. Notably, in small datasets, the OLS method yields superior performance, while in medium to high dimensional datasets, the SCAD-penalized method has a superior performance for parameter estimation and variable selection.

To further explore the performance of the proposed SCAD-penalized method in estimation and variable selection, we apply the method to two real data sets with comparison to the OLS, Lasso, adaptive Lasso methods. The first data set is about the Heart Valve replacement surgery data from the joineR package (Lim et al. 2008) in RStudio. The SCAD method holds the lowest absolute mean prediction error in both linear and sequential models. The second dataset is about the latest housing prices in Austin, TX (Pierce 2021). Although the assumptions are not perfectly satisfied, removing the extreme outliers or taking the log of the response variable will make the linear relationship between the predictors and the log of the housing price in Austin, TX, even stronger. The SCAD-penalized method demonstrates more stable performance in variable selection, incorporating most of the significant variables identified by the adaptive Lasso.

In the future study, we may consider analyzing data sets with multicollinearity and make a comparison among the methods of the OLS, Lasso, SCAD, and adaptive Lasso. Many real datasets often entail high correlation among variables, leading us to apply various penalized methods to the sequential linear modeling which involves selecting the significant lagged dependent variables from multivariate time series data simultaneously. We can also explore more hypothesis testings to determine the minimum sample size we need for a reliable model prediction in the sequential linear modeling. We can also estimate the sample size to ensure that the model is robust to significant variables and forecasting a good prediction performance in applications.

Author Contributions: Y.Y. and J.S. contributed to the conceptualization and methodology of the study; Y.Y. conducted all simulation and application analysis and wrote the original draft; J.S. supervised, revised and finalized the manuscript; C.G. critically reviewed, commented on, and edited the manuscript. All authors reviewed the final manuscript.

Conflicts of Interest: The authors declare no conflict of interests.

References

Brockwell, P.J. and R.A. Davis. 2002. *Introduction to Time Series and Forecasting*. Switzerland: Springer.

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

Fan, J. and R. Li. 2006. Statistical challenges with high-dimensionality: Feature selection in knowledge discovery. *Proceedings of the International Congress of Mathematicians*.

Gu, C. and S. Ratnasingam. 2025. Change point detection in scad-penalized dynamic panel models. *Sequential Analysis 44(4)*, 377–403. https://doi.org/https://doi.org/10.1080/07474946.2025.2510372.

Hastie, T., R. Tibshirani, and R. Tibshirani. 2020. Analysis and recommendations based on extensive comparisons. *Institute of Mathematical Statistics 35*, 579 – 592.

Keele, L. and N. J. Kelly. 2006. Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Political Analysis, 14(2)*, 186–205.

Keele, L. and N. J. Kelly. 2017. Dynamic models for dynamic theories: The ins and outs of lagged dependent variables. *Cambridge University Press 14*, 2.

 doi:10.20944/preprints202603.1439.v1

Kim, Y., H. Choi, and H.-S. Oh. 2008. Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association 103*, 1665–1673.

Lim, E., A. Ali, P. Theodoro, I. Sousa, H. Ashrafian, T. Chamageorgakis, A. Duncan, M. Henein, P. Diggle, and J. Pepper. 2008. Longitudinal study of the profile and predictors of left ventricular mass regression after stentless aortic valve replacement. *The Annals of thoracic surgery 85*, 2026–2029.

Nicholson, W. B., I. Wilms, J. Bien, and D. S. Matteson. 2020. High dimensional forecasting via interpretable vector autoregression. *Journal of Machine Learning Research 21(166)*, 1–52.

Pesaran, M.H. and Y. Shin. 1999. An autoregressive distributed lag modelling approach to cointegration analysis. *In: Strom, S., Ed., Chapter 11 in Econometrics and Economic Theory in the 20th Century the Ragnar Frisch Centennial Symposium, Cambridge University Press, Cambridge, 371-413.*.

Pierce, E. 2021. Austin, tx house listings (kaggle).

Qiu, J., D. Li, and J. You. 2015. Scad-penalized regression for varying-coefficient models with autoregressive errors. *Journal of Multivariate Analysis 137*, 100–118.

Shamszad, P., T. C. Slesnick, E. O. Smith, M. D. Taylor, and D. I. Feig. 2012. Association between left ventricular mass index and cardiac function in pediatric dialysis patients. *Pediatr Nephrol 27(5)*, 835–41.

Shestopaloff, K., M. Canizares, and J. D. Power. 2022. A sequential modeling approach for predicting clinical outcomes with repeated measures. *Communications in Statistics - Theory and Methods 15(20)*, 7465–7478.

Stock, James H. and Mark W. Watson. 2003. Introduction to econometrics. *New York: Prentice Hall*.

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

Xie, H. and J. Huang. 2009. Scad-penalized regression in high-dimensional partially linear models. *Institute of Mathematical Statistics 37*, 673–696.

Zhao, P. and B. Yu. 2006. On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101(476)*, 1418 – 1429.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2)*, 301–320.