mathematics

MDPI

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; cgu@citadel.edu
* 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. For estimation, variable selection, and high-accuracy response prediction, we propose the penalized method based on the Smoothly Clipped Absolute Deviation Penalty (SCAD) for sequential linear models. We conduct simulations comparing the SCAD-penalized method with ordinary least squares (OLS), Lasso, and adaptive Lasso in 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 datasets to further illustrate the performance of the SCAD-penalized method in sequential linear modeling.

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

MSC: 62F10; 62J07; 62-08

1. Introduction

In the clinical field, biological studies, and econometrics, much 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 [1]) model is a powerful tool and has gained significant traction for its ability to capture both short-term and long-term dependencies and to facilitate the exploration of complex temporal dynamics [2]. 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 is also correlated with time-varying predictors. For such situations, the sequential modeling approach of [3] can be adopted. This sequential linear modeling approach can effectively model the dependent variable using its last state and predictors at sequential time points. Designed to handle sequential information, sequential linear models are well-suited for absorbing the time-series process and predicting the dependent variable through time-varying lagged outcomes and predictors. For example, to cope with

Check for updates

Academic Editor: Heng Lian

Received: 16 March 2026
Revised: 24 April 2026
Accepted: 27 April 2026
Published: 29 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

sequential predictive modeling, ref. [3] proposed 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 by incorporating past values of the lagged dependent variable to demonstrate the influence of dynamic variables on the prediction process. However, ref. [2] pointed out that correlation effects arise because the lagged dependent variable causes the coefficients of the independent variables to be biased toward lower values. To deal with this problem in sequential time-series data, ref. [2] introduced the lagged dependent variable in OLS regression, where the lagged dependent variable coefficient indicates the timing effect relative to the independent variable. Ref. [4] combined estimated residuals with the SCAD penalty and Yule–Walker equations to determine the order of autoregressive errors and estimate the autoregressive parameters. Ref. [5] proposed a new class of hierarchical lag structures (HLag), which embeds lag selection into a convex regularizer and improves forecasting performance.

Motivated by the ADL and the above related work, we employ a sequential linear modeling approach and utilize sequential linear models to describe the dependent variable by incorporating the lagged dependent variable and time-varying independent variables. 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 [6] 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 sequential linear modeling. The SCAD method, introduced by [7], is known for its ability to handle large-scale variable selection problems while providing less bias in coefficient estimation compared to the 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 sequential linear modeling estimated with the SCAD penalty and to compare it with the Lasso and adaptive Lasso [8,9], which have been widely applied for their effectiveness in variable selection and regularization. Our study examines the strengths and weaknesses of these penalized regression approaches by comparing their ability in estimation and variable selection. Specifically, we assess whether the SCAD-penalized method outperforms the 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.

Although SCAD has been widely used since [7], its application in time-series and dynamic settings remains narrower than in purely cross-sectional regression. Ref. [4] studied SCAD in varying-coefficient models with autoregressive errors. Ref. [10] established SCAD consistency in partially linear models. Ref. [11] proposed data-driven tuning-parameter selection for SCAD via BIC-type criteria. Ref. [12] analyzed SCAD-type penalization in autoregressive models to distinguish stationary from nonstationary regimes. Ref. [13] derived oracle inequalities for adaptive penalized methods in high-dimensional correlated panel data. Ref. [14] used SCAD to perform change-point detection in dynamic panel models. The vector-autoregressive literature has also seen related hierarchical-lag regularization [5].

Among these works, the dynamic panel models of [13,14] are the broader class most closely related to sequential linear models. In both, however, the lag dimension is fixed, and the model is not refit at each time point. For this reason, the sequential linear model of [3] is not a special case of either framework. Its defining features are that (i) the dimension of the lagged-outcome coefficient vector $\boldsymbol{\gamma_{t-1}}$ grows with the time index t, and (ii) the model is refit at each time point with an updated history of dependent-variable measurements. These two features place the formulation outside the scope of existing SCAD theory for dynamic models. Extending SCAD theory to this setting is the gap addressed in the present paper. In particular, at each time point t, we establish the oracle property for the SCAD-penalized estimator in the high-dimensional regime $p + t > n$, where the nominal lag dimension may exceed the cross-sectional sample size.

The incremental contributions of this paper can be summarized as follows:

1. We extend the SCAD-penalized framework to sequential linear models in which the lagged-outcome coefficient vector has dimension growing with the time index, a setting not covered by existing SCAD theory for dynamic panel or autoregressive models.
2. We establish the oracle property for the proposed estimator (Theorem 1), including selection consistency and asymptotic normality on the active set, and we derive a corollary on lag-order selection consistency (Corollary 1). Our argument adapts the high-dimensional SCAD theory of [14,15] to the sequential-lag setting of model (2).
3. We demonstrate, through simulations across low-, medium-, and high-dimensional settings and through two real longitudinal applications, that the proposed estimator outperforms OLS, Lasso, and adaptive Lasso on mean absolute prediction error, relative risk, and empirical 95% coverage and that it recovers the correct effective lag order.

The remainder of the paper is organized as follows. Section 2 introduces the sequential linear models and the regularity conditions. Section 3 develops the SCAD-penalized method, states the oracle property theorem with a proof, gives the lag-order selection corollary, describes the coordinate-descent algorithm and tuning-parameter selection, and discusses algorithmic and statistical convergence. Section 4 presents the simulation study. Section 5 contains two applications. Section 6 concludes and discusses.

2. Sequential Linear Models

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

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

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

Before proceeding with the sequential modeling, we can visualize the setting through an example. Measurements on patients can be collected at different time points, and these measurements include outcomes such as a measure detecting an illness together with predictors that are all time-varying, such as the many factors that affect this illness. We can then naturally assume that the outcome at the current time point is predicted by the predictors at that time point, together with the state of previous outcomes.

The sequential linear models are as follows.

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

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

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

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

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

where $\epsilon_{i,2} \sim N(0, \sigma^2)$, and the outcome vector is $\boldsymbol{y_{i,1}} = (y_{i,1})^T$.

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

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

where $\epsilon_{i,3} \sim N(0, \sigma^2)$, and $\boldsymbol{y_{i,2}} = (y_{i,1}, y_{i,2})^T$.

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

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

where $\epsilon_{i,t} \sim N(0, \sigma^2)$ and $\boldsymbol{y_{i,t-1}} = (y_{i,1}, \ldots, y_{i,t-1})^T$. This recursion is applied until the terminal time point $t = T$.

We remark that sequential linear models involve lagged dependent variables across time points, so the key assumptions need to be verified before fitting the model and predicting. A conceptual description is given below. Precise mathematical statements follow after model (2) is introduced.

1. The relationship between the dependent variable and the independent variables is linear.
2. The error terms at each time point are independent and follow a normal distribution with constant variance.
3. The independent variables do not exhibit perfect collinearity at each time point.
4. The independent variables are uncorrelated with the error term.
5. A fixed effect incorporated for each observation in the sequential linear model is permitted.

After model (2) is introduced, these key assumptions are phrased in mathematical form for use in the proofs. For the sequential modeling in Equation (1), we see that this model is a combination of linear predictors and autoregressive time-series regressions. In Section 3, we propose penalized methods to conduct estimation. Such methods enjoy a sparsity property, meaning that adopting a suitable penalty term can shrink some parameter estimates to zero. Relying on this sparsity property, variable selection can be implemented simultaneously with estimation. Estimating the parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma_{t-1}}$ in model (1) targets selection of the predictors/features within $\boldsymbol{\beta}$ and the number of lags for the autoregressive part within $\boldsymbol{\gamma_{t-1}}$. Regarding the number of lags, for example, if the estimation finds that only $\gamma_{t-2,t}$ and $\gamma_{t-1,t}$ are significant (i.e., do not shrink to zero), we conclude that the response variable at the current time point can be predicted by those for the last two points, so the lag order equals 2. As time proceeds, the response variables typically become increasingly uncorrelated with distant past values, and the corresponding lag coefficients shrink to zero. Theoretically, it is interpretable that the response variable at the current time point depends only on recent time points, so the effective lag order cannot be a large integer. Our purpose is therefore to propose a well-functioning penalized method that selects the predictors in $\boldsymbol{\beta}$ and the lag order in $\boldsymbol{\gamma_{t-1}}$.

Model (1) serves as a combined linear regression and autoregressive time-series specification, with the lag order determined by the proposed method. For a succinct form, we arrange all predictors into an $n \times (p+1)$ matrix

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

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

The response/outcome variables up to time t are represented by

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

which is an $n \times t$ matrix.

To represent model (1), the outcome at time point t, we use the outcome $\boldsymbol{Y_{t-1}}$ at time point $t-1$ as predictors, and combine all predictors into one matrix $\boldsymbol{Z_t} = \left[\boldsymbol{X_t}, \boldsymbol{Y_{t-1}}\right]$. For any time point t in a succinct format, the sequential linear models in Equation (1) can be rewritten as

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

where $\boldsymbol{X_t}$ is the predictor matrix at time t, $\boldsymbol{y_t} = (y_{1,t}, \dots, y_{n,t})^T$ is the response/outcome vector at time t, and $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \dots, \epsilon_{n,t})^T$ is the error-term vector at time t. We assume that $\epsilon_{1,t}, \dots, \epsilon_{n,t}$ are independent and identically distributed (i.i.d.) with variance σ^2. Again, model (2) is repeated until $t = T$. At the initial time point $t = 1$, the model does not involve the term $\boldsymbol{Y_{t-1}}\boldsymbol{\gamma_{t-1}}$, and at later time points $\boldsymbol{\gamma_{t-1}} = (\gamma_{1,t-1}, \dots, \gamma_{t-1,t-1})^T$.

Specifically, in the sequential linear model (2), we partition the design matrix as $\boldsymbol{Z_t} = (\boldsymbol{Z_t}^{(1)}, \boldsymbol{Z_t}^{(2)})$, where $\boldsymbol{Z_t}^{(1)}$ collects the columns corresponding to nonzero regression coefficients and $\boldsymbol{Z_t}^{(2)}$ collects the columns corresponding to zero regression coefficients. We also write $\boldsymbol{Z_t} = (\boldsymbol{Z}_t^1, \dots, \boldsymbol{Z}_t^{p+t})$, where $\boldsymbol{Z}_t^j$ is the column corresponding to the jth parameter at time t. We also impose model assumptions for weak (or no) autocorrelation and absence of heteroscedasticity. Let s be a positive integer labeling the lags. Without loss of generality, and similar to those in [14,16], the assumptions are as follows.

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

A2. $\boldsymbol{y_t}$ and $\boldsymbol{y_{t-s}}$ become asymptotically uncorrelated as $s \to \infty$, i.e.,

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

A3. The process $\{(y_{i,t}, x_{i1,t}, \dots, x_{ip,t})\}_{t \geq 1}$ is weakly (covariance) stationary. That is, the mean vector $E[(y_{i,t}, x_{i1,t}, \dots, x_{ip,t})^T]$ is constant in t. The variances are finite and constant in t, i.e., $\mathrm{Var}(y_{i,t}) < \infty$ and $\mathrm{Var}(x_{ij,t}) < \infty$ for all j. The autocovariance $\mathrm{Cov}(y_{i,t}, y_{i,t-s})$ depends only on the lag s and not on t.

A4. The active-set Gram matrix $n^{-1}(\boldsymbol{Z_t}^{(1)})^\top \boldsymbol{Z_t}^{(1)}$ is almost surely positive definite for all n sufficiently large and each fixed $t = 1, \dots, T$.

A5. $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \epsilon_{2,t}, \dots, \epsilon_{n,t})^T$ consists of n i.i.d. random variables whose distribution does not vary with t, with $E(\epsilon_{1,t}) = 0$, $\mathrm{Var}(\epsilon_{1,t}) = \sigma^2 < \infty$, and $E(\epsilon_{1,t}^4) < \infty$.

We note that A1 assumes that the error term is mean-independent of the predictors and lags. A2 encodes asymptotic uncorrelatedness, weakening linear dependence over time rather than requiring full independence. A2 does not imply full statistical independence unless additional distributional assumptions are imposed. A3 is weak stationarity, which is a mild and commonly used assumption for linear time-series models. Any strictly stationary process with finite second moments automatically satisfies A3, so the strictly stationary construction explored in Remark 1 below suffices for this compatibility. A4 is the identifiability condition for the active-set parameters. It guarantees that the active-set Gram matrix is invertible, so the regression on the active set is well-posed at each n. In the high-dimensional regime $p + t > n$, the full Gram matrix $n^{-1}\boldsymbol{Z_t}^\top \boldsymbol{Z_t}$ is necessarily rank-deficient, so identifiability cannot be imposed on the full parameter vector. When we develop the SCAD-penalized method in Section 3.2, we impose a further assumption that strengthens A4 to a uniform lower bound on the minimum eigenvalue (A7). A4 and A5 together guarantee that the model is well-posed and its active-set parameters are estimable.

For the sequential models in Equations (1) and (2), we aim to conduct variable selection and estimation in order to identify the most appropriate model to describe the time-varying data, to predict the outcomes, and to perform further inferences.

After we estimate $\boldsymbol{\beta}$ and γ_{t-1} by $\hat{\boldsymbol{\beta}}$ and $\hat{\gamma}_{t-1}$ respectively, the prediction of the dependent variable for the next time point is

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

where $t > 1$. Equation (3) is the prediction formula. Each time point's predicted output is a function of the predictors at the current point and of the lagged response at previous time points, which is taken as the true $\boldsymbol{Y_{t-1}}$ when available and as the predicted $\hat{\boldsymbol{Y}}_{t-1}$ otherwise. In stable time-series settings, the entries of the lag coefficient vector γ_{t-1} typically decrease in magnitude as the lag index increases [17], which further motivates the lag-order selection carried out by the penalized method in Section 3.

3. SCAD-Penalized Method for Estimation and Variable Selection

Having introduced sequential linear models (1) and (2) in the previous section, we now develop parameter estimation and variable selection for the sequential modeling. To facilitate prediction with the most appropriate model, we adopt penalized methods to improve estimation and variable selection and to mitigate overfitting. One typical penalized method is the Lasso [6], which introduces an ℓ_1 penalty term. The Lasso estimator possesses consistency and sparsity, which makes the model more interpretable and efficient in predicting and handling dependencies in sequential modeling. Beginning with the Lasso, this section then introduces the adaptive Lasso and proposes the SCAD-penalized method in sequential linear models for estimation and variable selection.

3.1. Presentation of SCAD-Penalized Method for Estimation and Variable Selection

In what follows, we present the methods used for estimation and variable selection in this study. The basic idea is to penalize Equation (2) to obtain the estimators. Recall $\boldsymbol{Z_t}$ presents the design matrix containing all predictors $\boldsymbol{X_t}$ and previous outcomes $\boldsymbol{Y_{t-1}}$, i.e., $\boldsymbol{Z_t} = (\boldsymbol{X_t}, \boldsymbol{Y_{t-1}})$. Let $z_{i,t}$ denote the ith row vector of Z_t, for $i = 1, \ldots, n$. For the regression in model (2), let $\boldsymbol{\alpha_t}$ be the parameter vector combining $\boldsymbol{\beta}$ and γ_{t-1}, denoted by $\boldsymbol{\alpha_t} = (\boldsymbol{\beta^T}, \gamma_{t-1}^T)^{\boldsymbol{T}}$. The dimension of $\boldsymbol{\beta}$ is $p + 1$ and the dimension of γ_{t-1} is $t - 1$, so the length of $\boldsymbol{\alpha_t}$ is $p + t$. We denote each parameter element of $\boldsymbol{\alpha_t}$ by α_{jt}, for $j = 1, \ldots, p + t$.

To estimate the parameter vector $\boldsymbol{\alpha_t}$, we start with the Lasso method. According to model (2), the Lasso estimator is obtained by minimizing the penalized squared-error loss:

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

where α_{jt} is the jth parameter element at time point t, $j = 1, \ldots, p + t$, and λ is the tuning parameter.

The Lasso method was initiated by [6]. As established in the literature, the Lasso is a regularization method that simultaneously performs parameter estimation and variable selection, and its estimators are consistent and sparse. Specifically, the ℓ_1 penalty $\lambda \sum_j |\alpha_{jt}|$ shrinks the coefficients of non-significant variables to zero, giving the Lasso the sparsity property on which variable selection relies.

In sequential linear modeling, observed or generated data are modeled at each time point, creating a dependency between successive iterations. A common challenge in this framework is multicollinearity, arising from significant correlations between the predictors. The Lasso regression method helps handle multicollinearity by automatically selecting relevant predictors and potentially excluding correlated ones. It also addresses temporal dependencies by considering the sequential ordering of observations and selecting lagged features that contribute to predicting future values.

When different weights are assigned to different parameters in $\boldsymbol{\alpha_t}$, the procedure is called the adaptive Lasso [8], which is essentially a weighted ℓ_1 penalization method. Under certain regularity conditions, the estimators of $\boldsymbol{\alpha_t}$ are

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

where w_j is the weight for parameter element α_{jt} (e.g., the reciprocal of a preliminary estimate), and λ is the tuning parameter.

The Lasso and adaptive Lasso in Equations (4) and (5) are employed for comparison against the SCAD-penalized method introduced below.

We now apply the penalized method with the SCAD penalty [7] to estimate the parameters in sequential linear models (1) and (2). The estimator is defined as the minimizer of the sum of the squared error and the SCAD penalty:

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

where $|\cdot|$ denotes the absolute value, and the penalty function $P_\lambda^s(\cdot)$, following [7,18], is explicitly written as

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

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

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

where $(\cdot)_+ = \max\{0, \cdot\}$. Expression (8) shows that the SCAD penalty is continuously differentiable on $\mathbb{R}$ except at the origin, where it is singular. The derivatives vanish outside the interval $[-a\lambda, a\lambda]$.

3.2. Oracle Property

We now explore the oracle property of the SCAD-penalized estimator in the sequential linear model. For clarity, we write $\hat{\boldsymbol{\alpha}}_t^s = (\hat{\alpha}_{1t}^s, \ldots, \hat{\alpha}_{(p+t)t}^s)^\top$. Let $\mathcal{A}_t = \{j : \alpha_{jt} \neq 0\}$ be the true active set at time t, with cardinality $|\mathcal{A}_t| = K$. Let $\mathcal{A}_t^* = \{j : \hat{\alpha}_{jt}^s \neq 0\}$ be the index set of the significant SCAD estimators, with cardinality $|\mathcal{A}_t^*| = K^*$. We organize the true parameter vector $\boldsymbol{\alpha}_t$ as $(\boldsymbol{\alpha}_t^{(1)}, \boldsymbol{\alpha}_t^{(2)})$, where $\boldsymbol{\alpha}_t^{(1)} \in \mathbb{R}^K$ collects the nonzero coefficients and $\boldsymbol{\alpha}_t^{(2)} = \mathbf{0}$ collects the zero coefficients. Similarly, we organize $\hat{\boldsymbol{\alpha}}_t^s$ as $(\hat{\boldsymbol{\alpha}}_t^{s(1)}, \hat{\boldsymbol{\alpha}}_t^{s(2)})$, and set $\boldsymbol{M_t^1} = (\boldsymbol{Z_t}^{(1)})^T \boldsymbol{Z_t}^{(1)}/n$.

Following [14], to establish the oracle property within a high-dimensional and sparse framework, we impose the following supplementary regularity conditions in [19], in conjunction with A1–A5.

A6. There exists a positive constant C_1 such that,

$$(\boldsymbol{Z}_t^j)^\top \boldsymbol{Z}_t^j / n \leq C_1$$

for all $j = 1, \ldots, p + t$ and for all n, at time t.

A7. There exists a positive constant C_2 such that

$$\boldsymbol{\xi}^\top \boldsymbol{M}_t^1 \boldsymbol{\xi} \geq C_2$$

for all $\boldsymbol{\xi} \in \mathbb{R}^K$ with $\|\boldsymbol{\xi}\|_2^2 = 1$.

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

$$n^{(1-b_2)/2} \min_{j \in \mathcal{A}_t} |\alpha_{jt}| \geq C_3.$$

A6 bounds the squared column norms of $\boldsymbol{Z_t}$ and is trivial when the predictors and lagged outcomes are normalized. A7 is the restricted eigenvalue condition and ensures that the active-set Gram matrix $\boldsymbol{M_t^1}$ is strictly positive definite uniformly in n, so that the predictors and the lags of $\boldsymbol{y_t}$ are sufficiently spread out for correct variable selection even when the number of predictors and lags is much larger than n. A8 controls the divergence rate of the effective dimension. It ensures that the number of significant predictors K grows at a controlled polynomial rate relative to n and prevents overfitting. A9 mandates a gap of order $n^{-(1-b_2)/2}$ between the smallest nonzero coefficient and zero, ensuring that the signal is not overwhelmed by the stochastic error. See [14,15] for further background.

Before turning to the oracle theorem, we verify that A1–A9 are jointly satisfiable.

Remark 1 (Compatibility of Assumptions A1–A9)**.** *Assumptions A1–A9 are mutually consistent. A concrete construction satisfying all of them is as follows. Let* $\{\boldsymbol{x_{i,t}}\}$ *be a doubly-indexed sequence of i.i.d.* $N_p(\mathbf{0}, I_p)$ *vectors, independent across i and t. For simplicity, fix a maximum lag* $L \geq 1$ *and time-invariant lag coefficients* $\gamma_1, \ldots, \gamma_L$*, with* $\gamma_L \neq 0$*, satisfying*

$$1 - \sum_{m=1}^{L} \gamma_m \zeta^m \neq 0 \quad \textit{for all } \zeta \in \mathbb{C} \textit{ with } |\zeta| \leq 1.$$

Define the nonzero entries of $\boldsymbol{\gamma_{t-1}}$ *by* $\gamma_{t-m,t-1} = \gamma_m$ *for* $m = 1, \ldots, L$*, and* $\gamma_{s,t-1} = 0$ *for* $s < t - L$*. Let* $\epsilon_{i,t} \overset{\text{iid}}{\sim} N(0, \sigma^2)$*, independent across i and t. Then the process* $\{y_{i,t}\}$ *admits a*

strictly stationary causal solution [17], which is therefore also weakly stationary. A1–A9 all hold for any sparsity index $K = \mathcal{O}(n^{b_1})$ with $0 < b_1 < 1$. Note that an analogous construction applies to any nonzero-lag support $S \subseteq \{1, \ldots, L\}$ with $L = \max S$ (for instance, sparse supports with non-consecutive lags), provided the corresponding polynomial $1 - \sum_{m \in S} \gamma_m \zeta^m$ satisfies the same root condition.

Although the parameter vector $\boldsymbol{\gamma_{t-1}}$ has nominal dimension $t-1$ that grows with t, A8 requires all but a shrinking fraction of its entries to be zero. The resulting finite effective lag order is precisely what makes the weak stationarity in A3 compatible with the growing nominal dimension of $\boldsymbol{\gamma_{t-1}}$. The simulation study in Section 4, which uses $\gamma = 0.6$ with effective lag order $L = 1$, is an explicit instance of this construction.

We now state the oracle property as a formal theorem.

Theorem 1 (Oracle property of the SCAD-penalized estimator)**.** *Assume that model (2) holds, assumptions A1–A9 are satisfied, and the tuning parameter λ satisfies $\lambda \to 0$, $\sqrt{n}\,\lambda \to \infty$, and $n^{(1-b_2)/2}\lambda \to 0$ as $n \to \infty$. Fix $a > 2$. Assume further that there exists a positive-definite $K \times K$ matrix $\boldsymbol{\Sigma}_t$ such that $\boldsymbol{M}_t^1 \xrightarrow{p} \boldsymbol{\Sigma}_t$ as $n \to \infty$. Let $\hat{\boldsymbol{\alpha}}_t^s$ denote a local minimizer of the SCAD criterion in (6). Then, as $n \to \infty$, the following conclusions hold.*

(i) Selection consistency.

$$\Pr(\mathcal{A}_t^* = \mathcal{A}_t) \to 1.$$

(ii) Asymptotic normality on the active set. *Let $q \geq 1$ be a fixed integer and let $\boldsymbol{A}_n$ be a sequence of $q \times K$ matrices satisfying $\boldsymbol{A}_n \boldsymbol{A}_n^\top \to \boldsymbol{H}$ as $n \to \infty$, where $\boldsymbol{H}$ is a $q \times q$ positive-definite matrix. Then*

$$\sqrt{n}\,\boldsymbol{A}_n\,\boldsymbol{\Sigma}_t^{1/2}\Big(\hat{\boldsymbol{\alpha}}_t^{s(1)} - \boldsymbol{\alpha}_t^{(1)}\Big) \xrightarrow{d} N\Big(\boldsymbol{0},\, \sigma^2 \boldsymbol{H}\Big),$$

where $\boldsymbol{\Sigma}_t^{1/2}$ denotes the symmetric positive-definite square root of $\boldsymbol{\Sigma}_t$.

Proof. The proof proceeds by reduction to Theorem 2 of [20], which establishes the oracle property for SCAD-penalized estimators under a diverging number of parameters. We verify that the hypotheses of that theorem are satisfied by the sequential linear model (2) under A1–A9 and the tuning-parameter conditions of Theorem 1.

The asymptotic regime is fixed t with $n \to \infty$. At each fixed t, model (2) is an n-sample regression across i.i.d. subjects (by A5), with temporal dependence entering only through the regressors $\boldsymbol{Y}_{t-1}$ that are part of the design matrix $\boldsymbol{Z}_t$. This matches the i.i.d. across-subjects setting assumed in [20].

First, model (2) is the Gaussian linear model $\boldsymbol{y}_t = \boldsymbol{Z}_t \boldsymbol{\alpha}_t + \boldsymbol{\epsilon}_t$ with parameter vector $\boldsymbol{\alpha}_t \in \mathbb{R}^{p+t}$. This is a special case of the general likelihood setup ([20], Section 1.2) with log-density $\log f(z, y; \boldsymbol{\alpha}) = -\frac{1}{2\sigma^2}(y - \boldsymbol{z}^\top \boldsymbol{\alpha})^2 + \text{const}$. Under this specialization, the Fisher information reduces to $I_n(\boldsymbol{\alpha}_t) = \sigma^{-2} E[\boldsymbol{z}_{i,t} \boldsymbol{z}_{i,t}^\top]$, and the likelihood-regularity conditions (E), (F), and (G) of [20] translate into conditions on the design and error distribution of model (2).

The correspondence between A1–A9 of the present paper and conditions (A)–(H) of [20] is as follows.

- Penalty conditions (A)–(D) are satisfied by the SCAD penalty with $a > 2$ under the tuning-parameter rates in Theorem 1. In particular, $a_n = 0$ and $b_n = 0$ eventually ([20], Section 3.1.1).
- Condition (E) (i.i.d. observations with common density, identifiable model, score with mean zero at the truth) follows from A5 (i.i.d. subjects with Gaussian errors, giving common support), A4 (active-set Gram matrix positive definite, giving identifiability

on the active set), and A1 (conditional mean-zero errors, giving score mean zero under correct specification).

- Condition (F) (Fisher information positive definite with bounded eigenvalues and bounded fourth moments of score components) follows from A7 (active-set Gram matrix bounded below), A6 (column norms bounded above), and A5 (finite fourth moment of $\epsilon_{i,t}$).
- Condition (G) (third-derivative regularity of the log-density) is trivial for the Gaussian linear model because the log-density is quadratic.
- Condition (H) (minimum signal condition $\min_{j\in\mathcal{A}_t}|\alpha_{jt}|/\lambda \to \infty$) follows from A9 combined with the tuning-parameter rate $n^{(1-b_2)/2}\lambda \to 0$, which together give $\min_{j\in\mathcal{A}_t}|\alpha_{jt}|/\lambda \geq C_3/(n^{(1-b_2)/2}\lambda) \to \infty$.
- The sparsity rate A8 supplies the growth condition on the number of parameters required by [20], Theorem 2, provided b_1 is sufficiently small relative to the rate constraints imposed there.

The remaining conditions, A2 (asymptotic uncorrelatedness) and A3 (weak stationarity) are descriptive of the sequential linear model framework and support the interpretation of the assumptions.

Under these conditions, [20], Theorem 2, yields the existence of a local minimizer $\hat{\boldsymbol{\alpha}}_t^s$ of the SCAD criterion (6) satisfying $\Pr(\hat{\boldsymbol{\alpha}}_t^{s(2)} = \mathbf{0}) \to 1$ together with consistency on the active set at rate $O_p(\sqrt{K/n})$, which combined with A9 gives $\Pr(\hat{\alpha}_{jt}^s \neq 0) \to 1$ for all $j \in \mathcal{A}_t$ and hence conclusion (i) of Theorem 1. The same theorem also yields

$$\sqrt{n}\,\boldsymbol{A}_n\,\boldsymbol{\Sigma}_t^{1/2}\Big(\hat{\boldsymbol{\alpha}}_t^{s(1)} - \boldsymbol{\alpha}_t^{(1)}\Big) \xrightarrow{d} N\Big(\mathbf{0},\, \sigma^2\boldsymbol{H}\Big)$$

for $\boldsymbol{A}_n$ and $\boldsymbol{H}$ as in the statement of Theorem 1(ii), which is conclusion (ii).

The analogous argument for the finite-dimensional SCAD oracle property appears in [7], Theorems 1 and 2, and has been extended to high-dimensional settings by [15]. □

The selection-consistency conclusion of Theorem 1(i) automatically implies a lag-order selection guarantee, which we record as the following corollary.

Corollary 1 (Lag-order selection consistency)**.** *Define the true effective lag order at time t as*

$$L_t^* := \max\{s \in \{1,\ldots,t-1\} : \gamma_{t-s,t-1} \neq 0\},$$

and its SCAD-based estimator as

$$\hat{L}_t := \max\{s \in \{1,\ldots,t-1\} : \hat{\gamma}_{t-s,t-1}^s \neq 0\},$$

with the convention that the maximum of an empty set is zero. Under the conditions of Theorem 1, $\Pr(\hat{L}_t = L_t^*) \to 1$ *as* $n \to \infty$.

Proof. The lag coefficient $\gamma_{t-s,t-1}$ for $s \in \{1,\ldots,t-1\}$ corresponds to coordinate $j = p + t + 1 - s$ of $\boldsymbol{\alpha}_t$, since $\boldsymbol{\alpha}_t = (\boldsymbol{\beta}^\top, \boldsymbol{\gamma_{t-1}}^\top)^\top$ with $\dim(\boldsymbol{\beta}) = p + 1$. The map $s \mapsto p + t + 1 - s$ is a bijection between lag indices $\{1,\ldots,t-1\}$ and the lag coordinates $\{p+2,\ldots,p+t\}$ of $\boldsymbol{\alpha}_t$. Under this bijection, $\gamma_{t-s,t-1} \neq 0$ if and only if $(p + t + 1 - s) \in \mathcal{A}_t$, and analogously for the SCAD-selected counterpart. Hence, the event $\{\mathcal{A}_t^* = \mathcal{A}_t\}$ implies $\{\hat{L}_t = L_t^*\}$, and

$$\Pr(\hat{L}_t = L_t^*) \geq \Pr(\mathcal{A}_t^* = \mathcal{A}_t) \to 1$$

by Theorem 1(i). □

We note that Corollary 1 shows that the SCAD-penalized method performs predictor selection in $\boldsymbol{\beta}$ and lag-order selection in $\boldsymbol{\gamma_{t-1}}$ simultaneously and consistently within a single optimization, without requiring a separate information-criterion search over candidate lag orders. In the sequential setting where the nominal dimension of $\boldsymbol{\gamma_{t-1}}$ grows with t, this is a meaningful computational and statistical advantage.

3.3. Algorithmic Implementation

As a consequence of the sparsity property established in Theorem 1, the SCAD-penalized method selects the true model, yielding a sparse set of solutions and approximately unbiased coefficients for significant predictors and for significant lags of $\boldsymbol{y_t}$.

In the orthonormal special case, the SCAD objective (6) is separable in the coordinates, with closed-form minimizer

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

where $\operatorname{sgn}(\cdot)$ is the sign function and $\hat{\alpha}_{jt}$ denotes the unpenalized least-squares coefficient for the jth coordinate.

For the general non-orthonormal design that arises in model (2), Equation (9) is no longer an exact closed-form solution to problem (6). Instead, we compute $\hat{\boldsymbol{\alpha}}_t^s$ iteratively via coordinate-descent with SCAD thresholding: at each coordinate, the thresholding rule (9) is applied with all other coordinates held fixed (so $\hat{\alpha}_{jt}$ in (9) is the current partial-residual univariate least-squares update for coordinate j), and the algorithm iterates until the relative change in the objective falls below a convergence tolerance (we used 10^{-7} in our experiments). This is the solver implemented in the `ncvreg` R package 3.16.0, which we used throughout the simulations and applications. The convergence properties of this algorithm, together with the statistical rate of convergence of the resulting estimator, are discussed in Section 3.5.

The optimal pair (λ, a) can be obtained through a two-dimensional grid search using cross-validation, which is computationally expensive. Fan and Li [7] recommended setting $a = 3.7$ as a robust choice across diverse scenarios and showed that the choice of a does not materially affect overall performance relative to λ. We adopt $a = 3.7$ throughout and select λ by cross-validation.

In the context of sequential linear models, we employ the SCAD penalty as in Equations (7)–(9) to enhance variable selection and refine coefficient estimation. 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 [7] is more effective for parameter estimation and variable selection, especially in high-dimensional data, because it promotes sparsity, achieves asymptotic unbiasedness, and satisfies the oracle property. With advantages similar to the Lasso, the SCAD-penalized method removes variables with small coefficients. Because the SCAD penalty is non-concave, it applies heavy penalization to small coefficients while leaving large coefficients nearly unbiased. As the sample size grows, the SCAD-penalized method consistently identifies the true set of non-zero coefficients. By Corollary 1, this in particular identifies the true lag order in the sequential linear modeling and thereby improves predictive accuracy in models (1) and (2).

As shown in the simulation results in the next section, the SCAD penalty provides a better solution to penalized estimation and variable selection than the Lasso and adaptive Lasso. Rather than applying a uniform penalty to the coefficients, SCAD applies a non-

concave penalty that effectively shrinks small coefficients toward zero while preserving the large ones for more accurate estimation. Under assumptions A7 (restricted eigenvalue), A8 (sparsity rate), and A9 (minimum signal), the SCAD-penalized method achieves the oracle property by Theorem 1, even when the number of predictors exceeds the sample size. See also [21] for a broader overview of SCAD-type methods in high dimensions.

The regularization parameter λ controls shrinkage of the coefficients for the predictors $\boldsymbol{Z_t} = (\boldsymbol{X_t}, \boldsymbol{Y_{t-1}})$, and the coefficient vector $\boldsymbol{\alpha_t}$ combines the parameters for the independent variables $\boldsymbol{X_t}$ and the lagged coefficients for $\boldsymbol{Y_{t-1}}$.

Note that the sequential linear modeling often involves dependencies between observations at different time points, and the SCAD-penalized method helps mitigate multicollinearity issues by encouraging a parsimonious representation of the underlying patterns. Xie and Huang [10] 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.

3.4. Selection of Tuning Parameter

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

As discussed in the previous subsection, the SCAD-penalized method estimates the parameters in the sequential linear model by shrinking non-significant parameters to zero and reducing the bias of significant parameter estimates. Therefore, the SCAD-penalized method consistently selects the significant parameters. Fan and Li [7] showed that the SCAD method is applicable in both parametric and high-dimensional nonparametric settings. The convergence rate of the penalized-likelihood estimators depends on the regularization parameter, so an appropriately selected tuning parameter is needed for effective variable selection and accurate parameter estimation.

To facilitate tuning-parameter selection, cross-validation is commonly used. The sequential linear model involves additional complexities, such as the construction of lagged variables with time-effect coefficients, which requires a careful balance between penalizing coefficients and preserving important features. In this context, the SCAD penalty proves effective, which is also why its performance surpasses that of the other methods adopted here, as demonstrated in the simulations of the next section.

For the penalized methods with different penalty terms, we use cross-validation to choose the tuning parameter, which controls the weight of the shrinkage on parameter estimates. In the sequential linear model with the Lasso method, the ℓ_1 penalty function is

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

and the Lasso estimator admits the coordinate-wise thresholding rule

$$\hat{\alpha}_{jt}^{\,l} = \mathrm{sgn}(\hat{\alpha}_{jt})\left(|\hat{\alpha}_{jt}| - \lambda\right)_+,$$

where $\hat{\alpha}_{jt}$ denotes the unpenalized least-squares coefficient, $\mathrm{sgn}(\cdot)$ is the sign function, and $(\cdot)_+ = \max\{0, \cdot\}$.

With the adaptive Lasso method, the coordinate-wise penalty is

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

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 the adaptive Lasso to focus more on variables with larger estimated coefficients.

The purpose of using adaptive weights is to down-weight certain variables and up-weight others, allowing for more effective variable selection. Variables with larger Lasso estimates receive smaller weights, while variables with smaller Lasso estimates receive larger weights.

Before fitting the model via various methods, we split the data into training and test sets in the ratio 4:1. The training set is used to estimate the parameters, identify significant variables, and tune regularization parameters. The test set evaluates the model's predictive performance. One step before fitting is to tune the regularization parameter to minimize the mean squared error.

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

The prediction error serves as the criterion in cross-validation. With the training set fixed, cross-validation estimates the expected prediction error, and the validation set provides an assessment of the prediction model. In K-fold cross-validation, the available folds are used to fit the model, and the held-out fold is used to evaluate performance. The rule for selecting the best tuning parameter λ is to minimize the estimated prediction error.

According to Equation (7), the SCAD-penalized method selects variables with coefficients α_{jt} as follows.

1. When $|\alpha_{jt}|$ is small, the SCAD penalty behaves like the Lasso and shrinks the parameter to zero.
2. When $|\alpha_{jt}|$ is large, the SCAD penalty is constant, avoiding bias for large parameters.
3. When $|\alpha_{jt}|$ is moderate, the SCAD penalty attaches different weights to parameters.

3.5. Algorithmic Convergence and Rate of Convergence

Having defined the SCAD-penalized estimator, established its oracle property, described the coordinate-descent algorithm used to compute it, and specified the tuning-parameter selection procedure, we conclude the theoretical development of the method by addressing two further questions. First, does the coordinate-descent algorithm in Section 3.3 reliably produce an estimator to which Theorem 1 applies? Second, at what rate does this estimator approach the true parameter as the sample size grows? Because the SCAD objective (6) is non-convex, global optimization is generally intractable, so neither question is trivial. Both, nevertheless, admit theoretical answers that we summarize here.

Algorithmic convergence. The SCAD-penalized objective in (6) is minimized using the coordinate-descent algorithm of [22], implemented in the R package *ncvreg*. At each step, the algorithm updates a single coordinate by soft-thresholding a univariate quadratic majorant of the objective. Breheny and Huang [22], in their Proposition 1, show that the sequence of iterates produced by this procedure converges to a coordinate-wise minimum of (6), which is also a local minimum because the directional derivatives of the SCAD penalty exist everywhere. In our implementation, convergence is declared when the relative change in the objective falls below the tolerance 10^{-7}. Moreover, they [22] establish a convexity diagnostic identifying regions of the parameter space where the SCAD objective is locally convex despite the non-convexity of the penalty; within such a region, the local minimum

is the unique global minimum, and the coordinate-descent output coincides with the global optimizer there.

Selection of the oracle local minimum. The SCAD objective may possess multiple local minima, and the oracle property of Theorem 1 holds for one of them. A relevant question is therefore whether the specific local minimum returned by our coordinate-descent solver is the one that attains the oracle property. Fan, Xue, and Zou [23] provide the sharpest available answer. They show that, under a localizability condition on the oracle estimator and the SCAD penalty, the local linear approximation (LLA) algorithm initialized by a $\sqrt{n}$-consistent estimator, such as the Lasso, converges to the oracle estimator in a finite number of iterations with probability tending to one. Our implementation uses coordinate-descent rather than explicit LLA, but *ncvreg* constructs the regularization path through warm-starts, so each SCAD fit is initialized from a sparse, near-Lasso solution obtained at the previous point on the path. This is closely analogous to the $\sqrt{n}$-consistent initialization required by [23], and their theoretical guarantee is informative for the practical behavior of the solver.

Statistical rate of convergence. The oracle property stated in Theorem 1 directly implies a rate of convergence for the estimator on the active set. Specifically, the asymptotic normality statement in Theorem 1(ii) yields

$$\|\hat{\boldsymbol{\alpha}}_t^{s(1)} - \boldsymbol{\alpha}_t^{(1)}\|_2 = O_p\Big(\sqrt{K/n}\Big),$$

where $K = |\mathcal{A}_t|$ is the cardinality of the true active set. This is the classical parametric $n^{-1/2}$ rate, the same rate attained by the oracle least-squares estimator that knows the active set in advance. Combined with the selection-consistency conclusion of Theorem 1(i), this shows that the SCAD-penalized sequential linear model estimator matches the oracle estimator both in support recovery and in rate of convergence, as $n \to \infty$.

Taken together, these three observations provide the theoretical basis for the numerical performance reported in Section 4. The algorithm converges to a local minimum by [22], that local minimum coincides with the oracle estimator when the algorithm is initialized appropriately by [23], and the oracle estimator attains the parametric $n^{-1/2}$ rate by Theorem 1.

4. Simulations

4.1. Simulation Settings

To check the performance of the SCAD-penalized method in the sequential linear model (2), we conduct a simulation study with fixed sample size and different numbers of predictors, along with a set of continuous outcomes based on fifty time points $t = 1, \ldots, 50$. Based on each simulation setting, we compare the SCAD-penalized method with OLS, Lasso, and adaptive Lasso in traditional linear and sequential linear models. The optimal tuning parameter λ is selected in the Lasso, SCAD-penalized, and adaptive Lasso methods, respectively, via 10-fold cross-validation.

Consider a dataset with repeated measures of outcomes $y_{i,t}$ for $i = 1, \ldots, n$ through time points $t = 1, \ldots, T$. A model is trained to sequentially predict the outcome at each time point, with baseline data matrix $\boldsymbol{x_i} = (x_{i,1}, \ldots, x_{i,p})^T$ and parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$. 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 $\epsilon_{i,1} \sim N(0, \sigma^2)$.

For any subsequent time point t, the sequential linear model with time series is

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

where $\boldsymbol{y_t} = (y_{1,t}, \ldots, y_{n,t})^T$ and $\boldsymbol{\epsilon_t} = (\epsilon_{1,t}, \ldots, \epsilon_{n,t})^T$. Model (10) is iterated to generate data until $t = T$. At the initial time point $t = 1$, the model does not involve the term $\boldsymbol{Y_{t-1}\gamma_{t-1}}$.

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

The true regression coefficients are $\beta_1 = \beta_3 = 2$, $\beta_5 = \beta_7 = \beta_9 = -3$, with the remaining β_j equal to 0. We set the effective lag order to $L = 1$, with true time-effect coefficient $\gamma = 0.6$. We take $T = 50$. The data are split in a 4 : 1 ratio into training and test sets. This setup is an explicit instance of the compatible construction in Remark 1.

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

4.2. Simulation Results for Estimation and Variable Selection

The simulation results for the estimation of the coefficients and the mean prediction test errors are averaged over 1000 replicates.

For estimation, the average prediction error is used as a measure. Table 1 features the average prediction errors for each method.

According to [24], the relative risk (RR) is the accuracy metric for comparing methods. We expect the RR to be as small as possible, which denotes a more accurate coefficient estimate. The formula for the relative risk is

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

where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\boldsymbol{Z_t}$, and $\boldsymbol{Z_t} = (\boldsymbol{X_t}, \boldsymbol{Y_{t-1}})$. We compute RR at the terminal time point $t = T$. The perfect score is 0, which means $\hat{\boldsymbol{\alpha}}_t$ equals $\boldsymbol{\alpha_t}$, while RR = 1 means $\hat{\boldsymbol{\alpha}}_t$ 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 dimensions, the average absolute prediction error of the SCAD-penalized method is among the smallest across methods.

For the sequential linear model, Table 1 illustrates that the SCAD-penalized method has a relatively small mean absolute prediction error, indicating high prediction accuracy. For instance, in the medium dimension ($p = 100$) with time series, the estimated $\hat{\gamma}$ is 0.56. The average prediction errors for the Lasso, SCAD, and 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. Similarly, at high dimensions ($p = 1000$), the SCAD-penalized method attains the smallest MAPE in the SLM.

Table 2 also shows that, when fitted by the sequential linear model, the SCAD-penalized method generally achieves low mean RR values, confirming that it performs best in estimating parameters among all the methods considered.

Additionally, for the generated data with time series, the RR results in Table 2 and the MAPE results in Table 1 imply that sequential linear modeling functions better than a simple linear model. This confirms the advantages of sequential linear modeling in models (2) and (3). When fixed effects and time-series data are involved, sequential linear modeling is a good choice, mirroring the role of the ADL for dynamic data as discussed in the Introduction.

In the generated data, we also conduct simulations for variable selection. We set five non-zero true β_j values as the significant variables—the first, third, fifth, seventh, and ninth—while the rest are insignificant. The individual empirical 95% interval for each parameter is calculated as the range from the 2.5% to the 97.5% quantile of the estimates across the 1000 replications.

To examine 95% estimation accuracy of significant variables using the OLS, Lasso, SCAD-penalized, and adaptive Lasso methods, the empirical 95% intervals (hereafter abbreviated "ACI" following the original notation) for all significant variables and for the number of significant variables are summarized in Tables 3–5 for the three dimensions. The methods are fitted both by the traditional linear model (LM) and the sequential linear model (SLM).

Table 3 gives these intervals in the low-dimension setting. Based on the simulated data with $p = 10$, we expect the interval for the number of significant variables to be close to the truth (5 in the LM; 6 in the SLM, counting the lag coefficient). In the LM, the adaptive Lasso attains the tightest interval around the truth. However, in the SLM, the SCAD-penalized method is the best, with the tightest intervals for both the number of significant variables (6 to 9.53) and for the individual parameters, and the confidence interval for the lag coefficient is (0.59, 0.66), which is the tightest among all methods.

In the medium-dimension setting, we compare the three penalized methods. In the SLM, SCAD yields 95% empirical intervals of (6, 11) for the number of significant variables and (0.60, 0.68) for the lag coefficient, both the tightest among the three methods. Details are in Table 4.

Table 5 shows that in the high-dimension setting, the SCAD-penalized method again attains the best empirical 95% intervals. With $p = 1000$ predictors and 1000 repetitions, SCAD yields intervals of (5, 12.53) for the LM and (6, 17) for the SLM on the number of significant variables—the tightest among the three methods—and its interval for the individual coefficients is also close to the truth. Its interval for the lag-effect coefficient γ is (0.59, 0.68), again tighter than the Lasso's. The adaptive Lasso has a 95% empirical interval for γ that is outside the true value range.

Table 3. Individual 95% empirical 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)
$\beta_5 = -3$	$(-3.12, -2.86)$	$(-3.08, -2.83)$	$(-3.12, -2.86)$	$(-3.11, -2.85)$
$\beta_7 = -3$	$(-3.14, -2.85)$	$(-3.11, -2.80)$	$(-3.13, -2.85)$	$(-3.13, -2.83)$
$\beta_9 = -3$	$(-3.14, -2.87)$	$(-3.11, -2.83)$	$(-3.14, -2.87)$	$(-3.13, -2.85)$
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)

Table 4. Individual 95% empirical 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)
$\beta_5 = -3$	$(-3.05, -2.73)$	$(-3.16, -2.85)$	$(-3.15, -2.83)$
$\beta_7 = -3$	$(-3.04, -2.71)$	$(-3.15, -2.83)$	$(-3.15, -2.82)$
$\beta_9 = -3$	$(-3.04, -2.72)$	$(-3.16, -2.84)$	$(-3.15, -2.83)$
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. Individual 95% empirical 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)
$\beta_5 = -3$	$(-2.98, -2.67)$	$(-3.16, -2.85)$	$(-3.15, -2.83)$
$\beta_7 = -3$	$(-2.97, -2.69)$	$(-3.14, -2.88)$	$(-3.13, -2.84)$
$\beta_9 = -3$	$(-2.96, -2.60)$	$(-3.16, -2.82)$	$(-3.13, -2.80)$

Table 5. *Cont.*

	ACI (LM)		
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)
$\beta_5 = -3$	$(-2.97, -2.01)$	$(-3.28, -2.71)$	$(-3.01, -0.34)$
$\beta_7 = -3$	$(-2.94, -2.11)$	$(-2.70, -2.69)$	$(-2.85, -0.94)$
$\beta_9 = -3$	$(-2.80, -2.15)$	$(-3.24, -2.67)$	$(-2.83, -0.79)$
$\gamma = 0.6$	(0.52, 0.63)	(0.59, 0.68)	(0.00, 0.53)

5. Applications

In the previous section, the simulation results demonstrated that the SCAD-penalized method has advantages over the other penalized methods considered here. To further compare the effectiveness of these methods, we apply them to two longitudinal study datasets—Heart.valve from the joineR package [18], and Austin House Price from [25]—for estimation and variable selection.

5.1. Heart Valve Data

5.1.1. Data Description

The data applied here concern aortic valve replacement surgery. They come from the joineR package in R [18] and longitudinally measure heart function outcomes after surgery in an observational study. For each patient, the effective subset is the part where the patient's observation time point is greater than or equal to 5. Within this sub-dataset, the observation time point equal to 5 is used as the training set, and time points greater than 5 as the test set. Among the various preoperative predictors, log.lvmi is the response variable. The longitudinal column lvmi is 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 on the day of surgery, years), status (censoring indicator, 1 = died and 0 = lost at follow-up), prenyha (preoperative New York Heart Association 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), and emergence (operative urgency, 0 = elective, 1 = urgent, 3 = emergency).

5.1.2. Results for Estimation and Variable Selection

To evaluate the log of the left ventricular mass index based on patients' factors at the five follow-up visits, Figure 1 displays a boxplot of the log.lvmi values, showing quartiles and averages at each time point. The distribution of the standardized log of the left ventricular mass index at each follow-up visit after surgery is approximately symmetric. Only at time point 5 does the response variable show right skewness. The medians are approximately the same across time, indicating that the log.lvmi values at sequential time points are highly correlated and that the current response depends strongly on past time points.

From the medical side, the lvmi value indicates the left ventricular mass index (LVMI), discussed by [26], which is an important measure of the likelihood of heart morbidity and mortality among adults with hypertension. LVMI has been shown in [18] to be strongly correlated with post-surgery heart recovery.

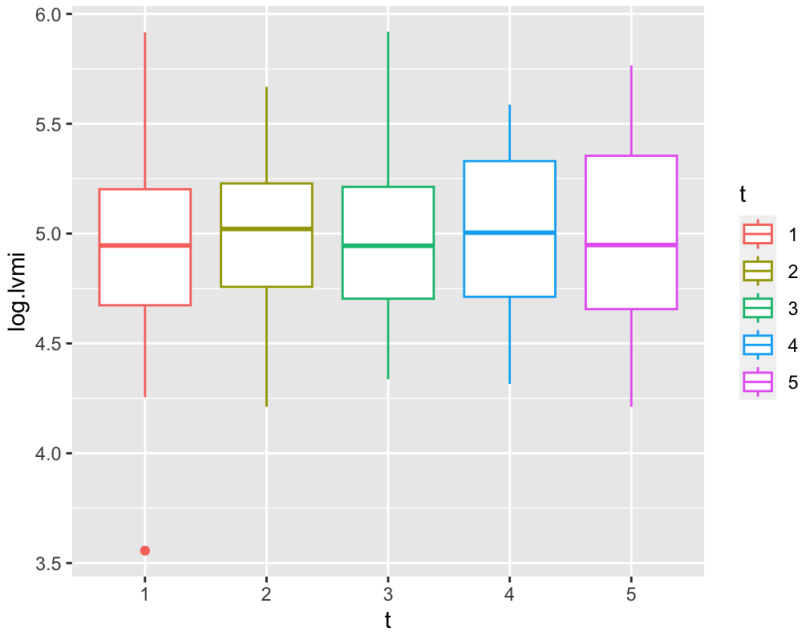


Figure 1. Boxplot of response variable (log.lvmi) for 5 time points.

Two types of models are applied here: the simple linear model and the sequential linear model of Equation (3). The methods compared in each are OLS, Lasso, SCAD-penalized, and adaptive Lasso. Table 6 reports the mean absolute prediction errors. The SCAD-penalized method (together with adaptive Lasso) attains an MAPE of 0.001 in the SLM, which is the lowest among all methods.

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 variable selection, the estimated coefficients for the variables selected by SCAD in the LM are summarized in Table 7, and in the SLM in Table 8. Table 7 shows the significant predictors of log.lvmi in the simple linear model. Table 8 shows that in the sequential linear model, log.lvmi as the response variable depends on the previous time point only through y_4 (the fourth lag variable). Because the lagged dependent variables at consecutive time points are highly correlated, the SCAD-penalized method removes the other lagged response variables to improve prediction accuracy, so that only one lagged response variable (y_4) enters the SLM.

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

Table 7. *Cont.*

Selected Variables	Estimated Coefficients (LM)
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

To visually demonstrate the estimation and variable selection in this dataset, Figure 2 shows the coefficient estimates for significant variables in the linear model. For the sequential linear model, only one variable is selected (y_4, with estimated coefficient 1.047, already listed in Table 8), so a separate coefficient figure is not informative and is omitted. The SCAD method in the LM detects thirteen significant variables with the estimated coefficients shown in Table 7, while the SCAD method in the SLM detects only one significant variable because the other lagged dependent variables are highly correlated. By removing unnecessary lagged dependent variables, the SCAD method improves prediction accuracy. See Table 6, which shows that the SLM achieves a dramatically better predictive behavior.

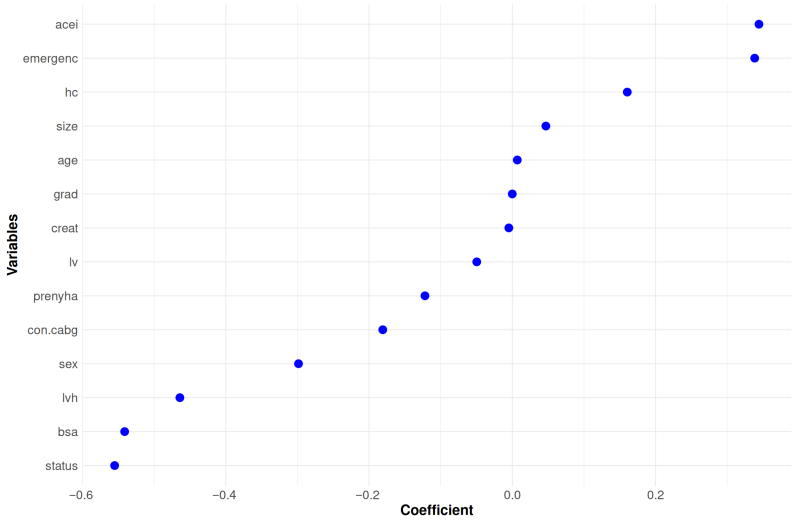


Figure 2. Coefficients estimates for significant variables in LM.

Last but not least, we check the regression assumptions for the SLM, with output plots shown in Figure 3. The residual plot shows that the error terms roughly follow a normal distribution with some skewness in both tails of the Q–Q residuals plot. The error terms

are spread approximately uniformly along the ranges of points in the Scale–Location plot, indicating approximately constant variance and no clear pattern. The significant predictors are independent. Based on these plots, the SLM assumptions are satisfied.

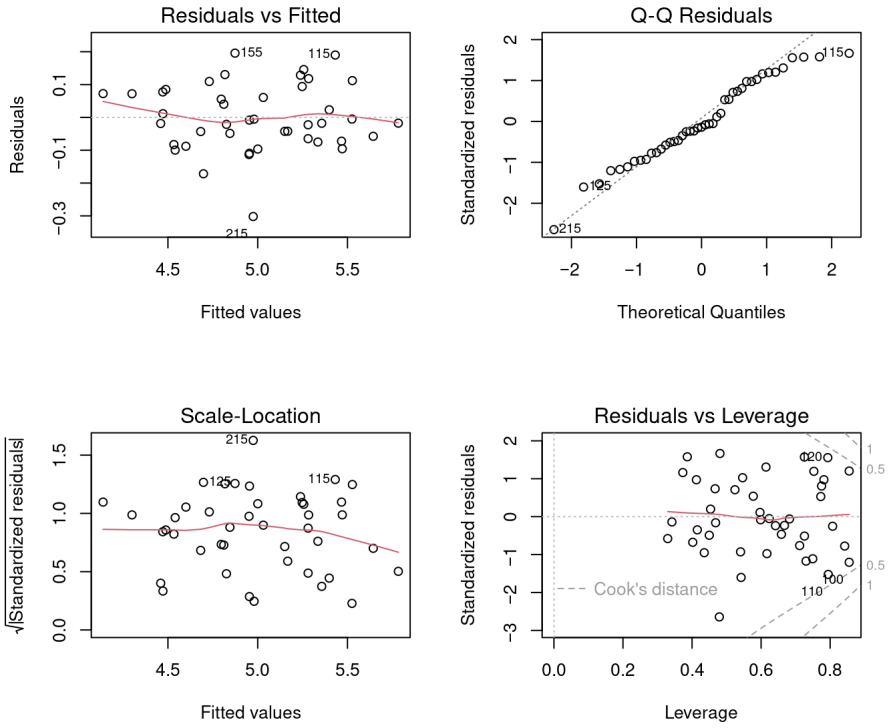


Figure 3. SLM diagnostics in Heart.valve data.

5.2. Austin House Price Data

5.2.1. Data Description

This dataset is collected from Kaggle [25], originally sourced from a project by Eric Pierce. In 2021, the Austin housing market was one of the most dynamic markets, and these listings provide insight into its evolution over recent years. The dataset includes various categories that affect housing prices. After cleaning irrelevant predictors, 38 variables remain in this application.

The data frame consists of 15,171 observations for each of 38 predictors, including hasAssociation (whether there is a Homeowners Association associated with the listing), hasSpa (whether the home has a spa), numOfPhotos (the number of photos in the Zillow listing), avgSchoolRating (the average school rating across all school types in the Zillow listing), and numOfBathrooms (the number of bathrooms in the 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. For data cleaning, we first convert character variables into factors in R. We then 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. We create a new data frame containing two variables, "yearbuilt" and "repeated numbers", which allows us to match the original dataset and select each segment of the data accurately. We split the data into training and test sets: the training set consists of duplicated yearbuilt values with a count less than or equal to 180, while the remaining data form the test set.

After this manipulation, we have 37 unique time points. The sequential linear model includes the 37th time point as the outcome and the previous 36 years' property prices

as predictor variables. We apply OLS, Lasso, SCAD-penalized, and adaptive Lasso in the sequential linear model to compare these methods.

Figure 4 plots the mean of the latest property prices over the 37 time points. The mean fluctuates considerably and exhibits an overall increasing trend with evident price changes over time, so the house price should have time-series dependent variables and other fixed-effect independent variables in sequential linear modeling.

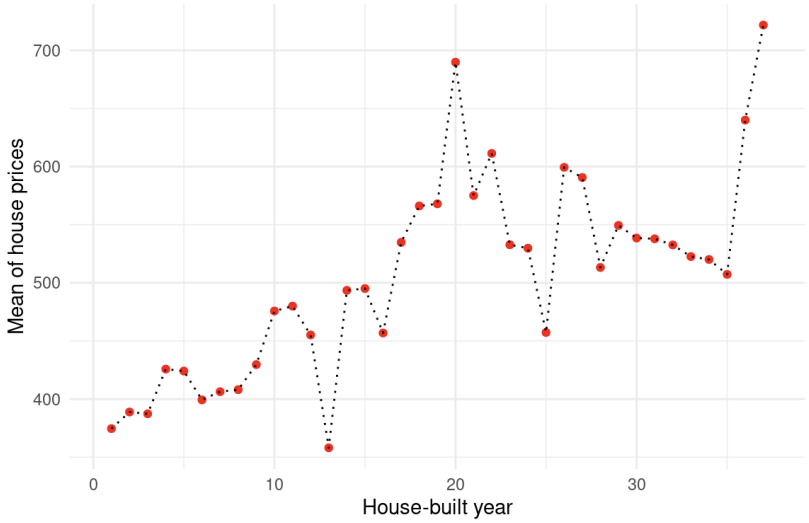


Figure 4. Mean of the latest house prices built over time points.

For estimation and variable selection, the SCAD-selected coefficient estimates in the SLM are summarized in Table 9. The current house price depends on the previous time point at $t = 25$ (y_{25}), and on several independent variables: hasAssociation, hasSpa, numOfPhotos, avgSchoolRating, and numOfBathrooms.

The map in Figure 5 shows the distribution of properties. Each point symbolizes a property, colored red (lowest recent prices), yellow (second-lowest), green (moderate), and blue (highest). The map reveals which areas command relatively higher property values.

As shown in Table 9, the selected variables—number of bathrooms, presence of a view, presence of a spa, and being part of an association—and one lagged dependent variable (y_{25}) appear. The other lagged dependent variables are removed from the sequential linear model due to high correlation. These selected variables significantly affect property prices in Austin, TX, in recent years. The number of bathrooms and the presence of a spa positively affect prices, while being part of an association negatively affects them.

Table 10 reports the mean absolute prediction errors. Again, the SCAD-penalized method achieves the smallest value (89.089) among the penalized methods.

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

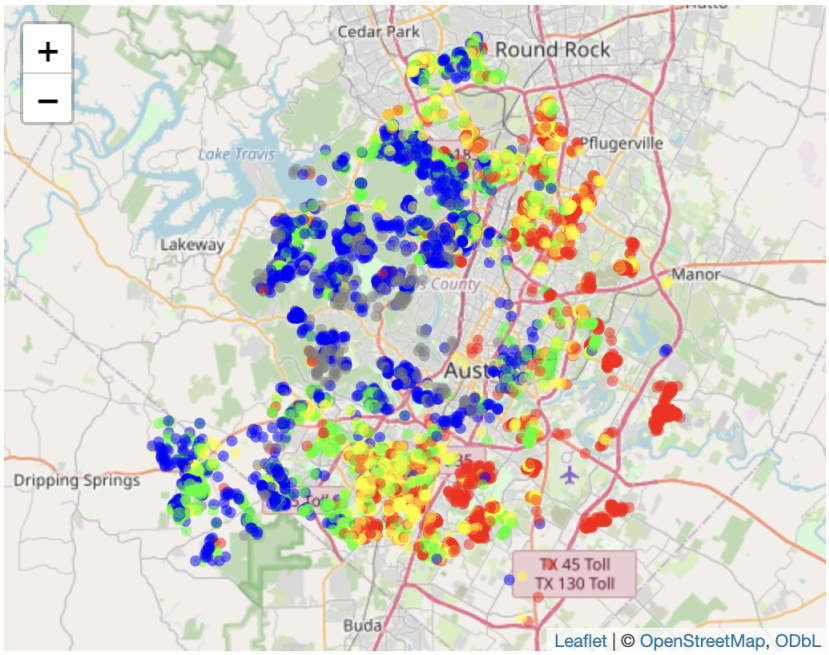


Figure 5. Distribution of properties' leaflet map.

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 significant variable selection, the Lasso coefficients in the sequential linear model are shown in Figure 6. In addition to the variables mentioned above (number of bathrooms, view, spa, association), the Lasso-selected model also includes several lagged dependent variables. Figure 6 clearly reflects the sparsity property: the estimated coefficients of insignificant variables are shrunk to zero.

From the coefficient plot for the adaptive Lasso in Figure 7, the selected variables are a subset of the Lasso's selection, and none of the lagged dependent variables is retained: building on the Lasso coefficients, the adaptive Lasso applies different penalty weights to each coefficient and removes all lagged dependent variables due to high correlation. The adaptive Lasso selects fewer significant variables than the Lasso, consistent with the Lasso over-fitting relative to the adaptive Lasso. With fewer significant variables, the adaptive Lasso outperforms the Lasso in prediction (see Table 10).

In Figure 8, the SCAD-selected variables include those selected by the adaptive Lasso and, additionally, the lagged dependent property price at the 25th time point. The number of bathrooms and the presence of a spa positively affect property prices, while being part of an association negatively affects them. We conclude that the SCAD method not only selects the predictors in the SLM but also selects the order of the lagged dependent variables, in accordance with Corollary 1. This joint selection property is what makes the SCAD method particularly well-suited to sequential linear models.

Among the penalized methods, the SCAD method attains the smallest mean absolute prediction error of 89.089 in Table 10.

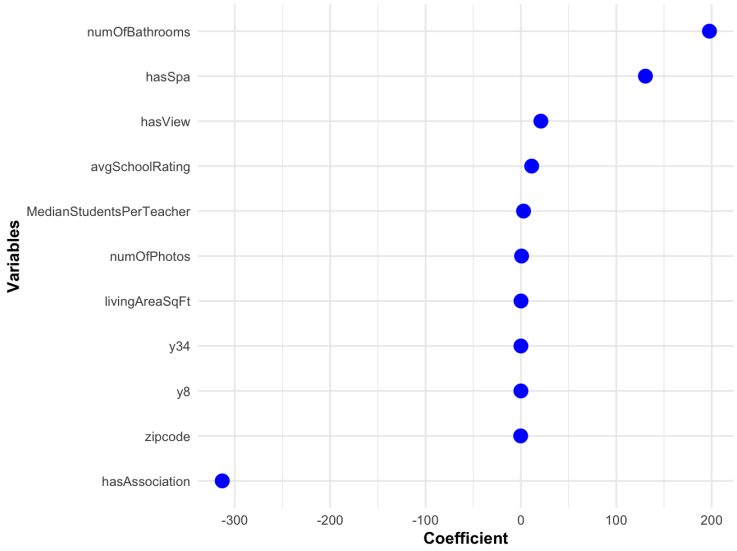


Figure 6. Coefficients estimates for significant variables via the Lasso method.

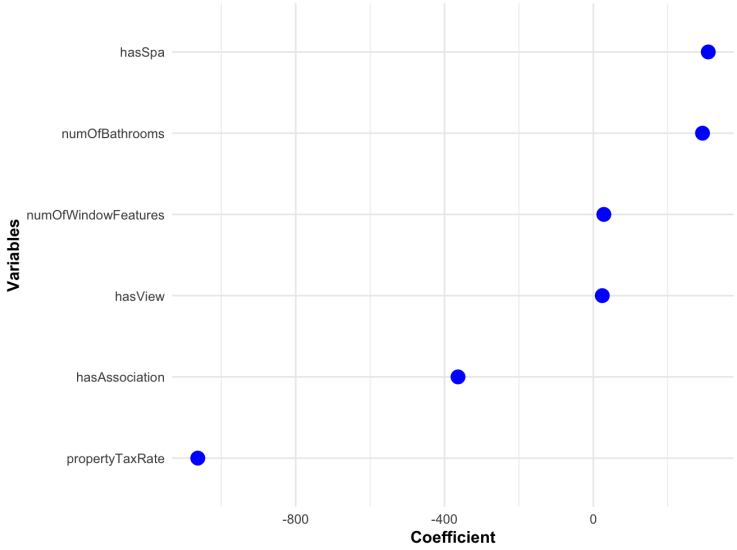


Figure 7. Coefficients estimates for significant variables via adaptive Lasso method.

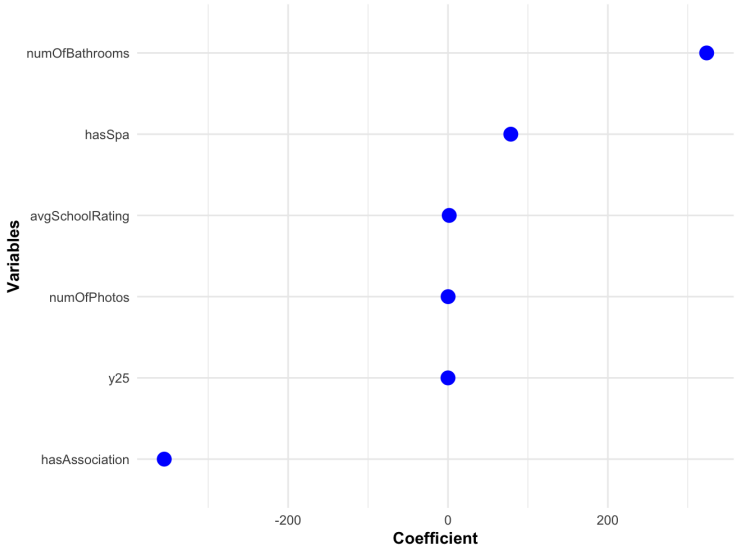


Figure 8. Coefficients estimates for significant variables via SCAD method.

Last but not least, we check the regression assumptions for the SLM. Output plots are shown in Figure 9. The first residual plot shows a nonlinear pattern, suggesting that some nonlinear predictors may be needed in modeling the response variable. The Q–Q plot exhibits an approximately normal distribution with slight skewness in both tails. In the Scale–Location plot, heteroscedasticity appears possible and should be considered in variance estimation. A few outliers are visible in the leverage plot. Overall, the disparities from the SLM assumptions are not severe, and the sequential linear models used here capture the trend of housing prices well. The estimation and selected variables are informative for model interpretation.

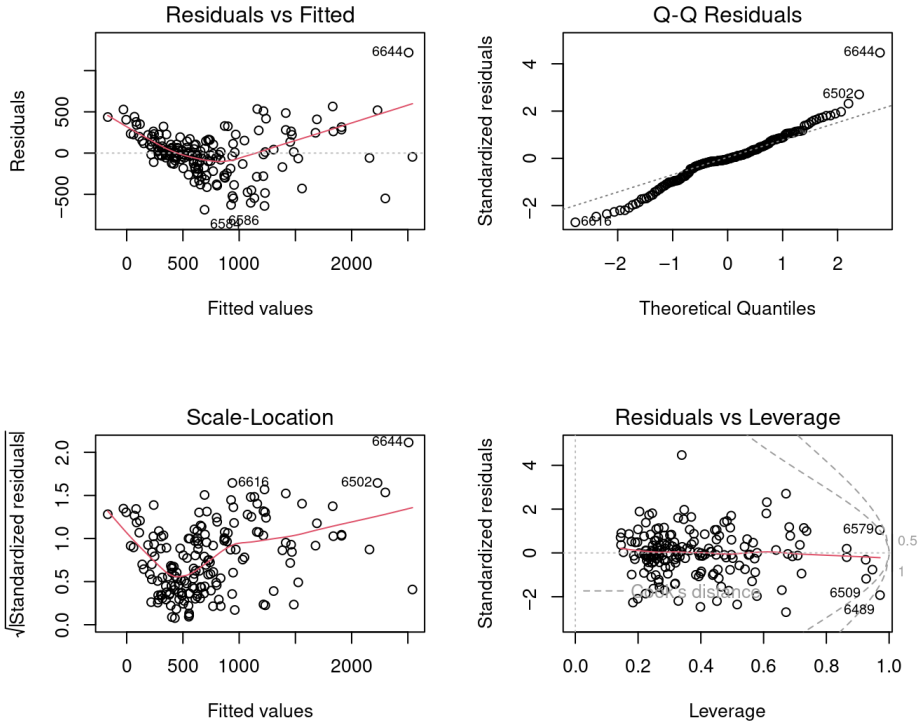


Figure 9. SLM diagnostics in Austin, TX house price data.

6. Concluding Remarks and Discussion

In sequential linear modeling, the measurements at each time point enable the prediction of data at intermediate time points, improving prediction accuracy because correlation structures are correctly captured. With the focus on the sequential linear regression model, lagged dependent variables are repeatedly included as input variables in the next predictive models. We therefore explore sequential linear modeling and its performance in estimation and variable selection.

For the sequential linear model proposed in this study, we introduce the SCAD-penalized method to improve estimation and variable selection. We establish the oracle property of the estimator (Theorem 1) and show that lag-order selection is consistent as a corollary (Corollary 1). The SCAD penalty is expected to perform better in estimation and variable selection than the Lasso and adaptive Lasso. The Lasso performs variable selection by shrinking coefficients to zero, useful in medium or high dimensions. The adaptive Lasso applies adaptive penalty weights to parameter estimates, improving over the Lasso. The SCAD-penalized method has different shrinkage at different coefficient magnitudes: for small coefficients, it behaves like the Lasso and shrinks to zero. For large coefficients, the penalty is constant, reducing bias and keeping significant signals. And for moderate coefficients, the penalty grows at a slower rate than the Lasso, reducing shrinkage.

We conduct simulations to explore the performance of the proposed SCAD-penalized method compared with Lasso and adaptive Lasso. OLS is used in low-dimension settings for reference. Across low, medium, and high dimensions, we generate base datasets with $p = 10, 100$, and 1000 with 1000 replications, and the error terms follow the standard normal distribution. The algorithmic and statistical convergence of the proposed estimator, including the $\sqrt{n}$-rate on the active set, are discussed theoretically in Section 3.5 by appeal to the results of [22,23].

By incorporating lagged dependent variables as predictors at each time point, we compare OLS, Lasso, SCAD, and adaptive Lasso. OLS applies only to low-dimensional settings and does not function well when p exceeds n.

The simulation results show that the SCAD-penalized method improves over the Lasso by reducing bias in parameter estimation and lowering prediction errors. The SCAD penalty improves estimation accuracy and variability, yielding better predictions in medium- and high-dimensional settings. Results are summarized through the MAPE, RR, and empirical 95% interval (ACI) outputs across methods and scenarios. The 95% empirical intervals show that the SCAD-penalized method yields results closest to the true settings, both for the number of significant variables and for the true parameter values. Notably, in small datasets OLS performs well, while in medium- to high-dimensional datasets the SCAD-penalized method is superior for parameter estimation and variable selection.

To further explore the performance of the proposed SCAD-penalized method, we applied it to two real datasets in comparison with OLS, Lasso, and adaptive Lasso. The first dataset is the Heart Valve replacement surgery data from the joineR package [18]. SCAD attains the lowest mean absolute prediction error in both linear and sequential models. The second dataset is the latest housing prices in Austin, TX [25]. Although the modeling assumptions are not perfectly satisfied, removing extreme outliers or taking the log of the response variable strengthens the linear relationship between the predictors and the log housing price. The SCAD-penalized method demonstrates more stable performance in variable selection, incorporating most of the significant variables also identified by the adaptive Lasso.

In future work, we will consider analyzing datasets with strong multicollinearity and comparing OLS, Lasso, SCAD, and adaptive Lasso in that setting. Many real datasets entail high correlation among variables, which motivates the application of various penalized methods to sequential linear modeling when selecting significant lagged dependent variables from the multivariate time series. We will also explore hypothesis tests to determine the minimum sample size needed for reliable model prediction in sequential linear modeling, and estimate the sample size required so that the model is robust to significant variables and yields 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 analyses and wrote the original draft; J.S. supervised, revised, and finalized the manuscript; C.G. critically reviewed, commented on, and revised the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are openly available in Austin house price data at https://www.kaggle.com/datasets/ericpierce/austinhousingprices (accessed on 12 April 2021) [25].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pesaran, M.H.; Shin, Y. An Autoregressive Distributed Lag Modelling Approach to Cointegration Analysis. In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium, Chapter 11, Econometric Society Monographs*; Strøm, S., Ed.; Cambridge University Press: Cambridge, UK, 1999; pp. 371–413. [CrossRef]
2. Keele, L.; Kelly, N.J. Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables. *Political Anal.* **2006**, *14*, 186–205. [CrossRef]
3. Shestopaloff, K.; Canizares, M.; Power, J.D. A Sequential Modeling Approach for Predicting Clinical Outcomes with Repeated Measures. *Commun. Stat. Theory Methods* **2022**, *52*, 7465–7478. [CrossRef]
4. Qiu, J.; Li, D.; You, J. SCAD-Penalized Regression for Varying-Coefficient Models with Autoregressive Errors. *J. Multivar. Anal.* **2015**, *137*, 100–118. [CrossRef]
5. Nicholson, W.B.; Wilms, I.; Bien, J.; Matteson, D.S. High Dimensional Forecasting via Interpretable Vector Autoregression. *J. Mach. Learn. Res.* **2020**, *21*, 1–52.
6. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodol.)* **1996**, *58*, 267–288. [CrossRef]
7. Fan, J.; Li, R. Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties. *J. Am. Stat. Assoc.* **2001**, *96*, 1348–1360. [CrossRef]
8. Zou, H. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
9. Zou, H.; Hastie, T. Regularization and Variable Selection via the Elastic Net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, *67*, 301–320. [CrossRef]
10. Xie, H.; Huang, J. SCAD-Penalized Regression in High-Dimensional Partially Linear Models. *Ann. Stat.* **2009**, *37*, 673–696. [CrossRef]
11. Wang, H.; Li, R.; Tsai, C.-L. Tuning Parameter Selectors for the Smoothly Clipped Absolute Deviation Method. *Biometrika* **2007**, *94*, 553–568. [CrossRef]
12. Caner, M.; Knight, K. An Alternative to Unit Root Tests: Bridge Estimators Differentiate between Nonstationary versus Stationary Models and Select Optimal Lag. *J. Stat. Plan. Inference* **2013**, *143*, 691–715. [CrossRef]
13. Kock, A.B. Oracle Inequalities, Variable Selection and Uniform Inference in High-Dimensional Correlated Random Effects Panel Data Models. *J. Econom.* **2016**, *195*, 71–85. [CrossRef]
14. Gu, C.; Ratnasingam, S. Change Point Detection in SCAD-Penalized Dynamic Panel Models. *Seq. Anal.* **2025**, *44*, 377–403. [CrossRef]
15. Kim, Y.; Choi, H.; Oh, H.-S. Smoothly Clipped Absolute Deviation on High Dimensions. *J. Am. Stat. Assoc.* **2008**, *103*, 1665–1673. [CrossRef]
16. Stock, J.H.; Watson, M.W. *Introduction to Econometrics*; Addison-Wesley Series in Economics; Addison-Wesley: Boston, FL, USA, 2003.
17. Brockwell, P.J.; Davis, R.A. *Introduction to Time Series and Forecasting*, 2nd ed.; Springer: New York, NY, USA, 2002. [CrossRef]
18. Lim, E.; Ali, A.; Theodorou, P.; Sousa, I.; Ashrafian, H.; Chamageorgakis, T.; Duncan, A.; Henein, M.; Diggle, P.; Pepper, J. Longitudinal Study of the Profile and Predictors of Left Ventricular Mass Regression after Stentless Aortic Valve Replacement. *Ann. Thorac. Surg.* **2008**, *85*, 2026–2029. [CrossRef] [PubMed]
19. Zhao, P.; Yu, B. On Model Selection Consistency of Lasso. *J. Mach. Learn. Res.* **2006**, *7*, 2541–2563.
20. Fan, J.; Peng, H. Nonconcave Penalized Likelihood with a Diverging Number of Parameters. *Ann. Stat.* **2004**, *32*, 928–961. [CrossRef]
21. Fan, J.; Li, R. Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery. *arXiv* **2006**, arXiv:math/0602133. [CrossRef]
22. Breheny, P.; Huang, J. Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection. *Ann. Appl. Stat.* **2011**, *5*, 232–253. [CrossRef]
23. Fan, J.; Xue, L.; Zou, H. Strong Oracle Optimality of Folded Concave Penalized Estimation. *Ann. Stat.* **2014**, *42*, 819–849. [CrossRef]
24. Hastie, T.; Tibshirani, R.; Tibshirani, R.J. Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons. *Stat. Sci.* **2020**, *35*, 579–592. [CrossRef]
25. Pierce, E. Austin, TX House Listings [Data Set]. Kaggle. 2021. Available online: https://www.kaggle.com/datasets/ericpierce/austinhousingprices (accessed on 12 April 2021).
26. Shamszad, P.; Slesnick, T.C.; Smith, E.O.; Taylor, M.D.; Feig, D.I. Association between Left Ventricular Mass Index and Cardiac Function in Pediatric Dialysis Patients. *Pediatr. Nephrol.* **2012**, *27*, 835–841. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.