Computational Statistics and Data Analysis 183 (2023) 107729

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Tuning parameter selection for penalized estimation via R^2

Julia C. Holter, Jonathan W. Stallrich *

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA

ARTICLE INFO

Article history:
Received 3 June 2022
Received in revised form 14 November 2022
Accepted 19 February 2023
Available online 28 February 2023

Keywords:
Cross validation
Model/variable selection
Functional data
Relaxed lasso

ABSTRACT

The tuning parameter selection strategy for penalized estimation is crucial to identify a model that is both interpretable and predictive. However, popular strategies (e.g., minimizing average squared prediction error via cross-validation) tend to select models with more predictors than necessary. A simple yet powerful cross validation strategy is proposed which is based on maximizing the squared correlation between the observed and predicted values, rather than minimizing squared error loss for the purposes of support recovery. The strategy can be applied to all penalized least-squares estimators and, under certain conditions, the metric implicitly performs a bias adjustment named the α-modification. When applied to the Lasso estimator, the α-modification is closely related to the relaxed Lasso estimator. The approach is demonstrated on a functional variable selection problem to identify optimal placement of surface electromyogram sensors to control a robotic hand prosthesis.

1. Introduction

Many statistical problems aim to build a predictive model from a large set of potential predictor variables.[1] Variable selection is often performed to select a predictive model that depends on as few predictor variables as possible. For example, Stallrich et al. (2020) discussed an important functional variable selection problem to develop a prosethesis controller (PC) for a robotic hand. Electromyogram (EMG) signals from surface sensors placed on the residual forearm muscles of an amputee were input into a PC and translated into movement of the robotic hand. For able-bodied subjects, it is known that certain movements are caused by contractions of only a few muscles, implying a predictive PC requires a few strategically-placed EMG sensors.

This paper concerns problems that are well approximated by a sparse linear model:

$$\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\boldsymbol{\beta}^*_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}\,, \tag{1}$$

where $E(\boldsymbol{\epsilon}) = \mathbf{0}$, $V(\boldsymbol{\epsilon}) = \boldsymbol{\Sigma}$, $\boldsymbol{\beta}^{*T} = (\beta^*_1, ..., \beta^*_{p^*}, 0, ..., 0)$, and p^* is the number of important variables. Without loss of generality, assume $\mathbf{y}$ and predictor variables, $\mathbf{X}$, are centered and the diagonals of $\mathbf{X}^T\mathbf{X}$ equal n. Let $\mathcal{M}^* = \{j : \beta^*_j \neq 0\}$ denote the support of $\boldsymbol{\beta}^*$. A predictive model's estimate for $\boldsymbol{\beta}^*$, denoted $\hat{\boldsymbol{\beta}}$, will ideally also have support $\mathcal{M}^*$ and will be close to $\boldsymbol{\beta}^*$ in some other sense, such as $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2 = \sum_j(\hat{\beta}_j - \beta^*_j)^2$.

* Corresponding author.
E-mail address: jwstalli@ncsu.edu (J.W. Stallrich).

[1] Note that more detailed proofs for all results, additional simulation results, the EMG data from the application, and all R code for this paper have been made available in the Supplementary Material.

https://doi.org/10.1016/j.csda.2023.107729
0167-9473/

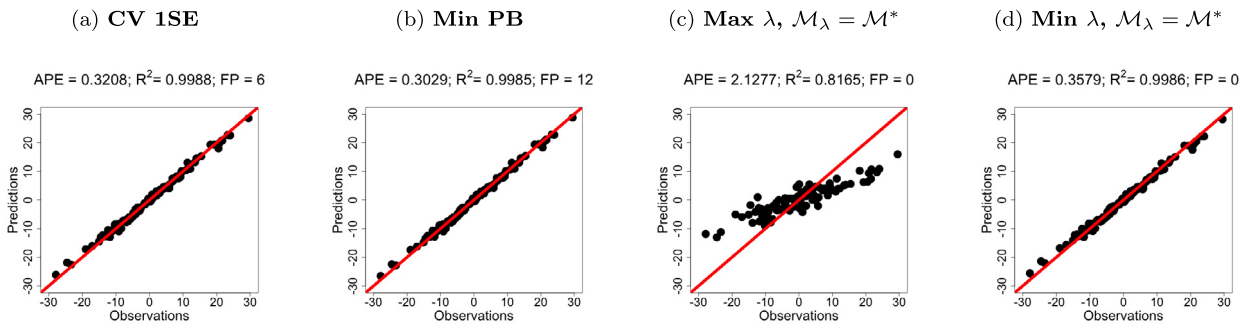

Fig. 1. In-sample observations (y_i) versus predictions ($\hat{y}_i$) for Lasso estimator at four values of λ for a data set with $n = 100$, $p = 100$, and $p^* = 5$. We also report average prediction error (APE) under 10-fold CV, in-sample squared prediction correlation (R^2), and the number of false positives (FP). The minimum APE plus-or-minus one standard error is 0.2844 ± 0.0262 and all estimates have $\mathcal{M}^* \subseteq \mathcal{M}_\lambda$.

For high-dimensional data like that in Stallrich et al. (2020), simultaneous support recovery and parameter estimation can be performed via penalized estimation. A penalized estimator is represented generally by $\hat{\boldsymbol{\beta}}_\lambda = \arg\min L(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta})$ where $L(\boldsymbol{\beta})$ is a loss function comparing $\mathbf{y}$ to its predicted values $\mathbf{X}\boldsymbol{\beta}$, and $P_\lambda(\boldsymbol{\beta})$ is a penalty function that depends on tuning parameter(s) $\lambda \geq 0$. Henceforth we let $L(\boldsymbol{\beta}) = (2n)^{-1}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2$. Penalty functions can take myriad forms but we are interested in those that increase as $\boldsymbol{\beta}$ moves away from $\mathbf{0}$. The Lasso (Tibshirani, 1996) penalty, $P_\lambda(\boldsymbol{\beta}) = \lambda||\boldsymbol{\beta}||_1 = \lambda\sum_{j=1}^{p}|\beta_j|$, is one such penalty that can force estimates to equal 0, thereby performing simultaneous variable selection and estimation. For such sparsity-inducing estimators, we are interested in comparing the estimated support, $\mathcal{M}_\lambda = \{j : \hat{\beta}_{\lambda,j} \neq 0\}$, to $\mathcal{M}^*$.

The chosen λ balances the importance of minimizing $P_\lambda(\boldsymbol{\beta})$ relative to $L(\boldsymbol{\beta})$, so it is recommended to explore the tuning parameter space to identify an "optimal" value. Potential criteria for an optimal value include identifying a $\hat{\boldsymbol{\beta}}_\lambda$ that minimizes $||\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}^*||_2^2$, minimizes $||\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda - \mathbf{X}\boldsymbol{\beta}^*||_2^2$, or has $\mathcal{M}_\lambda = \mathcal{M}^*$. The latter criterion is referred to as support recovery and is the primary focus of this paper. Even if a λ exists where $\mathcal{M}_\lambda = \mathcal{M}^*$, there is no guarantee that we will be able to correctly identify it. Popular approaches, such as minimizing information criteria (Akaike, 1974; Schwarz, 1978) or minimizing average squared prediction error from K-fold cross validation often choose a λ that overselects the number of important variables (Feng and Yu, 2013; Hastie et al., 2017), i.e., $\mathcal{M}^* \subset \mathcal{M}_\lambda$. Post-selection inference techniques (Lockhart et al., 2014; Zhang and Zhang, 2014) and multi-stage modifications (Zou, 2006; Meinshausen, 2007) can correct for this overselection, albeit with added computations and assumptions.

This paper proposes a new K-fold cross validation strategy that assesses the predictive quality of $\hat{\boldsymbol{\beta}}_\lambda$ by maximizing average squared prediction correlation rather than minimizing average squared prediction error. The scale invariance of correlation reduces the impact of the potential shrinkage when estimating large $|\beta_j^*|$ that would burgeon squared prediction error. To demonstrate, we simulated a dataset $\{\mathbf{y}, \mathbf{X}\}$ with $n = p = 100$, $p^* = 5$, and normally-distributed errors with $\boldsymbol{\Sigma} = \mathbf{I}$, and generated the entire solution path of the Lasso. The elements of $\mathbf{X}$ were independently sampled from a standard Normal distribution and then appropriately centered and scaled. The active coefficients in $\boldsymbol{\beta}^*$ were $\{2.13, 1.81, -2.46, -1.89, -2.51\}$. Many λ values had $\mathcal{M}_\lambda = \mathcal{M}^*$, so a wide range were optimal for support recovery. Fig. 1 plots $\mathbf{y}$ against their in-sample predictions, $\hat{\mathbf{y}}_\lambda$, under four different λ values, and summarizes the models' number of false positives (FP), average squared prediction error (APE) from 10-fold cross validation, and in-sample squared prediction correlation (R^2). Fig. 1(a) corresponds to the λ determined by a cross validation one-standard-error rule (CV 1SE) and Fig. 1(b) corresponds to the λ that minimizes $||\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||_2$ (Min PB). Among the λ having $\mathcal{M}_\lambda = \mathcal{M}^*$, we consider the largest and smallest in Figs. 1(c) and 1(d), respectively.

All models shown in Fig. 1 include the 5 important variables. Both the CV 1SE and Min PB models have small APE and large R^2, but many false positives. The model in Fig. 1(c) has no false positives, but has relatively large APE. The model under Fig. 1(d) also has no false positives and its in-sample R^2 approximately equals that of the Min PB estimate. This motivates a tuning parameter selection strategy based on an R^2 metric rather than APE to compromise between prediction error and variable selection.

After justifying the R^2 metric, we highlight and investigate an equivalence between the metric and a multiplicative adjustment on $\hat{\boldsymbol{\beta}}_\lambda$, referred to here as the α-modification. We argue that for $\hat{\boldsymbol{\beta}}_\lambda$ with certain statistical properties, the adjustment can reduce the bias of $\hat{\boldsymbol{\beta}}_\lambda$ thereby improving its predictive potential. We go on to study the α-modification for the Lasso, highlighting its similarities to the Nonnegative Garrote (Breiman, 1995) and relaxed Lasso (Meinshausen, 2007). Unlike these two methods, the α-modification can be applied to any penalized least-squares estimator, including non-convex penalties, without additional computational complexity.

This paper is organized as follows. In Section 2 we review classes of penalized estimators, popular tuning parameter selection strategies, and post-selection inference methods. Section 3 justifies the value in the R^2 metric and provides statistical properties for a general class of penalized estimators. Finite-sample properties are then derived for the α-modification for the Lasso in Section 4. Section 5 presents a simulation study of the new approaches and Section 6 applies our new methods

to the EMG data of Stallrich et al. (2020). Section 7 provides a discussion on the implications of our new framework for evaluating model fit and propose avenues of future research.

2. Background

2.1. Classes of penalized estimators

Consider the class of penalties $P_\lambda(\boldsymbol{\beta}) = ||\boldsymbol{\beta}||_q^q = \sum_j |\beta_j|^q$, $q > 0$, corresponding to the so-called bridge estimators (Frank and Friedman, 1993). It has been shown (Knight and Fu, 2000) that for $q \leq 1$ and large enough λ, the penalized estimate will have some $\hat{\beta}_{\lambda,j} = 0$, yielding a continuous approach to the intractable exploration of all submodels. Under appropriate regularity conditions, the limiting distributions of such $\hat{\beta}_{\lambda,j}$ whose corresponding $\beta_j^* = 0$ can have positive probability mass at 0 when p is fixed, meaning that the estimators are capable of support recovery. The result has been shown to hold as p and n grow to infinity, under certain growth rate conditions (Huang et al., 2008). However, the large λ necessary for this to occur may cause $|\hat{\beta}_{\lambda,j}| << |\beta_j^*|$ for $j \in \mathcal{M}^*$, thereby inflating criteria commonly used for tuning parameter selection. A λ yielding support recovery may then be ignored by popular tuning parameter selection strategies.

To address this shrinkage problem, the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001) and the Minimax Concave penalty (MCP) give continuous, nearly unbiased methods of penalized estimation (Zhang, 2010). The SCAD estimator has been shown to be support recovery consistent (i.e., $P(\mathcal{M}_{\lambda_n} = \mathcal{M}^*) \to 1$ as $n \to \infty$) for appropriately chosen λ (Fan and Li, 2001). MCP shares a similar result but again under certain conditions on λ (Zhang, 2010). These penalties are less likely to have $|\hat{\beta}_{\lambda,j}| << |\beta_j^*|$ but their tuning parameter selection problem is further complicated by having to explore a multidimensional space.

Another approach to prevent $|\hat{\beta}_{\lambda,j}| << |\beta_j^*|$ is to adjust bridge penalties. The Ridge penalty ($q = 2$) cannot shrink any coefficient estimate to exactly zero, so Wu (2021) recently proposed $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p \lambda_j \beta_j^2$, to allow for this behavior. The adaptive Lasso (Zou, 2006) is a similar adjustment but for the Lasso penalty. It follows a two-stage process: first $\hat{\boldsymbol{\beta}}$—a consistent estimator for $\boldsymbol{\beta}^*$ such as Ordinary Least Squares (OLS)—is calculated. Then, with $\gamma > 0$, adaptive Lasso estimates are found using the penalty $P_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^p |\hat{\beta}_j|^{-\gamma} |\beta_j|$. When $\gamma = 1$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$, the objective function reduces to the Nonnegative Garotte (Breiman, 1995). This approach is support recovery consistent when $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$ as $n \to \infty$. The adaptive Lasso is easily generalized to non-Lasso penalties, however the selection of γ and consistent estimation of $\boldsymbol{\beta}^*$ for high dimensional data can be difficult to establish.

The relaxed Lasso (Meinshausen, 2007) minimizes the objective function:

$$\frac{1}{2n}||\mathbf{y} - \mathbf{X}(\boldsymbol{\beta} \circ \mathbf{1}_{\mathcal{M}_\lambda})||_2^2 + \lambda\phi||\boldsymbol{\beta}||_1 \,. \tag{2}$$

where $\phi \in (0, 1]$ and $\boldsymbol{\beta} \circ \mathbf{1}_{\mathcal{M}_\lambda}$ is the Hadamard product of $\boldsymbol{\beta}$ with the support vector under $\mathcal{M}_\lambda$. In simulations, the relaxed Lasso returns sparse models with low bias. Moreover, the expected value of the loss function of the relaxed Lasso converges to 0 faster than the Lasso when p increases quickly relative to n, meaning that relaxed Lasso estimates tend to be closer to $\boldsymbol{\beta}^*$ for smaller n than the traditional Lasso. This result is again achieved assuming that λ is sufficiently large. The relaxed Lasso is computationally efficient to calculate but its extensions to more complicated penalties can be computationally intensive, and to our knowledge have not been well-studied.

The group Lasso (Yuan and Lin, 2006) penalty assumes $E(y_i) = \sum_{j=1}^p \mathbf{x}_{ij}^T \boldsymbol{\beta}_j$ where each $\mathbf{x}_{ij}$ is a $d_j \times 1$ vector corresponding with the ith observation of the jth covariate group and shrinks coefficients at the group level. This class of models includes general additive models (GAMs) and functional linear models. GAMs (Hastie and Tibshirani, 1986) have the form $y_i = \sum_{i=1}^p f_j(\mathbf{x}_{ij}) + \epsilon_i$ where the f_j's are functions of one or more covariates. Each $f(\cdot)$ is commonly approximated by a pre-specified basis expansion, such as B-splines, so that estimation of $f(\cdot)$ is equivalent to estimating the corresponding group of basis coefficients. The functional linear model (Ramsay and Silverman, 2005), an example of which may be found in Section 6, has functional covariates $x_{ij}(t)$ on domain $\mathcal{T}$ and models $y_i = \sum_{i=1}^p \int x_{ij}(t)\beta_j(t)dt + \epsilon_i$. The model is often approximated by imposing a basis expansion of the $\beta_j(\cdot)$. The group Lasso penalty for such models is $P_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p \lambda_j ||\boldsymbol{\beta}_j||_{K_j}$ where $\|\mathbf{z}\|_K = (\mathbf{z}^T\mathbf{K}\mathbf{z})^{\frac{1}{2}}$ and $\mathbf{K}_1, ..., \mathbf{K}_J$ are known positive definite matrices. The group of coefficient estimates, $\hat{\boldsymbol{\beta}}_{\lambda,j}$, are then either all zero or all nonzero. For the linear model, and under some regularity conditions, Nardi and Rinaldo (2008) proved that the group Lasso is support recovery consistent as long as $\sqrt{n}\lambda_j \to \infty$ for all $j \notin \mathcal{M}^*$.

2.2. Tuning parameter selection strategies

Desirable statistical properties of penalized estimators hold under certain λ values, making tuning parameter selection a pivotal step in the analysis. One popular approach is to choose the λ that minimizes an information criterion $IC(\lambda) = -2\log L(\hat{\boldsymbol{\beta}}_\lambda) + h(k_\lambda)$ where $L(\hat{\boldsymbol{\beta}}_\lambda)$ is the likelihood of the data under $\hat{\boldsymbol{\beta}}_\lambda$ and $h(\cdot)$ is a penalty to prevent overselection based on the $k_\lambda = |\mathcal{M}_\lambda|$. Two well-known criteria are AIC (Akaike, 1974), where $h(k_\lambda) = 2k_\lambda$, and BIC (Schwarz, 1978), where $h(k_\lambda) = k_\lambda \log(n)$. AIC often overselects, particularly for small sample sizes, so a corrected AIC (AICc), where $h(k_\lambda) = 2k_\lambda +$

$\frac{2k_\lambda^2+2k_\lambda}{n-k_\lambda-1}$ is recommended (Hurvich and Tsai, 1989). BIC, unlike AIC, is support recovery consistent when $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$ (Nishii, 1984). For Gaussian errors, $L(\hat{\boldsymbol{\beta}}_\lambda)$ involves σ^2 and for unknown σ^2, $-2\log L(\hat{\boldsymbol{\beta}}_\lambda) \propto n\log(\hat{\sigma}^2_\lambda)$ where $\hat{\sigma}^2_\lambda$ is the estimated model variance at λ. This can lead to overselection when $\hat{\sigma}^2_\lambda < 1$ (Buhlmann and van de Geer, 2011). When possible, σ^2 is substituted with $\hat{\sigma}^2$ from a presumed low bias model (Hastie et al., 2017), but this may be challenging to identify for high-dimensional data. The Extended Regularized Information Criterion (ERIC), where $h(k_\lambda) = 2\nu k_\lambda \log(n\hat{\sigma}^2_\lambda/\lambda)$ and $\nu > 0$, was proposed by Hui et al. (2015) specifically for tuning parameter selection of penalized estimators. ERIC outperformed popular tuning parameter selection approaches in their simulations for the adaptive Lasso, but the choice of ν is subjective and determines the balance between fit and sparsity.

Cross validation (CV), is the process of splitting data into training and validation sets, in which the models are fit on the training sets and overfitting is assessed by predicting observations in the validation sets. CV takes many forms, but K-fold CV (Geisser, 1975; Allen, 1974; Stone, 1974) is arguably the most common. In K-fold CV, the data are partitioned into K sets, or folds, of size n_k each. For each λ, K sets of estimates are generated using $K-1$ of the K folds and predictions are generated for the remaining fold, denoted $\hat{\mathbf{y}}_{\lambda,k}$. Prediction error is calculated for each λ and fold as $\frac{1}{n_k}||\mathbf{y}_k - \hat{\mathbf{y}}_{\lambda,k}||_2^2$ and is averaged across the K folds to give the average prediction error (APE) for each λ. The λ with the minimum APE is selected, or a 1SE rule—which chooses the largest λ within one standard error of the minimum APE—is implemented. The use of a 1SE rule is most common for a one-dimensional λ, but Stallrich et al. (2020) proposed a multidimensional extension where the $\hat{\boldsymbol{\beta}}_\lambda$ chosen lies within one standard error of the minimum and also minimizes some penalty function. K-fold CV still has a tendency towards overselection, even with a 1SE rule (Krstajic et al., 2014).

Generalized cross validation (GCV) is an efficient alternative to n-fold Cross Validation (Craven and Wahba, 1978; Golub et al., 1979). GCV is appropriate when the estimation procedure admits linear predictions $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for some matrix $\mathbf{S}$ (Hastie et al., 2017). For example, in Ridge regression, $\mathbf{S} = \mathbf{X}((\mathbf{X}^T\mathbf{X})^{-1} + \lambda\mathbf{I})\mathbf{X}^T$. GCV minimizes a function that divides $L(\hat{\boldsymbol{\beta}}_\lambda) = ||\mathbf{y} - \hat{\mathbf{y}}_\lambda||_2^2$ by a function of k_λ, the effective degrees of freedom. However, just like with other information criteria, GCV has been shown to lead to overselection (Homrighausen and McDonald, 2018).

2.3. Post-selection inference

Post-selection inference techniques carry out further variable selection after a tuning parameter value has been selected (van de Geer et al., 2014; Javanmard and Montanari, 2014; Taylor and Tibshirani, 2015; Lee et al., 2016; Shi et al., 2020). The debiased Lasso estimator (Zhang and Zhang, 2014) is a linear adjustment to the Lasso estimate $\frac{1}{n}\boldsymbol{\Theta}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda)$, where $\boldsymbol{\Theta}$ is an estimate of $(\mathbf{X}^T\mathbf{X})^{-1}$. This causes zero estimates to become nonzero, but the additive adjustment for such estimates tends to be small. The correction leads to an approximately Normal distribution of $\hat{\boldsymbol{\beta}}_\lambda$ so one may perform hypothesis testing and construct confidence intervals.

The Covariance Test (Lockhart et al., 2014) takes advantage of the LARS algorithm, which constructs the Lasso solution path by adding variables one at a time (Efron et al., 2004). It is distinctive in that it assesses model fit using covariance rather than squared error loss. The test requires estimation of σ^2 and its extension to non-Lasso penalties is not well studied, but it provides precedence for the use of correlation in the evaluation of model fit for variable selection.

In general, when the tuning parameter selection event can be written as $\{\mathbf{A}\mathbf{y} \leq \mathbf{b}\}$ for some matrix $\mathbf{A}$ and vector $\mathbf{b}$, there exists a general scheme for post-selection inference that gives exact confidence intervals and p-values for Gaussian errors. Choose $\boldsymbol{\eta}$ such that inference about $\boldsymbol{\eta}^T E[\mathbf{y}]$ is of interest. Using the polyhedral lemma for Gaussian errors, Lee et al. (2016) and Lockhart et al. (2014) represent this event in terms of $\boldsymbol{\eta}^T\mathbf{y}$ to perform conditional inference. This allows for inference upon multiple λ or a fixed λ. When used for successive steps of LARS, it is known as the Spacing Test (Tibshirani et al., 2014) and is a non-asymptotic version of the Covariance Test.

3. Methodology

3.1. Why correlation over squared prediction error?

Distinguish the magnitude of $\boldsymbol{\beta}^*$, denoted $\alpha^* = ||\boldsymbol{\beta}^*||_2$, from its direction, $\boldsymbol{\xi}^* = \boldsymbol{\beta}^*/\alpha^*$. Then $\boldsymbol{\xi}^*$ retains all information about $\mathcal{M}^*$. The same summaries may be computed from an estimate, $\hat{\boldsymbol{\beta}}$, denoted by $\tilde{\alpha}$ and $\tilde{\boldsymbol{\xi}}$, having support $\mathcal{M}$. Comparing $\mathcal{M}$ to $\mathcal{M}^*$ is equivalent to comparing the supports of $\tilde{\boldsymbol{\xi}}$ and $\boldsymbol{\xi}^*$. A tuning parameter selection strategy based on squared prediction error, however, concerns both magnitude and direction. Let $\{\mathbf{y}, \mathbf{X}\}$ denote a holdout sample where $\mathbf{y}$ and $\mathbf{X}$ have been centered. The squared prediction error for $y_i = \mathbf{x}_i^T\boldsymbol{\beta}^* + \epsilon_i$ is

$$\sum_i (y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})^2 = \sum_i (\epsilon_i + \mathbf{x}_i^T(\alpha^*\boldsymbol{\xi}^* - \tilde{\alpha}\tilde{\boldsymbol{\xi}}))^2 \,. \tag{3}$$

In the ideal situation where $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^*$, (3) will be inflated when $\tilde{\alpha} \neq \alpha^*$. Indeed, for penalized estimators with large λ, typically $\tilde{\alpha} < \alpha^*$. Therefore, it is possible for an estimate having $\tilde{\boldsymbol{\xi}} = \boldsymbol{\xi}^*$ to have a large APE and so would be unlikely to be chosen by an APE-based tuning parameter selection strategy.

Consider now the correlation between $\mathbf{y}$ and the predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. After some simplification, we get the expression

$$\text{Corr}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{(\mathbf{X}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}^*)^T \mathbf{X}\tilde{\boldsymbol{\xi}}}{||\mathbf{X}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}^*||_2 \, ||\mathbf{X}\tilde{\boldsymbol{\xi}}||_2}, \tag{4}$$

where $\boldsymbol{\epsilon}^* = \boldsymbol{\epsilon}/\alpha^*$ is a scaled error vector that does not depend on $\hat{\boldsymbol{\beta}}$. The $\tilde{\alpha}$ has no influence over this summary so this measure better compares the $\boldsymbol{\xi}^*$ and $\tilde{\boldsymbol{\xi}}$, and hence better assesses support recovery than squared prediction error.

Our proposed tuning parameter selection strategy, called AR2 CV, employs K-fold CV with folds $\{\mathbf{y}_k, \mathbf{X}_k\}$ but replaces APE with

$$\text{AR2} = \frac{1}{K}\sum_{k=1}^{K}[1 - \text{Corr}(\mathbf{y}_k, \hat{\mathbf{y}}_k)^2]. \tag{5}$$

The optimal λ may be chosen as the one that minimizes AR2, but we have found significant improvements in support recovery under an analogous 1SE rule. Applying AR2 CV with a 1SE rule to the toy example in Section 1, the optimal λ is that given in Fig. 1(d). The AR2 CV estimator was then able to compromise between support recovery and prediction error, while the APE-based CV prioritized prediction error. The coefficient estimates for the $j \in \mathcal{M}^*$ under the APE CV model exhibit less bias than those of the AR2 CV model. A potential drawback then of AR2 CV is that its indifference towards α^* may lead to a $\hat{\boldsymbol{\beta}}_\lambda$ that exhibits more shrinkage than is desired. A potential remedy is to follow selection of $\mathcal{M}_\lambda$ with unpenalized estimation for only predictors in $\mathcal{M}_\lambda$. For example, one could perform OLS on only the $j \in \mathcal{M}_\lambda$ to form the estimator $\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$ where $\hat{\beta}_{OLS,j}^{\mathcal{M}_\lambda} = \hat{\beta}_{OLS,j}$ when $j \in \mathcal{M}_\lambda$ and zero otherwise. We next discuss an alternative strategy that is related to AR2 CV that adjusts the shrinkage of $\hat{\boldsymbol{\beta}}_\lambda$.

3.2. The α-modification

Consider now the training data's $\mathbf{y}$ and their predictions, $\hat{\mathbf{y}}$. If $\hat{\mathbf{y}} \neq \mathbf{0}$, calculate the least-squares estimate $\hat{\alpha} = \arg\min_\alpha ||\mathbf{y} - \alpha\hat{\mathbf{y}}||_2^2 = (\hat{\mathbf{y}}^T\hat{\mathbf{y}})^{-1}\hat{\mathbf{y}}^T\mathbf{y}$. Note this $\hat{\alpha}$ likely differs from $\tilde{\alpha} = ||\hat{\boldsymbol{\beta}}||_2$. By definition, the modified predictions $\hat{\alpha}\hat{\mathbf{y}}$ will be as close or closer to $\mathbf{y}$ as $\hat{\mathbf{y}}$. The modified prediction is also equivalent to prediction under the adjusted penalized estimate, $\hat{\alpha}\hat{\boldsymbol{\beta}}$. Calculating these α-modified coefficient estimates and predictions is described in Algorithm 1.

Algorithm 1 α-Modification for the Linear Model.

1: **Given:** Suppose $\mathbf{y}$ and $\mathbf{X}$ are centered such that an intercept term is unnecessary. Let $\hat{\boldsymbol{\beta}}$ be the vector of coefficient estimates.
2: Calculate $\hat{\alpha} = (\hat{\mathbf{y}}^T\hat{\mathbf{y}})^{-1}\hat{\mathbf{y}}^T\mathbf{y} = (\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}})^{-1}\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{y}$.
3: Calculate α-modified estimates, $\hat{\alpha}\hat{\boldsymbol{\beta}}$, and α-modified predictions, $\hat{\alpha}\hat{\mathbf{y}} = \hat{\alpha}\mathbf{X}\hat{\boldsymbol{\beta}}$.

For penalized estimators, the α-modified estimate can be viewed as

$$\arg\min_{\alpha,\boldsymbol{\xi}} \frac{1}{2n}||\mathbf{y} - \alpha\mathbf{X}\boldsymbol{\xi}||_2^2 + \lambda P(\boldsymbol{\xi}), \tag{6}$$

by first fixing $\alpha = 1$ to get $\hat{\boldsymbol{\xi}} = \hat{\boldsymbol{\beta}}_\lambda$ and then minimizing the function for α given $\hat{\boldsymbol{\xi}}$. This is a slight abuse of notation because $\hat{\boldsymbol{\xi}}$ is not required to be a unit vector like $\boldsymbol{\xi}^*$, but it does help to show how the α-modified estimate is related to separating $\boldsymbol{\beta}^*$ into its magnitude and direction. As $\lambda \to 0$, $\hat{\alpha}_\lambda \to 1$ because the objective function focuses most of its attention on the loss function. Thus the impact of the α-modification will be more pronounced for larger values of λ, and ideally the α-modified estimate will correct the bias of $\hat{\boldsymbol{\beta}}_\lambda$ due to shrinkage.

The α-modification is similar to existing modifications to penalized estimators. First, one can view the modification as reversing the process of calculating the Nonnegative Garrote estimator, which starts with OLS estimates of $\boldsymbol{\beta}^*$ and then performs penalization. Zou and Hastie (2005) also recommended a multiplicative adjustment to the Elastic Net estimator, although the adjustment only involves one of the tuning parameters. For penalties that satisfy $P(\alpha\boldsymbol{\xi}) = \alpha P(\boldsymbol{\xi})$ for $\alpha > 0$, we may rewrite the penalty in (6) as $\lambda\alpha^{-1}P(\alpha\boldsymbol{\xi})$ which resembles the relaxed Lasso penalty, so long as $\alpha^{-1} \in (0, 1]$. Finally, the debiased Lasso tries to reduce bias through an additive adjustment, but this will cause some or all $\hat{\beta}_{\lambda,j} = 0$ to become nonzero, while a multiplicative adjustment will not change the support of $\hat{\boldsymbol{\beta}}_\lambda$.

Penalized estimates are typically shrunk towards zero so the α-modification will correct this type of bias only if $\hat{\alpha}_\lambda \geq 1$. This property is guaranteed for common penalties:

Theorem 1. *Suppose $P_\lambda(\boldsymbol{\beta}) = \sum_{\ell=1}^{L}\lambda_\ell g_\ell(\boldsymbol{\beta})$ where $g_\ell(\boldsymbol{\beta})$ is convex and minimized at $\mathbf{0}_p$. Then $\hat{\alpha}_\lambda \geq 1$ when $\hat{\boldsymbol{\beta}}_\lambda \neq \mathbf{0}$.*

Amplifying penalized estimates does not necessarily decrease bias. To evaluate the α-modification as a bias-reduction tool, we have the following result.

Lemma 1. *If there exists a λ where $P(\hat{\boldsymbol{\xi}}_\lambda = \boldsymbol{\xi}^*) = 1$, then $E(\hat{\alpha}_\lambda \hat{\boldsymbol{\beta}}_\lambda) = \boldsymbol{\beta}^*$.*

Lemma 1 conditions on an event that may have probability 0. The following lemma considers a broader condition, whereby the penalized estimate recovers the direction of $\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$, defined at the end of Section 3.1.

Lemma 2. *If $\hat{\boldsymbol{\beta}}_\lambda \propto \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$ then $\hat{\alpha}_\lambda \hat{\boldsymbol{\beta}}_\lambda = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$.*

There are multiple examples that satisfy the condition of Lemma 2. The OLS estimator itself qualifies as a scaled OLS estimator, where $\hat{\alpha}_\lambda = 1$. Ridge estimates when $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$ also take this form, having $\hat{\boldsymbol{\beta}}_\lambda = \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}_{OLS}$. Lemma 2 also applies whenever $\hat{\boldsymbol{\beta}}_\lambda$ contains exactly one non-zero entry. Finally, note that the relaxed Lasso always includes $\hat{\boldsymbol{\beta}}_{\lambda,\phi} = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$ among its solutions by setting $\phi = 0$. Lemma 2 shows that this can sometimes occur for the α-modified estimates as well.

The α-modification serves to improve predictions under a given $\hat{\boldsymbol{\beta}}_\lambda$ through a positive, multiplicative adjustment and so its ability to improve estimation depends on the properties of $\hat{\boldsymbol{\beta}}_\lambda$. While this paper is mainly concerned with tuning parameter selection for support recovery under finite sample sizes, properties of $\hat{\boldsymbol{\beta}}_\lambda$ are easier to study as $n \to \infty$ and the same is true for α-modified estimators. Concerning support recovery consistency, since the support of $\hat{\alpha}_\lambda \hat{\boldsymbol{\beta}}_\lambda$ equals the support of $\hat{\boldsymbol{\beta}}_\lambda$, the α-modified estimator is support recovery consistent if and only if $\hat{\boldsymbol{\beta}}_\lambda$ is support recovery consistent. Next, the following theorem establishes estimation consistency of α-modified estimators.

Theorem 2. *Fix p and suppose there exists a positive definite matrix $\boldsymbol{C}$ where $\frac{1}{n}\boldsymbol{X}_n^T\boldsymbol{X}_n = \frac{1}{n}\boldsymbol{C}_n \to \boldsymbol{C}$ as $n \to \infty$. For a $P_\lambda(\cdot)$, if there exists a λ_n where $\hat{\boldsymbol{\beta}}_{\lambda_n}$ converges in probability to $c\boldsymbol{\xi}^*$ for some $c > 0$, then $\hat{\alpha}_{\lambda_n}\hat{\boldsymbol{\beta}}_{\lambda_n}$ converges in probability to $\boldsymbol{\beta}^*$.*

The case of $c = \alpha^*$ in Theorem 2 says that if $\hat{\boldsymbol{\beta}}_\lambda$ converges in probability to $\boldsymbol{\beta}^*$ then so will its corresponding α-modified estimator. For example, Knight and Fu (2000) determined certain conditions for which the Lasso exhibits this property. However, the rate of convergence for the α-modified estimators may improve due to the relaxation of finding a λ_n where $\hat{\boldsymbol{\beta}}_{\lambda_n}$ converges in probability to any $c\boldsymbol{\xi}^*$. Theorem 2 also suggests there may be opportunities for other penalized estimators that are not estimation consistent to have an α-modified estimator that is estimation consistent.

3.3. α-Modified cross validation

In addition to AR2 CV, we propose α-modified CV based on average squared prediction error under the α-modified estimates:

$$\text{Mod APE} = \frac{1}{K}\sum_{k=1}^{K} ||\mathbf{y}_k - \hat{\alpha}_k\hat{\mathbf{y}}_k||_2^2 \,, \tag{7}$$

where $\hat{\alpha}_k$ is calculated from the training data. Returning to the toy example from Section 1, the λ chosen through α-modified CV with a 1SE rule returned a model with $\mathcal{M}_\lambda = \mathcal{M}^*$. Its APE was 0.3821, which is marginally larger than that of AR2 CV. This demonstrates the possibility that the two CV strategies may point to different optimal λ.

Another benefit of α-modified CV over traditional CV surprisingly derives from the potential drawbacks of the modification. We say that $\hat{\boldsymbol{\beta}}_\lambda$ recovers the direction of $\boldsymbol{\beta}^*$ when $\hat{\boldsymbol{\beta}}_\lambda / \|\hat{\boldsymbol{\beta}}_\lambda\|_2 = \boldsymbol{\xi}^*$. Because the modification unshrinks the $\hat{\boldsymbol{\beta}}_\lambda$ without changing its direction, the modification may exacerbate an estimate of poor quality, causing the Mod APE to exceed APE based on $\hat{\boldsymbol{\beta}}_\lambda$. It is unlikely then for such a λ to be selected as the optimal value. Similarly, when $\hat{\boldsymbol{\beta}}_\lambda$ recovers $\boldsymbol{\xi}^*$ but has small magnitude, Mod APE will decrease significantly over the traditional APE. Theorem 3 gives a theoretical justification for the latter situation.

Theorem 3. *Suppose that $\hat{\boldsymbol{\beta}}_\lambda$ recovers the direction of $\boldsymbol{\beta}^*$ and assume all ϵ_i are independent with constant variance σ^2. Then the expected value of the α-modified APE is less than the expected value of the traditional APE when:*

$$\frac{\alpha^{*2}}{E[(\|\hat{\boldsymbol{\beta}}_\lambda\|_2 - \alpha^*)^2]} \leq \frac{\boldsymbol{\beta}^{*T}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}^*}{\sigma^2} \,. \tag{8}$$

Mod APE is then expected to be smaller than the traditional APE as long as the signal-to-noise ratio, represented by $\boldsymbol{\beta}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}^*/\sigma^2$, is sufficiently large.

4. The α-modified Lasso

The methods in Section 3 generalize to many types of penalties, but to better understand their properties we must focus on a specific penalty. Due to its popularity, we explore the properties for the Lasso penalty and finite n. First, note

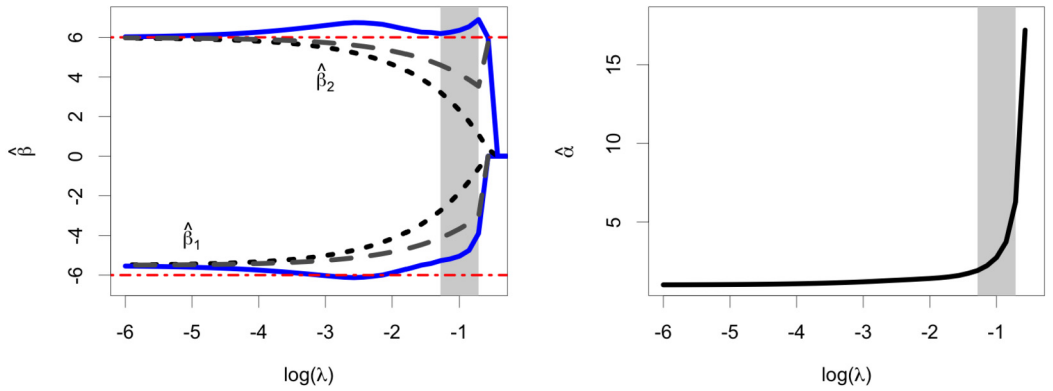

Fig. 2. The leftmost plot shows solution path for the Lasso (dotted line), α-modified Lasso (solid line), and the relaxed Lasso at $\phi = 0.5$ (dashed line) for β_1^* and β_2^*. The gray shaded area indicates the values of λ for which support recovery has occurred. The horizontal, dashed-dotted lines show the true values of β_1^* and β_2^*, -6 and 6 respectively. The plot on the right shows the $\hat{\alpha}_\lambda$ values.

that the α-modified Lasso may be thought of as a reverse Nonnegative Garotte in that it starts with shrunken estimates and uses a least squares modification to readjust and "un-shrink" them. The use of correlation to assess model quality is also consistent with the premise of the Covariance Test. The α-modified Lasso is most similar to the relaxed Lasso. The α-modification penalty is $\lambda\alpha^{-1}||\alpha\boldsymbol{\xi}||_1$ and the relaxed Lasso penalty is $\lambda\phi||\boldsymbol{\beta}||_1$ for $\phi \in (0, 1]$. Theorem 1 establishes $\alpha^{-1} \in (0, 1]$ but we calculate the minimum α directly while the relaxed Lasso treats ϕ as a tuning parameter. Our approach reduces computation, but the resulting estimators are less flexible than the relaxed Lasso. Specifically, the α-modified estimate cannot change the direction of $\hat{\boldsymbol{\beta}}_\lambda$.

Fig. 2 illustrates the distinction between the α-modified and relaxed Lasso estimator for an orthogonal $\mathbf{X}$ with $n = 100$, $p = 50$ and $p^* = 2$ where $(\beta_1^*, \beta_2^*) = (-6, 6)$. The α-modified Lasso and relaxed Lasso solutions were generated for a set of 100 λ and, for the relaxed Lasso, 20 ϕ values. The solution paths for $\hat{\beta}_{\lambda,1}$ and $\hat{\beta}_{\lambda,2}$ across all values of λ are shown for both the traditional and α-modified Lasso. The relaxed Lasso estimates for $\phi = 0.5$ are also plotted. Fig. 2 also includes a plot of the $\hat{\alpha}_\lambda$ for the full data. For the λ exhibiting perfect support recovery, we see that $\hat{\alpha}_\lambda > 1$, thereby demonstrating the value of the modification.

We now derive properties of the α-modified estimator for orthogonal $\mathbf{X}$. The Lasso estimator for such $\mathbf{X}$ is $\hat{\beta}_{\lambda,j} = \text{sign}(\hat{\beta}_{OLS,j})(|\hat{\beta}_{OLS,j}| - \lambda)_+$, where $\hat{\beta}_{OLS,j}$ is the OLS estimate of the β_j^* and $(z)_+ = \max(0, z)$. If the Lasso estimate for a j^* is nonzero, the closed form expression for $\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*}$ may be derived.

Lemma 3. *When $\boldsymbol{X}$ is orthogonal, the α-modified Lasso estimator has elements:*

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = w_1\hat{\beta}_{OLS,j^*} + (1 - w_1)\hat{\beta}_{\lambda,j^*} + w_2\hat{\beta}_{\lambda,j^*}\,, \tag{9}$$

where $w_1 = \frac{d_{j^*}^2}{\sum_{j=1}^p d_j^2}$, $w_2 = \frac{\lambda\sum_{j\neq j^*} d_j}{\sum_{j=1}^p d_j^2}$, *and* $d_j = (|\hat{\beta}_{OLS,j}| - \lambda)_+$.

If $d_{j^*} = 0$, then $\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = \hat{\beta}_{\lambda,j^*} = 0$. If $\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j} = 0$ for all $j \neq j^*$ and $d_{j^*} > 0$ then $\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = \hat{\beta}_{OLS,j}$ which is consistent with Lemma 2. When $|\mathcal{M}_\lambda| > 1$, (9) involves a convex combination of the OLS and Lasso estimates, as well as an additive term, $w_2\hat{\beta}_{j^*,\lambda}$, that may overcorrect the α-modified Lasso beyond the OLS estimate. This behavior is illustrated for a simple example in Fig. 3. In that example, for $\lambda < 10$ there are potential values of $\hat{\beta}_{OLS,j^*}$ where the α-modified Lasso exceeds $\hat{\beta}_{OLS,j^*}$. To better understand this behavior, Theorem 4 provides an upper bound for $|\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*}|$.

Theorem 4. *Suppose $\boldsymbol{X}$ is orthogonal and consider a given predictor, j^*, and λ where at least one $j \neq j^*$ has $|\hat{\beta}_{OLS,j}| > \lambda$. Let $\boldsymbol{x}_{j^*}$ denote column j^* of $\boldsymbol{X}$. Then $\hat{\beta}_{OLS,j^*} = \beta_{j^*}^* + n^{-1}\boldsymbol{x}_j^T\boldsymbol{\epsilon}$ where $n^{-1}\boldsymbol{x}_j^T\boldsymbol{\epsilon}$ is fixed implies*

$$|\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*}| \leq \lambda \times \max\left(1, \frac{\sqrt{u^2v + v^2} - v}{2v}\right)\,, \tag{10}$$

where $u = \sum_{j\neq j^*} d_j$ and $v = \sum_{j\neq j^*} d_j^2$. Moreover, $\lim_{|\beta_{j^*}^*|\to\infty} |\hat{\alpha}_\lambda\hat{\beta}_{j^*,\lambda} - \hat{\beta}_{OLS,j^*}| \to 0$.

The result of Theorem 4 is demonstrated in Fig. 3 for different values of λ. As $|\beta_1^*|$ increases in magnitude, thereby increasing $\hat{\beta}_{OLS,1}$, the α-modified Lasso estimate moves away from the traditional Lasso estimate and towards the $\hat{\beta}_{OLS,1}$.

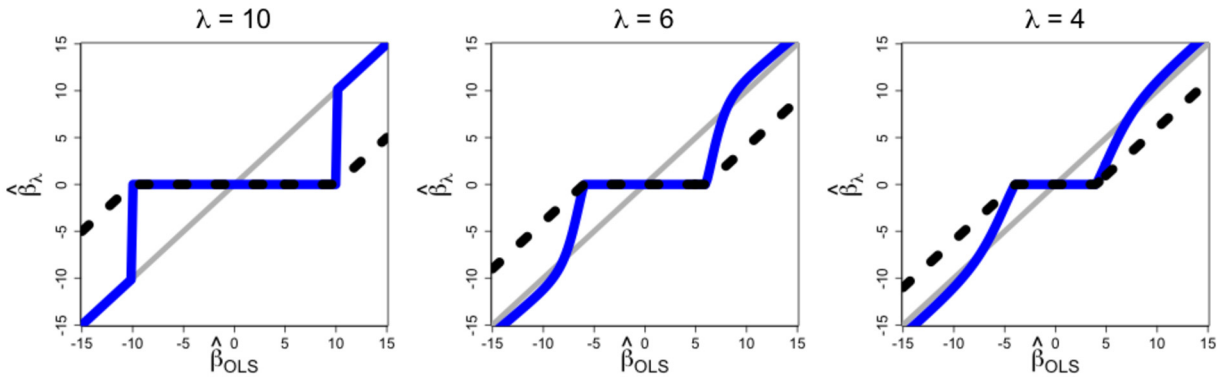

Fig. 3. The α-modified estimates of β_1^* under an orthogonal $\mathbf{X}$ as $\hat{\beta}_{OLS,1}$ changes, with all other nonzero OLS estimates fixed to $\hat{\beta}_{OLS,j} = (-8, 5, -3, 1)$ for $j = (2, 3, 4, 5)$. The dashed line gives the Lasso estimates and the dark solid line gives the α-modified estimates. The light solid line references when the penalized estimate equals the OLS estimate.

This happens instantaneously for $\lambda = 10$ and $|\hat{\beta}_{OLS,1}| > 10$ because $\hat{\beta}_{OLS,1}$ is the only estimate that exceeds λ. For $\lambda = 6$ and 4, when $(\sqrt{u^2v + v^2} - v)/2v > 1$ there exists some β_1^* in which the α-modified Lasso estimate overshoots $\hat{\beta}_{OLS,1}$. This occurs as u^2/v increases, meaning multiple d_j are nonzero and are close to 0.

To generalize Theorem 4 to an arbitrary $\mathbf{X}$ we condition on the event that the λ recovers the sign vector of $\boldsymbol{\beta}^*$ and consider the oracle OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} = (\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})^{-1}\mathbf{X}_{\mathcal{M}^*}^T\mathbf{y}$ where $\mathbf{X}_{\mathcal{M}^*}$ is the subset of columns of $\mathbf{X}$ corresponding to $\mathcal{M}^*$. Then $\hat{\boldsymbol{\beta}}_{OLS,j^*}^{\mathcal{M}^*} = \beta_{j^*}^* + (\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})_{j^*}^{-1}\mathbf{X}_{\mathcal{M}^*}^T\boldsymbol{\epsilon}$ where $(\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})_{j^*}^{-1}$ denotes the j-th row of $(\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})^{-1}$. Finally, let $\tilde{s}_{j^*} = (\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})_{j^*}^{-1}\mathbf{s}$ where $\mathbf{s}$ is the $|\mathcal{M}^*| \times 1$ vector of signs for the β_j^* in $\mathcal{M}^*$.

Theorem 5. *Suppose the Lasso estimate recovers the correct sign vector of $\boldsymbol{\beta}^*$. Let $G_{j^*} = \lambda(s_{j^*} - n\tilde{s}_{j^*})$. Then for a fixed $\boldsymbol{\epsilon}$, $\lim_{|\beta_{j^*}^*|\to\infty}|\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*}^{\mathcal{M}^*}| = |G_{j^*}|$. Moreover $|G_{j^*}| < |\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*}^{\mathcal{M}^*}|$ if and only if*

$$\left|1 - \frac{s_{j^*}}{n\tilde{s}_{j^*}}\right| < 1\,. \tag{11}$$

Note for an orthogonal design $n\tilde{s}_{j^*} = s_{j^*}$, making $G_{j^*} = 0$. Theorem 5 shows that the absolute difference between the α-modified Lasso estimate and the OLS estimate generally does not approach 0 as $|\beta_{j^*}^*| \to \infty$. Rather, it approaches a constant that is always smaller than $|\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*}^{\mathcal{M}^*}|$ so long as $\tilde{s}_{j^*}$ is not close to 0. This indicates the α-modified Lasso estimate generally improves the bias of $\hat{\beta}_{\lambda,j^*}$ for large $\beta_{j^*}^*$ when the sign is recovered.

5. Numerical results

5.1. α-Modified CV under poor bias adjustment

Theorems 4 and 5 show that the α-modification cannot guarantee bias reduction for all λ. However, a poor bias adjustment may be detected by CV. To illustrate, we performed a simulation study for the $\mathbf{X}$ used in Fig. 1 and considered two $\boldsymbol{\beta}^*$ vectors, displayed in Fig. 4. For each $\boldsymbol{\beta}^*$, we generated 500 $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})$ and averaged the three 10-fold CV metrics (APE, AR2, and Mod APE) for a range of λ. We also recorded the minimum and maximum λ values for which support recovery was achieved. The results are presented in Fig. 4. The gray shaded area was constructed by averaging the minimum and maximum λ achieving support recovery across all 500 data sets. Hence, the left of the gray area tended to lead to overselection of predictors (false positives) and values of λ to the right of the gray area tended to underselect predictors (false negatives).

As shown in Fig. 4(a), the Mod APE was less than or equal to APE for all λ, and especially so for larger λ that recover $\mathcal{M}^*$. The optimal λ's according to the minimum AR2 and Mod APE resulted in sparser models than the minimum APE. With a 1SE rule, traditional CV recovered the support in only 24.4% of the replications, whereas α-modified and AR2 CV did so in 91.6% and 93.6% of the replications, respectively.

Fig. 4(b) corresponds to $\boldsymbol{\beta}^*$ with larger and equal magnitude coefficients. Generally, the Mod APE was equal to or smaller than regular APE except for $\log(\lambda) \approx 4$, which had only one nonzero coefficient estimate. Mod APE highlights the poor bias adjustment and so would not recommend choosing this λ. The support recovery percentages were 19.8% for APE CV, 98.2% for α-modified CV, and 98.8% for AR2 CV. Surprisingly, the support recovery percentage decreased for APE CV despite increasing magnitude of $\boldsymbol{\beta}^*$.

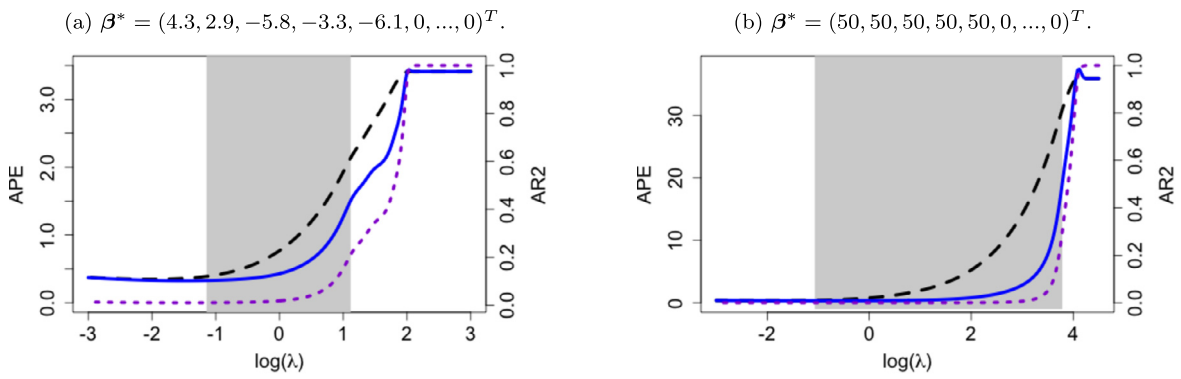

Fig. 4. Average 10-fold Cross Validation results across 500 replications of linear data. **X** was taken from the example in Section 1. Average values of the minimum and maximum λ achieving support recovery are represented by the gray shaded region, the dashed line gives the APE, the solid line gives α-modified APE, and the dotted line gives AR2. Standard errors from these simulations were too small to be depicted.

Table 1
Average Hamming Distance for 100 replications of CV 1SE. Standard errors given in parentheses.

p	50	100	200	400	800	50	100	200	400	800
n	$p^*=5$; $SNR=1.25$; $\bar{\sigma}=5.19$					$p^*=5$; $SNR=5$; $\bar{\sigma}=2.59$				
100	3.1 (0.3)	4.3 (0.4)	6 (0.6)	7.8 (1.1)	9.1 (1.2)	2.2 (0.2)	4 (0.4)	7.5 (1.2)	9.3 (1)	10.7 (1.2)
500	0.6 (0.1)	1 (0.2)	1.3 (0.2)	1.9 (0.5)	2.9 (0.5)	0.9 (0.2)	0.9 (0.1)	1.7 (0.3)	1.7 (0.3)	1.5 (0.3)
1000	0.2 (0)	0.4 (0.1)	0.4 (0.1)	0.6 (0.1)	0.9 (0.2)	0.1 (0)	0.4 (0.1)	0.5 (0.1)	0.7 (0.1)	0.9 (0.2)
n	$p^*=50$; $SNR=1.25$; $\bar{\sigma}=16.6$					$p^*=50$; $SNR=5$; $\bar{\sigma}=8.3$				
100	31.7 (1.3)	43.5 (0.6)	50.1 (0.5)	51.1 (0.4)	53.2 (1.1)	5 (0.4)	32 (0.6)	47 (0.7)	53.8 (1.1)	54.5 (1.1)
500	1.6 (0.2)	18.2 (0.5)	33.3 (0.8)	48.6 (1.3)	57.4 (1.8)	0 (0)	16.8 (0.5)	35.3 (0.9)	58.4 (1.6)	81.2 (2)
1000	0.4 (0.1)	14.5 (0.4)	28.5 (0.9)	43.5 (1.4)	57.6 (1.6)	0 (0)	14.2 (0.5)	28 (0.8)	44.8 (1.3)	64.2 (1.8)

5.2. Support recovery simulation study

To evaluate the new methods for the Lasso, a broader simulation study consisting of 100 replications of data from model (1) was conducted. Following Meinshausen (2007), the rows of **X** were drawn from a multivariate Normal distribution with mean **0** and covariance $\mathbf{\Sigma}_X$. Nonzero elements of $\boldsymbol{\beta}^*$ were drawn from a Gamma distribution with shape 10 and scale 0.25 with negative and positive signs randomly assigned with equal probability, updating the coefficient vector with every replication. For each **X** and $\boldsymbol{\beta}^*$, two independent error vectors were generated from $N(0, \sigma^2)$ distributions, with σ determined by a fixed signal-to-noise ratio $\text{SNR} = \boldsymbol{\beta}^{*T}\mathbf{\Sigma}_X\boldsymbol{\beta}^*/\sigma^2$. We considered $n \in \{100, 500, 1000\}$, $p \in \{50, 100, 200, 400, 800\}$, SNR $\in \{1.25, 5\}$ and $p^* = \{5, 6, \ldots, 50\}$.

Each simulated data set was analyzed using 10-fold CV with and without a 1SE rule under traditional APE (CV Min, CV 1SE), AR2 (AR2 Min, AR2 1SE), Mod APE (Mod Min, Mod 1SE), and the relaxed Lasso. For all methods, we considered 250 λ values equally spaced on the exponential scale from e^{-20} to e^{10}. For the relaxed Lasso, we also considered 100 ϕ values equally spaced from e^{-10} to 1. Selection under the relaxed Lasso was done using minimum APE as well as a 1SE rule based on Stallrich et al. (2020), wherein the optimal combination of λ and ϕ is the model with the smallest $||\hat{\boldsymbol{\beta}}_{\lambda,\phi}||_1$ among all models with an APE within one standard error of the minimum. Relaxed Lasso estimates were found using a coordinate descent algorithm, capable of admitting more than $n-1$ covariates into models, unlike LARS which was used by Meinshausen (2007). Support recovery was evaluated using the Hamming Distance (HD) between the 0/1 support vectors of $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}_\lambda$, which is the sum of the number of false negatives and false positives. Additional metrics including the false discovery rate (FDR), average number of false positives and negatives, and the average prediction bias, where prediction bias is $||\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda||_2$ or $||\mathbf{X}\boldsymbol{\beta}^* - \mathbf{X}\hat{\alpha}_\lambda\hat{\boldsymbol{\beta}}_\lambda||_2$ for the α-modification, can be found in the Supplementary Materials.

We first considered the case of $\mathbf{\Sigma}_X = \mathbf{I}_p$. Tables 1, 2, and 3 give the average HD between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}_\lambda$ for the traditional APE method with a 1SE Rule, AR2 with a 1SE rule, and the relaxed Lasso with a minimum APE approach. AR2 and Mod APE performed similarly; the latter results may be found in the Supplementary Materials. In general, AR2 CV had better variable selection than APE CV and, for small n and large p^*, the relaxed Lasso.

Results from increasing values of p^* were considered for $n = \{100, 500, 1000\}$ and $p = 100$ in Fig. 5. Additional methods and results may be found in the Supplementary Materials. For $n = 100$ all methods performed comparably, except for the relaxed Lasso with a minimum APE rule, which had a higher number of false positives than the other methods, leading to larger average HD. As n increased, Mod APE and AR2 CV performed comparably to the relaxed Lasso with a 1SE Rule, and had a consistently smaller average HD than CV 1SE. The new methods of CV performed similarly to Relaxed 1SE and outperformed CV 1SE in terms of support recovery.

Table 2
Average Hamming Distance for 100 replications of AR2 CV 1SE. Standard errors given in parentheses.

p	50	100	200	400	800	50	100	200	400	800
n	$p^*=5; SNR=1.25; \bar{\sigma}=5.19$					$p^*=5; SNR=5; \bar{\sigma}=2.59$				
100	2.1 (0.3)	2.8 (0.3)	3.3 (0.3)	4.9 (0.7)	5 (0.5)	0.6 (0.1)	0.6 (0.1)	1.9 (0.8)	1.8 (0.4)	2.6 (0.4)
500	0.1 (0)	0.2 (0)	0.1 (0)	0.2 (0)	0.4 (0.1)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
1000	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
n	$p^*=50; SNR=1.25; \bar{\sigma}=16.6$					$p^*=50; SNR=5; \bar{\sigma}=8.3$				
100	23.6 (1.3)	40.4 (0.5)	51.4 (0.8)	55.1 (1.2)	54.8 (1.2)	4.2 (0.4)	31.9 (0.5)	47.9 (0.8)	54.6 (1.3)	53.5 (0.9)
500	1.8 (0.2)	17.1 (0.5)	28.7 (0.8)	42.3 (1.3)	51.9 (2.2)	0.2 (0)	9.1 (0.4)	17.8 (0.7)	30 (1)	46.6 (1.3)
1000	0.7 (0.1)	9.1 (0.4)	15.4 (0.6)	23.1 (0.9)	32.5 (1.1)	0 (0)	3.9 (0.3)	6.3 (0.4)	11.1 (0.6)	16.9 (0.8)

Table 3
Average Hamming Distance for 100 replications of Relaxed Lasso. Standard errors given in parentheses.

p	50	100	200	400	800	50	100	200	400	800
n	$p^*=5; SNR=1.25; \bar{\sigma}=5.19$					$p^*=5; SNR=5; \bar{\sigma}=2.59$				
100	3.7 (0.4)	5.4 (0.8)	7.4 (1.1)	18.4 (3.4)	28.1 (4.2)	1.1 (0.3)	1.8 (0.5)	4.5 (1.7)	6.6 (2.3)	7.9 (2.6)
500	0.8 (0.3)	1.2 (0.3)	0.5 (0.2)	0.7 (0.4)	1.5 (0.6)	1.2 (0.3)	0.8 (0.2)	0.8 (0.4)	0.3 (0.1)	0.1 (0)
1000	0.6 (0.2)	0.5 (0.2)	0.7 (0.3)	0.4 (0.1)	1.9 (1.5)	0.6 (0.2)	0.8 (0.3)	0.4 (0.2)	0.4 (0.2)	0.4 (0.2)
n	$p^*=50; SNR=1.25; \bar{\sigma}=16.6$					$p^*=50; SNR=5; \bar{\sigma}=8.3$				
100	19.5 (1.3)	39.9 (0.5)	57.6 (1.7)	70.8 (3.1)	67.9 (3.2)	2.1 (0.3)	34.1 (0.5)	75.9 (2.5)	87.6 (3.3)	93.6 (3.8)
500	0.9 (0.1)	22.9 (0.9)	39.6 (1.8)	66.2 (3.4)	74.5 (4.2)	0.1 (0)	7.2 (0.5)	14.8 (1.2)	25.4 (2.3)	36.8 (2.7)
1000	0.2 (0)	12.3 (0.9)	17.5 (1.4)	27.9 (2.7)	32.7 (2.9)	0 (0)	2.4 (0.3)	3.1 (0.3)	4.9 (0.5)	5.5 (0.3)

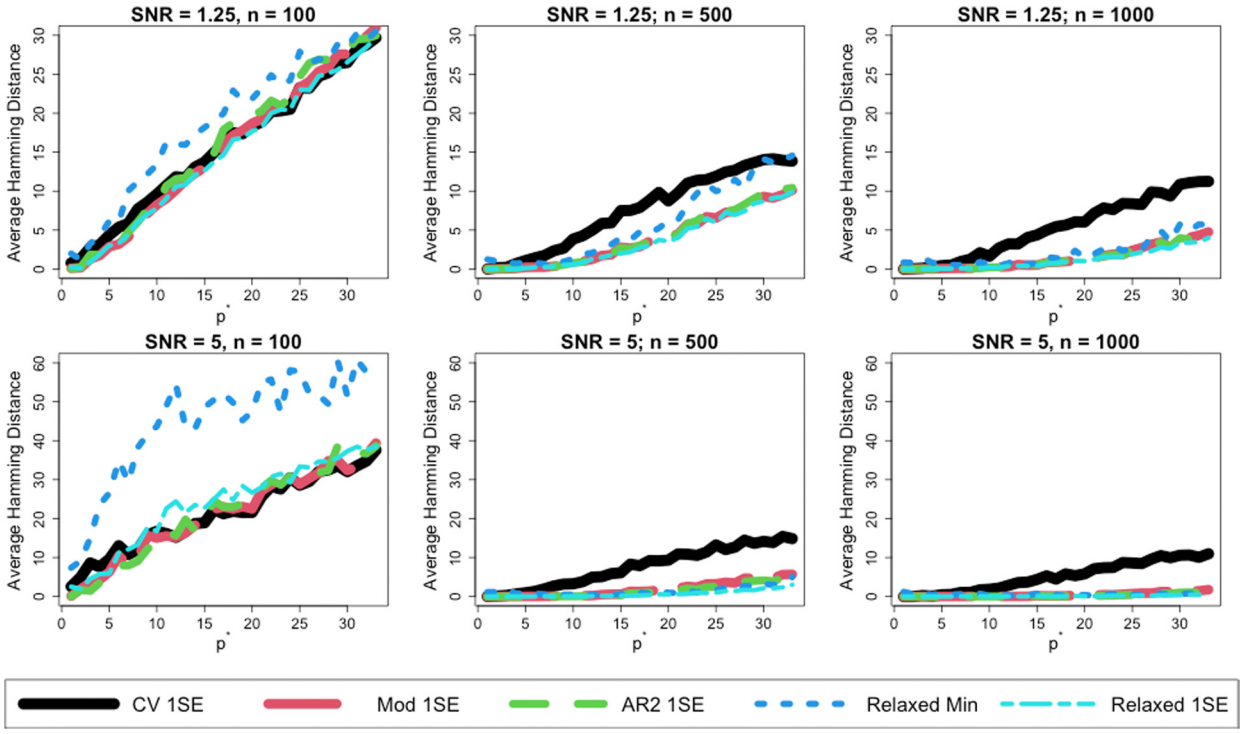

Fig. 5. Average Hamming Distance between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}_\lambda$ from 100 replications of simulated data with $n=\{100, 500, 1000\}$ and $p=100$ with independent predictors for both SNR $=1.25$ and SNR $=5$. Standard errors were too small to be plotted but can be seen in the Supplementary Material.

We next considered correlated predictors where $\boldsymbol{\Sigma}_X$ satisfied $\Sigma_{ii}=1$ and $\Sigma_{ij}=0.75$ whenever $i\neq j$. Fig. 6 gives the average HD for 100 replications with $p=100$ and $n=\{100, 500, 1000\}$, but further results can be found in the Supplementary Materials. Our two proposed CV methods were highly competitive with the Relaxed 1SE, whereas Relaxed Min tended to overselect as p^* increases, particularly for $n=100$. Similarly to the results from independent predictors, as n increased, the average HD for the relaxed Lasso and the two new methods of CV decreased for all considered p^*. Relaxed Lasso with a minimum APE rule struggled when n was small, even for a high signal-to-noise ratio, primarily due to overselection. Once

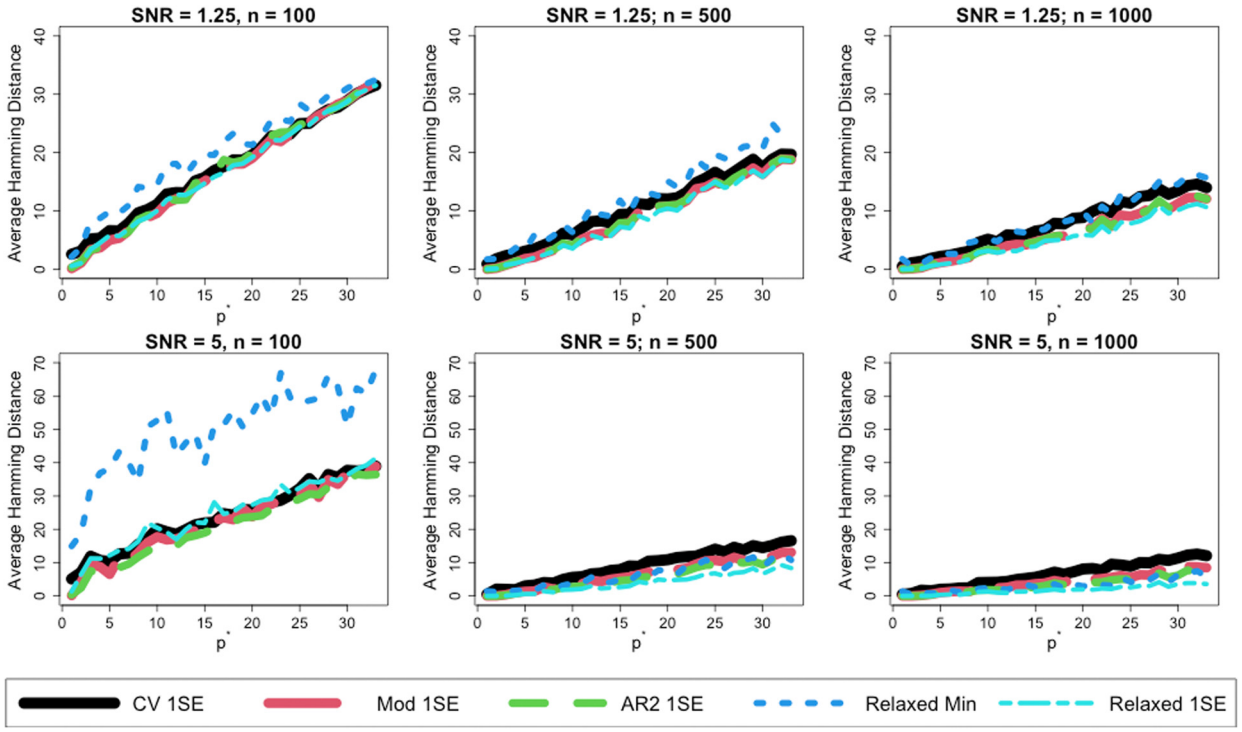

Fig. 6. Average Hamming Distance between $\boldsymbol{\beta}^*$ and $\hat{\boldsymbol{\beta}}_\lambda$ from 100 replications of simulated data with $n = \{100, 500, 1000\}$ and $p = 100$ for correlated predictors. Standard errors were too small to be plotted but can be seen in the Supplementary Material.

again the new methods of CV generally had a smaller average HD than CV 1SE and were similar in support recovery to the relaxed Lasso with a 1SE Rule.

Unlike the relaxed Lasso, the α-modification is straightforward to apply to other penalties. The α-modified Lasso was compared with the non-convex penalties SCAD and MCP, both traditionally and with an α-modification. As expected, the α-modification had a minimal effect on the predictive models. MCP and SCAD estimates were prone to underselection whereas the α-modified Lasso estimates were prone to overselection. Details may be found in the Supplementary Materials.

6. Optimal EMG placement for a robotic prosthesis controller

For the EMG application introduced in Section 1, we want to identify as few EMG sensors as needed to reliably predict hand movement. The data were collected from an able-bodied subject and consist of concurrent measurements of the subject's finger position and 16 EMG signals as predictors. A full description of the data and some of its challenges can be found in Stallrich et al. (2020). Six data sets were analyzed, corresponding to three consistent finger movement patterns (FC1, FC2, FC3) and three random patterns (FR1, FR2, FR3). As data were collected from an able-bodied subject, it was known that three of the 16 sensors, denoted X_5, X_7, and X_{12}, targeted muscles known to fully explain finger movement. Sensors X_5 and X_7 collected information from the same muscle, however, and so only one of the pair is necessary to predict finger position. An ideal model would thus include X_{12} and either X_5 or X_7, but recovery of all three sensors is also acceptable.

Due to known biomechanical features of hand movement, Stallrich et al. (2020) use finger velocity as the response and treat the recent past EMG signals as functional covariates. The model is

$$y_i = \sum_{j=1}^{16} \int_{-\delta}^{0} X_{ij}(t)\gamma_j(t, z_i)dt + \epsilon_i \,, \tag{12}$$

where y_i is the velocity, $X_{ij}(t)$ represents past EMG signals, $t \in [-\delta, 0]$ is the recent past time, and z_i is the recent finger position. Although this model is not the linear model introduced in (1), its approximation via basis expansion allows it to be treated similarly to the linear model. Following Gertheiss et al. (2013), Stallrich et al. (2020) proposed a penalized estimation procedure where the penalty accounted for both sparsity and smoothness of the $\gamma_j(\cdot, \cdot)$:

$$P_\lambda(\gamma_j) = \lambda(f_j||\gamma_j||_2^2 + g_j\lambda_t||\gamma''_{j,t}||_2^2 + h_j\lambda_z||\gamma''_{j,z}||_2^2)^{1/2} \,, \tag{13}$$

Table 4
Variable selection results for EMG finger movements without adaptive weighting. FP indicates the total number of false positives in the model, and Size is the total number of EMG signals contained in the model.

		FC1	FC2	FC3	FR1	FR2	FR3	Mean
APE	FP	2	1	2	2	1	1	1.5
	Size	4	3	4	5	4	4	4
AR2	FP	0	1	3	0	1	1	1
	Size	2	3	6	2	4	3	3.33
Mod	FP	1	1	2	2	1	1	1.33
	Size	3	3	4	5	4	4	3.83

where $||\gamma_j||_2^2 = \iint \gamma_j(t,z_i)^2 dtdz$ and $\gamma''_{j,t} = \partial^2\gamma_j(t,z_i)/\partial t^2$. There are three tuning parameters, λ, λ_t, and λ_z, and adaptive weights f_k, g_k, and h_k. To facilitate estimation, the $\gamma_j(\cdot,\cdot)$ were written using a tensor product basis expansion, leading to a group Lasso-type penalty; more details may be found in Stallrich et al. (2020) and the Supplementary Materials.

To perform variable selection, Stallrich et al. (2020) proposed Sequential Adaptive Functional Estimation (SAFE) that performs penalized estimation in stages. The first stage set $f_j = g_j = h_j = 1$ and chose optimal tuning parameters based on an APE 1SE rule following 10-fold CV. Let $\mathcal{M}_{\lambda,1}$ denote the support of this estimator. Adaptive weights were updated based on the estimates of the nonzero effects and penalized estimation was performed again using these weights and only those $j \in \mathcal{M}_{\lambda,1}$. The process was repeated for up to 5 stages. While effective in identifying the correct submodel, the analysis can be very time consuming. We modified their method based on AR2 CV and modified APE for this application in hopes to reduce the number of stages required to perform variable selection.

Table 4 gives the variable selection results for the three CV methods based on the initial stage; results from subsequent stages of SAFE can be found in the Supplementary Materials. AR2 CV and Mod APE generally give smaller models than traditional APE CV, with one exception: the Average R^2 method has a large model size for FC3. On average, however, both new methods have fewer false positives and smaller model sizes. Mod APE gives very similar results to APE, suggesting that the α-modified CV approach requires further study for the group Lasso. Although AR2 CV is not perfect in this application, in general it reduces model size and decreases false positives at no additional computational cost.

7. Discussion

In this paper, we proposed AR2 CV to choose tuning parameters to balance support recovery and prediction performance. This led to the α-modification, a multiplicative adjustment to predictions from penalized estimates which can also be used for α-modified CV. The α-modification is simple and efficient, making it an attractive option across many types of penalized estimators. A simulation study on the capabilities of AR2 and α-modified CV found that their variable selection results were highly competitive with—or, in some cases, better than—the relaxed Lasso. To ensure fair comparison, we introduced a 1SE Rule for the relaxed Lasso. Finally, we applied the approaches to a functional data problem in a demonstration of their flexibility.

The α-modification and the tuning parameter selection methods proposed here inspire several research questions. First, further theoretical analysis of the methods is of interest. Because the two new methods of tuning parameter selection are CV-based approaches, the theoretical properties of CV are central. Theoretical justifications for the use of CV for penalized estimators are relatively new and still evolving. Chetverikov et al. (2021) may be extended to show that α-modified and AR2 CV lead to estimates that are low bias and appropriately sparse. A second area of future research is the theoretical properties of α-modified estimates themselves. In Theorem 2, it was shown that any consistent estimator will still be consistent with the α-modification. We posit that the rates of convergence for α-modified estimators are faster than unmodified estimators, but a proof of this conjecture is the work of future research. Similarly, many of the results from this paper assume finite sample sizes. Further study is needed to determine more of the asymptotic properties of α-modified estimators and to adapt the specific results given for the Lasso penalty to other penalties.

There are also a few extensions and adaptations of the α-modification that may prove fruitful. We are currently expanding the α-modification to Generalized Linear Models (GLMs). This extension will require an iterative algorithm to find estimates of α because closed form solutions do not exist and an accommodation for the inclusion of an intercept term is necessary. Additionally, as noted in some of the theoretical results in this paper, the α-modification does not always reduce bias. We are interested in exploring a further penalty on α itself to ensure bias reduction. Finally, the calculation of $\hat{\alpha}_\lambda$ described here uses in-sample predictions and observations. As overspecification is a particular concern, the question of whether out-of-sample data can be used to find estimates of α is another subject of further research.

Acknowledgements

This work was partially supported by the National Science Foundation, grant IOS–2039226.

Appendix A

Proof of Theorem 1. Let $\hat{\boldsymbol{\beta}}_\lambda \neq 0$ and $g(\boldsymbol{\beta}) = \sum_{\ell=1}^{L} \lambda_\ell g_\ell(\boldsymbol{\beta})$. Denote the subgradient vector of $g(\boldsymbol{\beta})$ by $\nabla \mathbf{g}^*$. Then the KKT conditions give $\frac{1}{n}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_\lambda) = \nabla \mathbf{g}^*$ and there exists a generalized inverse $(\mathbf{X}^T\mathbf{X})^-$ where $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T\mathbf{X})^-(\mathbf{X}^T\mathbf{y} - n\nabla \mathbf{g}^*)$. It is easy to show $\hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda = \hat{\mathbf{y}}_\lambda^T\mathbf{y} - n\nabla \mathbf{g}^{*T}\hat{\boldsymbol{\beta}}_\lambda$. Therefore:

$$\hat{\alpha}_\lambda = 1 + \frac{n\nabla \mathbf{g}^{*T}\hat{\boldsymbol{\beta}}_\lambda}{\hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda}. \tag{A.1}$$

Because $\hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda$, n, and λ are greater than zero, we must show $\nabla \mathbf{g}^{*T}\hat{\boldsymbol{\beta}}_\lambda \geq 0$. As $\nabla \mathbf{g}^*$ is the subgradient of a convex function at $\hat{\boldsymbol{\beta}}_\lambda$, it satisfies

$$\nabla \mathbf{g}^{*T}\hat{\boldsymbol{\beta}}_\lambda \geq g(\hat{\boldsymbol{\beta}}_\lambda) - g(\boldsymbol{\beta}) + \nabla \mathbf{g}^{*T}\boldsymbol{\beta} \tag{A.2}$$

for all $\boldsymbol{\beta}$. For $\boldsymbol{\beta} = \mathbf{0}$, $g(\hat{\boldsymbol{\beta}}_\lambda) - g(\boldsymbol{\beta}) + \nabla \mathbf{g}^{*T}\boldsymbol{\beta} = g(\hat{\boldsymbol{\beta}}_\lambda) - g(\mathbf{0}) \geq 0$. Hence, $\nabla \mathbf{g}^{*T}\hat{\boldsymbol{\beta}}_\lambda \geq 0$. □

Proof of Lemma 1. When $\boldsymbol{\xi}^*$ is recovered, we have $\hat{\boldsymbol{\beta}}_\lambda = \|\hat{\boldsymbol{\beta}}_\lambda\|_2\boldsymbol{\xi}^*$. The response $\mathbf{y} = \alpha^*\mathbf{X}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}$, so:

$$\hat{\alpha}_\lambda\hat{\boldsymbol{\beta}}_\lambda = \alpha^*\boldsymbol{\xi}^* + \Big(\frac{\boldsymbol{\xi}^{*T}\mathbf{X}^T\boldsymbol{\epsilon}}{\boldsymbol{\xi}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\xi}^*}\Big)\boldsymbol{\xi}^*. \tag{A.3}$$

Thus $E[\hat{\alpha}_\lambda\hat{\boldsymbol{\beta}}_\lambda] = \alpha^*\boldsymbol{\xi}^* = \boldsymbol{\beta}^*$. □

Proof of Lemma 2. Fix λ and let $\hat{\boldsymbol{\beta}}_\lambda = a\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$ where $\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}$ is the OLS estimate for submodel $\mathcal{M}_\lambda$. Let $\mathbf{X}_{\mathcal{M}_\lambda}$ be the submatrix of $\mathbf{X}$ containing only the columns indexed by $\mathcal{M}_\lambda$. Then $\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda} = (\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{X}_{\mathcal{M}_\lambda})^{-1}\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{y}$. Now:

$$\hat{\alpha}_\lambda = \frac{a\mathbf{y}^T\mathbf{X}_{\mathcal{M}_\lambda}(\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{X}_{\mathcal{M}_\lambda})^{-1}\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{y}}{a^2\mathbf{y}^T\mathbf{X}_{\mathcal{M}_\lambda}(\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{X}_{\mathcal{M}_\lambda})^{-1}\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{X}_{\mathcal{M}_\lambda}(\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{X}_{\mathcal{M}_\lambda})^{-1}\mathbf{X}_{\mathcal{M}_\lambda}^T\mathbf{y}} = \frac{1}{a}. \tag{A.4}$$

Therefore:

$$\hat{\alpha}_\lambda\hat{\boldsymbol{\beta}}_\lambda = \frac{1}{a} \times a\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda} = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}_\lambda}. \quad \square \tag{A.5}$$

Proof of Theorem 2. Consider $\hat{\alpha}_{\lambda_n}$:

$$\hat{\alpha}_{\lambda_n} = \frac{\frac{1}{n}\hat{\boldsymbol{\beta}}_{\lambda_n}^T\mathbf{X}_n^T\mathbf{y}}{\frac{1}{n}\hat{\boldsymbol{\beta}}_{\lambda_n}^T\mathbf{X}_n^T\mathbf{X}_n\hat{\boldsymbol{\beta}}_{\lambda_n}}. \tag{A.6}$$

By Continuous Mapping Theorem, the denominator of (A.6) converges in probability to $c^2\boldsymbol{\xi}^{*T}\mathbf{C}\boldsymbol{\xi}^*$. Expanding $\mathbf{y} = \alpha^*\mathbf{X}\boldsymbol{\xi}^* + \boldsymbol{\epsilon}$ establishes the numerator converges to $\alpha c\boldsymbol{\xi}^{*T}\mathbf{C}\boldsymbol{\xi}^*$. Then $\hat{\alpha}_{\lambda_n} \to_p \frac{\alpha}{c}$ and $\hat{\alpha}_{\lambda_n}\hat{\boldsymbol{\beta}}_{\lambda_n} \to \frac{\alpha}{c}c\boldsymbol{\xi}^* = \alpha\boldsymbol{\xi}^* = \boldsymbol{\beta}^*$. □

Proof of Theorem 3. It is sufficient to show that $E[ModPE] < E[PE]$ where $PE = \sum_i (y_{k_i} - \mathbf{x}_{k_i}^T\hat{\boldsymbol{\beta}}_\lambda)^2$ and $ModPE = \sum_i (y_{k_i} - \hat{\alpha}_\lambda\mathbf{x}_{k_i}^T\hat{\boldsymbol{\beta}}_\lambda)^2$. We can write $\hat{\boldsymbol{\beta}}_\lambda = \|\hat{\boldsymbol{\beta}}_\lambda\|_2\boldsymbol{\xi}^*$. Expand $y_{k_i} = \alpha^*\mathbf{x}_{k_i}^T\boldsymbol{\xi}^* + \epsilon_i$ to find:

$$PE = \sum_i \big(\epsilon_i - \mathbf{x}_{k_i}^T\boldsymbol{\xi}^*(\|\hat{\boldsymbol{\beta}}_\lambda\|_2 - \alpha^*)\big)^2. \tag{A.7}$$

Similarly,

$$ModPE = \sum_i \left(\epsilon_i - \mathbf{x}_{k_i}^T\boldsymbol{\xi}^*\Big(\frac{\boldsymbol{\xi}^{*T}\mathbf{X}^T\boldsymbol{\epsilon}}{\boldsymbol{\xi}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\xi}^*}\Big)\right)^2. \tag{A.8}$$

For $V(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}$

$$E[PE] = n_k\sigma^2 + \sum_i (\mathbf{x}_{k_i}^T \boldsymbol{\xi}^*)^2 E[(\|\hat{\boldsymbol{\beta}}_\lambda\|_2 - \alpha)^2] \tag{A.9}$$

$$E[ModPE] = E\left[\sum_i \left(\epsilon_i - \mathbf{x}_{k_i}^T \boldsymbol{\xi}^* \left(\frac{\boldsymbol{\xi}^{*T}\mathbf{X}^T\boldsymbol{\epsilon}}{\boldsymbol{\xi}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\xi}^*}\right)\right)^2\right] \tag{A.10}$$

$$= n_k\sigma^2 + \frac{\sigma^2}{\boldsymbol{\xi}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\xi}^*}\sum_i (\mathbf{x}_{k_i}^T\boldsymbol{\xi}^*)^2 . \tag{A.11}$$

For $\boldsymbol{\beta}^* = \alpha^*\boldsymbol{\xi}^*$, $E[PE] \geq E[ModPE]$ whenever

$$E[(\|\hat{\boldsymbol{\beta}}_\lambda\|_2 - \alpha^*)^2] \geq \frac{\sigma^2}{\boldsymbol{\xi}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\xi}^*} \Leftrightarrow \frac{\alpha^{*2}}{E[(\|\hat{\boldsymbol{\beta}}_\lambda\|_2 - \alpha^*)^2]} \leq \frac{\boldsymbol{\beta}^{*T}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}^*}{\sigma^2} . \quad \square \tag{A.12}$$

Proof of Lemma 3. When $\mathbf{X}$ is orthonormal,

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = \left(\frac{\sum_{j=1}^p |\hat{\beta}_{OLS,j}|(|\hat{\beta}_{OLS,j}| - \lambda)_+}{\sum_{j=1}^p (|\hat{\beta}_{OLS,j}| - \lambda)_+^2}\right) \times \text{sign}(\hat{\beta}_{OLS,j^*}) \times (|\hat{\beta}_{OLS,j^*}| - \lambda)_+ . \tag{A.13}$$

Letting $d_j = (|\hat{\beta}_{OLS,j}| - \lambda)_+$ and $s_{j^*} = \text{sign}(\hat{\beta}_{OLS,j^*})$, we can rewrite this expression as

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = w_1\hat{\beta}_{OLS,j^*} + \hat{\beta}_{\lambda,j^*}\frac{\sum_{j\neq j^*}(d_j + \lambda)d_j}{\sum_{j=1}^p d_j^2} , \tag{A.14}$$

which uses $s_{j^*}|\hat{\beta}_{OLS,j^*}| = \hat{\beta}_{OLS,j^*}$, $w_1 = d_{j^*}^2/\sum_j d_j^2$, $\hat{\beta}_{\lambda,j^*} = s_{j^*}d_{j^*}$, and $|\hat{\beta}_{OLS,j}| = d_j + \lambda$ whenever $d_j > 0$. Hence we have the expression

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} = w_1\hat{\beta}_{OLS,j^*} + (1 - w_1)\hat{\beta}_{\lambda,j^*} + w_2\hat{\beta}_{\lambda,j^*} , \tag{A.15}$$

where $w_2 = \lambda\frac{\sum_{j\neq j^*} d_j}{\sum_{j=1}^p d_j^2}$. $\square$

Proof of Theorem 4. For $\hat{\beta}_{OLS,j^*} > \lambda$, applying Lemma 3 gives the expression

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}_{OLS,j^*} = \frac{\lambda(\sum_{j\neq j^*} d_j)\hat{\beta}_{OLS,j^*} - \lambda(\sum_{j\neq j^*} d_j^2 + \lambda\sum_{j\neq j^*} d_j)}{(\hat{\beta}_{OLS,j^*} - \lambda)^2 + \sum_{j\neq j^*} d_j^2} . \tag{A.16}$$

Let $u = \sum_{j\neq j^*} d_j$, $v = \sum_{j\neq j^*} d_j^2$, and $x = \hat{\beta}_{OLS,j^*}$. Note that $|\beta_{j^*}^*| \to \infty$ implies $|x| \to \infty$. We can express (A.16) as $f(x) = [\lambda ux - \lambda(v + \lambda u)]/[(x - \lambda)^2 + v]$ which is a differentiable function for $|x| \geq \lambda$. Between $x = \lambda$ and $x = (u\lambda + v + \sqrt{u^2v + v^2})/u \equiv x^*$, $f(x)$ is an increasing function bounded below by $-\lambda$ and bounded above by

$$\frac{\lambda}{2}\left(\sqrt{\frac{u^2}{v} + 1} - 1\right) . \tag{A.17}$$

For $x > x^*$, $f(x)$ is decreasing and it is easy to see $\lim_{x\to\infty} f(x) = 0 > -\lambda$. Similar arguments can be applied to the case of $x \leq -\lambda$. Therefore,

$$|f(x)| \leq \lambda \times \max\left(1, \frac{1}{2}\left(\sqrt{\frac{u^2}{v} + 1} - 1\right)\right) , \tag{A.18}$$

and $\lim_{x\to\infty} |f(x)| = 0$. $\square$

Proof of Theorem 5. When the sign vector of the true model, $\mathbf{s}$, is known and recovered by the lasso estimate, the nonzero coefficients have the expression $\hat{\boldsymbol{\beta}}_\lambda^{\mathcal{M}^*} = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} - n\lambda\tilde{\mathbf{s}}_j$. Then it is easy to show that both

$$\hat{\mathbf{y}}_\lambda^T\mathbf{y} = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*T}\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*}\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} - n\lambda\mathbf{s}^T\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} \tag{A.19}$$

$$\hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda = \hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*T}\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*}\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} - 2n\lambda\mathbf{s}^T\hat{\boldsymbol{\beta}}_{OLS}^{\mathcal{M}^*} + n^2\lambda^2\mathbf{s}^T(\mathbf{X}_{\mathcal{M}^*}^T\mathbf{X}_{\mathcal{M}^*})^{-1}\mathbf{s} , \tag{A.20}$$

are quadratic polynomials with respect to $\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ with quadratic coefficients of n (due to the scaling of the columns of $\boldsymbol{X}$). We also have

$$\hat{\mathbf{y}}_\lambda^T\mathbf{y} - \hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda = n\lambda s_{j^*}\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} + n\lambda\sum_{j\neq j^*} s_j\hat{\beta}^{\mathcal{M}^*}_{OLS,j} - n^2\lambda^2\mathbf{s}^T(\mathbf{X}^T_{\mathcal{M}^*}\mathbf{X}_{\mathcal{M}^*})^{-1}\mathbf{s}\,, \tag{A.21}$$

making $(\hat{\mathbf{y}}_\lambda^T\mathbf{y} - \hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda)\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ a quadratic polynomial with respect to $\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ with quadratic coefficient $n\lambda s_{j^*}$. Then

$$\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} = \hat{\alpha}_\lambda\big(\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} - n\lambda\tilde{s}_{j^*}\big) - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} \tag{A.22}$$

$$= \frac{(\hat{\mathbf{y}}_\lambda^T\mathbf{y} - \mathbf{y}_\lambda^T\hat{\mathbf{y}}_\lambda)\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} - n\lambda\tilde{s}_{j^*}\hat{\mathbf{y}}_\lambda^T\mathbf{y}}{\hat{\mathbf{y}}_\lambda^T\hat{\mathbf{y}}_\lambda}\,, \tag{A.23}$$

so $\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ is a ratio of quadratic polynomials with respect to $\hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ where the numerator and denominator quadratic coefficients are $n\lambda(s_{j^*} - n\tilde{s}_{j^*})$ and n, respectively. It follows that $\lim_{|\beta^*_{j^*}|\to\infty}\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*} = \lambda(s_{j^*} - n\tilde{s}_{j^*}) = G_{j^*}$ and so

$$\lim_{|\beta^*_{j^*}|\to\infty}|\hat{\alpha}_\lambda\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}| = |G_{j^*}|\,. \tag{A.24}$$

Clearly $-n\lambda\tilde{s}_{j^*} = \hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ so $|G_{j^*}| = |\lambda s_{j^*} + \hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}|$. If $\tilde{s}_{j^*} = 0$ then $\hat{\beta}_{\lambda,j^*} = \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}$ so $|G_{j^*}| = \lambda > 0 = |\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}|$. If $\tilde{s}_{j^*} \neq 0$ then

$$\lambda = \frac{-(\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*})}{n\tilde{s}_{j^*}} \tag{A.25}$$

which makes

$$|G_{j^*}| = \left|\left(1 - \frac{s_{j^*}}{n\tilde{s}_{j^*}}\right)(\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*})\right| = \left|1 - \frac{s_{j^*}}{n\tilde{s}_{j^*}}\right| \times |\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}| \tag{A.26}$$

Hence $|G_{j^*}| < |\hat{\beta}_{\lambda,j^*} - \hat{\beta}^{\mathcal{M}^*}_{OLS,j^*}|$ if and only if

$$\left|1 - \frac{s_{j^*}}{n\tilde{s}_{j^*}}\right| < 1\,. \quad \square \tag{A.27}$$

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2023.107729.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Allen, D.M., 1974. The relationship between variable selection and data agumentation and a method for prediction. Technometrics 16 (1), 125–127.
Breiman, L., 1995. Better subset regression using the nonnegative garrote. Technometrics 37 (4), 373–384.
Buhlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data. Methods, Theory and Applications. Springer Series in Statistics. Springer, Heidelberg.
Chetverikov, D., Liao, Z., Chernozhukov, V., 2021. On cross-validated Lasso in high dimensions. Ann. Stat. 49 (3), 1300–1317.
Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31 (4), 377–403.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32 (2), 407–499.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96 (456), 1348–1360.
Feng, Y., Yu, Y., 2013. Consistent cross-validation for tuning parameter selection in high-dimensional variable selection. In: 59th ISI World Statistics Congress, Session IPS046, pp. 435–465.
Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35 (2), 109–135.
Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70 (350), 320–328.
Gertheiss, J., Maity, A., Staicu, A.-M., 2013. Variable selection in generalized functional linear models. Stat 2 (1), 86–101.
Golub, G.H., Heath, M., Wahba, G., 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21 (2), 215–223.
Hastie, T., Tibshirani, R., 1986. Generalized additive models. Stat. Sci. 1 (3), 297–310.
Hastie, T., Friedman, J., Tisbshirani, R., 2017. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
Homrighausen, D., McDonald, D.J., 2018. A study on tuning parameter selection for the high-dimensional lasso. J. Stat. Comput. Simul. 88 (15), 2865–2892.
Huang, J., Horowitz, J.L., Ma, S., 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Stat. 36 (2), 587–613.
Hui, F.K.C., Warton, D.I., Foster, S.D., 2015. Tuning parameter selection for the adaptive lasso using ERIC. J. Am. Stat. Assoc. 110 (509), 262–269.
Hurvich, C.M., Tsai, C.-L., 1989. Regression and time series model selection in small samples. Biometrika 76 (2), 297–307.
Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. J. Mach. Learn. Res. 15, 2869–2909.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Stat. 28 (5), 1356–1378.

Krstajic, D., Buturovic, L., Leahy, D., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminform. 6, 10.
Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E., 2016. Exact post-selection inference, with application to the lasso. Ann. Stat. 44 (3), 907–927.
Lockhart, R., Taylor, J., Tibshirani, R.J., Tibshirani, R., 2014. A significance test for the lasso. Ann. Stat. 42 (2), 413–468.
Meinshausen, N., 2007. Relaxed lasso. Comput. Stat. Data Anal. 52 (1), 374–393.
Nardi, Y., Rinaldo, A., 2008. On the asymptotic properties of the group lasso estimator for linear models. Electron. J. Stat. 2, 605–633.
Nishii, R., 1984. Asymptotic properties of criteria for selection of variables in multiple regression. Ann. Stat. 12 (2), 758–765.
Ramsay, J.O., Silverman, B.W., 2005. Functional Data Analysis. Springer.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Shi, X., Liang, B., Zhang, Q., 2020. Post-selection inference of generalized linear models based on the lasso and the elastic net. Commun. Stat., Theory Methods 51 (14), 4739–4756.
Stallrich, J., Islam, M.N., Staicu, A.-M., Crouch, D., Pan, L., Huang, H., 2020. Optimal EMG placement for a robotic prosthesis controller with sequential, adaptive functional estimation (SAFE). Ann. Appl. Stat. 14 (3), 1164–1181.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc., Ser. B, Methodol. 36 (2), 111–147.
Taylor, J., Tibshirani, R.J., 2015. Statistical learning and selective inference. Proc. Natl. Acad. Sci. 112 (25), 7629–7634.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Methodol. 58 (1), 267–288.
Tibshirani, R.J., Taylor, J.E., Lockhart, R.A., Tibshirani, R., 2014. Exact post-selection inference for sequential regression procedures. J. Am. Stat. Assoc. 111, 600–620.
van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Stat. 42 (3).
Wu, Y., 2021. Can't ridge regression perform variable selection? Technometrics 63 (2), 263–271.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B 68, 49–67.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38 (2), 894–942.
Zhang, C.-H., Zhang, S.S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. J. R. Stat. Soc., Ser. B, Stat. Methodol. 76 (1), 217–242.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101 (476), 1418–1429.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B, Stat. Methodol. 67 (2), 301–320.