Title

Numerical characterization of support recovery in sparse regression with correlated design

Journal

Communications in Statistics - Simulation and Computation, 53(3)

ISSN

0361-0918

Authors

Kumar, Ankit

Bhattacharyya, Sharmodeep

Bouchard, Kristofer

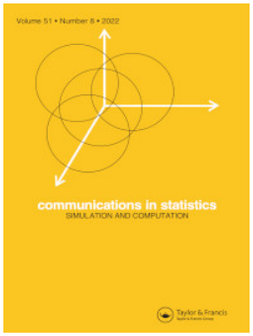
Publication Date

2024-03-03

DOI

10.1080/03610918.2022.2050392

Peer reviewed



Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Numerical characterization of support recovery in sparse regression with correlated design

Ankit Kumar, Sharmodeep Bhattacharyya & Kristofer Bouchard

To cite this article: Ankit Kumar, Sharmodeep Bhattacharyya & Kristofer Bouchard (2022): Numerical characterization of support recovery in sparse regression with correlated design, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2022.2050392

To link to this article: https://doi.org/10.1080/03610918.2022.2050392

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2022.2050392

Numerical characterization of support recovery in sparse regression with correlated design

Ankit Kumar[a,b,c], Sharmodeep Bhattacharyya[d], and Kristofer Bouchard[b,c,e,f]

[a]Department of Physics, University of California, Berkeley, Berkeley, California, USA; [b]Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, California, USA; [c]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; [d]Department of Statistics, Oregon State University, Corvallis, Oregon, USA; [e]Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA; [f]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, California, USA

ABSTRACT

Sparse regression is employed in diverse scientific settings as a feature selection method. A pervasive aspect of scientific data is the presence of correlations between predictive features. These correlations hamper both feature selection and estimation and jeopardize conclusions drawn from estimated models. On the other hand, theoretical results on sparsity-inducing regularized regression have largely addressed conditions for selection consistency via asymptotics, and disregard the problem of model selection, whereby regularization parameters are chosen. In this numerical study, we address these issues through exhaustive characterization of the performance of several regression estimators, coupled with a range of model selection strategies. These estimators and selection criteria were examined across correlated regression problems with varying degrees of signal to noise, distributions of non-zero model coefficients, and model sparsity. Our results reveal a fundamental tradeoff between false positive and false negative control in all regression estimators and model selection criteria examined. Additionally, we numerically explore a transition point modulated by the signal-to-noise ratio and spectral properties of the design covariance matrix at which the selection accuracy of all considered algorithms degrades. Overall, we find that SCAD coupled with BIC or empirical Bayes model selection performs the best feature selection across the regression problems considered.

ARTICLE HISTORY
Received 23 June 2021
Accepted 2 March 2022

KEYWORDS
Correlated variability; Model selection; Sparse regression; Information criteria; Compressed sensing

1. Introduction

In the last several decades, significant research in the mathematics and statistics communities has been directed at the problem of reconstructing a k-sparse vector from noisy, linear observations. In its simplest form, one is concerned with inference within the following model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \epsilon \tag{1}$$

with $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times p}$ and $\beta \in \mathbb{R}^p$ is a k-sparse vector. The noise is i.i.d, $\epsilon \in \mathbb{R}^n, \epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and the observational model is Gaussian, $y_i \sim \mathcal{N}(\boldsymbol{X}_i\beta, \epsilon_i)$. The sparse linear model is employed in

CONTACT Kristofer Bouchard kebouchard@lbl.gov Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, USA.

Supplemental data for this article is available online at https://doi.org/10.1080/03610918.2022.2050392

diverse scientific fields (Tibshirani 1997; Wright et al. 2010; Waldmann et al. 2013; Steyerberg and Vergouwe 2014; Satija et al. 2015). In real world applications, it is also commonly the case that the design or covariate matrix X is correlated, so that the columns of $\boldsymbol{X}$ cannot be taken to be i.i.d. In this setting, the correct identification of non-zero elements of β, which is crucial for scientific interpretability, is especially challenging. Yet, a systematic exploration of the effect of correlations between the covariates on the recoverability of β is lacking.

Statistically optimal sparse estimates of β within (1) are returned by the solution to the following constrained optimization problem:

$$\begin{aligned} &\min ||y - \boldsymbol{X}\beta||_2^2 \\ &||\beta||_0 \leq \lambda \end{aligned} \tag{2}$$

Finding the global minima of problem (2) is NP-hard, though recent progress has been made in computationally tractable approaches (Bertsimas, King, and Mazumder 2016; Zhu et al. 2020). The most common approach is to relax the l_0 regularization. In this work, we focus on the Lasso, Elastic Net, SCAD, MCP (Tibshirani 1996; Fan and Li 2001; Zou and Hastie 2005; Zhang 2010), and $\text{UoI}_{\text{Lasso}}$, an inference framework we introduced in (Bouchard et al. 2017) that combines stability selection and bagging approaches to produce low variance and nearly unbiased estimates. To select the regularization strength or otherwise compare between candidate models returned between these estimators, one must employ a model selection criteria such as cross-validation or BIC. While the literature on sparsity inducing estimators and model selection criteria is vast, studies that consider the interaction of particular choices of estimator and model selection criteria are lacking. In particular, no systematic exploration of the impact of choice of estimator *and* model selection criteria on the selection accuracy of the resulting procedure when the predictive features exhibit correlations has been carried out. In this work, we address this gap by performing systematic numerical investigations of the selection accuracy performance of several estimators and model selection criteria across a broad range of regression designs, including diverse correlated design matrices. Section 2 summarizes prior theoretical and empirical work on model selection and compressed sensing. We also discuss a scalar parameterization of signal strength in correlated sparse regression borrowed from (Wainwright 2009) that we call α. In Sec. 3, we outline the scope of this study and the evaluation criteria used. In Sec. 4 we present the main results. We characterize the impact of correlated design on the false negative and false positive discovery rates, as well as the magnitude of coefficients likely to be falsely set to zero or false assigned non-zero values. We reveal that estimators and selection methods display a remarkable degree of universality with respect to the correlation strength (quantified by α). We also identify the best performing combinations of estimator and selection methods under various signal conditions. Connections to prior theoretical work and concrete recommendations for practitioners are provided in Sec. 5.

2. Review of prior work

The statistical theory of the sparse estimators considered in this work is vast and we do not attempt to review it all here. Our particular focus is on characterizing finite sample selection accuracy, especially in the context of correlated design. The asymptotic oracular selection performance of the SCAD and MCP are well known (Fan and Li 2001; Zhang 2010) and require only mild conditions on the design matrix. For the Lasso, one must impose an irrepresentable condition to guarantee asymptotic selection consistency (Zhao and Yu 2006). The finite sample implications of these differing requirements have not been explored. A series of works have addressed the correlated design problem by devising regularizations that tend to assign correlated covariates similar model coefficients (Tibshirani et al. 2005; Bogdan et al. 2013; Bühlmann et al. 2013; Witten, Shojaie, and Zhang 2014; Figueiredo and Nowak 2016; Li et al. 2018). In fact, the Elastic Net was the first estimator introduced to exhibit this type of "grouping" effect (Zou and

Hastie 2005). However, this type of behavior can be undesirable in many real data applications where covariates may be correlated, yet still contribute heterogeneously to a response variable of interest.

When the true model generating the data is contained amongst the candidate model supports, the BIC and gMDL have asymptotic guarantees of selection consistency (Zhao and Yu 2006). Extensions of these results to the high dimensional case are available (Kim, Kwon, and Choi 2012), but fall outside the scope of this work. Implicit in these theoretical results is that one can evaluate the penalized likelihoods on all 2^p candidate model supports (Shao 1997). Practically, one first assembles a much smaller set of candidate model supports using a regularized estimators. To this end, the use of the BIC with SCAD has been shown to be selection consistent (Wang, Li, and Tsai 2007).

A more recent body of work has focused on non-asymptotic analyses of model (1) in the framework of compressed sensing rather than regression. Here, the sparsity level of β is a priori known, and the sensing matrix X is typically drawn from a random ensemble. In this setting, it is possible to establish sharp transitions in the mean square error distortion of the signal vector as a function of measurement density (i.e., asymptotic n/p ratio) (Donoho, Maleki, and Montanari 2009). Necessary and sufficient conditions on the number of samples needed for high probability recovery of the support of β by the Lasso was treated in (Wainwright 2009). Subsequently, a series of works examined the information theoretic limits on sparse support recovery by forgoing analysis of computationally tractable estimators in favor of establishing the sample complexity of exhaustive evaluation of all $\binom{p}{k}$ possible supports via maximum likelihood decoding (Wainwright 2009; Atia and Saligrama 2009; Aeron, Saligrama, and Zhao 2010; Rad 2011; Scarlett, Evans, and Dey 2013; Aksoylar and Saligrama 2014; Aksoylar, Atia, and Saligrama 2017; Scarlett and Cevher 2017). This approach provides information theoretic bounds on the selection performance of any inference algorithm, and a measure of the suboptimality of existing algorithms.

Of particular relevance to this work are (Wainwright 2009) and (Scarlett and Cevher 2017), whose analyses permit correlated sensing (i.e., design) matrices. Let $\beta_{\min}$ be the minimum non-zero coefficient of β, σ^2 be the additive noise variance, and Σ be the covariance matrix of the distribution from which columns of $\boldsymbol{X}$ are drawn. Denote the set of all subsets of $\{1, 2, \ldots, p\}$ of size k as $\mathcal{I}_k$. $\mathcal{I}_k$ indexes possible model supports. Given $S, T \in \mathcal{I}_k$ we define the matrix $\Gamma(S, T)$ to be the Schur complement of $\Sigma_{S\cup T, S\cup T}$ with respect to Σ_{TT}, $\Gamma(T, S) = \Sigma_{S\setminus T, S\setminus T} - \Sigma_{S\setminus T, T}(\Sigma_{TT})^{-1}\Sigma_{T, S\setminus T}$. Let $\rho(\Sigma, k)$ be the smallest eigenvalue this matrix can have for any T: $\rho(\Sigma, k) = \min_{T\in\mathcal{I}_k\setminus S}\lambda_{\min}(\Gamma(T, S))$. From these quantities, we define α:

$$\alpha = \frac{\beta_{\min}^2 \rho(\Sigma, k)}{\sigma^2} \tag{3}$$

In Theorem 1 of Wainwright (2009), sufficient conditions on the sample size required for an exhaustive search maximum likelihood decoder to recover the true model support with high probability are given in terms of p, k, and α:

Theorem 1. *Theorem 1 of* Wainwright (2009)*. Define the function* $g(c_1, p, k, \alpha)$*:*

$$g(c_1, p, k, \alpha) := (c_1 + 2048)\max\left\{\log\binom{p-k}{k}, \log(p-k)/\alpha\right\}$$

If the sample size n satisfies $n > g(c_1, p, k, \alpha)$ *for some* $c_1 > 0$*, then the probability of correct model support recovery exceeds* $1 - \exp(-c_1(n-k))$.

If $\alpha^{-1} > p\log(p-2k) + 2k/p$, then g, and therefore the sample complexity of support recovery, will be modulated by α for p large enough. Many of the design matrices considered in our numerical study (see Sec. 3) satisfy this condition.

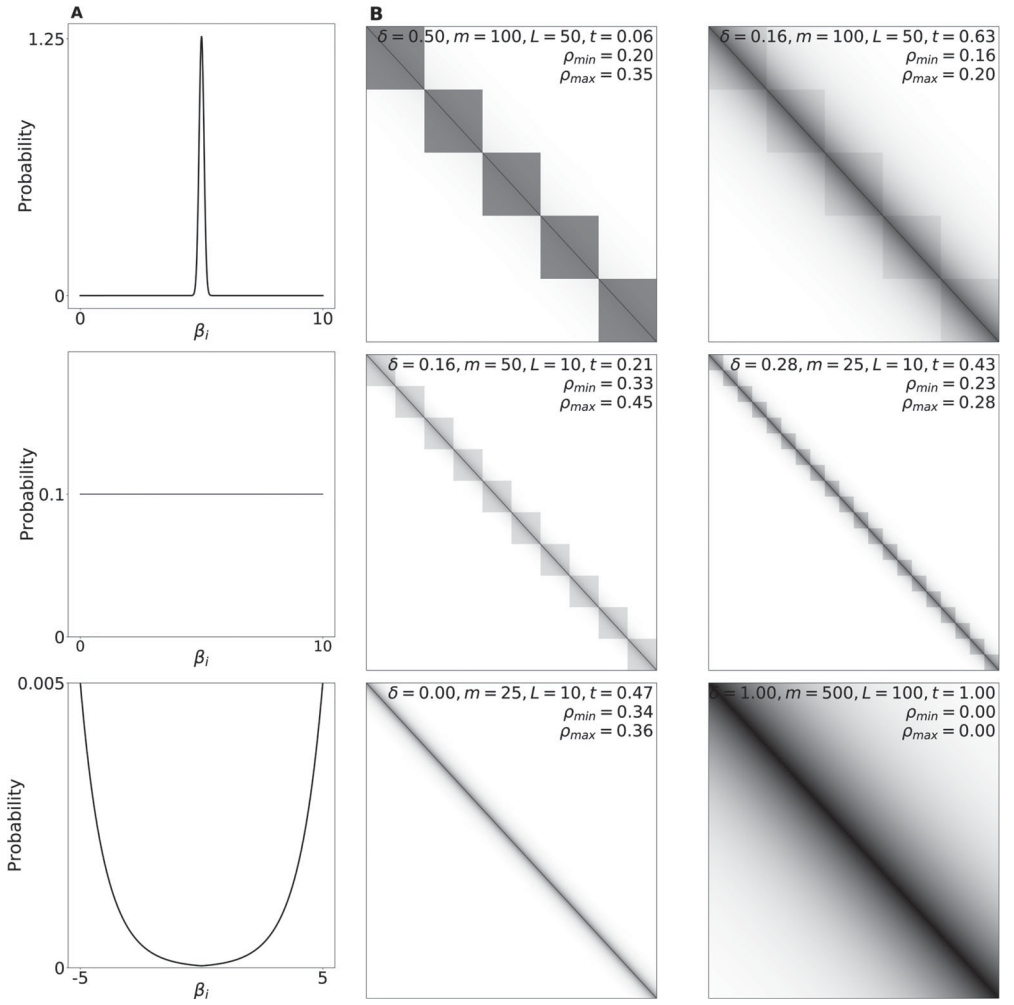


Figure 1. Design of Simulation Study. (a) (Right column) Coefficients β are drawn from a narrowly peaked Gaussian, uniform, and inverse exponential distribution. (b) (Left column) Design matrices are parameterized as $\Sigma = t\bigoplus_i \delta I_{m\times m} + (1-t)\Lambda(L)$ where $\Lambda(L)_{ij} = \exp(-|i-j|/L)$ and $I_{m\times m}$ is the m-dimensional identity matrix. Parameters δ, m, t and L are shown for each example design matrix. Also shown are bounds for the minimum and maximum $\rho(\Sigma, k)$ across k.

In contrast to compressed sensing, the sparsity level of β (i.e., k) is typically unknown in applications of regression. Furthermore, sufficient conditions on high probability theory such as Theorem 1 above rely on concentration inequalities, which may formally hold in the non-asymptotic setting, but are rarely tight. As a result, the applicability of these results for practitioners evaluating the robustness of support recovery in finite sample regression is unclear. The main contribution of this work is to address this gap through extensive numerical simulations. We find α to be a useful measure of the difficulty of a particular regression problem, and find selection accuracy performance to be modulated by α even when it does not satisfy the condition stated above.

Previous empirical works have evaluated the effects of collinearity on domain specific regression problems (Dormann et al. 2013; Vatcheva et al. 2016) and evaluate the efficacy of various information criteria for model selection (Schöniger et al. 2014; Brewer, Butler, and Cooksley 2016; Dziak et al. 2020). Finally, the performance scaling of a series of sparse estimators with sample size is evaluated in (Bertsimas, Pauphilet, and Van Parys 2020).

Table 1. (Top) Sparsity inducing regularized estimators. λ and γ denote regularization parameters. In this study, we keep γ for SCAD and MCP fixed to 3. (Bottom) Model selection criteria.

Estimator	Regularization
Lasso	$\lambda\|\beta\|_1$
Elastic Net	$\lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2$
SCAD	$\int_0^{\|\beta\|} dx(\lambda\mathbb{I}(\|\beta\| \leq \lambda) + \frac{(\gamma\lambda - x)_+}{(\gamma-1)\lambda}\mathbb{I}(\|\beta\| > \lambda))$
MCP	$\int_0^{\|\beta\|} dx(1 - \frac{x}{\gamma\lambda})_+$
$\text{UoI}_{\text{Lasso}}$	$\lambda\|\beta\|_1$ across bootstraps, see Bouchard et al. (2017)
Model Selection Criteria	
Cross-Validation	R^2 averaged over 5 folds
BIC	$2\log\|y - X\hat{\beta}\|_2^2 - \log(n)\|\hat{\beta}\|_0$
AIC	$2\log\|y - X\hat{\beta}\|_2^2 - 2\|\hat{\beta}\|_0$
gMDL (Hansen and Yu 2001)	$\begin{cases} \frac{\hat{k}}{2}\log\left(n - \hat{k}\hat{k}\frac{y^\top y - \|y-\hat{y}\|_2^2}{\|y-\hat{y}\|_2^2}\right) + \log n & \text{if } R^2 > \frac{\hat{k}}{n} \\ \frac{n}{2}\log\left(\frac{y^\top y}{n}\right) + \frac{1}{2}\log(n) & \text{otherwise} \end{cases}$
Empirical Bayes (George and Foster 2000)	$2\log\|y - X\hat{\beta}\|_2^2 - \begin{cases} \hat{k} + \hat{k}\log(\hat{y}^\top\hat{y}) - \hat{k} - 2((p-\hat{k})\log(p-\hat{k}) + \hat{k}\log\hat{k}) & \text{if } \hat{y}^\top\hat{y}/\hat{k} > 1 \\ \hat{y}^\top\hat{y} - 2((p-\hat{k})\log(p-\hat{k}) + \hat{k}\log\hat{k}) & \text{otherwise} \end{cases}$

Here and throughout, $\hat{k}$ refers to the estimated support size, $\hat{y}$ the model predictions of y, and p is the total number of features.

In contrast, we specifically consider the differing effects on selection accuracy of *joint* choices of estimators and model selection criteria. We demonstrate that the choice of model selection criteria significantly modulates the selection performance of estimators, and that there are empirically identifiable transition points in the value of α beyond which the selection performance of all inference procedures degrades.

3. Methods

We consider regression problems with 500 features with 15 different model densities (i.e., $|\beta|_0$) logarithmically distributed from 0.025 to 1. Additionally, we vary over the following design parameters:

1. 80 covariance matrices Σ of exponentially banded, block diagonal, or a structure that interpolates between the two (see Figure 1).
2. Three different β distributions: a sharply peaked Gaussian, a uniform, and an inverse exponential distribution (see Figure 1)
3. Signal to noise (SNR) ratios of 1, 2, 5, 10. We define signal to noise as $|X\beta|_2^2/\sigma^2$.
4. Sample to feature (n/p) ratios of 2, 4, 8, and 16.

To simplify the presentation, we often restrict the analysis to the following three combinations of SNR and n/p ratio that represent ideal signal and sample, SNR starved, and sample starved scenarios, respectively:

1. Case 1: SNR 10 and n/p ratio 16
2. Case 2: SNR 1, and n/p ratio 4
3. Case 3: SNR 5 and n/p ratio 2

A distinct model design is comprised of a particular model density, predictor covariance matrix, a coefficient distribution drawn from one of the three β-distributions, an SNR, an n/p ratio. Each distinct model is fit over 20 repetitions with each repetition being comprised of a new draw of $X \sim \mathcal{N}(0, \Sigma)$ and $\epsilon \sim \mathcal{N}(0, \sigma^2)$, with σ^2 set by the desired SNR. We use the term

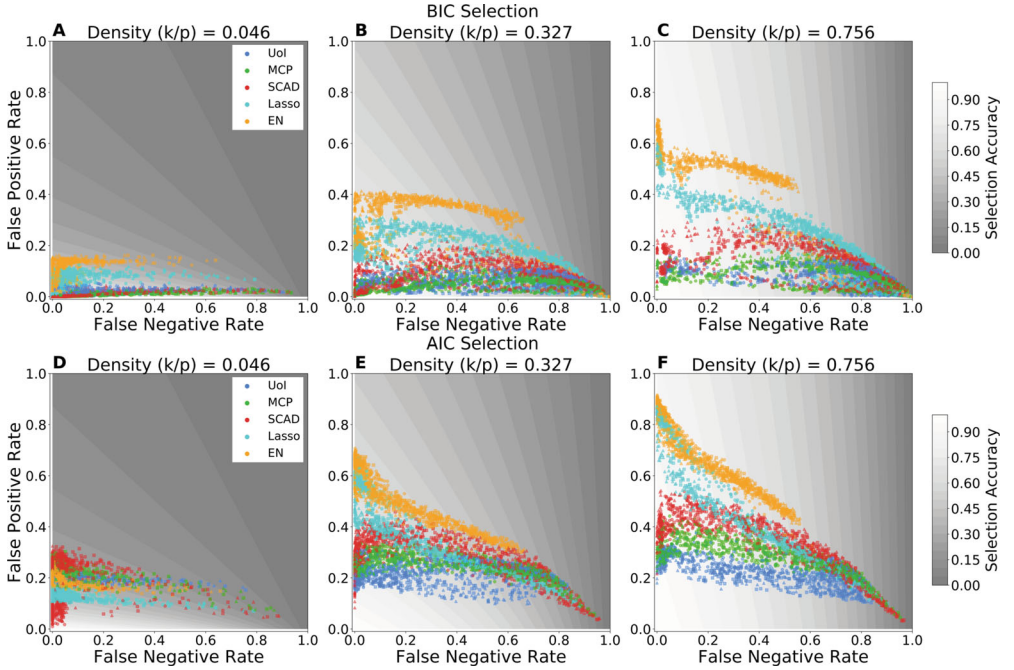


Figure 2. Scatter plots of the false negative rate vs. false positive rate for BIC selection (A-C) and AIC selection (D-F) across 3 different model densities (n/p ratio = 4, all signal to noise parameters included). Each scatter point represents a single fit. β distributions are encoded in marker shapes (square: uniform distribution, triangular: inverse exponential distribution, circular: Gaussian distribution). Shaded regions represent regions of equal selection accuracy. The orientation of these regions for different model densities illustrates the differing contributions of false negatives vs. false positives, with false positive control being far more important for sparser models, and conversely false negatives being more important for denser models. Estimators can be seen to be characterized by specific tradeoffs between the false positive and false negative control, with SCAD/BIC/UoI (red/green/blue) controlling the false positive rate most aggressively, whereas Elastic Net (orange) controls for false negatives more effectively. The tails of the scatter points extending toward the bottom right of the plot are comprised of the model designs with smallest α.

estimator to refer to a particular regularized solution to problem 1 (e.g., Lasso) and model selection criteria to refer to the method used to select regularization strengths (e.g., BIC). The estimators and model selection criteria we consider are listed in Table 1. We use the term inference algorithm to refer to particular choices of estimator and model selection criteria.

Let $S = \{i|\beta_i \neq 0\}$ in eq. 1, and $\hat{S} = \{i|\hat{\beta}_i \neq 0\}$, i.e., the true and estimated model supports. Then, we evaluate regression on the basis of selection accuracy ($1 - \frac{|(S\setminus\hat{S})\cup(\hat{S}\setminus S)|_0}{|S|_0+|\hat{S}|_0}$), false negative rate ($\frac{|S\setminus\hat{S}|_0}{|S|_0}$) and false positive rate ($\frac{|\hat{S}\setminus S|_0}{p-|S|_0}$). We use α to associate a single scalar to measure the difficulty of a regression problem. Smaller α correspond to harder regression problems. In practice, we do not calculate $\rho(\Sigma, k)$ explicitly, but rather lower bound it (Supplementary section S1). The parameter $\rho(\Sigma, k)$ becomes smaller with larger k.

4. Results

4.1. False positive/false negative characteristics

We first visualized support selection performance across estimators by scattering the false negative rate vs. false positive rate of each fit for several representative model densities (Figure 2 for BIC and AIC selection, Figure S1 for other criteria). Each scatter point represents the selection characteristics of fits to a distinct model design averaged over its 20 instantiations. The boundaries of the grayscale partitions of the false positive false negative rate plane correspond to contours of equal selection accuracy. The rotation of these contours with the true underlying model

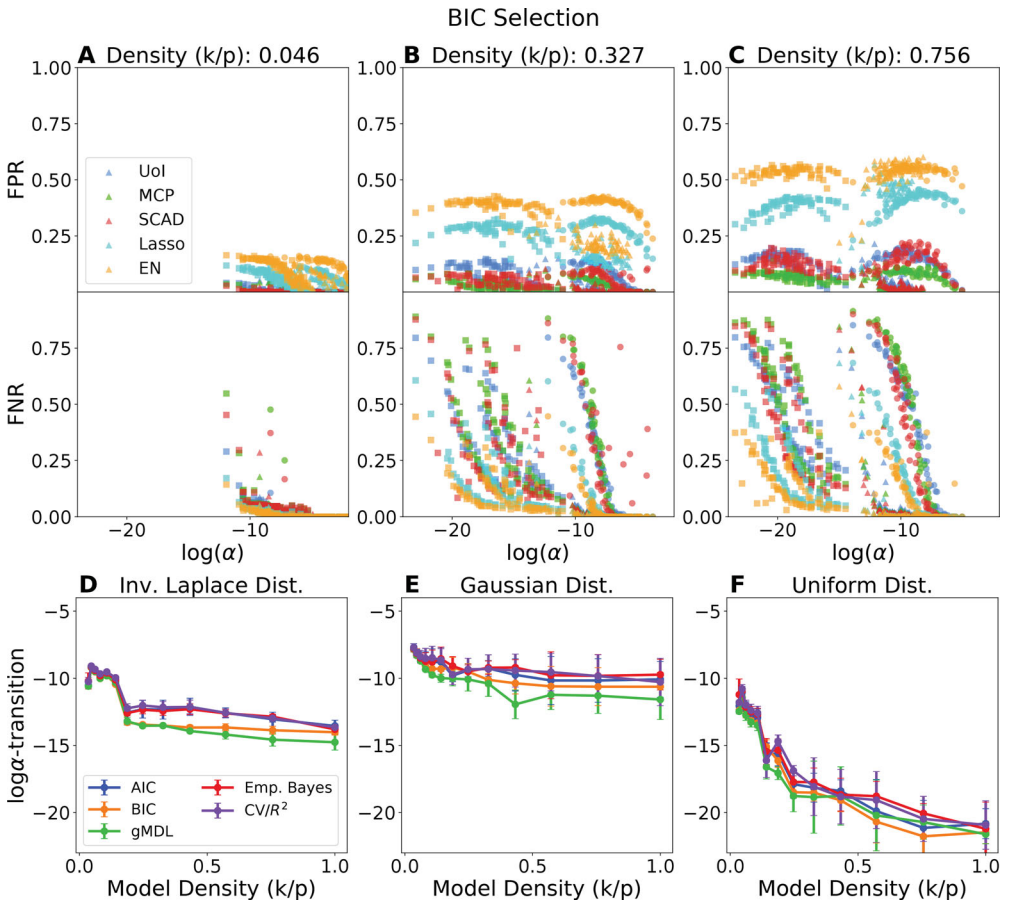


Figure 3. (A–C) Scatter plot of the false positive rate and the false negative rate vs. α for each estimator using BIC as a selection criteria for three different model densities. β distributions are encoded in marker shapes (square: uniform distribution, triangular: inverse exponential distribution, circular: Gaussian distribution). The false positive rate is only weakly modulated by α, but is ordered by estimator, with MCP/SCAD/UoI consistently having the lowest false positive rates, Lasso intermediate, and the Elastic Net having the highest false positive rates. The effect becomes pronounced at higher model densities (panel C). For the false negative rates, scatter points follow a characteristic sigmoidal profile, with a region of stable selection accuracy at low correlation (high α), followed by an transition point of α after which the selection accuracy monotonically decreases. The effect is again most visible at higher model densities. At model density 0.046 (panel A, bottom), all estimators are reasonably robust to decreasing alpha. (D–F) Plot of the α-transition point associated with an inference algorithm's false negative rate as a function of model density, separated by β distribution and selection method. Error bars are standard deviations taken across repetitions and estimator. The different numerical regimes of the α-transition (highest in panel E, intermediate in panel D, and lowest in panel F) are attributable to the different characteristic value of β_{min} for the different β distributions.

density reflects the relative importance of false negative and false positive control in modulating selection accuracy. Specifically, rotation toward the horizontal implies larger sensitivity to false positives, while conversely rotation toward the vertical implies greater sensitivity toward false negatives.

The accuracy of estimators exhibited clear structure that depends on the characteristics of the model design described above. We observe in panel A of Figure 2 that estimators that more aggressively promote sparsity (SCAD, MCP, UoI in red, green, and dark blue, respectively) featured better selection accuracy at low model densities (i.e., scatter points for these estimators lie in the white to light gray shaded regions), whereas those that control false negatives less aggressively, namely the Elastic Net (orange) and to a lesser extent the Lasso (cyan), fared better in denser true models (panel C). The scatter points for each estimator formed bands that span the false negative rate. This banding effect was most pronounced for SCAD/MCP/UoI.

Comparing the BIC selection (Figure 2A–C) to AIC (Figure 2D–F), these scatter plots also revealed that varying model selection methods also systematically shifted false negative & false positive characteristics of estimators. Selection methods with lower complexity penalties (i.e., AIC, CV) lifted the bands up along the false positive direction. Comparing the location of the blue/red/green scatter points between panels B and E, for example, we note that this effect was most dramatic for the set of estimators that most aggressively control false positives (SCAD/MCP/UoI). Consequently, similar tradeoffs as described before arose, with empirically better selection accuracy when models are dense obtained for AIC/CV, and vice versa for larger complexity penalties (BIC). The gMDL and eB methods behaved similarly to BIC (although there are a few exceptions to this, see Section S2). We conclude that the choice of estimator *and* model selection criteria are both important in determining the false positive/false negative rate behavior of inference strategies.

4.2. α-Dependence of false positives/false negatives

Recalling that the parameter α tunes the difficulty of the selection problem, we scattered the false positive and false negative rate vs. α for each inference algorithm across different model densities. A representative set of such plots for BIC selection is shown in Figure 3A–C; other selection methods are shown in Supplementary Figure S2. There was broadly large variation in performance modulated by the selection method employed. Furthermore, β-distributions are separately resolvable due to their different typical values of $\beta_{\min}$. For example, in the bottom axes of Figure 3C, for each estimator, the uniform distribution scatter points (squares) lie to the left of the inverse exponential distribution (triangular), which in turn lies to the left of the Gaussian distribution (circular).

In line with Figure 2, the false positive rate was not modulated by α (Figure 3A–C, top axes). In fact, for some estimators, the highest false positive rate was achieved for intermediate α, followed by a decline in false positive rate for smaller α (e.g., Lasso in Figure 3C). The false positive rate is instead a characteristic of each estimator. The SCAD/MCP/UoI class of estimators achieved lower false positives than Lasso, which in turn featured lower false positives than the Elastic Net. Model selection criteria can also be classified into a set that led to low false positive rates (gMDL, empirical Bayes, and BIC) vs. those that lead to high false positive rates (AIC, CV), although the Elastic Net with empirical Bayes selection featured the highest false positive rate of any inference algorithm (Figure S2, panels D-F).

On the other hand, the false negative rate scatter points, when separated by β-distribution, featured consistent behavior across inference algorithms. Focusing on BIC selection, all estimators achieved low false negative rates at the low model densities (Figure 3A). At intermediate model densities (Figure 3B), the false negative rate remained low until $\log \alpha$ became sufficiently small, at which point it rapidly increases. This value of $\log \alpha$ varied by β-distribution due to the differing characteristic values of $\beta_{\min}$, occurring around $\log \alpha \approx -7.5$ for the Gaussian distribution at model density 0.327, $\approx \log \alpha = -10$ for the inverse exponential distribution, and $\approx \log \alpha = -15$ for the uniform distribution. Otherwise, this transition point is fairly universal across inference algorithms.

To produce summary statistics of false negative rates across model densities, selection methods, and n/p ratio/SNR cases, we fit sigmoidal curves to data for each inference algorithm and for each β distribution. The sigmoid curve is described by 4 parameters:

$$S(\alpha) = c + \frac{a}{1 + \exp\left(-b(\alpha - \alpha_0)\right)}$$

In particular, we use the fitted value for the sigmoid midpoint α_0, which we refer to as the α-transition point, to quantify the value of α at which false negative rate has begun to increase

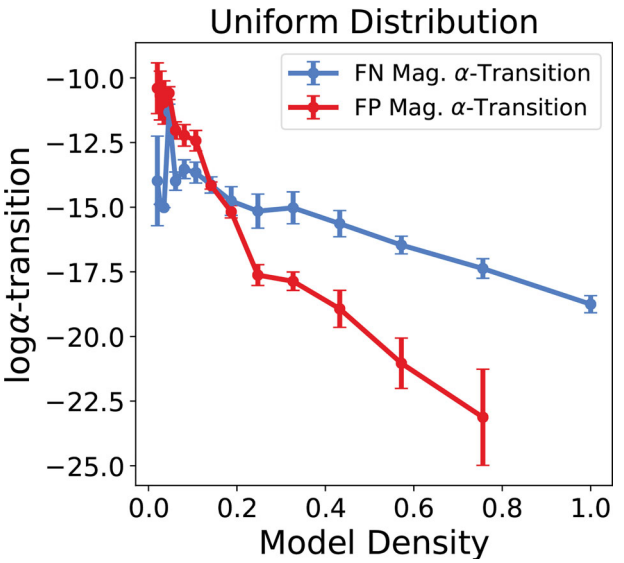


Figure 4. Plot of the average α transition point for estimation distortion across all inference algorithms and selection methods vs. model density for signal case 1. Error bars represent standard deviation. After a model density of > 0.15, the transition generally occurs at lower correlations (smaller α) for the false negative magnitude. Furthermore, the variance across inference algorithms is consistently smaller for false negatives as opposed to false positives.

appreciably. We found a large degree of universality in this transition point across estimators and selection methods. In Figure 3D–F, we have averaged curves across estimators and plotted the mean and standard deviation of the resulting α transition points. Colors now represent each selection method. The curves for each selection method were strikingly similar within a β distribution, with small standard deviations within each selection method indicating universality across estimators. The decrease of the α-transition point with increasing model density can be explained by the overall shift of α toward smaller values due to the increase of $\rho(\Sigma, k)$ with k.

In the preceding analysis we treated false positives and false negatives as hard thresholded quantities. On the other hand, one can ask whether false negatives primarily arise from setting support elements with small signal strength to zero, and conversely whether false positives are associated with small coefficient estimates. Thus, while exact model support recovery in most cases is unattainable, one would hope that support inconsistencies produce low distortion of the desired coefficient vector. To evaluate this supposition, we calculated the average magnitude of false negatives and false positives, and normalized these quantities by the average magnitude of ground truth β. In the case of false negative magnitudes, we focused on the uniform β distribution, as this provides the most "edge" cases of small coefficient magnitudes. Raw scatter plots of these quantities (not shown) revealed that at low correlations, the hoped for low distortion effect largely holds true, but that there is an α transition point for both false negative and false positives after which significantly larger ground truth β_i are selected out, and erroneously selected β_i are assigned much larger values relative to the true signal mean.

We again fit sigmoidal functions to the raw scatter points of normalized false negative & false positive magnitude vs. $\log \alpha$ and extracted the α-transition points as in Figures 3D–F. In Figure 4, we plot the transition point as a function of model density averaged across all estimators, selection criteria, and fit repetitions. For model densities > 0.15, the transition point occurs at much smaller correlation strengths for false negative distortions than for false positive distortions. The variance in the location of this transition point for false negative distortions is noticeably smaller than for false positive distortions. Nevertheless, similarly to the behavior exhibited by the α-transition points associated with the false negative rate, the α-transition points for false positive/false

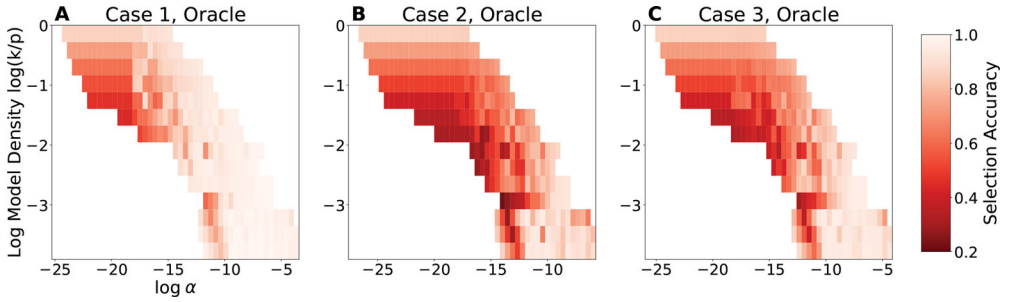


Figure 5. Oracle selection accuracy as a function of the log model density and α for each of the 3 signal cases described in Section 3. Each pixel in the colormap is the maximum oracle performance across all estimators for the particular combination of density and α. For ideal signal characteristics in Case 1 (panel A), near perfect support recovery is in principle possible for a broad range of correlation strengths for log model densities < -1.9. The similar oracle selection accuracies between cases 2 and 3 (panels B and C) suggest that the sample starved and signal starved regression problems behave similarly. As compared to Case 1, worst case performance for intermediate model densities ($\mathbf{log(k/p) > -2.3}$ and $< \mathbf{-0.69}$) is lower, especially for large correlations. For the densest models ($\mathbf{log(k/p) > 0.5}$), oracle performance is relatively insensitive to correlation strength, reflecting the insensitivity of the FPR to α. Near-perfect support recovery is empirically still possible for the sparsest models ($\mathbf{log(k/p) < -3}$).

negative coefficient magnitudes is saliently uniform across inference strategies. Overall, these results highlight the usefulness in the parameter α, which emerges out of tail bounds on the performance of the exhaustive maximum likelihood decoder, as a quantifier of the difficulty of a sparse regression problem.

4.3. Overall selection accuracy

An inference algorithm deployed in practice must employ both an inference estimator and model selection criteria. We have therefore determined what the best performing combination is as a function of underlying model density and α. To set an overall scale for these comparisons, one can use an oracle selection criteria that simply chooses the support along a regularization path of maximum selection accuracy. For each value of α and model density, the maximum of this oracular selection across all estimators gives a proxy for the best achievable selection accuracy in principle at finite sample size and SNR.

In Figure 5, we plot the oracle selector for each signal case. In the ideal signal and sample size case (case 1), the oracle selector was able to achieve near perfect selection accuracy in the fully dense models (top row, panel C) and those models with density < 0.14 (log model densities < -2) even in model designs with very small α. The oracle selector suffered moderate loss of selection accuracy in intermediate model densities for model designs with small α (darker orange regions of panel C). A similar structure is present in the adequate sample but high noise and low sample size but adequate SNR cases (cases 2 and 3 in panels B and C, respectively), but the magnitude of selection accuracy performance loss and regions of α and model densities for which the loss occurred expanded. In particular, only in the very sparsest models (density < 0.05, log model density < -3) with larger α was perfect selection possible in principle.

For each estimator and selection criteria combination, we take the average deviation of its selection accuracy from the oracular performance shown in Figure 5 as a measure of sub-optimality. We divide the analysis into an overall measure of sub-optimality, averaging over all model densities and α, as well as restricting the averaging to only sparse generative models (model densities <0.3). The results are summarized in Table 2. The best performing inference algorithms are bolded. When taken across all model densities, in signal case 1 (Table 2, top left), the SCAD with BIC selection and SCAD with empirical Bayesian selection emerged as the best inference algorithms with respect to feature selection. When restricted to low SNR or low sample sizes (cases 2

Table 2. Table of summed deviation in selection accuracy from oracular performance. (Top) Case 1 signal conditions (SNR 10, n/p ratio 16). (Middle) Case 2 Signal Conditions (SNR 1, n/p ratio 4). (Bottom) Case 3 Signal Conditions (SNR 5 and n/p ratio 2). (Left column) All model densities. (Right column) Sparse models only. Best performers are highlighted in bold.

	Case 1, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	27.000	19.228	25.214	14.483	11.964
Lasso	23.867	14.634	27.151	5.840	5.982
MCP	23.408	**4.325**	6.948	4.717	5.220
SCAD	16.947	**3.233**	8.051	**3.534**	**4.039**
Uol Lasso	22.163	5.659	33.795	5.020	5.134
	Case 2, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp.	Bayes gMDL
EN	22.615	22.473	18.382	13.092	13.581
Lasso	22.495	19.004	19.524	16.708	14.185
MCP	26.411	13.521	**11.869**	14.382	14.671
SCAD	26.453	**11.789**	12.003	12.628	12.266
Uol Lasso	23.366	17.505	27.575	15.959	13.351
	Case 3, All Densities				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	18.290	26.087	15.339	**10.420**	12.985
Lasso	19.424	19.653	17.955	17.632	15.227
MCP	23.590	16.526	14.600	17.211	18.156
SCAD	21.125	15.241	14.119	15.007	15.873
Uol Lasso	22.030	19.866	24.080	17.420	15.268
	Case 1, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	35.139	25.146	33.098	17.533	15.371
Lasso	30.622	18.402	35.220	4.601	5.046
MCP	30.319	**1.121**	7.511	**1.033**	**2.184**
SCAD	21.267	**0.815**	9.361	**0.728**	**1.756**
Uol Lasso	29.558	3.290	44.522	3.077	3.396
	Case 2, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	29.940	18.539	25.556	16.691	13.503
Lasso	26.464	14.120	22.749	9.593	8.965
MCP	30.789	**3.879**	5.971	4.659	8.013
SCAD	31.579	**3.485**	8.154	4.213	6.434
Uol Lasso	27.151	7.287	36.588	9.150	7.024
	Case 3, Sparse Models Only				
	Selection Method				
Estimator	AIC	BIC	CV/R^2	Emp. Bayes	gMDL
EN	22.596	15.357	21.305	13.346	11.668
Lasso	20.213	10.948	18.151	8.921	8.636
MCP	24.455	**5.059**	6.754	6.695	11.546
SCAD	22.070	**5.020**	9.135	**5.884**	9.905
Uol Lasso	21.729	7.765	31.733	9.615	7.832

and 3, Tables 2 middle and bottom, left), these strategies remained amongst the best performing, with cross-validated SCAD/MCP exhibiting robust selection in case 2, the Elastic Net with empirical Bayesian selection performing the best in case 3. When restricting to sparse models only, false positive control becomes paramount, and the Elastic Net was no longer competitive. Instead, the SCAD with BIC or empirical Bayes is near optimal in case 1 (Table 2, top right), and still the best performing in cases 2 and 3 (Table 2, top and middle, right). MCP exhibited similar performance, with UoI Lasso trailing slightly behind. Thus, in general, the SCAD estimator with BIC or empirical Bayesian model selection led to the most robust algorithm for feature selection.

5. Discussion

5.1. Connections to prior work

Our numerical work corroborates and extends several results from the statistical literature in a non-asymptotic setting. We found the frequently employed cross-validated Lasso to be amongst the worst performing selection strategies. It has been shown that using predictive performance as a criteria for regularization strength selection with the Lasso leads to inconsistent support recovery (Leng, Lin, and Wahba 2006). A necessary and sufficient condition for asymptotically consistent model selection by the Lasso is for the irrepresentable condition to hold (Zhao and Yu 2006). In the non-asymptotic setting of this study, we find that the parameter α is a more useful modulator of selection accuracy, and that the irrepresentable constant of Zhao and Yu (2006) tracks the selection accuracy of Lasso only insofar as it tracks α (section S5). We find that the SCAD/MCP and UoI Lasso select model supports more robustly in the presence of correlated design. It is known that the SCAD/MCP do not require any strong conditions on the design matrix for oracular properties to hold (Loh and Wainwright 2017), and neither does the BoLasso (Bach 2008), upon which the selection logic of UoI is partially based on. Our work demonstrates that the choice of model selection criteria is as important as the choice of estimator to achieve good selection accuracy. The model selection criteria we have considered can all be categorized as penalized likelihood methods. Cross-validation is known to behave asymptotically like the AIC (Shao 1997). The magnitude of this complexity penalty can be interpreted as a prior on the model size. We correspondingly find that the BIC performs best in sparse models, whereas the AIC and CV perform best in dense models. The tension between the BIC and AIC has been noted in the literature (Yang 2005). The asymptotic selection consistency of using BIC to select SCAD regularization strength has been noted in Wang, Li, and Tsai (2007). Our numerical investigations reveal that this remains one of the best extant selection strategies in non-asymptotic settings with mild correlated variability as well.

The empirical Bayesian and gMDL procedures were devised with complexity penalties nominally adaptive to the underlying model density. We find that these methods lead to good model selection performance across model densities, but only in ideal signal conditions (i.e., case 1) and low design matrix correlations. There is therefore possible room for methodological development of adaptive complexity penalties. We leave this for future work.

5.2. Best practices in real data

Proper model selection is essential for interpretability of parametric models. While sufficient conditions for model selection are available in the literature, they do not provide actionable results for the practitioner in real data. Our extensive numerical simulations reveal best practices. Nonconvex optimization estimators such as the SCAD and MCP generically perform better at selection than the Lasso and Elastic Net when the underlying model is sparse. This in line with both prior numerical work and the understanding that asymptotically, these estimators are oracular

selectors (Fan and Li 2001; Zhang 2010). Our work reveals that this performance gap remains even as design matrices become increasingly correlated. While the SCAD and MCP are nonconvex problems, recent work has shown that the statistical performance of all stationary points is nearly equivalent (Loh and Wainwright 2015). Furthermore, development of the optimization algorithms for these estimators has matured to the point where regularization paths for the SCAD and MCP can be computed in the same order of magnitude of time as the Lasso/Elastic Net (see for e.g., Zhao, Liu, and Zhang 2018). Our work provides further motivation for the adoption of these algorithms. The $\text{UoI}_{\text{Lasso}}$ algorithm has selection performance competitive with MCP and SCAD in many cases. Furthermore, as we show in section S4, the OLS-bagging procedure used in coefficient estimates in UoI leads to lower bias/variance estimates than SCAD/MCP.

There is a tradeoff between false positive and false negative control achieved by model selection strategies. False positive control is largely insensitive to the degree of design correlation. Practitioners seeking tight control of false negatives in model selection may be inclined to use the Elastic Net estimator. The presence of a number of fairly generic α transition points after which selection accuracy degrades, and false negative & positive magnitude inflates suggests a heuristic criteria that could be estimated from the sample covariance. Specifically, combining empirical estimates of the precision matrix with empirical estimates of β_{min} and σ^2 allows one to estimate α, and therefore have a rough sense of whether selection and estimation performance is likely to have degraded due to correlated covariates or low signal strength.

6. Conclusions and future work

Our empirical results reveal that the joint choice of sparse estimator and model selection criteria significantly modulates selection performance. Nevertheless, with the exception of the previously mentioned (Wang, Li, and Tsai 2007), theoretical results that capture non-asymptotic behavior of regularization strength selection via specific model selection criteria are lacking.

We found no inference algorithm to be dominant across underlying model density in the presence of correlated covariates, including the nominally adaptive empirical Bayes and gMDL selection criteria. Whether these reflect information theoretic constraints or methodological gaps is a potentially avenue of future work. We also believe our observation of a universal α-transition point across false negatives and coefficient distortion to be novel. This phenomena is reminiscent of the well-known reconstructability transition in compressed sensing as a function of noise level and sampling density (Donoho, Maleki, and Montanari 2009). An average case analysis of coefficient support distortion as a function of α or other spectral parameters of the design matrix will be the topic of future work.

Funding

This work was supported by NIH/NINDS under R01NS118648; DOE/ASCR under AWD00003162. The authors have no conflicts of interest to disclose.

References

Aeron, S., V. Saligrama, and M. Zhao. 2010. Information theoretic bounds for compressed sensing. *IEEE Transactions on Information Theory* 56 (10):5111–30. doi:10.1109/TIT.2010.2059891.

Aksoylar, C., G. K. Atia, and V. Saligrama. 2017. Sparse signal processing with linear and nonlinear observations: A unified Shannon-theoretic approach. *IEEE Transactions on Information Theory* 63 (2):749–76. doi:10.1109/TIT.2016.2605122.

Aksoylar, C., and V. Saligrama. 2014. Information-theoretic characterization of sparse recovery. In Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, April, edited by Samuel Kaski and Jukka Corander, 38–46. PMLR.

Atia, G. K., and V. Saligrama. 2009. Boolean compressed sensing and noisy group testing. arXiv:0907.1061.

Bach, F. R. 2008. Bolasso: Model consistent Lasso estimation through the Bootstrap. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, Helsinki, Finland, 33–40. New York, NY: Association for Computing Machinery.

Bertsimas, D., A. King, and R. Mazumder. 2016. Best subset selection via a modern optimization lens. *The Annals of Statistics* 44 (2):813–52. doi:10.1214/15-AOS1388.

Bertsimas, D., J. Pauphilet, and B. Van Parys. 2020. Sparse regression: Scalable algorithms and empirical performance. *Statistical Science* 35 (4):555–78.

Bogdan, M., E. van den Berg, W. Su, and E. Candes. 2013. Statistical estimation and testing via the sorted L1 norm. arXiv:1310.1969.

Bouchard, K., A., Bujan, F. Roosta-Khorasani, S. Ubaru, M. Prabhat, A. Snijders, J.-H. Mao, E. Chang, M. W. Mahoney, and a S. Bhattacharya. 2017. Union of intersections (UoI) for interpretable data driven discovery and prediction. In *Advances in neural information processing systems 30*, eds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 1078–86. Red Hook, NY: Curran Associates, Inc.

Brewer, M. J., A. Butler, and S. L. Cooksley. 2016. The relative performance of AIC, AICC and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution* 7 (6):679–92. doi:10.1111/2041-210X.12541.

Bühlmann, P., P. Rütimann, S. van de Geer, and C.-H. Zhang. 2013. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference* 143 (11):1835–58. doi:10.1016/j.jspi.2013.05.019.

Donoho, D. L., A. Maleki, and A. Montanari. 2009. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences* 106 (45):18914–9. doi:10.1073/pnas.0909892106.

Dormann, C. F., J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J. R. G. Marquéz, B. Gruber, B. Lafourcade, P. J. Leitão, et al. 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36 (1):27–46. doi:10.1111/j.1600-0587.2012.07348.x.

Dziak, J. J., D. L. Coffman, S. T. Lanza, R. Li, and L. S. Jermiin. 2020. Sensitivity and specificity of information criteria. *Briefings in Bioinformatics* 21 (2):553–65. doi:10.1093/bib/bbz016.

Fan, J., and R. Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348–60. doi:10.1198/016214501753382273.

Figueiredo, M. A. T., and R. D. Nowak. 2016. Ordered weighted L1 regularized regression with strongly correlated covariates: Theoretical aspects. In AISTATS.

George, E. I., and D. P. Foster. 2000. Calibration and empirical Bayes variable selection. *Biometrika* 87 (4):731–47. doi:10.1093/biomet/87.4.731.

Hansen, M. H., and B. Yu. 2001. Model selection and the principle of minimum description length. *Journal of the American Statistical Association* 96 (454):746–74. doi:10.1198/016214501753168398.

Kim, Y., S. Kwon, and H. Choi. 2012. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research* 13 (36):1037–57.

Leng, C., Y. Lin, and G. Wahba. 2006. A note on the Lasso and related procedures in model selection. *Statistica Sinica* 16 (4):1273–84.

Li, Y., B. Mark, G. Raskutti, and R. Willett. 2018. Graphbased regularization for regression problems with highly-correlated designs. arXiv:1803.07658.

Loh, P.-L., and M. J. Wainwright. 2015. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* 16 (1):559–616.

Loh, P.-L., and M. J. Wainwright. 2017. Support recovery without incoherence: A case for nonconvex regularization. *The Annals of Statistics* 45 (6):2455–82. doi:10.1214/16-AOS1530.

Rad, K. R. 2011. Nearly sharp sufficient conditions on exact sparsity pattern recovery. *IEEE Transactions on Information Theory* 57 (7):4672–9.

Satija, R., J. A. Farrell, D. Gennert, A. F. Schier, and A. Regev. 2015. Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology* 33 (5):495–502. doi:10.1038/nbt.3192.

Scarlett, J., and V. Cevher. 2017. Limits on support recovery with probabilistic models: An information-theoretic framework. *IEEE Transactions on Information Theory* 63 (1):593–620. doi:10.1109/TIT.2016.2606605.

Scarlett, J., J. S. Evans, and S. Dey. 2013. Compressed sensing with prior information: Information-theoretic limits and practical decoders. *IEEE Transactions on Signal Processing* 61 (2):427–39. doi:10.1109/TSP.2012.2225051.

Schöniger, A., T. Wöhling, L. Samaniego, and W. Nowak. 2014. Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resources Research* 50 (12):9484–513. doi:10.1002/2014WR016062.

Shao, J. 1997. An asymptotic theory for linear model selection. *Statistica Sinica* 7 (2):221–42.

Steyerberg, E. W., and Y. Vergouwe. 2014. Towards better clinical prediction models: Seven steps for development and an ABCD for validation. *European Heart Journal* 35 (29):1925–31. doi:10.1093/eurheartj/ehu207.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1):267–88. doi:10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R. 1997. The Lasso method for variable selection in the Cox model. *Statistics in Medicine* 16 (4): 385–95. doi:10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (1):91–108. doi:10.1111/j.1467-9868.2005.00490.x.

Varga, R. S. 2004. Geršgorin-type eigenvalue inclusion theorems. In *Geršgorin and his circles*, eds. R. S. Varga, 35–72. Berlin, Heidelberg: Springer Berlin Heidelberg.

Vatcheva, K. P., MJae Lee, J. B. McCormick, and M. H. Rahbar. 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale, Calif.)* 6 (2):227.

Wainwright, M. J. 2009. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory* 55 (12):5728–41. doi:10.1109/TIT.2009.2032816.

Wainwright, M. J. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using $l1$ -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory* 55 (5):2183–202. doi:10.1109/TIT.2009.2016018.

Waldmann, P., G. Mészáros, B. Gredler, C. Fuerst, and J. Sölkner. 2013. Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in Genetics* 4:270. doi:10.3389/fgene.2013.00270.

Wang, H., R. Li, and C.-L. Tsai. 2007. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94 (3):553–68. doi:10.1093/biomet/asm053.

Witten, D. M., A. Shojaie, and F. Zhang. 2014. The cluster elastic net for high-dimensional regression with unknown variable grouping. *Technometrics* 56 (1):112–22. doi:10.1080/00401706.2013.810174.

Wright, J., Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan. 2010. Sparse representation for computer vision and pattern recognition. *Proceedings of the IEEE* 98 (6):1031–44. doi:10.1109/JPROC.2010.2044470.

Yang, Y. 2005. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* 92 (4):937–50. doi:10.1093/biomet/92.4.937.

Zhang, C.-H. 2010. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38 (2):894–942. doi:10.1214/09-AOS729.

Zhao, P., and B. Yu. 2006. On model selection consistency of Lasso. *Journal of Machine Learning Research* 7: 2541–63.

Zhao, T., H. Liu, and T. Zhang. 2018. Pathwise coordinate optimization for sparse learning: Algorithm and theory. *The Annals of Statistics* 46 (1):180–218. doi:10.1214/17-AOS1547.

Zhu, J., C. Wen, J. Zhu, H. Zhang, and X. Wang. 2020. A polynomial algorithm for best-subset selection problem. *Proceedings of the National Academy of Sciences* 117 (52):33117–23. doi:10.1073/pnas.2014241117.

Zou, H., and T. Hastie. 2005. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.