Journal of Data Science 19 (3), 429–449
July 2021
DOI: 10.6339/21-JDS1015
Statistical Data Science

Sign-based Shrinkage Based on an Asymmetric LASSO Penalty

Eric S. Kawaguchi[1,*], Burcu F. Darst[1], Kan Wang[2], and David V. Conti[1]

[1]*Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA*
[2]*Google, Mountain View, California, USA*

Abstract

Penalized regression provides an automated approach to preform simultaneous variable selection and parameter estimation and is a popular method to analyze high-dimensional data. Since the conception of the LASSO in the mid-to-late 1990s, extensive research has been done to improve penalized regression. The LASSO, and several of its variations, performs penalization symmetrically around zero. Thus, variables with the same magnitude are shrunk the same regardless of the direction of effect. To the best of our knowledge, sign-based shrinkage, preferential shrinkage based on the sign of the coefficients, has yet to be explored under the LASSO framework. We propose a generalization to the LASSO, asymmetric LASSO, that performs sign-based shrinkage. Our method is motivated by placing an asymmetric Laplace prior on the regression coefficients, rather than a symmetric Laplace prior. This corresponds to an asymmetric ℓ_1 penalty under the penalized regression framework. In doing so, preferential shrinkage can be performed through an auxiliary tuning parameter that controls the degree of asymmetry. Our numerical studies indicate that the asymmetric LASSO performs better than the LASSO when effect sizes are sign skewed. Furthermore, in the presence of positively-skewed effects, the asymmetric LASSO is comparable to the non-negative LASSO without the need to place an *a priori* constraint on the effect estimates and outperforms the non-negative LASSO when negative effects are also present in the model. A real data example using the breast cancer gene expression data from The Cancer Genome Atlas is also provided, where the asymmetric LASSO identifies two potentially novel gene expressions that are associated with *BRCA1* with a minor improvement in prediction performance over the LASSO and non-negative LASSO.

Keywords *asymmetric Laplace distribution; high-dimensional statistics; penalized regression; quantile regularization; variable selection*

1 Introduction

Recent developments in data acquisition, collection, and storage have allowed researchers to obtain a large number of potential predictors in order to avoid missing important factors that may be associated with the outcome of interest. This is often the case in genomic studies, where the number of predictors collected is often larger than the sample size. Simultaneous variable selection and parameter estimation is an essential task in high-dimensional data analysis that aims to identify a smaller subset of important variables. Penalized regression methods accomplish this by shrinking the regression coefficients toward zero while setting some coefficients equal to zero. These methods estimate a sparse vector of regression coefficients by minimizing an objective function that is composed of both a loss function and a penalty function.

*Corresponding author. Email: eric.kawaguchi@med.usc.edu.

Received March 16, 2021; Accepted May 16, 2021

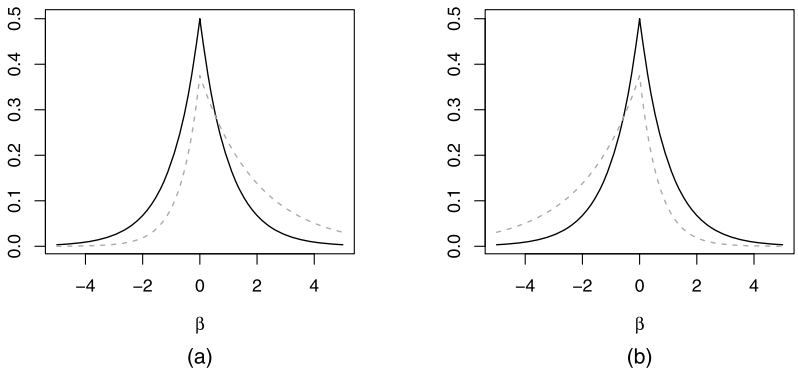

Figure 1: Density of the asymmetric Laplace distribution. Black solid line corresponds to the standard Laplace distribution ($\tau = 0.5$). Dotted grey lines correspond to (a): $\tau = 0.25$ and (b): $\tau = 0.75$.

One of, if not, the most popular penalized regression methods is the LASSO (Tibshirani, 1996). Since its conception in the mid-to-late 1990's, the LASSO framework has been extensively used in several different research areas including, but not limited to, signal processing (Angelosante et al., 2009), genomic studies (Huang and Pan, 2003; Ghosh and Chinnaiyan, 2005; Wu and Lange, 2008; Wu et al., 2009, 2011), finance (Wu et al., 2014; Pereira et al., 2016; Panagiotidis et al., 2018), and text mining (Li et al., 2014; Debortoli et al., 2016). LASSO performs estimation and selection by forcing the sum of the absolute value of the regression coefficients (the ℓ_1 norm) to be less than a non-negative fixed value which, consequently, forces some of the coefficients to zero. From a Bayesian perspective, the LASSO is motivated by placing a Laplace prior on the regression coefficients (see e.g., Tibshirani, 1996; Park and Casella, 2008; Hans, 2009). The density of the Laplace distribution is provided in Figure 1 (solid black line). The prior is symmetric around 0 implying that the degree of shrinkage for a particular magnitude is the same regardless of the direction of effect. Several extensions and improvements, both in estimation and computation, to the LASSO have been proposed in the literature (see e.g., Tibshirani, 1997; Fan and Li, 2001; Zou and Hastie, 2005; Tibshirani et al., 2005; Yuan and Lin, 2006; Meinshausen and Bühlmann, 2006; Friedman et al., 2008; Zou, 2006; Wu and Lange, 2008; Friedman et al., 2010; Zhang et al., 2010; Tibshirani et al., 2012).

While extensive research has been done to improve penalization by shrinking the *magnitude* of the coefficients differently (Zou, 2006), to the best of our knowledge, preferential shrinkage based on the sign of the coefficients has yet to be explored. Traditionally, LASSO and other penalized regression procedures shrink variables symmetrically around 0. That is, the degree of shrinkage for a particular magnitude is the same regardless of the direction of effect. There are several motivating scientific questions in which shrinking both positive and negative coefficients equally may not be preferred in certain situations. Most likely these studies leverage some previous knowledge in which we expect effects to be favored in one direction over another or if we are particularly interested in one effect direction. For example, *BRCA1* is a DNA damage repair gene that has been shown to have strong associations with breast and ovarian cancer risk (Welcsh et al., 2000; Welcsh and King, 2001). Based on this knowledge, we may be interested in identifying additional genes that are associated with *BRCA1* expression; in particular, genes with a positive association. These genes could implicate mechanisms through which *BRCA1* impacts cancer risk and warrant further investigation in future studies. Since we are interested

in identifying genes that have an elevated effect on *BRCA1*, we may want to focus our attention to selecting positive effects, while also allowing for the possibility of identifying strong negative effects. Additionally, in certain genomic studies designed for the construction of a polygenic risk score (PRS), there is an emphasis on identifying risk variants for certain diseases (Khera et al., 2018). Variable selection procedures, such as the LASSO, are often employed to identify relevant markers that are used in developing a PRS. Finally, certain biomarkers within known biological pathways may be suspected to be associated with elevated risk (i.e., a positive association with the disease outcome). In a metabolomic investigation we may be particularly interested in discovering additional biomarkers with smaller risk effects which may help elucidate the biological mechanisms of the disease.

Generalizations to the LASSO, such as the constrained LASSO, have been developed to augment the standard LASSO with linear equality and inequality constraints (Efron et al., 2004; James et al., 2012; Tibshirani and Taylor, 2011; Wu et al., 2014; Gaines et al., 2018). The non-negative LASSO is an example of the constrained LASSO that requires the LASSO coefficients to be nonnegative. At first glance, this formulation seems to solve the issue of preferential shrinkage since it forces effect estimates to be positive. However, these linear constraints must be specified *a priori* and can be problematic if negative effects are present. It would be ideal to develop a LASSO-based variable selection method that can perform preferential shrinkage without the need to place *a priori* constraints on the parameter space.

Motivated by this idea, we propose a new variation of LASSO penalization that accomplishes asymmetric shrinkage. Our proposed method, asymmetric LASSO, replaces the standard ℓ_1 penalty with an asymmetric ℓ_1 penalty. In doing so, the asymmetric LASSO performs preferential shrinkage through an auxiliary tuning parameter that controls the degree of asymmetry. While estimation is focused under a penalized regression framework, we provide a Bayesian interpretation that motivates the use of the asymmetric ℓ_1 penalty. Specifically, one can view the asymmetric LASSO as placing an asymmetric Laplace prior on the regression coefficients. We also show that the standard LASSO is a special case of the asymmetric LASSO. Since the objective function is convex, we employ an efficient optimization algorithm for our implementation.

The paper is organized as follows. In Section 2 we introduce the asymmetric LASSO under a generalized linear model framework. We provide insight into the behavior of the estimator under the ordinary least squares model with orthogonal design. Simulation studies are provided in Section 3 to explore the empirical properties of asymLASSO and compare its performance to both the traditional LASSO and non-negative LASSO across several scenarios. We provide a real data example using the breast cancer gene expression data from The Cancer Genome Atlas (TCGA) in Section 4. Finally, parting comments and future directions and discussed in Section 5.

2 Methodology

2.1 The Asymmetric LASSO

Let us consider the generalized linear model (GLM) with a response vector $\mathbf{y}$ and design matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, assume that the observations $\mathbf{v}_i = (\mathbf{x}_i^T, y_i)^T$, $i = 1, \ldots, n$, are mutually independent, and that, conditional on $\mathbf{x}_i$, y_i belongs to the exponential family with the following density

$$f_Y(y; \mathbf{x}, \phi) = \exp\left\{\frac{y\theta - a(\theta)}{b(\phi)} - c(y, \phi)\right\}, \tag{1}$$

where θ is defined as the canonical parameter, $\phi > 0$ is the scale (dispersion) parameter and $a(\phi)$, $b(\theta)$, and $c(y, \phi)$ are known functions whose values depend on the distribution (McCullagh and Nelder, 1983; Dobson and Barnett, 2018). If we assume that $a(\cdot)$ is twice differentiable, then Model (1) indicates that $E(y_i|\mathbf{x}_i) = \mu_i = a'(\theta_i)$ and $var(y_i|\mathbf{x}_i) = a''(\theta_i)b(\phi_i)$. Furthermore, the canonical parameter θ is connected to $\mathbf{x}_i$ through a prespecified link function $h(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ for some $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. Examples of commonly used GLMs with canonical link include linear regression, logistic regression, and Poisson regression. We can now define the likelihood function for $\boldsymbol{\beta}$,

$$L(\boldsymbol{\beta}; \mathbf{v}_i) \propto \prod_{i=1}^{n} \exp(y_i\theta_i - a(\theta_i)). \tag{2}$$

Consequently, the log-likelihood is defined as $l(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}; \mathbf{v}_i)$. The regression coefficients $\boldsymbol{\beta}$ are typically estimated through minimizing the negated log-likelihood function. Typically, not all of the p covariates that are included in the data are associated with the outcome and interest lies in estimating a sparse $\boldsymbol{\beta}$ (i.e., several values of $\boldsymbol{\beta}$ are 0). This is especially the case in the high-dimensional ($p > n$) setting. Penalized regression provides an automated approach to perform simultaneous variable selection and parameter estimation.

To conceptualize asymmetric penalization, we motivate the idea under a Bayesian framework where we propose to model the regression coefficients using an asymmetric Laplace prior

$$\pi(\beta_j|\lambda, \tau) = 2\lambda\tau(1-\tau)\exp\{-2\lambda\beta_j(\tau - I(\beta_j < 0))\}, \tag{3}$$

where $\lambda \geqslant 0$ is the scale parameter and $\tau \in (0, 1)$ is the skewness parameter that controls the asymmetry. Two examples of the asymmetric Laplace distribution are provided in Figure 1 (dotted grey line) for $\tau = 0.25$ (Figure 1a) and $\tau = 0.75$ (Figure 1b). In both figures, we see that the distribution is still concentrated at 0; however, the behavior of the tails is asymmetric. When $\tau = 0.25$, the left and right tail of the density are narrower and wider than the standard Laplace distribution, respectively. More mass is reserved for positive-valued $\boldsymbol{\beta}$ than for negative-valued $\boldsymbol{\beta}$. The converse is true when $\tau = 0.75$. We can allow that data to dictate the choice of τ, allowing us to perform sign-dependent shrinkage in a data-driven manner rather than a prespecified constraint as in the constrained LASSO.

In the context of penalized regression, the LASSO estimates are obtained by minimizing an objective function that is composed of the negated log-likelihood function plus an ℓ_1 penalty function. It is easy to show that the ℓ_1 penalty, $|\boldsymbol{\beta}|$, is proportional to the negated log-density of the standard Laplace distribution. Like LASSO, imposing an asymmetric Laplace prior on $\boldsymbol{\beta}$ has a direct correspondence to estimation using a penalized likelihood. By rewriting the check function $f(x) = x(\tau - I(x < 0)) = (|x| + (2\tau - 1)x)/2$, the asymmetric Laplace distribution will correspond to an asymmetric ℓ_1 penalty, $|\beta_j| + (2\tau - 1)\beta_j$ and our asymmetric LASSO (asymLASSO) estimator is defined as

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min_{\boldsymbol{\beta}} \left\{ -l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \left(|\beta_j| + (2\tau - 1)\beta_j \right) \right\}. \tag{4}$$

The use of the check function as the basis for the penalization term in (4) and placing an asymmetric Laplace prior on $\boldsymbol{\beta}$ are intrinsically connected to quantile estimation and quantile regression (Koenker and Basset, 1978; Yu and Moyeed, 2001; Yu and Zhang, 2005; Kozumi and Kobayashi, 2011; Takeuchi et al., 2006). Specifically if $\boldsymbol{\beta}$ follows an asymmetric Laplace

distribution with location parameter 0, scale parameter $1/(2\lambda)$, and skew parameter τ, as in (3), then $\Pr(\beta < 0) = \tau$ and $\Pr(\beta > 0) = 1 - \tau$, and therefore 0 can be interpreted as the τ-th quantile of the distribution. In the following section we show how τ impacts estimation when compared to the standard LASSO.

2.2 The Behavior of asymLASSO Under Orthogonal Design

To better understand the behavior of asymLASSO, we investigate the OLS model under an orthogonal design matrix (i.e., $X^T X = I_p$ with $p < n$). Under these conditions, asymLASSO leads to the following closed-form solution

$$\hat{\beta}_j(ols; \tau) = S(\hat{\beta}_j^{ols} - \lambda(2\tau - 1), \lambda), \tag{5}$$

where $S(a, b) = sgn(a)(|a| - b)_+$ is the soft-thresholding operator (Donoho and Johnstone, 1994) defined for $\lambda \geqslant 0$ and $\hat{\beta}_j^{ols} = \mathbf{x}_j^T \mathbf{y}$ is the OLS estimate. Equation (5) follows a modified version of the LASSO and, in fact, is equivalent to the LASSO when $\tau = 1/2$. Thus LASSO can be viewed as a special case of asymLASSO. Furthermore, as $\lambda \to 0$ we have that $\hat{\beta}_j(ols; \tau) \to \hat{\beta}_j^{ols}$, which implies that if $\lambda = o(1)$, then $\hat{\boldsymbol{\beta}}(ols; \tau)$ is a consistent estimator for all $\tau \in (0, 1)$.

Figure 2 illustrates the behavior of the asymmetric LASSO soft-thresholding operator provided in Equation (5) under orthogonal design with $\tau = 0.25, 0.5$ and 0.75. We can compare the three panels in terms of their effect on both bias and sparsity. Note that $\hat{\beta}_j(ols; \tau) = 0$ whenever $-2\lambda(1 - \tau) \leqslant \hat{\beta}_j^{ols} \leqslant 2\lambda\tau$. As discussed earlier, asymLASSO with $\tau = 0.5$ (Figure 2a) reduces to the LASSO. In this panel we see that $\hat{\beta}_j(ols; 0.5) = 0$ whenever $-\lambda \leqslant \hat{\beta}_j^{ols} \leqslant \lambda$. Furthermore, the nonzero values are penalized by a constant factor, λ, as indicated by the difference between the dotted gray line (true value of $\boldsymbol{\beta}$) and solid black line (asymLASSO shrinkage). Figure 2b illustrates asymLASSO with $\tau = 0.25$ and we see that the thresholding function is shifted to the right. When compared to the LASSO (Figure 2a) positive-valued estimates will be less biased and less likely to be shrunk to 0 compared to negative-valued estimates of the same magnitude. Therefore, scenarios where we expect more positive-valued (and smaller) effect estimates will benefit from asymLASSO over the standard LASSO. We see the opposite relationship in Figure 2c where we set $\tau = 0.75$. In this situation, asymLASSO favors negative-valued effect estimate over positive-valued effect estimates. Under the orthogonal design, we can explicitly quantify the shrinkage seen in Figure 2 for general τ. To understand these properties better, we can think about the solution path as two components:

Case 1: $\hat{\beta}_j^{ols} > 0$. For this case we are only concerned with the positive estimates of asymLASSO. Here $\hat{\beta}_j(ols; \tau) = 0$ whenever $\hat{\beta}_j^{ols} \in [0, 2\lambda\tau]$. Hence when $\tau < 1/2$, $\hat{\beta}_j(ols; \tau)$ is shrunk to 0 over a smaller interval than LASSO. In fact, estimates where $\hat{\beta}_j^{ols} \in (2\lambda\tau, \lambda]$ will be 0 for LASSO and nonzero for $\hat{\beta}_j(ols; \tau)$. Therefore, asymLASSO with $\tau < 1/2$ will select smaller positive effect estimates than the LASSO. When $\hat{\beta}_j^{ols} > 2\lambda\tau$, $\hat{\beta}_j(ols; \tau) = \hat{\beta}_j^{ols} - 2\lambda\tau$. Again when $\tau < 1/2$, $\hat{\beta}_j^{ols} < \hat{\beta}_j^{ols} - 2\lambda\tau < \hat{\beta}_j^{ols} - \lambda$, and asymLASSO provides a less biased estimate when compared to LASSO. However when $\tau > 1/2$, we have $\lambda < 2\lambda\tau$ and therefore asymLASSO tends to overshrink and produce more biased estimates compared to LASSO.

Case 2: $\hat{\beta}_j^{ols} < 0$. Focusing on negative estimates of asymLASSO, $\hat{\beta}_j(ols; \tau) = 0$ whenever $\hat{\beta}_j^{ols} \in [-2\lambda(1 - \tau), 0]$ and $\hat{\beta}_j(ols; \tau) = \hat{\beta}_j^{ols} + 2\lambda(1 - \tau)$ when $\hat{\beta}_j^{ols} < -2\lambda(1 - \tau)$. For $\tau < 1/2$ we have $\lambda < 2\lambda(1 - \tau)$ and when $\tau > 1/2$, $2\lambda(1 - \tau) < \lambda$. Therefore asymLASSO shrinks smaller negative effects to 0 and produces more biased estimates compared to LASSO for $\tau < 1/2$ and retains smaller negative effects and produces less biased estimates for $\tau > 1/2$.

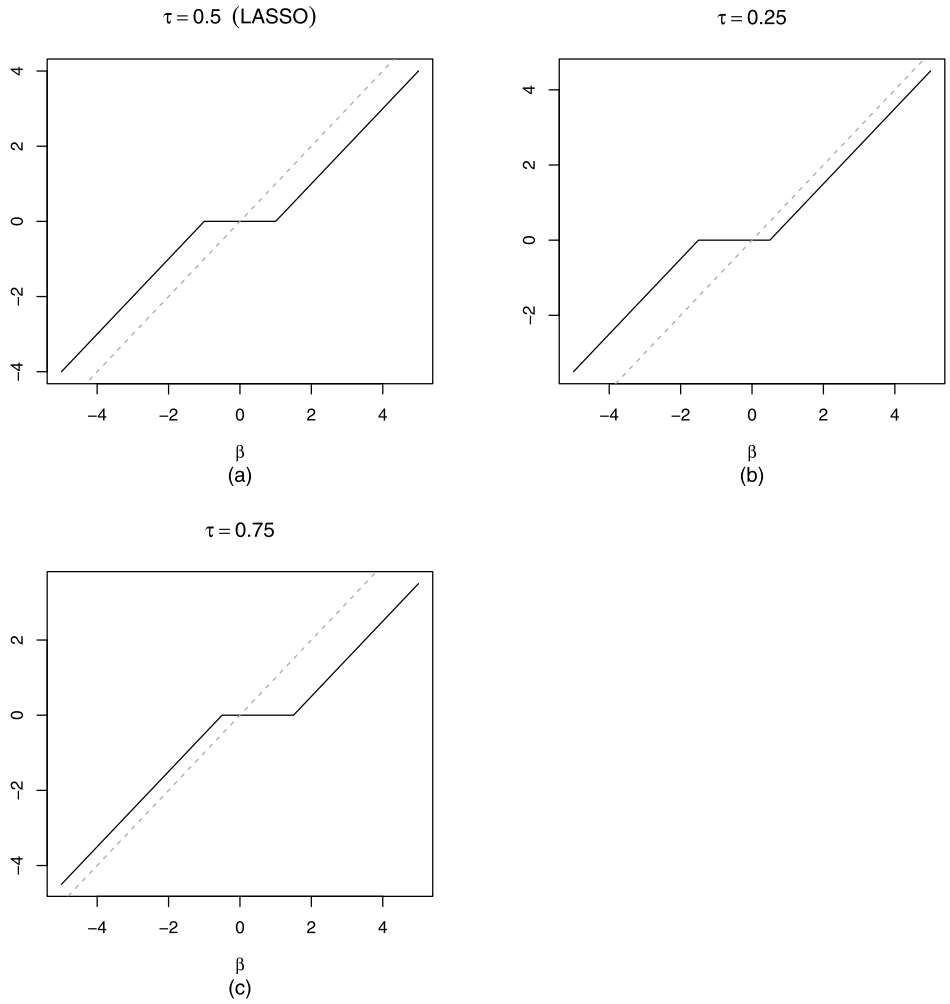

Figure 2: Behavior of the soft-thresholding function for ASYMLASSO under orthogonal design for: (a) $\tau = 0.5$ (LASSO), (b) $\tau = 0.25$, and (c) $\tau = 0.75$. The dotted grey line represents the true value of $\boldsymbol{\beta}$. Symmetry of sparsity and bias are dependent on the value of τ.

If we relax the orthogonal design matrix condition, we also have the following results for the ASYMLASSO estimator under the ordinary least squares model.

Lemma 2.1. *Suppose* $l(\boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2$, *as in the ordinary least squares model and let* A *be the event that* $\{\lambda(1 + |2\tau - 1|) \geqslant \left\|\boldsymbol{\epsilon}^T X\right\|_\infty\}$, *where* $\boldsymbol{\epsilon} = \mathbf{y} - X\boldsymbol{\beta}$ *and* $\|\cdot\|_\infty$ *is the uniform norm. Defining* $\hat{\boldsymbol{\beta}}$ *as the solution to (4), if* A *holds, then*

$$\left\|X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right\|_2^2 \leqslant 4(1 + |2\tau - 1|)\lambda \|\boldsymbol{\beta}\|_1 ,$$

under mild regularity conditions.

The proof is provided in the Online Supplementary Material and mirrors similarly to the proof for the ordinary LASSO estimator. Note that $(1 + |2\tau - 1|) \in (1, 2)$ and equals one when $\tau = 1/2$. Therefore these bounds can be larger (up to a constant) than the bounds for ordinary

LASSO estimator. Consequently, if $X^TX = I_p$, then we also have $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2 \leqslant 4(1+|2\tau-1|)\lambda\,\|\boldsymbol{\beta}\|_1$ since $\left\|X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\right\|_2 = \sqrt{(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T X^T X(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})} = \left\|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\right\|_2$.

2.3 Implementation via Cyclic Coordinate Descent

For notational convenience, we suppress the dependence of τ in $\hat{\boldsymbol{\beta}}$. Letting $\nabla l(\boldsymbol{\beta}) = \partial l(\boldsymbol{\beta})/\partial\boldsymbol{\beta} = X^T\mathbf{u}$ and $\nabla^2 l(\boldsymbol{\beta}) = \partial^2 l(\boldsymbol{\beta})/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T = X^T W X$, we approximate the log-likelihood based on a Taylor series expansion about the current iteration $\boldsymbol{\beta}^{(m)}$:

$$l(\boldsymbol{\beta}) \approx \frac{1}{2}(\tilde{\mathbf{y}} - X\boldsymbol{\beta})^T W(\tilde{\mathbf{y}} - X\boldsymbol{\beta}),$$

where $\tilde{\mathbf{y}}$ is the working response vector $\tilde{\mathbf{y}} = X\boldsymbol{\beta}^{(m)} + W^{-1}\mathbf{u}$. Note here that $\mathbf{u}$, W, and $\tilde{\mathbf{y}}$ are dependent on $\boldsymbol{\beta}^{(m)}$. With this approximation, efficient convex optimization algorithms can be used to minimize (4). We employ cyclic coordinate descent, a widely-used algorithm for penalization (Wu and Lange, 2008; Friedman et al., 2010; Breheny and Huang, 2011), for our implementation.

The algorithm starts by setting all p variables to some initial value (e.g. $\boldsymbol{\beta}^{(0)} = \mathbf{0}$). It then solves a one-dimensional optimization problem by setting the first variable ($j = 1$) to a value that minimizes the objective function while holding all other variables constant. This process is repeated for the second variable, third variable, and so on. When the algorithm cycles through all the variables, it returns to the first variable and repeats the cycling process until some convergence criterion is met. For asymLASSO, the one-dimensional update for the jth covariate at the $(m+1)^{th}$ iteration is

$$\beta_j^{(m+1)} \leftarrow \frac{S(r_j - \lambda(2\tau-1), \lambda)}{v_j}, \tag{6}$$

where v_j is the jth diagonal element of $V = X^T W X$ and r_j is the jth element of $\mathbf{r} = X^T W\mathbf{u} + V\boldsymbol{\beta}^{(m)}$.

Typically, we are interested in obtaining estimates for $\hat{\boldsymbol{\beta}}$ over a range of values between a maximum value λ_{max} for which all coefficient estimates are 0 to a minimum value λ_{min} at which the model becomes excessively large (saturated) or ceases to be identifiable. For the LASSO, $\lambda_{max} = \max_j\{|r_j|\}$ when the quadratic approximation is taken with respect to the intercept-only model (Friedman et al., 2010). This is due to the fact that the LASSO estimates are zero whenever $|r_j| \leqslant \lambda$ for all j. The asymLASSO estimates, however, are zero whenever $|r_j - \lambda(2\tau-1)| \leqslant \lambda$. This complicates finding a value for λ_{max} since shrinkage is not symmetric about 0. We propose to use a conservative value for λ_{max} given by $\lambda_{max} = \max_j\left\{\frac{|r_j|}{2\tau}, \frac{|r_j|}{2(1-\tau)}\right\}$. This bound is equivalent to the LASSO bound when $\tau = 1/2$ and larger otherwise.

2.4 Selection of τ and λ

Model complexity depends critically on the choice of the tuning parameters. As evident in Section 2.2, τ induces a "sign-specific shrinkage tradeoff" that determines whether emphasis is placed on shrinking positive or negative-valued effects. While one can consider specific biological scenarios in which τ can be selected a priori, as shown in Section 3.2, a naively prespecified value can lead to biased estimation and improper shrinkage. The penalization parameter λ dictates the degree of shrinkage and therefore must be carefully selected. In practice, one generally implements a penalization method across a grid of tuning parameters and selects the tuning

parameter that minimizes some criterion. Since estimating both τ and λ is of interest, we use a two-dimensional grid search to select the optimal pair $(\tau^{opt}, \lambda^{opt})$. Several criteria have been proposed in the literature including, but not limited to, k-fold cross validation, generalized cross validation (Golub et al., 1979), the Akaike information criterion (Akaike, 1974) and the Bayesian information criterion (Schwarz et al., 1978).

3 Numerical Studies

A series of simulations are conducted to illustrate the performance of asymLASSO under various design settings. All computations are carried out using the R programming language. The design matrix $X = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)$ is generated from a multivariate Gaussian distribution with mean 0 and variance-covariance matrix Σ. We allowed for mild correlation between covariates by specifying an autoregressive covariance structure, $\Sigma = 0.5^{|i-j|}$. The data are generated from a normal linear model via $\mathbf{y}|\mathbf{x} \sim N(\mu + \mathbf{x}^T\boldsymbol{\beta}^*, \sigma_y^2)$, where μ is the intercept term. Clarification of the simulation parameters, such as the structure of $\boldsymbol{\beta}^*$, is provided in the corresponding subsections.

3.1 Sensitivity to τ

As noted earlier Section 2.4, preferential shrinkage of positive or negative effects is dictated by τ. We investigate the effect τ has on the selection performance of asymLASSO. We used an evenly-spaced grid on the interval $[0.05, 0.95]$ for τ. For each value of τ, we used five-fold cross validation over a data-driven grid of 20 values to estimate λ. We set $n = 400$, $\mu = 0.10$, and $\boldsymbol{\beta}^* = (-0.03, 0, 0, -0.03, -0.03, 0.03, 0.03, 0, 0, 0.03)$ and vary $\sigma_y \in \{0.3, 0.5\}$. Furthermore, we let $\Sigma = I_{10}$ so that the covariates are independent. We compared the following approaches: 1) asymLASSO with fixed $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$ and 2) asymLASSO with τ also being estimated via cross validation, and evaluated their selection performance through the inclusion probability (P_j), the proportion of simulations that correctly identify β_j^* as non zero. We report our findings in Table 1 where the results are averaged over $B = 100$ Monte Carlo replicates.

We can see that for $\tau < 0.5$, asymLASSO has a higher probability of selecting the positive-signed effects (P_6, P_7, and P_{10}) over the negative-signed effects (P_1, P_4, and P_5) the degree to which is determined by the value of τ; whereas, the opposite is true for $\tau > 0.5$. By construction of the parameter vector, we do not expect to prefer shrinking positive effects over negative effects or vice versa. In fact, when $\tau = 0.5$ (i.e., the standard LASSO) we see that the inclusion probabilities for all six non-zero variables are comparable. Furthermore, the estimated optimal value for τ, averaged over all 100 simulations, is close to 0.5, suggesting that a data-driven method should be used to select τ rather than a prespecified value. We also assessed selection performance under correlated covariates (Tables S1 and S2 in the Online Supplementary Material). As expected, selection performance worsens when correlation is present; however, the conclusions generally remain consistent with what we observe in Table 1.

3.2 Finite Sample Performance Compared to the LASSO

In this section we study the finite sample performance of asymLASSO compared to both LASSO and the non-negative LASSO (nLASSO). We let $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_0, \mathbf{0}_{p-10})$, where we set $\mu = 0.10$ and

Table 1: Asymmetric LASSO (asymLASSO) with varying values for τ where $n = 400$, $\Sigma = I_{10}$, $\mu = 0.10$, and $\boldsymbol{\beta}^* = (-0.03, 0, 0, -0.03, -0.03, 0.03, 0.03, 0, 0, 0.03)$. The tuning parameter λ was selected using five-fold cross validation between an evenly-spaced grid $[0.05, 0.95]$. Results are averaged over 100 simulations. $\hat{\tau}_{CV}$ is the average value of τ selected via cross validation for each of the 100 simulations (P_j = proportion of simulations where β_j is correctly identified as non-zero). See Section 3.1 for more details.

σ_y	Method	P_1	P_4	P_5	P_6	P_7	P_{10}
0.5	$\hat{\tau}_{CV} = 0.41$	0.39	0.33	0.35	0.38	0.37	0.37
	$\tau = 0.05$	0.30	0.25	0.29	0.52	0.51	0.54
	$\tau = 0.25$	0.32	0.28	0.33	0.54	0.54	0.53
	$\tau = 0.50$	0.45	0.42	0.45	0.42	0.40	0.39
	$\tau = 0.75$	0.46	0.40	0.43	0.29	0.26	0.27
	$\tau = 0.95$	0.47	0.45	0.46	0.25	0.22	0.22
0.3	$\hat{\tau}_{CV} = 0.47$	0.70	0.73	0.69	0.78	0.77	0.81
	$\tau = 0.05$	0.61	0.61	0.59	0.88	0.87	0.90
	$\tau = 0.25$	0.63	0.66	0.64	0.88	0.86	0.92
	$\tau = 0.50$	0.82	0.87	0.82	0.86	0.79	0.85
	$\tau = 0.75$	0.87	0.88	0.84	0.73	0.70	0.74
	$\tau = 0.95$	0.88	0.90	0.86	0.70	0.68	0.70

$\sigma_y = 0.5$ and considered the following scenarios for $\boldsymbol{\beta}_0$:

$$\text{Model 1: } \boldsymbol{\beta}_0 = (0.03, 0, 0, 0.03, 0, 0.05, 0.08, 0, 0, 0.10)$$
$$\text{Model 2: } \boldsymbol{\beta}_0 = (-0.03, 0, 0, 0.03, 0, 0.05, 0.08, 0, 0, -0.10)$$
$$\text{Model 3: } \boldsymbol{\beta}_0 = (-0.03, 0, 0, -0.03, 0, 0.05, 0.08, 0, 0, 0.10)$$
$$\text{Model 4: } \boldsymbol{\beta}_0 = (0.03, 0, 0, 0.03, 0, 0.05, -0.08, 0, 0, -0.10).$$

For asymLASSO, we used an evenly-spaced grid on the interval $[0.05, 0.95]$ to select τ. Oracle estimates were retrieved from OLS regression using the underlying true model. Both LASSO and the non-negative LASSO were performed using the *glmnet* package (Friedman et al., 2010). A data-driven grid of 20 λ values was employed for all three methods and five-fold cross validation was used to select the final model.

We evaluate the approaches by their variable selection, parameter estimation, and prediction performance. For variable selection, we used the probability of inclusion measures defined in Section 3.1 as well as the mean number of false positives (FP) and mean number of false negatives (FN). Estimation bias is estimated using the mean squared bias, $MSB = \frac{1}{B}\sum_{j=1}^{B} \left\| \hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}^* \right\|_2$, where B is the number of simulations. Lastly, prediction performance is estimated using the predicted mean squared error (PMSE) derived from a test set of $n = 1{,}000$. Results are averaged over $B = 100$ Monte Carlo replicates and are presented in Table 2 for Model 1 when $n = 400$ and 800, $p = 50$ and 200, and $\Sigma = (0.5^{|i-j|})_{ij}$.

First, we observe that estimation and prediction performance between the three methods are comparable. However, both nLASSO and asymLASSO have better selection performance than the traditional LASSO across all five true non-zero coefficients, especially for the smaller

Table 2: Comparison of asymLASSO to LASSO and the non-negative LASSO (nLASSO) based on 100 Monte Carlo replicates. (MSB = mean squared bias; FP = mean number of false positives (out of 45); FN = mean number of false negatives (out of 5); P_j = proportion of simulations where β_j is correctly identified as non-zero; PMSE = Averaged predicted mean squared error.) See Section 3.2 for more details.

	n	Method	MSB	FP	FN	P_1	P_4	P_6	P_7	P_{10}	PMSE
$p = 50$	400	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
		LASSO	0.09	5.91	1.67	0.39	0.43	0.81	0.79	0.91	0.26
		nLASSO	0.08	4.33	1.49	0.40	0.47	0.86	0.83	0.95	0.26
		asymLASSO($\hat{\tau} = 0.20$)	0.08	5.58	1.49	0.41	0.47	0.85	0.83	0.95	0.26
	800	Oracle	0.04	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
		LASSO	0.07	7.28	0.92	0.57	0.62	0.96	0.93	1.00	0.25
		nLASSO	0.06	4.94	0.77	0.63	0.68	0.97	0.95	1.00	0.25
		asymLASSO($\hat{\tau} = 0.23$)	0.07	6.55	0.77	0.63	0.68	0.97	0.95	1.00	0.25
$p = 200$	400	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
		LASSO	0.09	6.69	2.57	0.11	0.36	0.61	0.58	0.77	0.26
		nLASSO	0.09	6.31	2.18	0.23	0.43	0.69	0.66	0.81	0.26
		asymLASSO($\hat{\tau} = 0.24$)	0.09	7.70	2.24	0.22	0.40	0.68	0.65	0.81	0.26
	800	Oracle	0.04	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
		LASSO	0.07	8.41	1.53	0.35	0.46	0.80	0.88	0.98	0.26
		nLASSO	0.07	8.13	1.30	0.45	0.54	0.84	0.88	0.99	0.26
		asymLASSO($\hat{\tau} = 0.22$)	0.07	9.48	1.34	0.43	0.54	0.83	0.87	0.99	0.26

effect sizes. For example, when $n = 400$, the probability of inclusion for $\beta^*_{04} = 0.03$ for the LASSO is 43% compared to 47% for nLASSO and asymLASSO. Our estimated value for τ using cross validation is $\hat{\tau} = 0.20 < 0.50$, which is expected since our true model is comprised of only positive signals. Furthermore, nLASSO tends to identify less false positives compared to LASSO and asymLASSO. As the sample size increases ($n = 800$), all three methods have improved overall performance but the patterns between them remain the same. In Table S3 of the Online Supplementary Material we repeat the same scenario but under two different correlation structures ($\Sigma = I$ and $\Sigma = \left(0.80^{1(i \neq j)}\right)_{ij}$). When the covariates are independent, the results mirror what we observe in Table 1. Surprisingly, under high equicorrelation, all three methods perform similarly in terms of selection while asymLASSO identifies slightly more false positives.

In our previous example, the performance of asymLASSO falls somewhere between the LASSO and nLASSO. asymLASSO had better selection performance than the LASSO but at the expense of identifying more false positive than nLASSO. Moreover, we expected nLASSO to perform well under the previous setting since the effect sizes were all positive. A predetermined constraint was placed to ensure that only positive effects were retained in the model for nLASSO; whereas asymLASSO allowed the data to dictate the shrinkage, which preferred selecting positive effects over negative ones. In models where negative effects are present, we would expect nLASSO to perform poorly. We further illustrate this in Table 3 where we allow

Table 3: Comparison of ASYMLASSO to LASSO and nLASSO based on 100 Monte Carlo replicates with $n = 400$, $p = 50$ and under various sign effects. Model 2: $\boldsymbol{\beta}_0 = (-0.03, 0, 0, 0.03, 0, 0.05, 0.05, 0, 0, -0.08)$; Model 3: $\boldsymbol{\beta}_0 = (-0.03, 0, 0, -0.03, 0, 0.05, 0.05, 0, 0, 0.08)$; Model 4: $\boldsymbol{\beta}_0 = (0.03, 0, 0, 0.03, 0, 0.05, -0.05, 0, 0, -0.08)$. (MSB = mean square bias; FP = mean number of false positives (out of 45); FN = mean number of false negatives (out of 5) P_j = proportion of simulations where β_j is correctly identified as non-zero; PMSE = Averaged predicted mean squared error.) See Section 3.2 for more details.

Model	Method	MSB	FP	FN	P_1	P_4	P_6	P_7	P_{10}	PMSE
2	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
	LASSO	0.09	5.08	2.21	0.20	0.37	0.81	0.61	0.80	0.26
	nLASSO	0.11	2.08	3.35	0.01	0.33	0.77	0.54	0.00	0.26
	ASYMLASSO($\hat{\tau} = 0.44$)	0.09	5.52	2.18	0.22	0.38	0.83	0.64	0.75	0.26
3	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
	LASSO	0.09	5.14	2.31	0.31	0.22	0.55	0.76	0.85	0.26
	nLASSO	0.09	3.59	2.70	0.01	0.00	0.60	0.81	0.88	0.26
	ASYMLASSO($\hat{\tau} = 0.26$)	0.09	4.92	2.36	0.23	0.12	0.62	0.80	0.87	0.26
4	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
	LASSO	0.10	5.03	2.98	0.35	0.30	0.34	0.28	0.75	0.26
	nLASSO	0.11	1.25	4.39	0.25	0.19	0.17	0.00	0.00	0.26
	ASYMLASSO($\hat{\tau} = 0.54$)	0.10	4.58	3.09	0.31	0.27	0.26	0.30	0.77	0.26

the coefficient estimates to vary in sign under Models 2, 3, and 4. In general, nLASSO produces sparser models than both LASSO and ASYMLASSO. Under Model 2, where the smallest and largest effects are negative, nLASSO fails to select the largest effect. Surprisingly, nLASSO selected the smallest negative effect but erroneously estimated its effect as positive. We see this same pattern in Models 3 and 4 where the two largest and two smallest effect sizes are negative, respectively. Focusing our attention to LASSO and ASYMLASSO, both methods have comparable performance and there is difficulty in preferring ASYMLASSO over LASSO and vice versa. For example, in Model 3, ASYMLASSO has better selection performance for the positive-signed coefficients compared to the LASSO but worse selection performance for the two negative-signed coefficients due, in part, by their magnitude. We also see that ASYMLASSO is aiming to balance the sign-specific shrinkage trade off based on the sign and magnitude of the effects that are present in the data, as reflected by the estimated value for τ for each model, providing a data-driven approach to asymmetric penalization that doesn't require an *a priori* constraint as in the constrained LASSO.

Lastly, we compare all three methods in a high-dimensional setting with $n = 400$, $p = 2{,}000$, $\Sigma = (0.5^{|i-j|})_{ij}$, and under Models 1 and 2. We increase the signal size of the smallest effect to 0.08 (Table 4). We notice similar results to what we have observed previously (Tables 2 and 3). While ASYMLASSO identifies slightly more false positives; the mean false positive rates (the number of false positives over 1,995) are comparable across all three methods.

Table 4: Comparison of asymLASSO to LASSO and nLASSO in a high-dimensional OLS setting. Results based on 100 Monte Carlo replicate with $n = 400$ and $p = 2{,}000$. Model 1: $\boldsymbol{\beta}_0 = (0.08, 0, 0, 0.08, 0, 0.10, 0.10, 0, 0, 0.15)$; Model 2: $\boldsymbol{\beta}_0 = (-0.08, 0, 0, 0.08, 0, 0.10, 0.10, 0, 0, -0.15)$. (MSB = mean square bias; FP = mean number of false positives (out of 1,995); FN = mean number of false negatives (out of 5); P_j = proportion of simulations where β_j is correctly identified as non-zero; PMSE = Averaged predicted mean squared error.) See Section 3.2 for more details.

Model	Method	MSB	FP	FN	P_1	P_4	P_6	P_7	P_{10}	PMSE
1	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
	LASSO	0.14	15.88	0.50	0.69	0.85	0.99	0.97	1.00	0.27
	nLASSO	0.13	13.82	0.43	0.74	0.87	0.99	0.97	1.00	0.27
	asymLASSO($\hat{\tau} = 0.26$)	0.14	16.85	0.40	0.75	0.89	0.99	0.97	1.00	0.27
2	Oracle	0.06	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.25
	LASSO	0.16	16.68	1.02	0.46	0.67	0.98	0.90	0.97	0.28
	nLASSO	0.20	6.86	2.56	0.00	0.61	0.96	0.87	0.00	0.29
	asymLASSO($\hat{\tau} = 0.48$)	0.16	18.06	1.18	0.36	0.65	0.97	0.92	0.92	0.28

3.3 Assessing Sign-Dependent Shrinkage

Our simulations from Section 3.2 show that asymLASSO outperforms LASSO in terms of variable selection when the true effects are in the same direction. In the presence of mixed-sign and mixed-magnitude effects, asymLASSO and LASSO have their own respective benefits and drawbacks. When the covariates in the model are independent (i.e., Σ is a diagonal matrix), simple transformations on the columns of the design matrix will change the magnitude and/or direction of the corresponding coefficient estimate. For example, negating the entries in a column will switch the sign of the estimate from positive to negative (or vice versa).

In the following study, we compare the performance of LASSO and nLASSO to asymLASSO under simple sign transformations of the design matrix. Our simulation setup follows similarly to Section 3.2 except that $\Sigma = I_p$ to ensure independence among the covariates and we set $\sigma_y = 0.5$. Furthermore, we set the two smallest signals to be negative, i.e., $\boldsymbol{\beta}_0 = (-0.03, 0, 0, -0.03, 0, 0.05, 0.05, 0, 0, 0.08)$ as in Model 2. Our interest lies in comparing the probability of inclusion for the two smallest signals $\beta_1^* = -0.01$ and $\beta_4^* = -0.01$ before and after we switch the signs of the first and fourth columns of X. In other words, we generate a design matrix X, simulate the outcome $\mathbf{y}|X$, and create a new design matrix $\tilde{X}$ such that for all $i = 1, \ldots, n$:

$$\tilde{X}_{ij} = \begin{cases} X_{ij} & j \notin \{1, 4\} \\ -X_{ij} & j \in \{1, 4\}. \end{cases}$$

Thus, $\tilde{X}$ only differs from X in the first and fourth column where the entries of $\tilde{X}$ are negated entries of X. In doing so, regressing the outcome $\mathbf{y}$ on $\tilde{X}$ will produce positive effect estimates for the first and fourth coefficients and thus will be in the same effect direction as the other non-zero values. Unlike LASSO, asymLASSO is sign variant and we believe that selection performance will improve when using $\tilde{X}$ as the design matrix in the model over X. We evaluate the selection performance of LASSO, nLASSO and asymLASSO when using either X or $\tilde{X}$ and present the results in Figure 3 when $n = 400$ and $p = 50$.

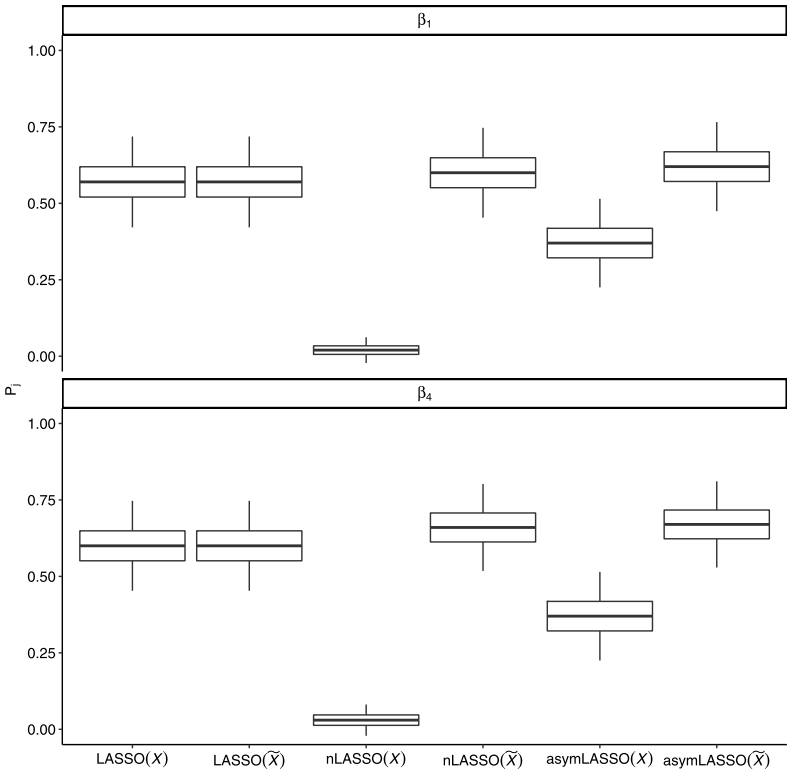

Figure 3: Box plot (mean ± 3 standard deviations) of the inclusion probability (P_j) for the first and fourth nonzero coefficients ($\beta_{01} = \beta_{04} = -0.03$) when $n = 400$, $p = 50$, and $\Sigma = I$. A grid search between [0.05, 0.95] is used to select τ for asymLASSO. Five-fold cross validation is used to select the final model for LASSO, nLASSO, and asymLASSO. Results are averaged over 100 Monte Carlo simulations. (X): Uses the design matrix X in the model fit; $(\tilde{X})$: Uses the design matrix $\tilde{X}$ in the model fit, where $\tilde{X}$ and X are identical except that the first and fourth columns are negated. See Section 3.3 for more details.

While the data generation scheme is slightly different, the results for all three methods using X as the design matrix reflect what we observed for Model 2 in Table 6. Specifically, the probabilities of inclusions, P_1 and P_4, are lower for asymLASSO than LASSO since asymLASSO prefers selection of positive effects ($\hat{\tau} = 0.23$). Similarly, nLASSO incorrectly assigns positive effect estimates to both β_1 and β_4. We see a drastic improvement in selection performance when we use $\tilde{X}$ as the design matrix for asymLASSO since the effects of interest are coded to be in the same direction as the other (larger) non-zero effects ($\hat{\tau} = 0.22$). The same is true for the nLASSO. Additionally, due to LASSO shrinking symmetrically around zero, the performance of LASSO is unchanged. We perform additional simulations (Figures S1 and S2 in the Online Supplementary Material) where we introduce correlation between the covariates. The overall conclusions are consistent to what we observe in Figure 3. Furthermore, as previously mentioned, the selection performance for asymLASSO is similar to the LASSO when the covariates are highly correlated ($\Sigma = \left(0.8^{1(i \neq j)}\right)_{ij}$).

These results show that by cleverly transforming the design matrix such that the expected effects are mostly (or all) in one direction, asymLASSO demonstrates better selection per-

Table 5: Comparison of ASYMLASSO to LASSO and nLASSO in a logistic regression setting. Results based on 100 Monte Carlo replicate with $n = 400$, $p = 50$, $\Sigma = (0.5^{|i-j|})_{ij}$. Model 1: $\boldsymbol{\beta}_0 = (0.08, 0, 0, 0.10, 0, 0.12, 0.15, 0, 0, 0.25)$; Model 2: $\boldsymbol{\beta}_0 = (-0.08, 0, 0, 0.10, 0, 0.12, 0.15, 0, 0, -0.25)$. (MSB = mean square bias; FP = mean number of false positives; FN = mean number of false negatives; P_j = proportion of simulations where β_j is correctly identified as non-zero; AUC = Area under curve estimate from the test set.) See Section 3.4 for more details.

Model	Method	MSB	FP	FN	P_1	P_4	P_6	P_7	P_{10}	AUC
1	Oracle	0.24	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.60
	LASSO	0.31	3.25	3.34	0.13	0.19	0.40	0.39	0.55	0.56
	nLASSO	0.30	3.12	2.99	0.21	0.29	0.46	0.47	0.58	0.56
	ASYMLASSO($\hat{\tau} = 0.29$)	0.32	3.80	3.10	0.18	0.29	0.43	0.45	0.55	0.56
2	Oracle	0.26	0.00	0.00	1.00	1.00	1.00	1.00	1.00	0.58
	LASSO	0.32	3.13	3.66	0.11	0.11	0.31	0.43	0.38	0.54
	nLASSO	0.34	1.84	4.07	0.03	0.15	0.33	0.42	0.00	0.53
	ASYMLASSO($\hat{\tau} = 0.38$)	0.33	3.38	3.53	0.10	0.19	0.33	0.45	0.40	0.54

formance of small effects compared to the LASSO. This is particularly applicable in genomics studies where covariates may be coded *a priori* in the risk direction where increases in the covariate value correspond to higher risk for the outcome and thus potentially allows for the discovery of small effects that may have been erroneously shrunken to zero by the LASSO.

3.4 Binary Outcome

To highlight the application within the GLM framework, we compared LASSO and nLASSO to ASYMLASSO under a binary outcome. Similar to Section 3.2, we set $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_0, \mathbf{0}_{p-10})$ and generated X from a multivariate Gaussian distribution with an autoregressive covariance structure. We simulated the outcome from the following logistic regression model $\mathbf{y}|\mathbf{x} \sim Bernoulli\{\pi(\mu + \mathbf{x}^T\boldsymbol{\beta}^*)\}$ where $\pi(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$. The intercept term $\mu = 0.50$ corresponded to a case rate of approximately 60%. We evaluated prediction performance using the area under the curve (AUC) in a test set of $n = 1{,}000$. The results comparing ASYMLASSO to LASSO for the logistic regression model are displayed in Table 5 under two models – Model 1: $\boldsymbol{\beta}_0 = (0.08, 0, 0, 0.10, 0, 0.12, 0.15, 0, 0, 0.25$; Model 2: $\boldsymbol{\beta}_0 = (-0.08, 0, 0, 0.10, 0, 0.12, 0.15, 0, 0, -0.25)$. Results were averaged over 100 Monte Carlo simulations. Our conclusions mimic what was reflected in the OLS scenario (see Tables 1 and 2) and we also observed consistent patterns, not reported, under different sample sizes, effect sizes, parameter dimensions, and correlation structures.

4 Real Data Analysis: Breast Cancer Gene Expression

BRCA1 is a DNA damage repair gene that produces tumor suppressor proteins. Pathogenic variants in *BRCA1* and *BRCA1* expression have been shown to have strong associations with breast and ovarian cancer risk (Welcsh et al., 2000; Welcsh and King, 2001). *BRCA1* is known to interact with many other genes, particularly in response to DNA damage. In this analysis, we aimed to identify genes associated with *BRCA1* expression, as such genes could implicate

Table 6: TCGA Breast Cancer Gene Expression Analysis. Number of positive and negative signals kept in the final model using the training ($n = 357$) set. Predicted R^2 was calculated using the test ($n = 179$) set. Ten-fold cross validation was performed across a grid of 20 λ values for all three methods. The skewness parameter τ was estimated across a grid of 19 values between $[0.05, 0.95]$.

Method	# Positive	# Negative	Pred. R^2
asymLASSO	101	21	0.574
LASSO	73	35	0.560
nLASSO	126	0	0.562

mechanisms through which *BRCA1* impacts cancer risk and warrant further investigation in future studies. We applied LASSO, nLASSO, and asymLASSO to identify such genes.

Gene expression data were available for 17,814 genes measured in breast cancer tissue samples from 536 women with breast cancer from The Cancer Genome Atlas (TCGA). The data are available at http://cancergenome.nih.gov and has been previously analyzed in Breheny (2019). We excluded 491 genes due to expression values missing in one or more women. Expression values of the remaining 17,322 genes were log-transformed and standardized. A broad grid between $[0.05, 0.95]$ was used to estimate τ for asymLASSO. Similar to our simulation study, LASSO, nLASSO, and asymLASSO were performed using 10-fold cross validation. We randomly split the data into both a training ($n = 357$) and test ($n = 179$) set. Table 6 summarizes the number of selected variables and predicted R^2 for each method.

The asymLASSO approach exhibits a minor improvement in the predicted R^2 when compared to both the LASSO and nLASSO. Furthermore, the number of variables retained in the model (122) is similar to both LASSO (108) and nLASSO (126). The asymLASSO prefers selecting positive effects over negative effects ($\hat{\tau} = 0.45$; Figure 4). As a comparison, we also performed LASSO by forcing only negative coefficients in the model (non-positive LASSO). The predicted R^2 (not reported in Table 6) of the model is 0.36, which is substantially worse than asymLASSO, LASSO, and nLASSO. Thus, one can infer that the positive estimates in the model are driving the predictive performance, which is in line with what we see in Figure 4 where the cross validation error is largest for asymLASSO when $\tau > 0.5$.

All three methods overlap in 57 of the gene expressions and nine were uniquely identified in asymLASSO (8 positive effects, 1 negative effect). Notable uniquely-identified gene expressions in this set include *MND1* and *JARID2*, which correspond to the two largest effects in this subset. *MND1* is a protein coding gene that has been shown to interact with the human oncogene *GT198*, which is located within the *BRCA1* locus (Ijichi et al., 2000; Ko et al., 2002; Tsubouchi and Roeder, 2002; Enomoto et al., 2004, 2006). The protein coding gene *JARID2* has been previously shown to be essential for the maintenance of tumor initiating cells in bladder cancer Zhu et al. (2017) and for ovarian cancer (Cao et al., 2017). Recently, *JARID2* has been shown to be widely expressed in various breast cancer cell lines and patients with *JARID2* mutation were shown to have a significantly shorter period of disease-free survival (Zhang et al., 2020).

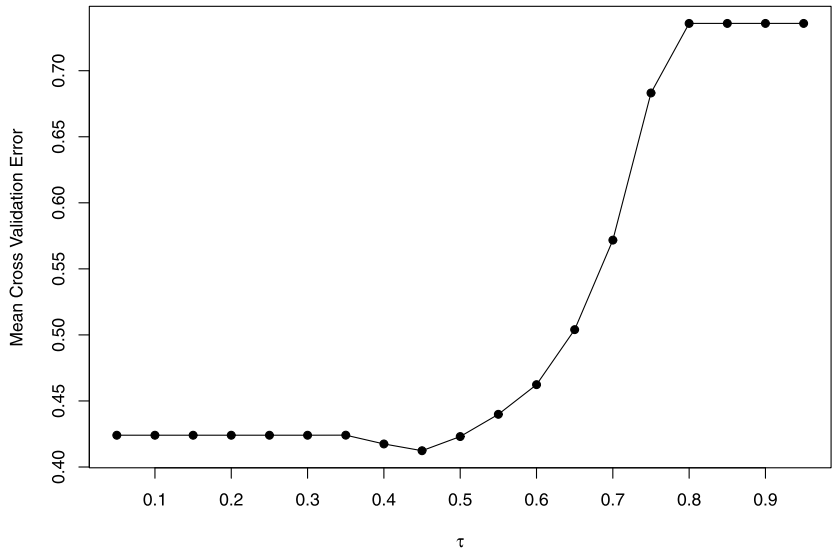

Figure 4: Plot of the mean cross validation error (with respect to optimizing λ) for asymLASSO across different values of τ used in the TCGA Breast Cancer Gene Expression analysis (Section 4). The final model selected for the analysis corresponds to the τ that minimizes the mean cross validated error ($\hat{\tau} = 0.45$).

5 Discussion

We develop a generalization to LASSO penalization that asymmetrically penalizes coefficients based on sign. We provide both a Bayesian and frequentist interpretation of our method. Under the Bayesian paradigm, shrinkage of the estimates is performed by placing an asymmetric Laplace prior on the regression coefficients. In doing so, the prior probability that a coefficient is less than (or greater than) zero is determined by the skew parameter $\tau \in (0, 1)$. Furthermore, the asymmetric Laplace prior corresponds to an asymmetric ℓ_1 penalty for penalized regression. To better understand the behavior of asymLASSO and its relation to the LASSO, we present a closed-form solution for the OLS model under orthonormal design. Preferential shrinkage of positive or negative effect estimates can be achieved based on the value of τ. Unlike the constrained LASSO, where constraints are predetermined, asymLASSO achieves asymmetric shrinkage through the tuning parameter τ, which can be estimated using the data. We implement our approach using cyclic coordinate descent.

Our simulations demonstrate that asymLASSO outperforms LASSO in selecting smaller signals when effect estimates are generally in the same direction for both low- and high-dimensional covariates at the expense of identifying slightly more false positives. While this may seem concerning at first, both LASSO and asymLASSO are not expected to be model selection consistent. To this end, the goal of asymLASSO is to provide a more flexible approach to the LASSO that allows for asymmetric shrinkage, potentially allowing for the discovery of smaller effects that may have been previously missed. Additionally, in the presence of mixed-sign effects, it is difficult to prefer one approach over the other. These challenges are due to the "sign-specific shrinkage tradeoff" that is inherent to asymLASSO providing both advantages and disadvantages. However, in mild circumstances, we observe that the selection performance of asymLASSO can be substantially improved if we have *a priori* knowledge about the direction

of effects and transform the design matrix accordingly. In general, while it may be difficult for practitioners to code the covariates accordingly in advance, the ability to improve selection performance by manipulating the design matrix is a unique benefit to asymmetric shrinkage when compared to the standard LASSO and non-negative LASSO.

We apply our approach to breast cancer gene expression data from the TCGA to identify genes associated with *BRCA1* expression and compare its performance with LASSO. Our method identified nine genes that were not identified by either LASSO or nLASSO. Two of these genes, *MND1* and *JARID2*, have been previously reported to be associated with *BRCA1* or with breast cancer progression and provides evidence to further understand their biological relationship within the context of *BRCA1* gene expression.

We envision several paths to improve asymmetric penalization. While the motivation of the asymmetric LASSO is derived from a Bayesian perspective, parameter estimation and variable selection is performed through minimizing a penalized log-likelihood. We are currently investigating the performance of the asymmetric LASSO under a fully Bayesian framework. The asymmetric ℓ_1 penalty does not overcome some of the theoretical and practical shortcomings that are well known to LASSO penalization. The LASSO has been shown to exhibit model selection consistency under strict conditions on the design matrix (Zhao and Yu, 2006). We conjecture that these results hold for the asymLASSO under certain assumptions on τ. Another approach to ensure model selection consistency is to extend asymmetric penalization to oracle-based procedures (Fan and Li, 2001). In Section 2.2, we show that asymLASSO can produce more biased estimates than the LASSO for certain coefficient estimates based on the value of τ. Similar to overcoming the bias issue for larger estimates for the LASSO, we can weight the shrinkage parameter for each coefficient differently and perform adaptive asymmetric LASSO penalization. We provide a graph (Figure 5) of the soft thresholding function under orthogonal design for both asymLASSO (solid black line) and adaptive asymLASSO (dashed black line) which clearly shows that adaptive asymLASSO reduces the bias for larger estimates in both directions. Proving the oracle property for asymmetric versions of, for example, adaptive LASSO (Zou, 2006), SCAD (Fan and Li, 2001), and MCP (Zhang et al., 2010), will require additional conditions on τ and λ. Lastly, when dealing with high-dimensional data, strong rules (Ghaoui et al., 2010; Tibshirani et al., 2012; Zeng et al., 2021) to safely and effectively discard large number of inactive predictors have been implemented for computational efficiency. These rules have been well studied for symmetric penalties around zero and we expect that modifications to these rules can be generally implemented for asymmetric penalization.

Supplementary Material

The following supplemental material are provided: R files necessary to reproduce the simulation results reported in this manuscript, and PDF providing supplemental tables and figures and the proof of Lemma 2.1.

Acknowledgements

We thank the referees, the associate editor, and the editor for their helpful comments that improved the presentation and quality of the article.

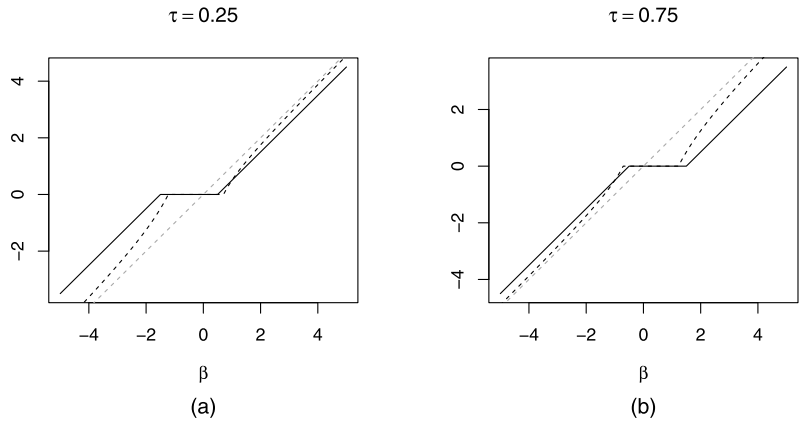

Figure 5: Behavior of the soft-thresholding function for adaptive asymLASSO (dotted black line) under orthogonal design for: (a) $\tau = 0.25$ and (b) $\tau = 0.75$. The dotted grey line represents the true value of $\boldsymbol{\beta}$. Solid black line: The soft thresholding operator for the standard asymLASSO.

Funding

Eric S. Kawaguchi's work is partially supported through the National Institutes of Health (NIH) grant T32ES013678. Burcu F. Darst's work is partially supported through the National Cancer Institute (NCI) grant K99CA246063 and the Achievement Rewards for College Scientists Foundation Los Angeles Founder Chapter. The research of David V. Conti is partly supported by the NIH grants P01CA196569, R01HG010297, R01CA241410, R01CA257328, and P30CA014089.

References

Akaike H (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723.

Angelosante D, Giannakis GB, Grossi E (2009). Compressed sensing of time-varying signals. In: *2009 16th International Conference on Digital Signal Processing*, 1–8. IEEE.

Breheny P, Huang J (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1): 232–253.

Breheny PJ (2019). Marginal false discovery rates for penalized regression models. *Biostatistics*, 20(2): 299–314.

Cao J, Li H, Liu G, Han S, Xu P (2017). Knockdown of jarid2 inhibits the proliferation and invasion of ovarian cancer through the pi3k/akt signaling pathway. *Molecular Medicine Reports*, 16(3): 3600–3605.

Debortoli S, Müller O, Junglas I, vom Brocke J (2016). Text mining for information systems researchers: An annotated topic modeling tutorial. *Communications of the Association for Information Systems*, 39(7): 111–135.

Dobson AJ, Barnett AG (2018). *An Introduction to Generalized Linear Models (4th ed.)*. Routledge.

Donoho DL, Johnstone JM (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3): 425–455.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression. *The Annals of Statistics*, 32(2): 407–499.

Enomoto R, Kinebuchi T, Sato M, Yagi H, Kurumizaka H, Yokoyama S (2006). Stimulation of dna strand exchange by the human tbpip/hop2-mnd1 complex. *Journal of Biological Chemistry*, 281(9): 5575–5581.

Enomoto R, Kinebuchi T, Sato M, Yagi H, Shibata T, Kurumizaka H, Yokoyama S (2004). Positive role of the mammalian tbpip/hop2 protein in dmc1-mediated homologous pairing. *Journal of Biological Chemistry*, 279(34): 35263–35272.

Fan J, Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456): 1348–1360.

Friedman J, Hastie T, Tibshirani R (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3): 432–441.

Friedman J, Hastie T, Tibshirani R (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22.

Gaines BR, Kim J, Zhou H (2018). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4): 861–871.

Ghaoui LE, Viallon V, Rabbani T (2010). Safe feature elimination for the lasso and sparse supervised learning problems. arXiv preprint: https://arxiv.org/abs/1009.4219.

Ghosh D, Chinnaiyan AM (2005). Classification and selection of biomarkers in genomic data using lasso. *Journal of Biomedicine and Biotechnology*, 2005(2): 147.

Golub GH, Heath M, Wahba G (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2): 215–223.

Hans C (2009). Bayesian lasso regression. *Biometrika*, 96(4): 835–845.

Huang X, Pan W (2003). Linear regression and two-class classification with gene expression data. *Bioinformatics*, 19(16): 2072–2078.

Ijichi H, Tanaka T, Nakamura T, Yagi H, Hakuba A, Sato M (2000). Molecular cloning and characterization of a human homologue of tbpip, a brca1 locus-related gene. *Gene*, 248(1–2): 99–107.

James GM, Paulson C, Rusmevichientong P (2012). The constrained lasso. In: *Refereed Conference Proceedings*, volume 31, 4945–4950. Citeseer.

Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, Natarajan P, Lander ES, Lubitz SA, Ellinor PT, Kathiresan S (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50(9): 1219–1224.

Ko L, Cardona GR, Henrion-Caude A, Chin WW (2002). Identification and characterization of a tissue-specific coactivator, gt198, that interacts with the dna-binding domains of nuclear receptors. *Molecular and Cellular Biology*, 22(1): 357–369.

Koenker R, Basset G (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association*, 73(363): 618–622.

Kozumi H, Kobayashi G (2011). Gibbs sampling methods for bayesian quantile regression. *Journal of Statistical Computation and Simulation*, 81(11): 1565–1578.

Li Y, Algarni A, Albathan M, Shen Y, Bijaksana MA (2014). Relevance feature discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, 27(6): 1656–1669.

McCullagh P, Nelder JA (1983). *Generalized Linear Models (2nd ed.)*. Routledge.

Meinshausen N, Bühlmann P (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3): 1436–1462.

Panagiotidis T, Stengos T, Vravosinos O (2018). On the determinants of bitcoin returns: A lasso approach. *Finance Research Letters*, 27: 235–240.

Park T, Casella G (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482): 681–686.

Pereira JM, Basto M, da Silva AF (2016). The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, 39: 634–641.

Schwarz G, et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2): 461–464.

Takeuchi I, Le Q, Sears T, Smola A (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7: 1231–1264.

Tibshirani R (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1): 267–288.

Tibshirani R (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4): 385–395.

Tibshirani R, Bien J, Friedman J, Hastie T, Simon N, Taylor J, Tibshirani RJ (2012). Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2): 245–266.

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

Tibshirani RJ, Taylor J (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3): 1335–1371.

Tsubouchi H, Roeder GS (2002). The mnd1 protein forms a complex with hop2 to promote homologous chromosome pairing and meiotic double-strand break repair. *Molecular and Cellular Biology*, 22(9): 3078–3088.

Welcsh PL, King MC (2001). Brca1 and brca2 and the genetics of breast and ovarian cancer. *Human Molecular Genetics*, 10(7): 705–713.

Welcsh PL, Owens KN, King MC (2000). Insights into the functions of brca1 and brca2. *Trends in Genetics*, 16(2): 69–74.

Wu L, Yang Y, Liu H (2014). Nonnegative-lasso and application in index tracking. *Computational Statistics & Data Analysis*, 70: 116–126.

Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6): 714–721.

Wu TT, Gong H, Clarke EM (2011). A transcriptome analysis by lasso penalized cox regression for pancreatic cancer survival. *Journal of Bioinformatics and Computational Biology*, 9(supp01): 63–73.

Wu TT, Lange K (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, 2(1): 224–244.

Yu K, Moyeed RA (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4): 437–447.

Yu K, Zhang J (2005). A three-parameter asymmetric Laplace distribution and its extension. *Communications in Statistics—Theory and Methods*, 34(9–10): 1867–1879.

Yuan M, Lin Y (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1): 49–67.

Zeng Y, Yang T, Breheny P (2021). Hybrid safe–strong rules for efficient optimization in lasso-type problems. *Computational Statistics & Data Analysis*, 153: 107063.

Zhang CH, et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2): 894–942.

Zhang X, Li J, Yang Q, Li X WY, Liu Y, Shan B (2020). Tumor mutation burden and jarid2 gene alteration are associated with short disease-free survival in locally advanced triple-negative breast cancer. *The Annals of Translational Medicine*, 8(17): 1052.

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

Zhu XX, Yan YW, Ai CZ, Jiang S, Xu SS, Niu M, et al. (2017). Jarid2 is essential for the maintenance of tumor initiating cells in bladder cancer. *Oncotarget*, 8(15): 24483–24490.

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

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