The Lasso and Ridge Regression Yield Biased Estimates of Imbalanced Binary Features

Larsson, Johan; Wallin, Jonas

2024

Document Version:
Other version

Link to publication

Citation for published version (APA):
Larsson, J., & Wallin, J. (2024). The Lasso and Ridge Regression Yield Biased Estimates of Imbalanced Binary Features. Unpublished.

Total number of authors:
2

The Lasso and Ridge Regression Yield Biased Estimates of Imbalanced Binary Features

Johan Larsson *johan.larsson@stat.lu.se*
Deparment of Statistics
Lund University

Jonas Wallin *jonas.wallin@stat.lu.se*
Department of Statistics
Lund University

Abstract

Many regularized methods, such as the lasso and ridge regression, are sensitive to the scales of the features in the data. As a consequence, it has become standard practice to normalize (center and scale) features such that they share the same scale. For continuous data, the most common strategy is standardization: centering and scaling each feature by its mean and and standard deviation, respectively. For binary data, especially when it is high-dimensional and sparse, the most common strategy, however, is to not scale at all. In this paper, we show that this choice has dramatic effects for the estimated model in the case when the binary features are imbalanced and that these effects, moreover, depend on the type regularization (lasso or ridge) used. In particular, we demonstrate the size of a feature's corresponding coefficient in the lasso is directly related to its class imbalance and that this effect depends on the normalization used. We suggest possible remedies for this problem and also discuss the case when data is mixed, that is, contains both continuous and binary features.

1 Introduction

When the data you want to model is high-dimensional, that is, the number of features p exceed the number of observations n, it is impossible to apply classical statistical models such as standard linear regression since the design matrix $\boldsymbol{X}$ is no longer of full rank. A common remedy to this problem is to *regularize* the model by adding a term to the objective function that punishes models with large coefficients ($\boldsymbol{\beta}$). If we let $g(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$ be the original objective function—which when minimized improves the model's fit to the data $(\boldsymbol{X}, \boldsymbol{y})$—then

$$f(\beta_0, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = g(\beta_0, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) + h(\boldsymbol{\beta})$$

is a composite function within which we have added a penalty term $h(\boldsymbol{\beta})$. In contrast to g, this penalty depends only on $\boldsymbol{\beta}$. The intercept, β_0, is not typically penalized.

Some of the most common penalties are the ℓ_1 norm and squared ℓ_2 norm penalties, that is $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ or $h(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$[1], which, if h is the standard ordinary least-squares objective, represent lasso (Tibshirani, 1996; Santosa & Symes, 1986; Donoho & Johnstone, 1994) and ridge (Tikhonov) regression respectively. Other common penalities include SLOPE (Bogdan et al., 2013; 2015), the minimax-concave penalty (MCP) (Zhang, 2010), hinge loss (used in support vector machines (Cortes & Vapnik, 1995)) and smoothly-clipped absolute deviation (SCAD) (Fan & Li, 2001). Many of these penalities—indeed all of the previously mentioned ones—shrink coefficients in proportion to their sizes.

The issue with this type of shrinkage is that it is typically sensitive to the scales of the features in $\boldsymbol{X}$. A common remedy is to *normalize* the features before fitting the model by translating and dividing each column

[1] Division by two in this case is used only for convenience.

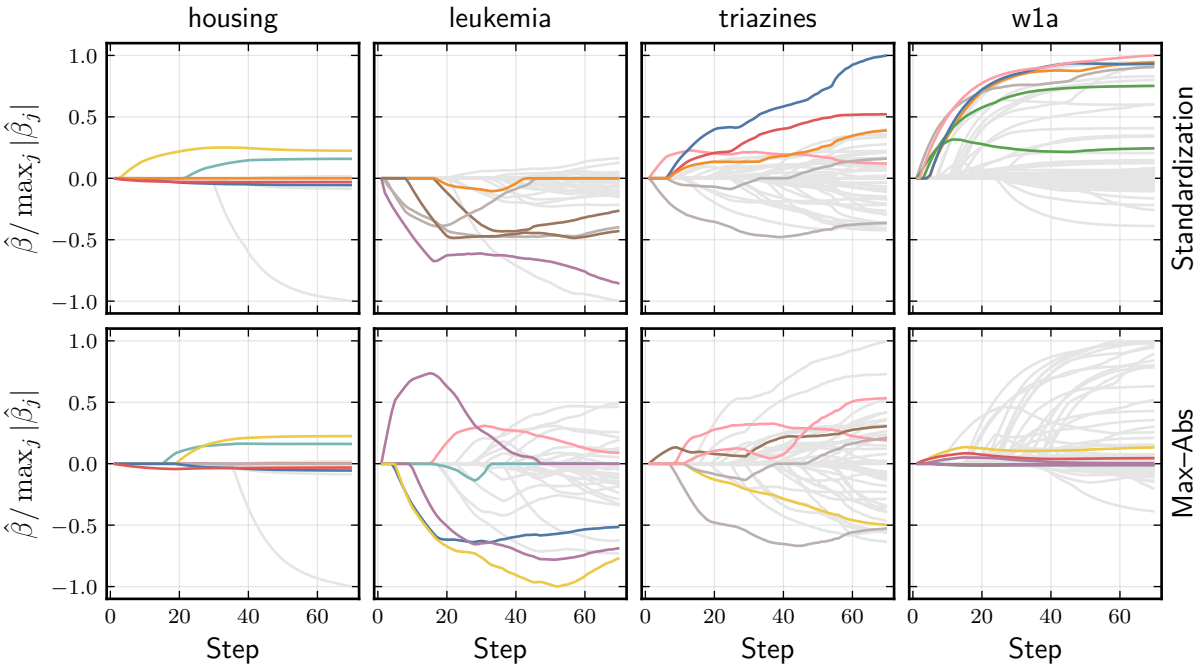


Figure 1: Lasso paths for real datasets using two types of normalization: standardization and maximum absolute value scaling (max–abs). We have fit the lasso path to four different datasets: `housing` (Harrison & Rubinfeld, 1978), `leukemia` (Golub et al., 1999), `triazines` (King), and `w1a` (Platt, 1998). For each dataset, we have colored the coefficients if they were among the first five features to become active in under either of the two types of normalization schemes. We see that the paths differ with regards to the size as well as the signs of the coefficients, and that, in addition, the coefficients to become active first differ between the normalization types.

by respective translation and scaling factors. For some problems, such factors may arise naturally from knowledge of the problem at hand. A researcher may for instance have collected data on coordinates within a limited area and know that the coordinates are measured in meters. Often, however, these scaling factors must be estimated from data. The most popular choices for this type of scaling are based only on the marginal distributions of the features. Some types of normalization, such as that applied in the adaptive lasso[2] (Zou, 2006), however, are based on the conditional distributions of the features and the response. After fitting the model, the estimated coefficients are then usually returned to their original scale. Another reason for normalizing the features is to improve the performance and stability of optimization algorithms used to fit the model. We will not cover this aspect in this paper, but note that it is an important one.

In most sources and discussions on regularized methods, normalization is typically treated as a preprocessing step—separate from modeling. As we will show in this paper, however, the type of normalization used can have a critical effect on the estimated model, sometimes leading to entirely different conclusions with regard to feature importance as well as predictive performance. As a first example of this, consider Figure 1, which displays the lasso paths for four real data sets and two different types of normalization. Each panel shows the union of the first five predictors picked under either normalization scheme. The choice of normalization can have a significant impact on the estimated model. In the case of the `leukemia` data set, for instance, the models are starkly different with respect to both the identities of the features selected as well as their signs and magnitudes.

In addition, discussions on the choice of normalization are often focused on computational aspects and data storage requirements, rather than on the statistical properties of the choice of normalization. In our paper,

[2]The adaptive lasso typically uses estimates of the regression coefficients, typically from ordinary-least squares or ridge regression, to scale the features with.

we argue that normalization should rather we considered as an integral part of the model and that it is problematic to base the choice of normalization on the type of data storage, which implicitly encodes the belief that the information in a data set is different if it is stored in a sparse viz-a-viz dense format. At the time of writing, for instance, the popular machine learning library `scikit-learn` (scikit-learn developers, 2024) recommends max–abs scaling in the case of sparse data.

2 Preliminaries

Throughout this paper, we assume that the data is generated from a linear model, that is,

$$y_i = \beta_0^* + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}^* + \varepsilon_i \quad \text{for} \quad i \in \{1, 2, \dots, n\},$$

where we use β_0^* and $\boldsymbol{\beta}^*$ to denote the true intercept and coefficients, respectively, and ε_i to denote measurement noise. $\boldsymbol{X}$ is the $n \times p$ design matrix with columns $\boldsymbol{x}_j$ and $\boldsymbol{y}$ the $n \times 1$ response vector. Furthermore, we use $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}$ to denote our estimates of the intercept and coefficients and use β_0 and β to refer to corresponding variables in the optimization problem. Unless otherwise stated, we assume $\boldsymbol{X}$, β_0^*, and $\boldsymbol{\beta}^*$ to be fixed.

There is ambiguity regarding many of the key terms in the field of normalization. *Scaling*, *standardization*, and *normalizaton* are for instance used interchangeably throughout the literature. Here, we define *normalization* as the process of centering and scaling the feature matrix, which we formalize in Definition 2.1.

Definition 2.1 (Normalization)**.** Let $\tilde{\boldsymbol{X}}$ be the normalized feature matrix, with elements

$$\tilde{x}_{ij} = \frac{x_{ij} - c_j}{s_j},$$

where x_{ij} is an element of the (unnormalized) feature matrix $\boldsymbol{X}$ and c_j and s_j are the *centering* and *scaling* factors respectively.

Some authors refer to this procedure as *standardization*, but here we define standardization only as the case when centering with the arithmetic mean and scaling with the (uncorrected) standard deviation. Also note that normalization is sometimes defined as the process of scaling the *samples*, rather than the features. We will not consider this type of normalization in this paper.

2.1 Types of Normalization

There are many different strategies for normalizing the design matrix. We list a few of the most common choices in Table 1.

Table 1: Common ways to normalize a matrix of features

Normalization	Centering (c_j)	Scaling (s_j)
Standardization	$\frac{1}{n}\sum_{i=1}^n x_{ij}$	$\sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$
Max–Abs	0	$\max_i(\lvert x_{ij}\rvert)$
Min–Max	$\min_i(x_{ij})$	$\max_i(x_{ij}) - \min_i(x_{ij})$
Norm Scaling	0	$\lVert \boldsymbol{x}_j \rVert_p$, $p \in \{1, 2, \dots\}$
Adaptive Lasso	0	β_j^{OLS}

Standardization is perhaps the most common type of normalization, at least in the field of statistics. It is sometimes known as *z-scoring* or *z-transformation*. One of the benefits of using standardization is that it simplifies certain aspects of fitting the model. For instance, the intercept term $\hat{\beta}_0$ is equal to the mean of the response $\boldsymbol{y}$. For regularized methods, it is typically the case that we standardize with the uncorrected sample standard deviation (division by n). The downside of standardization is that it involves centering by

the mean, which typically destroys sparsity in the data structure. This is not a problem when the data is stored as a dense matrix; but when the data is sparse, it can lead to a significant increases in memory usage and processing time.

A common alternative to standardization, particularly when data is sparse, is to scale the features by their maximum absolute value (max–abs normalization). This method has no impact on binary data[3], and therefore retains sparsity. For other types of data, it scales the features to take values in the range $[-1, 1]$. Since the scaling is determined by a single value for each feature, the method is naturally sensitive to outliers. In addition, it is for many types of continuous data, such as normally distributed data, the case that the sample maximum depends on the sample size, which makes the method problematic for much continuous data. In Theorem A.1 (Appendix A), we study how this effect comes into play in the case when the feature is normally distributed.

Min-max normalization scales the data to lie in $[0, 1]$. As with maximum absolute value scaling, min-max normalization retains sparsity and also shares its sensitivity to outliers and sample size. Unlike max–abs scaling, min–max scaling is not sensitive to the *location* of the data, only its *spread*. Norm-scaling, scaling by a norm, is seldom used in practice and more often encountered in theoretical work. The norm can be any p-norm, and the choice of p will determine the scaling. Standard choices are $p = 1$, when the scaling is the sum of the absolute values of the features, and $p = 2$, where it is the Euclidean norm. A special case of normalization is the adaptive lasso (Zou, 2006), which is a two-step procedure. In the first step, a model, often ordinary least-squares regression (OLS) or ridge regression, is fit to the data. The estimated coefficients from the model are then used to scale the features.

2.2 The Lasso and Ridge Regression

From now on, we will direct our focus on ridge regression and the lasso. Both of these models are special cases of the elastic net (Zou & Hastie, 2005), which is the ordinary-least squares regression objective regularized by a combination of the ℓ_1 and squared ℓ_2 norms. For the normalized feature matrix $\tilde{\boldsymbol{X}}$, the elastic net is represented by the following convex optimization problem:

$$\underset{\beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p}{\text{minimize}} \left(f(\beta_0, \boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}, \lambda_1, \lambda_2) = \frac{1}{2} \|\boldsymbol{y} - \beta_0 - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2 \right). \tag{1}$$

We define $(\hat{\beta}_0^{(n)}, \hat{\boldsymbol{\beta}}^{(n)})$ as a solution to the optimization problem in Equation (1). When $\lambda_1 > 0$ and $\lambda_2 = 0$, the elastic net is equivalent to the lasso, and when $\lambda_1 = 0$ and $\lambda_2 > 0$, it is equivalent to ridge regression. Expanding f in Equation (1), we have

$$\frac{1}{2} \left(\boldsymbol{y}^\intercal \boldsymbol{y} - 2(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \beta_0)^\intercal \boldsymbol{y} + (\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \beta_0)^\intercal (\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \beta_0) \right) + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta}\|_2^2.$$

Taking the subdifferential with respect to $\boldsymbol{\beta}$ and β_0, the KKT stationarity condition yields the following system of equations:

$$\begin{cases} \tilde{\boldsymbol{X}}^\intercal (\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \beta_0 - \boldsymbol{y}) + \lambda_1 g + \lambda_2 \boldsymbol{\beta} \ni \boldsymbol{0}, \\ n\beta_0 + (\tilde{\boldsymbol{X}}\boldsymbol{\beta})^\intercal \boldsymbol{1} - \boldsymbol{y}^\intercal \boldsymbol{1} = 0, \end{cases} \tag{2}$$

where g is a subgradient of the ℓ_1 norm that has elements g_i such that

$$g_i \in \begin{cases} \{\operatorname{sign} \beta_i\} & \text{if } \beta_i \neq 0, \\ [-1, 1] & \text{otherwise.} \end{cases}$$

2.3 Orthogonal Features

If the features of the normalized design matrix are orthogonal, that is, $\tilde{\boldsymbol{X}}^\intercal \tilde{\boldsymbol{X}} = \operatorname{diag}\left(\tilde{\boldsymbol{x}}_1^\intercal \tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_p^\intercal \tilde{\boldsymbol{x}}_p\right)$, then Equation (2) can be decomposed into a set of $p + 1$ conditions:

$$\begin{cases} \tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j \beta_j + \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{1} \beta_0 - \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y} + \lambda_2 \beta_j + \lambda_1 g \ni 0, \quad j = 1, \ldots, p, \\ n\beta_0 + (\tilde{\boldsymbol{X}}\boldsymbol{\beta})^\intercal \boldsymbol{1} - \boldsymbol{y}^\intercal \boldsymbol{1} = 0. \end{cases}$$

[3]Except in the extreme case when all values are 0.

The inclusion of an intercept, β_0, ensures that the locations of the features (their means) does not affect the solution (except for the intercept itself). Therefore, we will from now on assume that the features are mean-centered, that is, $c_j = \bar{\boldsymbol{x}}_j$ for all j and therefore $\tilde{\boldsymbol{x}}_j^\intercal \mathbf{1} = 0$. A solution to the system of equations is then given by the following set of equations (Donoho & Johnstone, 1994):

$$\hat{\beta}_j^{(n)} = \frac{\mathrm{S}_{\lambda_1}\left(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y}\right)}{\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j + \lambda_2}, \qquad \hat{\beta}_0^{(n)} = \frac{\boldsymbol{y}^\intercal \mathbf{1}}{n},$$

where S is the soft-thresholding operator, defined as

$$\mathrm{S}_\lambda(z) = \operatorname{sign}(z)\max(|z| - \lambda, 0) = \mathrm{I}_{|z|>\lambda}\left(z - \operatorname{sign}(z)\lambda\right).$$

2.4 Rescaling Regression Coefficients

Normalization changes the optimization problem and therefore its solution, the coefficients, which will now be on the scale of the normalized features. We, however, are interested in $\hat{\boldsymbol{\beta}}$: the coefficients on the scale of the original problem. To obtain these, we transform the coefficients from the normalized poblem, $\hat{\beta}_j^{(n)}$, back via

$$\hat{\beta}_j = \frac{\hat{\beta}_j^{(n)}}{s_j} \quad \text{for} \quad j = 1, 2, \ldots, p. \tag{3}$$

There is a similar transformation for the intercept which we omit here since we are not interested in it.

3 Bias and Variance of the Elastic Net Estimator

Now, assume that $\boldsymbol{X}$ is fixed and that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where ε_i is identically and independently distributed noise with mean zero and finite variance σ_ε^2. As in the previous section, we assume that the feature vectors are orthogonal. We are interested in the expected value of Equation (3), $\mathrm{E}\,\hat{\beta}_j$. Let

$$Z = \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y} = \tilde{\boldsymbol{x}}_j^\intercal (\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}) = \tilde{\boldsymbol{x}}_j^\intercal (\boldsymbol{x}_j \beta_j^* + \boldsymbol{\varepsilon}) \qquad \text{and} \qquad d_j = s_j(\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j + \lambda_2)$$

so that $\hat{\beta}_j = \mathrm{S}_{\lambda_1}(Z)/d_j$. Since d_j is fixed under our assumptions, we will direct most of our focus towards $S_{\lambda_1}(Z)$. First observe that

$$\begin{aligned}
\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j &= \frac{1}{s_j^2}(\boldsymbol{x}_j - c_j)^\intercal (\boldsymbol{x}_j - c_j) = \frac{\boldsymbol{x}_j^\intercal \boldsymbol{x}_j - nc_j^2}{s_j^2} = \frac{nv_j}{s_j^2}, \\
\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{x}_j &= \frac{1}{s_j}(\boldsymbol{x}_j^\intercal \boldsymbol{x}_j - \boldsymbol{x}_j^\intercal \mathbf{1} c_j) = \frac{nv_j}{s_j},
\end{aligned}$$

where v_j is the uncorrected sample variance of $\boldsymbol{x}_j$. This means that

$$Z = \frac{\beta_j^* n v_j - \boldsymbol{x}_j^\intercal \boldsymbol{\varepsilon}}{s_j} \qquad \text{and} \qquad d_j = s_j\left(\frac{nv_j}{s_j^2} + \lambda_2\right). \tag{4}$$

For the expected value and variance of Z we then have

$$\begin{aligned}
\mathrm{E}\,Z = \mu &= \mathrm{E}\left(\tilde{\boldsymbol{x}}_j^\intercal (\boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon})\right) = \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{x}_j \beta_j, \\
\operatorname{Var} Z = \sigma^2 &= \operatorname{Var}\left(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{\varepsilon}\right) = \tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j \sigma_\varepsilon^2.
\end{aligned}$$

The expected value of the soft-thresholding estimator is

$$\begin{aligned}
\mathrm{E}\,\mathrm{S}_\lambda(Z) &= \int_{-\infty}^{\infty} \mathrm{S}_\lambda(z) f_Z(z)\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} \mathrm{I}_{|z|>\lambda}(z - \operatorname{sign}(z)\lambda) f_Z(z)\,\mathrm{d}z \\
&= \int_{-\infty}^{-\lambda} (z + \lambda) f_Z(z)\,\mathrm{d}z + \int_{\lambda}^{\infty} (z - \lambda) f_Z(z)\,\mathrm{d}z.
\end{aligned}$$

And then the bias of $\hat{\beta}_j$ with respect to the true coefficient β_j^* is

$$\mathrm{E}\,\hat{\beta}_j - \beta_j^* = \frac{1}{d_j}\,\mathrm{E}\,\mathrm{S}_\lambda(Z) - \beta_j^*.$$

Finally, we note that the variance of the soft-thresholding estimator is

$$\operatorname{Var} S_\lambda(Z) = \int_{-\infty}^{-\lambda} (z+\lambda)^2 f_Z(z)\,\mathrm{d}z + \int_{\lambda}^{\infty} (z-\lambda)^2 f_Z(z)\,\mathrm{d}z - (\mathrm{E}\,\mathrm{S}_\lambda(Z))^2 \tag{5}$$

and that the variance of the elastic net estimator is therefore

$$\operatorname{Var}\hat{\beta}_j = \frac{1}{d_j^2}\operatorname{Var}\mathrm{S}_\lambda(Z). \tag{6}$$

3.1 Normally Distributed Noise

Next, we add the additional assumption that $\boldsymbol{\varepsilon}$ is normally distributed. Then

$$Z \sim \mathrm{Normal}\left(\mu = \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{x}_j \beta_j, \sigma^2 = \tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j \sigma_\varepsilon^2\right).$$

Let $\theta = -\mu - \lambda_1$ and $\gamma = \mu - \lambda_1$. Then the expected value of soft-thresholding of Z is

$$\begin{aligned}\mathrm{E}\,\mathrm{S}_{\lambda_1}(Z) &= \int_{-\infty}^{\frac{\theta}{\sigma}} (\sigma u - \theta)\,\phi(u)\,\mathrm{d}u + \int_{-\frac{\gamma}{\sigma}}^{\infty} (\sigma u + \gamma)\,\phi(u)\,\mathrm{d}u \\ &= -\theta\,\Phi\left(\frac{\theta}{\sigma}\right) - \sigma\,\phi\left(\frac{\theta}{\sigma}\right) + \gamma\,\Phi\left(\frac{\gamma}{\sigma}\right) + \sigma\,\phi\left(\frac{\gamma}{\sigma}\right)\end{aligned} \tag{7}$$

where $\phi(u)$ and $\Phi(u)$ are the probability density and cumulative distribution functions of the standard normal distribution, respectively.

Next, we consider what the variance of the elastic net estimator looks like. Starting with the first term on the left-hand side of Equation (5), we have

$$\begin{aligned}\int_{-\infty}^{-\lambda_1} (z+\lambda_1)^2 f_Z(z)\,\mathrm{d}z &= \sigma^2 \int_{-\infty}^{\frac{\theta}{\sigma}} y^2\,\phi(y)\,\mathrm{d}y + 2\theta\sigma \int_{-\infty}^{\frac{\theta}{\sigma}} y\,\phi(y)\,\mathrm{d}y + \theta^2 \int_{-\infty}^{\frac{\theta}{\sigma}} \phi(y)\,\mathrm{d}y \\ &= \frac{\sigma^2}{2}\left(\operatorname{erf}\left(\frac{\theta}{\sigma\sqrt{2}}\right) - \frac{\theta}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\theta^2}{2\sigma^2}\right) + 1\right) + 2\theta\sigma\,\phi\left(\frac{\theta}{\sigma}\right) + \theta^2\,\Phi\left(\frac{\theta}{\sigma}\right).\end{aligned} \tag{8}$$

Similar computations for the second term on the left-hand side of Equation (5) yield

$$\int_{\lambda_1}^{\infty} (z-\lambda_1)^2 f_Z(z)\,\mathrm{d}z = \frac{\sigma^2}{2}\left(\operatorname{erf}\left(\frac{\gamma}{\sigma\sqrt{2}}\right) - \frac{\gamma}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2\sigma^2}\right) + 1\right) + 2\gamma\sigma\,\phi\left(\frac{\gamma}{\sigma}\right) + \gamma^2\,\Phi\left(\frac{\gamma}{\sigma}\right). \tag{9}$$

Plugging Equations (7) to (9) into Equation (6) yields the variance of the estimator. Consequently, we can also compute the mean-squared error via the bias-variance decomposition

$$\mathrm{MSE}(\hat{\beta}_j, \beta_j^*) = \operatorname{Var}\hat{\beta}_j + \left(\mathrm{E}\,\hat{\beta}_j - \beta_j^*\right)^2.$$

3.2 Binary Features

The main focus in this paper is the case when $\boldsymbol{x_j}$ is a binary feature with class balance $q = \bar{\boldsymbol{x}}_j$, that is, $x_{ij} \in \{0, 1\}$ for all i and $\sum_{i=1}^n x_{ij} = nq$. In this case, inserting $v_j = (q - q^2)$ (the uncorrected sample variance for a binary feature) into Equation (4), we have

$$Z = \frac{\beta_j^* n(q-q^2) - \boldsymbol{x}_j^\intercal \boldsymbol{\varepsilon}}{s_j}, \qquad d_j = s_j\left(\frac{n(q-q^2)}{s_j^2} + \lambda_2\right),$$

and consequently

$$\mu = \frac{\beta_j^* n(q-q^2)}{s_j} \qquad \text{and} \qquad \sigma^2 = \frac{\sigma_\varepsilon^2 n(q-q^2)}{s_j^2}.$$

We will allow ourselves to abuse notation and overload the definitions of μ, σ^2, and d_j as functions of q. Then, an expression for the expected value of the elastic net estimate with respect to q can be obtained by plugging in μ and σ into Equation (7).

The presence of the factor $q - q^2$ in μ, σ^2, and d_j means that there is a relationship between class balance and the elastic net estimator and that this relationship is mediated by the scaling factor s_j. To achieve some initial intuition for this relationship, we begin by considering the noiseless case ($\sigma_\varepsilon = 0$) in which, inserting μ and d_j into Equation (3) yields

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y})}{s_j\left(\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j + \lambda_2\right)} = \frac{\mathrm{S}_{\lambda_1}\left(\frac{\beta_j^* n(q-q^2)}{s_j}\right)}{s_j\left(\frac{n(q-q^2)}{s_j^2} + \lambda_2\right)}. \tag{10}$$

This expression shows that the class balance, q, directly affects the estimator. For values of q close to 0 or 1, the input into the soft-thresholding part of the estimator will diminish and consequently force the estimate to zero, that is, unless we use the scaling factor $s_j = (q - q^2)$, in which case the soft-thresholding part will be unaffected by class imbalance. This choice will not, however, mitigate the impact of class imbalance on the ridge part of the estimator, for which we would instead need $s_j = \sqrt{q - q^2}$. For any other choices of δ, such as $\delta = 0$, q will affect the estimator through both the ridge and lasso parts.

Based on these facts, we will consider the scaling parameterization $s_j = (q - q^2)^\delta$, $\delta \geq 0$. This includes the cases that we are primarily interested in, that is, $\delta = 0$ (no scaling), $\delta = 1/2$ (standard-deviation scaling), and $\delta = 1$ (variance scaling). Note that the last of these types, variance scaling, is not a standard type of normalization; yet, as we have already seen, it has some interesting properties in the context of binary features.

Another interesting fact about Equation (10), which holds also in the noisy situation, is that even when the binary feature is balanced ($q = 1/2$), normalization will still have an effect on the estimator. Using $\delta = 0$, for instance, leads the true coefficient β_j^* in the input to S_λ to be scaled by $n(q - q^2) = n/4$. For $\delta = 1$, there would be, in contrast, be no scaling in the class-balanced case. And for $\delta = 1/2$, the scaling factor is $n/2$. Generalizing this, we see that to achieve equivalent scaling in the class-balanced case for all types of normalization, under our parameterization, we would need to use

$$s_j = 4^{\delta-1}(q - q^2)^\delta.$$

This only resolves the issue for the lasso. To achieve a similar effect for ridge regression, we would need another (but similar) modification. Since all features are binary under our current assumptions, however, we will for now just assume that we scale λ_1 and λ_2 to account for this effect,[4] which is equivalent to modifying s_j. We will return to this issue later in Section 3.3 where we consider mixes of binary and normally distributed features in, in which case this has significant implications.

We now leave the noise-less scenario and proceed to consider how class balance affects the probability of selection, bias, and variance of the elastic net estimator, starting with the first of these. A consequence of the normal error distribution and consequent normal distribution of Z is that the probability of selection in the

[4] We do this in all of the following examples.

elastic net problem is given analytically by

$$
\begin{aligned}
\Pr\left(\hat{\beta}_j \neq 0\right) &= \Pr\left(\mathrm{S}_{\lambda_1}(Z) \neq 0\right) \\
&= \Pr\left(Z > \lambda_1\right) + \Pr\left(Z < -\lambda_1\right) \\
&= \Phi\left(\frac{\mu - \lambda_1}{\sigma}\right) + \Phi\left(\frac{-\mu - \lambda_1}{\sigma}\right). \\
&= \Phi\left(\frac{\beta_j^* n(q-q^2)^{1/2} - \lambda_1 (q-q^2)^{\delta - 1/2}}{\sigma_\varepsilon \sqrt{n}}\right) + \Phi\left(\frac{-\beta_j^* n(q-q^2)^{1/2} - \lambda_1 (q-q^2)^{\delta - 1/2}}{\sigma_\varepsilon \sqrt{n}}\right).
\end{aligned}
$$

Letting $\theta = -\mu - \lambda_1$ and $\gamma = \mu - \lambda_1$, we can express the probability of selection in the limit as $q \to 1^+$ as

$$
\lim_{q \to 1^+} \Pr(\hat{\beta}_j \neq 0) = \begin{cases} 0 & \text{if } 0 \leq \delta < \frac{1}{2}, \\ 2\,\Phi\left(-\frac{\lambda_1}{\sigma_\varepsilon \sqrt{n}}\right) & \text{if } \delta = \frac{1}{2}, \\ 1 & \text{if } \delta > \frac{1}{2}. \end{cases}
$$

In Figure 2, we plot this probability for various settings of δ for a single feature. Our intuition from the noise-less case holds: δ mitigates the influence of class imbalance on selection probability. The lower the value of δ, the larger the effect of class imbalance becomes. Note that the probability of selection initially decreases also in the case when $\delta \geq 1$. This is a consequence of increased variance of Z dues to the scaling factor that scales the measurement noise σ_ε^2 upwards. Then, as q approaches 1, the probability picks up again and eventually approaches 1 for these $\delta \in \{1, 1.5\}$. The reason for this is that the variance of Z eventually explodes (again due to the scaling), which ultimately removes the soft-thresholding effect altogether. Note that the selection probability is unaffected by λ_2 (the ridge penalty), so these results hold for any value of it.

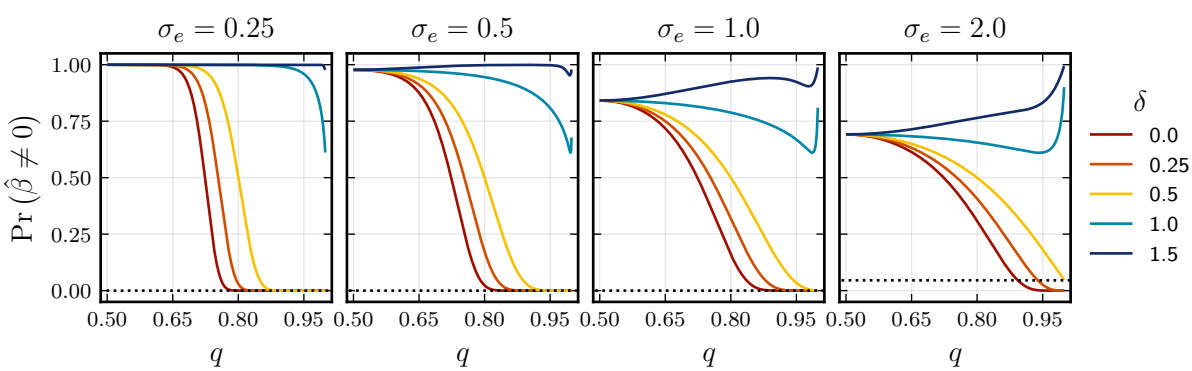


Figure 2: Probability of selection in the lasso given a measurement noise level σ_ε, a regularization parameter λ_1, and a class balance q. The scaling factor is parameterized by $s_j = (q - q^2)^\delta$, $\delta \geq 0$. The dotted line represents the asymptotic limit for the standardization case, $\delta = 1/2$.

Now we turn to the impact of class imbalance on bias and variance of the elastic net estimator. We begin, in Theorem 3.1, by considering the expected value of the elastic net estimator in the limit as $q \to 1^+$.

Theorem 3.1. *If $\boldsymbol{x}_j$ is a binary feature with class balance $q \in (0,1)$, $\lambda_1 \in (0, \infty)$, $\lambda_2 \in [0, \infty)$, $\sigma_\varepsilon > 0$, and $s_j = (q - q^2)^\delta$, $\delta \geq 0$ then*

$$
\lim_{q \to 1^+} \mathrm{E}\,\hat{\beta}_j = \begin{cases} 0 & \textit{if } 0 \leq \delta < \frac{1}{2}, \\ \frac{2n\beta_j^*}{n + \lambda_2}\,\Phi\left(-\frac{\lambda_1}{\sigma_\varepsilon \sqrt{n}}\right) & \textit{if } \delta = \frac{1}{2}, \\ \beta_j^* & \textit{if } \delta > \frac{1}{2}. \end{cases}
$$

Theorem 3.1 shows that the bias of the elastic net estimator when $0 \leq \delta < 1/2$ approaches $-\beta_j^*$ as $q \to 1^+$. Interestingly, when $\delta = 1/2$ (standardization), the estimate does not in fact tend to zero. Instead, it

approaches the true coefficient scaled by the probability that a standard normal variable is smaller than $\beta_j^* \sqrt{n}\sigma_\varepsilon^{-1}$. For $\delta > 1/2$, the estimate is unbiased asymptotically, which is related to the scaled variance of the error term. Note that this unbiasedness is paralleled by a surge in variance and therefore also a rise in mean-squared error, and only serves to demonstrate that the cost of the decoupling of q is unbearable in the large noise–large imbalance scenario. In Theorem 3.2, we continue by studying the variance in the limit as $q \to 1^+$.

Theorem 3.2. *If $\boldsymbol{x}_j$ is a binary feature with class balance $q \in (0, 1)$ and $\lambda_1, \lambda_2 \in (0, \infty)$, $\sigma_\varepsilon > 0$, and $s_j = (q - q^2)^\delta$, $\delta \geq 0$, then*

$$\lim_{q \to 1^+} \operatorname{Var} \hat{\beta}_j = \begin{cases} 0 & \text{if } 0 \leq \delta < \frac{1}{2}, \\ \infty & \text{if } \delta \geq \frac{1}{2}. \end{cases}$$

Corollary 3.2.1 (Variance in Ridge Regression)**.** *Assume the conditions of Theorem 3.2 hold, except that $\lambda_1 = 0$. Then*

$$\lim_{q \to 1^+} \operatorname{Var} \hat{\beta}_j = \begin{cases} 0 & \text{if } 0 \leq \delta < 1/4, \\ \frac{\sigma_\varepsilon^2 n}{\lambda_2^2} & \text{if } \delta = 1/4, \\ \infty & \text{if } \delta > 1/4. \end{cases}$$

Theorem 3.2 formally proves the asymptotic variance effects of our scaling parameter s_j which we have already discussed in the context of selection probability and bias. Taken together with the results from Theorem 3.1, this suggests that the choice of scaling parameter, at least in the case of our specific parameterization, introduces a bias–variance tradeoff with respect to δ: to reduce bias (with respect to q), we need to pay the cost of increased variance.

In Figure 3, we now visualize bias, variance, and mean-squared error for ranges of class balance and various noise-level settings for a lasso problem. The figure demonstrates the bias–variance tradeoff that our asymptotic results suggested and indicates that the optimal choice of δ is related to the noise level in the data. Since this level is unknown for most data sets, it suggests there might be value in selecting δ through hyper-optimization as is typically done for the other hyper-parameters in the elastic net (λ_1, λ_2)

So far, we have only considered a single binary feature. But under the assumption of orthogonal features, it is straightforward to introduce multiple binary features. In a first example, we study how the power of correctly detecting $k = 10$ signals under q linearly spaced in $[0.5, 0.99]$ (Figure 5a). We set $\beta_j^* = 2$ for each of the signals, use $n = 100\,000$, and let $\sigma_\varepsilon = 1$. The level of regularization is set to $\lambda_1 = n4^\delta/10$. As we can see, the power is directly related to q and for unbalanced features stronger the higher the choice of δ is.

We also consider a version of the same setup, but with p linearly spaced in $[20, 100]$ to compute the normalized mean-squared error (NMSE) and false discovery rate (FDR) (Figure 5b). As before, we let $k = 10$ and consider three different levels of class imbalance. The remaining $p - k$ features have class balances spaced evenly on a logarithmic scale from 0.5 to 0.99. Unsurprisingly, the increase in power gained from selecting $\delta = 1$ imposes increased false discovery rates. The mean-squared error depends on the class balance. For class-balanced signals, $\delta \in \{0, 1/2\}$ proves to b the best choice, while for unbalanced signals, $\delta = 1$ is the best choice. In the case when $q = 0.99$, the model under scaling with $\delta = 0$ is altogether unable to detect any of the true signals, instead picking up on the noisy, but better-balanced, features.

In Section 4, we will continue to study binary features in simulated experiments. For now, however, we will turn to the case of mixed data.

3.3 Mixed Data

In this section, we consider the case where the features are made up of a mix of continuous and and binary features. Throughout the section, we will continue to assume that $\boldsymbol{X}$ is fixed and that the features are orthogonal to one another. As in our theoretical results, we will also restrict our focus to the case where the continuous features are normally distributed.

A fundamental problem in the context of mixed data is how to put the binary and normal features on the same scale, which we need to do in order for regularization to be, roughly speaking, "fair", given that the

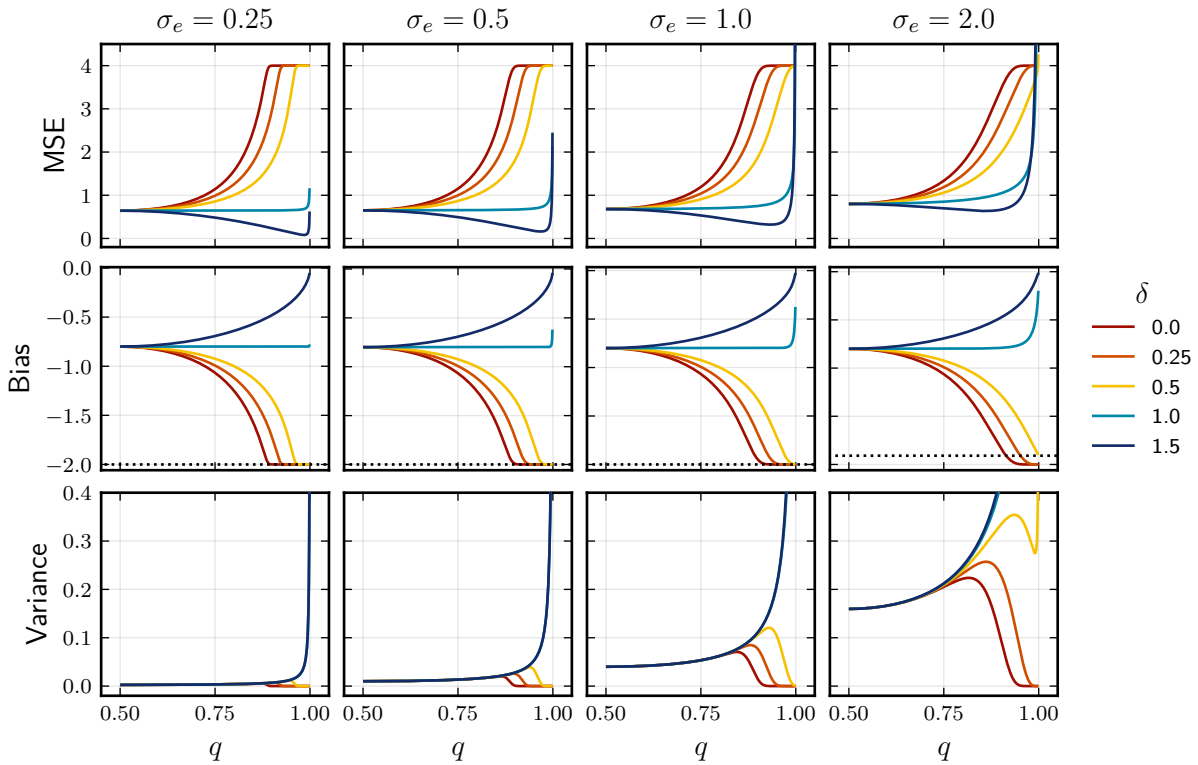


Figure 3: Bias, variance, and mean-squared error for a one-dimensional lasso problem. We show these measures for various noise levels (σ_ε), class balances (q), and scaling factors (δ). The dotted lines represent the asymptotic bias of the lasso estimator in the case of $\delta = 1/2$.

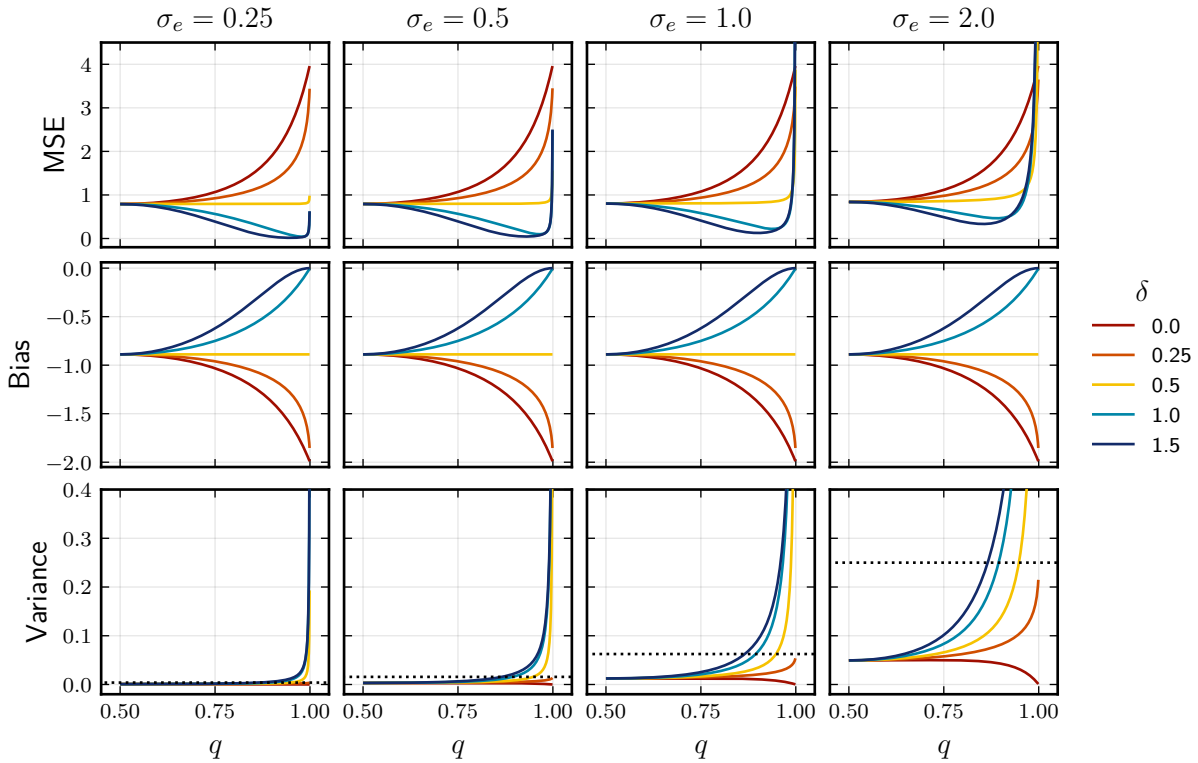
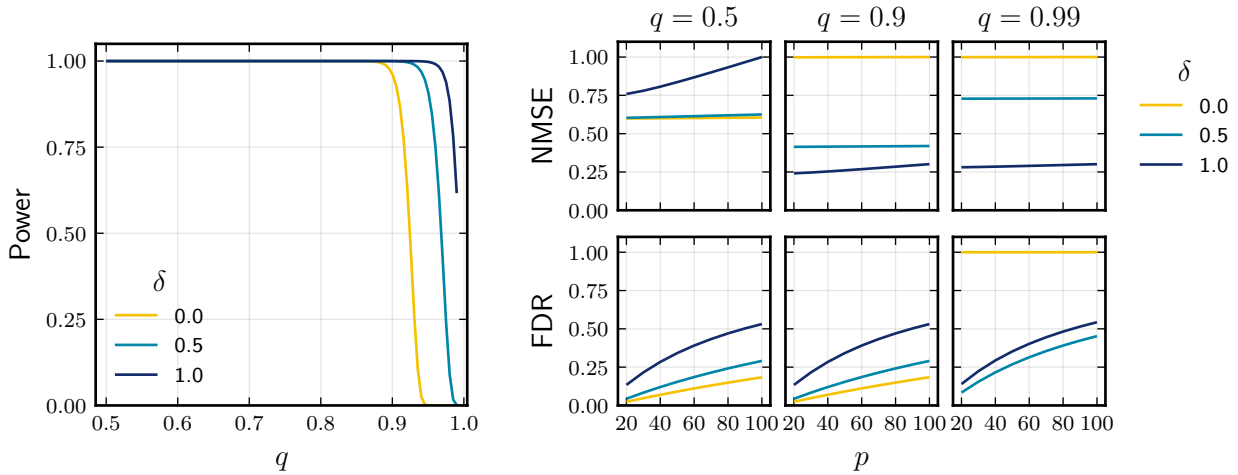


Figure 4: Bias, variance, and mean-squared error for one-dimensional ridge regression. We show these measures for various noise levels (σ_ε), class balances (q), and scaling factors (δ). The dotted lines represent the asymptotic bias of the lasso estimator in the case of $\delta = 1/2$.



(a) The power (probability of detecting all true signals) of the lasso. In our orthogonal setting, power is constant over p, which is why we have omitted the parameter in the plot.

(b) NMSE and FDR: the rate of coefficients incorrectly set to non-zero (false discoveries) to the total number of estimated coefficients that are nonzero (discoveries).

Figure 5: Normalized mean-squared error (NMSE), false discovery rate (FDR), and power for a lasso problem with $k = 10$ true signals (nonzero β_j^*), varying p, and $q \in [0.5, 0.99]$. The noise level is set at $\sigma_\varepsilon = 1$ and $\lambda_1 = 0.02$.

solution is sensitive to the scale of the features. In essence, we need to say something about how an effect associated with a one-unit change in the binary feature (a flip) relates to a one-unit change in the continuous feature. Since we assume our continuous feature to be normal, however, we will instead reason about change in terms of standard deviations of the normal feature.

To setup this situation more formally, we will say that the effect of a binary feature $\boldsymbol{x}_1$ and a normal feature $\boldsymbol{x}_2$ are *comparable* if

$$\beta_1^* = \kappa\sigma_2\beta_2^*,$$

where σ_2 is the standard deviation of $\boldsymbol{x}_2$ and $\kappa > 0$ is a scaling factor that represents the number of standard deviations (of the continuous feature) we consider achieves comparability between the features' effects. (Note that $\sigma_2\beta_2^*$ is just the standardized coefficient for the normal feature.) We illustrate this notion of comparability by a couple of examples.

Example 3.1. Assume $\kappa = 2$. If $\boldsymbol{x}_2$ is sampled from $\text{Normal}(\mu, 1/2^2)$, then the effects of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are comparable if $\beta_1^* = \beta_2^*$.

Example 3.2. Assume $\kappa = 1$. If $\boldsymbol{x}_2$ is sampled from $\text{Normal}(\mu, 2^2)$, then the effects of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are comparable if $\beta_1^* = 2\beta_2^*$.

Note that this definition refers to the data-generating mechanism, and not the regularized estimates. What we ultimately want for comparability, however, is for the following relationship to hold:

$$\hat{\beta}_1 = \kappa\sigma_2\hat{\beta}_2.$$

Put plainly, we want the effects of regularization to be distributed evenly across the estimates. The crux of the problem is how to choose the scaling factor s_j for the binary features in order to achieve this effect for a given κ. Let us assume that we have two features, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, where $\boldsymbol{x}_1$ is binary and $\boldsymbol{x}_2$ is normally distributed and that their effects are comparable in the sense given above. Then it should hold that

$$\begin{aligned}
\hat{\beta}_1 &= \kappa\sigma_2\hat{\beta}_2 && \Longrightarrow \\
\frac{\mathrm{S}_{\lambda_1}(\tilde{\boldsymbol{x}}_1^\intercal \boldsymbol{y})}{s_1\left(\tilde{\boldsymbol{x}}_1^\intercal\tilde{\boldsymbol{x}}_1 + \lambda_2\right)} &= \frac{\kappa\sigma_2\,\mathrm{S}_{\lambda_1}(\tilde{\boldsymbol{x}}_2^\intercal \boldsymbol{y})}{s_2\left(\tilde{\boldsymbol{x}}_2^\intercal\tilde{\boldsymbol{x}}_2 + \lambda_2\right)} && \Longrightarrow \\
\frac{\mathrm{S}_{\lambda_1}\left(\frac{n\beta_1^*(q-q^2)}{s_1}\right)}{s_1\left(\frac{n(q-q^2)}{s_1^2} + \lambda_2\right)} &= \frac{\kappa\,\mathrm{S}_{\lambda_1}\left(\frac{n\beta_1^*}{\kappa}\right)}{n + \lambda_2} && (11)
\end{aligned}$$

since we strandadize he normal feature and therefore $s_2 = \sigma_2$. For the lasso ($\lambda_2 = 0$) and ridge regression ($\lambda_1 = 0$), we observe that $s_1 = \kappa(q - q^2)$ and $s_1 = (q - q^2)^{1/2}$, respectively, are the values for which Equation (11) hold. In other words, we can achieve comparability in the lasso by scaling each binary feature with its variance times κ, the number of standard deviations we consider achieves comparability between the features' effects. And for ridge regression, we can achieve comparability by scaling with standard deviation, irrespective of κ.

For any other choices of s_1, equality can only hold for a specific level of class balance. If we let this level be q_0, then, to achieve equality for $\lambda_2 = 0$, we need $s_1 = \kappa(q_0 - q_0^2)^{1-\delta}(q - q^2)^\delta$. Similarly, for $\lambda_1 = 0$, we need $s_1 = (q_0 - q_0^2)^{1-2\delta}(q - q^2)^\delta$. In the sequel, we will assume that $q_0 = 1/2$, to have effects be equivalent for the class-balanced case.

Note that this also means that there is an implicit relationship between the strength of penalization for binary and normal features, which depends on the level of class balance and normalization type. This means, for instance, that even in the class-balanced case ($q = 1/2$), we have to account for the type of normalization if we want binary and normal features to be treated equally. For example, if we were to use $\delta = 0$ and fit the lasso, then Equation (11) for a binary feature with $q = 1/2$ becomes

$$\frac{4\,\mathrm{S}_{\lambda_1}\left(\frac{n\beta_1^*}{4}\right)}{n} = \frac{\kappa\,\mathrm{S}_{\lambda_1}\left(\frac{n\beta_1^*}{\kappa}\right)}{n},$$

which then implies $\kappa = 4$, which may or may not agree with our assumptions about comparability between these features' effects.

For the rest of this paper, we will use $\kappa = 2$. That is, we will say that the effects are comparable if the effect of a flip in the binary feature equals the effect of a two-standard deviation change in the normal feature. We base this argument on the discussion by Gelman (2008), who argues that the classical approach of comparing standardized coefficients[5] awards effects of continuous features undue strength for most real data, since a change from, for instance, the lower to the upper 16% of the distribution will equal approximately twice the effect of a change in the binary feature. Using two standard deviations as a comparability factor would, in contrast, equivocate this change with the flip of the binary feature, which we believe is a better default. We want to stress that the choice of κ should, if possible, in general be made on a case-by-case (feature-by-feature) basis, using all available knowledge about the data at hand. But, irrespective of this, we also want to emphasize that the choice should be made. If you do not make it explicitly, then it will be implicitly dictated through the combination of normalization and penalization types you use.

Finally, note that the reasoning of comparability above rests on the assumption of no noise. And we are, in fact, in general instead more interested in the expected value of the estimators, which depend on the noise level. In the case of large class-imbalances and large noise, for instance, our previous results (see Figure 3 for instance), suggest that the estimators for normally distributed and binary features will not be comparable in this case.

4 Experiments

In the following sections, we present the results of our experiments. We begin by examining the variability and bias in the estimates of the regression coefficients. We then move on to predictive performance and hyperparameter selection. We also consider the effect of class imbalance on the estimates of the regression coefficients. Finally, we look at the effect of interactions between features on the estimates of the regression coefficients.

In all cases where we use simulated data, we generate our response vector according to

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

with $\boldsymbol{\varepsilon} \sim \text{Normal}(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I})$, where $\boldsymbol{X}$ is the design matrix, $\boldsymbol{\beta}^*$ is the vector of true regression coefficients, and σ_ε^2 is the noise level.

We consider two types of features: binary and quasi-normal features. To generate binary vectors, we sample $\lceil qn \rceil$ indexes uniformly at random without replacement from $\{1, 2, \dots, n\}$ and set the corresponding elements to one and the remaining ones to zero. To generate quasi-normal features, we generate a linear sequence $\boldsymbol{w}$ with n values from 10^{-4} to $1 - 10^{-4}$, and set

$$x_{ij} = \Phi^{-1}(w_i)$$

and then shuffle the elements of $\boldsymbol{x}_j$ uniformly at random.

In each case, we fit either the lasso (the elastic net with $\lambda_1 = \alpha\lambda$) or ridge (the elastic net with $\lambda_2 = (1-\alpha)\lambda$). To normalize the data, we use standardization for all quasi-normally distributed features and otherwise

$$s_j = (q - q^2)^\delta,$$

which is equivalent to the (uncorrected) sample variance raised to the power of δ.

Throughout the experiments, we have used the Lasso.jl package (Kornblith, 2024) to fit lasso or ridge regression, which implements the coordinate descent algorithm by citetfriedman2010. All experiments were coded using the Julia programming language (Bezanson et al., 2017) and the code is availabe at https://github.com/jolars/normreg.

[5] Coefficients multiplied by the standard deviation of the respective feature.

4.1 Variability and Bias in Estimates

In our first experiment, we consider fitting the lasso to a simulated data set with $n = 500$ observations and $p = 1000$ features, out of which the first 20 features correspond to signals, with β_j^* decreasing linearly from 1 to 0.1. We introduce dependence between the features by copying the first $\lceil \rho n/2 \rceil$ values from the first feature to each of the following features. In addition, we set the class balance of the first 20 features so that it decreases linearly on a log-scale from 0.5 to 0.99. We estimate the regression coefficients using the lasso, setting $\lambda_1 = 2\sigma_\varepsilon\sqrt{2\log p}$ and compare the estimates to the true coefficients. We run the experiment for 50 iterations in each case and aggregate the results by reporting means and standard deviations.

The results (Figure 6) show that there is a considerable effect of class balance, particularly in the case of no scaling ($\delta = 0$), which corroborates our theoretical results from Section 3.2. At $q = 0.99$, for instance, the estimate ($\hat{\beta}_{20}$) is consistently zero when $\delta = 0$. There is a similar effect also in the case of standardization ($\delta = 1/2$), but it is less pronounced. For $\delta = 1$ (variance scaling), we see that the effect of class balance on the estimates is, if anything, the reverse when the class imbalance is severe. What is also clear is that the variance of the estimates increase with class imbalance and that this effect increases together with δ. The level of correlation between the features introduces additional variance in the estimates but also seems to increase the effect of class imbalance in the cases when $\delta = 0$ or $1/2$.

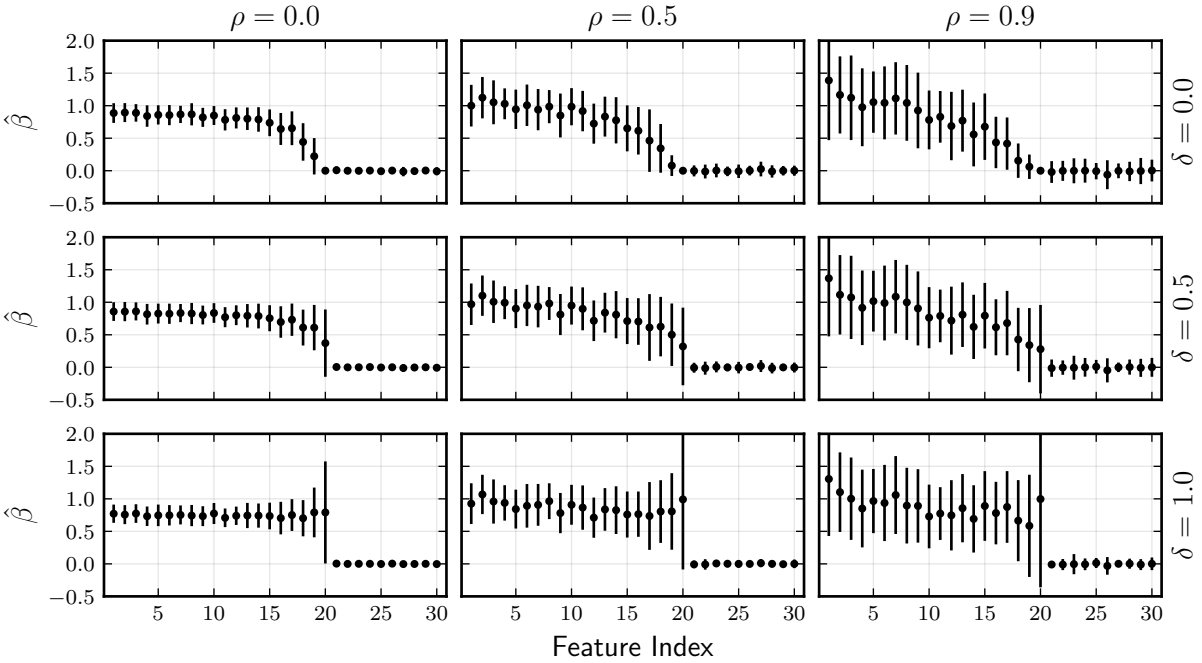


Figure 6: Estimates of the regression coefficients from the lasso, $\hat{\boldsymbol{\beta}}$, for the first 30 coefficients in the experiment. All of the features are binary and the first 20 features correspond to true signals with $\beta_j^* = 2$ and geometrically decreasing class balance from 0.5 to 0.99. The remaining features have a class balance $q_j \in [0.5, 0.99]$, distributed linearly among the features. The plot shows means and standard deviations averaged over 50 iterations.

4.2 Predictive Performance

In this experiment, we consider predictive performance in terms of mean-squared error of the lasso given different levels of class balance ($q \in \{0.5, 0.9, 0.99\}$), signal-to-noise ratio, and normalization (δ). As in the previous section, all of the features are binary, but here we have used $n = 300$, $p = 1000$. The $k = 10$ first features correspond to true signals with $\beta_j^* = 1$ and all have class balance q. To set signal-to-noise ratio levels, we rely on the same choice as in Hastie et al. (2020) and use a log-spaced sequence of values from 0.05 to 6.

To estimate prediction performance, we use a standard hold-out validation method equal splits for the training, validation, and test sets. We fit a full lasso path, parameterized by a log-spaced grid of 100 values[6], from $\lambda_{\max}$ (the value of λ at which the first feature enters the model) to $10^{-2}\lambda_{\max}$ on the training set and pick a λ based on validation set error. Then we compute the hold-out test set error and aggregate the results across 100 iterations.

The results (Figure 7) show that the optimal normalization type in terms of prediction power depends on the class balance of the true signals. If the imbalance is severe, then we gain from using $\delta = 1/2$ or 1, which gives a chance of recovering the true signals. If everything is balanced, however, then we do better by not scaling at all. In general, $\delta = 1/2$ works well for these specific combinations of settings.

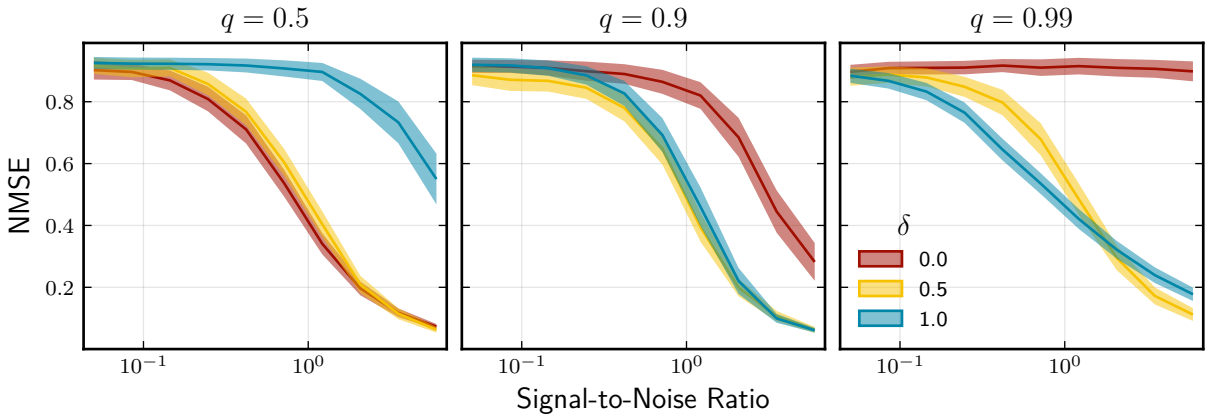


Figure 7: Normalized mean-squared prediction error in a lasso model for different types of normalization (δ), types of class imbalances (q), and signal-to-noise ratios (0.05 to 6) in a data set with $n = 300$ observations and $p = 1000$ features. The error is aggregated test-set error from hold-out validation with 100 observations in each of the training, validation, and test sets. The plot shows means and Student's t-based 95% confidence intervals.

4.3 Normalization as a Hyperparameter

Our previous results (particularly those from Section 4.2) suggest that the choice of normalization matters for predictive performance. These results have relied on knowledge of the measurement error (signal-to-noise ratio), which we do not have reliable estimates of in practice (at least not in the high-dimensional context). An alterative that, however, comes naturally as a consequence of our particular parameterization using δ, is to treat the choice of normalization as a hyperparameter and optimize over it. This is the approach we take in this experiment.

We set up a grid of λ values as in Section 4.2 and, in addition, also create a linearly spaced grid of δ values in $[0, 1]$. We split the data into a 50/50 training/validation set split and for each point in this two-dimensional grid fit the lasso or ridge to the training set and compute a hold-out validation set error. We do this for three data sets: `a1a` (Becker & Kohavi, 1996), `rhee2006` (Rhee et al., 2006), and `w1a` (Platt, 1998).

Table 2: Details of the real datasets used in the experiments

Dataset	n	p	Response
`w1a`	2477	300	Binary
`a1a`	1605	123	Binary
`rhee2006`	842	361	Continuous

[6]This is a standard choice of grid, used for instance by Friedman et al. (2010)

We show estimated level-curves of validation set error, in terms of normalized mean-squared error (NMSE), in Figure 8. For `a1a`, the lasso is generally quite insensitive to the type of normalization, even if the optimal value is around 0.2. For ridge regression, lower values of δ clearly work better. With the `w1a` data set, however, the relationship is flipped in the case of ridge regression and the optimal value is approximately 0.8. In the case of the lasso (for `w1a`), a value around 0.5 is optimal and low values (little scaling) yield worse prediction errors. Finally, for `rhee2006`, the lasso is again insensitive to normalization type. This is not the case for ridge, however, where a value around 0.2 is optimal and high values of δ yield worse prediction errors.

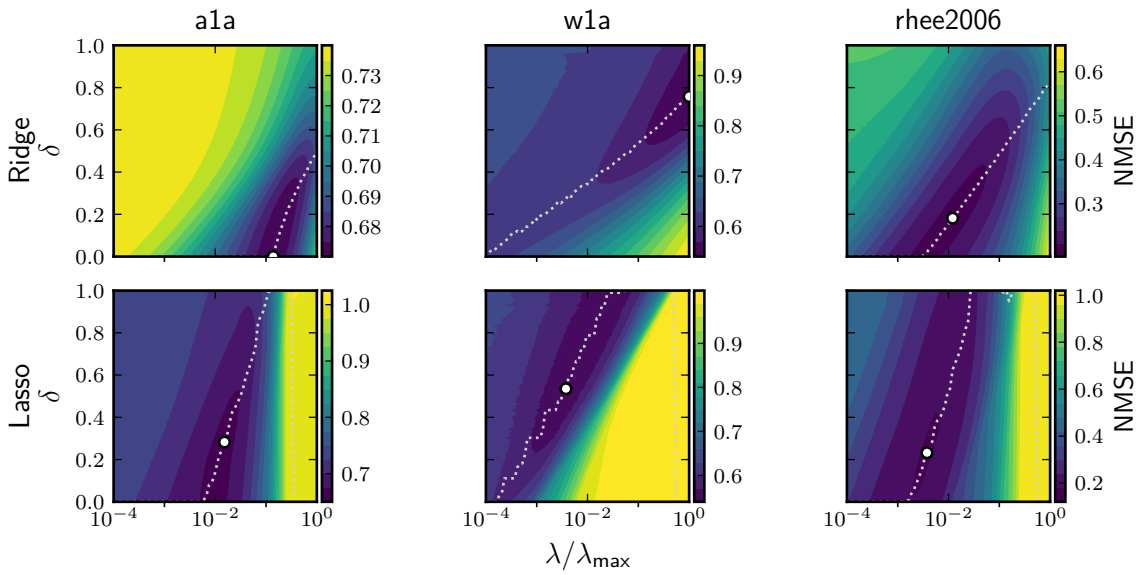


Figure 8: Contour plots of normalized mean-squared error (NMSE) for the hold-out validation set across a grid of δ and λ values for ridge regression and the lasso. The dotted path shows the smallest NMSE as a function of λ. The dot marks the combination with the smallest error.

We would like to point out that there is a dependency between λ and δ here that make it difficult to interpret the relationship between them and the error. This comes fro mthe fact that scaling with a smaller value (as in $\delta = 1$) increases the sizes of the vectors, which means that the level of penalization is relaxed, relative speaking.

In Figure 9, we have, in addition to NMSE on the validation set, also plotted the size of the support of the lasso (cardinality of the set of features that have corresponding nonzero coefficients). Here, however, we only show results for $\delta \in \{0, 1/2, 1\}$. It is clear that $\delta = 1/2$ works quite well for all of these three data sets, being able to attain a value close to the mininum for each of the three data sets. This is not the case for $\delta \in \{0, 1\}$, for which the best possible prediction error is considerably worse. This is particularly the case with $\delta = 0$ and the `w1a` data set. The dependency between λ and δ is also visible here by looking at the support size.

4.4 Mixed Data

In Section 3.3, we discovered that extra care needs to be taken when normalizing mixed data. In this experiment, we construct a quasi-normal feature with mean zero and standard deviation 1/2 and a binary feature with varying class balance q. We set the signal-to-noise ratio to 0.5 and generate our response vector $\boldsymbol{y}$ as before, with $n = 1000$. These features are constructed so that their effects are comparable under the notion of comparability that we introduce in Section 3.3, using $\kappa = 2$. In order to preserve the comparability for the baseline case $q_0 = 1/2$, we use the scaling introduced in Section 3.3, which leads to $s_j = 2 \times (1/4)^{1-\delta}(q - q^2)^{\delta}$. For the lasso, we set the level of penalization to $\lambda_{\max}/2$ and for ridge regression, we set the level of penalization to $2\lambda_{\max}$.[7]

[7]This makes the level of regularization comparable between the two cases.

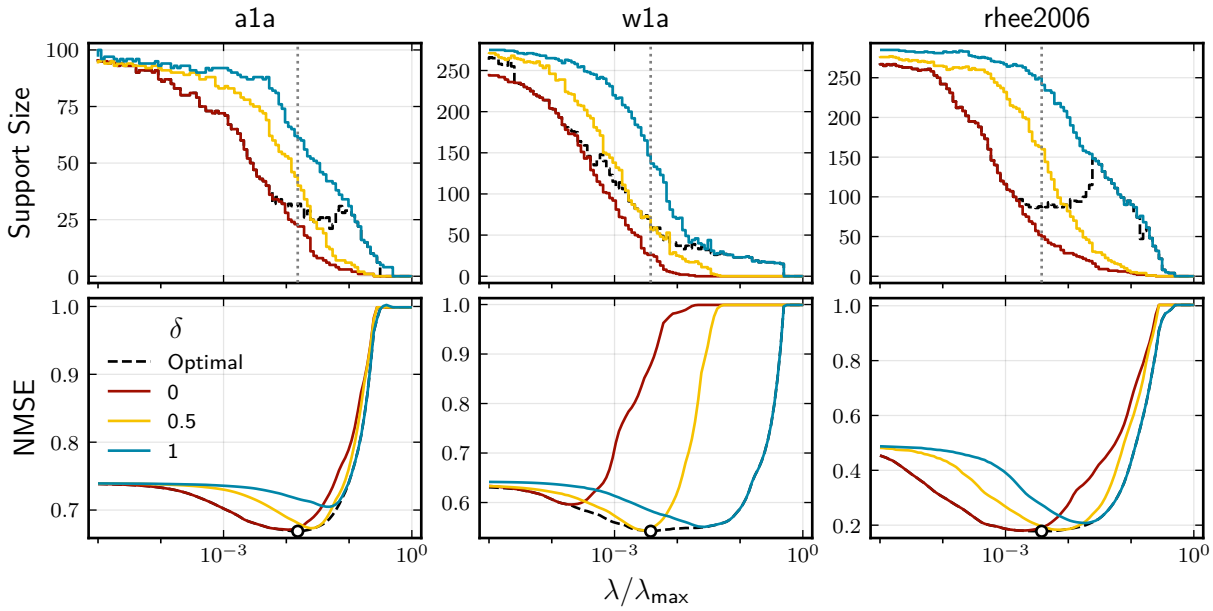


Figure 9: Support size and normalized mean-squared error (NMSE) for the validation set for the lasso fit to datasets `a1a`, `w1a`, and `rhee2006` across combinations of δ and λ. The optimal δ is marked with dashed black lines and the best combination of δ (among 0, 1/2, and 1) and λ is shown as a dot.

The results (Figure 10) reflect our theoretical results from Section 3. In the case of the lasso, we need $\delta = 1$ to avoid the effect of class imbalance, whereas for ridge we instead need $\delta = 1/2$ (standardization). As our theory suggests, this extra scaling mitigates this class-balance dependency at the cost of added variance.

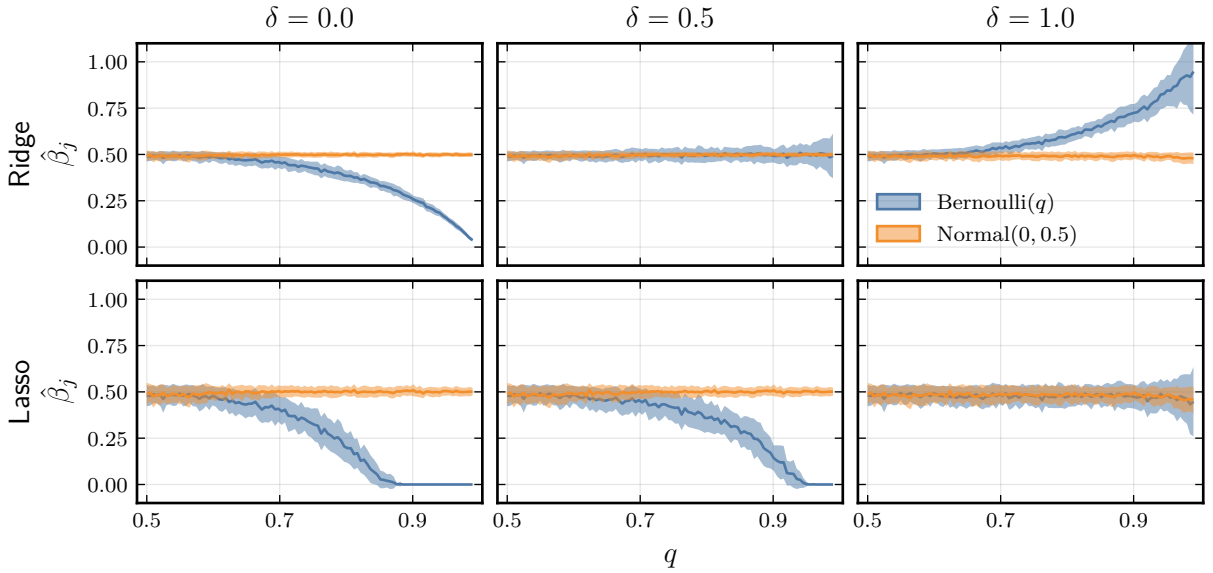


Figure 10: Lasso and ridge estimates for a two-dimensional problem where one feature is a binary feature with class balance q, Bernoulli(q), and the other is a quasi-normal feature with standard deviation 1/2, Normal$(0, 0.5)$. Here, we have $n = 1000$ observations. The signal-to-noise ratio is 0.5 In every case, we standardize the normal feature. The binary feature, meanwhile, is centered by its mean and scaled by $(q - q^2)^\delta$. The experiment is run for 50 iterations and we aggregate and report means and standard deviations of the estimates.

Note that we do not see the bias reduction that we observed in our theoretical results for high q values and $\delta \geq 1/2$ in Figure 10. This is related to the error term (signal-to-noise ratio) and level of q. Typically, we would need stronger class imbalance and larger error for the effect to show up in our experiments.

4.5 Interactions

In our final experiment, we study the effect of normalization and class balance on interactions when using the lasso. Our example consists of a two-feature problem with an added interaction term given by $x_{i3} = x_{i1}x_{i2}$. The first feature is binary with class balance $q = 0.9$ and the second quasi-normal with standard deviation 0.5. We set $n = 1000$ and specify $\lambda_1 = n/4$ as the level of regularization. Note that we normalize *after* the interaction term is added.

The results (Section 4.5) show, as before, that class balance (which, recall, is set to 0.9 here) has a dramatic effect on estimates of the binary feature when $\delta \in \{0, 1/2\}$. Somewhat surprisingly, however, the interaction term does not seem to be affected by the normalization type for any of the cases in which it is present.

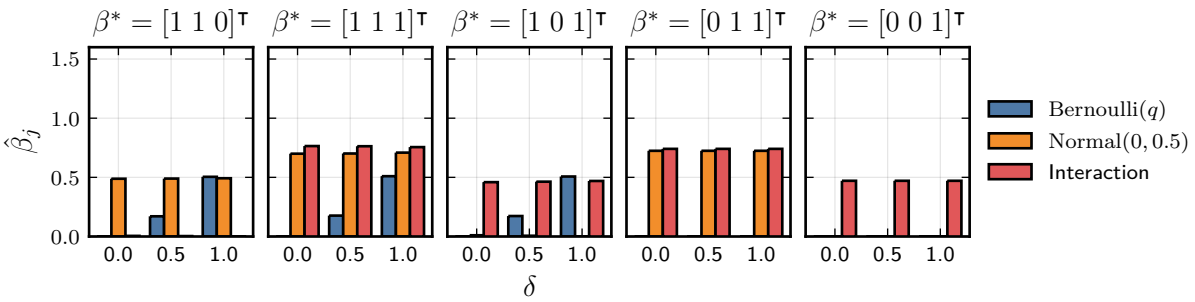


Figure 11: Lasso estimates for a three-feature problem where the third feature is an interaction term between the first two features. The first feature is binary (quasi-Bernoulli) with class balance $q = 0.9$ and the second is quasi-normal with standard deviation 0.5. The signal-to-noise ratio is 0.5. The experiment is run for 50 iterations and we aggregate and report means across all iterations.

Note that the interaction in this experiment naturally introduces correlation between the features and that this has an effect on the lasso estimates since we, for instance, can penalize the main effect whilst still retaining information about it in the interaction term.

5 Discussion

In this paper, we have studied the effects of normalization in ridge regression and the lasso for features that are binary—an issue that has so far been treated with disregarded in the literature. We have discovered the class imbalance of binary features—the proportion of ones and zeros in the features—have a pronounced effect on both lasso and ridge estimates, and that this effect depends on the type of normalization used. For the lasso, for instance, our results show that features with large class imbalances will be regularized heavily, and provided that λ is large enough might stand little chance of being selected, even if the true effect of the feature on the response is large.

We have, however, found that scaling binary features with standard deviation in the case of ridge regression and variance in the case of the lasso mitigates this effect, but that doing so comes at the price of increased variance. This effectively means that the choice of normalization constitutes a bias–variability trade-off with respect to imbalanced binary features.

To study these effects theoretically and in practice, we have introduced the scaling parameterization

$$s_j = (q - q^2)^\delta,$$

which, for instance, includes the cases $\delta = 0$ (no scaling), $\delta = 1/2$ (standard deviation scaling), and $\delta = 1$ (variance scaling). These, in turn, correspond to standard choices of normalization types for this kind of data.

The common variants max–abs and min–max normalization, for instance, in practice correspond to $\delta = 0$ in the case of binary data, whilst standardization corresponds to $\delta = 1/2$. As far as we know, scaling with $\delta = 1$ have previously not been considered in the literature nor to any extent that we are aware of in practice.

Our results demonstrate, however, that the choice of δ affects the lasso and ridge estimates heavily in many cases. This is particularly true with respect to selective inference, in which case $\delta = 0$ scaling will reduce the chances of finding the true model via the lasso in class-imbalanced settings (Section 4.1). But it will also bias the regression coefficients in both the lasso and ridge, which may also lead to suboptimal predictive performance (Section 4.2).

Both our theoretical results (Section 3.2) and experiments (Section 4.1) show that the optimal choice of δ may depend on the error in the data-generating process, which is typically unknown. As an alternative, we investigated choosing δ in a data-driven manner by optimizing over δ as if it were a hyperparameter (Section 4.3).

We have also studied the case of mixed data: designs that consist of both binary and normally distributed features. In this setting, our first finding is that there is an implicit relationship between the choice of normalization and the manner in which regularization affects binary viz-a-viz normally distributed features. For instance, the choice of max–abs normaliation carries a specific assumption about how the effect of a binary feature should be compared to that of a normally distributed feature. There is still much uncertainty about how to best handle the mixed data case and no ground truth given that a binary feature can mean any number of things—few of which are directly comparable to a continuous feature.

In our experimental results, we touch briefly on the case of interactions. In this case, it seems that the interaction term between a normal feature and a binary one is more-or-less unaffected by the class balance of the latter (Section 4.5). An interesting avenue for future research could be to study this in more detail, both theoretically and empirically. One particular problem with interactions is that the interaction term depends on the location, and not just the scale, of the normal feature (in this two-feature setting), which may call for conditional normalization strategies. Much remain to be explored in this area.

Finally, note that our theoretical results are limited by several assumptions: 1) a fixed feature matrix $\boldsymbol{X}$, 2) orthogonality between the features, and 3) normal and idependent errors. Future work could relax these assumptions to study the effects of normalization in more general settings. For instance, the assumption of orthogonality could be relaxed to allow for correlated features, which is often the case in practice. This would allow for a more general understanding of the effects of normalization in regularized regression models. We have also limited ourselves to the case of the lasso and ridge regression. Investigating to which extent, if any, the effects we observe generalize to other models as well would yield valuable insights. We have also focused on the case of binary and continuous features here, but we are convinced that the case of categorical features is also of interest and might raise additional challenges with respect to normalization.

References

Barry Becker and Ronny Kohavi. Adult, 1996.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, February 2017. ISSN 0036-1445. doi: 10.1137/141000671.

Małgorzata Bogdan, Ewout van den Berg, Weijie Su, and Emmanuel J. Candès. Statistical estimation and testing via the sorted L1 norm, October 2013.

Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès. SLOPE – adaptive variable selection via convex optimization. *The annals of applied statistics*, 9(3):1103–1140, September 2015. ISSN 1932-6157. doi: 10.1214/15-AOAS842.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, September 1995. ISSN 1573-0565. doi: 10.1007/BF00994018.

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3): 425–455, August 1994. ISSN 0006-3444. doi: 10.2307/2337118.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, December 2001. ISSN 0162-1459. doi: 10/fd7bfs.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, January 2010. doi: 10.18637/jss.v033.i01.

Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, 27 (15):2865–2873, July 2008. ISSN 02776715, 10970258. doi: 10.1002/sim.3107.

T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999. ISSN 0036-8075. doi: 10.1126/science.286.5439.531.

David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978. ISSN 0095-0696. doi: 10.1016/0095-0696(78)90006-2.

Trevor Hastie, Robert Tibshirani, and Ryan Tibshirani. Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. *Statistical Science*, 35(4):579–592, November 2020. ISSN 0883-4237. doi: 10.1214/19-STS733.

Ross King. Qualitative structure activity relationships.

Simon Kornblith. Lasso.jl, March 2024.

Haikady N. Nagaraja and Herbert A. David. *Order Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc, Hoboken, N.J, 3 edition, July 2003. ISBN 978-0-471-38926-2.

John C. Platt. Fast training of support vector machines using sequential minimal optimization. In Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.), *Advances in Kernel Methods: Support Vector Learning*, pp. 185–208. MIT Press, Boston, MA, USA, 1 edition, January 1998. ISBN 978-0-262-28319-9. doi: 10.7551/mitpress/1130.003.0016.

Soo-Yon Rhee, Jonathan Taylor, Gauhar Wadhera, Asa Ben-Hur, Douglas L. Brutlag, and Robert W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46):17355–17360, November 2006. doi: 10.1073/pnas.0607274103.

Fadil Santosa and William W. Symes. Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1307–1330, October 1986. ISSN 0196-5204. doi: 10.1137/0907087.

scikit-learn developers. 6.3. Preprocessing data. https://scikit-learn/stable/modules/preprocessing.html, February 2024.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996. ISSN 0035-9246. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, April 2010. ISSN 0090-5364. doi: 10/bp22zz.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101 (476):1418–1429, December 2006. ISSN 0162-1459. doi: 10.1198/016214506000000735.

Hui Zou and Trevor Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 1369-7412.

A Additional Theory

A.1 Why Maximum–Absolute and Min–Max Scaling are Unsuitable for Normally Distributed Data

In Theorem A.1, we show that the scaling factor in the max–abs method converges in distribution to a Gumbel distribution.

Theorem A.1. *Let $X_1, X_2, \ldots, X_n$ be a sample of normally distributed random variables, each with mean μ and standard deviation σ. Then*

$$\lim_{n\to\infty} \Pr\left(\max_{i\in[n]} |X_i| \leq x\right) = G(x),$$

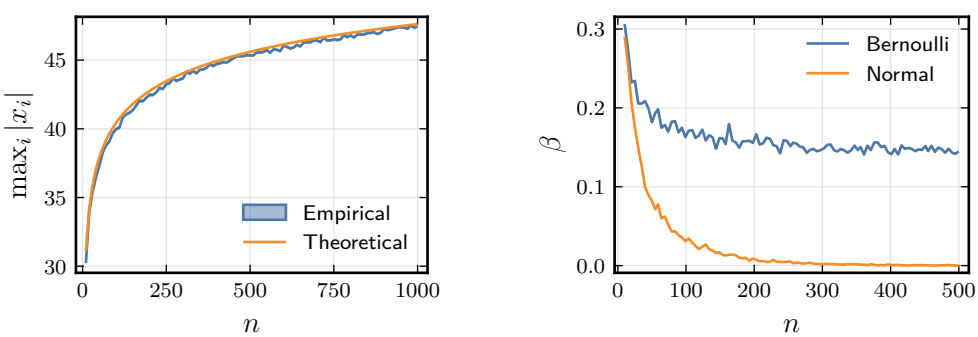
where G is the cumulative distribution function of a Gumbel distribution with parameters

$$b_n = F_Y^{-1}(1 - 1/n) \quad \textit{and} \quad a_n = \frac{1}{n f_Y(\mu_n)},$$

where f_Y and F_Y^{-1} are the probability distribution function and quantile function, respectively, of a folded normal distribution with mean μ and standard deviation σ.

The gist of Theorem A.1 is that the limiting distribution of $\max_{i\in[n]} |X_i|$ has expected value $b_n + \gamma a_n$, where γ is the Euler-Mascheroni constant. This indicates that the scaling factor strongly dependent on the sample size. In Figure 12a, we observe empirically that the limiting distribution agrees well with the empirical distribution in expected value even for small values of n.

In Figure 12b we show the effect of increasing the number of observations, n, in a two-feature lasso model with max-abs normalization applied to both features. The coefficient corresponding to the Normally distributed feature shrinks as the number of observation n increases. Since the expected value of the Gumbel distribution diverges with n, this means that there's always a large enough n to force the coefficient in a lasso problem to zero with high probability.



(a) Theoretical versus empirical distribution of the maximum absolute value of normally distributed random variables.

(b) Estimation of mixed features under maximum absolute value scaling

Figure 12: Effects of maximum absolute value scaling.

For min–max scaling, the situation is similar and we omit the details here. The main point is that the scaling factor is strongly dependent on the sample size, which makes it unsuitable for normally distributed data in several situations, such as on-line learning (where sample size changes over time) or model validation with uneven data splits.

B Proofs

B.1 Proof of Theorem A.1

If $X_i \sim \text{Normal}(\mu, \sigma)$, then $|X_i| \sim \text{FoldedNormal}(\mu, \sigma)$. By the Fisher–Tippett–Gnedenko theorem, we know that $(\max_i |X_i| - b_n)/a_n$ converges in distribution to either the Gumbel, Fréchet, or Weibull distribution,

given a proper choice of $a_n > 0$ and $b_n \in \mathbb{R}$. A sufficient condition for convergence to the Gumbel distribution for a absolutely continuous cumulative distribution function (Nagaraja & David, 2003, Theorem 10.5.2) is

$$\lim_{x\to\infty} \frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = 0.$$

We have

$$\begin{aligned}\frac{1-F_Y(x)}{f_Y(x)} &= \frac{1-\frac{1}{2}\operatorname{erf}\left(\frac{x-\mu}{\sqrt{2\sigma^2}}\right)-\frac{1}{2}\operatorname{erf}\left(\frac{x+\mu}{\sqrt{2\sigma^2}}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}+\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-(x+\mu)^2}{2\sigma^2}}} \\ &= \frac{2-\Phi\left(\frac{x-\mu}{\sigma}\right)-\Phi\left(\frac{x+\mu}{\sigma}\right)}{\frac{1}{\sigma}\left(\phi\left(\frac{x-\mu}{\sigma}\right)+\phi\left(\frac{x+\mu}{\sigma}\right)\right)} \\ &\to \frac{\sigma(1-\Phi(x))}{\phi(x)} \text{ as } n\to n,\end{aligned}$$

where ϕ and Φ are the probability distribution and cumulative density functions of the standard normal distribution respectively. Next, we follow Nagaraja & David (2003, example 10.5.3) and observe that

$$\frac{d}{dx}\frac{\sigma(1-\Phi(x))}{\phi(x)} = \frac{\sigma x(1-\Phi(x))}{\phi(x)} - \sigma \to 0 \text{ as } x\to\infty$$

since

$$\frac{1-\Phi(x)}{\phi(x)} \sim \frac{1}{x}.$$

In this case, we may take $b_n = F_Y^{-1}(1-1/n)$ and $a_n = \left(nf_Y(b_n)\right)^{-1}$.

B.2 Proof of Theorem 3.1

Since $s_j = (q-q^2)^\delta$, we have

$$\begin{aligned}
\mu &= \beta_j^* n(q-q^2)^{1-\delta} & \frac{\theta}{\sigma} &= -a\sqrt{q(1-q)} - b(q-q^2)^{\delta-1/2}, \\
\sigma &= \sigma_\varepsilon\sqrt{n}(q-q^2)^{1/2-\delta}, & \frac{\gamma}{\sigma} &= a\sqrt{q(1-q)} - b(q-q^2)^{\delta-1/2}, \\
d_j &= n(q-q^2)^{1-\delta} + \lambda_2(q-q^2)^\delta, & \frac{\theta}{d_j} &= -\beta_j^* - \frac{\lambda_1(q-q^2)^{\delta-1}}{n}, \\
\theta &= -\beta_j^* n(q-q^2)^{1-\delta} - \lambda_1, & \frac{\gamma}{d_j} &= \beta_j^* - \frac{\lambda_1(q-q^2)^{\delta-1}}{n}, \\
\gamma &= \beta_j^* n(q-q^2)^{1-\delta} - \lambda_1,
\end{aligned}$$

with

$$a = \frac{\beta_j^*\sqrt{n}}{\sigma_\varepsilon} \qquad \text{and} \qquad b = \frac{\lambda_1}{\sigma_\varepsilon\sqrt{n}}.$$

We are interested in

$$\lim_{q\to 1^+} \mathrm{E}\,\hat{\beta}_j = \lim_{q\to 1^+}\frac{1}{d}\left(-\theta\,\Phi\left(\frac{\theta}{\sigma}\right) - \sigma\,\phi\left(\frac{\theta}{\sigma}\right) + \gamma\,\Phi\left(\frac{\gamma}{\sigma}\right) + \sigma\,\phi\left(\frac{\gamma}{\sigma}\right)\right). \tag{12}$$

Before we proceed, note the following limits, which we will make repeated use of throughout the proof.

$$\lim_{q\to 1^+}\frac{\theta}{\sigma} = \lim_{q\to 1^+}\frac{\gamma}{\sigma} = \begin{cases} -\infty & \text{if } 0\le\delta<\frac{1}{2}, \\ -b & \text{if } \delta=\frac{1}{2}, \\ 0 & \text{if } \delta>\frac{1}{2}, \end{cases} \tag{13}$$

Starting with the terms involving Φ inside the limit in Equation (12), for now assuming that they are well-defined and that the limits of the remaining terms also exist seperately, we have

$$\begin{aligned}
&\lim_{q\to 1^+}\left(-\frac{\theta}{d}\,\Phi\left(\frac{\theta}{\sigma}\right)+\frac{\gamma}{d_j}\,\Phi\left(\frac{\gamma}{\sigma}\right)\right)\\
&=\lim_{q\to 1^+}\Bigg(\left(\frac{\beta_j^* n}{n+\lambda_2(q-q^2)^{2\delta-1}}+\frac{\lambda_1}{n(q-q^2)^{1-\delta}+\lambda_2(q-q^2)^{\delta}}\right)\Phi\left(\frac{\theta}{\sigma}\right)\\
&\qquad+\left(\frac{\beta_j^* n}{n+\lambda_2(q-q^2)^{2\delta-1}}-\frac{\lambda_1}{n(q-q^2)^{1-\delta}+\lambda_2(q-q^2)^{\delta}}\right)\Phi\left(\frac{\gamma}{\sigma}\right)\Bigg)\\
&=\lim_{q\to 1^+}\frac{\beta_j^* n}{n+\lambda_2(q-q^2)^{2\delta-1}}\left(\Phi\left(\frac{\theta}{\sigma}\right)+\Phi\left(\frac{\gamma}{\sigma}\right)\right)\\
&\quad+\lim_{q\to 1^+}\frac{\lambda_1}{n(q-q^2)^{1-\delta}+\lambda_2(q-q^2)^{\delta}}\left(\Phi\left(\frac{\theta}{\sigma}\right)-\Phi\left(\frac{\gamma}{\sigma}\right)\right).
\end{aligned}\tag{14}$$

Considering the first term in Equation (14), we see that

$$\lim_{q\to 1^+}\frac{\beta_j^* n}{n+\lambda_2(q-q^2)^{2\delta-1}}\left(\Phi\left(\frac{\theta}{\sigma}\right)+\Phi\left(\frac{\gamma}{\sigma}\right)\right)=\begin{cases}0 & \text{if } 0\le\delta<1/2,\\ \frac{2n\beta_j^*}{n+\lambda_2}\,\Phi(-b) & \text{if } \delta=1/2,\\ \beta_j^* & \text{if } \delta>1/2.\end{cases}$$

For the second term in Equation (14), we start by observing that if $\delta=1$, then $q(1-q)^{\delta-1}=1$, and if $\delta>1$, then $\lim_{q\to 1^+}(q-q^2)^{\delta-1}=0$. Moreover, the arguments of Φ approach 0 in the limit for $\delta\ge 1$, which means that the entire term vanishes in both cases ($\delta\ge 1$).

For $0\le\delta<1$, the limit is indeterminite of the form $\infty\times 0$. We define

$$f(q)=\Phi\left(\frac{\theta}{\sigma}\right)-\Phi\left(\frac{\gamma}{\sigma}\right)\qquad\text{and}\qquad g(q)=n(q-q^2)^{1-\delta}+\lambda_2(q-q^2)^{\delta},$$

such that we can express the limit as $\lim_{q\to 1^+} f(q)/g(q)$. The corresponding derivatives are

$$\begin{aligned}
f'(q)&=\left(-\frac{a}{2}(1-2q)(q-q^2)^{-1/2}-b(\delta-1/2)(1-2q)(q-q^2)^{\delta-3/2}\right)\phi\left(\frac{\theta}{\sigma}\right)\\
&\quad-\left(-\frac{a}{2}(1-2q)(q-q^2)^{-1/2}-b(\delta-1/2)(1-2q)(q-q^2)^{\delta-3/2}\right)\phi\left(\frac{\gamma}{\sigma}\right),\\
g'(q)&=n(1-\delta)(1-2q)(q-q^2)^{-\delta}+\lambda_2\delta(1-2q)(q-q^2)^{\delta-1}
\end{aligned}$$

Note that $f(q)$ and $g(q)$ are both differentiable and $g'(q)\neq 0$ everywhere in the interval $(1/2,1)$. Now note that we have

$$\begin{aligned}
\frac{f'(q)}{g'(q)}&=\frac{1}{n(1-\delta)(q-q^2)^{1/2-\delta}+\lambda_2\delta(1-2q)(q-q^2)^{\delta-1/2}}\\
&\quad\times\left(\left(-\frac{a}{2}-b(\delta-1/2)(q-q^2)^{\delta-1}\right)\phi\left(\frac{\theta}{\sigma}\right)-\left(\frac{a}{2}-b(\delta-1/2)(q-q^2)^{\delta-1}\right)\phi\left(\frac{\gamma}{\sigma}\right)\right).
\end{aligned}\tag{15}$$

For $0\le\delta<1/2$, $\lim_{q\to 1^+} f'(q)/g'(q)=0$ since the exponential terms of ϕ in Equation (15) dominate in the limit.

For $\delta=1/2$, we have

$$\lim_{q\to 1^+}\frac{f'(q)}{g'(q)}=-\frac{a}{n+\lambda_2}\lim_{q\to 1^+}\left(\phi\left(\frac{\theta}{\sigma}\right)+\phi\left(\frac{\gamma}{\sigma}\right)\right)=-\frac{a}{n+\lambda_2}\,\phi(-b)$$

so that we can use L'Hôpital's rule to show that the second term in Equation (14) becomes

$$-\frac{2\beta_j^*\lambda_1\sqrt{n}}{\sigma_\varepsilon(n+\lambda_2)}\,\phi\left(\frac{-\lambda_1}{\sigma_\varepsilon\sqrt{n}}\right).\tag{16}$$

For $\delta > 1/2$, we have

$$\begin{aligned}
\lim_{q\to 1^+} \frac{f'(q)}{g'(q)} &= \lim_{q\to 1^+} \frac{-\frac{a}{2}\left(\phi\left(\frac{\theta}{\sigma}\right) + \phi\left(\frac{\gamma}{\sigma}\right)\right)}{n(1-\delta)(q-q^2)^{1/2-\delta} + \lambda_2\delta(1-2q)(q-q^2)^{\delta-1/2}} \\
&\quad + \lim_{q\to 1^+} \frac{b(\delta-1/2)\left(\phi\left(\frac{\gamma}{\sigma}\right) - \phi\left(\frac{\theta}{\sigma}\right)\right)}{n(1-\delta)(q-q^2)^{3/2-2\delta} + \lambda_2\delta(1-2q)(q-q^2)^{1/2}} \\
&= 0 + \lim_{q\to 1^+} \frac{b(\delta-1/2)e^{-\frac{1}{2}\left(a^2(q-q^2)+b^2(q-q^2)^{2\delta-1}\right)}\left(e^{-ab(q-q^2)^\delta} - e^{ab(q-q^2)^\delta}\right)}{\sqrt{2\pi}\left(n(1-\delta)(q-q^2)^{3/2-2\delta} + \lambda_2\delta(1-2q)(q-q^2)^{1/2}\right)} \\
&= 0
\end{aligned}$$

since the exponential term in the numerator dominates.

Now we proceed to consider the terms involving ϕ in Equation (12). We have

$$\lim_{q\to 1^+} \frac{\sigma}{d}\left(\phi\left(\frac{\gamma}{\sigma}\right) - \phi\left(\frac{\theta}{\sigma}\right)\right) = \sigma_\varepsilon\sqrt{n} \lim_{q\to 1^+} \frac{\phi\left(\frac{\gamma}{\sigma}\right) - \phi\left(\frac{\theta}{\sigma}\right)}{n(q-q^2)^{1/2} + \lambda_2(q-q^2)^{2\delta-1/2}} \tag{17}$$

For $0 \le \delta < 1/2$, we observe that the exponential terms in ϕ dominate in the limit, and so we can distribute the limit and consider the limits of the respective terms individually, which both vanish.

For $\delta \ge 1/2$, the limit in Equation (17) has an indeterminate form of the type $\infty \times 0$. Define

$$u(q) = \phi\left(\frac{\gamma}{\sigma}\right) - \phi\left(\frac{\theta}{\sigma}\right) \qquad \text{and} \qquad v(q) = n(q-q^2)^{1/2} + \lambda_2(q-q^2)^{2\delta-1/2}$$

which are both differentiable in the interval $(1/2, 1)$ and $v'(q) \neq 0$ everywhere in this interval. The derivatives are

$$\begin{aligned}
u'(q) &= -\phi\left(\frac{\gamma}{\sigma}\right)\frac{\gamma}{\sigma}\left(\frac{1}{2}\left(a(1-2q)(q-q^2)^{-1/2}\right) - b(\delta-1/2)(1-2q)(q-q^2)^{\delta-3/2}\right) \\
&\quad + \phi\left(\frac{\theta}{\sigma}\right)\frac{\theta}{\sigma}\left(-\frac{1}{2}\left(a(1-2q)(q-q^2)^{-1/2}\right) - b(\delta-1/2)(1-2q)(q-q^2)^{\delta-3/2}\right), \\
v'(q) &= \frac{n}{2}(1-2q)(q-q^2)^{-1/2} + \lambda_2(2\delta-1/2)(1-2q)(q-q^2)^{2\delta-3/2}.
\end{aligned}$$

And so

$$\begin{aligned}
\frac{u'(q)}{v'(q)} = \frac{1}{n + \lambda_2(4\delta-1)(q-q^2)^{2\delta-1}}\Bigg(&-\left(a - b(2\delta-1)(q-q^2)^{\delta-1}\right)\phi\left(\frac{\gamma}{\sigma}\right)\frac{\gamma}{\sigma} \\
&- \left(a + b(2\delta-1)(q-q^2)^{\delta-1}\right)\phi\left(\frac{\theta}{\sigma}\right)\frac{\theta}{\sigma}\Bigg).
\end{aligned} \tag{18}$$

Taking the limit, rearranging, and assuming that the limits of the separate terms exist, we obtain

$$\begin{aligned}
\lim_{q\to 1^+} \frac{u'(q)}{v'(q)} &= -a \lim_{q\to 1^+} \frac{1}{n + \lambda_2(4\delta-1)(q-q^2)^{2\delta-1}}\left(\phi\left(\frac{\gamma}{\sigma}\right)\frac{\gamma}{\sigma} + \phi\left(\frac{\theta}{\sigma}\right)\frac{\theta}{\sigma}\right) \\
&\quad + b(2\delta-1) \lim_{q\to 1^+} \frac{1}{n + \lambda_2(4\delta-1)(q-q^2)^{2\delta-1}}\Bigg(\phi\left(\frac{\gamma}{\sigma}\right)\left(a(q-q^2)^{\delta-1/2} - b(q-q^2)^{2\delta-3/2}\right) \\
&\qquad - \phi\left(\frac{\theta}{\sigma}\right)\left(-a(q-q^2)^{\delta-1/2} - b(q-q^2)^{2\delta-3/2}\right)\Bigg).
\end{aligned} \tag{19}$$

For $\delta = 1/2$, we have

$$\lim_{q\to 1^+} \frac{u'(q)}{v'(q)} = -\frac{a}{n+\lambda_2}\left(-b\,\phi(-b) - b\,\phi(-b)\right) + 0 = 2ab\,\phi(-b) = \frac{2\beta_j^*\lambda_1}{\sigma_\varepsilon^2(n+\lambda_2)}\,\phi\left(\frac{-\lambda_1}{\sigma_\varepsilon\sqrt{n}}\right).$$

Using L'Hôpital's rule, the second term in Equation (17) must consequently be

$$\frac{2\beta_j^* \lambda_1 \sqrt{n}}{\sigma_\varepsilon (n + \lambda_2)} \, \phi\left(\frac{-\lambda_1}{\sigma_\varepsilon \sqrt{n}}\right).$$

which cancels with Equation (16).

For $\delta > 1/2$, we first observe that the first term in Equation (19) tends to zero due to Equation (13) and the properties of the standard normal distribution. For the second term, we note that this is essentially of the same form as Equation (15) and that the limit is therefore 0 here.

B.3 Proof of Theorem 3.2

The variance of the elastic net estimator is given by

$$\begin{aligned}\operatorname{Var} \hat{\beta}_j = \frac{1}{d^2} \Bigg(\frac{\sigma^2}{2} \left(2 + \operatorname{erf}\left(\frac{\theta}{\sigma\sqrt{2}}\right) - \frac{\theta}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right) + \operatorname{erf}\left(\frac{\gamma}{\sigma\sqrt{2}}\right) - \frac{\gamma}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\gamma^2}{2\gamma^2}\right) \right) \\ + 2\theta\sigma\, \phi\left(\frac{\theta}{\sigma}\right) + \theta^2\, \Phi\left(\frac{\theta}{\sigma}\right) + 2\gamma\sigma\, \phi\left(\frac{\gamma}{\sigma}\right) + \gamma^2\, \Phi\left(\frac{\gamma}{\sigma}\right) \Bigg) - \left(\frac{1}{d} \operatorname{E} \hat{\beta}_j\right)^2. \quad (20)\end{aligned}$$

We start by noting the following identities:

$$\begin{aligned}
\theta^2 &= \left(\beta_j^* n\right)^2 (q - q^2)^{2-2\delta} + \lambda_1^2 + 2\lambda_1 \beta_j^* n (q - q^2)^{1-\delta}, \\
d^2 &= n^2 (q - q^2)^{2-2\delta} + 2n\lambda_2 (q - q^2) + \lambda_2^2 (q - q^2)^{2\delta}, \\
\theta\sigma &= -\sigma_\varepsilon \left(\beta_j^* n^{3/2} (q - q^2)^{3/2-2\delta} + \sqrt{n}\lambda_1 (q - q^2)^{1/2-\delta}\right), \\
\frac{\theta^2}{\sigma^2} &= a^2 (q - q^2) + b^2 (q - q^2)^{2\delta-1} + 2ab (q - q^2)^{\delta}, \\
\frac{\sigma}{d} &= \frac{\sigma_\varepsilon \sqrt{n}}{n (q - q^2)^{\frac{1}{2}} + \lambda_2 (q - q^2)^{2\delta - 1/2}}.
\end{aligned}$$

Expansions involving γ, instead of θ, have identical expansions up to sign changes of the individual terms. Also recall the definitions provided in the proof of Theorem 3.1.

Starting with the case when $0 \leq \delta < 1/2$, we write the limit of Equation (20) as

$$\begin{aligned}
&\lim_{q \to} \operatorname{Var} \hat{\beta}_j \\
&= \sigma_\varepsilon^2 n \lim_{q \to 1^+} \frac{1}{\left(n(q - q^2)^{1/2} + \lambda_2 (q - q^2)^{2\delta - 1/2}\right)^2} \left(1 + \operatorname{erf}\left(\frac{\theta}{\sigma\sqrt{2}}\right) - \frac{\theta}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right)\right) \\
&\quad + \sigma_\varepsilon^2 n \lim_{q \to 1^+} \frac{1}{\left(n(q - q^2)^{1/2} + \lambda_2 (q - q^2)^{2\delta - 1/2}\right)^2} \left(1 + \operatorname{erf}\left(\frac{\gamma}{\sigma\sqrt{2}}\right) - \frac{\gamma}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\gamma^2}{2\sigma^2}\right)\right) \\
&\quad + \lim_{q \to 1^+} \frac{2\theta\sigma}{d^2}\, \phi\left(\frac{\theta}{\sigma}\right) + \lim_{q \to 1^+} \frac{\theta^2}{d^2}\, \Phi\left(\frac{\theta}{\sigma}\right) + \lim_{q \to 1^+} \frac{2\gamma}{d^2} \sigma\, \phi\left(\frac{\gamma}{\sigma}\right) + \lim_{q \to 1^+} \frac{\gamma^2}{d^2}\, \Phi\left(\frac{\gamma}{\sigma}\right) \\
&\quad - \left(\lim_{q \to 1^+} \frac{1}{d} \operatorname{E} \hat{\beta}_j\right)^2,
\end{aligned}$$

assuming, for now, that all limits exist. Next, let

$$\begin{aligned}
f_1(q) &= 1 + \operatorname{erf}\left(\frac{\theta}{\sigma\sqrt{2}}\right) - \frac{\theta}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\theta^2}{2\sigma^2}\right), \\
f_2(q) &= 1 + \operatorname{erf}\left(\frac{\gamma}{\sigma\sqrt{2}}\right) - \frac{\gamma}{\sigma}\sqrt{\frac{2}{\pi}} \exp\left(-\frac{\gamma^2}{2\sigma^2}\right), \\
g(q) &= \left(n^2 (q - q^2) + 2n\lambda_2 (q - q^2)^{2\delta} + \lambda_2^2 (q - q^2)^{4\delta - 1}\right)^2.
\end{aligned}$$

And

$$
\begin{aligned}
f_1'(q) &= \frac{\theta^2}{\sigma^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\theta^2}{2\sigma^2}\right), \\
f_2'(q) &= \frac{\gamma^2}{\sigma^2}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2\sigma^2}\right), \\
g'(q) &= (1-2q)\left((q-q^2)^{-1} + 4n\delta\lambda_2(q-q^2)^{2\delta-1} + \lambda_2^2(4\delta-1)(q-q^2)^{4\delta-2}\right).
\end{aligned}
$$

f_1, f_1 and g are differentiable in $(1/2, 1)$ and $g'(q) \neq 0$ everywhere in this interval. f_1/g and f_2/g are indeterminate of the form $0/0$. And we see that

$$
\lim_{q\to 1^+} \frac{f_1'(q)}{g'(q)} = \lim_{q\to 1^+} \frac{f_2'(q)}{g'(q)} = 0
$$

due to the dominance of the exponential terms as θ/σ and γ/σ both tend to $-\infty$. Thus f_1/g and f_2/g also tend to 0 by L'Hôpital's rule.

Similar reasoning shows that

$$
\lim_{q\to 1^+} \frac{2\theta\sigma}{d^2}\,\phi\left(\frac{\theta}{\sigma}\right) = \lim_{q\to 1^+} \frac{\theta^2}{d^2}\,\Phi\left(\frac{\theta}{\sigma}\right) = 0.
$$

The same result applies to the respective terms involving γ.

And since we in Theorem 3.1 showed that $\lim_{q\to 1^+} \frac{1}{d}\,\mathrm{E}\,\hat{\beta}_j = 0$, the limit of Equation (20) must be 0.

For $\delta = 1/2$, we start by establishing that

$$
\lim_{q\to 1^+} \int_{-\infty}^{-\lambda} (z+\lambda)^2 f_Z(z)\,\mathrm{d}z = \lim_{q\to 1^+} \left(\sigma^2 \int_{-\infty}^{\frac{\theta}{\sigma}} y^2\,\phi(y)\,\mathrm{d}y + 2\theta\sigma \int_{-\infty}^{\frac{\theta}{\sigma}} y\,\phi(y)\,\mathrm{d}y + \theta^2 \int_{-\infty}^{\frac{\theta}{\sigma}} \phi(y)\,\mathrm{d}y\right)
$$

is a positive constant since $\theta/\sigma \to -b$, $\sigma = \sigma_\varepsilon\sqrt{n}$, $\theta \to -\lambda$, and $\theta\sigma \to -\sigma_\varepsilon\sqrt{n}\lambda$. An identical argument can be made in the case of

$$
\lim_{q\to 1^+} \int_{\lambda}^{\infty} (z-\lambda)^2 f_Z(z)\,\mathrm{d}z.
$$

We then have

$$
\lim_{q\to 1^+} \frac{1}{d^2} \int_{-\infty}^{-\lambda} (z+\lambda)^2 f_Z(z)\,\mathrm{d}z = \frac{C^+}{\lim_{q\to 1^+} d^2} = \frac{C^+}{0} = \infty,
$$

where C^+ is some positive constant. And because $\lim_{q\to 1^+} \frac{1}{d}\,\mathrm{E}\,\hat{\beta}_j = \beta_j^*$ (Theorem 3.1), the limit of Equation (20) must be ∞.

Finally, for the case when $\delta > 1/2$, we have

$$
\begin{aligned}
&\lim_{q\to 1^+} \frac{1}{d^2}\left(\sigma^2 \int_{-\infty}^{\frac{\theta}{\sigma}} y^2\,\phi(y)\,\mathrm{d}y + 2\theta\sigma \int_{-\infty}^{\frac{\theta}{\sigma}} y\,\phi(y)\,\mathrm{d}y + \theta^2 \int_{-\infty}^{\frac{\theta}{\sigma}} \phi(y)\,\mathrm{d}y\right) \\
&\qquad = \lim_{q\to 1^+} \left(\frac{n\sigma^2}{\left(n(q-q^2)^{1/2} + \lambda_2(q-q^2)^{2\delta-1/2}\right)^2} \int_{-\infty}^{\frac{\theta}{\sigma}} y^2\,\phi(y)\,\mathrm{d}y \right. \\
&\qquad\qquad - \frac{2\sigma_\varepsilon\sqrt{n}\left(\beta_j^* n(q-q^2)^{1-\delta} - \lambda_1\right)}{\left(n(q-q^2)^{3/4-\delta/2} + \lambda_2(q-q^2)^{3\delta/2-1/4}\right)^2} \int_{-\infty}^{\frac{\theta}{\sigma}} y\,\phi(y)\,\mathrm{d}y \\
&\qquad\qquad \left. + \left(\frac{-\beta_j^* n(q-q^2)^{1-\delta} - \lambda_1}{n(q-q^2)^{1-\delta} + \lambda_2(q-q^2)^{\delta}}\right)^2 \int_{-\infty}^{\frac{\theta}{\sigma}} \phi(y)\,\mathrm{d}y\right).
\end{aligned}
$$

Inspection of the exponents involving the factor $(q-q^2)$ shows that the first term inside the limit will dominate. And since the upper limit of the integrals, $\theta/\sigma \to 0$ as $q \to 1^+$, the limit must be ∞.

B.4 Proof of Corollary 3.2.1

We have

$$\lim_{q\to 1^+} \operatorname{Var}\hat{\beta}_j = \lim_{q\to 1^+} \frac{\sigma^2}{d_j^2}\left(\frac{\sigma_\varepsilon\sqrt{n}(q-q^2)^{1/2-\delta}}{n(q-q^2)^{1-\delta}+\lambda_2(q-q^2)^{\delta}}\right)^2 = \frac{\sigma_\varepsilon^2 n}{\lambda_2^2}\lim_{q\to 1^+}(q-q^2)^{1-4\delta},$$

from which the result follows directly.