Formatting Instructions for TMLR Journal Submissions

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

The abstract paragraph should be indented 1/2 inch on both left and right-hand margins. Use 10 point type, with a vertical spacing of 11 points. The word **Abstract** must be centered, in bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

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 $h(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y})$ be the original objective function—which when minimized improves the model's fit to the data $(\boldsymbol{X}, \boldsymbol{y})$—then

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

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

Some of the most common penalties are the ℓ_1 and ℓ_2 penalties, that is $g(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$ or $g(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2/2$[1], which, if h is the standard ordinary least-squares objective, represent lasso and ridge (Tikhonov) regression respectively. Other common penalities include SLOPE, MCP, hinge loss (used in support vector machines) and SCAD. 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 and locations of the features in $\boldsymbol{X}$. A common remedy is to *normalize* the features before fitting the model by translating and dividing each column 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 and location 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], 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.

[1] Division by two in this case is used only for convenience.
[2] The adaptive lasso typically uses ordinary least square estimates of the regression coefficients to scale the features with.

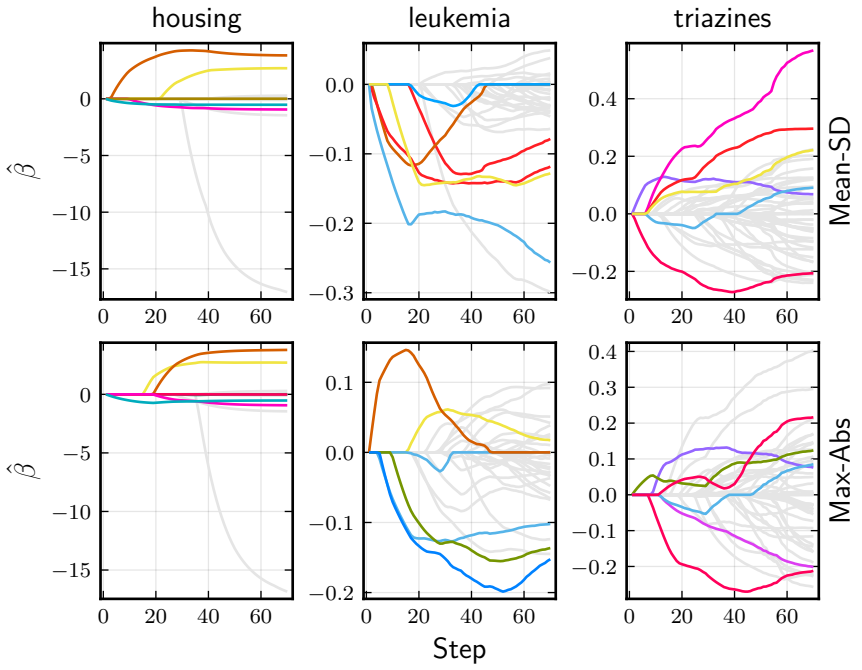

Figure 1: A display over the first predictors selected by the lasso for each type of normalization. Each panel shows the union of the first five predictors picked by either type of normalization.

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 three real data sets and three different types of normalization. Each panel shows the union of the first five predictors picked by either type of normalization. As we can see, 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, we will argue that normalization should rather we considered as an integral part of the model. And that it for instance is unreasonable to base the choice of normalization on the type of data storage, which implicitly encodes the belief that a data set stored as a sparse matrix is somehow fundamentally different from a data set stored as a dense matrix.

2 Preliminaries

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

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

with $[n] = \{1, 2 \dots, n\}$ and 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 $\boldsymbol{S}$ be the *scaling matrix*, which is a $p \times p$ diagonal matrix with entries $s_1, s_2, \dots, s_p$. Let $\boldsymbol{C}$ be the *centering matrix*, which is an $n \times p$ matrix with each row equal to $[c_1, c_2, c_n]^\intercal$. Then the *normalized design matrix* $\tilde{\boldsymbol{X}}$ is defined as $\tilde{\boldsymbol{X}} = (\boldsymbol{X} - \boldsymbol{C})\boldsymbol{S}^{-1}$.

Some authors refer to this procedure as *standardization* or *scaling*, but here we define scaling only as multiplication with the inverse of the scaling matrix and standardization as the case when scaling and centering with standard deviations and means respectively. Also note that normalization is sometimes defined as the process of scaling the samples (rather than the features).

2.1 Rescaling Regression Coefficients

Normalization changes the optimization problem as well as 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 estimates of these, we transform the coefficients from the normalized poblem, which we denote by $\hat{\beta}_j^{(n)}$, back using the following formulae.

$$\hat{\beta}_j = \frac{\hat{\beta}_j^{(n)}}{s_j}.$$

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

2.2 Types of Normalization

There are many different strategies for normalizing the design matrix. In this paper, we will focus on the ones outlined in Table 1. In the following sections, we will discuss some basic properties of these normalization strategies that will be useful in subsequent sections of the paper.

Table 1: Common ways to normalize a matrix of features

Normalization	Centering (c_{1j})	Scaling (s_j)
Standardization	$\frac{1}{n}\sum_{i=1}^n x_{ij}$	$\sqrt{\frac{1}{n}\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$
Min–Max	$\min_i(x_{ij})$	$\max_i(x_{ij}) - \min_i(x_{ij})$
Unit Vector (L2)	0	$\sqrt{\sum_{i=1}^n x_{ij}^2}$
Max Abs	0	$\max_i(\lvert x_{ij}\rvert)$
Adaptive Lasso	0	β_j^{OLS}

2.2.1 Standardization

Standardization is perhaps the most common type of normalization, at least in the field of statistics. It is also 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, this can lead to a significant increase in memory usage and computational time.

2.2.2 Maximum Absolute Value Scaling

A common alternative to standardization is to scale the features by their maximum absolute value. This is sometimes called *max–abs* scaling. This type of scaling typically has no impact on binary data (since the maximum absolute value is usually 1), and therefore retains sparsity. For other types of data, it scales the features to take values in the range $[-1, 1]$. This type of scaling is naturally sensitive to outliers, since they will single-handedly determine the scaling factor.

For many types of continuous data, such as normally distributed data, the sample maximum, and therefore the level of scaling in the max–abs method, depends on the sample size Theorem 2.1. This, in addition to the sensitivity to outliers, makes maximum absolute value scaling unsuitable for such data. As a result, we will only discuss it in the context of binary features.

Theorem 2.1. *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.

Proof. 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, 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, 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 $a_n = \left(n f_Y(b_n)\right)^{-1}$. □

As a result of Theorem 2.1, 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. In Figure 2, we verify that the limiting distribution agrees well with the empirical distribution in expected value even for small values of n.

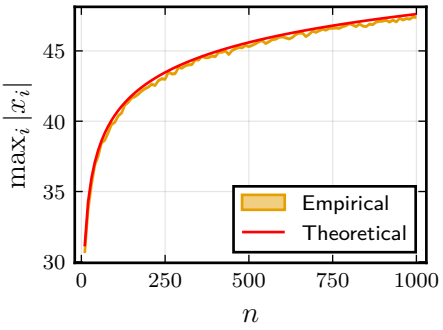

Figure 2: Theoretical versus empirical distribution of the maximum absolute value of normally distributed random variables.

In Figure 3 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 make the coefficient zero with high probability.

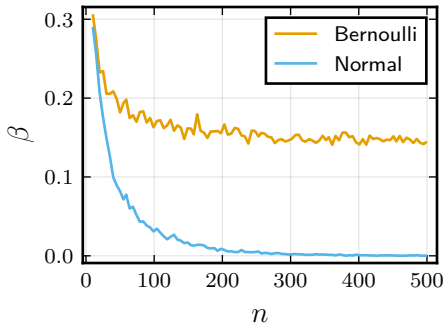

Figure 3: Effects of maximum absolute value scaling.

2.2.3 Min-Max Normalization

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.

2.3 The Elastic Net

From now on, we will direct our focus on the elastic net, which is a combination of the ℓ_1 and ℓ_2 penalties, that is,

$$\frac{1}{2}\|\boldsymbol{y} - \beta_0 - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2. \tag{1}$$

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 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.

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

Here, g is a subgradient of the ℓ_1 norm, which has elements g_i such that

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

2.4 Orthogonal Features

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

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

The inclusion of an intercept, β_0, ensures that the location 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 $\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{1} = 0$. A solution to the system of equations is then given by the following set of equations (Donoho & Johnstone):

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y}\right)}{s_j\left(\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j + \lambda_2\right)}, \qquad \hat{\beta}_0 = \frac{\boldsymbol{y}^\intercal \boldsymbol{1}}{n}, \tag{3}$$

where S is the soft-thresholding operator, defined as

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

3 Bias-Variance Tradeoffs in Data with Binary Features

Now, assume that $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are 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. We are interested in the expected value of Equation (3). Let $Z = \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{y} = \tilde{\boldsymbol{x}}_j^\intercal(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$ and $d_j = s_j\left(\tilde{\boldsymbol{x}}_j^\intercal \tilde{\boldsymbol{x}}_j + \lambda_2\right)$ so that $\hat{\beta}_j = \mathrm{S}_{\lambda_1}(Z)/d_j$. We start by focusing on the numerator, since the denominator, d_j, is fixed. First observe that

$$\mathbb{E}Z = \mu = \mathbb{E}\left(\tilde{\boldsymbol{x}}_j^\intercal(\boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon})\right) = \tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{x}_j \beta_j.$$

And for the variance, we have

$$\operatorname{Var} Z = \sigma^2 = \operatorname{Var}\left(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{\varepsilon}\right) = \sigma_\varepsilon^2 \|\tilde{\boldsymbol{x}}_j\|_2^2.$$

The expected value of the soft-thresholding estimator is

$$\begin{aligned} \mathbb{E}\,\mathrm{S}_\lambda(Z) &= \int_{-\infty}^{\infty} \mathrm{S}_\lambda(z) f_Z(z)\,\mathrm{d}z \\ &= \int_{-\infty}^{\infty} \mathrm{I}_{|z|>\lambda}(z - \operatorname{sign}(z)\lambda) f_Z(z)\,\mathrm{d}z \\ &= \int_{-\infty}^{-\lambda} (z + \lambda) f_Z(z)\,\mathrm{d}z + \int_{\lambda}^{\infty} (z - \lambda) f_Z(z)\,\mathrm{d}z. \end{aligned} \tag{4}$$

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

$$\mathbb{E}\hat{\beta}_j - \beta_j^* = \frac{1}{d_j}\mathbb{E}\,\mathrm{S}_\lambda(Z) - \beta_j^*. \tag{5}$$

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

$$\operatorname{Var} S_\lambda(Z) = \int_{-\infty}^{-\lambda} (z+\lambda)^2 f_Z(z)\,\mathrm{d}z + \int_{\lambda}^{\infty} (z-\lambda)^2 f_Z(z)\,\mathrm{d}z - \left(\mathbb{E}\,\mathrm{S}_\lambda(Z)\right)^2 \tag{6}$$

and that the variance of the elastic net estimator is therefore

$$\operatorname{Var}\hat{\beta}_j = \frac{1}{d_j^2}\operatorname{Var}\mathrm{S}_\lambda(Z). \tag{7}$$

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

$$Z \sim \operatorname{Normal}\left(\tilde{\boldsymbol{x}}_j^\intercal \boldsymbol{x}_j \beta_j, \sigma_\varepsilon^2 \|\tilde{\boldsymbol{x}}_j\|_2^2\right).$$

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

$$\begin{aligned}
\mathbb{E}\,\mathrm{S}_{\lambda_1}(Z) &= \int_{-\infty}^{\frac{\theta}{\sigma}} (\sigma u - \theta)\,\phi(u)\,\mathrm{d}u + \int_{-\frac{\gamma}{\sigma}}^{\infty} (\sigma u + \gamma)\,\phi(u)\,\mathrm{d}u \\
&= -\theta\,\Phi\left(\frac{\theta}{\sigma}\right) - \sigma\,\phi\left(\frac{\theta}{\sigma}\right) + \gamma\,\Phi\left(\frac{\gamma}{\sigma}\right) + \sigma\,\phi\left(\frac{\gamma}{\sigma}\right)
\end{aligned} \tag{8}$$

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 (6), we have

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

Similar computations for the second term on the left-hand side of Equation (6) yield

$$\begin{aligned}
&\int_{\lambda_1}^{\infty} (z-\lambda_1)^2 f_Z(z)\,\mathrm{d}z \\
&\qquad = \frac{\sigma^2}{2}\left(\operatorname{erf}\left(\frac{\gamma}{\sigma\sqrt{2}}\right) - \frac{\gamma}{\sigma}\sqrt{\frac{2}{\pi}}\exp\left(-\frac{\gamma^2}{2\sigma^2}\right) + 1\right) + 2\gamma\sigma\,\phi\left(\frac{\gamma}{\sigma}\right) + \gamma^2\,\Phi\left(\frac{\gamma}{\sigma}\right).
\end{aligned} \tag{10}$$

Plugging Equations (8) to (10) into Equation (7) yields the variance of the estimator. Consequently, we can also compute the mean-squared error via the bias-variance decomposition

$$\operatorname{MSE}(\hat{\beta}_j, \beta_j^*) = \operatorname{Var}\hat{\beta}_j + \left(\mathbb{E}\hat{\beta}_j - \beta_j^*\right)^2. \tag{11}$$

Our results have so far covered the general case where we have made no assumptions on $\boldsymbol{X}$, except for being non-random. But our main focus in this paper is the case when $\boldsymbol{x_j}$ is a binary feature with class balance q, that is, $x_{ij} \in \{0, 1\}$ for all i and $\sum_{i=1}^n x_{ij} = nq$. In this case, we observe that

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

And consequently

$$\mu = \frac{\beta_j^* nq(1-q)}{s_j}, \qquad \sigma^2 = \frac{\sigma_\varepsilon^2 nq(1-q)}{s_j^2}, \qquad d_j = \frac{nq(1-q)}{s_j} + \lambda_2 s_j.$$

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 (8).

We are mainly interested in examining the case when $\boldsymbol{x}_j$ is imbalanced and what effect various approaches to normalizing the features has on the elastic net estimator. In order to study this, we will parameterize the scaling factors s_j for $j \in [p]$ by $s_j = (q - q^2)^\delta$, $\delta \geq 0$. This includes the cases that we are primary interested in, that is,

- $\delta = 0$: no scaling, which includes min–max and max–abs normalization,
- $\delta = 1/2$: standardization, and
- $\delta = 1$: scaling by the variance.

Note that the last of these cases does not correspond to a standard type of normalization. But as we will see, it has some interesting properties in the case of binary features.

A natual consequence of the normal distribution of Z is that the probability of selection in the elastic net problem is given by

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

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 Theorem 3.1, we show what the bias is under mean-centering and scaling with $s_j = (q - q^2)^\delta$, $\delta \geq 0$.

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*

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

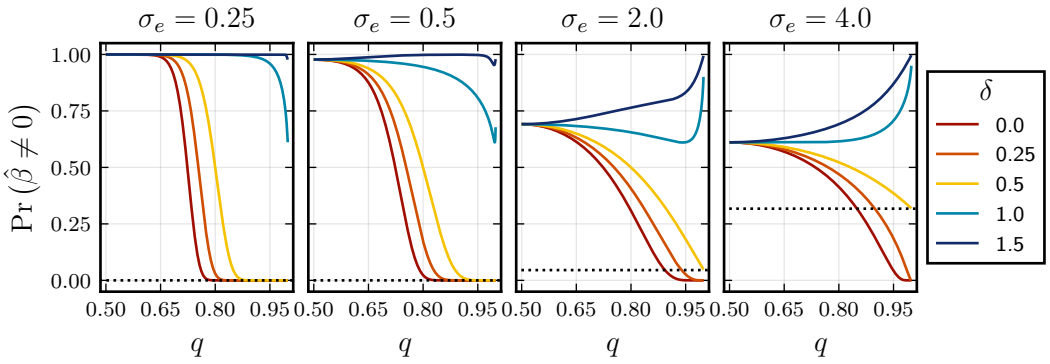

Figure 4: Probability of selection in the elastic net problem given a measurement noise level σ_ε, a regularization parameter λ_1, and a class balance q. The scaling factor is parameterized by $s_j = (q - q^2)^\delta$, $\delta \geq 0$. The dotted line represents the asymptotic limit for the standardization case, $\delta = 1/2$.

Proof. 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

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

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

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

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

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

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

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

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

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

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. □

Corollary 3.1.1 (Bias in Ridge Regression)**.** *Asume the conditions of Theorem 3.1 but that* $\lambda_1 = 0$*. Then*

$$\lim_{q \to 1^+} \mathbb{E}\hat{\beta}_j = \begin{cases} 0 & \text{if } 0 \le \delta < 1/2, \\ \frac{\beta_j^*}{1 + \frac{\lambda_2}{n}} & \text{if } \delta = 1/2, \\ 0 & \text{if } \delta > 1/2. \end{cases}$$

Proof. If $\lambda_1 = 0$, we have $\mathbb{E}\, \mathrm{S}_{\lambda_1}(Z) = \mu$ and consequently

$$\mathbb{E}\hat{\beta}_j = \frac{\beta_j^*}{1 + \frac{\lambda_2}{n}(q - q^2)^{2\delta - 1}}.$$

Then

$$\lim_{q \to 1^+} \mathbb{E}\hat{\beta}_j = \frac{\beta_j^*}{1 + \frac{\lambda_2}{n} \lim_{q \to 1^+}(q - q^2)^{2\delta - 1}}.$$

The result follows by noting that

$$\lim_{q \to 1^+} (q - q^2)^{2\delta - 1} = \begin{cases} \infty & \text{if } 0 \le \delta < 1/2, \\ 1 & \text{if } \delta = 1/2, \\ \beta_j^* & \text{if } \delta > 1/2. \end{cases}$$

□

Theorem 3.1 shows that the bias of the elastic net estimator when $0 \le \delta < 1/2$, which includes to the case of min–max and max–abs normalization (no scaling), approaches $-\beta_j^*$ as $q \to 1^+$. When $\delta = 1/2$ (standardization), the lasso estimate does not vanish completely. 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 $\delta \ge 1$, the estimate is unbiased asymptotically. The last fact may seem somewhat counterintuitive but is a consequence of the variance of the distribution of the estimator exploding 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* $s_j = (q - q^2)^\delta$*,* $\delta \ge 0$*, then*

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

Proof. The variance of the elastic net estimator is given by

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

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 $0 \le \delta < 1/2$, we write the limit of Equation (20) as

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

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 $\lim_{q\to 1^+} \frac{1}{d}\mathbb{E}\hat{\beta}_j = 0$, 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 $\lim_{q\to 1^+} \frac{1}{d}\mathbb{E}\hat{\beta}_j = \beta_j^*$ (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 ∞.

□

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

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

Proof. 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. □

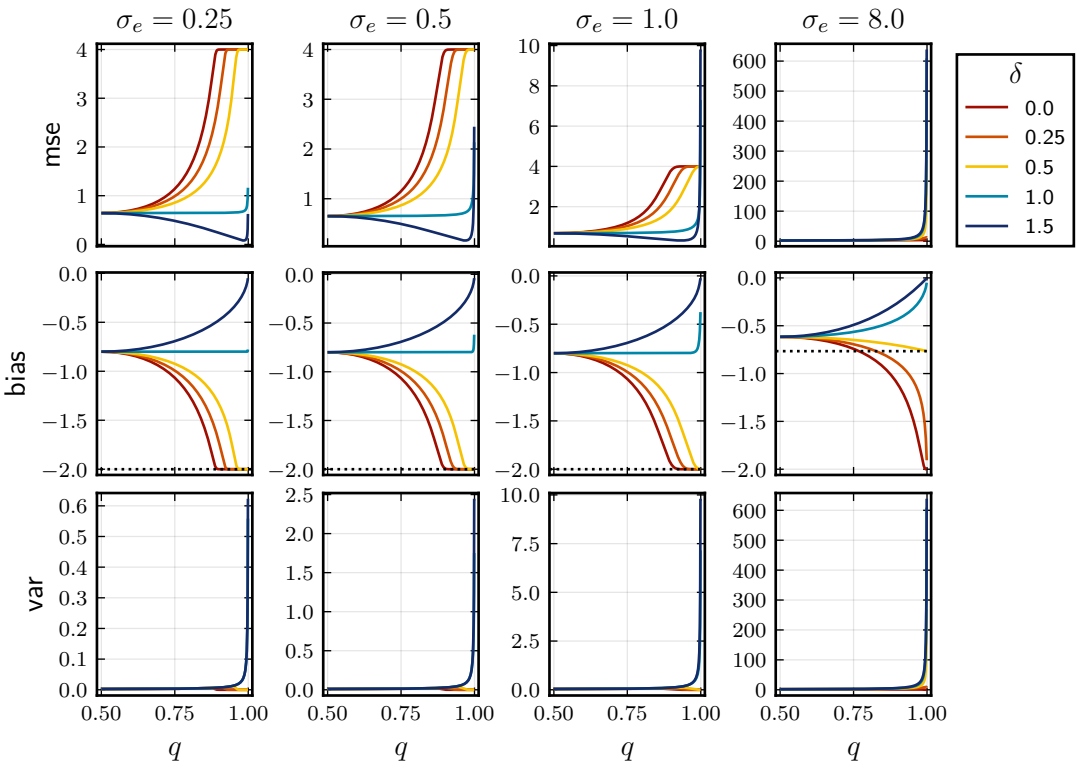

Figure 5: Bias, variance, and mean-squared error for a one-dimensional lasso problem. Note that $\delta = 0$ corresponds to the case of no scaling, $\delta = 1/2$ corresponds to standardization, and $\delta = 1$ corresponds to scaling with the variance. The dotted lines represent the asymptotic bias of the lasso estimator in the case of $\delta = 1/2$.

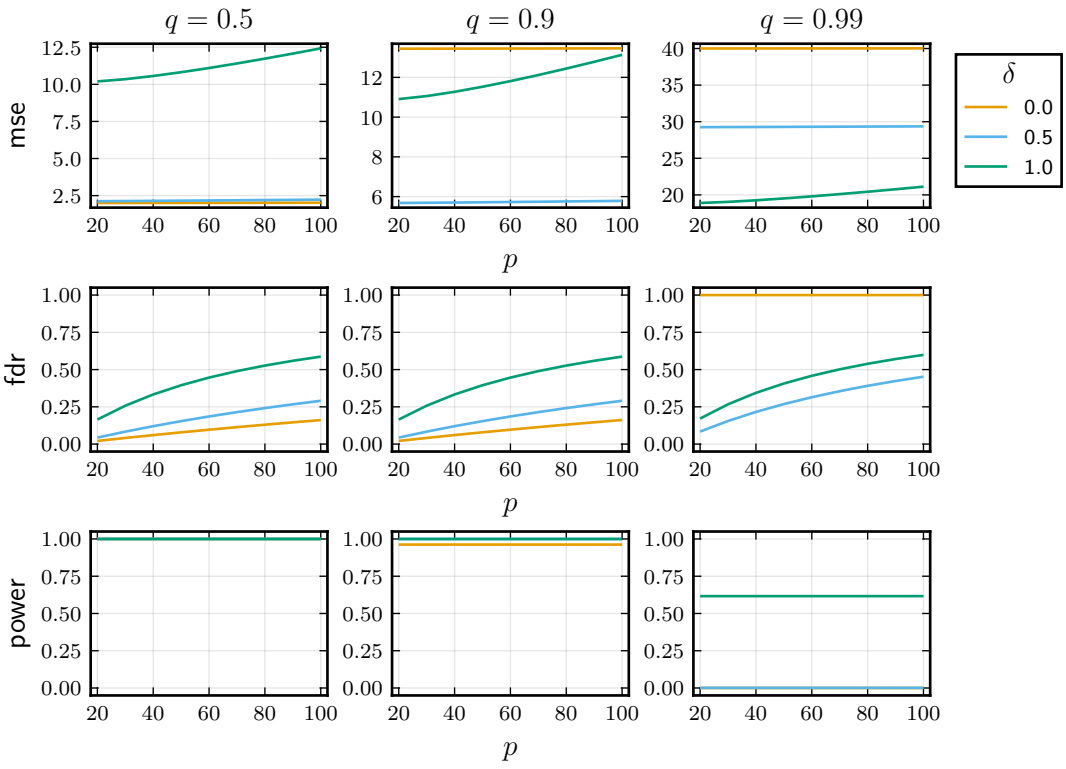

Figure 6: Mean-squared-error, false discovery rate (FDR), and power for a lasso problem with $k = 10$ true signals (nonzero β_j^*), varying p, and $q \in [0.5, 0.9, 0.99]$. The noise level is set at $\sigma_\varepsilon = 1$.

3.1 Multiple Predictors

4 Mixed Data

A natural follow-up topic to the discussion in the previous section is to consider the case where the features are of mixed type, that is, some are continuous and some are binary. To be able to compare normalization methods with respect to these cases, we need to construct problems in which the coefficients of the continuous and binary features are, in some sense, comparable. In this section, we will discuss what it means for a continuous and binary feature to have *comparable* effects and how the choice of normalization needs to be adapted to ensure that our penalized estimates respect this notion of comparability.

In this pape, we will focus on normally distrubted continuous features. We acknowledge that this is a limiting choice, but leave it to future papers to approach this issue for other types of distributions.

We will assume that the effect of a change in the binary variable (going from 0 to 1) corresponds to a difference of two standard deviations in the normally distributed variable. We base this choice on the reasoning by Gelman. In other words, if the regression coefficient of the binary variable is β_1^*, then the effect corresponding to a normally distributed random variable is equivalent if $\beta_2^* = (2\sigma)^{-1}\beta_1^*$.

Example 4.1. If $\boldsymbol{x}_2$ is sampled from Normal$(\mu, 2)$, then the effects of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are equivalent if $\beta_1^* = 1$ and $\beta_2^* = 0.25$.

Our particular choice of two standard deviations is not critical for our results, which hold for any other choice, as long it is linear with respect to the standard deviation of the normally distributed variable.

On the other hand, we also assume that the effects are equivalent irrespective of the class balance of the binary feature. In other words, we say that two binary features $\boldsymbol{x}_1$ and $\boldsymbol{x}_3$ have equivalent effects as long as $\beta_1^* = \beta_3^*$, even if the values in $\boldsymbol{x}_1$ are spread evenly between zeros and ones and those of $\boldsymbol{x}_3$ are all zeros except for one. We will see that this is a fundamental assumption upon which our results hinge entirely.

We will cover cases where the continuous feature is not normally distributed on a case-by-case basis as we proceed through the paper.

5 Experiments

5.1 Relative Size of Predictors in Model

The next question we now ask ourselves is: given that both features are in the model, what are their respective sizes given differences in class balance (q)?

To begin to answer this question, we conduct simulations on a two-dimensional problem. Along with our previous reasoning, we sample one feature from Normal$(0, 0.5)$ and the other from Bernoulli(q), varying q in $[0.5, 0.99]$ to simulate the effect of class imbalance on the estimates from the model. We compare four different strategies of normalization:

Mean-Std Standardization

Mean-StdVar Mean centering and scaling the normal feature by standard deviation and the binary feature by variance

Mean-Var Mean centering and scaling each feature by its variance

None No normalization

The results (Figure 7) show that when it comes to ridge, standardization creates class balance-insensitive estimates, whereas for the lasso, this is not the case. For the lasso, it is instead the Mean-StdVar and Mean-Var normalization methods that generate estimates that are insensitive to class imbalances.

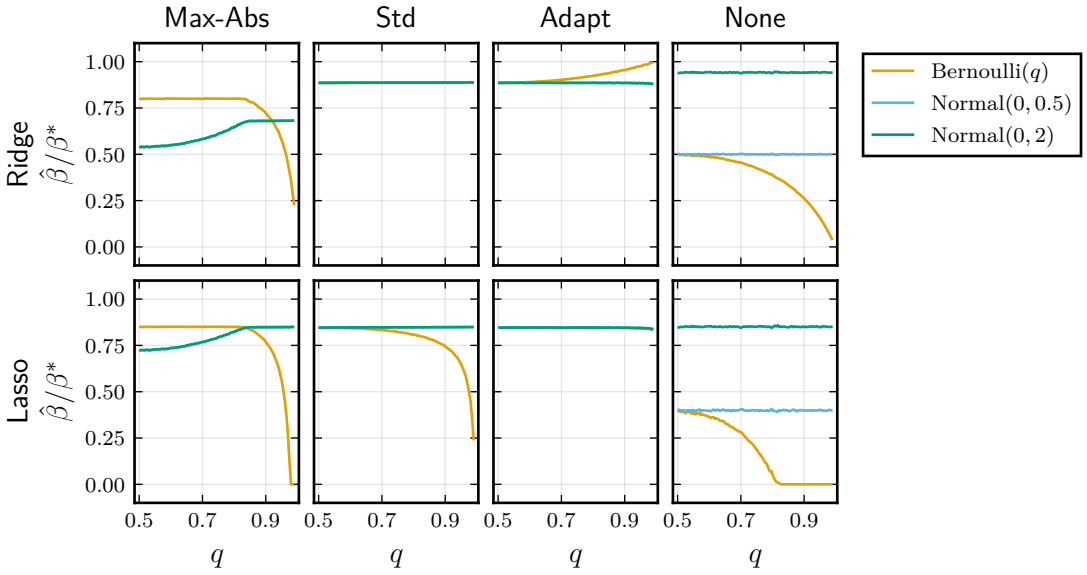

Figure 7: Comparison between lasso and ridge estimators for a two-dimensional problem where one feature is generated from Bernoulli(q) and the other from Normal$(0, 0.5)$ and the features are normalized in various ways.

5.2 Varying Class Imbalances

Here, we conduct an experiment on a 300×500 design matrix, where the first 20 features are binary and the remaining ones are normally distributed with standard deviation 0.5. We consider four different cases for the class balances:

Balanced All of the signals have a class balance of 0.5.

Unbalanced All of the signals have a class balance of 0.9.

Very Unbalanced All of the signals have a class balance of 0.99.

Decreasing The class balance of the signals decreases geometrically from 0.5 to 0.99.

To conduct the experiment, we generate random data and split it in a 50/50 training/test set split. Then, we select *lambda* using 10-fold cross validation on the training set and finally compute mean-squared error on the test set. We repeat this procedure 50 times for each combination of normalization type and class balance behavior.

The results (Figure 8) show that standardization performs best among the different types of normalization strategies.

5.3 Mixed Data

References

David L. Donoho and Iain M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. 81(3):425–455. ISSN 0006-3444. doi: 10.2307/2337118. URL https://www.jstor.org/stable/2337118.

Andrew Gelman. Scaling regression inputs by dividing by two standard deviations. 27(15):2865–2873. ISSN 02776715, 10970258. doi: 10.1002/sim.3107. URL https://onlinelibrary.wiley.com/doi/10.1002/sim.3107.

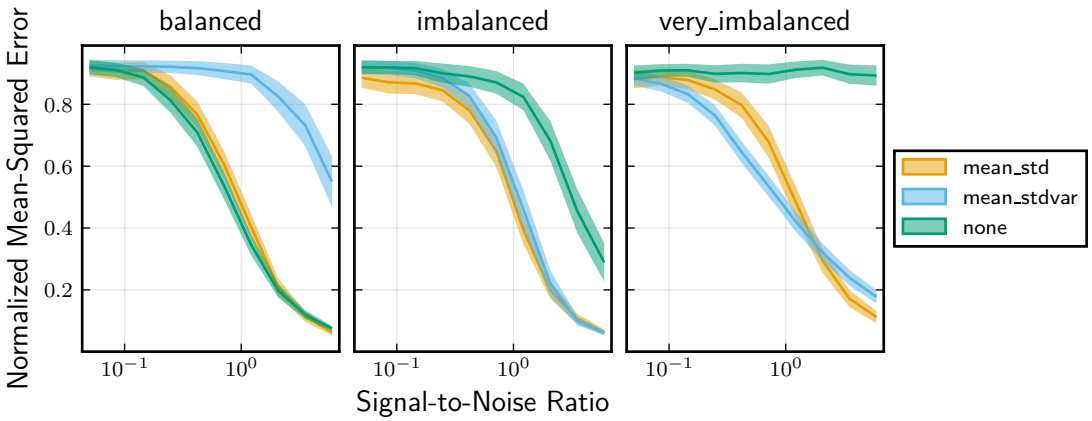

Figure 8: Mean-squared error of $y - \hat{y}$ for different types of normalizaion and types of class imbalances in a data set with only binary features.

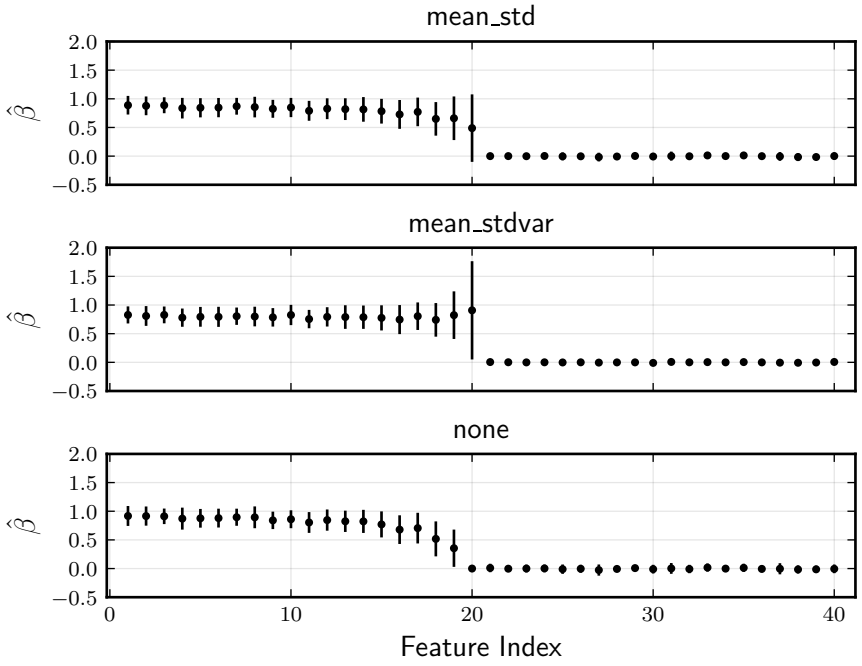

Figure 9: Estimates of the regression coefficients, $\hat{\beta}$, for the first 40 coefficients in the experiment. All of the features are binary and the first 20 features correspond to true signals, with a geometrically decreasing class balance from 0.5 to 0.99. The remaining features have a class balance that's randomly sampled from a uniform distribution with parameters 0.5 and 0.99.

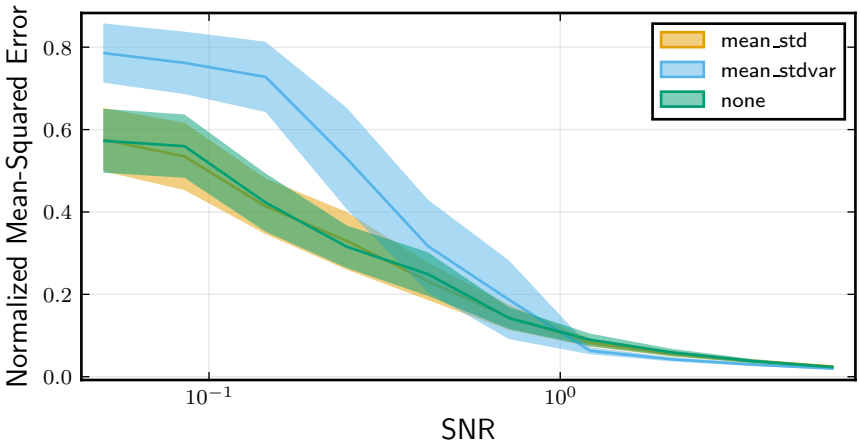

Figure 10: Prediction performance of an experiment with geometrically decreasing class balances for signals and varying signal to noise ratios.

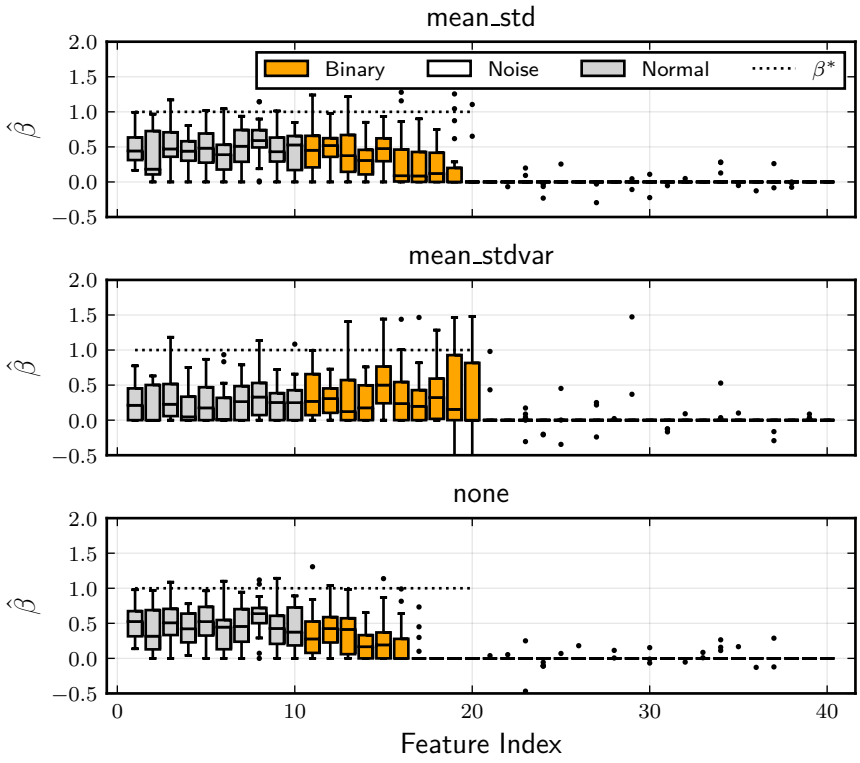

Figure 11: An experiment with mixed (normal and Bernouli-distributed) data.

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

A Appendix