UNIVERSITY OF
COPENHAGEN

The Choice of Normalization Influences Shrinkage in Regularized Regression

TMLR 2025

Johan Larsson Jonas Wallin

https://jolars.co, @jolars@mastodon.social

Department of Mathematical Sciences, Copenhagen University

November 11, 2025

The Elastic Net

Linear regression plus a combination of the ℓ_1 and ℓ_2 penalties:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \underbrace{\lambda_1 \|\beta\|_1}_{\text{lasso}} + \underbrace{\frac{\lambda_2}{2} \|\beta\|_2^2}_{\text{ridge}} \right)$$

The Elastic Net

Linear regression plus a combination of the ℓ_1 and ℓ_2 penalties:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{2} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \underbrace{\lambda_1 \|\beta\|_1}_{\text{lasso}} + \underbrace{\frac{\lambda_2}{2} \|\beta\|_2^2}_{\text{ridge}} \right)$$

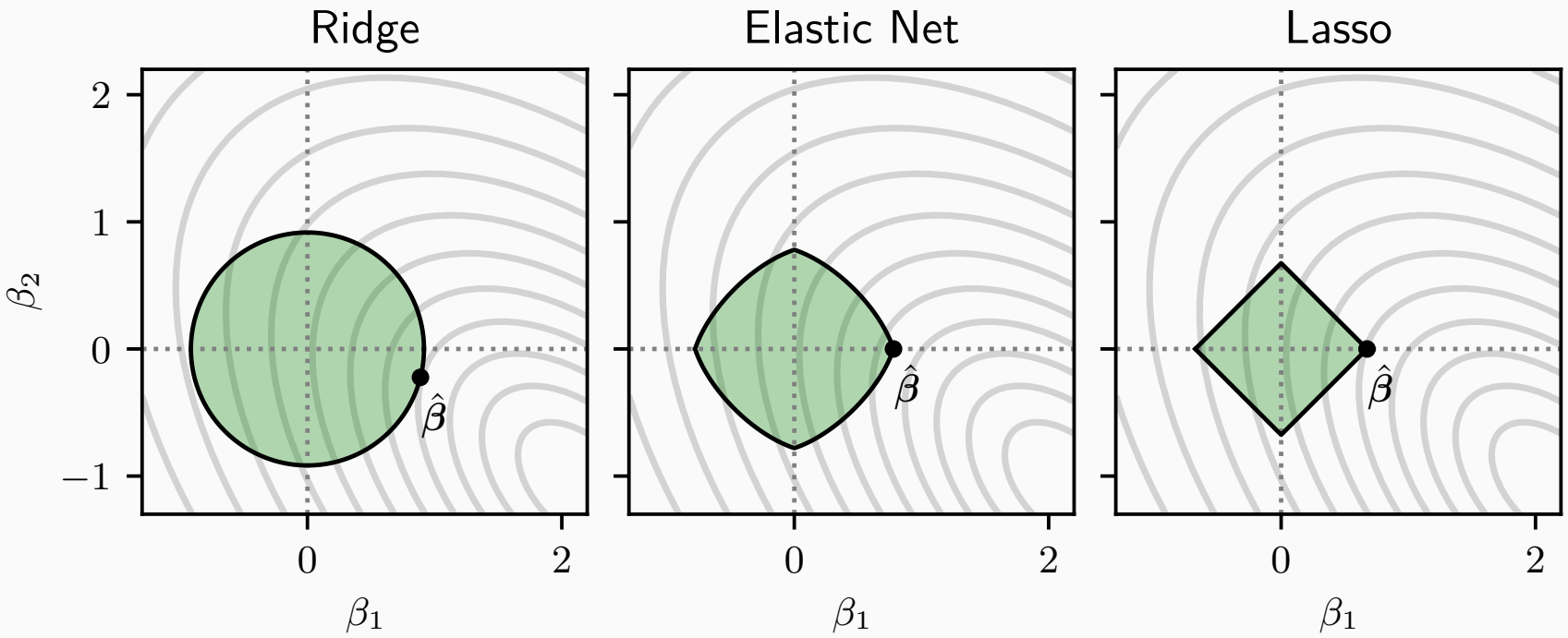

Figure 1: Elastic net is a combination of the lasso and ridge.

The Elastic Net Path

We don't know optimal λ_1 and λ_2 in advance; instead we use hyper-parameter optimization (e.g. cross-validation).

The Elastic Net Path

We don't know optimal λ_1 and λ_2 in advance; instead we use hyper-parameter optimization (e.g. cross-validation).

Common parametrization:

$$\lambda_1 = \alpha\lambda,$$
$$\lambda_2 = (1-\alpha)\lambda$$

with $\alpha \in [0, 1]$.

The Elastic Net Path

We don't know optimal λ_1 and λ_2 in advance; instead we use hyper-parameter optimization (e.g. cross-validation).

Common parametrization:

$$\lambda_1 = \alpha\lambda,$$
$$\lambda_2 = (1-\alpha)\lambda$$

with $\alpha \in [0, 1]$.

For each α, solve the elastic net over a sequence of λ: the **elastic net path**.

The Elastic Net Path

We don't know optimal λ_1 and λ_2 in advance; instead we use hyper-parameter optimization (e.g. cross-validation).

Common parametrization:

$$\lambda_1 = \alpha\lambda,$$
$$\lambda_2 = (1-\alpha)\lambda$$

with $\alpha \in [0, 1]$.

For each α, solve the elastic net over a sequence of λ: the **elastic net path**.

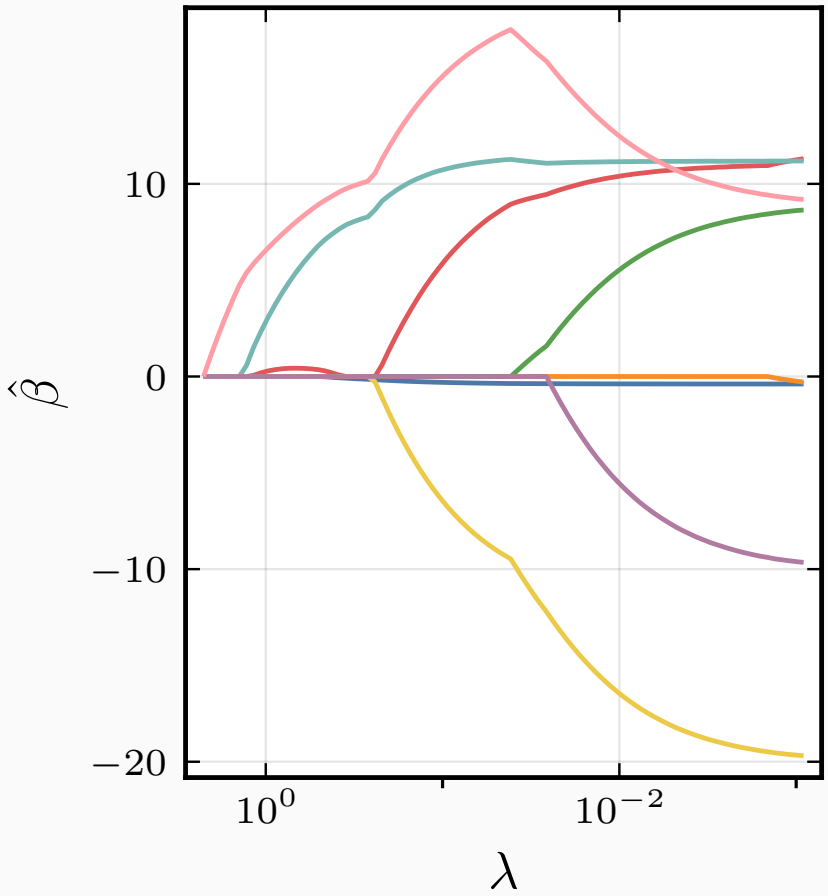

Figure 2: The elastic net path

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \text{Normal}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}4 & 0\\0 & 1\end{bmatrix}\right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix}\frac{1}{2}\\1\end{bmatrix}$$

and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Model	$\hat{\boldsymbol{\beta}}^{(n)}$	$\hat{\boldsymbol{\beta}}_{\text{std}}$
OLS	$\begin{bmatrix} 0.50 & 1.00 \end{bmatrix}^\top$	$\begin{bmatrix} 1.00 & 1.00 \end{bmatrix}^\top$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Model	$\hat{\boldsymbol{\beta}}^{(n)}$	$\hat{\boldsymbol{\beta}}_{\text{std}}$
OLS	$[0.50 \quad 1.00]^\top$	$[1.00 \quad 1.00]^\top$
Lasso	$[0.38 \quad 0.50]^\top$	$[0.74 \quad 0.50]^\top$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Model	$\hat{\boldsymbol{\beta}}^{(n)}$	$\hat{\boldsymbol{\beta}}_{\text{std}}$
OLS	$[0.50 \quad 1.00]^\top$	$[1.00 \quad 1.00]^\top$
Lasso	$[0.38 \quad 0.50]^\top$	$[0.74 \quad 0.50]^\top$
Ridge	$[0.37 \quad 0.41]^\top$	$[0.74 \quad 0.41]^\top$

Sensitivity to Scale

Lasso and ridge penalties are **norms**—feature scales matter!

Example

Assume

$$\boldsymbol{X} \sim \text{Normal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} \right), \qquad \boldsymbol{\beta}^* = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}$$

and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon$, with $\varepsilon \sim \text{Normal}(0, \sigma_\varepsilon^2)$.

Model	$\hat{\boldsymbol{\beta}}^{(n)}$	$\hat{\boldsymbol{\beta}}_{\text{std}}$
OLS	$[0.50 \quad 1.00]^\top$	$[1.00 \quad 1.00]^\top$
Lasso	$[0.38 \quad 0.50]^\top$	$[0.74 \quad 0.50]^\top$
Ridge	$[0.37 \quad 0.41]^\top$	$[0.74 \quad 0.41]^\top$

Large scale means less penalization because the size of β_j can be smaller for an equivalent effect (on $\boldsymbol{y}$).

Normalization

Scale sensitivity can be mitigated by normalizing the features. Let $\tilde{\boldsymbol{X}}$ be the normalized feature matrix, with elements

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

Normalization

Scale sensitivity can be mitigated by normalizing the features. Let $\tilde{\boldsymbol{X}}$ be the normalized feature matrix, with elements

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

After fitting, we transform the coefficients back to their original scale via

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

where $\hat{\beta}_j^{(n)}$ is a coefficient from the normalized problem.

Table 1: Common ways to normalize $\boldsymbol{X}$

Normalization	c_j	s_j
Standardization	$\bar{x}_j$	$\frac{1}{\sqrt{n}}\|\boldsymbol{x}_j - \bar{x}_j\|_2$
ℓ_1-Normalization	$\bar{x}_j$	$\frac{1}{\sqrt{n}}\|\boldsymbol{x}_j - \bar{x}_j\|_1$
Max–Abs	0	$\max_i \lvert x_{ij}\rvert$
Min–Max	$\min_i(x_{ij})$	$\max_i(x_{ij}) - \min_i(x_{ij})$

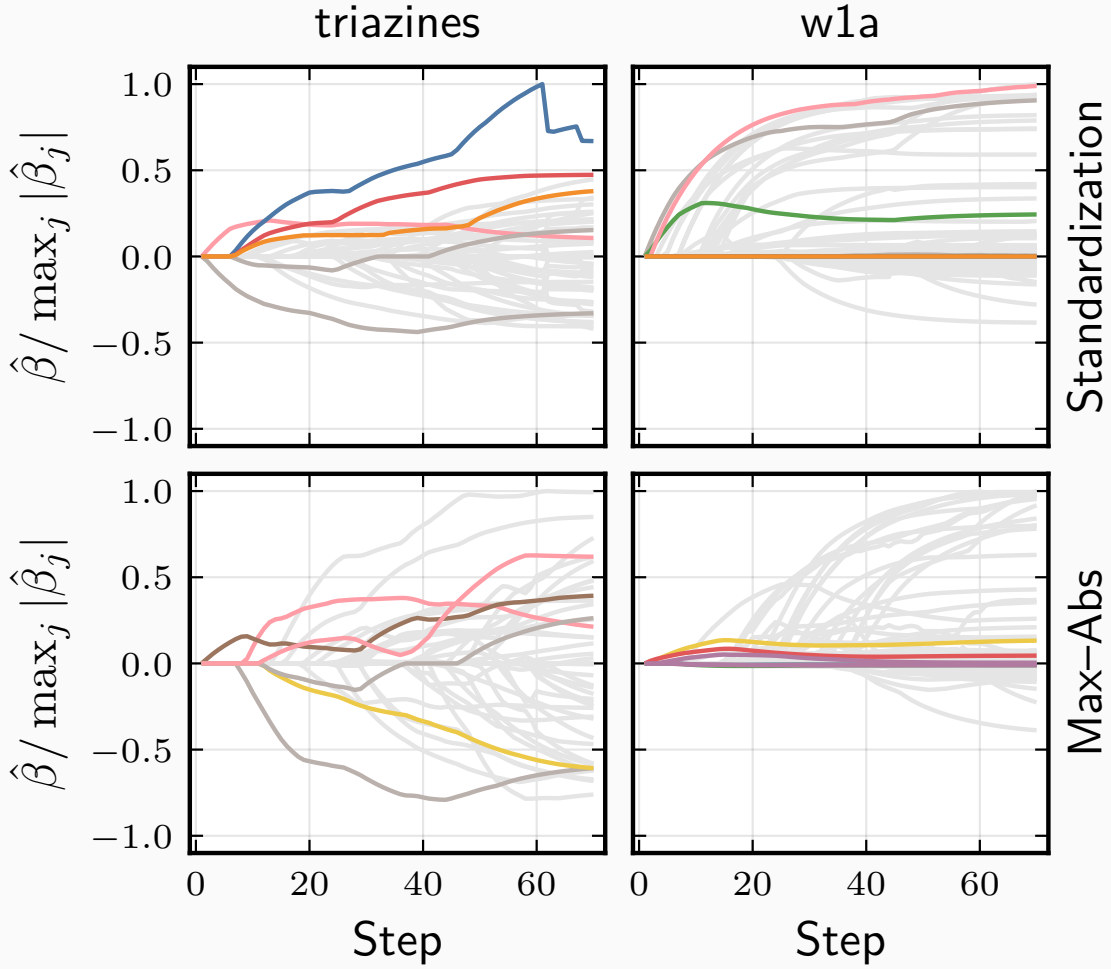

Figure 3: Normalization matters. Lasso paths under two different types of normalization (standardization and max–abs normalization). The union of the first five features selected in any of the settings are colored.

Table 2: Lasso coefficients on test sets, with λ set from 5-fold cross-validation repeated 5 times. We show the five largest coefficients in magnitude for each data set for the standardization setting and corresponding coefficients for the max–abs normalization setting.

`housing`		`triazines`		`w1a`	
$\hat{\beta}_{\text{std}}$	$\hat{\beta}_{\text{max–abs}}$	$\hat{\beta}_{\text{std}}$	$\hat{\beta}_{\text{max–abs}}$	$\hat{\beta}_{\text{std}}$	$\hat{\beta}_{\text{max–abs}}$
-0.63	-0.68	0.17	0.0	1.8	0.0
-1.4	-0.78	0.069	0.0	1.8	0.78
0.27	0.0	0.028	0.0	1.8	0.63
-0.99	-0.34	0.071	0.0	1.4	0.080
2.8	3.1	0.029	0.0	1.7	0.0

Paper Summary

Motivation
Normalization matters but there is no research into this.

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

The meaning of "standard" depends on field!

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

The meaning of "standard" depends on field!

What We Show
Normalization influences shrinkage for the elastic net for binary features.

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

The meaning of "standard" depends on field!

What We Show
Normalization influences shrinkage for the elastic net for binary features.

There's a bias–variance tradeoff that depends on the normalization.

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

The meaning of "standard" depends on field!

What We Show
Normalization influences shrinkage for the elastic net for binary features.

There's a bias–variance tradeoff that depends on the normalization.

For the elastic net, you should scale the penalty weights—not the features.

Paper Summary

Motivation
Normalization matters but there is no research into this.

Everyone agrees you need to normalize, but how to do so is usually motivated by being "standard".

The meaning of "standard" depends on field!

What We Show
Normalization influences shrinkage for the elastic net for binary features.

There's a bias–variance tradeoff that depends on the normalization.

For the elastic net, you should scale the penalty weights—not the features.

When mixing binary and normal features, normalization implicitly weighs their importance.

A Little Theory

Orthogonal Features

There is no explicit solution to the elastic net problem in general.

[1]We have also assumed that the features are mean-centered here.

Orthogonal Features

There is no explicit solution to the elastic net problem in general.

But if we assume that the features are orthogonal, that is

$$\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \operatorname{diag}(\tilde{\boldsymbol{x}}_1^\top \tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_p^\top \tilde{\boldsymbol{x}}_p),$$

then there is:[1]:

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\tilde{\boldsymbol{x}}_j^\top \boldsymbol{y}\right)}{s_j\left(\tilde{\boldsymbol{x}}_j^\top \tilde{\boldsymbol{x}}_j + \lambda_2\right)},$$

where

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

[1] We have also assumed that the features are mean-centered here.

Orthogonal Features

There is no explicit solution to the elastic net problem in general.

But if we assume that the features are orthogonal, that is

$$\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \operatorname{diag}(\tilde{\boldsymbol{x}}_1^\top \tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_p^\top \tilde{\boldsymbol{x}}_p),$$

then there is:[1]:

$$\hat{\beta}_j = \frac{\mathrm{S}_{\lambda_1}\left(\tilde{\boldsymbol{x}}_j^\top \boldsymbol{y}\right)}{s_j\left(\tilde{\boldsymbol{x}}_j^\top \tilde{\boldsymbol{x}}_j + \lambda_2\right)},$$

where

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

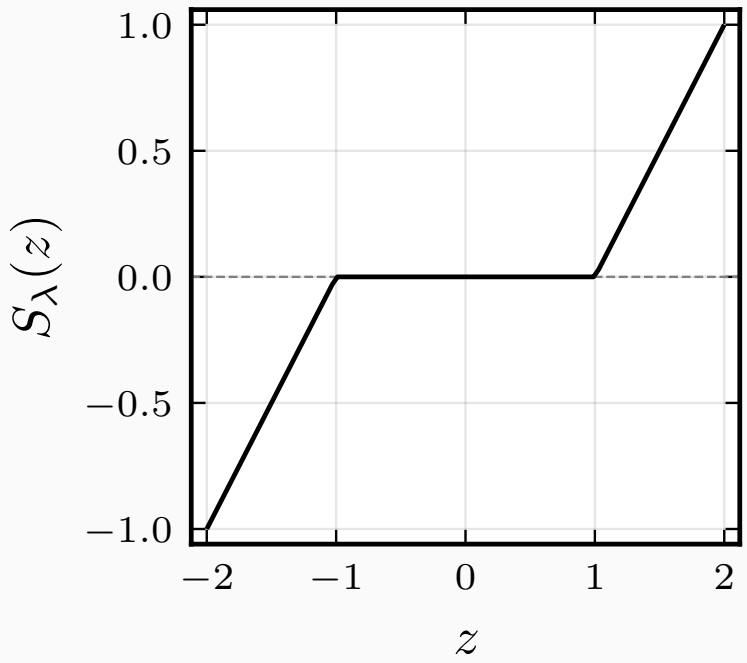

Figure 4: Soft thresholding

[1] We have also assumed that the features are mean-centered here.

Binary Features

Assume we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$. Let $q \in [0, 1]$ be the class balance of this feature, that is: $q = \bar{\boldsymbol{x}}_j$.

Binary Features

Assume we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$. Let $q \in [0, 1]$ be the class balance of this feature, that is: $q = \bar{\boldsymbol{x}}_j$.

Noiseless Case

In the noiseless case, we have

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

Binary Features

Assume we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$. Let $q \in [0, 1]$ be the class balance of this feature, that is: $q = \bar{\boldsymbol{x}}_j$.

Noiseless Case

In the noiseless case, we have

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

Means that the elastic net estimator depends on class balance (q).

Binary Features

Assume we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$. Let $q \in [0, 1]$ be the class balance of this feature, that is: $q = \bar{\boldsymbol{x}}_j$.

Noiseless Case
In the noiseless case, we have

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

Means that the elastic net estimator depends on class balance (q).

$s_j = q - q^2$ for lasso and $s_j = \sqrt{q - q^2}$ for ridge removes effect of q, which suggests the parametrization

$$s_j = (q - q^2)^\delta, \qquad \delta \geq 0.$$

Binary Features

Assume we have a binary feature $\boldsymbol{x}_j$, such that $x_{ij} \in \{0, 1\}$. Let $q \in [0, 1]$ be the class balance of this feature, that is: $q = \bar{\boldsymbol{x}}_j$.

Noiseless Case
In the noiseless case, we have

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

Means that the elastic net estimator depends on class balance (q).

$s_j = q - q^2$ for lasso and $s_j = \sqrt{q - q^2}$ for ridge removes effect of q, which suggests the parametrization

$$s_j = (q - q^2)^\delta, \qquad \delta \geq 0.$$

There is no (simple) s_j that will work for the elastic net.

Probability of Selection

Since $\boldsymbol{X}$ is fixed and ε is normal, we can compute the probability of selection.

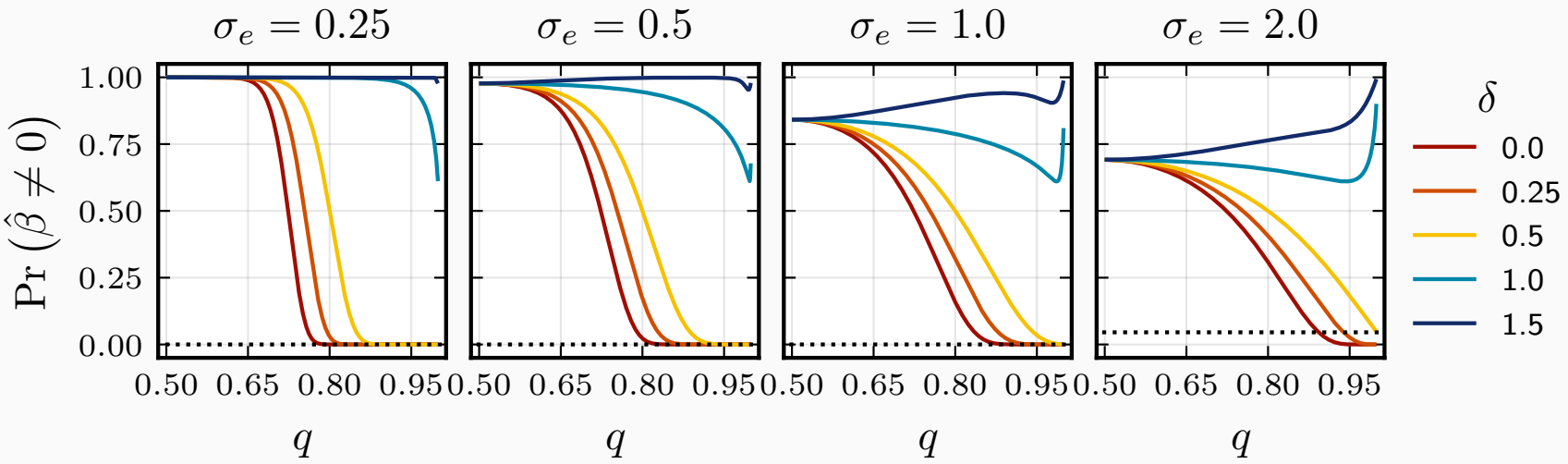

Figure 5: Probability that the elastic net selects a feature across different noise levels (σ_ε), types of normalization (δ), and class balance (q). The dashed line is asymptotic behavior for $\delta = 1/2$. Scaling used is $s_j \propto (q - q^2)^\delta$.

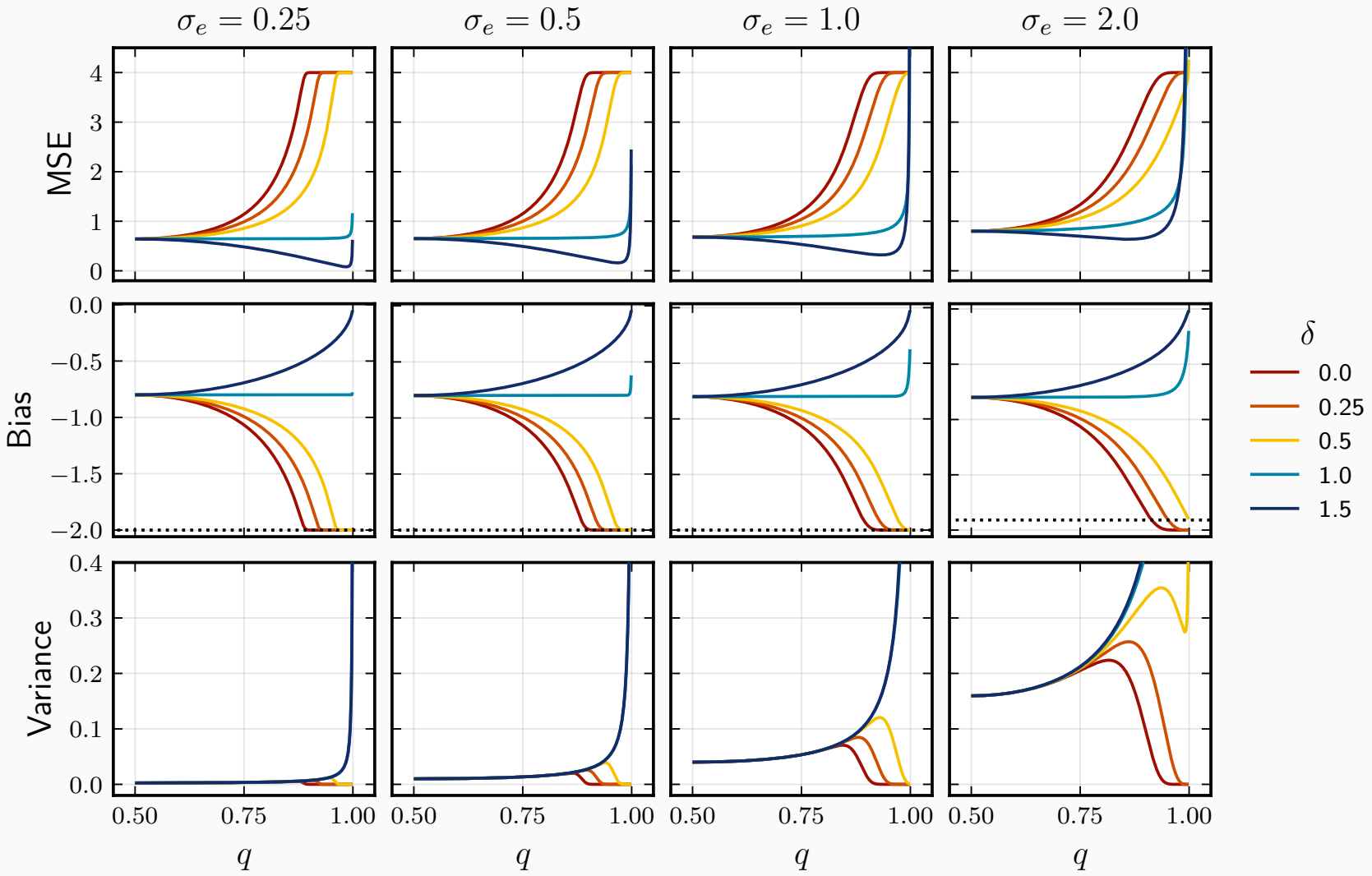

Figure 6: A bias variance tradeoff. Bias, variance, and mean-squared error for a one-dimensional lasso problem. Theoretical result for orthogonal features. Dotted line is asymptotic result or $\delta = 1/2$. Scaling used is $s_j \propto (q - q^2)^\delta$.

Experiments

Binary Features (Decreasing q)

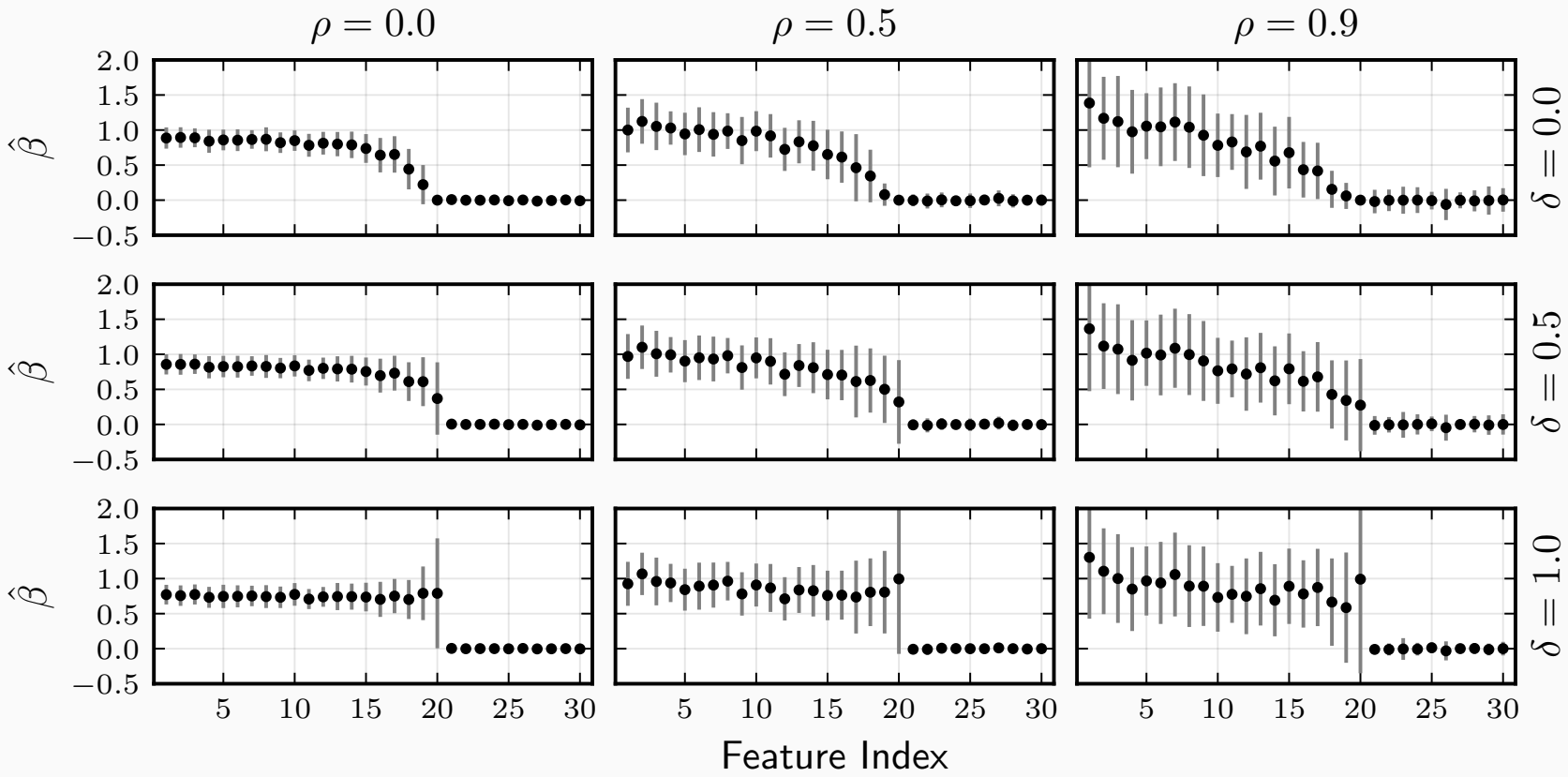

Figure 7: Lasso estimates for first 30 coefficients in an example where $n = 500$ and $p = 1000$. The first 20 features are true signals with a geometrically decreasing class balance from 0.5 to 0.99. ρ is a measure of autocorrelation.

Mixed Data

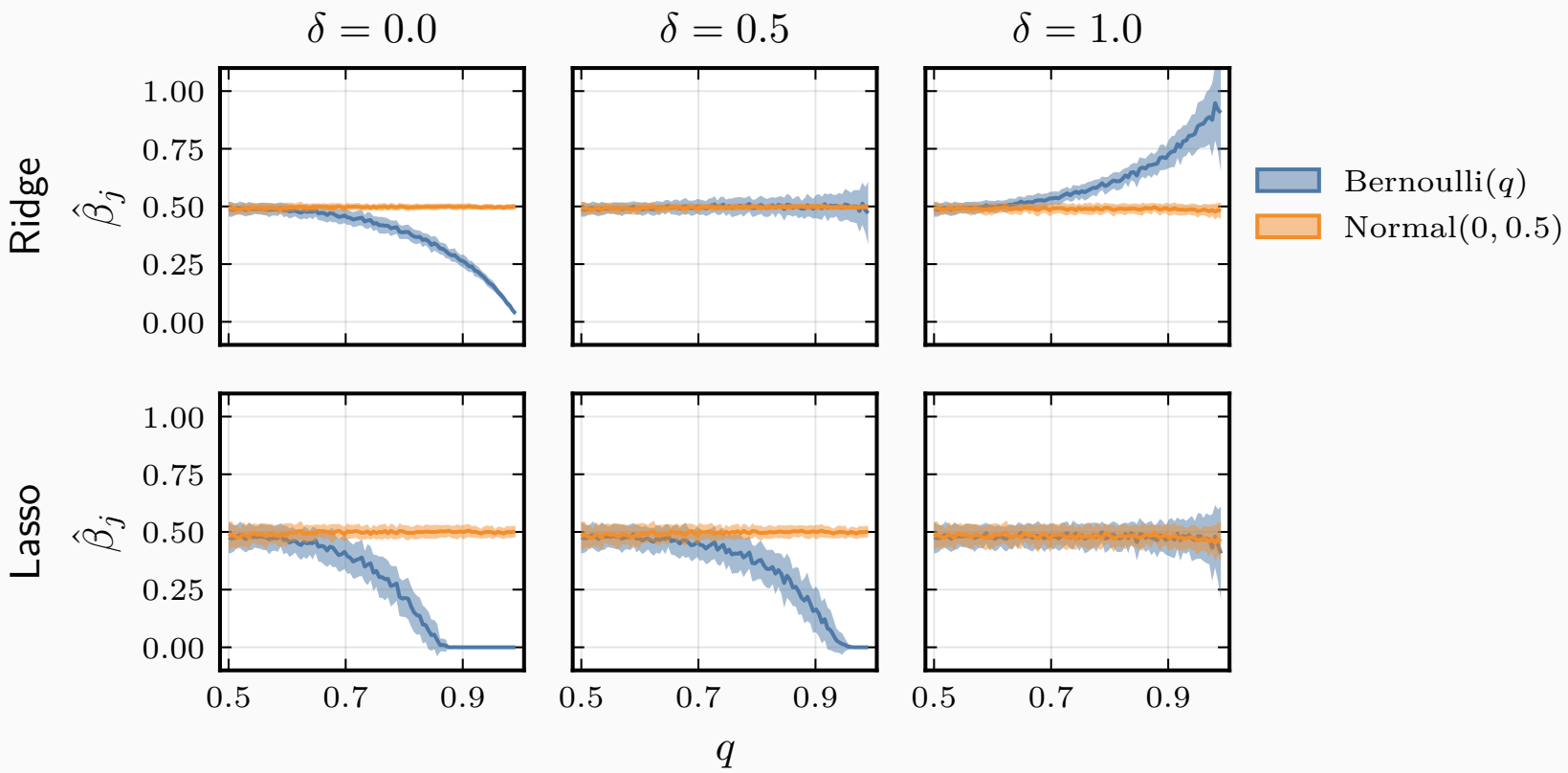

Figure 8: Comparison between different normalization strategies on a two-feature problem with one binary and one normal feature. The normal feature is always standardized, while the binary feature is scaled with $s_j \propto (q - q^2)^\delta$.

Hyperparameter Optimization

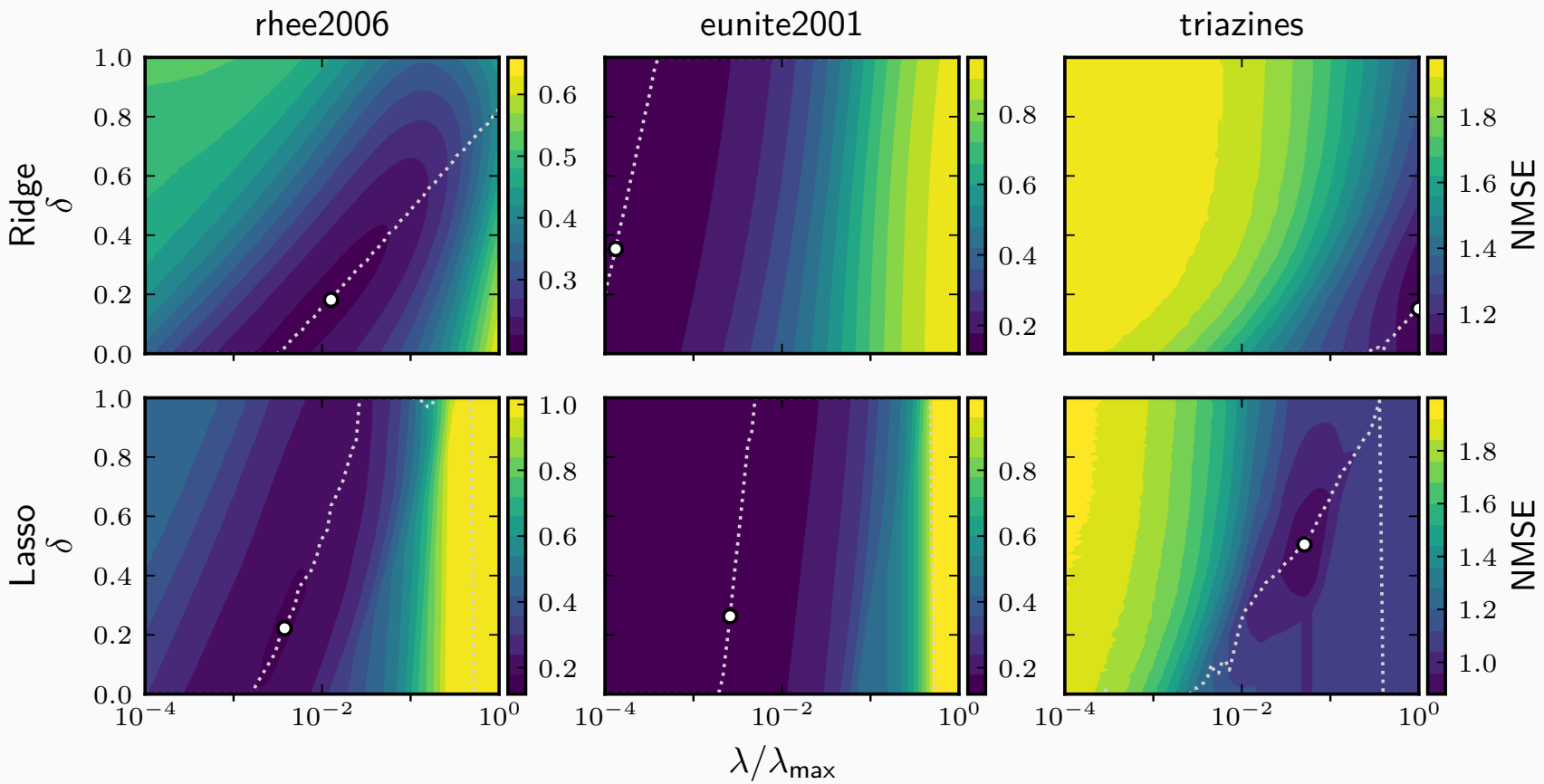

Figure 9: Contour plots of hold-out (validation set) error across a grid of δ and λ values for the lasso and ridge.

Summary

Class balance plays a crucial role when using regularized regression on binary data.

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

We introduced a new scaling approach to deal with class-imbalanced binary features.

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

We introduced a new scaling approach to deal with class-imbalanced binary features.

More Details in Paper

- Mixed data

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

We introduced a new scaling approach to deal with class-imbalanced binary features.

More Details in Paper

- Mixed data
- Interactions

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

We introduced a new scaling approach to deal with class-imbalanced binary features.

More Details in Paper

- Mixed data
- Interactions
- The Weighted Elastic Net

Summary

Class balance plays a crucial role when using regularized regression on binary data.

This is the first paper to investigate the interplay between normalization and regularization.

We introduced a new scaling approach to deal with class-imbalanced binary features.

More Details in Paper

- Mixed data
- Interactions
- The Weighted Elastic Net
- Many more experiments on real and simulated data

Thank you!