Testing and Estimation for Sparsity-Inducing Power Penalties

Maryclare Griffin

Department of Mathematics and Statistics
University of Massachusetts Amherst

November 15, 2023

Introduction

Introduction

Introduction

Introduction

Introduction

Introduction

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$
\begin{aligned}
&\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}(\mathbf{0}, \boldsymbol{I}_n) \\
&p(\beta_j) \propto \exp\{-(\lambda |\beta_j|^q\}
\end{aligned}
$$

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{I}_n)$$
$$p(\beta_j) \propto \exp\{-(\lambda |\beta_j|^q\}$$

- $p(\beta_j)$ is density of the exponential power distribution

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right) \\ &p\left(\beta_j\right) \propto \exp\left\{-\left(\lambda\left|\beta_j\right|^q\right\}\right. \end{aligned}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{I}_n)$$
$$p(\beta_j) \propto \exp\{-(\lambda |\beta_j|^q\}$$

- $p(\beta_j)$ is density of the exponential power distribution
- Special cases include:

  $q = 1$: Laplace model/lasso penalty

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$
$$p\left(\beta_j\right) \propto \exp\left\{-\left(\lambda \left|\beta_j\right|^q\right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

  $q = 1$: Laplace model/lasso penalty
  $q = 2$: Normal distribution/ridge penalty

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

Sparsity Inducing Power Penalties

Power penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$

Posterior mode of $\boldsymbol{\beta}$ under:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$
$$p\left(\beta_j\right) \propto \exp\left\{-\left(\lambda \left|\beta_j\right|^q\right\}\right.$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

  $q = 1$: Laplace model/lasso penalty
  $q = 2$: Normal distribution/ridge penalty
  $q \leq 1$: Bridge penalty

Frank and Friedman (1993); Polson et al. (2014); Subbotin (1923); Box (1953)

What is this talk going to be about?

"Testing Sparsity-Inducing Penalties"

What is this talk going to be about?

"Testing Sparsity-Inducing Penalties"

Testing when $q = 1$ is a "good" choice, picking a "better" q

What is this talk going to be about?

“Testing Sparsity-Inducing Penalties”

Testing when $q = 1$ is a “good” choice, picking a “better” q

“Improved Pathwise Coordinate Descent for Power Penalties”

What is this talk going to be about?

"Testing Sparsity-Inducing Penalties"

Testing when $q = 1$ is a "good" choice, picking a "better" q

"Improved Pathwise Coordinate Descent for Power Penalties"

Solving $\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda \, ||\boldsymbol{\beta}||_q^q$ for a range of q, λ

“Testing Sparsity-Inducing Penalties”

Problem

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1$

- ☺ Computationally simple

"Testing Sparsity-Inducing Penalties"

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$
- ☺ Same as finding posterior mode of $\boldsymbol{\beta}$ under

$$\begin{aligned} &\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right) \\ &p\left(\beta_j\right) \propto \exp\left\{-\lambda\left|\beta_j\right|\right\} \end{aligned}$$

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$
- ☺ Same as finding posterior mode of $\boldsymbol{\beta}$ under

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{I}_n)$$
$$p(\beta_j) \propto \exp\{-\lambda |\beta_j|\}$$

- ☹ Can perform poorly if $\boldsymbol{\beta}$ is very sparse or not sparse at all

"Testing Sparsity-Inducing Penalties"

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$
- ☺ Same as finding posterior mode of $\boldsymbol{\beta}$ under

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}(\boldsymbol{0}, \boldsymbol{I}_n)$$
$$p(\beta_j) \propto \exp\{-\lambda |\beta_j|\}$$

- ☹ Can perform poorly if $\boldsymbol{\beta}$ is very sparse or not sparse at all

Objectives

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$
- ☺ Same as finding posterior mode of $\boldsymbol{\beta}$ under

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\mathbf{0}, \boldsymbol{I}_n\right)$$
$$p\left(\beta_j\right) \propto \exp\left\{-\lambda\left|\beta_j\right|\right\}$$

- ☹ Can perform poorly if $\boldsymbol{\beta}$ is very sparse or not sparse at all

Objectives

- Test Laplace prior appropriateness

“Testing Sparsity-Inducing Penalties”

Problem

Lasso penalized regression solves $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$

- ☺ Computationally simple
- ☺ Yields possibly sparse estimates of $\boldsymbol{\beta}$
- ☺ Same as finding posterior mode of $\boldsymbol{\beta}$ under

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$
$$p\left(\beta_j\right) \propto \exp\left\{-\lambda \left|\beta_j\right|\right\}$$

- ☹ Can perform poorly if $\boldsymbol{\beta}$ is very sparse or not sparse at all

Objectives

- Test Laplace prior appropriateness
- Adaptively choose a better prior

Why bother treating the lasso penalty as a prior?

Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components

Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components, $\lambda = 2/\tau^2$

Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components, $\lambda = 2/\tau^2$
  - Variance components are interpretable

Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components, $\lambda = 2/\tau^2$
  - Variance components are interpretable
  - Variance component estimation has a long history

Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components, $\lambda = 2/\tau^2$
  - Variance components are interpretable
  - Variance component estimation has a long history
- Posterior summaries have decision theoretic justifications

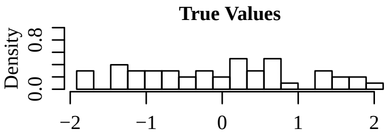
Why bother treating the lasso penalty as a prior?

- Tuning parameters become variance components, $\lambda = 2/\tau^2$
  - Variance components are interpretable
  - Variance component estimation has a long history
- Posterior summaries have decision theoretic justifications

$$\overbrace{\mathbb{E}\left[\boldsymbol{\beta}|\boldsymbol{X},\boldsymbol{y},\sigma^2,\lambda\right]}^{\text{Posterior Mean}} = \text{argmin}_{\boldsymbol{b}} \underbrace{\int ||\boldsymbol{\beta}-\boldsymbol{b}||_2^2\, p\left(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2\right) p\left(\boldsymbol{\beta}|\lambda\right) d\boldsymbol{\beta}}_{\text{Weighted average squared error loss}}$$

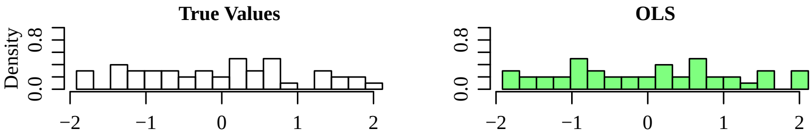
What if $\boldsymbol{\beta}$ is not sparse at all?

Simulated example with $n = 50$, $p = 45$, $\sigma^2 = \tau^2 = 1$



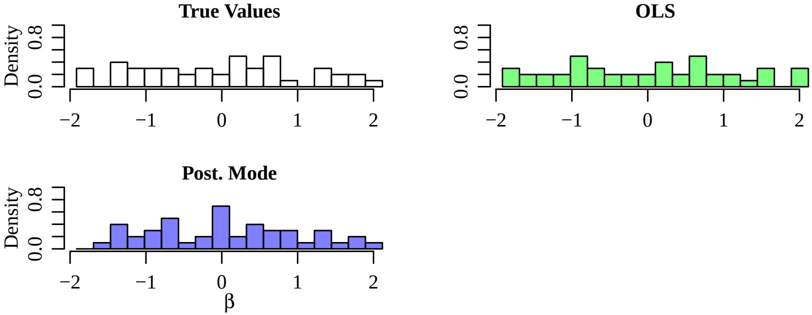
What if $\boldsymbol{\beta}$ is not sparse at all?

Simulated example with $n = 50$, $p = 45$, $\sigma^2 = \tau^2 = 1$



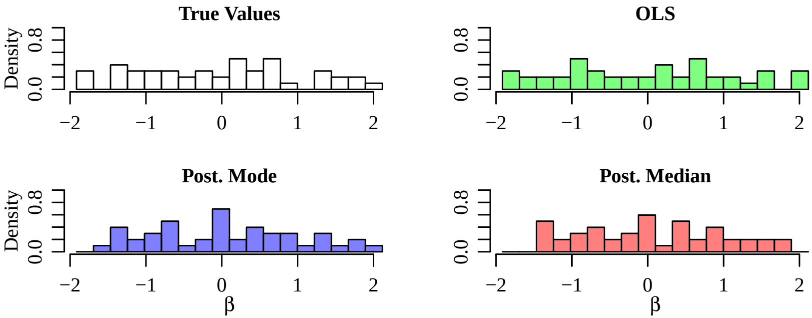
What if $\boldsymbol{\beta}$ is not sparse at all?

Simulated example with $n = 50$, $p = 45$, $\sigma^2 = \tau^2 = 1$



What if β is not sparse at all?

Simulated example with $n = 50$, $p = 45$, $\sigma^2 = \tau^2 = 1$



What if $\boldsymbol{\beta}$ is sparse?

What if $\boldsymbol{\beta}$ is sparse?

- Sparsity rate of mode under Laplace depends on σ^2/τ^2

What if $\boldsymbol{\beta}$ is sparse?

- Sparsity rate of mode under Laplace depends on σ^2/τ^2
- $\boldsymbol{\beta}$ could be spike-and-slab distributed

What if $\boldsymbol{\beta}$ is sparse?

- Sparsity rate of mode under Laplace depends on σ^2/τ^2
- $\boldsymbol{\beta}$ could be spike-and-slab distributed
  - β_j is exactly equal to zero with probability $1 - \pi$

What if $\boldsymbol{\beta}$ is sparse?

- Sparsity rate of mode under Laplace depends on σ^2/τ^2
- $\boldsymbol{\beta}$ could be spike-and-slab distributed
  - β_j is exactly equal to zero with probability $1-\pi$
  - $\beta_j \sim \text{normal}\left(0, \tau^2/\pi\right)$ distribution otherwise

What if $\boldsymbol{\beta}$ is sparse?

- Sparsity rate of mode under Laplace depends on σ^2/τ^2
- $\boldsymbol{\beta}$ could be spike-and-slab distributed
  - β_j is exactly equal to zero with probability $1 - \pi$
  - $\beta_j \sim$ normal $\left(0, \tau^2/\pi\right)$ distribution otherwise
- Sparsity rate of mode independent of sparsity rate of $\boldsymbol{\beta}$!

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Tricky because:

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Tricky because:

- $\boldsymbol{\beta}$ is latent

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p\left(\boldsymbol{\beta}\right)$

Largely appropriateness of normal random effects models

Tricky because:

- $\boldsymbol{\beta}$ is latent
- Estimates of $\boldsymbol{\beta}$ often reflect assumed model

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Tricky because:

- $\boldsymbol{\beta}$ is latent
- Estimates of $\boldsymbol{\beta}$ often reflect assumed model

Existing tests we could consider may:

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Tricky because:

- $\boldsymbol{\beta}$ is latent
- Estimates of $\boldsymbol{\beta}$ often reflect assumed model

Existing tests we could consider may:

- Depend on unknown parameters, σ^2, τ^2

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

Existing Literature on Testing Specification of $p(\boldsymbol{\beta})$

Largely appropriateness of normal random effects models

Tricky because:

- $\boldsymbol{\beta}$ is latent
- Estimates of $\boldsymbol{\beta}$ often reflect assumed model

Existing tests we could consider may:

- Depend on unknown parameters, σ^2, τ^2
- Not offer guidance given evidence of misspecification

Verbeke and Lesaffre (1996); Drikvandi et al. (2017)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{-\left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)}\right)^{-q/2} \left|\frac{\beta_j}{\tau}\right|^q\right\}$$

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\mathbf{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{-\left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)}\right)^{-q/2}\left|\frac{\beta_j}{\tau}\right|^q\right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{-\left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)}\right)^{-q/2}\left|\frac{\beta_j}{\tau}\right|^q\right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma \boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{ - \left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)} \right)^{-q/2} \left| \frac{\beta_j}{\tau} \right|^q \right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

  $q = 1$: Laplace model/lasso penalty

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{-\left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)}\right)^{-q/2}\left|\frac{\beta_j}{\tau}\right|^q\right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

  $q = 1$: Laplace model/lasso penalty
  $q = 2$: Normal distribution/ridge penalty

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

$$p\left(\beta_j\right) \propto \exp\left\{ - \left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)} \right)^{-q/2} \left| \frac{\beta_j}{\tau} \right|^q \right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

$q = 1$: Laplace model/lasso penalty
$q = 2$: Normal distribution/ridge penalty
$q \leq 1$: Bridge penalty

Subbotin (1923); Box (1953); Box and Tiao (1973)

A More General Model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{z}, \qquad \boldsymbol{z} \sim \text{normal}\left(\boldsymbol{0}, \boldsymbol{I}_n\right)$$

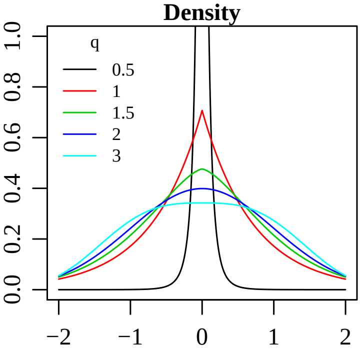
$$p\left(\beta_j\right) \propto \exp\left\{-\left(\frac{\Gamma\left(1/q\right)}{\Gamma\left(3/q\right)}\right)^{-q/2}\left|\frac{\beta_j}{\tau}\right|^q\right\}$$

- $p\left(\beta_j\right)$ is density of the exponential power distribution
- Special cases include:

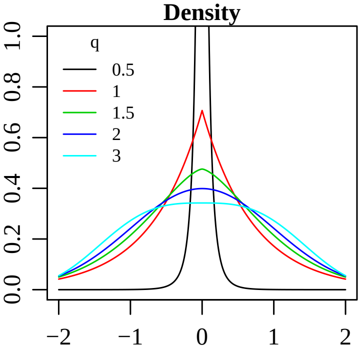
  $q = 1$: Laplace model/lasso penalty
  $q = 2$: Normal distribution/ridge penalty
  $q \leq 1$: Bridge penalty
  $q \to \infty$: Uniform distribution on $\left(-\sqrt{3}\tau, \sqrt{3}\tau\right)$

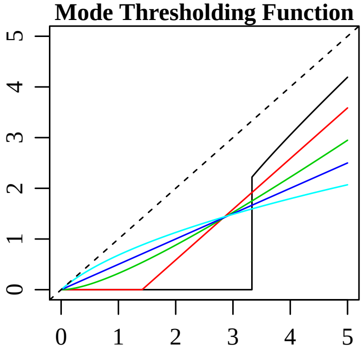
Subbotin (1923); Box (1953); Box and Tiao (1973)

Flexibility of the Exponential Power Distribution



Flexibility of the Exponential Power Distribution





How does the exponential power distribution help?

Relates uninterpretable q to interpretable

How does the exponential power distribution help?

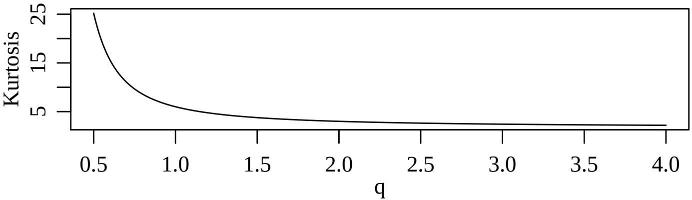
Relates uninterpretable q to interpretable

$$\underbrace{\kappa + 3 = \mathbb{E}[\beta_j^4]/\mathbb{E}[\beta_j^2]^2}_{\text{kurtosis}}$$

How does the exponential power distribution help?

Relates uninterpretable q to interpretable

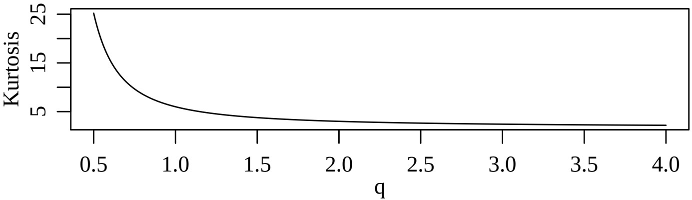
$$\underbrace{\kappa + 3 = \mathbb{E}[\beta_j^4]/\mathbb{E}[\beta_j^2]^2}_{\text{kurtosis}}$$



How does the exponential power distribution help?

Relates uninterpretable q to interpretable

$$\underbrace{\kappa + 3 = \mathbb{E}[\beta_j^4]/\mathbb{E}[\beta_j^2]^2}_{\text{kurtosis}}$$



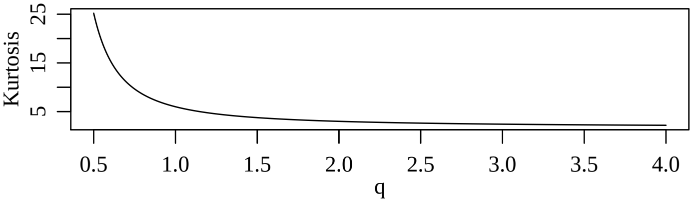
Facilitates:

- Testing $q = 1$

How does the exponential power distribution help?

Relates uninterpretable q to interpretable

$$\underbrace{\kappa + 3 = \mathbb{E}[\beta_j^4]/\mathbb{E}[\beta_j^2]^2}_{\text{kurtosis}}$$



Facilitates:

- Testing $q = 1$
- Specifying a “better” model if evidence suggests $q \neq 1$

Is it easy to test the null hypothesis that $q = 1$?

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

- Needs estimation of σ^2 and τ^2 under Laplace prior

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

- Needs estimation of σ^2 and τ^2 under Laplace prior
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

- Needs estimation of σ^2 and τ^2 under Laplace prior
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

- Needs estimation of σ^2 and τ^2 under Laplace prior
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Relationship between kurtosis and $q \rightarrow$

Is it easy to test the null hypothesis that $q = 1$?

Likelihood Ratio Test

- Needs maximum marginal likelihood estimation of σ^2, τ^2, q
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Score Test

- Needs estimation of σ^2 and τ^2 under Laplace prior
- Elements of $\boldsymbol{y}$ are <u>not</u> marginally independent
- Null distribution of test statistic may depend on σ^2, τ^2

Relationship between kurtosis and $q \rightarrow$ Moment-Based Test

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
  - Distribution of ψ is independent of τ^2, can be simulated

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
  - Distribution of ψ is independent of τ^2, can be simulated

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
  - Distribution of ψ is independent of τ^2, can be simulated
- Reject when $\psi \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
  - Distribution of ψ is independent of τ^2, can be simulated
- Reject when $\psi \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

Approximate Test: Construct $\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\hat{\psi} = \frac{\bar{b}^{(4)}}{\left(\bar{b}^{(2)}\right)^2}$

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
  - Distribution of ψ is independent of τ^2, can be simulated
- Reject when $\psi \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

Approximate Test: Construct $\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\hat{\psi} = \frac{\bar{b}^{(4)}}{\left(\bar{b}^{(2)}\right)^2}$

- Reject when $\hat{\psi} \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
    - Distribution of ψ is independent of τ^2, can be simulated
- Reject when $\psi \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

Approximate Test: Construct $\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$, $\hat{\psi} = \frac{\bar{b}^{(4)}}{\left(\bar{b}^{(2)}\right)^2}$

- Reject when $\hat{\psi} \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$
- Valid when $\operatorname{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \to 0$

An Approximate Level-α Test for $q = 1$, Full Rank $\boldsymbol{X}$

Oracle Test: Given $\boldsymbol{\beta}$, use $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ as test statistic

- Obtain quantiles $s_{\alpha/2}$, $s_{1-\alpha/2}$ of ψ under null
    - Distribution of ψ is independent of τ^2, can be simulated
- Reject when $\psi \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$

Approximate Test: Construct $\boldsymbol{b} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y}$, $\hat{\psi} = \frac{\bar{b}^{(4)}}{\left(\bar{b}^{(2)}\right)^2}$

- Reject when $\hat{\psi} \notin \left(s_{\alpha/2}, s_{1-\alpha/2}\right)$
- Valid when $\mathrm{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \to 0$
    - $\mathrm{tr}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) = \sum_{j=1}^{p} \frac{1}{\eta_j}$, i.e. $\boldsymbol{X}'\boldsymbol{X}$ not ill-conditioned

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\text{diag}\left\{\boldsymbol{X}'\boldsymbol{X}\right\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\operatorname{diag}\{\boldsymbol{X}'\boldsymbol{X}\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\text{diag}\{\boldsymbol{X}'\boldsymbol{X}\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$
- $\boldsymbol{b}_{\delta,\boldsymbol{X}} - \underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{\beta}_{\delta,\boldsymbol{X}}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\right)$

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\operatorname{diag}\{\boldsymbol{X}'\boldsymbol{X}\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$
- $\boldsymbol{b}_{\delta,\boldsymbol{X}} - \underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{\beta}_{\delta,\boldsymbol{X}}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\right)$

Approximate Test: Define $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\text{diag}\left\{\boldsymbol{X}'\boldsymbol{X}\right\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$
- $\boldsymbol{b}_{\delta,\boldsymbol{X}} - \underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{\beta}_{\delta,\boldsymbol{X}}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\right)$

Approximate Test: Define $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$

- Obtain quantiles of $\psi_{\delta,\boldsymbol{X}}$ under null

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\text{diag}\left\{\boldsymbol{X}'\boldsymbol{X}\right\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$
- $\boldsymbol{b}_{\delta,\boldsymbol{X}} - \underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{\beta}_{\delta,\boldsymbol{X}}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\right)$

Approximate Test: Define $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$

- Obtain quantiles of $\psi_{\delta,\boldsymbol{X}}$ under null
  - Distribution of $\psi_{\delta,\boldsymbol{X}}$ still independent of τ^2, can simulate

Extension: $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Let $\boldsymbol{V} = \sqrt{\text{diag}\left\{\boldsymbol{X}'\boldsymbol{X}\right\}}$ and $\boldsymbol{C} = \boldsymbol{V}^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{V}^{-1}$

- $\boldsymbol{b}_{\delta,\boldsymbol{X}} = \overbrace{\boldsymbol{V}^{-1}(\boldsymbol{C} + \delta^2\boldsymbol{I}_p)^{-1}\boldsymbol{V}^{-1}}^{\boldsymbol{D}}\boldsymbol{X}'\boldsymbol{y}$
- $\boldsymbol{b}_{\delta,\boldsymbol{X}} - \underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}}_{\boldsymbol{\beta}_{\delta,\boldsymbol{X}}} \sim \text{normal}\left(\boldsymbol{0}, \sigma^2\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\right)$

Approximate Test: Define $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$

- Obtain quantiles of $\psi_{\delta,\boldsymbol{X}}$ under null
  - Distribution of $\psi_{\delta,\boldsymbol{X}}$ still independent of τ^2, can simulate
- Compute $\hat{\psi}_{\delta,\boldsymbol{X}} = \frac{\bar{b}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{b}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$, compare to quantiles of $\psi_{\delta,\boldsymbol{X}}$

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved
- ☹ $\boldsymbol{b}_{\delta,\boldsymbol{X}} \to \mathbf{0}$, power decreases

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved
- ☹ $\boldsymbol{b}_{\delta,\boldsymbol{X}} \to \mathbf{0}$, power decreases

Our Recommendation:

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved
- ☹ $\boldsymbol{b}_{\delta,\boldsymbol{X}} \to \mathbf{0}$, power decreases

Our Recommendation:

Letting $\eta_1, \ldots, \eta_p$ be eigenvalues of $\boldsymbol{C}$

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved
- ☹ $\boldsymbol{b}_{\delta,\boldsymbol{X}} \to \mathbf{0}$, power decreases

Our Recommendation:

Letting $\eta_1, \ldots, \eta_p$ be eigenvalues of $\boldsymbol{C}$

$$\delta^2 = \left(1 - \min_j \eta_j\right)_+$$

Choice of δ^2 for $n < p$, Ill-Conditioned $\boldsymbol{X}'\boldsymbol{X}$

As $\delta \to \infty$

- ☺ $\hat{\psi}_{\delta,\boldsymbol{X}} \to \psi_{\delta,\boldsymbol{X}}$, nominal level more likely to be achieved
- ☹ $\boldsymbol{b}_{\delta,\boldsymbol{X}} \to \boldsymbol{0}$, power decreases

Our Recommendation:

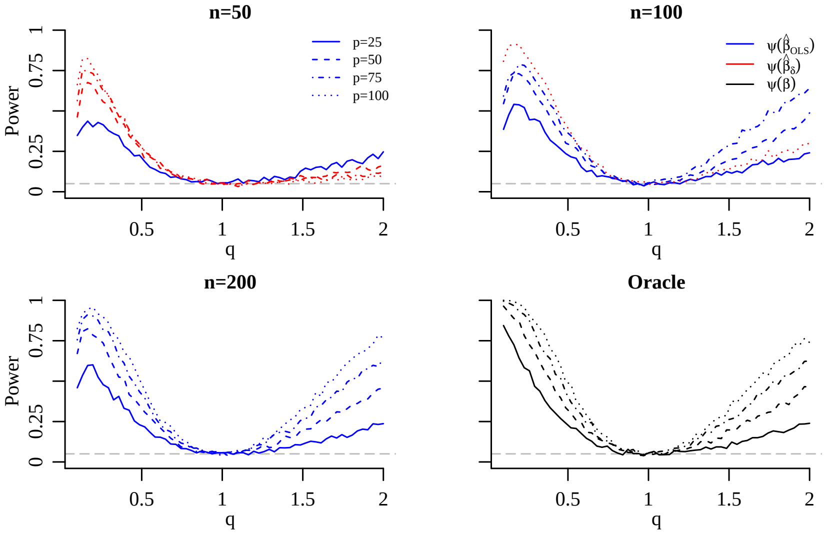
Letting $\eta_1, \ldots, \eta_p$ be eigenvalues of $\boldsymbol{C}$

$$\delta^2 = \left(1 - \min_j \eta_j\right)_+$$

If columns of $\boldsymbol{X}$ standardized to have norm n,

$$\mathrm{tr}\left(\underbrace{\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}}_{\text{Variance of } \boldsymbol{\beta}_{\delta,\boldsymbol{X}}}\right) \leq \frac{1}{n}\sum_{j=1}^{p}\frac{1}{\delta^2 + \eta_j}$$

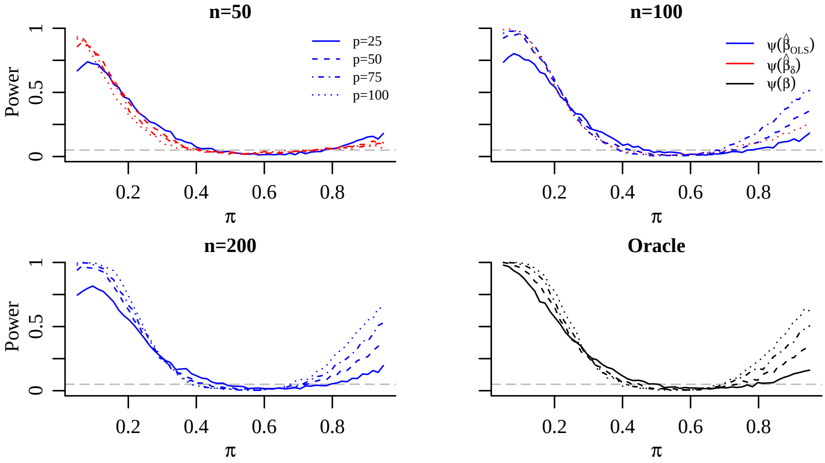
Test Performance: Exponential Power β



Test Performance: Spike-and-Slab $\boldsymbol{\beta}$

Laplace prior often used when $\boldsymbol{\beta}$ may be **sparse**

$$p\left(\beta_j \neq 0\right) = \pi,\ \beta_j | \beta_j \neq 0 \sim \text{normal}\left(0, \tau^2/\pi\right),\ \kappa = 3\left(\frac{1-\pi}{\pi}\right)$$



Specifying a “Better” Model After Rejection

1. Compute $\hat{q}$

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a "Better" Model After Rejection

1. Compute $\hat{q}$
2. Compute unbiased $\hat{\tau}^2$, $\hat{\sigma}^2$ from $\boldsymbol{b}_{\delta,\boldsymbol{X}}$ and residuals

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a "Better" Model After Rejection

1. Compute $\hat{q}$
2. Compute unbiased $\hat{\tau}^2$, $\hat{\sigma}^2$ from $\boldsymbol{b}_{\delta,\boldsymbol{X}}$ and residuals
3. Compute $\hat{\boldsymbol{\beta}}$ by:

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a “Better” Model After Rejection

1. Compute $\hat{q}$
2. Compute unbiased $\hat{\tau}^2$, $\hat{\sigma}^2$ from $\boldsymbol{b}_{\delta,\boldsymbol{X}}$ and residuals
3. Compute $\hat{\boldsymbol{\beta}}$ by:
    - Solving for posterior mode

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a “Better” Model After Rejection

1. Compute $\hat{q}$
2. Compute unbiased $\hat{\tau}^2$, $\hat{\sigma}^2$ from $\boldsymbol{b}_{\delta,\boldsymbol{X}}$ and residuals
3. Compute $\hat{\boldsymbol{\beta}}$ by:
   - Solving for posterior mode (**nonconvex** when $q < 1$)

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a "Better" Model After Rejection

1. Compute $\hat{q}$
2. Compute unbiased $\hat{\tau}^2$, $\hat{\sigma}^2$ from $\boldsymbol{b}_{\delta,\boldsymbol{X}}$ and residuals
3. Compute $\hat{\boldsymbol{\beta}}$ by:
    - Solving for posterior mode (**nonconvex** when $q < 1$)
    - Obtain posterior mean/median via Gibbs sampling

Marjanovic and Solo (2014); Polson et al. (2014)

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

Specifying a “Better” Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

  $\rightarrow$ Use $\hat{\psi} = \widehat{\kappa + 3}$

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

  $\rightarrow$ Use $\hat{\psi} = \widehat{\kappa + 3}$

When we use $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$, need to correct for bias of $\boldsymbol{\beta}_{\delta,\boldsymbol{X}}$:

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

  $\rightarrow$ Use $\hat{\psi} = \widehat{\kappa + 3}$

When we use $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$, need to correct for bias of $\boldsymbol{\beta}_{\delta,\boldsymbol{X}}$:

$$\widehat{\kappa + 3} = \left(\frac{\alpha^2_{\delta,\boldsymbol{X}}}{\gamma_{\delta,\boldsymbol{X}}}\right)\left(\hat{\psi}_{\delta,\boldsymbol{X}} - \frac{\omega_{\delta,\boldsymbol{X}}}{\alpha^2_{\delta,\boldsymbol{X}}}\right)$$

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

  $\rightarrow$ Use $\hat{\psi} = \widehat{\kappa + 3}$

When we use $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$, need to correct for bias of $\boldsymbol{\beta}_{\delta,\boldsymbol{X}}$:

$$\widehat{\kappa + 3} = \left(\frac{\alpha^2_{\delta,\boldsymbol{X}}}{\gamma_{\delta,\boldsymbol{X}}}\right)\left(\hat{\psi}_{\delta,\boldsymbol{X}} - \frac{\omega_{\delta,\boldsymbol{X}}}{\alpha^2_{\delta,\boldsymbol{X}}}\right)$$

- Reduces to $\hat{\psi}$ when $\delta = 0$

Specifying a "Better" Model After Rejection

Idea: Use $\hat{\psi}_{\delta,\boldsymbol{X}}$ to estimate $\kappa + 3$

- $\psi = \frac{\bar{\beta}^{(4)}}{\left(\bar{\beta}^{(2)}\right)^2}$ is a method of moments estimator of $\kappa + 3$

  $\rightarrow$ Use $\hat{\psi} = \widehat{\kappa + 3}$

When we use $\psi_{\delta,\boldsymbol{X}} = \frac{\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}}{\left(\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}\right)^2}$, need to correct for bias of $\boldsymbol{\beta}_{\delta,\boldsymbol{X}}$:

$$\widehat{\kappa + 3} = \left(\frac{\alpha^2_{\delta,\boldsymbol{X}}}{\gamma_{\delta,\boldsymbol{X}}}\right)\left(\hat{\psi}_{\delta,\boldsymbol{X}} - \frac{\omega_{\delta,\boldsymbol{X}}}{\alpha^2_{\delta,\boldsymbol{X}}}\right)$$

- Reduces to $\hat{\psi}$ when $\delta = 0$
- Can obtain $\hat{q}$ via $\widehat{\kappa + 3} = \Gamma(\frac{1}{\hat{q}})\Gamma(\frac{5}{\hat{q}})/\Gamma(\frac{3}{\hat{q}})^2$

Estimating Variance Components

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge
  - Scales slowly with p

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge
  - Scales slowly with p
  - Estimates of σ^2, τ^2 will depend on q

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge
  - Scales slowly with p
  - Estimates of σ^2, τ^2 will depend on q
- Moment-based estimators can be negative

Estimating Variance Components

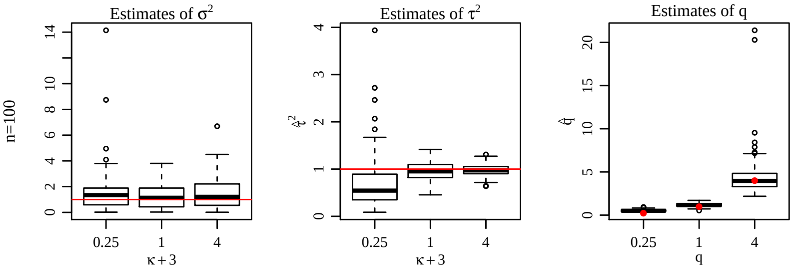
- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge
  - Scales slowly with p
  - Estimates of σ^2, τ^2 will depend on q
- Moment-based estimators can be negative
- Approximate σ^2, τ^2 obtained via MML for normal $\boldsymbol{\beta}$

Estimating Variance Components

- Gibbs-within-EM needed for maximum marginal likelihood
  - Can be slow to converge
  - Scales slowly with p
  - Estimates of σ^2, τ^2 will depend on q
- Moment-based estimators can be negative
- Approximate σ^2, τ^2 obtained via MML for normal $\boldsymbol{\beta}$

$$\min_{\tau^2,\sigma^2} \log\left(\left|\boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \boldsymbol{I}_n\sigma^2\right|\right) + \mathrm{tr}\left(\boldsymbol{y}\boldsymbol{y}^\top\left(\boldsymbol{X}\boldsymbol{X}^\top\tau^2 + \boldsymbol{I}_n\sigma^2\right)^{-1}\right)$$

Adaptive Simulations: Exponential Power β

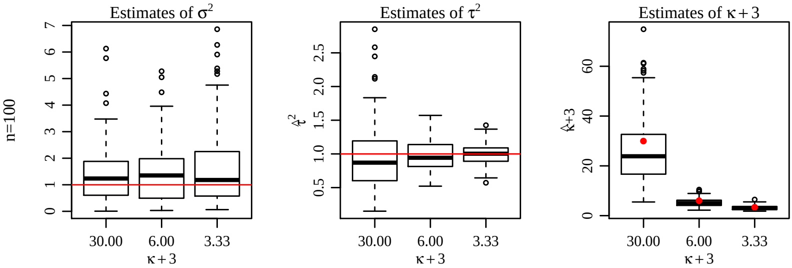


Adaptive Simulations: Exponential Power $\boldsymbol{\beta}$

p	100		
q	1/4	1	4
Null Hypothesis Rejected	83	5	69
MSE APMe $\leq$ MSE LPMe	**97**	**97**	**95**
MSE APMe $\leq$ MSE EPMe	**85**	**85**	50
MSE APMe $\leq$ MSE BB	**100**	**100**	49
MSE APMe $\leq$ MSE CV	**52**	**52**	**63**

Polson et al. (2014); Friedman et al. (2010)

Adaptive Simulations: Spike-and-slab $\boldsymbol{\beta}$



Adaptive Simulations: Spike-and-slab $\boldsymbol{\beta}$

p	100		
π	0.1	0.5	0.9
Pow. ADMo $\geq$ Pow. LPMo	**68**	**98**	**100**
Pow. ADMo $\geq$ Pow. EPMo	**100**	42	31
Pow. ADMo $\geq$ Pow. CV	**84**	**97**	**95**
FDR ADMo $\leq$ FDR LPMo	**100**	**100**	**75**
FDR ADMo $\leq$ FDR EPMo	**93**	**70**	**97**
FDR ADMo $\leq$ FDR CV	**70**	9	24
CS ADMo $\geq$ CS LPMo	**100**	**100**	**98**
CS ADMo $\geq$ CS EPMo	**93**	**70**	31
CS ADMo $\geq$ CS CV	**70**	16	**94**

Polson et al. (2014); Friedman et al. (2010)

Applications: Canonical Examples

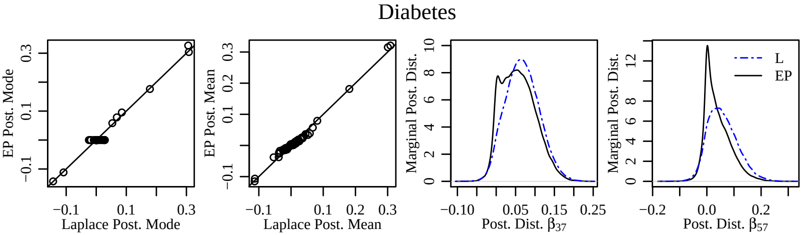
Dataset	n	p	$\psi_{\delta,0.025}$	$\psi_{\delta,0.975}$	$\psi(\hat{\boldsymbol{\beta}}_\delta)$
Diabetes	422	64	2.31	7.68	10.36
Boston Housing	506	104	1.97	7.61	6.57
Motif	287	195	2.87	10.35	5.77
Glucose	68	72	2.31	7.06	7.99

Efron et al. (2004); Park and Casella (2008); Polson et al. (2014); Bühlmann and van de Geer (2011); Priami and Morine (2015)

Application: Diabetes Data with Interactions

- $\boldsymbol{y}$ is a measure of diabetes progression for $n = 442$
- $p = 64$ main effects and interactions for 10 covariates

	Par. Ests.			Mode Sparsity	
Dataset	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{q}$	L	EP
Diabetes	0.4720	0.0071	0.5505	50.0%	87.5%

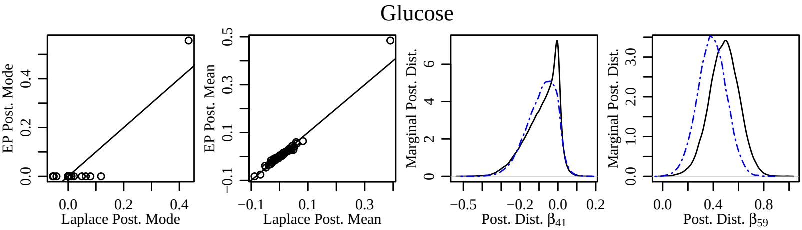


Efron et al. (2004)

Application: Glucose Data

- $\boldsymbol{y}$ is blood glucose concentration $\boldsymbol{y}$ for $n = 68$ subjects
- $p = 72$ health indicators, e.g. metabolite measurements

	Par. Ests.			Mode Sparsity	
Dataset	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{q}$	L	EP
Glucose	0.4754	0.0070	0.5939	83.3%	98.6%



Priami and Morine (2015)

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty
- Test statistic can guide model selection after rejection

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty
- Test statistic can guide model selection after rejection

Our approach has natural extensions to:

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty
- Test statistic can guide model selection after rejection

Our approach has natural extensions to:

- Testing appropriateness of different distributions for $\boldsymbol{\beta}$

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty
- Test statistic can guide model selection after rejection

Our approach has natural extensions to:

- Testing appropriateness of different distributions for $\boldsymbol{\beta}$
- Generalized linear models for $\boldsymbol{y}$

Recap: Testing Laplace Model/Lasso Penalty

We've shown:

- We can construct a test of the Laplace model/lasso penalty
- Test statistic can guide model selection after rejection

Our approach has natural extensions to:

- Testing appropriateness of different distributions for $\boldsymbol{\beta}$
- Generalized linear models for $\boldsymbol{y}$
- Nonnormal $\boldsymbol{z}$

“Improved Pathwise Coordinate Descent for Power…”

Problem

Mazumder et al. (2011)

“Improved Pathwise Coordinate Descent for Power…”

Problem

Solving $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_q^q$ is hard

---

Mazumder et al. (2011)

“Improved Pathwise Coordinate Descent for Power…”

Problem

Solving $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$ is hard

- ☹ No closed form solution for arbitrary q, even when $\boldsymbol{\beta} = \beta$

Mazumder et al. (2011)

"Improved Pathwise Coordinate Descent for Power…"

Problem

Solving $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$ is hard

- ☹ No closed form solution for arbitrary q, even when $\boldsymbol{\beta} = \beta$
- ☹ Solutions for arbirary q aren't nested

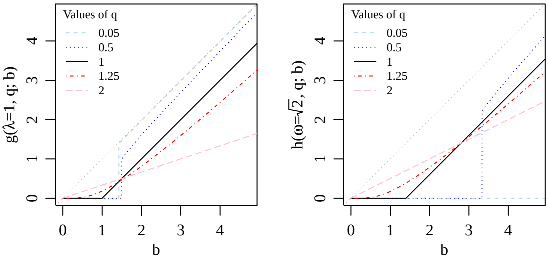
Mazumder et al. (2011)

"Improved Pathwise Coordinate Descent for Power. . ."

Problem

Solving $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} \left|\left|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right|\right|_2^2 + \lambda \left|\left|\boldsymbol{\beta}\right|\right|_q^q$ is hard

- ☹ No closed form solution for arbitrary q, even when $\boldsymbol{\beta} = \beta$
- ☹ Solutions for arbirary q aren't nested



Solution

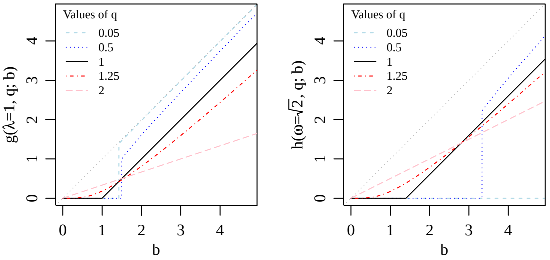
Mazumder et al. (2011)

“Improved Pathwise Coordinate Descent for Power…”

Problem

Solving $\min_{\boldsymbol{\beta}} \frac{1}{2\sigma^2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \lambda ||\boldsymbol{\beta}||_q^q$ is hard

- ☹ No closed form solution for arbitrary q, even when $\boldsymbol{\beta} = \beta$
- ☹ Solutions for arbirary q aren't nested



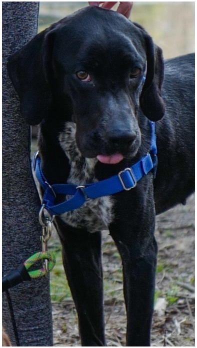
Solution

- ☺ Reparametrization to $\frac{1}{2} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||_2^2 + \left(\frac{\omega^{2-q}}{q}\right) ||\boldsymbol{\beta}||_q^q$

Mazumder et al. (2011)

Acknowledgements

- My advisor Peter Hoff
- NSF Grants DGE-1256082, DMS-1505136, & DMS-2113079
- My canine assistants



Box, G. E. P. (1953). A Note on Regions for Tests of Kurtosis. *Biometrika 40*(3/4), 465–468.

Box, G. E. P. and G. C. Tiao (1973). Bayesian Assessment of Assumptions. In *Bayesian Inference in Statistical Analysis*, Chapter 3, pp. 149–202. Reading, MA: Addison-Wesley Pub. Co.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Heidelberg: Springer.

Drikvandi, R., G. Verbeke, and G. Molenberghs (2017). Diagnosing misspecification of the random-effects distribution in mixed models. *Biometrics* (March), 63–71.

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

Frank, I. E. and J. H. Friedman (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics 35*(2), 109.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software 33*(1).

Marjanovic, G. and V. Solo (2014). lq Sparsity Penalized Linear Regression with Cyclic Descent. *IEEE Transactions on Signal Processing 62*(6), 1464–1475.

Mazumder, R., J. H. Friedman, and T. Hastie (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association 106*, 1125–1138.

Park, T. and G. Casella (2008). The Bayesian Lasso. *Journal of the American Statistical Association 103*(482), 681–686.

Polson, N. G., J. G. Scott, and J. Windle (2014). The Bayesian bridge. *Journal of the Royal Statistical Society. Series B: Statistical Methodology 76*(4), 713–733.

Priami, C. and M. J. Morine (2015). *Analysis of Biological Systems*. London: Imperial College Press.

Subbotin, M. T. (1923). On the Law of Frequency of Error. *Matematicheskii Sbornik 31*(2), 296–301.

Verbeke, G. and E. Lesaffre (1996). A Linear Mixed-Effects Model With Heterogeneity in the Random-Effects Population. *Journal of the American Statistical Association 91*(433), 217–221.

Details for Estimating $\hat{q}$

$$\left.\begin{aligned} \mathbb{E}[\bar{\beta}^{(2)}_{\delta,\boldsymbol{X}}] &= \alpha_{\delta,\boldsymbol{X}}\tau^2 \\ \mathbb{E}[\bar{\beta}^{(4)}_{\delta,\boldsymbol{X}}] &= \gamma_{\delta,\boldsymbol{X}}\left(\kappa+3\right)\tau^4 + \omega_{\delta,\boldsymbol{X}}\tau^4 \end{aligned}\right\} \to \widehat{\kappa+3} = \left(\frac{\alpha^2_{\delta,\boldsymbol{X}}}{\gamma_{\delta,\boldsymbol{X}}}\right)\left(\hat{\psi}_{\delta,\boldsymbol{X}} - \frac{\omega_{\delta,\boldsymbol{X}}}{\alpha^2_{\delta,\boldsymbol{X}}}\right)$$

$$\begin{aligned} \alpha_{\delta,\boldsymbol{X}} &= \operatorname{tr}\left(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\right)/p \\ \gamma_{\delta,\boldsymbol{X}} &= \sum_{j=1}^{p}\left(\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\right)^4_{jj}/p \\ \omega_{\delta,\boldsymbol{X}} &= 3\left(\sum_{j=1}^{p}\left(\boldsymbol{X}'\boldsymbol{X}\boldsymbol{D}\boldsymbol{D}\boldsymbol{X}'\boldsymbol{X}\right)^2_{jj}/p - \gamma_{\delta,\boldsymbol{X}}\right) \end{aligned}$$