The LASSO

CSci 8980: ML at Large Scale and High Dimensions

Instructor: Arindam Banerjee

January 29, 2014

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy
- Alternatives

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy
- Alternatives
  - Subset selection: Unstable, sensitive to small changes

Regression with OLS

- Given training data $(y_i, \mathbf{x}_i), i = 1, \ldots, n, \mathbf{x}_i \in \mathbb{R}^p$
- Ordinary least squares (OLS)

$$\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2$$

- Issues/challenges with OLS
  - Accuracy: low bias, high variance
  - Interpretation: All coefficients are non-zero
  - Cannot determine small subsets with strong effects
- Shrinking coefficients
  - Increases bias, lowers variance, improves accuracy
- Alternatives
  - Subset selection: Unstable, sensitive to small changes
  - Ridge regression: Shrinks coefficients, but not to 0

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \underset{(\alpha, c)}{\text{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t.} c_j \geq 0, \sum_j c_j \leq t$$

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \underset{(\alpha,c)}{\text{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t.} c_j \geq 0, \sum_j c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \underset{(\alpha,c)}{\text{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t.} c_j \geq 0, \sum_j c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \text{ s.t. } \sum_j |\beta_j| \leq t$$

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \underset{(\alpha,c)}{\text{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t.} c_j \geq 0, \sum_j c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \text{ s.t. } \sum_j |\beta_j| \leq t$$

- Parameter $t < t_0$ will cause shrinkage

The LASSO

- Let $\hat{\beta}^0$ be the OLS solution, and $t_0 = \sum_{j=1}^{p} |\hat{\beta}^0|$
- The non-negative garotte estimator (Breiman, 1996)

$$(\hat{\alpha}, \hat{c}) = \underset{(\alpha,c)}{\text{argmin}} \sum_{i=1}^{n} (y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \text{ s.t.} c_j \geq 0, \sum_j c_j \leq t$$

- Relies on OLS $\hat{\beta}^0$: may be problematic in certain settings
- Least absolute shrinkage and selection operator (LASSO)

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 \text{ s.t. } \sum_j |\beta_j| \leq t$$

- Parameter $t < t_0$ will cause shrinkage
  - Some coefficients will become 0

Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$

Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks k largest coefficients

Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks k largest coefficients
- For a suitable constant γ, the LASSO solution is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks k largest coefficients
- For a suitable constant γ, the LASSO solution is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

- Ridge regression shrinks the coefficients

$$\hat{\beta}_j^{\text{ridge}} = \frac{1}{1+\gamma}\hat{\beta}_j^0$$

Orthonormal Design Case

- Design matrix $X \in \mathbb{R}^{n \times p}$, assume $X^T X = I \in \mathbb{R}^{p \times p}$
- Best subset selection picks k largest coefficients
- For a suitable constant γ, the LASSO solution is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

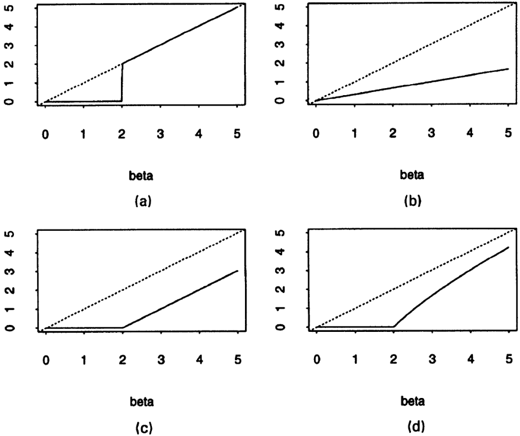
- Ridge regression shrinks the coefficients

$$\hat{\beta}_j^{\text{ridge}} = \frac{1}{1+\gamma} \hat{\beta}_j^0$$

- Garotte estimates

$$\hat{\beta}_j^{\text{garotte}} = \left(1 - \frac{\gamma}{(\hat{\beta}_j^0)^2}\right)_+ \hat{\beta}_j^0$$

Orthonormal Design Case



Shrinkage due to (a) subset selection, (b) ridge regression, (c) the lasso, and (b) the garotte

Geometry of LASSO

- Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)$$

Geometry of LASSO

- Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X (\beta - \hat{\beta}^0)$$

- Level sets of the contour intersects with L_q norm ball

Geometry of LASSO

- Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0)$$

- Level sets of the contour intersects with L_q norm ball
  - $q = 2$: Ridge regression, shrinkage but no sparsity

Geometry of LASSO

- Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0)$$

- Level sets of the contour intersects with L_q norm ball
  - $q = 2$: Ridge regression, shrinkage but no sparsity
  - $q = 1$: Lasso, shrinkage and sparsity

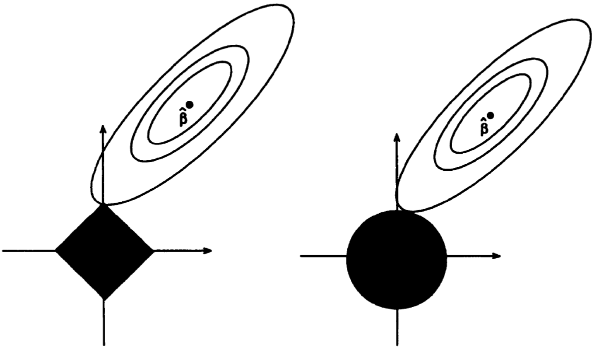
Geometry of LASSO

- Elliptical contour of the objective

$$(\beta - \hat{\beta}^0)^T X^T X(\beta - \hat{\beta}^0)$$

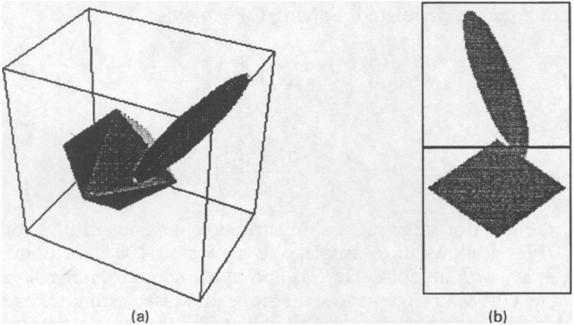
- Level sets of the contour intersects with L_q norm ball
  - $q = 2$: Ridge regression, shrinkage but no sparsity
  - $q = 1$: Lasso, shrinkage and sparsity
- Ridge vs Lasso: Can the sign change from OLS estimate?

Geometry of LASSO: $p = 2$



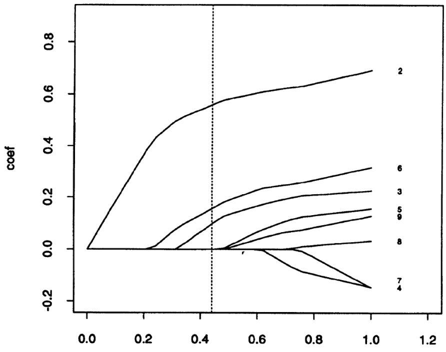
Estimation in (a) the lasso, and (b) ridge regression

Geometry of LASSO: $p > 2$



Sign change in LASSO vs OLS is possible for $p > 2$

Example: Regularization Path



Shrinkage of parameters over $s = \frac{t}{\sum_j \hat{\beta}_j^0}$

Estimating "t"

- The 'regularized' version of Lasso

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

Estimating "t"

- The 'regularized' version of Lasso

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

- Cross-validation over λ (or t)

Estimating "t"

- The 'regularized' version of Lasso

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

- Cross-validation over λ (or t)
  - Pick the value that leads to smallest error

Estimating "t"

- The 'regularized' version of Lasso

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin}_{(\alpha,\beta)} \sum_{i=1}^{n} (y_i - \alpha - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$$

- Cross-validation over λ (or t)
  - Pick the value that leads to smallest error
- Resampling based estimates, e.g., stability selection

Generalized Regression Models

- General regression problem formulation

Generalized Regression Models

- General regression problem formulation
  - Constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ L(y, X, \beta) \ \ \text{s.t.}\ \|\beta\|_1 \leq t$$

Generalized Regression Models

- General regression problem formulation
  - Constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; L(y, X, \beta) \;\; \text{s.t.} \; \|\beta\|_1 \leq t$$

  - Regularized version

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; L(y, X, \beta) + \lambda \|\beta\|_1$$

Generalized Regression Models

- General regression problem formulation
  - Constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ L(y, X, \beta) \ \text{ s.t. } \|\beta\|_1 \leq t$$

  - Regularized version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ L(y, X, \beta) + \lambda \|\beta\|_1$$

  - The other constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ \|\beta\|_1 \ \text{ s.t. } L(y, X, \beta) \leq a$$

Generalized Regression Models

- General regression problem formulation
  - Constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; L(y, X, \beta) \;\; \text{s.t.} \; \|\beta\|_1 \leq t$$

  - Regularized version

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; L(y, X, \beta) + \lambda \|\beta\|_1$$

  - The other constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \; \|\beta\|_1 \;\; \text{s.t.} \; L(y, X, \beta) \leq a$$

- Examples: logistic regression, generalized linear models, etc.

Generalized Regression Models

- General regression problem formulation
  - Constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ L(y, X, \beta) \ \text{ s.t. } \|\beta\|_1 \leq t$$

  - Regularized version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ L(y, X, \beta) + \lambda \|\beta\|_1$$

  - The other constrained version

$$\hat{\beta} = \underset{\beta}{\text{argmin}}\ \|\beta\|_1 \ \text{ s.t. } L(y, X, \beta) \leq a$$

- Examples: logistic regression, generalized linear models, etc.
- We will consider efficient algorithms for such general problems

Analysis of Risk: Donoho et al.

- Consider orthonormal design $X^T X = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

Analysis of Risk: Donoho et al.

- Consider orthonormal design $X^T X = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

- Let β be the 'true' parameter:

$$y = \beta^T \mathbf{x} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Analysis of Risk: Donoho et al.

- Consider orthonormal design $X^T X = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

- Let β be the 'true' parameter:

$$y = \beta^T \mathbf{x} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Consider risk

$$R(\hat{\beta}, \beta) = E\|\hat{\beta} - \beta\|^2$$

Analysis of Risk: Donoho et al.

- Consider orthonormal design $X^T X = I$, so Lasso estimate is

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$

- Let β be the 'true' parameter:

$$y = \beta^T \mathbf{x} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

- Consider risk

$$R(\hat{\beta}, \beta) = E\|\hat{\beta} - \beta\|^2$$

- Let R_{DP} be the loss of the 'optimal' predictor

$$T_{DP}(\hat{\beta}^0, \delta) = (\delta_j \hat{\beta}_j^0), \quad \delta_j = I(|\beta_j| > \sigma) \in \{0, 1\}$$

Analysis of Risk: Donoho et al.

- Consider orthonormal design $X^T X = I$, so Lasso estimate is
$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$$
- Let β be the 'true' parameter:
$$y = \beta^T \mathbf{x} + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$
- Consider risk
$$R(\hat{\beta}, \beta) = E\|\hat{\beta} - \beta\|^2$$
- Let R_{DP} be the loss of the 'optimal' predictor
$$T_{DP}(\hat{\beta}^0, \delta) = (\delta_j \hat{\beta}_j^0), \quad \delta_j = I(|\beta_j| > \sigma) \in \{0, 1\}$$
- T_{DP} needs knowledge of β, not practical

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}^0_j I(|\hat{\beta}^0_j| > \gamma)$

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}^0_j I(|\hat{\beta}^0_j| > \gamma)$
  - Has risk

$$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk
  $$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
  - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk
$$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
  - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk
    $$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
  - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
  - With $\gamma = \sigma(2 \log n)^{1/2}$, has same behavior

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk
  $$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
  - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
  - With $\gamma = \sigma(2 \log n)^{1/2}$, has same behavior
- General design matrices

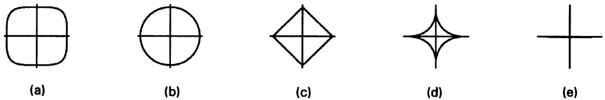
Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
    - Has risk
    $$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
    - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
    - With $\gamma = \sigma(2 \log n)^{1/2}$, has same behavior
- General design matrices
    - Lasso estimator continues to have good properties

Bounds on the Risk: Donoho et al.

- Hard threshold estimator $\tilde{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma)$
  - Has risk
  $$R(\tilde{\beta}, \beta) \leq (2 \log p + 1)(\sigma^2 + R_{DP})$$
  - Threshold $\gamma = \sigma(2 \log n)^{1/2}$ to get smallest asymptotic risk
- Soft threshold estimator $\hat{\beta}_j = \text{sign}(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)_+$
  - With $\gamma = \sigma(2 \log n)^{1/2}$, has same behavior
- General design matrices
  - Lasso estimator continues to have good properties
  - Generalized to other sparsity inducing norms

Norm level sets



L_q norm level sets: (a) $q = 4$, (b) $q = 2$, (c) $q = 1$, (d) $q = 0.5$, (e) $q = 0.1$