Topic 1: Linear regression and regularization

Eric B. Laber

Department of Statistical Science, Duke University

Statistics 561

The world needs another penalized regression method.
—Nobody (circa 2010).[1]

Penalization is like the word 'myself,' it's used too much (often incorrectly) by people wishing to seem more intelligent than they are.
—Joel Vaughan

[1] Relax, I'm 98.5 percent joking.

Warm-up (5 minutes)

- Explain to your group
  - Why is penalization/regularization used with predictive models?
  - How does penalization fit into the bias-variance trade-off?
  - What's the connection between ridge-regression and a Bayesian linear model?
- True or false
  - Masking occurs when there is a single large outlier
  - The Gauss-Markov theorem says the least squares estimator minimizes MSE
  - The owner of "the regulator bookshop" on 9th street thinks it's hilarious when you open the door and holler ''REGULATORS!!" like in the Warren G song

Roadmap

- Review and reminders
- All subsets (you don't want the truth)
- Ridge regression
- Lasso

Roadmap

- **Review and reminders**
- All subsets (you don't want the truth)
- Ridge regression
- Lasso

Review: fitting a linear model

- Ordinary least squares estimator as

$$\widehat{\beta}_n = \arg\min_{\beta \in \mathbb{R}^p} \mathbb{P}_n (Y - \boldsymbol{X}^\intercal \beta)^2 = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^\intercal \beta)^2$$

- Alternatively, view $\widehat{\beta}_n$ as solution to

$$\mathbb{P}_n \boldsymbol{X} (Y - \boldsymbol{X}^\intercal \beta) = 0$$

- If $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\intercal$ is invertible $\widehat{\beta}_n = (\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\intercal)^{-1} \mathbb{P}_n \boldsymbol{X} Y$

Review: fitting a linear model cont'd

- What to do if $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ is (nearly) singular?
  -
  -
  -
  -

Estimation and approximation error

- Suppose $Y = f(\boldsymbol{X}) + \epsilon$
  - Let $\beta^* = \arg\min_{\beta \in \mathbb{R}^p} P(Y - \boldsymbol{X}^\mathsf{T}\beta)^2$
- Decompose mean squared error as

$$\begin{aligned}\mathbb{E}(Y - \boldsymbol{X}^\mathsf{T}\widehat{\beta}_n)^2 &= \mathrm{Var}(\epsilon) + \mathbb{E}\left\{f(\boldsymbol{X}) - \boldsymbol{X}^\mathsf{T}\beta^*\right\}^2 + \mathbb{E}\left\{\boldsymbol{X}^\mathsf{T}(\widehat{\beta}_n - \beta^*)\right\}^2 \\ &= \text{Noise} + \text{Approx Err} + \text{Est. Err}\end{aligned}$$

- A complex model reduces the approximation error but increases the estimation error

Parsimonious models

- Predictive models used to inform decision making
  - Generate interesting hypotheses for further study
  - Forecasts weighed by stakeholders
  - Drive automatic action, i.e., decide if and what type of push notification to send patient in mHealth
- Need to build trust with stakeholders[2]
  - Models must be validated in domain context
  - Interpretable models often required

[2]Note* this might be you! If you don't trust your own models it will be hard to make progress.

Parsimonious models cont'd

- Justification for parsimonious models
  - Occam's Razor
  - (Medicine) true optimal decision rule simple
  - More nuanced mathematical arguments[3]
- Linear models aren't really that interpretable
  - Easy when abstracted away from the problem
  - Notion of one variable moving while all others held fix can be nonsensical to domain experts in some contexts[4]
  - But the fewer terms they have, the more interpretable they tend to be (we'll explore lists and trees later in this course)

[3] Duke's own Cynthia Rudin has some excellent work in this area. Check out https://impact.duke.edu/story/whats-in-the-box and her papers on this topic.

[4] Imagine increasing the square-footage of a grocery store without changing the products, back-of-house inventory, layout, etc.

It is our experience and strong belief that better models and a better understanding of ones data result from focused data analysis, guided by substantive theory.
–Gwyneth Paltrow[5]

Automatic model-building procedures should be avoided at all cost. –D.R. Cox[6]

[5] For the best in psychic vampire repellent see:
https://goop.com/paper-crane-apothecary-psychic-vampire-repellent/p/

[6] This is a real one.

Roadmap

- Review and reminders
- **All subsets (you don't want the truth)**
- Ridge regression
- Lasso

Finding the 'true' model

- Assume that $\beta_j^* = 0$ for $j \in \mathcal{J}$
  - Only need to regress Y on X_j where $j \notin \mathcal{J}$
  - Let $\mathbf{X}_{\mathcal{J}^c}$ denote relevant predictors
  - Question: suppose an oracle gave you $\mathcal{J}$ is $\widehat{\beta}_n^{\mathcal{J}^c} = \arg\min_{\boldsymbol{\beta}} \mathbb{P}_n(Y - \boldsymbol{X}_{\mathcal{J}^c}^\intercal \boldsymbol{\beta})^2$ the optimal estimator?
- `R` code example: growingLinearModel.R[7]

[7] We'll mostly use python in this course but sometimes a guy already has some R code written. Geez, get off my back.

Finding the 'true' model

- If you want optimal predictions, you may not want to use the 'true' model even if it were available to you
  - Estimating small effects inflates variance but does little to improve prediction
  - What constitutes a small effect depends on residual variance and sample size (not an absolute)
- Exercise: suppose $X_1, \ldots, X_n \sim_{i.i.d.} (\mu, \sigma^2)$, derive the MSE for the estimators $\widetilde{\mu}_n \equiv 0$ and $\widehat{\mu}_n = \overline{X}_n$. When is $\widetilde{\mu}_n$ preferable to $\widehat{\mu}_n$?

Blank page for notes

Hard- and soft-thresholding

- Consider our toy example $X_1, \ldots, X_n \sim_{i.i.d.} (\mu, \sigma^2)$
- Idea: reduce MSE by adaptively shrinking our estimator $\overline{X}_n$
  - Hard-thresholding $\widehat{\mu}_n^H = \overline{X}_n 1_{f(|\overline{X}_n|) \geq \tau}$ for some function $f : \mathbb{R}_+ \to \mathbb{R}_+$
  - Soft-thresholding $\widehat{\mu}_n^H = \overline{X}_n g(|\overline{X}_n|)$ for some function $g : \mathbb{R}_+ \to \mathbb{R}_+$

Soft-thresholding: try it at home!

- Consider an estimator of the form $\widetilde{\mu}_n = \alpha \overline{X}_n$, derive the value of α^{opt} that minimizes MSE over α

Hooray! A new estimator of the mean?

- $\alpha^{\text{opt}}\overline{X}_n$ resembles an empirical Bayes estimator of the mean [8]
  - Optimal shrinkage depends on signal-to-noise ratio
  - Common theme for adaptive shrinkage methods, i.e, adaptive ridge and adaptive lasso
- Not very exciting for a 1D example, but can be extremely useful/effective in higher dimensions
- Warning: hard-thresholding estimators are often non-regular which means we cannot estimate their sampling distributions uniformly $\Rightarrow$ standard inference procedures, i.e., bootstrap or series approximations, perform poorly

[8] you will derive the analog for hard-thresholding estimator $\overline{X}_n 1_{|\overline{X}_n| \geq \alpha}$ in HW2.

Quick aside: all subsets

- Before moving on to penalized regression, we should mention the obvious approach of looking at all models and choosing the 'best one' based on some criterion
- An intuitive approach is to examine all possible models
  - Choose model with lowest BIC/AIC, etc.
  - p variables $\Rightarrow 2^p$ possible models
  - Branch-and-bound algorithms can reduce the search space making all subsets search feasible for $p \leq 50$[9]
- R code example: allSubsets.R

[9] For a fun paper on this topic see Furnival, George M., and Robert W. Wilson. "Regressions by leaps and bounds." Technometrics 42.1 (2000): 69-79.

Roadmap

- Review and reminders
- All subsets (you don't want the truth)
- **Ridge regression**
- Lasso

People say sometimes that Beauty is superficial. That may be so. But at least it is not so superficial as Thought is. To me, Beauty is the wonder of wonders. It is only shallow people who do not judge by appearances. The true mystery of the world is the visible, not the invisible.
—Excerpt "What to Expect When You're Expecting."

Ridge regression: a superficial first look

- Least squares estimator is unbiased and has minimum variance among all unbiased estimators
  - As we've seen, this does not mean it minimizes MSE
  - Recall MSE = bias^2 + variance $\Rightarrow$ small increase in bias + big reduction in variance = smaller MSE
- Shrink regression coefficients toward zero by solving

$$\widehat{\beta}_n^{\lambda} = \arg\min_{\beta} \mathbb{P}_n(Y - \boldsymbol{X}^{\mathsf{T}})^2 + \lambda||\beta||^2,$$

where $\lambda \geq 0$ is a tuning parameter

Historical side-note

- Proposed by Hoerl and Kennard (1970)
  - Goal was to stabilize OLS estimator
  - Proved there always exists λ^* s.t.

$$\mathbb{E}||\widehat{\beta}_n^{\lambda^*} - \beta^*||^2 < \mathbb{E}||\widehat{\beta}_n - \beta^*||^2,$$

    but λ^* depends on β^*, suggested looking at 'ridge trace' and picking values at each coefficients appear to 'stabilize'[10] we'll look at other data-driven tuning methods

- Fact: Hoerl is a fun name to say, try it out

[10] the ridge trace plot is what we might call a solution path today (nothing is new)

Ridge regression: orthogonal case

- Ridge regression estimator can give dramatic reduction in MSE, especially when the dimension p is large
- Consider first the orthogonal case $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T} = I_p$

$$\widehat{\beta}^{\lambda}_{n,j} = \frac{\widehat{\beta}_{n,j}}{1+\lambda}$$

thus $\mathbb{E}(\widehat{\beta}^{\lambda}_{n}) = \beta^*_j/(1+\lambda)$ and variance is $\mathrm{Var}(\widehat{\beta}_{n,j})/(1+\lambda)^2$

- Exercise: suppose $p = 1$, find the value λ that minimizes $\mathrm{MSE}(\widehat{\beta}^{\lambda}_{n,1})$. What is the optimal for general p?

Blank page for notes

Ridge regression cont'd

- In the non-orthogonal case

$$\begin{aligned}
\widehat{\beta}_n^{\lambda} &= \left\{\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\intercal} + n^{-1}\lambda I_p\right\}^{-1} \mathbb{P}_n \boldsymbol{X} Y \\
&= \left\{I + n^{-1}\lambda \left(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\intercal}\right)^{-1}\right\}^{-1} \left(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\intercal}\right)^{-1} \mathbb{P}_n \boldsymbol{X} Y \\
&= \left\{I + n^{-1}\lambda \left(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^{\intercal}\right)^{-1}\right\}^{-1} \widehat{\beta}_n
\end{aligned}$$

- How to choose $\lambda \geq 0$ in this case?
  - Information criteria AIC, BIC, etc.
  - Cross-validation (generalized cross-validation)
  - You'll explore these criteria in lab

Tuning λ_n (one more criterion)

- Can use empirical Bayes (EB) to select λ_n
- EB in a nutshell
  - Posit Bayesian model
  - Marginal distribution to obtain frequentist estimators of hyper-parameters[11]
- Pro-tip: if you want to derive an estimator that performs well in practice, posit a Bayesian model, derive posterior mean, call this a frequentist estimator and hide all evidence you ever considered Bayes approach[12]

[11] This is a rich area but we don't have the bandwidth to cover it in depth in this class.

[12] Credit to Derek Bingham for this nugget of wisdom.

Tuning λ_n with EB

- Assume linear model correct $Y = \boldsymbol{X}^\mathsf{T}\beta + \epsilon$ where $\epsilon \sim \mathrm{Normal}(0, \sigma^2)$ and $\beta \sim \mathrm{Normal}(\mu, \tau^2)$ and $\beta \perp \epsilon$

- Posterior mean is

$$\widehat{\beta}_n = (\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T} + \lambda_n I)^{-1} \mathbb{P}_n(\boldsymbol{X}Y + \lambda_n \mu),$$

  where $\lambda_n = \sigma^2/(n\tau^2)$

- Setting $\mu = 0$ yields ridge estimator, given estimators $\widehat{\sigma}_n^2$ and $\widehat{\tau}_n^2$ yield a plug-in estimator of λ_n, i.e., $\widehat{\lambda}_n = \widehat{\sigma}_n^2/(n\widehat{\tau}_n^2)$

Tuning λ_n with EB cont'd

- Marginal distribution[13] $\mathbb{E}Y^2 = \tau^2 \mathrm{trace}(\boldsymbol{X}\boldsymbol{X}^T) + \sigma^2$
- MOM estimator matches

$$\mathbb{P}_n Y^2 \approx \tau^2 \mathrm{trace}(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T}) + \sigma^2$$

plugging in $\widehat{\sigma}_n^2 = \mathbb{P}_n(Y - \boldsymbol{X}^\mathsf{T}\widehat{\boldsymbol{\beta}}_n)^2$ yields

$$\widehat{\tau}_n^2 = \frac{\mathbb{P}_n Y^2 - \widehat{\sigma}_n^2}{\mathrm{trace}(\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T})}$$

and thus $\widehat{\lambda}_n = \widehat{\sigma}_n^2/(\widehat{\tau}_n^2 n)$

[13] To align with classic derivations, we're conditioning on $\boldsymbol{X}$ and treating it as fixed here. Which makes the expression $\mathbb{P}_n \boldsymbol{X}\boldsymbol{X}^\mathsf{T}$ a bit of an abuse of notation.

Tuning λ_n with EB more notes

- Note that $\widehat{\lambda}_n$ is a well-defined statistic without reference to the Bayes model
  - We can use this estimator and analyze it from a frequentist point-of-view
  - Bayesian connection is just icing
- Related idea: when we talk about RL, we might use a Bayesian framework for exploration-exploitation but not require these models to be correct when analyzing alg performance

Blank page for notes

Ridge regression: BIC

- Let $\mathbb{X}$ denote the design matrix, effective degrees of freedom

$$df(\lambda) = \text{trace}\left\{\mathbb{X}^\intercal\left(\mathbb{X}^\intercal\mathbb{X} + \lambda I\right)^{-1}\mathbb{X}\right\}$$

to compute this efficiently we use the SVD of $\mathbb{X}$

$$\mathbb{X} = UDV^\intercal \Rightarrow df(\lambda) = \sum_{j=1}^{p}\frac{d_j^2}{d_j^2+\lambda}$$

- BIC for each value of λ is

$$\text{BIC}(\lambda) = \log\left\{\text{RSS}(\lambda)\right\} + df(\lambda)\log(n)/n$$

choose $\widehat{\lambda}^{\text{BIC}} = \arg\min_{\lambda}\text{BIC}(\lambda)$

Asymptotic behavior

- Under what conditions on $\{\lambda_n\}_{n \geq 1}$ will:
  - $\widehat{\beta}_n^{\lambda_n} \to \beta^*$
  - $\sqrt{n}\left(\widehat{\beta}_n^{\lambda_n} - \beta^*\right)$ converge to a Gaussian limit
- Before we mathematize anything, what does your intuition say? Should λ_n shrink? Grow? Converge to a constant?

Asymptotic behavior notes

Ridge regression: prostate data example

- Example: prostate cancer data set
  - $n = 97$ patients, response prostate-specific antigen (`lspa`)
  - Predictors:
    - `lcavol`: log (cancer volume)
    - `lweight`: log(prostate weight)
    - `age`: age
    - `lbph`: log (benign prostatic hyperplasia)
    - `svi`: seminal vesicle invasion
    - `lcp`: log (capsular penetration)
    - `gleason`: Gleason score
    - `pgg45`: percent of Gleason scores 4 or 5
  - `R` code example: ridgeExample.R

Ridge regression and dropout

- Dropout is popular heuristic to reduce overfitting in machine learning models (esp. nnets)
  - Randomly 'zero-out' some components in the training input
  - Lots of heuristic explanations for why this works[14], with a linear model we can obtain a more rigorous answer
- Let $\boldsymbol{Z} \in \{0, 1, \}^p$ be a vector of independent Bernoulli random variables so that $P(Z_j = 0) = \phi$ (and $P(Z_j = 1) = 1 - \phi$))
  - Suppose $\boldsymbol{Z}$ is ind of $(\boldsymbol{X}, Y)$
  - Random dropout $\boldsymbol{X} \odot \boldsymbol{Z} = (X_1 Z_1, \ldots, X_p Z_p)^\intercal$

[14] Search reddit or quora for 'why dropout works'.

Ridge regression and dropout cont'd

- Imagine for each $i = 1, \ldots, n$ we generate a bajillion[15] values of $\boldsymbol{Z}$, say $\boldsymbol{Z}_i^1, \ldots, \boldsymbol{Z}_i^B$, and subsequently generate new dataset $\left\{ \left[Y_i, \boldsymbol{X}_i \odot \boldsymbol{Z}_i^k/(1-\phi) \right] k = 1, \ldots, B, i = 1, \ldots, n \right\}$

- Now suppose we fit a linear model using this data

$$\begin{aligned} \widetilde{\beta}_n &= \arg\min_{\beta} \sum_{k=1}^{B} \sum_{i=1}^{n} \left\{ Y_i - (\boldsymbol{X}_i \odot \boldsymbol{Z}_i^k)^\intercal \beta/(1-\phi) \right\}^2 \\ &\approx \arg\min_{\beta} \mathbb{E}_Z \mathbb{P}_n \left\{ Y - (\boldsymbol{X} \odot \boldsymbol{Z})^\intercal \beta/(1-\phi) \right\}^2, \end{aligned}$$

  where we've approximated the sample average with expectation (recall we generate as many $\boldsymbol{Z}$'s as we want)

[15] By a bajillion, I mean a lot. Also, this lecture is sponsored by "Bajillion Dollar Properties," watch now on Amazon Prime.

Ridge regression and dropout cont'd

- Suppose data has been centered and scaled derive a closed form expression for $\widetilde{\beta}_n$

Ridge regression and dropout discussion

- Reducing info available to the model to make a prediction at any given point ($\boldsymbol{X} \mapsto \boldsymbol{X} \odot \boldsymbol{Z}$) has a regularizing effect
  - General strategy can be used with models that are harder to penalize explicitly, e.g., this is used when building trees in random forests (we'll talk more about this later)
  - Smoother version of dropout can be obtained by replacing Bernoulli's with any unit-mean r.v.'s

Ridge regression and noise addition

- Another way to reduce info available to the model is to replace $\boldsymbol{X}$ with $\boldsymbol{X} + \boldsymbol{Z}$, w/ $\boldsymbol{Z}$ vector of independent $(0, \lambda)$ r.v.'s
- As with dropout, note that this is a general strategy that can be applied with any ML algorithm (for supervised learning)
- In-class exercise: compute

$$\widetilde{\beta} = \arg\min_{\beta} \mathbb{E}_{\boldsymbol{Z}} \mathbb{P}_n \{Y - (\boldsymbol{X} + \boldsymbol{Z})^\intercal \beta\}^2$$

Blank page for notes

Ridge regession: discussion

- Reduces MSE by shrinking regression coefficients, very effective in a wide range of settings (battle-tested)
- Does not perform variable selection, i.e., all variables are kept in the model, this can be a problem if parsimony is critical but we can always threshold small values to zero
- Choose amount of penalization using information criteria (BIC, AIC, etc.) or cross-validation

Principal components regression

- A closely related alternative to ridge regression is principal components regression
- Review: principal components[16]
  - Write $\mathbb{X}^\intercal \mathbb{X}/n = VD^2V^\intercal$ where $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p]$
  - $Z_j = \boldsymbol{X}^\intercal \boldsymbol{v}_j$ is the called the *jth* principal component of $\boldsymbol{X}$
  - $\mathrm{Var}(Z_1) \geq \mathrm{Var}(Z_2) \geq \ldots \geq \mathrm{Var}(Z_p)$
- Idea: regression Y on $Z_1, \ldots, Z_q$ for some $q \leq p$ instead of regressing on $\boldsymbol{X}$
  - Capture important features of $\boldsymbol{X}$
  - Reduce dimension $\Rightarrow$ bias-variance trade-off

[16] Assume predictors have been centered

Principal components regression cont'd

- Matrix of principal components is $\mathbb{Z} = \mathbb{X}V$
  - PR regression design matrix is orthogonal $\mathbb{Z}^\intercal\mathbb{Z} = nD^2$ (why?)
  - Compute regression of Y on $Z_1, \ldots, Z_q$ using series of univariate regressions
  - Tuning parameter q can be chosen using BIC/AIC, cross-validation, etc.
- `R` code example: pcr.R

Principal components regression cont'd

- Pros:
  - Reduce MSE relative to standard least squares
  - Sometimes interpret principal components (see pcr.R)
  - Parallel computation possible
- Cons:
  - Interpretation of principal components subjective
  - Doesn't involve Y in construction of features (pc's)

Roadmap

- Review and reminders
- All subsets (you don't want the truth)
- Ridge regression
- **Lasso**

A nice relaxing quiz

- Explain to your group
  - What does the acronym LASSO stand for?
  - What is the problem of inference after model selection?
  - What goes wrong in regression when $p \gg n$?
- True or false
  - Maximum likelihood is limited to parametric models
  - Large coefficient std errors can be sign of near collinearity
  - The world's most expensive Donkey cheese sells for several thousand dollars a pound

Lasso: superficial overview

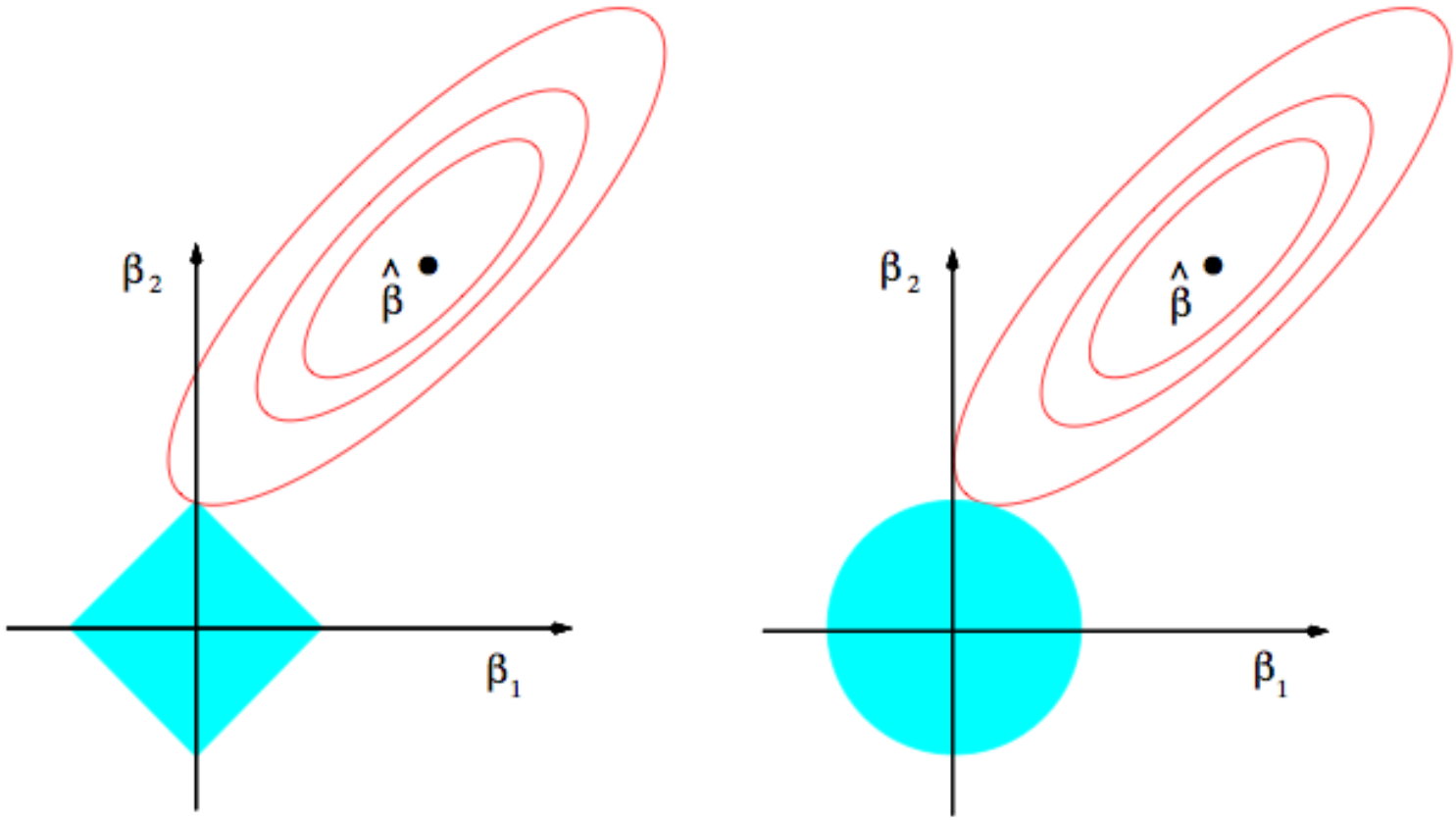
- Simultaneous estimation and model selection via penalization

$$\widehat{\beta}_n^{\tau} = \arg\min_{\beta} \mathbb{P}_n(Y - \boldsymbol{X}^{\intercal}\beta)^2 + \tau||\beta||_1,$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$ and $\tau > 0$ is a tuning parameter

- Looks like ridge but use of L1 norm yields sparse solutions

Lasso vs. ridge (a stolen picture)



Lasso: orthogonal predictors

- Suppose $\mathbb{X}^\intercal\mathbb{X} = I$ then

  $$\mathbb{P}_n(Y - \boldsymbol{X}^\intercal\beta)^2 + \tau||\beta||_1 = \mathbb{P}_n Y^2 - ||\widehat{\beta}_n||^2 + ||\widehat{\beta}_n - \beta||^2 + \tau||\beta||_1,$$

  we can re-write the lasso solution as

  $$\begin{aligned} \widehat{\beta}_n^\tau &= \arg\min_\beta ||\widehat{\beta}_n - \beta||^2 + \tau||\beta||_1 \\ &= \arg\min_\beta \sum_{j=1}^p \left\{ (\widehat{\beta}_{n,j} - \beta_j)^2 + \tau|\beta_j| \right\}, \end{aligned}$$

  we can look at each component separately

- `R` code example: lasso.R

Lasso: orthogonal predictors

- In orthogonal case we can compute $\widehat{\beta}^{\tau}_{n,j}$ explicitly

$$\widehat{\beta}^{\tau}_{n,j} = \mathrm{sgn}(\widehat{\beta}_{n,j})(|\widehat{\beta}_{n,j}| - \tau/2)_+,$$

  where $(u)_+ = \max(0, u)$

- Can give us further insight into lasso soln (back to lasso.R)

Computing the lasso solution

- Can reformulate lasso objective as quadratic program, however, faster iterative algorithms exist
- As with ridge, select tuning parameter using BIC/AIC, cross-validation, etc.
  - Problem: $\widehat{\beta}_n^\tau$ is not linear estimator, how to define degrees of freedom?
  - Approximate degrees of freedom is the number of nonzero components of $\widehat{\beta}_n^\tau$

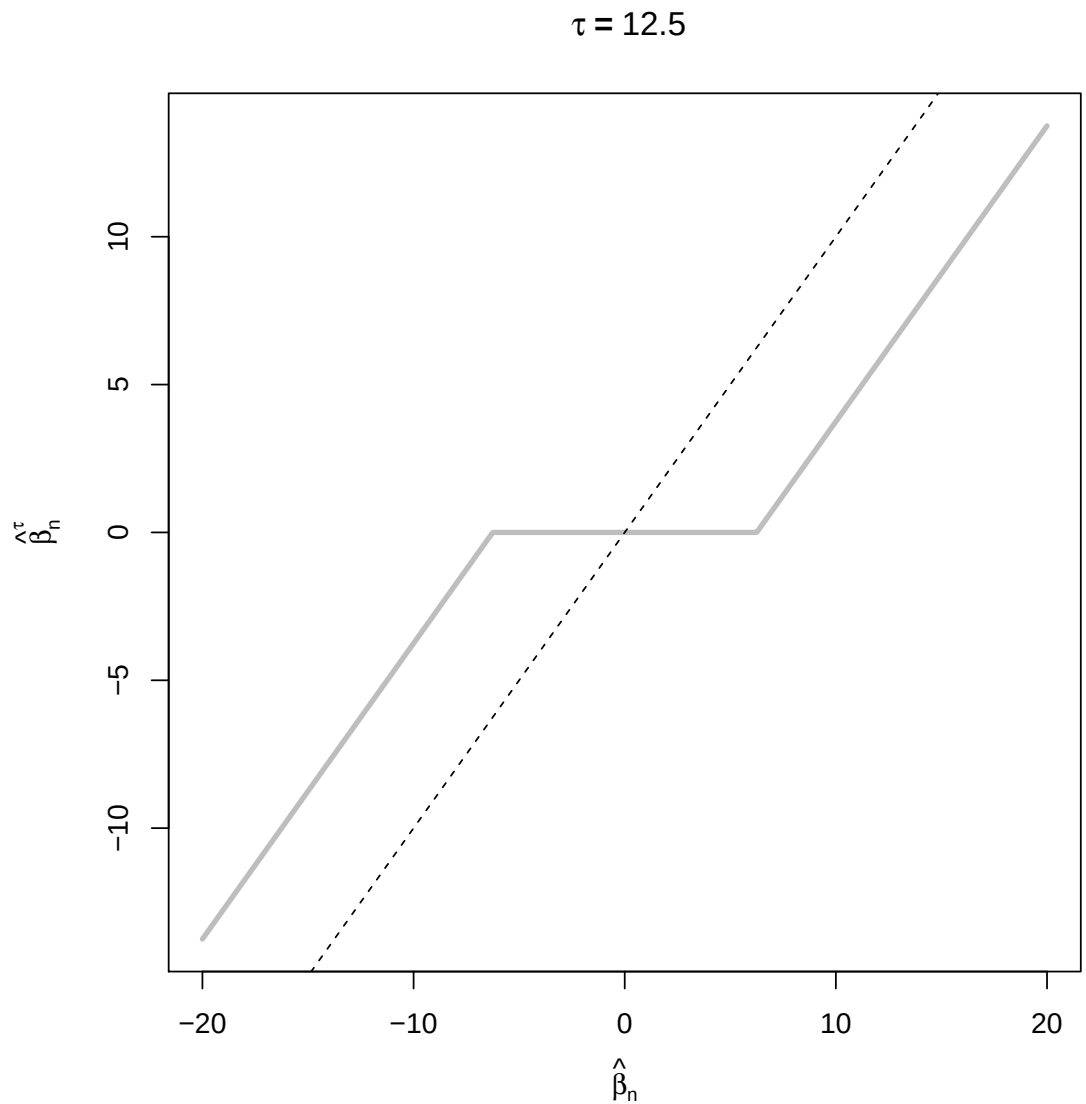
$$df(\tau) = \sum_{j=1}^{p} 1_{\widehat{\beta}_{n,j}^\tau \neq 0}$$

Lasso discussion

- Tool for simultaneous variable selection and estimation
  - Can be solved very quickly for large problems
  - Leads to parsimonious models $\Rightarrow$ iterpretable
  - Variants exist, i.e., for glm, cox-ph, etc. (more on this later)
- Potential problems
  - Shrinks all the coefficients, even those that are not close to zero, can lead to excessive bias
  - Select at most $\min(n, p)$ variables, problematic in some settings

Adaptive lasso

- Recall nature of shrinkage for lasso



Adaptive lasso cont'd

- Lasso shrinks all coefficients toward zero
  - Introduces excessive bias when true regression coef is large
  - Potentially better strategy is to shrink more aggressively when coefficients are small and less aggressively if they are large
  - How can we do this if true coefficients are unknown?
- Idea: use ordinary least squares estimates as surrogates

Adaptive lasso cont'd

- Adaptive lasso estimator

$$\widehat{\beta}_n^{\delta} = \arg\min_{\beta} \mathbb{P}_n(Y - \boldsymbol{X}^{\intercal}\beta)^2 + \delta \sum_{j=1}^{p} |\beta_j| / |\widehat{\beta}_{n,j}|$$

- If $|\widehat{\beta}_{n,j}|$ is small $\Rightarrow$ more shrinkage is applied
- Note* if p is large can use ridge estimator as weights

Adaptive lasso: orthogonal case

- In the orthogonal case the adaptive lasso estimator is

$$\widehat{\beta}^{\delta}_{n,j} = \operatorname{sgn}(\widehat{\beta}_{n,j}) \left(|\widehat{\beta}_{n,j}| - \frac{\delta}{2|\widehat{\beta}_{n,j}|} \right)_+$$

  for comparison

$$\begin{aligned} \text{Lasso:} \quad \widehat{\beta}^{\tau}_{n,j} &= \operatorname{sgn}(\widehat{\beta}_{n,j}) \left(|\widehat{\beta}_{n,j}| - \frac{\tau}{2} \right)_+ \\ \text{Ridge:} \quad \widehat{\beta}^{\lambda}_{n,j} &= \frac{\widehat{\beta}_{n,j}}{1+\lambda} \end{aligned}$$

- `R` code example: adaptiveLasso.R

Adaptive lasso: fitting non-orthogonal case

- Adaptive lasso objective can be recast as quadratic program
  - We can use existing lasso software by modifying design matrix
  - Write

$$\mathbb{P}_n(Y - \boldsymbol{X}^\intercal \beta)^2 + \delta \sum_{j=1}^{p} |\beta_j| / |\widehat{\beta}_{n,j}| = \mathbb{P}_n(Y - \widetilde{\boldsymbol{X}}^\intercal \widetilde{\beta})^2 + \delta \sum_{j=1}^{p} |\widetilde{\beta}_j|,$$

where $\widetilde{X}_j = \widehat{\beta}_{n,j} X_j$ and $\widetilde{\beta}_j = \beta_j / \widehat{\beta}_{n,j}$

Adaptive lasso: fitting non-orthogonal case cont'd

- Pseudo-code (assumes centered, scaled data)
  1. Fit ordinary least squares, say $\widehat{\beta}_n$
  2. Create scaled design matrix $\widetilde{\mathbb{X}} = \mathbb{X}\mathrm{diag}(\widehat{\beta}_n)$
  3. For each δ under consideration compute lasso estimator

$$\widetilde{\beta}_n^{\delta} = \arg\min_{\beta} \mathbb{P}_n(Y - \widetilde{X}^{\mathsf{T}}\beta)^2 + \delta||\beta||_1,$$

     compute adaptive lasso estimator $\widehat{\beta}_{n,j}^{\delta} = \widehat{\beta}_{n,j}\widetilde{\beta}_{n,j}^{\delta}$
  4. Select δ using BIC/AIC, etc.

Adaptive lasso discussion

- Reduce excessive bias due to overshrinkage of large coefficients
  - An alternative is to refit the model selected by lasso using least squares, this generally works well in practice
- Can use existing statistical software, generally much more computationally efficient than alternative proposals (SCAD, etc.)

Robustness and the lasso

- Let $c_1, \ldots, c_p$ be positive constants and define

$$\mathcal{M} = \left\{ \boldsymbol{\Delta} \in \mathbb{R}^{n \times p} \, : \, \sqrt{\sum_{i=1}^{n} \Delta_{i,j}^2} \leq c_j \, , \, j = 1, \ldots, p \right\},$$

  i.e., the set of $n \times p$ matrices where norm of jth column bounded by c_j for $j = 1, \ldots, p$

- Robust lin regression estimator under perturbations $\mathcal{M}$ is

$$\widetilde{\beta}_n = \arg\min_{\beta} \left\{ \max_{\boldsymbol{\Delta} \in \mathcal{M}} ||\mathbb{Y} - (\mathbb{X} + \boldsymbol{\Delta})\beta||_2 \right\}$$

Robustness and lasso equivalence thm

- Claim: lasso and robust regression soln are equivalent in that

$$\begin{aligned}\widetilde{\beta}_n &= \arg\min_{\beta}\left\{\max_{\mathbf{\Delta}\in\mathcal{M}}||\mathbb{Y}-(\mathbb{X}+\mathbf{\Delta})\beta||_2\right\} \\ &= \arg\min_{\beta}\left\{||\mathbb{Y}-\mathbb{X}\beta||_2+\sum_{j=1}^{p}c_j|\beta_j|\right\}\end{aligned}$$

Blank page for notes

Robustness and lasso discussion

- Showed equivalence between so-called square-root lasso and robust regression
  - Set $c_j \equiv \tau$ yields standard (square-root) lasso
  - Setting $c_j \equiv 1/|\widehat{\beta}_{n,j}|$ yields adaptive (square-root) lasso
- The square-root lasso has some desirable properties in terms of tuning (see papers by A. Belloni here at Duke), however, the solution paths are the same

Penalization and regularization discussion

- Bias-variance trade-off $\Rightarrow$ regularization needed to improve predictive performance
- Ridge regression smoothly (soft) penalizes coefficients but can combine with thresholding to obtain sparse solutions
- Lasso automatically yields sparse solutions but may perform poorly if true signal is dense and comprised of weak signals
- Showed that several information deletion/distortion methods are equivalent to penalized regression $\Rightarrow$ more general strategy for regularizing complex estimators

Thank you.

eric.laber@duke.edu

laber-labs.com