Regularized Regression

COS 424/524, SML 302: Fundamentals of Machine Learning

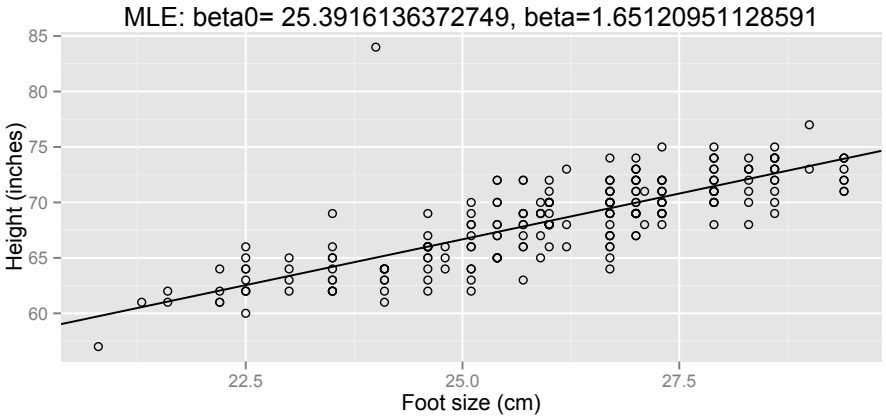
Professor Engelhardt

COS424/524, SML 302

Lecture 8

Regression

In the last lecture, we learned about linear regression



$$y_i = \beta_0 + \beta x_i + \epsilon_i$$

In this lecture, we will discuss *regularized regression*

What is regularization?

We have discussed finding maximum likelihood estimates (MLE) of parameters to regression models; recall the normal equation. MLE estimates are *unbiased*.

MLE approaches often have problems on real data, as we saw:

- sensitive to outliers;
- overfit data (i.e., fits data noise);
- cannot solve mathematically underconstrained problems (e.g., $p \gg n$).

Idea behind *regularization*: reduce the complexity or magnitude of parameter estimates.

Regularization history

Regularization has a long history in statistics:

- Occham's Razor
- Bayesian statistics
- Regularizing ill-posed regression problems *[Tikhonov 1943]*
- Stein's phenomenon, and the James-Stein estimator *[Stein 1956]*: biased estimator has lower mean squared error on average than MLE
- LASSO regression *[Tibshirani 1996]*

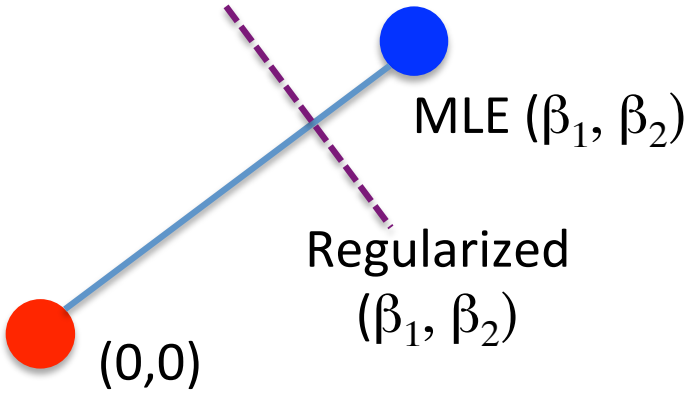
Why do we use regularization?

Regularization has big advantages, if done properly:

- *Outliers*: we can protect parameter estimates from effects of outliers.
- *Interpretability*: parameter estimates can be used to determine when a specific predictor is not useful.
- *Stability*: even with a large number of predictors, we get a robust and stable solution.
- *Generalization*: regularized parameter estimates avoid overfitting.
- *Adding domain-specific information*: background knowledge about system may be included.

Regularization: intuition

Mathematically, regularization is often constructed as the weighted sum of the MLE parameters and a fixed point.

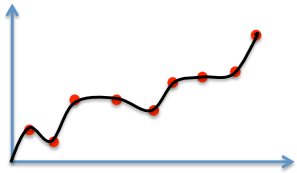


For this reason, regularization is often referred to as *statistical shrinkage*.

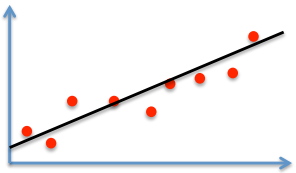
Regularization: intuition

Regularization can be interpreted as *adding bias* to the parameter estimates.

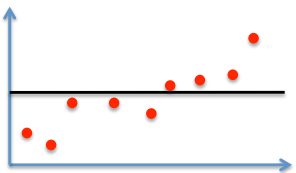
Recall the bias-variance tradeoff. For an arbitrarily complex model:



Large variance



Some bias, variance



Large bias

Regularization: practically

There are a number of ways to practically perform regularization.

Bayesian priors

One approach to regularization is to use a Bayesian framework, and put a prior distribution on the parameters; recall:

$$p(\theta \mid \mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta).$$

Estimates of parameter θ according to posterior probability (*maximum a posterior*, or MAP, estimates) will be proportional to the sum of the maximum likelihood estimates and the prior distribution (in log space).

Let's look at an example for regression.

Simplest form of regularized regression: ridge regression

Likelihood for multivariate linear regression is a conditional probability:

$$y_i \mid \mathbf{x_i}, \beta, \sigma^2 \sim \mathcal{N}(\mathbf{x_i^T}\beta, \sigma^2).$$

A Gaussian prior for our coefficients specifies *ridge regression*:

$$p(\beta|\mu, \tau^2) = \prod_{\mathbf{j=1}}^{\mathbf{p}} \mathcal{N}(\beta_\mathbf{j}|\mathbf{0}, \tau^2)$$

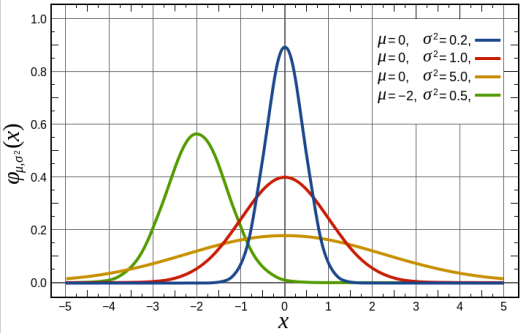


Figure: `wikipedia`

Ridge regression

We can write the log posterior distribution using Bayes rule:

$$
\begin{aligned}
\log p(\beta|\mathcal{D}) &\propto \log\left(p(\mathcal{D} \mid \beta)\mathbf{p}(\beta)\right) \\
&= \log\left[\left(\prod_{i=1}^{n}\mathcal{N}(y_i \mid x_i^T\beta, \sigma^2)\right)\prod_{j=1}^{p}\mathcal{N}(\beta_j \mid \mathbf{0}, \tau^2)\right] \\
&= \log\left[\left(\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}(y_i-\beta^{\mathbf{T}}\mathbf{x_i})^2\right\}\right)\prod_{j=1}^{p}\mathcal{N}\left(\beta_j|0,\tau^2\right)\right] \\
&= \sum_{i=1}^{n}\log\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{\frac{-1}{2\sigma^2}(y_i-\beta^{\mathbf{T}}\mathbf{x_i})^2\right\}\right]+\sum_{j=1}^{p}\log\mathcal{N}\left(\beta_j|0,\tau^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(y_i-\beta^{\mathbf{T}}\mathbf{x_i}\right)^2+\frac{\sigma^2}{\tau^2}\sum_{p=1}^{j}\beta_j^2
\end{aligned}
$$

Now we can find the value of β that maximizes the posterior probability.

Ridge regression parameter updates

The solution for $\hat{\beta}_{MAP}$ is referred to as *ridge regression*:

$$\hat{\beta}_{MAP} = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \beta^{\mathbf{T}} \mathbf{x_i} \right)^2 + \frac{\sigma^2}{\tau^2} \sum_{p=1}^{j} \beta_j^2$$

We compute the MAP estimate for β using the same strategy as we used to find the MLE (letting $\lambda = \frac{\sigma^2}{\tau^2}$):

MAP estimates for regression coefficients

$$\hat{\beta}_{\text{MAP}} = (X^T X + \lambda I_p)^{-1} X^T \mathbf{y}$$

Compare to normal equation: $\hat{\beta}_{\text{MLE}} = (X^T X)^{-1} X^T \mathbf{y}$.

Why does ridge regression work?

Notice here that $(X^T X + \lambda I_p)^{-1}$ exists as the matrix $(X^T X + \lambda I_p)$ is generally invertible (i.e., non-singular). Why?

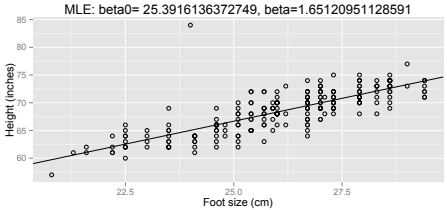
Aside: Mathematical trick to making matrices full rank

A simple way to make a singular matrix invertible is to add a very small value to the diagonal.

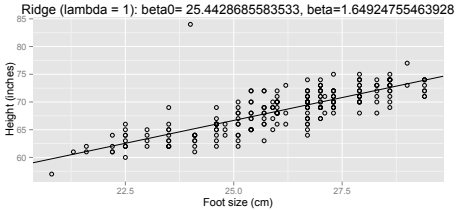
The small diagonal terms (artificially) create a full rank (and hence invertible) matrix.

Should we regularize the regression intercept term, or just slope?

Ridge regression versus standard regression

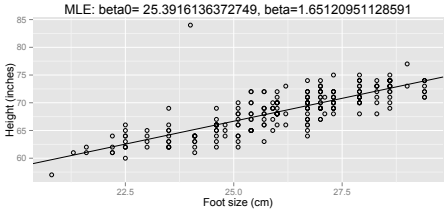


Estimate of regression parameters using normal equation (MLE).

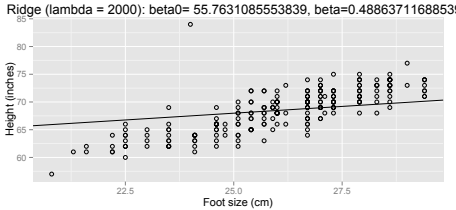


Ridge regression ($\lambda = 1$, corresponding to diffuse Gaussian distribution).

Ridge regression versus standard regression



Estimate of regression parameters using normal equation (MLE).

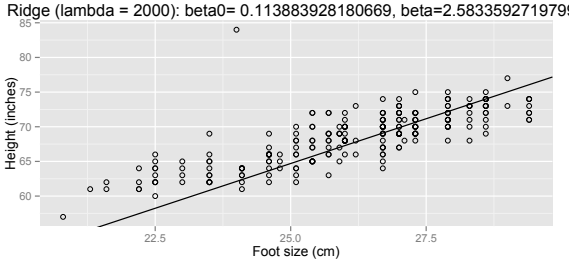


Ridge regression ($\lambda = 2000$, peaked Gaussian distribution).

Ridge regression: a note about the intercept

Do you regularize the intercept β_0?

- The ridge regression framework naturally regularizes the intercept.
- That said, *it is generally unwise to regularize the intercept*.
- The intercept term is sensitive to data translations
- Pushing intercept toward zero has undesirable effect on coefficients.



Ridge regression ($\lambda = 2000$, peaked Gaussian distribution).

Sparse regularization

One type of regularization is *sparse regularization*:

- recall that the number of parameters β scale with the number of features (predictors) p:

$$\mathbf{y} = \mathbf{X}\beta + \epsilon$$

- *sparse regularization* or *sparsity* shrinks parameters so as to remove subsets of features from the model;
- in regression, this means that regression coefficients β_j are zero.

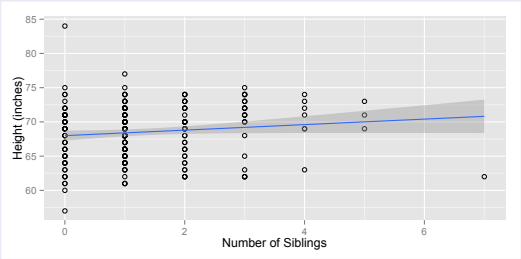
Sparsity

- Sparsity has been a "hot topic" in statistics and machine learning since the LASSO paper *[Tibshirani 1996]*.
- Sparsity is currently a very active area of research
- Although the tools I present here are motivated by (and presented for) regression models, sparsity can be included in any model: classifiers, factor analysis, mixture models, etc.
- Sparsity often presented as *feature selection* or *model selection*.

Advantages of sparsity in regression

1. **Prevents over-fitting.**
   - Values of β close to zero often represent sample-specific noise that is being modeled, instead of true signal.
   - If irrelevant features are removed, we avoid fitting noise.

Siblings predictive of height example



How can we tell whether or not number of siblings is predictive of height?

Advantages of sparsity in regression

2. **Fewer parameters to estimate.**

- If you have p features (predictors) and p is very large with respect to n samples, problem is underconstrained
- Sparsity reduces p so that the effective dimensionality is small
- Sparsity immediately allows gains in making the problem solvable
- Gains in computational speed up have not fully been realized

Advantages of sparsity in regression

3. **Interpretability.**
   - Having a sparse model makes interpreting the underlying natural phenomenon easier.

Doctor diagnosing a patient

- You are a doctor and would like to determine whether your patient has heart disease
- You have hundreds of clinical predictors in patient's medical charts
- Science does not fully understand the mechanisms of heart disease
- How can you make a precise prediction using a few predictors?

Types of sparsity

We will discuss three approaches to (sparse) regularization:

- Classical (frequentist) statistics: penalize parameters with non-zero values
- Bayesian statistics: put a prior distribution on parameters to encourage zero values
- Greedy approaches: iteratively add parameters to the model, starting from zero

Multivariate linear regression

Recall the general model of multivariate regression, with variables $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $\beta \in \mathbb{R}^{\mathbf{p}}$ (**sparse** vector), $\epsilon \in \mathbb{R}^n$.

We write a linear model with p features and n samples as:

$$y = X\beta + \epsilon.$$

Sparse linear models

Recall that, in this linear regression model,

- $\mathbf{y}$ are the n response terms
- X are the $n \times p$ observed predictors or covariates
- β coefficients (unobserved), weight each of p features in X depending how well feature j can be used to predict Y
- ϵ describes additive Gaussian noise, $\epsilon \in \Re^{n \times p}$, where $\epsilon_{i,j} \sim \mathcal{N}\left(0, \sigma^2\right)$.

We can put a sparsity-inducing prior on the coefficients β that will allow us to choose a subset of the given features for our prediction model.

Sparsity, ideally

If we were to idealize the correct model for sparse regression, we might envision Bayesian variable selection.

We may add indicator variables $\gamma_j, j = 1, \ldots, p$, which are zero when a predictor is "out", and one when a predictor is "in"

Specifically,

$$\gamma_j = \begin{cases} 1, & \text{feature } j \text{ is in} \\ 0, & \text{feature } j \text{ is out} \end{cases}.$$

Bayesian variable selection

There are a few drawbacks with this approach:

1. Computational: $\gamma \in \{0,1\}^p$, we will need to search a set of size 2^p, which grows exponentially with p.
2. Non-sparse point estimates: The MAP estimate of γ will (likely) not have zeros
3. Unstable point estimates: A single point estimate of γ does not always represent the full posterior distribution well; unstable

Example: Well-correlated covariates

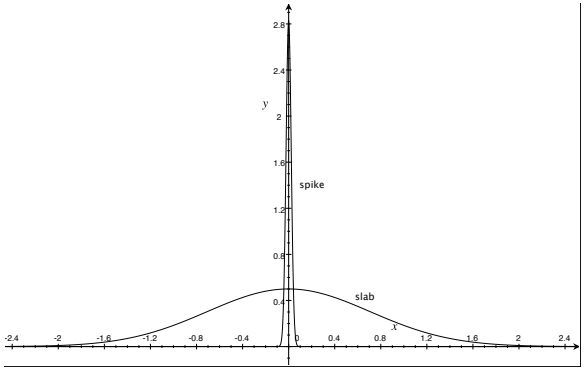
When point estimate of γ maximizes posterior probability, including different correlated covariates may not affect the posterior much.

Spike and Slab distribution

By Bayes rule, the posterior distribution may be written as the prior times the likelihood:

$$p(\gamma|\mathcal{D}) \propto p(\gamma)p(\mathcal{D}|\gamma).$$

One way to induce sparsity is to use a "spike and slab" prior distribution



Spike and Slab distribution

$\beta \in \mathbb{R}^p$ depend on $\gamma \in \mathbb{R}^p$, assigns class to each predictor j

- If $\gamma_j = 0$ then $\beta_j = 0$ (noise)
- If $\gamma_j = 1$, then β_j is drawn from a Gaussian distribution (signal)

- Each γ_j is one of p i.i.d. draws from Bernoulli with parameter π_0, i.e.,

$$p(\gamma_j|\pi_0) \quad = \quad \pi_0^{\mathbb{1}(\gamma=1)}(1-\pi_0)^{\mathbb{1}(\gamma=0)},$$

- Then the spike and slab prior can be written as:

$$\beta_j \sim \pi_0 \mathcal{N}(0, \sigma^2) + (1-\pi_0)\delta(0).$$

Spike and slab prior regression: example

Let's predict height from all of our class data.

Response variable y = height

Predictor variables X = gender, shoe_size, month, digit, sleep, siblings, handed, thumb, registered, The_Imitation_Game, Pulp_Fiction, Gone_Girl, Avatar, Matrix, Frozen, Silver_Linings_Playbook, The_Hunger_Games, Slumdog_Millionaire, The_Princess_Bride, Monty_Python_and_the_Holy_Grail, Ferris_Buellers_Day_Off, Love_Actually, Fight_Club, Shawshank_Redemption, The_Social_Network, The_Dark_Knight, Gravity, 500_Days_of_Summer

We would like to determine which of these predictors are most predictive of height.

Spike and slab prior regression: example

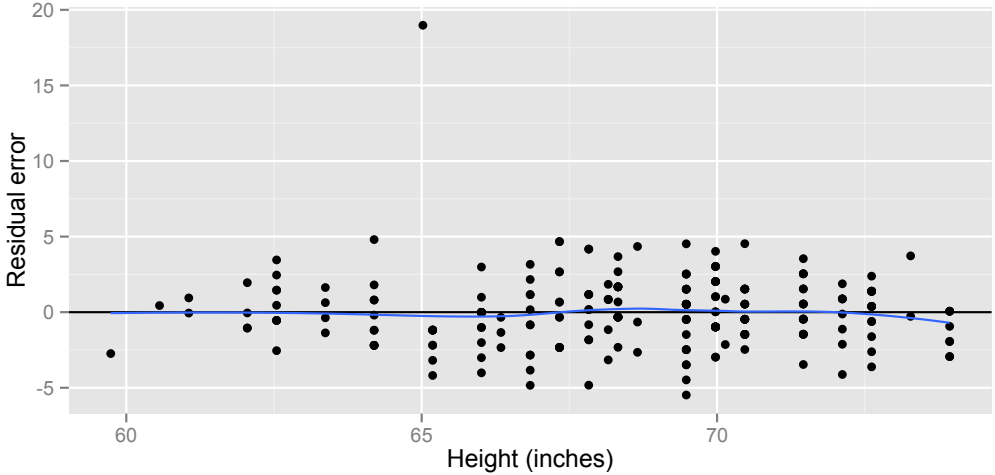
Predictor	slope
shoe_size	2.176
gender	-1.060
Gone_Girl	-0.625
The_Social_Network	0.538
X500_Days_of_Summer	0.455
Pulp_Fiction	0.367
month	-0.323
Love_Actually	0.268
Monty_Python_and_the_Holy_Grail	-0.206
digit	0.178
handed	0.094
Avatar	-0.063

What does it mean to have multiple predictors in the regression?

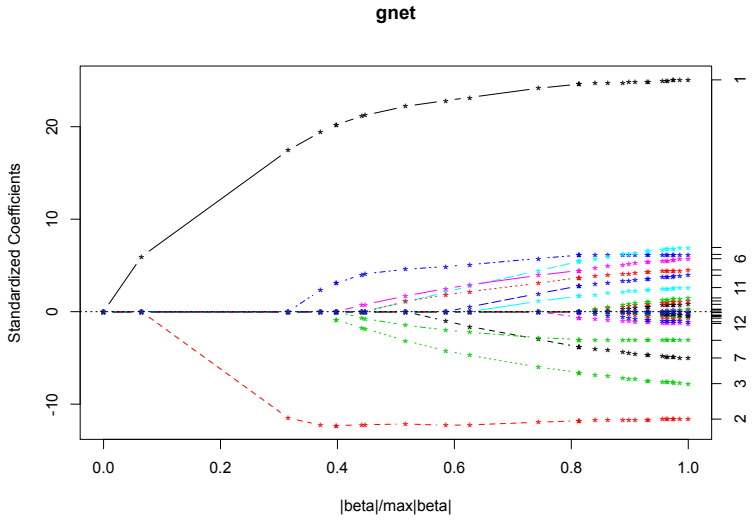
Consider residual:

$$\mathbf{y_{height}} - \beta_0 - \beta_{\mathbf{shoe\_size}}\mathbf{x_{shoe\_size}}.$$

Spike and slab prior regression: example



Spike and slab prior regression: example



Greedy approaches to sparse regression

Greedy methods can solve optimization problems with sparsity and many local optima.

Greedy methods are simple: decompose a multi-parameter problem into a series of single-parameter subproblems (i.e., the marginal problems)

Marginal subproblems do not have the associated combinatorial explosion of the joint problem (p versus 2^p)

Detour: model selection

What does it mean to be the 'best' model?

Model scores are statistics by which we can compare different models.

Intuitively, model score consists of:

- likelihood of generating our current data set $\mathcal{D}$ given a specific model

More parameters and likelihood

- What happens to the likelihood $p(\mathcal{D}|\Theta)$ when we add additional parameters?
- likelihood will improve because we check the likelihood of data used to fit the parameters.
- More parameters will remove bias and add variance, leading to overfitting

- penalty term for additional parameters

Model selection: BIC

Bayesian Information Criteria (BIC) for model selection.

The goal is to minimize:

$$\text{BIC} = -2 \log p(\mathcal{D}|\hat{\theta}_{ML}, M) + K \log(n),$$

where

- K = number of parameters to estimate (e.g., coefficients)
- n = sample size
- $\hat{\theta}_{ML}$ is the maximum likelihood parameter estimates,
- M refers to a model.

Smaller values of the BIC indicate superior models.

Model selection: BIC

BIC definition:

$$\mathrm{BIC} = -2 \log p(\mathcal{D}|\hat{\theta}_{ML}, M) + K \log(n),$$

- The first term of the BIC score quantifies the likelihood.
- The second term is a penalty for the number of parameters (K) scaled by log sample size (n): additional samples increase the relative penalty of each additional parameter.
- When K is equal for two models M_1 and M_2, the BIC is a function of data likelihood.

Greedy methods for model selection

- Greedy methods rely on a measure of comparison between models.
- We may use model scores including
  - Akaike information criterion (AIC)
  - Bayesian Information Criterion (BIC)
  - Minimum Description Length (MDL).
- The model parameters are the regression intercept and coefficients of the included covariates (but not the excluded covariates).

Forward Selection

(Also: Forward Stepwise Search or Forward Stepwise Regression (FSR)).

Forward Stepwise Regression algorithm

- Start with $\gamma^0 = [0]'$ vector
- Until model score is stable:
  - Consider all possible single bit changes of $\gamma^0 = [0]'$
  - Include the covariate that improves the score maximally.
  - Stop when no single bit change improves the score.

- Example: compare BIC score for: $\gamma = [110]'$ to score for $\gamma = [111]'$
- if $\gamma = [111]'$ produces best score, add third covariate.

Forward Reverse Selection

Forward Reverse Stepwise Regression algorithm

- Start with $\gamma^0 = [0]'$ vector
- Until model score is stable:
  - Consider all possible additions of variables to γ
  - Include the covariate that improves the score maximally.
  - Stop when no single bit change improves the score.
- Until model score is stable:
  - Consider all possible eliminations of predictors from γ
  - Exclude the covariate that improves the score maximally.
  - Stop when no single bit change improves the score.

Forward stepwise regression: example

Predictor	BIC
shoe_size	611.0849
gender	588.9522
The_Social_Network	580.5625
Gone_Girl	568.8447
Love_Actually	567.3306
Pulp_Fiction	564.4156
Monty_Python_and_the_Holy_Grail	564.1016
500_Days_of_Summer	564.0084
month	563.2116

Forward and reverse stepwise regression: example

Predictor	BIC
shoe_size	611.0849
gender	588.9522
The_Social_Network	580.5625
Gone_Girl	568.8447
Love_Actually	567.3306
Pulp_Fiction	564.4156
Monty_Python_and_the_Holy_Grail	564.1016
500_Days_of_Summer	564.0084
month	563.2116

When we perform a reverse step, we do not lose any predictors.

Single best replacement (SBR)

Single best replacement (SBR) algorithm

- Start with $\gamma^0 = [0]'$ vector
- Until model score is stable:
  - Consider all possible single bit changes of $\gamma^0 = [0]'$
  - Include or exclude the covariate that improves the score maximally.
  - Stop when no single bit change improves the score.

- Example: compare BIC score for: $\gamma = [110]'$ to score for γs $[010]'$, $[100]'$, and $[111]'$
- if $\gamma = [010]'$ produces best score, remove first covariate.

Orthogonal Matching Pursuit (OMP)

Idea: a covariate that is correlated with the residual should explain remaining lack-of-fit.

Orthogonal matching pursuit (OMP)

- Start with $\gamma^0 = [0]'$ vector
- Until model score is stable:
  - Pick next covariate to include j^* in the current model by finding the one most correlated with residual

$$\begin{aligned} j^* &= \arg\min_{j \neq \gamma_t} \left[\min_{\beta_t} \|y - \underbrace{X\beta_t}_{\hat{y}_t} - \beta_j X_j\|^2 \right] \\ &= \arg\max_j X_j^T r_t \end{aligned}$$

  - Stop when no new covariate improves the score.

Lasso Regression

Ideal approach to sparsity is to find parameters that maximize the likelihood with an ℓ_0 norm (number of non-zero elements) penalty for coefficients β: penalize the number of covariates.

In practice, ℓ_1 norm is used as an approximation of the ℓ_0 norm, making computation easier.

What is a norm? What is the optimization problem we are solving?

Norms

We have discussed the ℓ_0 norm, which is the number of non-zero elements:

$$||\beta||_0 = \sum_{j=1}^{p} \mathbb{1}(|\beta_j| > 0).$$

ℓ_0 norm can be approximated by the ℓ_1 norm, the absolute distance of each element from zero:

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|.$$

Ridge regression can be framed as the ℓ_2 norm, the Euclidean distance of each element to zero:

$$||\beta||_2 = \left(\sum_{j=1}^{p} \beta_j^2 \right)^{\frac{1}{2}}$$

Other norms

The ℓ_∞ norm, or *max norm*, penalizes based on element with the largest absolute value:

$$||\beta||_\infty = \max(|\beta_1|, |\beta_2|, \ldots, |\beta_p|).$$

Generally, the ℓ_q norm is:

$$||\beta||_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{\frac{1}{q}}$$

Penalizing the residual sum of squares

The residual sum of squares (unregularized regression):

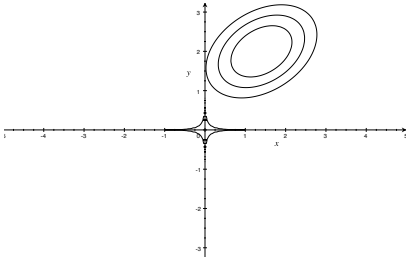
$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \right\}$$

ℓ_0 regression as penalized regression

Ideal sparse regression adds ℓ_0 penalty to the minimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \mathbb{1}(|\beta_j| > 0) \right\}$$

This equation penalizes specific values of β that are non-zero by adding the number of β that are non-zero to the RSS for that β. Contour plot for this optimization problem:

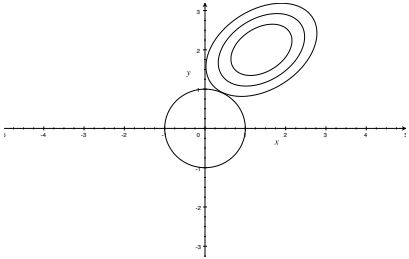


Ridge regression as penalized regression

Ridge regression adds an ℓ_2 penalty to this minimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \right\}$$

This equation penalizes specific values of β that are far from zero by adding the Euclidean distance of β from zero to the RSS for that β. Contour plot for this optimization problem:

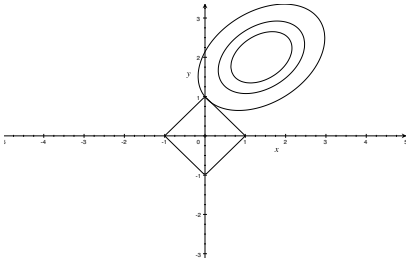


Lasso regression as penalized regression

Lasso regression adds an ℓ_1 penalty to this minimization problem:

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

This equation penalizes specific values of β that are far from zero by adding the absolute distance of β from zero to the RSS for that β. Contour plot for this optimization problem:

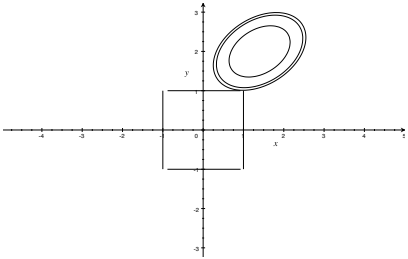


Regularized regression as penalized regression

We may penalize with the ℓ_∞ penalty:

$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} \max(|\beta_j|) \right\}$$

This equation penalizes the largest (in magnitude) value of β by adding this term to the RSS for that β. Contour plot for this optimization problem:



Lasso regression

Let's summarize Lasso regression

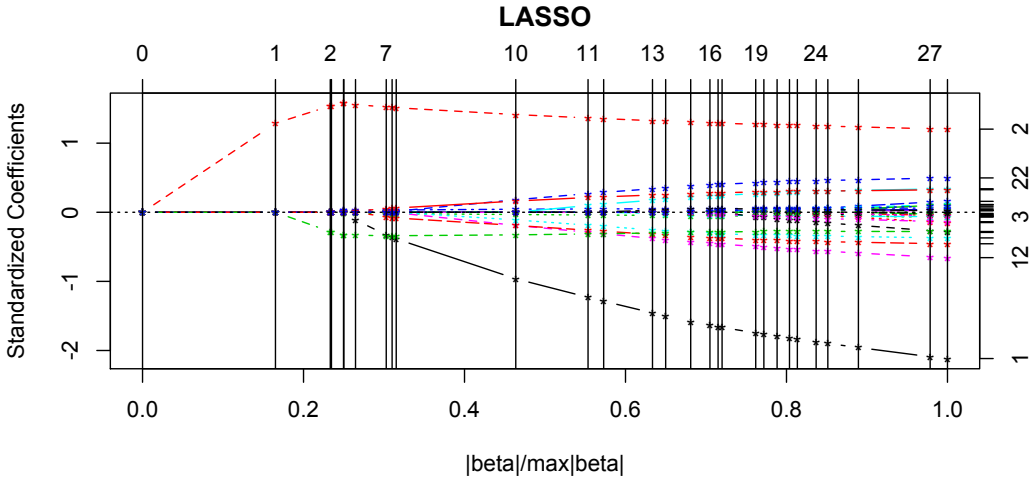
$$\hat{\beta} = \arg\min_{\beta \in \mathcal{B}} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

- parameter λ dictates sparsity; higher λ creates a sparser solution
- the choice of λ depends on data and analysis problem
- this optimization problem is called Lasso *[Tibshirani 1996]*
- a fast, exact solution to this problem across all λ values is called least angle regression (LARS) *[Efron et al. 2004]*
- we may choose a λ using cross validation, BIC

Lasso regression: example

Predictor	Order
shoe_size	1
digit	2
The_Social_Network	3
month	4
Pulp_Fiction	5
Gone_Girl	6
500_Days_of_Summer	7
gender	8
Monty_Python_and_the_Holy_Grail	9
Love_Actually	10
Avatar	11
handed	12
Fight_Club	13
siblings	14
sleep	15
Slumdog_Millionaire	16
The_Imitation_Game	17

Lasso regression: example



Summary: regularized regression

- There are many benefits to regularizing regression:
  - control outliers
  - robust solutions
  - solution to underdetermined problems
  - interpretability
- Three basic ways to regularize: Bayesian priors, greedy approaches, or penalty terms
- Each approach, and associated priors, parameters, norms, etc., produce different results
- Useful to select the approach most appropriate for your analysis problem
- Avoid overinterpreting the *included* predictors without cross validation, bootstrapping

Additional concepts

There are many related concepts for the interested student:

- Extensions of Lasso include group Lasso and fused Lasso
- Model averaging or stability selection
- Other Bayesian one-group priors that promote sparsity but are computationally tractable
- Extending sparsity to other models, e.g., factor analysis

Additional Resources

- MLAPA Chapter 13
- *Elements of Statistical Learning* Chapter 7, 10

- Lasso and LARS papers: *[Tibshirani 1996] [Efron et al. 2004]*
- Metacademy: *LASSO*
- Metacademy: *regularization*