Modern Variable Selection

Dominic D LaRoche

September 28, 2015

Outline

- What are “Modern” methods? How do they differ?
- Penalization Methods (regularizations)
- Model Selection / Averaging

Why Modern Variable Selection

Traditional (generally p-value based) methods have a number of shortcomings.

- Traditional Selection (backward, forward, and best subset selection):
  - Suffer from α inflation
  - Do not address problem of correlated predictors
  - Difficult to decide the appropriate number of predictors
  - Do not have very good out-of-sample performance
  - Cannot handle a large number of predictors
  - See: Breiman (1996) **Heuristics of instability and stabilization in model selection.** *Annals of Statistics* 24 (6)
  - Also: *The Elements of Statistical Learning* by Hastie, Tibshirani, and Friedman

Modern Methods

Two main approaches:

1. Penalization:
    - Shrink the all the coefficients simultaneously
    - Anything that isn't zero is important

Modern Methods

Two main approaches:

1. Penalization:
    - Shrink the all the coefficients simultaneously
    - Anything that isn't zero is important
2. Model selection using information criterion:
    - Start with a set of *plausible* models
    - Determine which model is most likely *given the data*
    - Includes a penalty for adding complexity
    - Can also average accross models*

Modern Methods

Two main approaches:

1. Penalization:
   - Shrink the all the coefficients simultaneously
   - Anything that isn't zero is important
2. Model selection using information criterion:
   - Start with a set of *plausible* models
   - Determine which model is most likely *given the data*
   - Includes a penalty for adding complexity
   - Can also average accross models*

*There is some current debate about the validity of variable importance from model averaging

Modern Methods

Modern methods for selection avoid many of the problems of tradtional methods.

- Avoid the multiple testing problem
- Deal better with correlated predictors
- Can accomodate very high dimensional data (large number of predictors)*

Modern Methods

Modern methods for selection avoid many of the problems of tradtional methods.

- Avoid the multiple testing problem
- Deal better with correlated predictors
- Can accomodate very high dimensional data (large number of predictors)*

However,

- Still don't know the correct number of predictors
- Penalization methods introduce a new (unknown) parameter λ

*Only penalization methods

Penalization Methods

Many methods rely on the same basic principal.

- l_0 regularization
  - Ridge regression (Hoerl and Kennard, 1970))

Penalization Methods

Many methods rely on the same basic principal.

- l_0 regularization
  - Ridge regression (Hoerl and Kennard, 1970))
- l_1 regularization
  - LASSO (Tibshirani 1996)
  - LARS (Efron et al. 2004)
  - Elastic net (Zou and Hastie 2005)

Penalization Methods

Many methods rely on the same basic principal.

- l_0 regularization
  - Ridge regression (Hoerl and Kennard, 1970))
- l_1 regularization
  - LASSO (Tibshirani 1996)
  - LARS (Efron et al. 2004)
  - Elastic net (Zou and Hastie 2005)
- Things I don't know about (among others)
  - SCAD
  - Non-negative garotte
  - COSSO

LASSO

LASSO works through the application of a penalty parameter (λ) to the fitted coefficients (β).
The l_1 regularization:

$$\min_{\beta} \mathit{Loss}\left(\beta; \mathbf{y}, \mathbf{X}\right) + \lambda J\left(\beta\right)$$

LASSO

LASSO works through the application of a penalty parameter (λ) to the fitted coefficients (β).
The l_1 regularization:

$$\min_{\beta} Loss\left(\beta; \mathbf{y}, \mathbf{X}\right) + \lambda J\left(\beta\right)$$

- The framework is generalizable through the loss function so that it can be applied to linear regression, logistic regession, Cox proportional hazards models, etc.

LASSO

LASSO works through the application of a penalty parameter (λ) to the fitted coefficients (β).
The l_1 regularization:

$$\min_{\beta} Loss\left(\beta; \mathbf{y}, \mathbf{X}\right) + \lambda J\left(\beta\right)$$

- The framework is generalizable through the loss function so that it can be applied to linear regression, logistic regession, Cox proportional hazards models, etc.
- J is the penalty function wich (for l_1) takes the form:

$$J\left(\beta\right) = \sum_{j=1}^{d} |\beta_j|$$

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

- What about scale?

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

- What about scale?
  - Coefficients are relative to the scale of the predictor
  - Scale the predictors ($X_s = X/SD_x$)

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

- What about scale?
  - Coefficients are relative to the scale of the predictor
  - Scale the predictors ($X_s = X/SD_x$)
- What about categorical predictors?

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

- What about scale?
  - Coefficients are relative to the scale of the predictor
  - Scale the predictors ($X_s = X/SD_x$)
- What about categorical predictors?
  - How can you interpret?
  - Would be better to eliminate categories together

LASSO

If we apply a single penalty (λ) to all of the coefficients simultaneously what should we consider?

- What about scale?
  - Coefficients are relative to the scale of the predictor
  - Scale the predictors ($X_s = X/SD_x$)
- What about categorical predictors?
  - How can you interpret?
  - Would be better to eliminate categories together
    - Grouped LASSO
    - Elastic net

Elastic Net

Modification of the l_1 regularization to a quadratic form:

$$\min_{\beta} Loss\left(\beta; \mathbf{y}, \mathbf{X}\right) + (1-\alpha)||\beta||^2 + \alpha||\beta||^2$$

with $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$

- Encourages grouping effect on related variables (not guaranteed, but not necessary to specify the groups)
- Can select more variables than rows in the data ($k > N$)
- Stabilizes the regularization path

Sparsity

When we shrink coefficients we want them to *go away*!

Sparsity

When we shrink coefficients we want them to *go away*!

- The choice of regularization will influence how quickly coefficients are truncated to 0
- For Ridge regression this will never happen!
- LASSO and Elastic net both produce "sparse" results.

Sparsity

From: *Elements of Statistical Learning* Hastie, Tibshirani, and Friedman

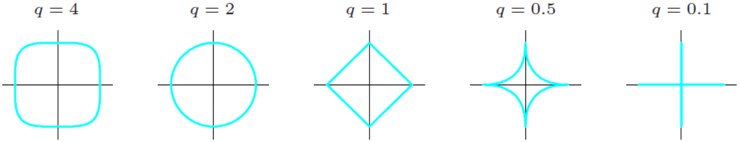


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

To get 0's the function must not be differentiable at the axes

Sparsity

From: *Elements of Statistical Learning* Hastie, Tibshirani, and Friedman

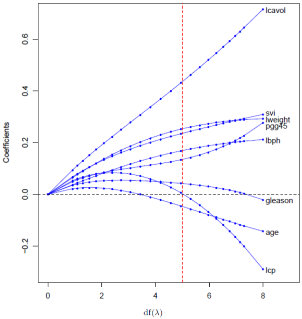


FIGURE 3.8. *Profiles of ridge coefficients for the prostate cancer example, as the tuning parameter λ is varied. Coefficients are plotted versus $\mathrm{df}(\lambda)$, the effective degrees of freedom. A vertical line is drawn at $\mathrm{df} = 5.0$, the value chosen by cross-validation.*

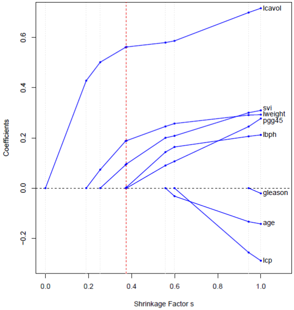
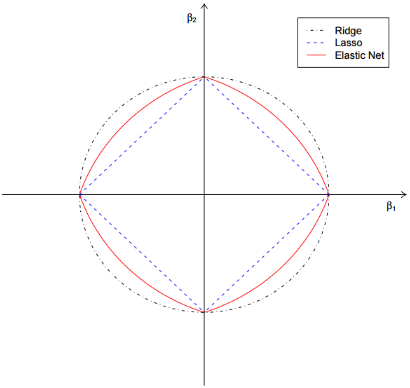


FIGURE 3.10. *Profiles of lasso coefficients, as the tuning parameter t is varied. Coefficients are plotted versus $s = t/\sum_1^p |\hat\beta_j|$. A vertical line is drawn at $s = 0.36$, the value chosen by cross-validation. Compare Figure 3.8 on page 65; the lasso profiles hit zero, while those for ridge do not. The profiles are piece-wise linear, and so are computed only at the points displayed; see Section 3.4.4 for details.*

Sparsity

From: Regularization and Variable Selection via the Elastic Net (Zou and Hastie 2005)



Elastic net regularization represents a compromise between LASSO and Ridge.

Cross-validation

The penalty parameter, λ, is considered "tuning" parameter.
How do we choose λ?

Cross-validation

The penalty parameter, λ, is considered "tuning" parameter.
How do we choose λ?

- Can be estimated from the data but this is generally not done.
- Select *lambda* based on prediction error using cross-validation

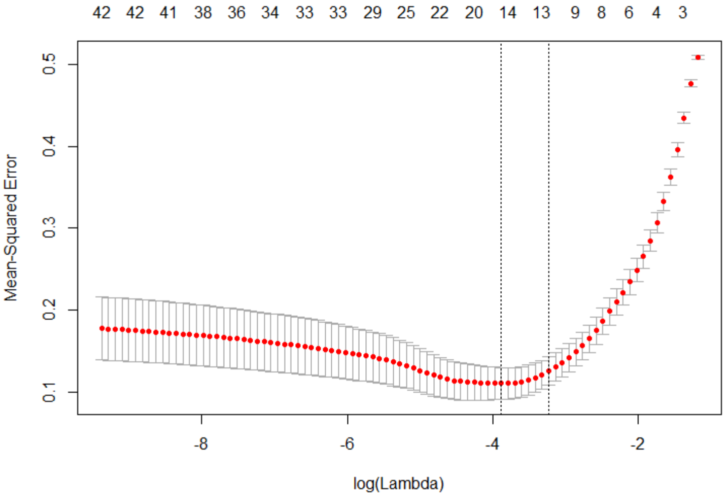
Cross-validation

K-fold Cross-validation steps: For each value of λ

1. Divide the data into two parts
2. "Train" the model on the first data set
3. Predict the outcome for the data in the second and calculate the error
4. Repeat k times for each value of λ

Cross-validation

Find the λ with the smallest prediction error and variance.



Cross-validation

What are some potential problems with this approach?

Cross-validation

What are some potential problems with this approach?

- Need a substantial amount of data.

Cross-validation

What are some potential problems with this approach?

- Need a substantial amount of data.
- Results may be very specific to the original data.

Cross-validation

What are some potential problems with this approach?

- Need a substantial amount of data.
- Results may be very specific to the original data.
- Interpretation of shrunken coefficients.

LASSO Implementation

SAS:
PROC GLMSELECT SEED=123;
PARTITION= ...; MODEL= .../ SELECTION=LASSO ;
(or SELECTION=LAR)
RUN;

YouTube Tutorial (click here)

R:
Package “glmnet” (created by Trevor Hastie and Junyang Qian)

cv.glmnet(x=predictors, y=response, family="gaussian", alpha=1, nlambda=100, nfolds=k)

glmnet(x=predictors, y=response, family="gaussian", alpha=1, lambda=best.lambda)

Model Selection

Variable selection finds important variables but what if you want to find important *relationships?*

Model Selection

Variable selection finds important variables but what if you want to find important *relationships?*

Model Selection:

Model Selection

Variable selection finds important variables but what if you want to find important *relationships?*

Model Selection:

- Start with a set of *plausible* models (must have some theory!)

Model Selection

Variable selection finds important variables but what if you want to find important *relationships?*

Model Selection:

- Start with a set of *plausible* models (must have some theory!)
- Fit and rank the models using some Information Criteria
  - AIC, BIC, KIC, AICc, qAIC....

Model Selection

Variable selection finds important variables but what if you want to find important *relationships?*

Model Selection:

- Start with a set of *plausible* models (must have some theory!)
- Fit and rank the models using some Information Criteria
  - AIC, BIC, KIC, AICc, qAIC....
- Select the top ranked model **or** Average the models to incorporate model selection uncertainty.

Information Criteria

All the information criteria follow the same general formula:

$$.IC = \text{-log(likelihood)} + \text{complexity penalty}$$

- $AIC = 2k - 2ln(L)$
- $BIC = -2ln(L) + k \cdot ln(n)$
- $KIC = -2\text{penalized log-likelihood} + C\left(\hat{\Sigma}_{\hat{\theta}}\right)*$

*I won't subject you to this function!

Model Averaging

How confident are we about the “best” model?

Model Averaging

How confident are we about the “best” model?

How much better does the best model have to be?

Model Averaging

How confident are we about the "best" model?

How much better does the best model have to be?

What about incorporating information from all the models?

Model Averaging

How confident are we about the "best" model?

How much better does the best model have to be?

What about incorporating information from all the models?

This is called multi-model inference.

Model Averaging

Model averaging (also called multi-model inference) has two main flavors:

Model Averaging

Model averaging (also called multi-model inference) has two main flavors:

- Bayesian Model Averaging (BMA)
- Frequentist Model Averaging (FMA)

Model Averaging

Model averaging (also called multi-model inference) has two main flavors:

- Bayesian Model Averaging (BMA)
- Frequentist Model Averaging (FMA)

These both follow the same general formula but have different philosophical motivations and interpretations

Model Averaging

The basic mechanics:

- Fit each model and assign an information score (AIC, BIC, etc.)

Model Averaging

The basic mechanics:

- Fit each model and assign an information score (AIC, BIC, etc.)
- Calculate a weight for each model based on the information score

$$w_i = \frac{e^{-\frac{1}{2}.IC_i}}{\sum e^{-\frac{1}{2}.IC_i}}$$

where $.IC_i$ is your favorite information criteria for model i.

Model Averaging

The basic mechanics:

- Fit each model and assign an information score (AIC, BIC, etc.)
- Calculate a weight for each model based on the information score

$$w_i = \frac{e^{-\frac{1}{2}.IC_i}}{\sum e^{-\frac{1}{2}.IC_i}}$$

  where $.IC_i$ is your favorite information criteria for model i.
- Use calculated weights to created a weighted average of the models

Model Averaging

The basic mechanics:

- Fit each model and assign an information score (AIC, BIC, etc.)
- Calculate a weight for each model based on the information score

$$w_i = \frac{e^{-\frac{1}{2}.IC_i}}{\sum e^{-\frac{1}{2}.IC_i}}$$

where $.IC_i$ is your favorite information criteria for model i.
- Use calculated weights to created a weighted average of the models
- Adjust standard errors for model uncertainty

Multi-model Inference vs Regularization

- Must have plausable models for model averaging (already know what variables are important)
- Regularization methods can handle very high dimensions ($k > n$)
- Model averaging incorporates the uncertainty in the final model
- Both methods shrink coefficients and improve prediction.

Questions?

Regularization:
Stanford open course on statistical learning (you will learn R at the same time) by Tevor Hastie, R. Tibshirani, and others.

Model Averaging:
Burnham and Anderson. 2004. Multimodel Inference. *Socio. Meth. and Research* 33(2)