Contents

Model Building Process

- Essential components of practicing statistics:
  - Study design
  - Data collection
  - Data analysis: Model selection, Model Fitting, Model Diagnostics
  - Conclusion and interpretation
- Building regression model
  - Data collection and preparation
  - Reduction of predictor variables (for observational studies)
  - Model refinement and selection (via model diagnostics and remedial measures)
  - Model validation and assessment

Controlled

- Controlled experiment:
  - The experimenter chooses the predictors to be examined.
  - The experimenter controls the levels of the explanatory variables and assigns a treatment, which is a combination of the levels of the explanatory variables, to each experimental unit and observes the response.
  - The explanatory variables are 'factors' or 'controlled variables'.
  - Reduction of explanatory variables is usually not too important.
- Controlled experiment with covariates:
  - Supplemental information impossible to incorporate into the design of an experiment, may be incorporated into the regression model as uncontrolled variables (or, covariates) in the model.
  - E.g., can't control gender or smoking status.
  - Inclusion of covariates can be helpful in reducing the error variance.

Uncontrolled

- Confirmatory observational studies:
  - Observational (not experimental) studies that are intended to test hypotheses derived from previous studies or intuition.
  - Covariates used to account for known influences on the response, e.g., 'risk factors'
  - Cast a wide net when choosing variables that could be included in the model
- Exploratory observational studies:
  - When it is not possible to conduct controlled experiments and there lacks adequate knowledge to conduct confirmatory observational studies, investigators search for explanatory variables that might be related to the response variable.
  - Determine potentially useful explanatory variables.
  - Screen out some of these explanatory variables.

More on Exploratory Observational Studies

- After the initial screening, further reduction of explanatory variables is often still needed.
- Some challenging issues are how to
  - Identify "good" subsets of explanatory variables.
  - Determine functional form of the regression (linear, quadratic, etc.)
  - Determine if interaction terms are to be included.
- Some considerations are
  - Omission of key explanatory variables $\Rightarrow$ Increased bias.
  - Inclusion of unimportant explanatory variables $\Rightarrow$ Increased variance.
- Different "best" subsets serve different purposes (descriptive versus predictive).
- For a given purpose, several subsets may be equally "good".

Purposes of Model Selection

Consider $y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j x_{ij} + \epsilon_i$, can any of the $p-1$ predictor variables be dropped?

- If the purpose is description/explanation/understanding, then
  - Empirical/data driven models (background).
  - Parsimony is a key priority– Occam's razor.
- If the purpose is prediction, then
  - Theoretical/physical models, e.g., engineering, physics, etc.
  - No model selection may be required.
  - Models are evaluated by predictive accuracy/power.
- Other purposes:
  - Choose predictor variables for further study.
  - Save resources, money, and/or effort.

Motivations for Model Selection

Consider $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$ for $i = 1, \ldots, n$. Without loss of generality, we center the response that $\sum_{i=1}^n y_i = 0$ by subtracting $\bar{y}$ and the predictors are similarly standardized so that $\sum_{i=1}^n x_{ij} = 0$ and $\sum_{i=1}^n x_{ij}^2 = 1$.

- Consider $A_0 = \{1, \ldots, p_0\} \subset A = \{1, \ldots, p\}$ with $p_0 \leq p$. The the design matrix becomes

$$\boldsymbol{X} = [\boldsymbol{X}_{A_0} \boldsymbol{X}_{A_0}^c]$$

  and the true coefficient are $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_{A_0}^{*,'}, \mathbf{o}')'$.

- If the correct choice of variables were known, then OLS of $\boldsymbol{\beta}_{A_0}^*$ is

$$\widehat{\boldsymbol{\beta}}_{A_0}^{ols} = (\boldsymbol{X}'_{A_0}\boldsymbol{X}_{A_0})^{-1}\boldsymbol{X}'_{A_0}\boldsymbol{y}$$

- We can show (HW) that for any $\boldsymbol{x} \in \mathbb{R}^p$

$$\mathbb{E}(\boldsymbol{x}'\widehat{\boldsymbol{\beta}}_{A_0}^{ols}) = \boldsymbol{x}'\boldsymbol{\beta}^* = \boldsymbol{x}'_{A_0}\boldsymbol{\beta}_{A_0}^* = \mathbb{E}(\boldsymbol{x}'_{A_0}\widehat{\boldsymbol{\beta}}_{A_0}^{ols})$$

  and

$$\text{Var}(\boldsymbol{x}\widehat{\boldsymbol{\beta}}_{A_0}^{ols}) \geq \text{Var}(\boldsymbol{x}'_{A_0}\widehat{\boldsymbol{\beta}}_{A_0}^{ols})$$

- The inclusion of extraneous variables can lead to inflated estimates of coefficients and wider confidence and prediction intervals.
- In practice, a perfectly true model is rarely known, so as more variables are added to the model, reduced bias is trade off against increased variance.
- If added variable is not in the true model, then the increase in prediction variance will compromise the reduction in bias if any.

Classical approaches

- Let $\mathcal{S}$ be a variable varying over subsets of A. Then *classical model selection* seeks an $\mathcal{S}$ of a certain size that achieves a relatively small SSE.
- One important measure of how well a given model specified by $\mathcal{S}$ describes the data is

$$R_k^2(\mathcal{S}) = 1 - \frac{SSE_k}{SSTO} \uparrow \quad \Rightarrow \quad R_k^2(\mathcal{S}) \uparrow$$

  where k is the number of parameters or $|\mathcal{S}|$
- Possible procedures:
  - For fixed k, find the model with the largest R_k^2, or
  - Find the model to which adding any single additional predictor variable leads to a very small increase in R_k^2. This approach can be very subjective.

Classical approaches (Cont'd)

- One main disadvantage of R^2 is that it favors large models as its value decreases monotonically as variables are added to the model.
- Recall adjusted coefficient of multiple determination

$$R^2_{a,k} = 1 - \left(\frac{n-1}{n-k}\right)\frac{SSE_p}{SSTO}$$

  and it is possible that $R^2_{a,k}$ could increase or decrease when $k \uparrow$.
- Possible procedures:
  - Select the models with or close to the maximum $R^2_{a,k}$ or,
  - For fixed p, find the model with the largest $R^2_{a,k}$, or
  - Find the model to which adding more predictor variables leads to a small increase or even decrease in $R^2_{a,k}$.

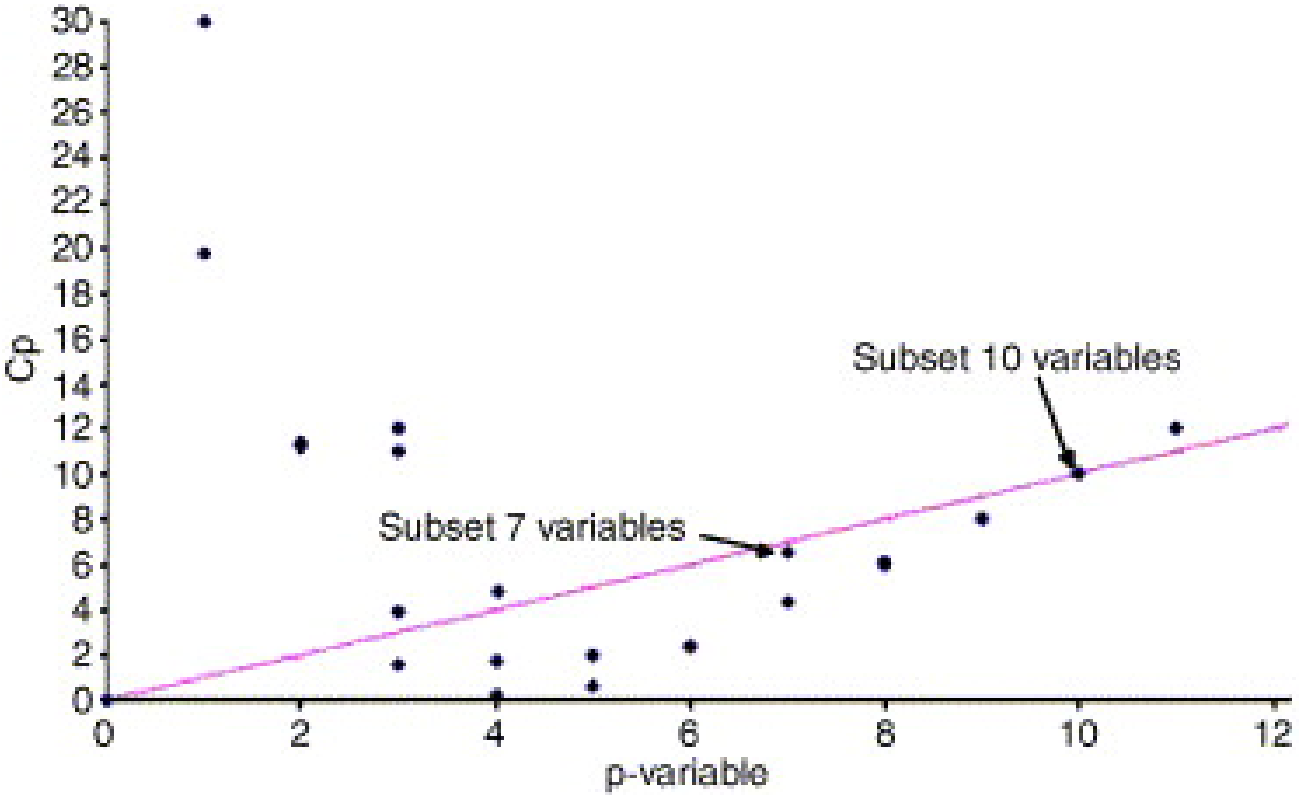
Classical approaches (Cont'd): Mallows' C_p Criterion

- Mallows (1964,1973) proposed a useful criterion to determine the optimal number of variables to be retained in a model.
- The idea is to search for the correct $\mathcal{S}$ by finding the best model of the best size k.
- Mallows' criterion is called Mallows' C_p, and is defined as

$$C_k = SSE/\widehat{\sigma}^2 - (n - 2k)$$

  where $\widehat{\sigma}^2$ is the residual mean square from the *full model* with p covariates.
- Rule of thumbs:
  - Plot C_k vs. k for various ks and add the line for $C_k = k$.
  - Select a model that has $C_k \approx$ or $\leq k$.
  - Models with substantial bias tend to fall above the line $C_p = p$.

Example C_k plot



Highlights

- Standard model/variable selection procedures typically fall into one of two categories: *subset selection* and *variable rankings*.
  - Subset selection favors those $\mathcal{S}$ with relatively small SSEs. General subset selection method are just slightly more sophisticated than using $R^2_{a,k}$, MSE or Mallows' C_p directly.
  - It is important to remember *though SSE is a good assessment of fit, it says little about predictions*.
  - Variable ranking tries to assign a worth to each variable x_j for their associations with response y. It is a marginal approach.
- Modern model/variable selection procedures focus on shrinkage method by using constrained optimization techniques.
- Bayesian model selection/average provides the third branch of this topic.

Subset selections

- In principle, an exhaustive search could be done by fitting $2^p - 1$ candidate models and selecting the best one under some criterion, such ash the largest $R^2_{a,k}$, smallest MSE or Mallows' C_p.
- A variety algorithm has been proposed to overcome computational cost: main idea is to *identify the best subsets by a greedy algorithm so large number of suboptimal subsets can be ruled out.*
- Hocking and Leslie (1967) observation: *when the SSE due to eliminating a set of variables for which the maximum subscript j is less than the SEE due to eliminating the variable $(j+1)$, then no subset including any variables with subscripts greater than j can result in a smaller reduction*, that is (Furnival (1971))

$$\mathcal{S}_1 \subset \mathcal{S}_2 \Rightarrow SSE(\mathcal{S}_1) \geq SSE(\mathcal{S}_2)$$

Subset selections (Cont'd)

- Furnival (1971) proposed a computationally efficient implementation of the "leaps and bounds" procedure by Hocking and Leslie.
- Both methods below to a class of optimization method called branch and bound.
- Large numbers of candidate models are rule out by using estimated upper and lower bounds on the quantity being optimized, like SSE or Mallows' Cp etc.
- Works well under sparse, and independent predictors.
- Still, exponential running time (NP Hard), only effective for up to 50 or so variables.

Sequential search method

- Fit a sequence of regression models and at each step, add or delete one predictor variable.
- Including forward, backward, and stepwise regression.
- Forward selection: *selecting variables based on their partial correlations*
  - Choose a measure for the jth predictor, e.g., $|t_j^*|$ or F-statistic or $SSE_{k+1}(j)$.
  - Algorithm
    1. Start with the null model.
    2. Fit a simple linear regression model for each of the $p-1$ predictor variables (separately). Compute $t_k^* = \hat{\beta}_k \Big/ s\{\hat{\beta}_k\}$ for each added variable and add the variable with the largest $|t_k^*|$ provided that it is significant at a pre-specified α level.
    3. Conditional on the current model, fit all models that add a single one of the remaining variables. Among these, add the variable with the largest $|t_k^*|$, provided that it is significant.
    4. Repeat the previous step until there are no variables that can be added to the current model that are significant. The resulting model is chosen.

Backwards Elimination

- **Backward elimination**: reverse the forward selection.
  - Start with the full model with p variables, and at each step remove the variable making the smallest contributions.
  - Suppose that there are k variables with $k \leq p$ in the current model. The corresponding design matrix is $\boldsymbol{X}_k$. Then the new SSE from deleting the jth $(1 \leq j \leq k)$ variable from the current model is (HW)

$$SSE_{k-1}(j) = SSE_k + (\widehat{\beta}_j)^2 / s_{jj}$$

    where $\widehat{\boldsymbol{\beta}}_k = (\widehat{\beta}_1, \ldots, \widehat{\beta}_k)'$ is the vector of the current regression coefficients and s_{jj} is the jth diagonal of $(\boldsymbol{X}_k' \boldsymbol{X}_k)^{-1}$.
  - Deletion continues until it starts to harm the fit. For example, the variable X_j is removed if

$$F_j = \min_j \left(\frac{SSE_{k-1} - SSE_k}{SSE_k/(n-k)} \right) < F_{out}$$

    where $F_{out} = F(\alpha; 1, n-k)$. Repeat the previous step until there are no variables that can be removed because all ones remaining in the model are significant.

Stepwise Selection

- One problem with forward or backward is that once a decision is made, it is never reversed.
- **Stepwise selection:** (Efroymson (1960))
  1. Begin with the null model.
  2. Do one step of forward selection.
  3. Do another step of forward selection (since backward elimination here is pointless).
  4. Do one step of backward elimination.
  5. Do one step of forward selection.
  6. Etc. Keep alternating.
  7. If a stopping rule is triggered, skip that step in the current cycle.
  8. If *both* stopping rules are triggered, stop because no further variables can be added or deleted.

Comments About Stepwise Methods

- The criterion $|t_k^*|$ can be replaced with F^*, R_a^2, SSE, etc.
- Dramatically reduce computational cost.
- Drawbacks of stepwise selection are:
  - It is not guaranteed to find the global optimal set.
  - Instability: the process is a discrete process so small change in data may cause large change in variable selection (Brieman (1996)).
  - Sensitive to the criterion, but could be fixed by using robust risk.
- A rule of thumb: stepwise is preferred over forward or backward selection.

Variable ranking

- When p grows, traditional subset selection method becomes computationally difficult.
- When $p > n$, all of them break down completely because OLS estimates are not defined (for very extreme sparse assumption, forward selection might remain valid).
- It is helpful to *screen* variables to eliminate those redundant or noisy. Such screening is often done by ranking the variables on the basis of some criterion and eliminateing all variables that do not have a high enough score.
- To overcome the *Curve of Dimensionality*, most ranking methods are based on marginal models: that is on a univariate model for y using a single x_j.
- Various coefficients reflecting associations between y and x_j have been used including Pearson correlation, t-statistics, p-values, Kendall's τ and Spearman correlation.

Sure independence screening (SIS)

- In 2008, Fan and Lv proposed a screening method suitable for ultra-high dimensional cases that is $p \gg n$.
- The method can asymptotically identify the correct model, which is termed as *consistent selection*.
- It is based on the marginal correlations of single x_j and y.
  - The vector of marginal correlations of individual predictors x_j and y scaled by the standard deviation of y is $\boldsymbol{\omega} = \boldsymbol{X}'\boldsymbol{y}$. item For a given $\gamma \in (0, 1)$, SIS sorts the p componentwise magnitude $\boldsymbol{\omega}$ into decreasing order to define submodels of the form

$$\{\mathcal{S}_\gamma = \{1 \leq j \leq p| \ |\omega_j| \text{ is the one of the } \lfloor \gamma n \rfloor \text{ largest entires in } \boldsymbol{\omega}\}$$

  - As γ increases, $\mathcal{S}$ increases so smaller γs give smaller models.
  - Under several regularity conditions, it is shown that if the distribution of $\boldsymbol{Z} = \boldsymbol{X}\text{Cov}(\boldsymbol{X})^{-1/2}$ is spherically symmetric for normal data for normal data having concentration properties, then SIS can captures all the important variables with probability tending to one as n increases.

Variable ranking

- Variable ranking based on marginal information may not work well in the presence of collinearity: it is possible that many unimportant variables are highly correlated with important ones and so may be likely to be selected over important predictors with weaker marginal signals.
- Also marginal information ignores interaction effects: it is possible that any important variables that are marginally uncorrelated with y bu jointly correlated with response being ruled out.
- When the predictors have nonlinear effects on y, then linear correlations is not likely to work. Recall the generalized additive model

$$y = \sum_{j=1}^{p} f_j(x_j) + \epsilon$$

Still an open problem: natural approach is to employ marginal smoothing method and goodness-of-fit statistics

Information criterion

- Recall that for any subset index $\mathcal{S} \subset \{1, \ldots, p\}$, the multiple regression or linear model $y = \sum_{j \in \mathcal{S}} \beta_j x_j + \epsilon$ can be fit and best subset selection means finding the $\mathcal{S}$ to minimize some criterion.
- The first type of criteria will be information-theoretic, that is based on penalizing a log likelihood by the model complexity, rather than based on residual error as in the traditional settings.
  - For example, Mallows' C_p, AIC and BIC etc.
- The second type of criteria is based on cross-validation, that evaluates the model performance using internal validation data set.
- These two type of criteria are asymptotically equivalent under some condition.

Information criterion: formulations

- Let $g(y|\boldsymbol{x}, \boldsymbol{\beta})$ be the density of the response y. Then for a sample of n observations, the log-likelihood is

$$\log L = \sum_{i=1}^{n} \log g(y_i|\boldsymbol{x}_i, \boldsymbol{\beta})$$

- Typically, an *informative criterion* is of the form

$$IC_k = -2[\log(L_k(\widehat{\boldsymbol{\beta}}_{MLE})) - \phi(n)k]$$

where $\log(L_k(\widehat{\boldsymbol{\beta}}_{MLE}))$ is the maximized log likelihood of a subset model containing a choice of k variables and $\phi(n)$ is a factor specifying the penalty on model dimension.

  - $\phi(n)$ increases in n and may take different forms for different criteria
  - IC_k can be interpreted as a combination of goodness of fit and model complexity, or as *bias and variance*

Information criterion: Gaussian and consistency

- Assume $g(y|\boldsymbol{x}, \boldsymbol{\beta})$ is $N(\boldsymbol{x}'\boldsymbol{\beta}, \sigma^2)$, then

$$IC_k = SSE/\sigma^2 + 2\phi(n)k$$

  - Estimate σ^2 by $\widehat{\sigma}^2 = SSE_p/n - p)$ under the full model
  - Mallows' C_p is an information criterion with $\phi(n) = 1$

- The best model have size $\widehat{k}$ with $\widehat{k} = \arg\min_{1 \le k \le p} IC_k$
- Assume the true model has size $p_0 \le p$, then
  - $k < p_0$ is called misspecified
  - $k > p_0$ is called correctly specified but overparameterized
  - An information criterion is *consistent* if

$$\lim_{n \to \infty} \mathbb{P}(\widehat{k} = p_0) = 1$$

- As long as $\lim_{n\to\infty} \phi(n)/n = 0$, IC_k is unlikely to lead a misspecified model asymptotically.
  - Actually, we can show

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}(\widehat{k} < p_0) \\
&\leq \limsup_{n\to\infty} \mathbb{P}(IC_{p_0} > IC_k \text{ for some } k < p_0) \\
&= \limsup_{n\to\infty} \mathbb{P}(-2\log L_{p_0}/n + 2p_0\phi(n)/n > -2\log L_k/n + 2k\phi(n)/n \text{ for some } k < p_0) \\
&= \limsup_{n\to\infty} \mathbb{P}(\log L_{p_0}/n - \log L_k/n < (p_0 - k)\phi(n)/n \text{ for some } k < p_0) \\
&\leq \sum_{k<p_0} \limsup_{n\to\infty} \mathbb{P}(\log L_{p_0}/n - \log L_k/n < 0) = 0
\end{aligned}$$

- This *one-sided property* in general hold for AIC, BIC, etc as long as $(p_0 - k)\phi(n)/n = o(1)$ (which implies sparsity automatically)
- *Consistency of model selection based on information criterion, however, requires* $\mathbb{P}(\widehat{k} > p_0) \to 0$ *as* $n \to \infty$. This is not held in general, need extra conditions.
  - Asymptotic results do not guarantee a satisfactory performance for finite data
  - Standard errors for parameter estimation and prediction post model selection will necessarily be increased due to the sample distribution of the model selection.

AIC Criteria

- Akaike (1973) proposed a selection criterion as

$$AIC_k = -2[\log(L_k(\widehat{\boldsymbol{\beta}}_{MLE})) - k]$$

  - AIC is IC with $\phi(n) = 1$
  - For normal data, AIC is $AIC_k = n\log(SSE_p) - n\log(n) + 2k$
  - AIC is information-theoretic as it can be derived from the Kullback-Leibler distance (STAT730)
  - AIC is a relative measure: only AIC values from the same data should be compared

- AIC is not consistent (Shibata (1983))! That is

$$\lim_{n\to\infty} \mathbb{P}(\widehat{k}_{AIC} \geq p_0) = 1 \text{ and } \lim_{n\to\infty} \mathbb{P}(\widehat{k}_{AIC} > p_0) > 0$$

  - The penalty in AIC is relatively small, not dependent on n, so AIC tends to permit larger models and therefore is better for prediction than other variable selection methods.
  - AIC, however, is minimax optimal (Barron (1999), Yang and Barron (1999))

BIC_k Criteria

- Schwartz (1978) proposed and refined a Bayesian information criterion

$$BIC_k = -2[\log(L_k) - \log(n)k]$$

  - It can be viewed as letting $\phi(n) = \log(n)$
  - For normal data, $BIC_k = n\log(SSE_k) - n\log(n) + k\log(n)$
  - BIC was obtained by maximizing the posterior probability of a model being selected from $2^p - 1$ candidates via maximizing

$$\mathbb{P}(\mathcal{S}_k|\boldsymbol{y}) \propto W(\mathcal{S}_k)\int \mathbb{P}(\boldsymbol{y}|\boldsymbol{\beta}_k, \mathcal{S}_k)\mathbb{P}(\boldsymbol{\beta}_k|\mathcal{S}_k)d\boldsymbol{\beta}_k = \log(\mathbb{P}(\boldsymbol{y}|\widehat{\boldsymbol{\beta}}_k^{mle}, \mathcal{S}_k) - \frac{k}{2}\log(n) + O(1)$$

    which is obtained through Laplace approximation on the integrals.
  - BIC is also information-theoretic as it can be derived from Kullback-Leibler distance
  - BIC can also be interpreted as Bayes Factor (BF)
  - BIC is minimum description length (Barron and Cover (1991))

- BIC is consistent!
  - *A key fact: most Bayesian procedures for different statistical inference problems have consistency when the true model is in the support of the priors.*
  - Actually, we can see that as $\phi(n) \to \infty$,

$$\limsup_{n\to\infty} -2(\log L_{p_0} - \log L_k)/\phi(n) = 0$$

and

$$\begin{aligned}
&\limsup_{n\to\infty}(BIC_{p_0} - BIC_k)/\phi(n) \\
=&\limsup_{n\to\infty} -2(\log L_{p_0} - \log L_k)/\phi(n) + 2(p_0 - k) \\
=&p_0 - k \leq -1
\end{aligned}$$

so that

$$\lim_{n\to\infty} \mathbb{P}(\widehat{h} > p_0) = \lim_{n\to\infty} \mathbb{P}(BIC_{p_0} \geq BIC_{k+1}) = 0.$$

- BIC, however, in practice may not be consistent in finite samples.
- BIC is NOT minimax optimal (Forster and George (1994)), that the BIC may require a lot of sample to get the true model selected, more than other criterion.

AIC and BIC

- General guideline for when to use what is hard, and more or less requires experience.
- Still, some guidelines.
  - A rule of thumb: models ≤ 2 AIC units apart are not much different.
    - ⋆ Therefore, good models are those that are within 2 AIC units of the lowest AIC value.
    - ⋆ Models with more than 10 AIC units above the lowest AIC value are generally not considered.
  - Between AIC and BIC.
    - ⋆ If the true model is simple, extremely sparse, or finite dimensional, then higher penalties such as BIC should be used.
    - ⋆ If the true model is complex or infinite dimensional, then smaller penalties such as AIC should be used.
    - ⋆ AIC is slower than the BIC for identifying the right model as n increases.
    - ⋆ However, the comparative efficiency of AIC allows it to be more robust and better for predictions.

- Yang (2005) showed that the strengths of AIC and BIC cannot be shared.
  - Any model selection criterion is consistent like BIC, it cannot be minimax-optimal, i.e. it must have a worse mean average squared error than AIC.
- A reconciliation was proposed by Erven et al. (2008), and the procedure switch from AIC (for searching) to BIC (for model identification) at some stage in sequential. It was shown that the method outperforms both AIC and BIC.

Other criteria

- For small sample size n, Hurvich and Tsai (1989) define AICc as

$$AIC_{c,k} = AIC_k + \frac{2k(k+1)}{n-k-1} = -2\left(\log L_k - \frac{n}{n-p-1}\right).$$

  - AIC can result in overfitting, but AICc protects against overfitting.
  - $\phi(n) = n/(n-k-1) > 1$, AICc tends to have a sharper cutoff and model with smaller subsets will be selected when p is large compared with n.
  - AICc was recommended when $n/p < 40$.

- Hannan and Quinn (1979) considered selecting time series models in which k increasing in n, and for the selection of the order of an autoregressive model they consider

$$HQ_k = -2[\log L_k - \log\log(n)k]$$

- Deviance information criterion (DIC) arises from calling $D(\boldsymbol{\theta}) = -2\log g(\boldsymbol{y}|\boldsymbol{\theta})$ a deviance and setting $\bar{D} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}} D(\boldsymbol{\theta})$ and $\bar{\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{\theta}|\boldsymbol{y}}\boldsymbol{\theta}$ so

$$DIC = 2\bar{D} - D(\bar{\boldsymbol{\theta}})$$

  where $k_D = \bar{D} - D(\bar{\boldsymbol{\theta}})$ is the effective number of parameters.

Model Validation

- Model validation refers to checking a selected model against independent data. There are several possible approaches.
  - Collect new data as validation data set and check
    - ⋆ stability of regression coefficient estimation
    - ⋆ accuracy of prediction
  - Compare results with theory or simulations.
  - Split data into training and validation set, and check
    - ⋆ stability of regression coefficient estimation
    - ⋆ accuracy of prediction
    - ⋆ This is called cross validation (bootstrap is also one special case of it)

Cross Validation

- CV is a blackbox tool for choosing model based on their *predictive ability*.
- CV is routinely used to select the important subsets of variables in linear models, select and build the architecture of neural networks and trees, choose the regularization parameters for smoothing splines or other penalized methods, and to select the bandwidth for kernels.
- Given a data set with n observations, one common measure for assessing the predictive performance of a fitted model $\widehat{f}$ is the *mean squared prediction error (MSPE)*

$$MSPE = \mathbb{E}_{\boldsymbol{X},y}[y - \widehat{f}(\boldsymbol{X})]^2$$

  - Since the distribution is unknown in real practice, it is necessary to have test data samples independent of the original data and compute

$$\widehat{MSPE} = \frac{1}{n^*}\sum_{i=1}^{n^*}(y_i^* - \hat{y}_i(\boldsymbol{x}_i^*))^2.$$

  - If $\widehat{MSPE} \approx MSE$, then the model is probably adequate, otherwise if $\widehat{MSPE} \gg MSE$, then the model may not be very useful for general use.

- In practice, however, independent data set are frequently difficult or expensive to obtain.
- Also, it is undesirable to hold back data from the original data to use for a separate test because it weakens the training data.
- CV will split the data set $\mathcal{D}$ into two complement parts as $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$.
  - The first part contains n_1 samples used for constructing a model
  - The second part contains $n_2 = n - n_1$ samples for assessing the predictive power of the model
  - This procedure repeats many times, and there are $\binom{n}{n_1}$ ways of partitioning the data.
  - The CV score is the average prediction error based on the different ways the data were partitioned.

Leave-one-out CV (LOOCV)

- Mosteller and Tukey (1968) introduced the idea of LOOCV that $n_2 = 1$ so there are n ways splitting the original data.
- Similar to the deleted residuals discussed for MLR.
- The average of n prediction errors is called LOOCV score

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}[y_i - \widehat{f}_{(-i)}(\boldsymbol{x}_i)]^2$$

- As we have seen for MLR, it is not necessarily to fit the model n times for linear model (smoother that $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$) and we have

$$LOOCV = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \widehat{f}(\boldsymbol{x}_i)}{1 - s_{ii}}\right]^2$$

  for linear model $f(\boldsymbol{x})$ and s_{ii} is the diagonal of the linear operator specified by f (leverage for MLR).
- LOOCV is approximately *unbiased* for the true prediction error but may suffer *high variance* because n training sets are very similar to each other.

K-Fold Cross Validation

- The K-fold CV (Brieman et al. (1984)) removes observations in groups not only one of them.
- $n_2 = n/K$ and there are K way to partition the data
- Procedure is as following.
  1. Partition the data into K roughly equal-sized groups
  2. For each $k = 1, \ldots, K$, one fit the model using all data but the kth group.
  3. Repeat this K times, with each group used exactly once as the validation set.
  4. The K prediction error are then averaged to define CV error

$$CV_K = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i)]^2$$

  where $m : \{1, \ldots, n\} \to \{1, \ldots, K\}$
- If the models are indexed by a parameter $\alpha \in \Lambda$, such as regularization parameter or bandwidth etc, then for each α

$$CV_K(\alpha) = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i; \alpha)]^2$$

  and $CV_K(\alpha)$ provides an estimated test error curve that optimal parameter $\widehat{\alpha}_{opt} = \arg\min_\alpha CV_K(\alpha)$
- In practice, $K = 5$ or $K = 10$.

K-Fold Cross Validation

- The K-fold CV (Brieman et al. (1984)) removes observations in groups not only one of them.
- $n_2 = n/K$ and there are K way to partition the data
- Procedure is as following.
  1. Partition the data into K roughly equal-sized groups
  2. For each $k = 1, \ldots, K$, one fit the model using all data but the kth group.
  3. Repeat this K times, with each group used exactly once as the validation set.
  4. The K prediction error are then averaged to define CV error

$$CV_K = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i)]^2$$

  where $m : \{1, \ldots, n\} \to \{1, \ldots, K\}$

- If the models are indexed by a parameter $\alpha \in \Lambda$, such as regularization parameter or bandwidth etc, then for each α

$$CV_K(\alpha) = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i; \alpha)]^2$$

  and $CV_K(\alpha)$ provides an estimated test error curve that optimal parameter $\widehat{\alpha}_{opt} = \arg\min_\alpha CV_K(\alpha)$
- In practice, $K = 5$ or $K = 10$.

K-Fold Cross Validation

- The K-fold CV (Brieman et al. (1984)) removes observations in groups not only one of them.
- $n_2 = n/K$ and there are K way to partition the data
- Procedure is as following.
  1. Partition the data into K roughly equal-sized groups
  2. For each $k = 1, \ldots, K$, one fit the model using all data but the kth group.
  3. Repeat this K times, with each group used exactly once as the validation set.
  4. The K prediction error are then averaged to define CV error

$$CV_K = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i)]^2$$

  where $m : \{1, \ldots, n\} \to \{1, \ldots, K\}$
- If the models are indexed by a parameter $\alpha \in \Lambda$, such as regularization parameter or bandwidth etc, then for each α

$$CV_K(\alpha) = \frac{1}{n} \sum i = 1^n [y_i - \widehat{f}_{(-m(i))}(\boldsymbol{x}_i; \alpha)]^2$$

  and $CV_K(\alpha)$ provides an estimated test error curve that optimal parameter $\widehat{\alpha}_{opt} = \arg\min_\alpha CV_K(\alpha)$
- In practice, $K = 5$ or $K = 10$.

Generalized CV (GCV)

- The computation for LOOCV is often expensive since the whole process requires fitting the model n times unless for linear model.
- *GCV provides a convenient way to approximate LOOCV for linear fitting methods under the squared error loss.*

$$GCV = \frac{1}{n}\sum_{i=1}^{n}\frac{[y_i - \widehat{f}(\boldsymbol{x}_i)]^2}{[1 - \text{tr}(\boldsymbol{S})/n]^2}$$

where $\boldsymbol{S}$ the matrix for $\widehat{y} = \boldsymbol{S}\boldsymbol{y}$

  - $\text{tr}(\boldsymbol{S})$ is called the effective number of parameter
  - GCV works well if s_{ii} are not very different from each other
  - GCV is weighted version of LOOCV with weights $[(1 - s_{ii})/(1 - \text{tr}(\boldsymbol{S})/n)]^2$
  - GCV is close to Mallows' C_p and AIC asymptotically

Remarks for CV

- Careful use of CV includes two steps.
  - The standard error of the CV score is important. It partially justifies the one-standard-error rule to avoid underfitting for linear models.
  - CV strongly favors smaller models, choosing the most parsimonious model whose error is no more than one standard error above the optimal one is a reasonable fix.
  - Second, it is important to look at histogram of $y_i - \widehat{f}(\boldsymbol{x}_i)$. If it is roughly normal, then CV is reliable; otherwise, it reflect bad fits.
- When the number of predictors increased as n increases, LOOCV is both consistent and optimal.

Shrinkage methods: introduction

- Penalized method or regularization method, or shrinkage, are usually based on adding a penalty term to the objective function.
- This can be done for a wide variety of model classes, not just linear models.
- Conceptually, shrinkage method aim to make an ill-posed problem well-posed, such as ridge regression.
- A good procedure for model selection need
  1. filter out unimportant variables
  2. estimate the regression coefficient of the important ones consistently with a high level of efficiency (small noise, like $n^{-1/2}$)

  Such variable selection procedure is called "oracle".
- A variable selection procedure for linear models is oracle if with probability tending to one,
  1. Consistent selection: $\widehat{\beta}_j \neq 0$ for $j \in A_0$ and $\widehat{\beta}_j = 0$ for $j \in A_0^c$
  2. Optimal estimation: $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{A_0} - \boldsymbol{\beta}^*_{A_0}) \to N(\mathbf{o}, \boldsymbol{\Sigma}^*)$ in distribution

- Why we bother introducing these fancy methods? Stepwise methods are all non-oracle!
- Oracle procedure performs as well asymptotically as a procedure knowing the true model, and uncover it so quickly that the convergence of parameter estimation is not affected by being based on the wrong, preconvergence model.
- Stepwise is not oracle as
  1. Greedy method tend to seek local rather than global optimal
  2. Subset selection is discrete and calls both efficiency and consistency into question
  3. Sensitivity to the data perturbation or outliers

Ridge Regression

- Ridge regression is introduced to address multicollinearity problems, and it solves

$$\widehat{\boldsymbol{\beta}}^{ridge} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \beta_j^2$$

  - $\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage: the larger the λ the greater the amount of shrinkage.
  - An equivalent way to write the ridge problem is

- The ridge solution is

- The ridge solution adds a positive constant to the diagonal of $\boldsymbol{X}'\boldsymbol{X}$ before inversion, which makes the problem nonsingular even if $\boldsymbol{X}'\boldsymbol{X}$ is not of full rank (Hoerl and Kennard, 1970).
- $\widehat{\boldsymbol{\beta}}^{ridge} \to \widehat{\boldsymbol{\beta}}^{OLS}$ as $\lambda \to 0$.
- If X is orthogonal, then $\widehat{\boldsymbol{\beta}}^{ridge}$ is

- Ridge regression estimator is a shrinkage estimator that shrinks the OLS estimator toward zero.
- When $\lambda > 0$, the estimated $\widehat{\boldsymbol{\beta}}^{ridge}$ is more stable, but biased.

- Let the singular value decomposition of the design matrix be $\boldsymbol{X} = \boldsymbol{U}\Lambda^{1/2}\boldsymbol{V}'$ such that

  then

$$\widehat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{V}\Lambda\boldsymbol{V}' + \lambda \boldsymbol{V}\boldsymbol{V}')^{-1}\boldsymbol{V}\Lambda^{1/2}\boldsymbol{U}'\boldsymbol{Y}$$
$$= $$

- The amount of bias increases as λ increases, but the variance component of MSE decreases.

$$MSE(\lambda) = \mathbb{E}\left\{(\widehat{\boldsymbol{\beta}}^{ridge} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}}^{ridge} - \boldsymbol{\beta})\right\} = \text{Var}(\lambda) + \text{Bias}^2(\lambda)$$

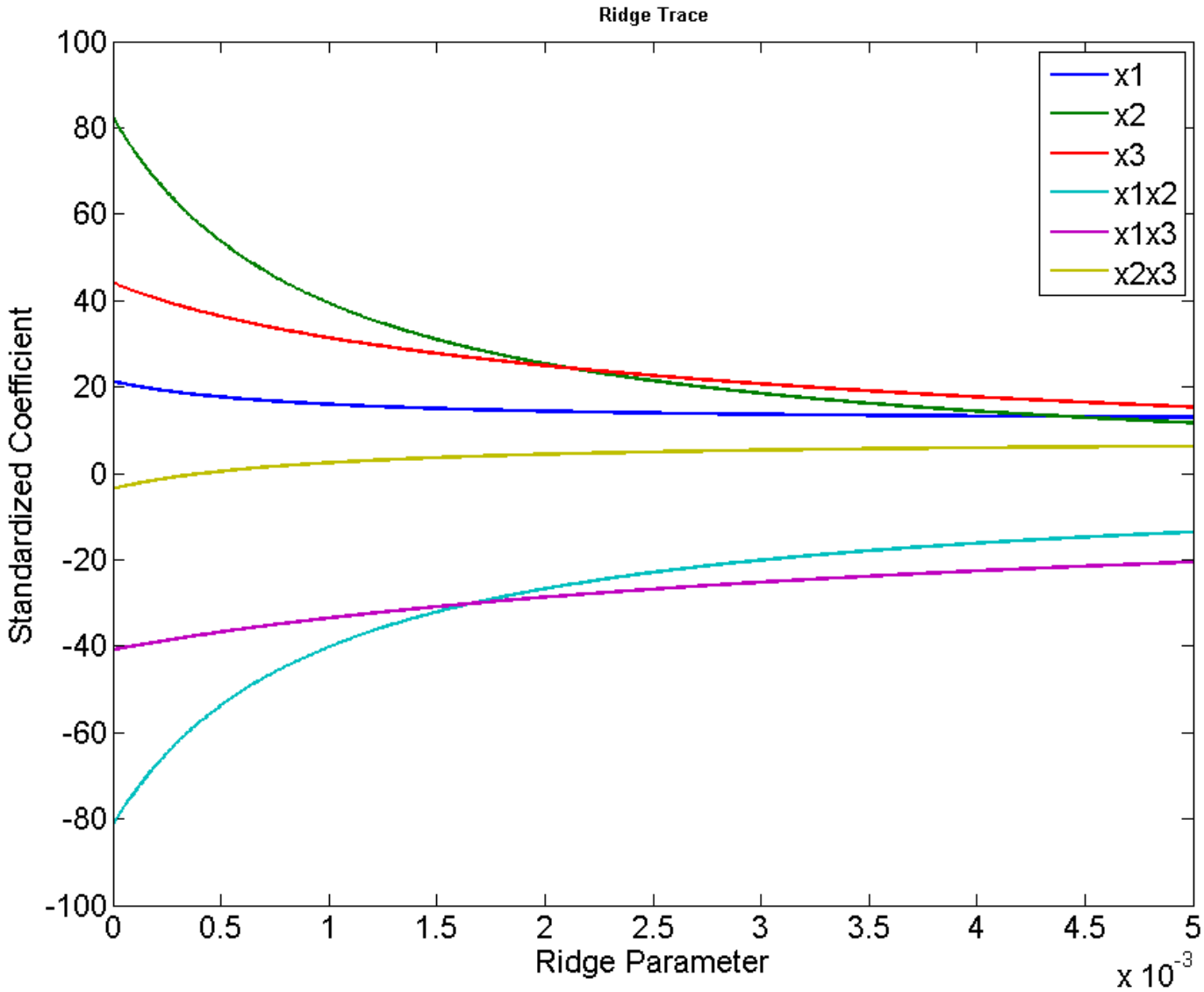
$$=$$

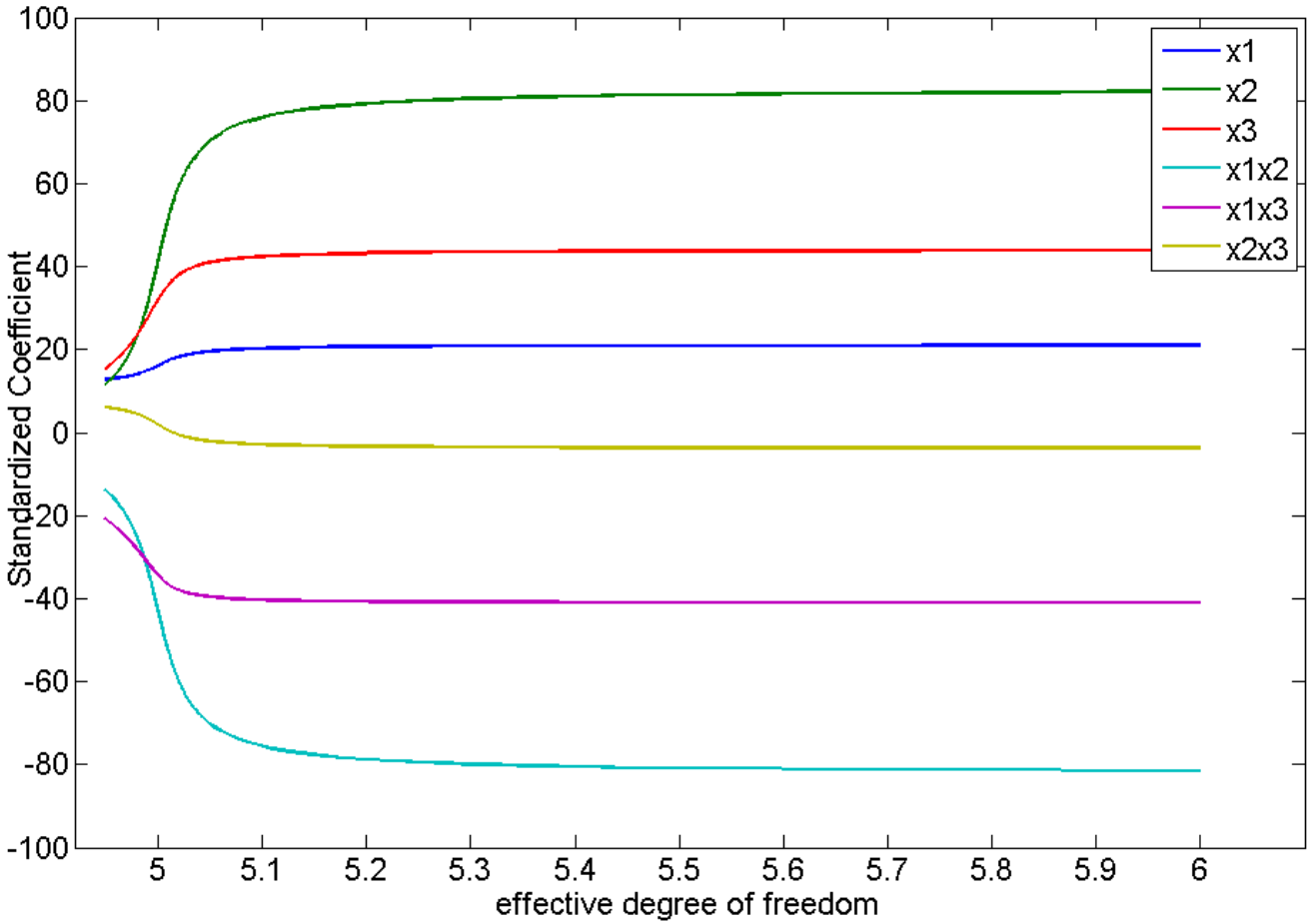
Choice of λ

- Selection of λ is to balance the variance versus bias, i.e. minimize MSE.
- Hoerl and Kennard (1970) recommended use of the ridge trace to graphically display all components of $\widehat{\boldsymbol{\beta}}^{ridge}$ against a range of values of λ.
  - As λ controls the amount of bias in the ridge estimator, the value of λ is estimated by the smallest value at which the trace stabilizes for all coefficients.
- Hastie, Tibshirani and Friedman consider $\widehat{\boldsymbol{\beta}}^{ridge} = W(\lambda)\widehat{\boldsymbol{\beta}}^{OLS}$ (what is $W(\lambda)$?) so that one can plot the components of $\widehat{\boldsymbol{\beta}}^{ridge}$ against the *effective degrees of freedom*

$$\text{df}(\lambda) = \text{tr}(W(\lambda)) = \sum_{j=1}^{r} \omega_j/(\omega_j + \lambda)$$

so that $W(\lambda)$ is a shrinking operator to shrink $\widehat{\boldsymbol{\beta}}^{OLS}$.


Ridge Trace
Standardized Coefficient
Ridge Parameter
x 10^-3
x1
x2
x3
x1x2
x1x3
x2x3
100
80
60
40
20
0
-20
-40
-60
-80
-100
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5


100
80
60
40
20
0
-20
-40
-60
-80
-100
Standardized Coefficient
5
5.1
5.2
5.3
5.4
5.5
5.6
5.7
5.8
5.9
6
effective degree of freedom
x1
x2
x3
x1x2
x1x3
x2x3

LASSO

- Tibshirani (1996) introduced L_1 penalty on the least squares error and solves the convex optimization problem

$$\widehat{\boldsymbol{\beta}}^{lasso} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\lambda_j|$$

  - LASSO put the unimportant variables exactly zero rather than moving them closer to zero as with RR since the contours of L_1 penalty is lozenge-shaped
  - $\widehat{\boldsymbol{\beta}}^{lasso} \to \boldsymbol{\beta}^{ols}$ as $\lambda \to 0$
  - From Bayesian point view, L_1 penalty put a double exponential prior with parameter λ on $\boldsymbol{\beta}$ and as λ increases, the prior put more of its mass near zero
  - LASSO combines variable selection with shrinkage on the regression function together, e.g. for orthogonal design

$$\widehat{\beta}_j^{lasso} = \text{sign}(\widehat{\beta}_j^{ols})(|\widehat{\beta}_j^{ols}| - 2\lambda)_+$$

while

$$\widehat{\beta}_j^{ridge} = (1+\lambda)^{-1}\widehat{\beta}_j^{ols}$$

- Unlike OLS or ridge, there is no theoretical solution to get standard error for $\widehat{\boldsymbol{\beta}}^{lasso}$ even for normal data. It can be however derived using bootstrap or CV.
- Post selection inference has been studied very recently by Tibshirani et al. (2014), Zhang and Zhang (2014); pre-selection inference is unknown.
- LASSO is NOT oracle: it is not always consistent and tends to select over parameterized models unless certain irrepresentable condition satisfied (Zhao and Yu (2006))
- In practice, LASSO always outperform AIC, BIC, subset selection and RR in predictive modeling (in the sense of predictive errors), which is due to large reduction on variance
- LASSO, however, perform poorly if
  1. true model is not sparse
  2. $|\{\widehat{\beta}_j^{lasso} \neq 0\}| \leq n$
  3. LASSO tends to pick random variables if they are highly correlated, and not consistent; while ridge performs better
  4. LASSO is very sensitive to the choice of λ: if CV is used, too many variables may be included and the bias can be very high

Bridge regression

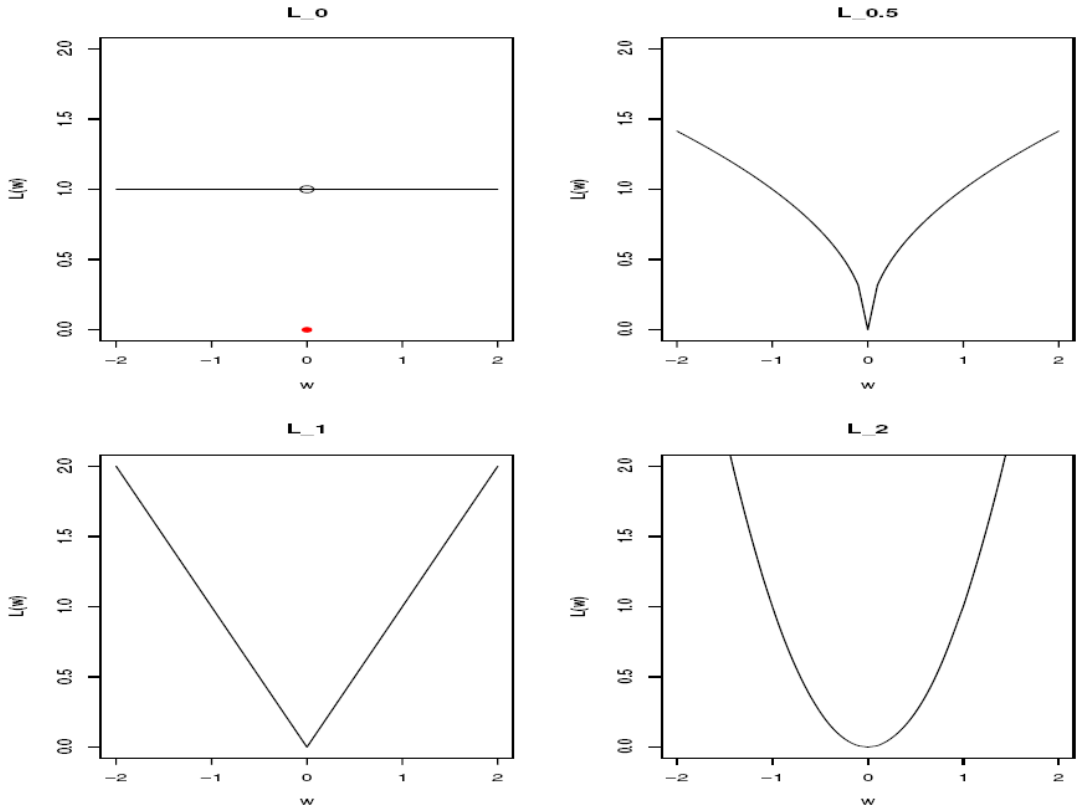
- For $r \geq 0$, Frank and Friedman (1993) suggested

$$\widehat{\boldsymbol{\beta}}^{bridge} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^r$$

  - $r = 1, 2$ correspond to LASSO and ridge
  - $r = 0$ is $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ and called hard thresholding (ideal but NP hard: nonconvex and discontinuous). For orthogonal design matrix

$$\widehat{\beta}_j^0 = \text{sign}(\widehat{\beta}_j^{ols}) \mathbb{I}(|\widehat{\beta}_j^{ols}| > \sqrt{\lambda})$$

  - $r \in (0, 1]$ is called soft thresholding
  - $r = \infty$ corresponds to $\max_j |\beta_j|$



- Bridge procedures are consistent and have asymptotic normality under some conditions (Knight and Fu (2000)).

Other methods

- Adaptive LASSO (Zou (2006)), permits different weights for different parameters

$$\widehat{\boldsymbol{\beta}}^{alasso} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} w_j |\beta_j|$$

  where $w_j = |\tilde{\beta}_j|^{-\gamma}$ with $\tilde{\boldsymbol{\beta}}$ a root-n consistent estimator of $\boldsymbol{\beta}$.
  - ALasso is near-minimax optimal
  - ALasso is oracle

- Grouped LASSO (Yuan and Lin (2007)), partition p variables into J mutually disjoint groups so $\boldsymbol{y} = \sum_{j=1}^{J} \boldsymbol{Z}_j \boldsymbol{\beta}_j + \boldsymbol{\varepsilon}$ where $\boldsymbol{Z}_j$ are $n \times p_j$ matrices. Then

$$\widehat{\boldsymbol{\beta}}^{glasso} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{Z}_j \boldsymbol{\beta}_j)'(\boldsymbol{y} - \sum_{j=1}^{J} \boldsymbol{Z}_j \boldsymbol{\beta}_j) + \lambda \sum_{j=1}^{J} p_j (\boldsymbol{\beta}_j' \boldsymbol{\beta}_j)^{1/2}$$

  - Outperform stepwise selection in factor selection problems

- Elastic net (Zou and Hastie (2005)), combine the benefits of both L_1 and L_2 regularizations and simultaneously ensure that related X_js get comparable sized coefficients.

$$\widehat{\boldsymbol{\beta}}^{enet} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2$$

  - To correct extra bias, the elastic net coefficient is defined as a rescaled one by $(1 + \lambda_2)\widehat{\boldsymbol{\beta}}^{enet}$
  - Elastic net has automatic grouping effect that the highly correlated variables tend to have similar coefficient estimates

- Consider generalized additive model $y = f(\boldsymbol{x}) = b_0 + \sum_{j=1}^{p} f_j(x_j) + \sum_{j<\ell} f_{jk}(x_j, x_\ell) + \ldots + \epsilon$, Gu (2002) proposed multidimensional smoothing splines by minimizing

$$n^{-1} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \sum_{j=1}^{p} \theta_j^{-1} ||\boldsymbol{P}_j f||_{\mathcal{H}}^2$$

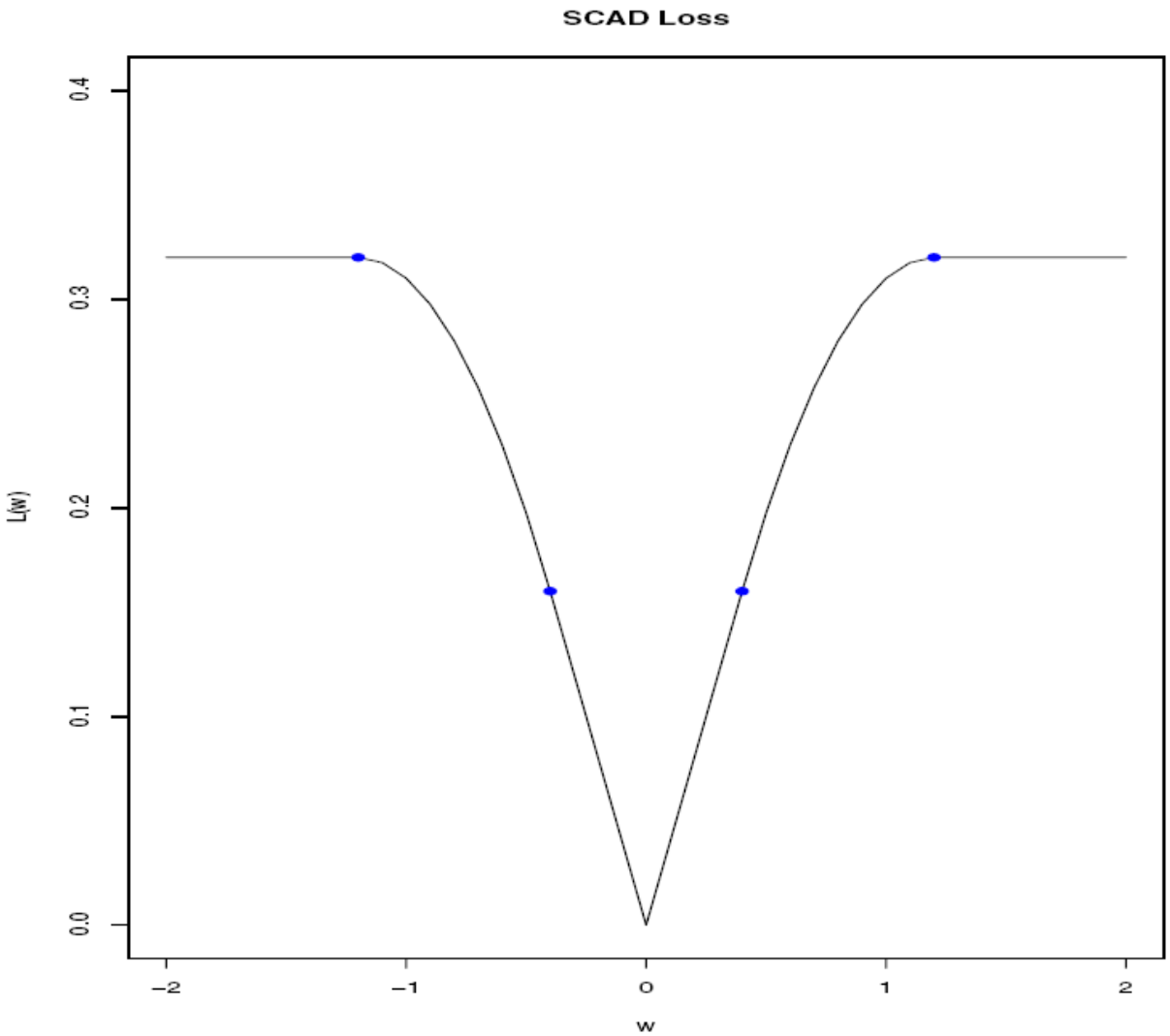
where $\boldsymbol{P}_j f$ is the orthogonal projection of f onto subspace $\mathcal{H}_j$ of function space.

Smooth Clipped Absolute Deviation (SCAD)

- An ideal penalty function should lead to a sparse, nearly unbiased estimator that is continuous in data and converges rapidly.
  - Ridge is not sparse, LASSO can be biased, L_0 is not fast.
- Fan and Li (2001) proposed a penalty called SCAD, that satisfies all the desired conditions. SCAD penalty is

$$q_\lambda(|\beta|) = \begin{cases} \lambda|\beta| & \text{if } |\beta| \leq \lambda \\ -\frac{(|\beta|^2 - 2a\lambda|\beta| + \lambda^2)}{2(a-1)} & \text{if } \lambda < |\beta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2} & \text{if } |\beta| > a\lambda \end{cases}$$

  - $q_\lambda(|\beta|)$ is a symmetric quadratic spline with knots at λ and $a\lambda$
  - It has continuous first order derivative


SCAD Loss
L(w)
0.0
0.1
0.2
0.3
0.4
−2
−1
0
1
2
w

- The SCAD estimator solves

$$\widehat{\boldsymbol{\beta}}^{scad} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} q_\lambda(|\beta_j|)$$

- SCAD is oracle
- SCAD has bias decreasing fast (faster than ALasso)
- SCAD is non-convex optimization problems. Some algorithms:

Bayesian Model Averaging

- Bayesian inference in two sentences: Let $\boldsymbol{\theta}$ be parameters, $\boldsymbol{y}$ be data, $p(\boldsymbol{y}|\boldsymbol{\theta})$ be the likelihood and $p(\boldsymbol{\theta})$ be your subjective belief ("the prior") about $\boldsymbol{\theta}$ before observing the data.
- Then $p(\boldsymbol{\theta}|\boldsymbol{y})$ ("the posterior") represents your subjective belief about $\boldsymbol{\theta}$ after incorporating the evidence from the data, where

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- Straightforward application of Bayes' Theorem if you are willing to treat parameters as random and inference as subjective.

Bayesian Model Averaging

- Let Δ=quantity of interest, and suppose K models (M_k for $k = 1, \ldots, K$).
- Weighted average of posterior distribution of Δ under each model, weighted by posterior model probabilities:

$$p(\Delta|\boldsymbol{y}) = \sum_{k=1}^{K} p(\Delta|M_k, \boldsymbol{y}) p(M_k|\boldsymbol{y})$$

- Recall $p(M_k|\boldsymbol{y}) \propto p(\boldsymbol{y}|M_k)p(M_k)$
  - $p(M_k)$ is tricky.