Model Selection

Yuhong Yang
School of Statistics
University of Minnesota

Outline

- Model selection is almost everywhere in statistical applications
- The different approaches to model selection
- Theoretical properties of model selection methods
- A proper use of cross-validation

Model selection diagnostics

The science community is facing the irreproducibility crisis! Model selection uncertainty surely contributes to the problem.

Model selection diagnostic tools should be built to properly assess reliability of the outcomes of model selection methods.

When model selection uncertainty is high, we may choose model averaging/combination instead.

Some examples

1. Regression

$$Y_i = f(x_i) + \varepsilon_i, 1 \leq i \leq n,$$

Possible candidate models to consider:

- linear models
- nonlinear models
- splines or wavelets
- various dimension reduction methods (e.g., neural nets, single-index model, SIR)
- local polynomial models
- deep learning with different network structures

One may compare the different types of models, and/or different specifications within the same type.

2. ANOVA: J factors, m replicates

$$Y_{i_1,\ldots,i_J,k} = \mu_{i_1,\ldots,i_J} + \varepsilon_{i_1,\ldots,i_J,k}, 1 \leq k \leq m,$$

Consider different models on $\mu_{i_1,\ldots,i_J}$ for estimating μ.

3. Time series forecasting

AR(1): $Y_i = \phi Y_{i-1} + e_i$, $i = 1, 2, ...$

More generally, consider $ARMA(p, q)$ model: $\phi(B)Y_i = \theta(B)e_i$, where B is the back-shift operator, $\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$.

4. Classification: Observe (X_i, Y_i), $i = 1, ..., n$ with $Y_i \in \{0, 1\}$

Candidates:

- LDA
- nearest neighbor rules
- neural networks
- classification trees
- margin based methods (SVM, ψ-learning,...)

A simple starting example: Effects of smoking

- A 20-year follow-up study to understand the effects of smoking on survival (see, Appleton, French and Vanderpump, 1996; Faraway, 2006)
- 1314 women were included in this analysis

	dead	survival	smoker	age
1	2	53	yes	18-24
2	1	61	no	18-24
3	3	121	yes	25-34
4	5	152	no	25-34
5	14	95	yes	35-44
6	7	114	no	35-44
7	27	103	yes	45-54
8	12	66	no	45-54

9	51	64	yes	55-64
10	40	81	no	55-64
11	29	7	yes	65-74
12	101	28	no	65-74
13	13	0	yes	75+
14	64	0	no	75+

Proportion of subjects dead at the end of the study:

```
                dead
smoker       yes        no
   yes   0.2388316 0.7611684
   no    0.3142077 0.6857923
```

What is going on?

Searching for a good model:

```
> m<-glm(cbind(survival,dead)~smoker*age,smoke,family=binomial)
> drop1(m,test="Chi")

Model:
cbind(survival, dead) ~ smoker * age
           Df Deviance    AIC    LRT Pr(Chi)
<none>        5.27e-10 74.996
smoker:age  6    2.381 65.377  2.381  0.8815
```

```
> m1<-glm(cbind(survival,dead)~smoker+age,smoke,family=binomial)
> drop1(m1,test="Chi")

Model:
cbind(survival, dead) ~ smoker + age
       Df Deviance    AIC    LRT Pr(Chi)
<none>        2.38  65.38
smoker  1     8.33  69.32   5.95 0.01475
age     6   632.30 683.29 629.92 < 2e-16
---
```

Proportion of subjects dead at the end of the study for each age group:

```
      dead (age=="18-24")
smoker        yes         no
   yes 0.03636364 0.96363636
   no  0.01612903 0.98387097
```

```
      dead (age=="25-34")
smoker        yes         no
   yes 0.02419355 0.97580645
   no  0.03184713 0.96815287
```

```
      dead (age=="35-44")
smoker        yes         no
   yes 0.12844037 0.87155963
   no  0.05785124 0.94214876
```

```
      dead (age=="45-54")
smoker       yes        no
   yes 0.2076923 0.7923077
   no  0.1538462 0.8461538
```

```
      dead (age=="55-64")
smoker       yes        no
   yes 0.4434783 0.5565217
   no  0.3305785 0.6694215
```

```
      dead (age=="65-74")
smoker       yes        no
   yes 0.8055556 0.1944444
   no  0.7829457 0.2170543
```

```
      dead (age=="75+")
smoker yes no
   yes   1  0
   no    1  0
```

Comments on the example

- The sequential test-based aproach to compare models worked very well in this case
- The Neyman-Pearson testing framework focuses on control of probability of type I error, which is great for confirmatory analysis
- When there are many factors, things become more complicated with the hypothesis testing approach
  - the multiple testing issue
  - dealing with non-nested models
  - controlling probability of type I error may not be the goal of the analysis
- General model selection methods that target specific objectives are desired

Why do not we fit a large model to be safe?

- Philosophic perspective: "entities must not be multiplied beyond necessity" (Occam's razor)
- Statistical perspective: A larger model fits better, but it does not necessarily behave better
  - It is difficult to interpret a large model
  - The large model may be very poor due to estimating many parameters
- Even if we know the true model, it may not be the best to use

There are different goals for model selection or comparison

- Find a parsimonious model for insight and interpretation
- Confirm a scientific understanding
- Estimate the underlying probability distribution or important quantities
- Predict future

It would be great if all these goals can be served at the same time. Unfortunately, that is not possible!

Model comparison methods

- graphical and diagnostic methods
- tests
- information criteria
  - AIC (Akaike, 1973)
  - BIC (Schwarz, 1978)
  - MDL (Rissanen, 1978)
- LASSO (Tibshirani, 96), Adaptive LASSO (Zou, 07), SCAD (Fan and Li, 01), MCP (Zhang, 10)
- cross validation (or related methods such as GCV) (e.g., Allen, 1974; Stone, 1974; Geisser, 1975)
- bootstrap-based methods

Now let's focus on regression models.

Suppose the data genearing process is:

$$Y_i = f(X_i) + \varepsilon_i, \quad 1 \leq i \leq n,$$

- $(X_i, Y_i)_{i=1}^n$ are independent observations with X_i being d-dimensional
- f is the true regression function
- ε_i are the random errors with $E(\varepsilon_i | X_i) = 0$

AIC: Seek the model that minimizes a criterion that is based on Kullback-Leibler divergence

- General form:

$$\text{-2loglikelihood}_k + 2m_k,$$

  where k is the model index and m_k is the number of parameters in the model

- In normal linear regression: $n \log \hat{\sigma}_k^2 + 2m_k$
- If σ^2 is known: $RSS_k + 2m_k\sigma^2$
- AICc: $n \log \hat{\sigma}_k^2 + \frac{n(n+m_k)}{n-m_k+2}$

BIC:

- General form:

$$\text{-2loglikelihood}_k + m_k \log n$$

- In normal linear regression: $n \log \hat{\sigma}_k^2 + m_k \log n$

- If σ^2 is known: $RSS_k + m_k \sigma^2 \log n$

CV:

- Leave-one-out or delete-one:

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_{i(i),k})^2$$

- delete-d

- v-fold

Methods based on penalization in terms of the coefficients

- LASSO (l_1 penalty)

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- adaptive LASSO

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - ... - \beta_p x_{ip})^2 + \lambda \sum_{j=1}^{p} |\beta_j / \hat{\beta}_j|$$

- SCAD and MCP (continuous but non-differentiable penalty functions)

- Non-differentiability of the penalty function forces some coefficients to be exactly zero

Two scenarios

- **Parametric scenario**: the regression function is assumed to be linear in the predictors (with possible transformations).
- **Nonparametric scenario**: the regression function is infinite-dimensional, but can be approximated by linear models.
- The behaviors of model selection methods depend on which scenario is proper for the data at hand.

Many nice theoretical results have been obtained on information criteria

- consistency: *Have you found the right model?*
- asymptotic efficiency: *Does your model perform as well as the best one in estimation?*
- risk bounds for minimax optimaility

Consider a parametric scenario.

- Let k^* be the true model. A model selection method δ is said to be consistent if the resulting selected model $\widehat{k}_n = \widehat{k}_n(\delta)$ satisfies

$$P\left(\widehat{k}_n = k^*\right) \to 1 \quad \text{when } n \to \infty.$$

- A model k is said to be a correct model if k^* is a sub-model of k
- Let Γ_c be the set of all correct models and let Γ_r be the set of all other models
- The model selection method is said to be under-fitting if

$$P\left(\widehat{k}_n \in \Gamma_r\right) \text{ does not go to } 0 \quad \text{when } n \to \infty.$$

- The model selection method is said to be over-fitting if

$$P\left(\widehat{k}_n \in \Gamma_c - \{k^*\}\right) \text{ does not go to } 0 \quad \text{when } n \to \infty.$$

- The familiar model selection methods (including AIC and BIC) do not have under-fitting problem in the above sense and over-fitting is the main issue.

Asymptotic efficiency:

Consider a countable list of models $k \in \Gamma$. A model selection method δ is said to be asymptotically efficient if the resulting selected model $\widehat{k}_n = \widehat{k}_n(\delta)$ satisfies

$$\frac{\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \widehat{f}_{\widehat{k}_n}(X_i)\right)^2}{\inf_k \frac{1}{n}\sum_{i=1}^{n}\left(f(X_i) - \widehat{f}_k(X_i)\right)^2} \to 1 \quad \text{in probability.}$$

The concept makes sense for both parametric and nonparametric scenarios.

In the parametric scenario, if a model selection method is consistent, and the true model is the best among the candidates, then the selected model is asymptotically efficient (also called asymptotically optimal).

A heuristic understanding of AIC (C_p)

For simplicity, assume σ^2 is known.

Let

- $Y^n = (Y_1, Y_2, ..., Y_n)^T$, $f_n = (f(X_1), ..., f(X_n))^T$, $e_n = (\epsilon_1, ..., \epsilon_n)^T$
- $\hat{Y}_k$ be the projection of Y^n into the space spanned by the columns of the design matrix of model k
- projection matrix M_k
- $\bar{f}_k = M_k f_n$ be the projection of f_n into the column space of the design matrix
- The RSS $\|Y^n - \hat{Y}_k\|^2$ is not the right target
- Instead consider the squared error $\|f_n - \hat{Y}_k\|^2$

One can easily show that $\|Y_n - \hat{Y}_k\|^2$ equals

$$\|f_n - \hat{Y}_k\|^2 - 2m_k\sigma^2 + e_n^{'}e_n + 2e_n^{'}(f_n - M_k f_n) + 2(m_k\sigma^2 - e_n^{'}M_k e_n)$$

- The remainder terms $e_n^{'}(f_n - M_k f_n)$ and $(m_k\sigma^2 - e_n^{'}M_k e_n)$ are hopefully asymptotically negligible
- $e_n^{'}e_n$ can be ignored because it is the same for all the models
- $\|Y_n - \hat{Y}_k\|^2 - e_n^{'}e_n$ substantially under-estimates $\|f_n - \hat{Y}_k\|^2$
- $2m_k\sigma^2$ is the bias-correction term
- AIC (C_p) of the form $\|Y_n - \hat{Y}_k\|^2 + 2m_k\sigma^2 - n\sigma^2$ provides an unbiased risk estimate

- If we can show that the remainder terms are uniformly negligible compared to $\|f_n - \hat{Y}_k\|^2$, then we can establish asymptotic efficiency and also minimax rate optimality.

  - the true model does not belong to the set of candidates
  - there cannot be too many models

- In the parametric case, for a correct model k, $\|Y_n - \hat{Y}_{k^*}\|^2 - \|Y_n - \hat{Y}_k\|^2$ is of order 1 and thus a constant penalty term (e.g., $2m_k\sigma^2$) is not sufficient to prevent over-fitting.

- A sufficient and necessary condition for consistency is basically that the penalty $p_{k,n}$ in $\|Y_n - \hat{Y}_k\|^2 + p_{k,n}\sigma^2$ goes to ∞ as $n \to \infty$ and $p_{k,n}/n \to 0$.

- In general, AIC tries to provide an asymptotically unbiased estimate of the Kullback-Leibler divergence between the true underlying density of the data and the estimate from a model. Asymptotic normality of the MLE is used in the derivation.

- BIC tries to select the model with highest posterior probability.
  - Assign prior probability on each model
  - Assign a prior distribution on the parameters in each model
  - Use Laplace approximation of the posterior probability and ignore small order terms

- BIC is consistent if the true model stays fixed and is among the candidates and BIC is also asymptotically efficient in this case
- AIC is asymptotically efficient when the true model is infinite-dimensional
- Delete-one CV behaves like AIC and delete-d CV with $d/n \to 1$ behaves like BIC

Conflicting properties of AIC and BIC

Consider a traditional situation of using a sequence of nested models for regression modeling.

- When one of the candidate models holds, BIC is consistent in selection, but AIC is not. BIC is also asymptotically efficient, but AIC is not.
- When none of the candidate models holds, AIC is asymptotically efficient, but BIC is not.
- In the parametric case, BIC is pointwise-risk adaptive, but AIC is minimax-rate adaptive. The price of being consistent is the inflation of worse-case risk by a factor of

$$\log(1/\text{prob. of over fitting}).$$

AIC or BIC?

- Some researchers rule out BIC based on the argument that parametric models cannot be right.
- A popular saying is that: Use AIC for a nonparametric situation and use BIC for a parametric situation.
- A more accurate statement is: Use AIC for a *practically* nonparametric situation and use BIC for a *practically* parametric situation.
- A parametricness index can help us decide which scenario we are in (Liu and Yang, 2012; Ding, Tarokh, and Yang, 2017+).

Stopping the war between AIC and BIC

Can we design a super criterion to resolve the conflicts?

- Asymptotic efficiency of BIC and AIC shown for different situations can be shared (Yang, 2007; Ing, 2007; Liu and Yang, 2012; Zhang and Yang, 2015; Ding, Tarokh, and Yang, 2017+)
  - Based on choosing between AIC and BIC;
  - Or using a parametricness index
- Consistency of BIC and minimax-rate optimality of AIC cannot be resolved (Yang, 2005). See also Erven, Grunwald and de Rooij (2012).

Many issues are subtle. A bottom line is: You may not hit two hoops in one shoot!

Minimax perspective

Consistency and asymptotic efficiency focus on asymptotic behaviors as the sample size tends to infinity.

Minimax view is different: It emphasizes a uniform performance. Minimax rates of convergence play a central role in statistical and machine learning theory.

Let's understand some basics of minimax theory.

An example of oracle inequalities or index of resolvability bounds:

For AIC, when the number of models of each dimension does not grow exponentially fast, we have

$$\frac{1}{n}\sum_{i=1}^{n} Ef(X_i) - \widehat{f}_{\widehat{k}_n}(X_i))^2 \leq C \inf_k \left(\frac{1}{n}\sum_{i=1}^{n} E\left(f(X_i) - \widehat{f}_k(X_i)\right)^2 \right),$$

where the constant C can be made 1 in various situations with an additional term in the risk bound.

It has minimax implications.