Section 9

Frequentist properties of Bayesian methods

Calibrated Bayes

- So far we have discussed Bayesian methods as being separate from the frequentist approach
- However, in many cases methods with frequentist properties are desirable
- For example, we may want a method with Type I error control or 80% power
- You can design Bayesian methods to achieve these frequentist properties
- In this view, Bayesian methods generate procedures/algorithms for further study
- Often Bayesian methods are very competitive with frequentist methods using frequentist criteria

Outline

These notes cover Chapter 7

- Decision theory
- Bias-variance tradeoff
- Asymptotics
- Simulation studies

Should Bayesians care about frequentist properties?

What if a Bayesian weather forecaster made a 95% prediction interval for temperature every day for a year but the interval only included the actual temperature 40% of the time?

Little in *Little, 2011, Stat Sci*

- Bayesian statistics is strong for inference under an assumed model, but relatively weak for the development and assessment of models
- Frequentist statistics provides useful tools for model development and assessment, but has weaknesses for inference under an assumed model
- If this summary is accepted, then the natural compromise is to use frequentist methods for model development and assessment, and Bayesian methods for inference under a model
- This capitalizes on the strengths of both paradigms, and is the essence of the approach known as Calibrated Bayes

Rubin in *Little, 2011, Stat Sci*

- The applied statistician should be Bayesian in principle and calibrated to the real world in practice - appropriate frequency calculations help to define such a tie
- Frequency calculations are useful for making Bayesian statements scientific, scientific in the sense of capable of being shown wrong by empirical test
- Here the technique is the calibration of Bayesian probabilities to the frequencies of actual events

Bayes as a procedure generator

- A Bayesian analysis produces a posterior distribution which summarize our uncertainty after observing the data
- However, if you have to give a one-number summary as an estimate you might pick the posterior mean

$$\hat{\theta}_B = \mathrm{E}(\theta|\mathbf{Y})$$

- This estimator $\hat{\theta}_B$ can be evaluated along with MLE or method of moments estimators
- Is it biased? Consistent? How does its MSE compare with the MLE?
- These are all frequentist properties of the Bayesian estimator

Bayes as a procedure generator

- Similarly, if we have to give an interval estimate, we might use the 95% posterior credible set
- In practice, this interval is motivated by the one data set we observed
- But we could view this as a procedure for constructing an interval and inspect its frequentist properties
- If we analyzed many datasets, each time computing a 95% posterior interval, how many would contain the true value?
- A Bayes test is to reject H_o if Prob$(H_o|\mathbf{Y}) < c$
- What are the Type I and Type II errors of this test?
- Can we pick the threshold c to control Type I error?

Bayesian decision theory

- Before studying the frequentist properties of Bayesian estimtors and hypothesis tests, we should determine the "best" Bayesian method
- For example, should we take the estimator to be the posterior mean, median, or mode?
- Defining "best" requires a scoring system
- We call this the loss function $l(\hat{\theta}, \theta)$
- Squared error loss is $l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$
- Absolute loss is $l(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$

Bayesian decision theory

- The summary of the posterior that minimizes the expected (posterior) loss is the **Bayes rule.**
- Squared error loss implies we should use the posterior mean for $\hat{\theta}$
- Absolute loss implies we should use the posterior median for $\hat{\theta}$
- Hypothesis test requires are more complicated loss function
- For proofs see the online derivations

Bias/variance trade-off

- Assume $Y_1, ..., Y_n \sim \text{Normal}(\mu, \sigma^2)$
- Estimator 1: $\hat{\mu}_1 = \bar{Y}$
- Estimator 2: $\hat{\mu}_2 = c\bar{Y}$ where $c = \frac{n}{n+m}$
- $\hat{\mu}_2$ is the posterior mean under prior $\mu \sim \text{Normal}(0, \frac{\sigma^2}{m})$
- Compute the bias and variance of each estimator
- Compute the mean squared error (recall MSE = bias^2+variance)
- Which estimator is preferred?

Properties of Bayesian estimators

Broadly speaking, the following comparisons between Bayes and MLE hold:

- Bayesian estimators have smaller standard errors because the prior adds information
- Bayesian estimators are biased if the prior is not centered on the truth
- Depending on this bias/variance trade-off, Bayes estimators may have smaller MSE than the MLE
- If the prior is weak the methods are similar
- For any prior that does not depend on the sample size, as n increases the prior is overwhelmed by the likelihood and the posterior approaches the MLE's sampling distribution

Bayesian central limit theorem

- Assumptions:
  - the usual MLE conditions on the likelihood
  - the prior does not depend on n and puts non-zero probability on the true value θ_0
- Then
  $$p(\theta|\mathbf{Y}) \to \mathrm{N}\left[\theta_0, I(\theta_0)^{-1}\right]$$
  where I is the information matrix
- Therefore, for large datasets the posterior is approximately normal
- Bayes methods are asymptotically unbiased

Bayesian central limit theorem

- This implies that Bayes and MLE will be equivalent in large samples
- What a relief!
- However, the interpretation is different
- We can use the Bayesian interpretation like Prob($\mathcal{H}_0|\mathbf{Y}$) and Prob($3.4 < \theta < 5.6$)
- The Bayesian CLT gives a way to approximate ($n \to \infty$) the posterior without MCMC
- Most still use MCMC with the hope that it better approximates ($S \to \infty$) the exact posterior
- The CLT is useful for initial values and tuning

Methods for studying frequentist properties

- Theoretical studies of Bayesian estimators use the same basic approaches as frequentist methods
- Theorems and proofs (of consistency etc.) are ideal
- When the math is intractable, simulation studies are used
- In a simulation study you generate many datasets with known parameters values
- You apply the Bayesian method to each dataset (so you may have to run MCMC several times)
- You then see how you did, e.g., what proportion of the 95% credible sets included the true value?
- The course website has code for a simulation study of the Bayesian LASSO regression (BLR)

Methods for studying frequentist properties

			MSE		Coverage	
n	p_0	p_1	OLS	BLR	OLS	BLR
40	20	0	5.40	0.03	94.7	100.0
	15	5	5.71	3.45	93.8	96.0
	0	20	5.40	9.47	93.7	91.6
100	20	0	1.17	0.02	95.8	100.0
	15	5	1.27	0.98	94.5	95.5
	0	20	1.22	1.26	96.0	95.6

- n is the sample size
- p_0 is the number of null covariates with $\beta_j = 0$
- p_1 is the number of non-null covariates with $\beta_j \neq 0$

Methods for studying frequentist properties

Conclusions:

- When the model is sparse (p_1 is small), BLR is has much smaller MSE than OLS
- When the model is dense (p_0 is small), OLS has smaller MSE, but for large n the methods are similar
- Both methods generally have reasonable coverage
- BLR's coverage is low when n is small and the model is dense, i.e., when its assumptions are grossly violated