Frequentist properties of Bayesian methods

Patrick Breheny

February 14

Introduction

- Today's lecture is a brief departure from our Bayesian paradigm
- If an unobservable parameter θ truly is random, then using Bayes rule to obtain a posterior is an unavoidable mathematical fact; anything else is incoherent
- However, even if we don't believe in θ being random, we may still be interested in using Bayesian methods since they usually prove to have good frequentist properties as well

An informal Bernstein-von Mises Theorem

- To begin with, we demonstrate that, given the same likelihood, the Bayesian and frequentist answers approach equivalency in an asymptotic sense (as $n \to \infty$)
- **Theorem:** Suppose $Y_1, Y_2, \ldots | \theta \sim p(y|\theta_0)$ and that our prior places positive density in a neighborhood surrounding θ_0. Then, assuming the same regularity conditions that are required for asymptotic likelihood theory, we have that

  $$\theta | \mathbf{y} \dot{\sim} N(\theta_0, \mathcal{I}(\theta_0)^{-1}),$$

  where $\mathcal{I}$ is the Fisher information

Remarks

- Note, however, the difference between the result on the previous slide and likelihood theory result: the previous slide describes the posterior distribution of θ, while the likelihood theory result describes the sampling distribution of $\hat{\theta}$
- Note that the posterior distribution is somewhat more complicated than a sampling distribution, in that it is a conditional, and hence stochastic, distribution
- For this reason, the theorem on the previous slide is intentionally a bit loose in its convergence statement
- It can, however, be made more rigorous, as well as extended to the case of multivariate $\boldsymbol{\theta}$; the result is known as the *Bernstein-von Mises theorem*

Implications

The Bernstein-von Mises theorem has a number of powerful implications:

- Bayesian methods are consistent: Letting B denote a ball – of any radius, no matter how small – encompassing θ_0, the posterior probability that $\theta \in B$ will always go to 1 as $n \to \infty$ (this is sometimes referred to as *concentration of the posterior*)
- Bayesian posteriors are asymptotically normal: distinctions between posterior modes, means, medians, central intervals and HPD intervals all become irrelevant as n grows large
- Inferences from each paradigm will eventually become equivalent: Not merely will both frequentist and Bayesian procedures converge to the truth, but confidence intervals will eventually coincide with posterior intervals

Likelihood-based versus procedural methods

- Thus, for all the fundamental philosophical differences between Bayesian and frequentist methods, they actually produce pretty similar results given enough data
- However, this conclusion only applies to parametric models with fully specified likelihoods
- A number of frequentist methods are nonparametric, and do not necessarily specify any sort of likelihood or model for the data (*e.g.*, Wilcoxon rank-sum tests, classification trees); our textbook calls these approaches "procedural", as opposed to model-based
- There is such a thing as "Bayesian nonparametrics", although it is (a) quite a bit different, conceptually, from a Wilcoxon rank-sum test, and (b) fairly advanced and beyond the scope of this course (although see section 11.8 if you are interested)

Caveats

So to summarize, Bayesian and frequentist methods often produce similar conclusions, with the following caveats:

- The frequentist approach permits the use of likelihood-free procedures like permutation tests that have no real Bayesian analogue
- As we have remarked previously, there is typically no direct Bayesian analogue to the p-value, and even when there is (*i.e.*, with a mixture prior), there is no guarantee of agreement
- Agreement is only guaranteed for large sample sizes

Introduction

- To follow up on the final caveat, we now look at a few examples involving small/finite sample sizes
- As we will see, Bayesian methods typically have satisfactory small-sample performance – indeed, often superior to that of likelihood-based alternatives

Regression

- Suppose that we fit a linear regression model with the following prior on $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{0}, \omega_0 \mathbf{I});$$

  let $\widehat{\boldsymbol{\beta}}^{\text{Bayes}}$ denote the posterior mean
- **Theorem:** There always exists a value of ω_0 such that the MSE of $\widehat{\boldsymbol{\beta}}^{\text{Bayes}}$ is less than the MSE of $\widehat{\boldsymbol{\beta}}^{\text{OLS}}$

The "many normal means" problem

- A related problem is the following: Suppose $Y_{ij} \sim \mathrm{N}(\theta_i, \sigma^2)$ and we are interested in estimating $\boldsymbol{\theta}$
- The obvious estimator is $\bar{\mathbf{y}}$, the observed means
- However, the theorem on the previous slide implies that we can always choose a prior $\theta_i \sim \mathrm{N}(0, \omega_0^{-1})$ such that the estimator

$$\hat{\theta}_i = \frac{\bar{y}_i}{1+\lambda},$$

where $\lambda = \omega_0 \sigma^2 / n_i$, has a lower MSE than $\bar{y}_i$

Shrinkage

- As remarked earlier, typically in ridge regression we do not penalize the intercept; this leads to the estimator

$$\hat{\theta}_i = \bar{y} + \frac{\bar{y}_i - \bar{y}}{1 + \lambda},$$

  where $\bar{y}$ is the overall ("grand") mean; this estimator can also be shown to be superior to $\bar{y}$ for a certain range of λ values
- In words, we can always obtain superior estimation accuracy by shrinking individual means towards the common mean

The James-Stein estimator

- An even more remarkable result was shown by Charles Stein and Willard James, who derived an empirical choice for λ
- Letting $\hat{\boldsymbol{\theta}}^{\mathrm{JS}}$ denote this estimator, James & Stein showed that $\hat{\boldsymbol{\theta}}^{\mathrm{JS}}$ uniformly dominates $\bar{\mathbf{y}}$ in terms of MSE (*i.e.*, has a lower MSE for all values of $\boldsymbol{\theta}_0$
- In the case where all samples have the same number of observations n, the James-Stein shrinkage factor is $(p-3)/(np-p)$:

n	p	Shrinkage factor
6	5	0.08
20	5	0.02
2	100	0.97

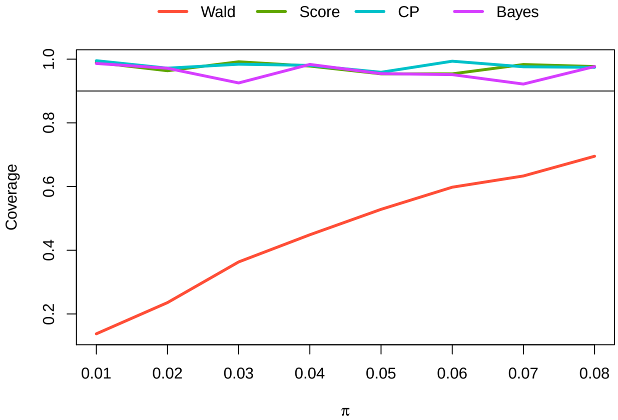
Empirical Bayes

- The James-Stein estimator is not a purely Bayesian approach, in that it uses the observed data to specify a prior (which is obviously not "prior")
- Instead, it falls under the category of what is known as *emprical Bayes*, which allows the use of data to specify what are considered to be nuisance parameters in priors, thereby in some sense combining ideas from frequentist and Bayesian analysis
- The advantage of these methods, of course, is that they are easy to apply and do not require one to think about priors; the disadvantage is that they treat estimates as known quantities in specifying priors, and thus ignore some sources of variability

Binomial coverage

- The previous examples have focused on estimation; we now turn provide an example dealing with coverage
- Consider the problem of obtaining a confidence interval for a binomial proportion
- What is the frequentist coverage of the Bayesian HPD interval? We will compare it with three frequentist methods: the Wald, Score, and Clopper-Pearson methods

Simulation results, $n = 15$



No method can achieve perfect coverage here, but the Bayes approach is generally closest to the nominal coverage of 90%

Final remarks

In summary,

- You don't necessarily have to believe in the Bayesian paradigm to employ a Bayesian analysis (and vice versa)
- With enough data, the two frameworks provide equivalent answers, and with smaller data sets, Bayes approaches can have attractive frequentist properties
- Furthermore, MCMC/BUGS often makes it easy to implement unconventional models and handle the complications of real data and inferences regarding functions of parameters