Bayesian Data Analysis Ch.4

Asymptotics and connections to non-Bayesian approaches

Kyeongwon Lee

January 21, 2019

Note

This slide is based on[1] the book "Bayesian Data Analysis" by Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin.

[1] It is more like plagiarism

1. Normal approximations to the posterior distribution Normal
2. Large-sample theory
3. Counterexamples to the theorems
4. Frequency evaluations of Bayesian inferences
5. Bayesian interpretations of other statistical methods

Introduction

- Many simple Bayesian analyses based on noninformative prior distributions give similar results to standard non-Bayesian approaches
  (For example, the posterior t interval for the normal mean with unknown variance)
- It is clear that as the sample size n increases, the influence of the prior distribution on posterior inferences decreases.

Normal approximations to the posterior distribution Normal

Normal approximation to the joint posterior distribution

- If the posterior $p(\theta|y)$ is unimodal and symmetric, It can be approximated to the normal distribution.
- i.e. the logarithm of the posterior density $\log p(\theta|y)$ is approximated by a quadratic function of θ.

사후분포의 근사

Let $\hat{\theta}$ be a posterior mode. If the log-posterior $logp(\theta|y)$ is differentiable, we can approximate it as follows from the fact $\log p(\hat{\theta}|y) = 0$.

$$\log p(\theta|y) = \log p(\hat{\theta}|y) + \frac{1}{2}(\theta - \hat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \cdots, \tag{4.1}$$

Normal approximation to the joint posterior distribution (cont.)

Theorem 1

From Eq. (4.1),

$$p(\theta|y) \approx N\left(\hat{\theta}, \left[I(\hat{\theta})\right]^{-1}\right), \tag{4.2}$$

where $I(\theta) = -\dfrac{d^2}{d\theta^2} \log p(\theta|y)$ *is an information matrix.*

Example 1 (Gelman et al. (2013) pp.84)

$$\log p(\mu, \log\sigma|y) = \textit{const.} - n\log\sigma - \frac{1}{2\sigma^2}\left((n-1)^2 s^2 + n(\bar{y}-\mu)^2\right)$$
$$\approx N\left(\begin{pmatrix}\bar{y}\\ \log\hat{\sigma}\end{pmatrix}, \begin{pmatrix}\hat{\sigma}^2/n & 0\\ 0 & 1/(2n)\end{pmatrix}\right).$$

Interpretation of the posterior density function relative to its maximum

- In addition to its direct use as an approximation, the multivariate normal distribution provides a benchmark for interpreting the posterior density function and contour plots.
- see Gelman et al. (2013) pp.85.

Summarizing posterior distributions by point estimates and standard errors

- If n is sufficiently large and Theorem 1 holds, $\hat{\theta} \pm 2I(\hat{\theta})$ is an approximated 95% interval for θ.
- Note the posterior distribution should be unimodal and symmetric for a quick approximation.
- If not, we can approximate the posterior distribution using proper transformation and delta method.

Data reduction and summary statistics

- From the normal approximation, we can summarize the posterior distribution using the posterior mode and information matrix.
- For example, in Section 5.5, each of a set of eight experiments is summarized by a point estimate and a standard error estimated from an earlier linear regression analysis.
- However, under the poor approximation, inferences based on approximated posterior can lead wrong results.

Lower-dimensional normal approximations

Note

- *For a finite sample size n, the normal approximation is typically more accurate for condiotional and marginal distributions of components of θ than for the full joint distribution.*
- *Note that if a joint distribution is multivariate normal, all its margins are normal, but the converse is not true. (see Lauritzen (2011) pp.8)*

If θ is high-dimensional, two situations commonly arise. (see Gelman et al. (2013) section 13.5)

1. The marginal distributions of many individual components of θ can be approximately normal.
2. θ can be partitioned into two subvectors, $\theta = (\theta_1,\ \theta_2)$, for which $p(\theta_2|y)$ is not necessarily close to normal, but $p(\theta_1|\theta_2, y)$ is, perhaps with mean and variance that are functions of θ_2.

Lower-dimensional normal approximations (cont.)

Example 2 (Gelman et al. (2013) pp.86)

We illustrate the normal approximation for the model and data from the bioassay experiment of Section 3.7. The sample size in this experiment is relatively small, only twenty animals in all, and we find that the normal approximation is close to the exact posterior distribution but with important differences.

Large-sample theory

Notation and mathematical setup

- The results apply most directly to observations $y_1, \cdots, y_n$ that are independent outcomes sampled from a common distribution, $f(y)$.
- Suppose the data are modeled by a parametric family, $p(y|\theta)$, with a prior distribution $p(\theta)$.
- If the true data distribution is included in the parametric family[2] then, in addition to asymptotic normality, the property of consistency[3] holds.
- When the true distribution is not included in the parametric family, there is no longer a true value θ_0, but its role in the theoretical result is replaced by a value θ_0 that makes the model distribution, $p(y|\theta_0)$, closest to the true distribution in a technical sense involving Kullback-Leibler divergence, as is explained in Appendix B.

[2] if $f(y) = p(y|\theta_0)$ for some θ_0.

[3] the posterior distribution converges to a point mass at the true parameter value, θ_0, as $n \to \infty$

Asymptotic normality and consistency

Theorem 2

Let $J(\theta)$ be an Fisher information in (2.20). Under regular conditions,

$$p(\theta|y) \approx N(\theta_0, (nJ(\theta_0))^{-1}).$$

Proof.

see Appendix B. □

Asymptotic normality and consistency (cont.)

summary

- In the limit of large n, in the context of a specified family of models, the posterior mode, $\hat{\theta}$, approaches θ_0, and the curvature (the observed information or the negative of the coefficient of the second term in the Taylor expansion) approaches $nJ(\hat{\theta})$ or $nJ(\theta_0)$.
- As $n \to \infty$, the likelihood dominates the prior distribution, so we can just use the likelihood alone to obtain the mode and curvature for the normal approximation.
- More precise statements of the theorems and outlines of proofs appear in Appendix B.

Counterexamples to the theorems

Counterexamples to the theorems

- A good way to understand the limitations of the large-sample results is to consider cases in which the theorems fail.
- The normal distribution is usually helpful as a starting approximation, but one must examine deviations, especially with unusual parameter spaces and in the extremes of the distribution.
- The counterexamples to the asymptotic theorems generally correspond to situations in which the prior distribution has an impact on the posterior inference, even in the limit of infinite sample sizes.

Counterexamples to the theorems

Example 3 (Gelman et al. (2013) pp.89-91)

- *Underidentified models and nonidentified parameters*
- *Number of parameters increasing with sample size: hierarchical model*
- *Aliasing: Identifiability*
- *Unbounded likelihoods*
- *Improper posterior distributions*
- *Prior distributions that exclude the point of convergence*
- *Convergence to the edge of parameter space*
- *Tails of the distribution*

Frequency evaluations of Bayesian inferences

Frequency evaluations of Bayesian inferences

- Just as the Bayesian paradigm can be seen to justify simple 'classical' techniques, the methods of frequentist statistics provide a useful approach for evaluating the properties of Bayesian inferences when these are regarded as embedded in a sequence of repeated samples.
- We have already used this notion in discussing the ideas of consistency and asymptotic normality.
- The notion of stable estimation, which says that for a fixed model, the posterior distribution approaches a point as more data arrive is based on the idea of repeated sampling.

summary

Validate Bayesian inferences with frequentist's approach (for a fixed model, check wether the posterior distribution approaches a point as more data arrive or not)

Large-sample correspondence

Theorem 3 (Central Limit Theorem)

Let $\hat{\theta}$ be the maximum likelihood estimate for the fixed θ, then

$$\left[I(\hat{\theta})\right]^{1/2}(\theta - \hat{\theta})|\theta \sim N(0, I) \tag{4.3}$$

Theorem 4 (variation of the Theorem 1)

Let $\hat{\theta}$ be the posterior mode, then

$$\left[I(\hat{\theta})\right]^{1/2}(\theta - \hat{\theta})|y \sim N(0, I) \tag{4.4}$$

Large-sample correspondence (cont.)

Remark

These results mean that, for any function of $\theta - \hat{\theta}$, the posterior distribution derived from Theorem 4 is asymptotically the same as the repeated sampling distribution derived from Theorem 3.

Example 4

A 95% central posterior interval[4] for θ will cover the true value 95% of the time under repeated sampling with any fixed true θ.

[4] equal tailed credible interval

consistency

Definition (consistent)

A point estimate is said to be consistent in the sampling theory sense if, as samples get larger, it converges to the true value of the parameter that it is asserted to estimate.

Remark (asymptotic unbiasedness)

A point estimate $\hat{\theta}$ is said to be asymptotic unbiasedness if $(\mathbb{E}\left[\hat{\theta}|\theta_0\right] - \theta_0)/sd(\hat{\theta}|\theta_0)$ converges to 0.

Example 5

- *sample mean $\bar{x}$ is consistent and unbiased.*
- *When the truth is included in the family of models being fitted, the posterior mode $\hat{\theta}$, and also the posterior mean and median, are consistent and asymptotically unbiased under mild regularity conditions.*

Confidence coverage

Definition (confidence region)

If a region $C(y)$ includes θ_0 at least $100(1-\alpha)\%$ of the time (given any value of θ_0) in repeated samples, then $C(y)$ is called a $100(1-\alpha)\%$ confidence region for the parameter θ.

Remark

At the Example 4, we saw that asymptotically a $100(1-\alpha)\%$ central posterior interval for θ has the property that, in repeated samples of y, $100(1-\alpha)\%$ of the intervals include the value θ_0.

Bayesian interpretations of other statistical methods

Bayesian interpretations of other statistical methods

Note

- *Bayesian methods are often similar to other statistical approaches in problems involving large samples from a fixed probability model*
- *Even for small samples, many statistical methods can be considered as approximations to Bayesian inferences based on particular prior distributions; as a way of understanding a statistical procedure, it is often useful to determine the implicit underlying prior distribution.*
- *Some methods from classical statistics (notably hypothesis testing) can give results that differ greatly from those given by Bayesian methods*

Bayesian interpretations of other statistical methods (cont.)

Example 6

- *Maximum likelihood and other point estimates*
- *Unbiased estimates: regression to the mean*
- *Confidence intervals*
- *Hypothesis testing*
- *Multiple comparisons and multilevel modeling*
- *Nonparametric methods, permutation tests, jackknife, bootstrap*

References

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall CRC.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer Science & Business Media.

Lauritzen, S. (2011). CIMPA Summerschool, Hammamet 2011, Lecture Note: Gaussian Graphical Models. URL: `http://www.stats.ox.ac.uk/~steffen/teaching/cimpa/gauss.pdf`. Last visited on 2019/01/18.

김우철 (2012). *수리통계학*. 민영사, 서울.