Math 459: Lecture 12

Todd Kuffner

Approximate Bayesian Inference

Exact analytic calculation of posterior quantities often not practical.

Alternatives:

1. asymptotic (large-sample) approximations
2. analytic integral approximations
3. numerical integration

Basics of Parametric Bayesian Asymptotics

consistency convergence to point mass

asymptotic normality Bernstein-von Mises theorems

agreement with frequentist intervals first-order likelihood and Bayesian asymptotics agree

Consistency of the Posterior

Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f(x|\theta_0)$. Denote the posterior by $\Pi(\cdot|X^{(n)})$.

Definition

The sequence of posteriors $\Pi(\cdot|X^{(n)})$ is **consistent** at a point $\theta_0 \in \Theta$ if for every neighborhood U of θ_0, we have that $\Pi(U|X^{(n)}) \to 1$ as $n \to \infty$ almost surely (with respect to the distribution under θ_0).

This implies that the usual estimators such as the posterior mean are consistent in the usual sense.

Reminder: Asymptotic Normality of MLE

Recall the Cramér-Rao conditions: assume $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P_\theta$ (probability distribution) with density $f(x|\theta)$, and

1. θ is identifiable, i.e. $\theta_1 = \theta_2 \Leftrightarrow P_{\theta_1} = P_{\theta_2}$
2. $\theta \in \Theta =$ an open interval in the real line
3. $S = \{x : f(x|\theta) > 0\}$ does not depend on θ (the support doesn't depend on the parameter)
4. for all $x \in S$, $\frac{d}{d\theta} f(x|\theta)$ exists (the likelihood depends smoothly on the parameter)

A model satisfying the above is called a regular parametric model.

Under the Cramér-Rao conditions, there exists a sequence of roots of the likelihood equation $L'(\theta) = 0$ that is consistent and satisfies

$$\sqrt{n}(\theta - \hat{\theta}_n) \to_d \mathcal{N}(0, I^{-1}(\theta_0))$$

where $I(\theta)$ is the Fisher information matrix.

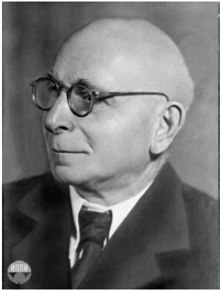
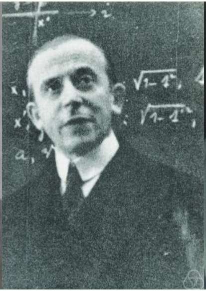
Multivariate Normal

A p-dimensional random vector X with mean vector μ and non-singular covariance matrix Σ has a **multivariate normal distribution** if its p-variate probability density function is

$$(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}$$

- written as $X \sim \mathcal{N}_p(\mu, \Sigma)$
- when Σ is singular, this distribution is still defined but it does not have a density in the usual sense

Bernstein-von Mises theorem



- Sergei Natanovich Bernstein (1880-1968); solved one of Hilbert's problems in his PhD thesis
- Richard von Mises (1883-1953); proposed the 'birthday problem'
- Joseph Doob (1910-2004); theory of martingales
- Lucien Le Cam (1924-2000); local asymptotic normality

(Corollary of) BvM theorem

Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} f(x|\theta)$ and let $\theta \sim \pi(\theta)$, a density function.

Assumptions regularity conditions

- Θ an open set, likelihood function sufficiently smooth
- $0 < I(\theta) < \infty$ (Fisher information)
- prior $\pi(\theta)$ sufficiently smooth in neighborhood of true value θ_0

Theorem Let $\hat{\theta}_n$ be a strongly consistent sequence of roots of the likelihood equation, and let $\Omega(x)$ be the CDF of a normal random variable with mean 0 and variance $I^{-1}(\theta_0)$. Then

$$\sup_{-\infty < x < \infty} |P(\sqrt{n}(\theta - \hat{\theta}_n) \leq x | X^{(n)}) - \Omega(x)| \to_{a.s.} 0$$

Note: θ is random here, $\hat{\theta}_n$ is not; more generally, we have $\sqrt{n}(\theta - \hat{\theta}_n) \in A$ where $A \subseteq \Theta$; then this is multivariate normal

Note: could use $I^{-1}(\hat{\theta}_n)$ instead of $I^{-1}(\theta_0)$ and it would still hold

Asymptotic Agreement of Bayesian and Frequentist Inference

One may guess that since both the posterior and MLE are asymptotically normal, then posterior credible sets and likelihood-based confidence sets might agree asymptotically.

- formally this means that all smooth priors are probability matching to order $O(n^{-1/2})$
- that is, the frequentist repeated sampling probability coverage of a $100(1-\alpha)$-credible set is $1-\alpha+O(n^{-1/2})$
- philosophically, would a Bayesian care about this?

Conditionality Principle (not controversial)

Suppose that cocaine smuggler Sally wants to measure the purity (e.g. concentration) of a shipment. She can use one of two labs:

Lab A more accurate, with standard deviation of 1

Lab B less accurate, with standard deviation of 10

The more accurate lab is available with probability 1/2.

- is it reasonable to argue that since the probability the more accurate lab was used is 50%, then the standard deviation of the measurement is

$$\sqrt{0.5 \cdot 1^2 + 0.5 \cdot 10^2} = 7.1$$

No, the standard deviation of the lab that was *not* used is not relevant. We should *condition* our inference on which lab was actually used.

Definition (Conditionality Principle)

If an experiment for inference about θ is chosen independently from a collection of different possible experiments, then any experiment not chosen is irrelevant to the inference about θ.

Sufficiency Principle (not controversial)

Suppose $X \sim f(x|\theta)$.

- a function (statistic) $T(x)$ is sufficient for a model $\{f(x|\theta),\ \theta \in \Theta\}$ if the conditional distribution of x given $T(x) = t$ does not depend on θ
- $T(x)$ contains all the information contained in x regarding θ

Definition (Sufficiency Principle)

If there are two observations x and y such that $T(x) = T(y)$ for a sufficient statistic T, then any conclusion about θ should be the same for x and y

Comment

The **factorization theorem** says that under certain regularity conditions

$$f(x|\theta) = g(T(x)|\theta)h(x|T(x))$$

- Rao-Blackwell theorem: optimal estimators under convex loss depend only on sufficient statistics
- if $g(X)$ is an estimator of θ, and the loss is convex, then $\delta(X) = E[g(X)|T(X)]$ has lower risk (Jensen's inequality)

Example

Consider n i.i.d. observations from $\mathcal{N}(\mu, \sigma^2)$.

- this is an exponential family; the sufficient statistic for the parameter $\theta = (\mu, \sigma^2)$ is the two-dimensional statistic $T(X) = (\bar{X}, S^2)$ given by

$$\bar{X} = n^{-1} \sum_{i=1}^{n} X_i, \quad S^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- sufficiency principle: inference on $\theta = (\mu, \sigma^2)$ should be based only on $T(x)$

Birnbaum's Theorem (1962)

The conditionality and sufficiency principles imply the likelihood principle.

Definition (Likelihood Principle)

If there are two different experiments for inference about the same parameter θ and if the outcomes x and y from the two experiments are such that the likelihood functions differ only by a multiplicative constant, then the inference should be the same

Example

Consider some event with unknown probability p and we wish to test $H_0 : p \leq 0.5$ vs. $H_1 : p > 0.5$.

- one approach: repeat the trial n times and observe number X of trials where the event happened; X is random, n is fixed

$$P_p(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- another approach: repeat the experiment a random number N times until the event has happened a fixed number of times x; N is random and x is fixed

$$P_p(N = n) = \binom{n-1}{x-1} p^x (1-p)^{n-x}$$

Both likelihoods are proportional to $p^x(1-p)^{n-x}$.

Likelihood principle $\Rightarrow$ inferences about p should be based on $p^x(1-p)^{n-x}$, regardless of how the sampling took place.

Comment

This principle is violated by many frequentist procedures (tests and confidence intervals).

- admissibility
- unbiasedness
- probability matching (priors)

These concepts (principles) depend on observations *not yet taken* (i.e. they require averaging over the sample space).

$\Rightarrow$ contravenes the likelihood principle

Overall message: though repeated sampling performance, admissibility, even asymptotics are not terribly important to applied Bayesians for their own sake, these criteria are still viewed as useful tools for evaluating Bayesian procedures