Bayesian data analysis – 4

Aki Vehtari

Chapter 4

Outline of the chapter 4

- 4.1 Normal approximation (Laplace's method)
- 4.2 Large-sample theory
- 4.3 Counter examples
- 4.4 Frequency evaluation (not part of the course, but interesting)
- 4.5 Other statistical methods (not part of the course, but interesting)

Normal approximation is used often used as part of posterior computation (more about this in Ch 13, which is not a part of the course).

Demos

- esim4_1: Bioassay example

Find all the terms and symbols listed below. When reading the chapter, write down questions related to things unclear for you or things you think might be unclear for others.

- sample size
- asymptotic theory
- normal approximation
- quadratic function
- Taylor series expansion
- observed information
- positive definite
- why $\log \sigma$?
- Jacobian of the transformation
- point estimates and standard errors
- lower-dimnsional approximations
- large-sample theory
- asymptotic normality
- consistency
- underidentified
- nonidentified
- number of parameters increasing with sample size
- aliasing
- unbounded likelihood
- improper posterior
- edge of parameter space
- tails of distribution

Normal approximation

Other Gaussian posterior approximations are discussed in Chapter 13. For example, variational and expectation propagation methods improve the approximation by global fitting instead of just the curvature at the mode. The Gaussian approximation at the mode is often also called the Laplace method, as Laplace used it first.

Several researchers have provided partial proofs that posterior converges towards Gaussian distribution. In the mid 20th century Le Cam was first to provide a strict proof.

Observed information

When $n \to \infty$, the posterior distribution approaches Gaussian distribution. As the log density of the Gaussian is a quadratic function, the higher derivatives of the log posterior approach zero. The curvature at the mode describes the information only in the case if asymptotic normality. In the case of the Gaussian distribution, the curvature describes also the width of the Gaussian. Information matrix is a *precision matrix*, which inverse is a covariance matrix.

Aliasing

In Finnish: valetoisto.

Aliasing is a special case of under-identifiability, where likelihood repeats in separate points of the parameter space. That is, likelihood will get exactly same values and has same shape although possibly mirrored or otherwise projected. For example, the following mixture model

$$p(y_i|\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = \lambda \,\mathrm{N}(\mu_1, \sigma_1^2) + (1 - \lambda)\,\mathrm{N}(\mu_2, \sigma_2^2),$$

has two Gaussians with own means and variances. With a probability λ the observation comes from $\mathrm{N}(\mu_1, \sigma_1^2)$ and a probability $1 - \lambda$ from $\mathrm{N}(\mu_2, \sigma_2^2)$. This kind of model could be used, for example, for the Newcomb's data, so that the another Gaussian component would model faulty measurements. Model does not state which of the components 1 or 2, would model good measurements and which would model the faulty measurements. Thus it is possible to interchange values of (μ_1, μ_2) and (σ_1^2, σ_2^2) and replace λ with $(1 - \lambda)$ to get the equivalent model. Posterior distribution then has two modes which are mirror images of each other. When $n \to \infty$ modes will get narrower, but the posterior does not converge to a single point.

If we can integrate over the whole posterior, the aliasing is not a problem. However aliasing makes the approximative inference more difficult.

Frequency property vs. frequentist

Bayesians can evaluate frequency properties of Bayesian estimates without being frequentist. For Bayesians the starting point is the Bayes rule and decision theory. Bayesians care more about efficiency than unbiasedness. For frequentists the starting point is to find an estimator with desired frequency properties and quite often unbiasedness is chosen as the first restriction.

Transformation of variables

See p. 21 for the explanation how to derive densities for transformed variables. This explains, for example, why uniform prior $p(\log(\sigma^2)) \propto 1$ for $\log(\sigma^2)$ corresponds to prior $p(\sigma^2) = \sigma^{-2}$ for $sigma^2$.

On derivation

Here's a reminder how to integrate with respect to $g(\theta)$. For example

$$\frac{d}{d\log\sigma}\sigma^{-2} = -2\sigma^{-2}$$

is easily solved by setting $z = \log\sigma$ to get

$$\frac{d}{dz}\exp(z)^{-2} = -2\exp(z)^{-3}\exp(z) = -2\exp(z)^{-2} = -2\sigma^{-2}.$$