Large-sample inference (chapter 4)

- Normal approximation
  - Taylor series expansion of log-posterior
  - aka Laplace approximation
- Counterexamples
- Frequency evaluations

Slide 1

Normal approximation

- If the posterior distribution is unimodal and roughly symmetric
  - it can be approximated by a normal distribution

$$p(\theta|y) \approx \frac{1}{\sqrt{2\pi}\sigma_\theta} \exp\left(-\frac{1}{2\sigma_\theta^2}(\theta - \hat{\theta})^2\right)$$

  - i.e. log-posterior $\log p(\theta|y)$ can be approximated by a quadratic function

Slide 2

$$\log p(\theta|y) \approx \alpha(\theta - \hat{\theta})^2 + C$$

Taylor series

- Taylor series expansion at $x = a$

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \dots$$

Slide 3

- Generalizes to multidimensional

$$f(x_1,\dots,x_n) = \sum_{j=0}^{\infty} \left\{ \frac{1}{j!} \left[\sum_{k=1}^{n} (x_k - a_x) \frac{\partial}{\partial x'_k} \right]^j f(x'_1,\dots,x'_n) \right\}_{x'_1=a_1,\dots,x'_n=a_n}$$

Normal approximation

- Taylor series expansion of log-posterior around at the posterior mode $\hat{\theta}$

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

where linear term is zero and higher terms are small when θ close to $\hat{\theta}$ and n large (see appendix B)

Slide 4

- Multivariate normal $\propto |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\theta-\hat{\theta}^T)\Sigma^{-1}(\theta-\hat{\theta})\right)$

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

where $I(\theta)$ is *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

Normal approximation

- $I(\theta)$ is *observed information*

$$I(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y)$$

  - $I(\hat{\theta})$ is the second derivative of the log posterior at the mode
  - if the mode is inside the parameter space, $I(\hat{\theta})$ is positive
  - if θ is vector, $I(\theta)$ is matrix

Slide 5

Normal approximation - example

- Normal distribution, unknown mean and variance
  - uniform prior on $(\mu, \log \sigma)$
  - normal approximation of posterior of $(\mu, \log \sigma)$

$$\log p(\mu, \log \sigma|y) = \text{constant} - n \log \sigma - \frac{1}{2\sigma^2}[(n-1)s^2 + n(\bar{y} - \mu)^2]$$

first derivatives

$$\begin{aligned} \frac{d}{d\mu} \log p(\mu, \log \sigma|y) &= \frac{n(\bar{y} - \mu)}{\sigma^2}, \\ \frac{d}{d(\log \sigma)} \log p(\mu, \log \sigma|y) &= -n + \frac{(n-1)s^2 + n(\bar{y} - \mu)^2}{\sigma^2}, \end{aligned}$$

Slide 6

from which posterior mode is easy calculate

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2\right)\right)$$

Normal approximation - example

- Normal distribution, unknown mean and variance

  first derivatives

$$\begin{aligned}\frac{d}{d\mu}\log p(\mu,\log\sigma|y) &= \frac{n(\bar{y}-\mu)}{\sigma^2},\\ \frac{d}{d(\log\sigma)}\log p(\mu,\log\sigma|y) &= -n+\frac{(n-1)s^2+n(\bar{y}-\mu)^2}{\sigma^2}\end{aligned}$$

Slide 7

  second derivatives

$$\begin{aligned}\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) &= -\frac{n}{\sigma^2},\\ \frac{d^2}{d\mu d(\log\sigma)}\log p(\mu,\log\sigma|y) &= -2n\frac{\bar{y}-\mu}{\sigma^2},\\ \frac{d^2}{d(\log\sigma)^2}\log p(\mu,\log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2+n(\bar{y}-\mu)^2)\end{aligned}$$

Normal approximation - example

- Normal distribution, unknown mean and variance

  second derivatives

$$\begin{aligned}\frac{d^2}{d\mu^2}\log p(\mu,\log\sigma|y) &= -\frac{n}{\sigma^2},\\ \frac{d^2}{d\mu(\log\sigma)}\log p(\mu,\log\sigma|y) &= -2n\frac{\bar{y}-\mu}{\sigma^2},\\ \frac{d^2}{d(\log\sigma)^2}\log p(\mu,\log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2+n(\bar{y}-\mu)^2)\end{aligned}$$

Slide 8

  matrix of second derivatives evaluated at $(\hat{\mu},\log\hat{\sigma})$

$$\begin{pmatrix}-n/\hat{\sigma}^2 & 0\\ 0 & -2n\end{pmatrix}$$

Normal approximation - example

- Normal distribution, unknown mean and variance

  mode of the posterior

$$(\hat{\mu}, \log \hat{\sigma}) = \left(\bar{y}, \frac{1}{2} \log \left(\frac{n-1}{n} s^2\right)\right)$$

  matrix of second derivatives evaluated at $(\hat{\mu}, \log \hat{\sigma})$

Slide 9

$$\begin{pmatrix} -n/\hat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}$$

  normal approximation

$$p(\mu, \log \sigma | y) \approx \mathrm{N}\left(\begin{pmatrix} \mu \\ \log \sigma \end{pmatrix} \middle| \begin{pmatrix} \bar{y} \\ \log \hat{\sigma} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2/n & 0 \\ 0 & 1/(2n) \end{pmatrix}\right)$$

Normal approximation

- Useful if
  - the posterior similar to normal
    - depends on the model and parametrisation how fast posterior approaches normality when n increases
  - inference not sensitive to the imperfections in the approximation
    - e.g. mean is less sensitive than extreme quantiles

Slide 10

- Approximation can be often improved with transformation of variables
  - e.g. use $\log \sigma$ instead of σ
  - posterior of σ and $\log \sigma$ approaches normality, but with finite n approximation is better for $\log \sigma$

Normal approximation

- Approximation can be made to marginal distribution
  - marginals are always closer to normal
  - requires that marginal is relatively easy to compute
  - Integrated Nested Laplace Approximation (INLA)
    - recent method for efficiently evaluating many marginals for latent Gaussian models (guest lecture 13.11. 16:00 Exactum B120!)

Slide 11

- Approximation can be made for conditional distribution
  - approximative Rao-Blackwellisation

Normal approximation

- Easy to compute
  - HPD
  - mean, median, mode, intervals
- Can be used as a starting guess for MCMC-methods
- Can be used as a proposal distribution in importance sampling

Slide 12

Normal approximation

- Can be computed numerically
  - derivatives can be computed using finite-difference (with small number of parameters)
  - minimize the negative log-poster: minimum is the mode and Hessian at the minimum is the observed information
  - e.g. with Matlab
    [w,fval,exitflag,output,g,H]=fminunc(@nlogp,w0,opt,x,y,n);

Slide 13

Bioassay

Dose, x_i (log g/ml)	Number of animals, n_i	Number of deaths, y_i
-0.86	5	0
-0.30	5	1
-0.05	5	3
0.73	5	5

Slide 14

- $y_i|\theta_i \sim \mathrm{Bin}(n_i, \theta_i)$
- Logistic regression $\mathrm{logit}(\theta_i) = \alpha + \beta x_i$
- Likelihood

$$p(y_i|\alpha, \beta, n_i, x_i) \propto [\mathrm{logit}^{-1}(\alpha + \beta x_i)]^{y_i}[1 - \mathrm{logit}^{-1}(\alpha + \beta x_i)]^{n_i - y_i}$$

- Posterior

$$p(\alpha, \beta|y, n, x) \propto p(\alpha, \beta) \prod_{i=1}^{n} p(y_i|\alpha, \beta, n_i, x_i)$$

- esim5_1.m, ex 4.2

Bioassay

- Hint for ex 4.2
- Likelihood

$$
\begin{aligned}
p(y_i|\alpha,\beta,n_i,x_i) &\propto [\operatorname{logit}^{-1}(\alpha+\beta x_i)]^{y_i}[1-\operatorname{logit}^{-1}(\alpha+\beta x_i)]^{n_i-y_i} \\
&\propto \theta^{y_i}[1-\theta]^{n_i-y_i}
\end{aligned}
$$

Slide 15

- Write log-poster in neat form
- denote $\theta = \operatorname{logit}^{-1}(\phi)$ and $\phi = \alpha + \beta x_i$, and use chain rule in derivation
- See logit and $\operatorname{logit}^{-1}$ at page 24
- Recognize familiar forms, rearrange terms and keep it simple
- Compare to numerical result (esim5_1.m) (*Hessian*)

Large sample theory

- In this course only superficially
  - see appendix B for some more
- Assume ”true” data distribution $f(y)$
  - observations $y_1, \ldots, y_n$ independent samples from $f(y)$
  - ”true” distribution $f(y)$ is not clear concept
  - Bayesians can say, that we proceed *as if* there were ”true” distribution $f(y)$
  - for large sample theory the exact form of $f(y)$ is not important, as long as some regularity conditions hold

Slide 16

Large sample theory

- Asymptotic normality
- Consistency
  - if $f(y) = p(y|\theta_0)$ for some θ_0, then posterior converges to single point θ_0, as $n \to \infty$
- if $f(y) \neq p(y|\theta_0)$
  - posterior converges to θ_0 for which $p(y|\theta_0)$ is closest to $f(y)$ measured with *Kullback-Leibler information*

$$H(\theta_0) = \int f(y_i) \log \left(\frac{f(y_i)}{p(y_i|\theta_0)} \right) dy_i$$

Slide 17

Kullback-Leibler information

$$H(\theta_0) = \int f(y_i) \log \left(\frac{f(y_i)}{p(y_i|\theta_0)} \right) dy_i$$

- Divergence measure
  - not distance, since non-symmetric
  - if $\log_2$, divergence measured in bits
  - if $\log_e$, divergence measured in nats

Slide 18

Asymptotic normality and consistency

- If certain regularity conditions hold for the likelihood
  - e.g. continuous function of θ and θ_0 not on the boundary of the parameter space

  then posterior of θ approaches normality

$$\mathrm{N}(\theta_0, (nJ(\theta_0))^{-1}),$$

Slide 19

  where $J(\theta)$ is Fisher's information

- Compare

$$\begin{aligned} \text{observed information} \quad I(\theta) &= -\frac{d^2 \log p(\theta|y)}{d\theta^2} \\ \text{Fisher's information} \quad J(\theta) &= -\,\mathrm{E}\left[\frac{d^2 \log p(y|\theta)}{d\theta^2}\Big|\theta\right] \end{aligned}$$

Asymptotic normality and consistency

- Observed information

$$I(\theta) = -\frac{d^2 \log p(\theta|y)}{d\theta^2}$$

  if for posterior $p(\theta|y)$ given specific observation y

- Fisher's information

Slide 20

$$J(\theta) = -\,\mathrm{E}\left[\frac{d^2 \log p(y|\theta)}{d\theta^2}\Big|\theta\right]$$

  if for likelihood $p(y|\theta)$ expectation over distribution of y given θ (not for specific y)

- When $n \longrightarrow \infty$ these approaches same value

- Can be interpreted using Taylor series expansion

Asymptotic normality and consistency

- Taylor series expansion at the mode of the posterior $\hat{\theta}$

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

- When $n \longrightarrow \infty$, mass of the posterior concentrates in smaller and smaller neighborhoods of θ_0:n and $|\hat{\theta} - \theta_0| \longrightarrow 0$, (consistency)

Slide 21

- Write quadratic term as

$$\left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\hat{\theta}} = \left[\frac{d^2}{d\theta^2} \log p(\theta)\right]_{\theta=\hat{\theta}} + \sum_{i=1}^{n} \left[\frac{d^2}{d\theta^2} \log p(y_i|\theta)\right]_{\theta=\hat{\theta}}$$

  as function of θ this is a constant plus the sum of n terms, each of whose expected value under the true sampling distribution $p(y|\theta_0)$ is approximately $-J(\theta_0)$, as long as $\hat{\theta}$ is close to θ_0

- For large n, the curvature of the log posterior density can be approximated by Fisher information evaluated at either $\hat{\theta}$ or θ_0 (only $\hat{\theta}$ available in practice)

Normal approximation

- In practice useful only for some models
  - often n not large enough
  - also several counter examples even if $n \longrightarrow \infty$
  - approximation can be evaluated, e.g., using importance sampling
  - other methods
    - use for conditionals or marginals
    - use for some marginals with Integrated Nested Laplace Approximation (INLA)
    - t-distribution, skewed-t-distribution
    - variational methods, expectation propagation
    - Monte Carlo methods

Slide 22

- Despite of limitations essential part of Bayesian toolkit

Normal approximation - counter examples

Slide 23

- Under- and non-identifibiality
- Number of parameters increasing with sample size
- Aliasing
- Unbounded likelihood
- Improper posterior
- Prior distribution excludes the point of convergence
- Convergence to the edge of parameter space
- Tails of the distribution

Large sample theory - counterexamples

Slide 24

- Theory does not always hold even if $n \longrightarrow \infty$
- Under- ja nonidentified
  - model is under-identified, if data can not update uncertainty related to some parameters or parameter combinations
  - no single convergence point θ_0
  - eg. if only one of u or v is observed from each pair (u, v) and model is

    $$\begin{pmatrix} u \\ v \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

    the ρ is nonidentified
  - eg. u is height of a student v is weight of a student
  - problematic for MC-methods, too

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Number of parameters increasing with sample size
  - in many models the number of parameters depends on the number of observations
  - eg. spatial models $y_i \sim \mathrm{N}(\theta_i, \sigma^2)$ and θ_i is has spatial prior
  - posterior of θ_i does converge to a point, if new data do not bring enough information about θ_i

Slide 25

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Aliasing
  - Special case of underidentified parameters in which the same likelihood function repeats at a discrete set of points
  - eg. normal mixture model

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

    if we interchange each of (μ_1, μ_2) and (σ_1^2, σ_2^2), and replace λ with $(1 - \lambda)$, the likelihood of the data remains same
    the posterior generally has at least two modes that are mirror images of each other; it does not converge to a single point
  - in general not a problem for MC-methods , but makes convergence diagnostics more difficult
  - can be eliminated by restricting parameter space; eg. in previous example by restricting $\mu_1 \leq \mu_2$

Slide 26

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Unbounded likelihood
  - if likelihood function is unbounded, then there might be no posterior mode within parameter space
  - eg. the previous normal mixture model; assume known λ (not 0 or 1); if $\mu_1 = y_i$ for any y_i and $\sigma_1^2 \longrightarrow 0$, then likelihood$\longrightarrow \infty$
  - as $n \longrightarrow \infty$, the number of modes of the likelihood increases
  - if the prior is uniform on σ_1^2 and σ_1^2 near zero $\longrightarrow$ the number of modes the likelihood increases
  - problematic also for, e.g. MC-methods
  - the problem can be solved by restricting the model to plausible set of distributions
  - note, that vague priors and finite n may have almost unbounded posterior

Slide 27

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Improper posterior distribution
  - asymptotic results require probabilities to sum to one
  - eg. Binomial with prior $\mathrm{Beta}(0, 0)$ and data $y = n$
    - posterior $p(\theta|n, 0) = \theta^{n-1}(1-\theta)^{-1}$
    - if $\theta \longrightarrow 1$, then $p(\theta|n, 0) \longrightarrow \infty$
  - problematic also for, eg.. MC methods
  - the problem can be solved by using proper prior
  - note, that vague priors may produce almost improper posterior

Slide 28

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Prior distribution excludes the point of convergence
  - if in discrete case $p(\theta_0) = 0$ or in continuous case $p(\theta) = 0$ in a neighborhood about θ_0 , then the convergence results do not hold
  - not a problem for MC methods
  - the solution is to give positive prior density to all values that are even remotely possible

Slide 29

Large sample theory - counterexamples

- Theory does not always hold even if $n \longrightarrow \infty$
- Convergence to the edge of parameter space
  - if θ_0 is on the boundary of the parameter space, then the Taylor series expansion must be truncated and approximation will not necessarily be appropriate
  - eg. $y_i \sim \mathrm{N}(\theta, 1)$ with the restriction $\theta \geq 0$ and assume that $\theta = 0$ is true value
  - θ's posterior is normal with $\mu = \bar{y}$ and truncated to be positive
  - in the limit as $n \longrightarrow \infty$ posterior is half of normal distribution
  - not a problem for MC methods

Slide 30

Large sample theory - counterexamples

- Tails of the distribution
  - normal approximation can hold for essential all the of the posterior distribution but still not be accurate in the tails
  - eg. parameter that is restricted to be positive, with finite n normal approximation gives positive density to negative values
- MC has also problems with tails, although different type

Slide 31

Frequency evaluation

- Frequentist methods are based on repeated sampling ie frequencies
- Frequency evaluation of Bayesian inference is also based on frequencies, but Bayesian interpretation is preserved although terms and analysis tools are borrowed from frequentists
- Although often in description of Bayesian theory it is emphasised the possibility of examining probability of a single event, there is not obstacle to examine repeated event
- Normal approximation and consistency are based also on repeated sampling
- Frequency evaluation examines the properties of the methods, by considering what would happen if experiment were repeated

Slide 32

Frequency evaluation

- Asymptotic calibration of posterior intervals
- Consistency
- Asymptotic unbiasedness $[\mathrm{E}(\hat{\theta}|\theta_0) - \theta_0]/\,\mathrm{sd}(\hat{\theta}|\theta_0) \rightarrow 0$
- Asymptotic efficiency

Slide 33

Frequency evaluation

- Asymptotic results nice, but usually behavior with finite n more interesting
- Usually Bayesian estimates are biased
  - estimate is biased due to prior information
  - since truth is not usually known, prior is probably somewhat wrong, which causes bias
  - bias is not problem if variance is reduced (better efficiency)
  - slightly wrong prior causes small bias, but may reduce variance greatly
- Bias-variance dilemma
  - by increasing bias, variance may be reduced
  - increasing prior-information may increase bias, but benefit is in reduced variance

Slide 34