Bayesian inference (continued)

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

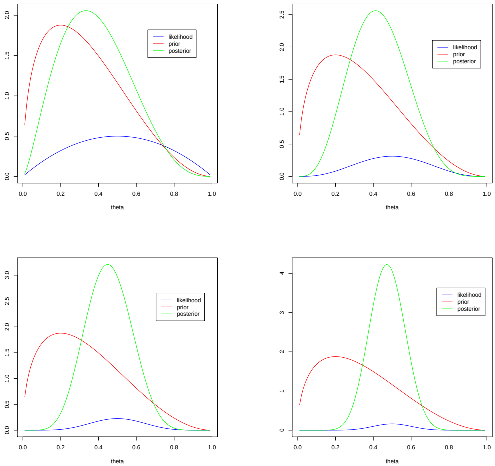
April 13, 2021

Today: selected topics in Bayesian statistics.

Asymptotics, REML, improper Gaussian

Influence of number of observations/convergence of posterior - binomial-beta

Beta-prior ($\alpha = 1.5$ $\beta = 3$). Observations $x = n/2$, $n = 2, 6, 12, 24$. Posterior mode 0.33, 0.41, 0.44, 0.47.



Note: posterior appears to converge to a normal density !

Bayesian asymptotics

Consider posterior of θ given observations $y_1, \dots, y_n$. Let $\hat{\theta}_n$ and $i_n(\theta)$ denote the MLE and Fisher information based on $y_1, \dots, y_n$.

Under appropriate regularity conditions, as $n \to \infty$,

$$\sup_A |P(i_n(\hat{\theta}_n)^{-1/2}(\theta - \hat{\theta}_n) \in A | y_1, \dots, y_n) - P(Z \in A)| \to 0$$

where $Z \sim N(0, I)$.

That is, posterior distribution of $i_n(\hat{\theta}_n)^{-1/2}(\theta - \hat{\theta}_n)$ converges in *total variation* distance (and hence in distribution) to the standard normal distribution. Note: given $y_1, \dots, y_n$, $\hat{\theta}_n$ is fixed !

Standard frequentist theory gives $i_n(\hat{\theta})^{-1/2}(\hat{\theta}_n - \theta)$ converges in distribution to a standard normal distribution but in this case θ represents fixed 'true' value while randomness of $\hat{\theta}_n$ due to sampling variation.

REML as marginal likelihood

Bayesian derivation of REML.

Consider linear mixed model $Y \sim N(X\beta, V(\psi))$.

Assume improper prior $p(\beta|\psi) \propto 1$.

Then REML is obtained by integrating out β in 'joint density' of (Y, β):

$$\texttt{REML} = f(y; \psi) = \int f(y|\beta, \psi) p(\beta|\psi) \mathrm{d}\beta$$

To show this we first compute $f(y;\psi)$ and compare it with REML.

Let $V(\psi) = LL^{\mathsf{T}}$ and $\tilde{Y} = L^{-1}Y$. Then $\tilde{Y}|\beta \sim N(\tilde{X}\beta, I)$ where $\tilde{X} = L^{-1}X$. Moreover (applying Pythagoras),

$$f(\tilde{y}|\psi) = \int f(\tilde{y}|\beta,\psi)\mathrm{d}\beta = (2\pi)^{(p-n)/2}|\tilde{X}^{\mathsf{T}}\tilde{X}|^{-1/2}\exp(-\frac{1}{2}\|\tilde{Y}-\tilde{X}\hat{\beta}\|^2)$$

where $\hat{\beta}$ is the MLE. Thus (using the transformation theorem)

$$f(y|\psi) = \frac{(2\pi)^{(p-n)/2}}{|V(\psi)|^{1/2}|X^{\mathsf{T}}V^{-1}(\psi)X|^{1/2}}\exp[-\frac{1}{2}\tilde{Y}^{\mathsf{T}}(I-\tilde{X}(\tilde{X}^{\mathsf{T}}\tilde{X})^{-1}\tilde{X}^{\mathsf{T}})\tilde{Y}]$$

Moreover,

$$\tilde{Y}^{\mathsf{T}}(I-\tilde{X}(\tilde{X}^{\mathsf{T}}\tilde{X})^{-1}\tilde{X}^{\mathsf{T}})\tilde{Y} = \|\tilde{Y}-\tilde{X}\hat{\beta}|^2 = (Y-X\hat{\beta})^{\mathsf{T}}V^{-1}(Y-X\hat{\beta})$$

(which may explain why REML can be short for "residual MLE")

REML is likelihood of $A^{\mathsf{T}} Y \sim N(0, A^{\mathsf{T}} V(\psi) A)$. Let $\tilde{A} = L^{\mathsf{T}} A$. Then $\tilde{A}^{\mathsf{T}} \tilde{X} = 0$ and

$$\tilde{A}(\tilde{A}^{\mathsf{T}} \tilde{A})^{-1} \tilde{A}^{\mathsf{T}} = I - \tilde{X}(\tilde{X}^{\mathsf{T}} \tilde{X})^{-1} \tilde{X}^{\mathsf{T}}.$$

Thus

$$\begin{aligned}\mathtt{REML} = |A^{\mathsf{T}} V(\psi) A|^{-1/2} \exp[-\frac{1}{2} Y^{\mathsf{T}} A (A^{\mathsf{T}} V(\psi) A)^{-1} A^{\mathsf{T}} Y] = \\ |A^{\mathsf{T}} V(\psi) A|^{-1/2} \exp[-\frac{1}{2} \tilde{Y}^{\mathsf{T}} (I - \tilde{X}(\tilde{X}^{\mathsf{T}} \tilde{X})^{-1} \tilde{X}^{\mathsf{T}}) \tilde{Y}]\end{aligned}$$

We hence just need to show that

$$|A^{\mathsf{T}} V(\psi) A| = \text{const} |V(\psi)| |X^{\mathsf{T}} V^{-1}(\psi) X|$$

This follows from

$$|A^\mathsf{T}A||X^\mathsf{T}X||V| = \left|\begin{bmatrix} A^\mathsf{T}A & 0 \\ 0 & X^\mathsf{T}X \end{bmatrix} V\right| = \left|\begin{bmatrix} A^\mathsf{T} \\ X^\mathsf{T} \end{bmatrix} \begin{bmatrix} A & X \end{bmatrix} V\right| =$$

$$\left|\begin{bmatrix} A^\mathsf{T} \\ X^\mathsf{T} \end{bmatrix} V \begin{bmatrix} A & X \end{bmatrix}\right| = \left|\begin{bmatrix} A^\mathsf{T}VA & A^\mathsf{T}VX \\ X^\mathsf{T}VA & X^\mathsf{T}VX \end{bmatrix}\right| =$$

$$|A^\mathsf{T}VA||X^\mathsf{T}VX - X^\mathsf{T}VA(A^\mathsf{T}VA)^{-1}A^\mathsf{T}VX| =$$

$$|A^\mathsf{T}VA||X^\mathsf{T}X(X^\mathsf{T}V^{-1}X)^{-1}X^\mathsf{T}X| = |A^\mathsf{T}VA||X^\mathsf{T}X|^2|X^\mathsf{T}V^{-1}X|^{-1}$$

(recall for partitioned matrix B, $|B| = |B_{11}||B_{22} - B_{21}B_{11}^{-1}B_{12}|$ and $A(A^\mathsf{T}VA)^{-1}A^\mathsf{T} = V^{-1} - V^{-1}X(X^\mathsf{T}V^{-1}X)^{-1}X^\mathsf{T}V^{-1}$)

Thus

$$|A^\mathsf{T}V(\psi)A| = \frac{|A^\mathsf{T}A|}{|X^\mathsf{T}X|}|V(\psi)||X^\mathsf{T}V^{-1}(\psi)X|$$

Pairwise difference prior

Suppose we observe $Y_i \sim N(\theta_i, 1)$, $i = 1, \ldots, n$ and we want to infer $\theta_1, \ldots, \theta_n$. Prior information: θ_i and θ_{i+1} "similar".

Consider stationary AR(1) prior ($\tau_1^2 = \tau^2/(1-a)$):

$$f(\theta; a) \propto \exp(-\frac{1}{\tau_1^2}\theta_1^2) \prod_{i=2}^{n} \exp[-\frac{1}{2\tau^2}(\theta_i - a\theta_{i-1}]^2)$$

Consider $a \to 1$. Then "limit" of right hand side is

$$f(\theta) \propto \exp[-\frac{1}{2\tau^2}\sum_{i=2}^{n}(\theta_i - \theta_{i-1})^2] = \exp[-\frac{1}{2\tau^2}\theta^\mathsf{T} Q\theta]$$

for which

$$Q = \begin{bmatrix} 1 & -1 & 0 & \ldots \\ -1 & 2 & -1 & \ldots \\ \vdots & \vdots & \vdots & \vdots \\ \ldots & -1 & 2 & -1 \\ \ldots & 0 & -1 & 1 \end{bmatrix}$$

Note $Q1_n = 0$ so Q does not have full rank.

On the other hand $Qx = 0$ implies $x = a1_n$ for some $a \in \mathbb{R}$. This follows since $Q = D^T D$ where D is the $n-1 \times n$ matrix

$$D = \begin{bmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots \\ \dots & 0 & -1 & 1 \end{bmatrix}$$

Hence the null space N_Q of Q is the span of 1_n. Note 1_n is the last eigenvector of Q with eigenvalue 0.

$f(\theta)$ not a proper density on $\mathbb{R}^n$ since Q is not positive definite.

Limiting posterior may nevertheless still be proper (exercise).

$f(\theta)$ is invariant to addition of a constant to all elements of θ (filters constants). Hence does not imply prior assumptions about 'level' of data $Y_i|\theta_i \sim N(\theta_i, 1)$. Only need to choose prior parameter τ^2 (smoothness parameter).

Prediction

Suppose we want to predict Y_2 given Y_1 where both depend on θ.

Frequentist approach: use $f(y_2|y_1, \hat{\theta})$ where $\hat{\theta}$ estimate based on y_1. This in general ignores extra uncertainty due to replacing θ by an estimate.

Bayesian approach offers a systematic way to take into account uncertainty of parameters in prediction by integrating out unknown parameters:

$$f(y_2|y_1) = \int f(y_2|y_1, \theta) p(\theta|y_1) \mathrm{d}\theta = \mathbb{E}_{\theta|y_1} f(y_2|y_1, \theta)$$

Predictive density is 'weighted' average of predictive densities $f(y_2|y_1, \theta)$ where 'weights' given by posterior density $p(\theta|y_1)$ reflects uncertainty of θ.

Nice solution in principle but in practice the computation may not be straightforward.

Difficult posteriors

Except for the simple examples with conjugate priors the posterior is often intractable - closed form expressions for posterior quantities like expectations, variances, quantiles etc. often not available.

Non-normal example: logistic regression with normal prior

$\beta \sim N(0, \tau^2)($ normal prior $)$
$Y_j|\beta \sim \text{binomial}(n_j, p_j)$ conditionally independent given β $j = 1, \ldots, n_j$
$\log(p_j/(1-p_j)) = \eta_j = x_j^{\mathsf{T}}\beta$
$p_j = \exp(\eta_j)/(1+\exp(\eta_j))$

Likelihood function:

$$f(y|\beta) = \prod_j p_j^{y_j}(1-p_j)^{1-y_j} = \prod_j \frac{\exp(x_j^{\mathsf{T}}\beta)^{y_j}}{(1+\exp(x_j^{\mathsf{T}}\beta))^{n_j}}$$

Marginal density $f(y)$:

$$\int_{\mathbb{R}} f(y|\beta)f(\beta;\tau^2)\mathrm{d}\beta = \int_{\mathbb{R}} \prod_j \frac{\exp(x_j^{\mathsf{T}}\beta)^{y_j}}{(1+\exp(x_j^{\mathsf{T}}\beta))^{n_j}} \frac{\exp(-\beta^2/(2\tau^2))}{\sqrt{2\pi\tau^2}}\mathrm{d}\beta$$

Integral can not be evaluated in closed form.

Laplace/Gaussian approximation

Let $g(\beta) = \log(f(y|\beta)f(\beta))$ and choose $\hat{\beta}$ so $g'(\hat{\beta}) = 0$ ($\hat{\beta} = \arg\max g(\beta)$).

Note: $\hat{\beta}$ is MAP (maximum a posteriori) estimate. Not MLE.

Taylor expansion around $\hat{\beta}$:

$$g(\beta) \approx \tilde{g}(\beta) =$$
$$g(\hat{\beta})+(\beta-\hat{\beta})g'(\hat{\beta})+\frac{1}{2}(\beta-\hat{\beta})^2 g''(\hat{\beta}) = g(\hat{\beta})-\frac{1}{2}(\beta-\hat{\beta})^2\left(-g''(\hat{\beta})\right)$$

I.e. $\exp(\tilde{g}(\beta))$ proportional to normal density $N\left(\mu_{LP}, \sigma^2_{LP}\right)$, $\mu_{LP} = \hat{\beta}$ $\sigma^2_{LP} = -1/g''(\hat{\beta})$.

Since

$$p(\beta|y) \approx \exp(g(\beta)) \approx \exp(\tilde{g}(\beta))$$

it follows

$$\beta|Y = y \approx N(\hat{\beta}, -1/g''(\hat{\beta}))$$

Regarding marginal density of y:

$$f(y) = \int_{\mathbb{R}} \exp(g(\beta))\mathrm{d}\beta \approx \int_{\mathbb{R}} \exp(\tilde{g}(\beta))\mathrm{d}\beta$$
$$= \exp(g(\hat{\beta})) \int_{\mathbb{R}} \exp\big(-\frac{1}{2\sigma_{LP}^2}(\beta - \mu_{LP})^2\big)\mathrm{d}\beta = \exp(g(\hat{\beta}))\sqrt{2\pi\sigma_{LP}^2}$$

Note: these kinds of arguments basis of asymptotic results for posterior distributions.

Other approaches

Numerical integration (Gaussian quadrature), Monte Carlo, importance sampling, Markov chain Monte Carlo,....

Enough material for a whole course.

Why/when is Bayesian inference useful

- obvious if prior information is available
- for highly complex models maximum likelihood inference is difficult (multimodality, evaluation of likelihood). Computation of posterior expectations and probabilities numerically more simple.
- can compute posterior distributions of complicated parameters whose distribution may be hard to obtain in the MLE setting.
- natural approach to take into account parameter uncertainty in prediction.

Exercises

1. (hidden AR(1) model) assume that Y_i given θ are independent $N(\theta_i, 1)$ and that $\theta = (\theta_1, \ldots, \theta_n)$ follows a stationary AR(1) process prior with known autoregression parameter a and noise variance τ^2.
   1.1 Compute the posterior distribution of θ (e.g. use the previous results for conditional distributions in general linear mixed models).
   1.2 What is the limiting posterior when $a \to 1$?
   1.3 is the limiting prior proper ? is the limiting posterior proper ?

2. Consider data $X_1, \ldots, X_n$ from a zero-mean AR(1) process. Consider the *conditional likelihood* of $X_2, \ldots, X_n$ given (X_1, a, τ^2).
   2.1 Show that the posterior distribution of (a, τ^2) obtained by combining the conditional likelihood with the (improper) prior $p(a, \tau^2) \propto 1/\tau^2$ is equivalent to the posterior for a linear normal model with observation vector $(X_2, \ldots, X_n)^{\mathsf{T}}$ and design matrix given by the column $(X_1, \ldots, X_{n-1})^{\mathsf{T}}$.
   2.2 use the previous results to compute the predictive mean and variance of X_{n+1} given $X_1, \ldots, X_n$ (again using the conditional likelihood instead of the usual likelihood of $(X_1, \ldots, X_n)$).

   NB: rather than using the conditional likelihood we could instead assume $X_1 \sim N(\mu_1, 1)$ and use the usual likelihood of $(X_1, \ldots, X_n)$ given (μ_1, a, τ^2) combined with the prior $p(\mu_1, a, \tau^2) \propto 1/\tau^2$. This would give the same posterior inference for (a, τ^2) as by using the conditional likelihood.