Bayesian Inference I Labs, AY 2024-2025

Roberto Trotta, labs led by Andre Scaffidi

November 2024

1. We investigate here in more detail the coin tossing problem. A coin is tossed N times and heads come up H times. Let θ denote the probability of heads in one flip.

   (a) Compute analytically the posterior probability for θ, for the case of a Jeffreys' prior and a uniform prior on θ. This integral will prove useful:

   $$\int_0^1 d\theta \theta^N (1-\theta)^M = \frac{\Gamma(N+1)\Gamma(M+1)}{\Gamma(N+M+2)}.$$

   (b) Determine the posterior mean and posterior standard deviation of θ for both choices of priors.

   (c) Plot the posterior (for both choices of prior) as a function of θ, assuming that the true value of the parameter is $\theta_t = \{0.1, 0.5, 0.8\}$, and for $N = \{10, 100, 1000\}$. Compute numerically the 68.3% (i.e., 1σ) highest posterior density (HPD) interval in each case, and compare it with the posterior standard deviation found above.

   (d) By simulating a large number of pseudo-data, compare the coverage properties of the 68.3% (1σ) and the 95.4% (2σ) HPD intervals for both choices of prior, as well as with the coverage properties of the likelihood-based MLE with Wald confidence intervals. Do this for the same choices of θ_t and N as above.

2. We wish to obtain a numerical posterior over all the variables of the errors-in-variables model presented in the lecture. The model has the following conditional structure, with known σ_x, σ_y (and no intrinsic scatter) for $i = 1, \ldots, N$:

   $$y_i^{\rm obs} \mid \boldsymbol{\theta}, x_i \sim \mathcal{N}\left(\theta_0 + \theta_1 x_i, \sigma_y^2\right), \tag{1}$$

   $$x_i^{\rm obs} \mid x_i \sim \mathcal{N}\left(0, \sigma_x^2\right), \tag{2}$$

   $$x_i \sim \mathcal{N}(x_0, R_x^2), \tag{3}$$

   $$x_0 \sim \mathcal{N}(\mu_{x_0}, \sigma_{x_0}^2), \tag{4}$$

   $$R_x^2 \sim \text{Inv-Gamma}(\alpha_R, \beta_R), \tag{5}$$

   where we have chosen conjugate priors for the population-level parameters x_0, R_x^2 (with fixed hyper-parameters $\mu_{x_0}, \sigma_{x_0}^2, \alpha_R, \beta_R$) to enable analytical computation of the relevant conditional distributions.

Choose fiducial values of the variables $\boldsymbol{\theta}, x_0, R_x$ (and sensible values of σ_x^2, σ_y^2, ensuring that $\sigma_x \approx R_x$) and generate synthetic data from the model for $N = 100$. Then perform numerical inference on *all* variables, using one of the two following approaches:

(a) Sample from the 4-dimensional marginal posterior $\Pr\{\boldsymbol{\theta}, x_0, R_x | \boldsymbol{d}\}$ after analytical marginalization over the latent x_i; then obtain marginal posterior distributions for x_i via conditional sampling. Simple Metropolis-Hastings should suffice in this case, but you can go overboard and use e.g. the `emcee` package[1].

(b) Sample from the full $4 + N$ dimensional posterior using Gibbs sampling. The explicit sampling scheme is given below.

Plot posterior marginals for all parameters, and compare the posterior for the latent variables with the likelihood.

Gibbs sampling for the errors-in-variables model

The Gibbs sampler proceeds as follows (remembering that after each conditional sampling step, the value of the variable just sampled is updated in the conditional for the next step):

(a) Sample the latent variables, x_i ($i = 1, \ldots, N$): the conditional distribution for x_i is obtained as, with $\boldsymbol{d}_i = \{x_i^{\rm obs}, y_i^{\rm obs}\}$, $\Xi = \{\boldsymbol{\theta}, x_0, R_x\}$:

$$\Pr\{x_i \mid \boldsymbol{d}, \Xi\} \propto \Pr\{x_i, \boldsymbol{d}, \Xi\} \propto \Pr\{\boldsymbol{d} | x_i, \Xi\} \Pr\{x_i, \Xi\} = \Pr\{\boldsymbol{d} | x_i, \boldsymbol{\theta}\} \Pr\{x_i | x_0, R_x\},$$

where the first term can be understood as a 'likelihood' (in that it pulls the conditional value of x_i towards the observations), and the second as a 'prior'. The first term comes from the edges connecting the x_i node to its children, and is given by

$$\Pr\{\boldsymbol{d} | x_i, \boldsymbol{\theta}\} = \mathcal{N}_{y_i^{\rm obs}}\left(\theta_0 + \theta_1 x_i, \sigma_y^2\right) \mathcal{N}_{x_i^{\rm obs}}\left(x_i, \sigma_x^2\right)$$

where in the above the value of x_i being conditioned upon is that from the previous iteration of the Gibbs' sampler. Altogether, the full conditional for the latent x_i takes the form:

$$\Pr\{x_i \mid \text{rest}\} \propto \mathcal{N}(y_i^{\rm obs} \mid \theta_0 + \theta_1 x_i, \sigma_y^2) \cdot \mathcal{N}(x_i^{\rm obs} \mid x_i, \sigma_x^2) \cdot \mathcal{N}(x_i \mid x_0, R_x^2).$$

This can be recast as:

$$x_i \sim \mathcal{N}\left(\mu_x, \sigma_x^2\right),$$

where:

$$\sigma_x^2 = \left(\frac{\theta_1^2}{\sigma_y^2} + \frac{1}{\sigma_x^2} + \frac{1}{R_x^2}\right)^{-1},$$

[1] Available as a Python implementation at `https://emcee.readthedocs.io/en/stable/`.

$$\mu_x = \sigma_x^2 \left(\frac{\theta_1 (y_i^{\text{obs}} - \theta_0)}{\sigma_y^2} + \frac{x_i^{\text{obs}}}{\sigma_x^2} + \frac{x_0}{R_x^2} \right).$$

(b) By a similar argument, the regression coefficients θ_0 and θ_1 only depend on their parent node (the top-level prior) and their children node, y_i^{obs} (y_i can be eliminated as it is deterministically related to x_i and $\boldsymbol{\theta}$).Therefore, in the second step of the sampler the regression coefficient are sampled jointly (in blocked step) from:

$$\Pr\{\boldsymbol{\theta} \mid \text{rest}\} \propto \prod_i \mathcal{N}(y_i^{\text{obs}} \mid \theta_0 + \theta_1 x_i, \sigma_y^2) \cdot \mathcal{N}(\theta_0 \mid \mu_\theta, \sigma_\theta^2) \cdot \mathcal{N}(\theta_1 \mid \mu_\theta, \sigma_\theta^2)$$

where we have chosen a Normal prior for both θ_0 and θ_1, out of analytical convenience. Using the results for the Gaussian linear model, this distribution can be written as bivariate normal:

$$\Pr\{\boldsymbol{\theta} \mid \text{rest}\} \propto \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right),$$

where:

$$\boldsymbol{\mu}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \left(\frac{1}{\sigma_y^2} \boldsymbol{A}^\top \boldsymbol{y}^{\text{obs}} + \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 \right),$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \left(\frac{1}{\sigma_y^2} \boldsymbol{A}^\top \boldsymbol{A} + \boldsymbol{\Sigma}_0^{-1} \right)^{-1},$$

where the normalized design matrix has entries $A_{ij} = (\delta_{i1} + \delta_{i2} x_i)/\sigma_y^2$ for $j = 1, \ldots, N$, $\boldsymbol{\mu}_0 = \{\mu_\theta, \mu_\theta\}$ and $\boldsymbol{\Sigma}_0 = \text{diag}(\sigma_\theta^2, \sigma_\theta^2)$.

(c) Sample the population-level mean of x_i, x_0:

$$\Pr\{x_0 \mid \text{rest}\} \propto \prod_i \mathcal{N}(x_i \mid x_0, R_x^2) \cdot \mathcal{N}(x_0 \mid \mu_{x_0}, \sigma_{x_0}^2),$$

which can be written as:

$$x_0 \sim \mathcal{N}(\mu'_{x_0}, \sigma'^2_{x_0}),$$

where:

$$\sigma'^2_{x_0} = \left(\frac{N}{R_x^2} + \frac{1}{\sigma_{x_0}^2} \right)^{-1},$$

$$\mu'_{x_0} = \sigma'^2_{x_0} \left(\frac{\sum_i x_i}{R_x^2} + \frac{\mu_{x_0}}{\sigma_{x_0}^2} \right).$$

(d) Sample the variance of x_i, R_x^2:

$$\Pr\{R_x^2 \mid \text{rest}\} \propto \prod_i \mathcal{N}(x_i \mid x_0, R_x^2) \cdot \text{Inv-Gamma}(\alpha_R, \beta_R),$$

which is:

$$R_x^2 \sim \text{Inv-Gamma}\left(\alpha_R + \frac{N}{2}, \beta_R + \frac{1}{2} \sum_i (x_i - x_0)^2 \right).$$

3. Consider again the coin tossing problem in question 1, but this time from the point of view of Bayesian model comparison: we wish to compare a model M_0 of a fair coin ($\theta = 1/2$) with an alternative M_1, where θ is a free parameter (with a prior).

   (a) Using a conjugate Beta(α, β) prior for θ, compute the Bayes factor after N flips. Starting from equal model probabilities, plot the posterior probability for model M_1 as a function of the number of successes r for $N = \{10, 100, 1000\}$ for the following choices of hyperparameters: $\alpha = \beta = 1$ (uniform prior); $\alpha = \beta = 1/2$ (Jeffreys prior); $\alpha = 2, \beta = 1$ (skewed prior).

   (b) Wilks' theorem says that, asymptotically, the likelihood ratio test statistics

   $$\lambda = -2 \ln \left(\frac{L(\theta = 1/2)}{L(\theta_{\mathrm{MLE}})} \right),$$

   is approximately distributed as a χ^2 distribution with one degree of freedom. Use this to construct a hypothesis test for $H_0 : \theta = 1/2$ vs $H_1 : \theta \neq 1/2$. Compare the strength of evidence against H_0 as measured by the p-value to the posterior model probability for M_1 (for the case of a uniform prior) by plotting the latter vs the former for the above cases. Comment on the difference, paying particular attention to whether you believe that the asymptotic limit has been reached (how can you check for that?).

   (c) To compare the two approaches, proceed as follows (assuming a uniform prior for θ under M_1), repeating the below procedure for $N = \{10, 100, 1000\}$:

      i. generate a 'true' value of θ by drawing from the equal mixture model

      $$\theta_t \sim \frac{1}{2} \delta \left(\frac{1}{2} - \theta \right) + \frac{1}{2} \Pr\{\theta \mid M_1\};$$

      ii. for each θ_t generate $K = 100$ mock data realizations from Binomial(N, θ); for each data realization, compute the Bayes factor and the likelihood ratio.

      iii. for the logarithm of the Bayes factor, select a decision threshold $\ln B_{\mathrm{th}}$ along a uniform grid in the range $[-5, 5]$; for the likelihood ratio, select a significance value α along a log-uniform grid in the range $\alpha \in [10^{-5}, \ldots, 10^{-1}]$. Use each decision threshold (Bayesian) or significance value (Frequentist) to decide whether to favour M_0 or M_1 (Bayesian) or whether to reject/fail to reject the null (Frequentist).

      iv. Once you have made the decision for all the $K = 100$ data, count the number of true positive and false positive. Move to a different decision threshold and repeat. Plot a few examples of true positive rate (i.e., the number of true positive divided by K) vs false positive rate as a function of the decision threshold, both Bayesian and Frequentist. This is the Receiver-operating characteristic (ROC) curve for each method, and the area under the curve (AUC) is a measure of the method's performance: the closer to 1, the better. Save the AUC

for each θ_t and for each method (if you want to produce a cool plot at the end, save the whole ROC curve for each).

v. repeat for several random choices of θ_t – this performs an average under the Bayesian prior. Compare the resulting average AUC $\pm$ standard deviation for each method (if you want, you can also compare the average ROC curves and their standard deviation). Comment on which method is best.