Statistical Methods - Bayesian Inference

Lecture 3

Mattias Villani

Sveriges Riksbank and Stockholm University

March 18, 2010

Prediction

- We may use the estimated model for forecasting a future observation $\tilde{y}$.
- *Posterior predictive distribution* (y denotes available data at the time of forecasting)

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta, y)p(\theta|y)d\theta = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta$$

  where the last step holds if $p(\tilde{y}|\theta, y) = p(\tilde{y}|\theta)$.
- The uncertainty that comes from not knowing θ is represented in $p(\tilde{y}|y)$ by averaging over $p(\theta|y)$.

Prediction Bernoulli data

- Let $y = \sum_{i=1}^{n} y_i$ and $\tilde{y}$ the outcome of the next trial

$$
\begin{aligned}
p(\tilde{y} &= 1|y) = \int_\theta p(\tilde{y} = 1|\theta) p(\theta|y) d\theta \\
&= \int_\theta \theta p(\theta|y) d\theta = E_{\theta|y}(\theta) = \frac{\alpha + y}{\alpha + \beta + n}.
\end{aligned}
$$

- Uniform prior ($\alpha = \beta = 1$)

$$
p(\tilde{y} = 1|y) = \frac{y+1}{n+2}.
$$

Prediction Normal data with known variance

- Assume the uniform prior $p(\theta) \propto c$.

$$p(\tilde{y}|y) = \int_\theta p(\tilde{y}|\theta)p(\theta|y)d\theta$$

where

$$\begin{aligned} \theta|y &\sim N(\bar{y}, \sigma^2/n) \\ \tilde{y}|\theta &\sim N(\theta, \sigma^2) \end{aligned}$$

Simulate from the predictive distribution - Normal model

1. Generate a posterior draw of θ ($\theta^{(1)}$) from $N(\bar{y}, \sigma^2/n)$
2. Generate a draw of $\tilde{y}$ ($\tilde{y}^{(1)}$) from $N(\theta^{(1)}, \sigma^2)$ (note the mean)
3. Repeat steps 1 and 2 a large number of times (N) with the result:
   - Sequence of posterior draws: $\theta^{(1)}, ..., \theta^{(N)}$
   - Sequence of predictive draws: $\tilde{y}^{(1)}, ..., \tilde{y}^{(N)}$.

Predictive distribution - Normal model and uniform prior

- $\theta^{(1)} = \bar{y} + \varepsilon^{(1)}$, where $\varepsilon^{(1)} \sim N(0, \sigma^2/n)$. (Step 1).
- $\tilde{y}^{(1)} = \theta^{(1)} + \upsilon^{(1)}$, where $\upsilon^{(1)} \sim N(0, \sigma^2)$. (Step 2).
- $\tilde{y}^{(1)} = \bar{y} + \varepsilon^{(1)} + \upsilon^{(1)}$.
- $\varepsilon^{(1)}$ and $\upsilon^{(1)}$ are independent.
- The sum of two normal random variables follows a normal distribution, so $\tilde{y}$ follows a normal distribution with

$$\begin{aligned} E(\tilde{y}|y) &= E(\tilde{y}|y) = \bar{y} \\ V(\tilde{y}|y) &= \frac{\sigma^2}{n} + \sigma^2 = \sigma^2\left(1 + \frac{1}{n}\right). \end{aligned}$$

- Note that the estimation uncertainty (σ^2/n) is typically much less important than the intrinsic population uncertainty, σ^2.

Predictive distribution - Normal model and normal prior

- It easy to see that the predictive distribution is normal.
- The mean can be obtained from

$$E_{\tilde{y}|\theta}(\tilde{y}|\theta) = \theta$$

and then remove the conditioning on θ by averaging over θ

$$E(\tilde{y}|y) = E_{\theta|y}(\theta) = \mu_n \text{ (Posterior mean of } \theta\text{).}$$

- The predictive variance of $\tilde{y}$ can be obtained from the conditional variance formula

$$\begin{aligned} V(\tilde{y}|y) &= E_{\theta|y}[V_{\tilde{y}|\theta}(\tilde{y}|\theta)] + V_{\theta|y}[E_{\tilde{y}|\theta}(\tilde{y}|\theta)] \\ &= E_{\theta|y}(\sigma^2) + V_{\theta|y}(\theta) \\ &= \sigma^2 + \tau_n^2 \\ &= \text{(Population variance + Posterior variance of } \theta\text{).} \end{aligned}$$

- In summary:

$$\tilde{y}|y \sim N(\mu_n, \sigma^2 + \tau_n^2).$$

Marginalization

- Models usually contains several parameter $\theta_1, \theta_2, ...$ Examples: $x_i \overset{iid}{\sim} N(\theta, \sigma^2)$; multiple regression ...
- The Bayesian computes the joint posterior distribution

$$p(\theta_1, \theta_2, ..., \theta_p|y) \propto p(y|\theta_1, \theta_2, ..., \theta_p)p(\theta_1, \theta_2, ..., \theta_p).$$

... or in vector form:

$$p(\theta) \propto p(y|\theta)p(\theta).$$

- Complicated to graph the joint posterior.
- Some of the parameters may not be of direct interest (nuisance parameters), but are nevertheless needed in the model.
- No problem: just integrate them out (marginalize with respect to, average over) all nuisance parameters.
- Example: $\theta = (\theta_1, \theta_2)'$, where θ_2 is a nuisance. We are interested in the marginal posterior of θ_1

$$p(\theta_1|y) = \int p(\theta_1, \theta_2|y)d\theta_2 = \int p(\theta_1|\theta_2, y)p(\theta_2|y)d\theta_2.$$

Normal model with unknown variance - Uniform prior

- Model:

$$y, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

- Prior

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$$

- Posterior:

$$\mu|\sigma^2, y \sim N\left(\bar{y}, \frac{\sigma^2}{n}\right)$$
$$\sigma^2|y \sim \mathrm{Inv} - \chi^2(n-1, s^2),$$

where

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

is the usual sample variance.

Normal model with unknown variance - Uniform prior, cont.

- Simulating the posterior of the normal model with non-informative prior:
    1. Draw $X \sim \chi^2(n-1)$
    2. Compute $\sigma^2 = \frac{(n-1)s^2}{X}$ (this a draw from $\text{Inv-}\chi^2(n-1, s^2)$)
    3. Draw a μ from $N\left(\bar{y}, \frac{\sigma^2}{n}\right)$ conditional on the previous draw σ^2
    4. Repeat step 1-3 many times.
- The sampling is implemented in the R program NormalNonInfoPrior.R
- We may derive the marginal posterior analytically as

$$\mu | y \sim t_{n-1}\left(\bar{y}, \frac{s^2}{n}\right).$$

- Normal model with unknown variance:

$$y, ..., y_n \overset{iid}{\sim} N(\mu, \sigma^2)$$

- Prior

$$\mu \sim N\left(\mu_0, \tau_0^2\right)$$
$$\sigma^2 \sim \mathrm{Inv} - \chi^2(\nu_0, \sigma_0^2)$$

- We can no longer obtain the posterior using analytical methods ...
- ... but we do know the two **conditional** posteriors:

$$\mu | y, \sigma^2 \sim N\left(\mu_n, \tau_n^2\right)$$
$$\sigma^2 | y, \mu \sim \mathrm{Inv} - \chi^2(\nu_n, \sigma_n^2).$$

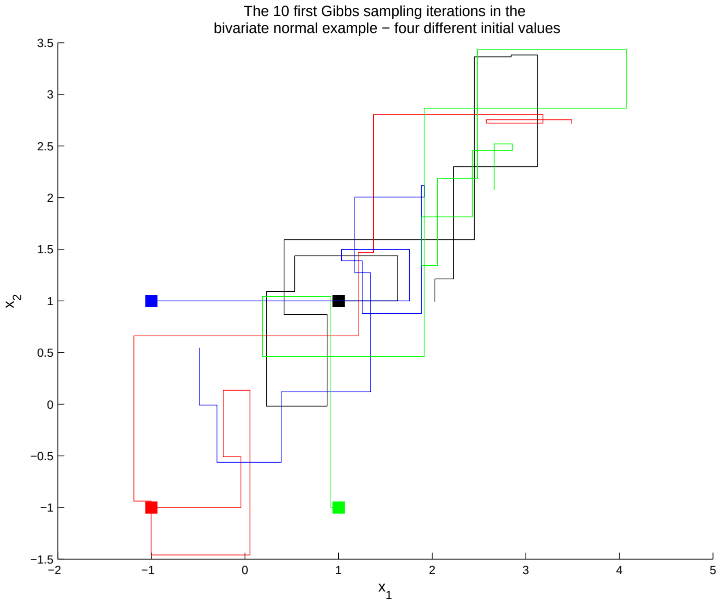
Gibbs sampling the Normal model with semi-conjugate prior

- Idea of Gibbs sampling: simulate iteratively from the two conditional posteriors:

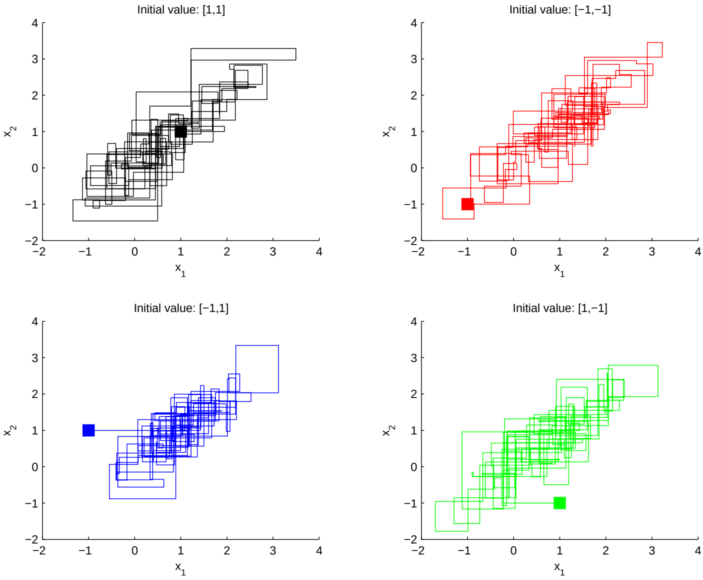
$$\mu|y,\sigma^2 \sim N\left(\mu_n,\tau_n^2\right)$$
$$\sigma^2|y,\mu \sim \mathrm{Inv}-\chi^2(\nu_n,\sigma_n^2).$$

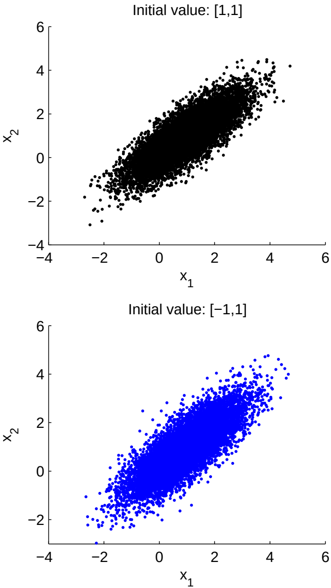
- General case with more than two blocks of parameters: Same idea, simulate from the posterior conditional on **all** other parameters.
- **Gibbs sampling algorithm**
  1. Initialize $\sigma^2_{(0)}$ with s^2.
  2. Draw $\mu_{(1)}$ from the conditional posterior $N\left(\mu_n,\tau_n^2\right)$, conditioning on $\sigma^2_{(0)}$.
  3. Draw $\sigma^2_{(1)}$ from the conditional posterior $\mathrm{Inv}\text{-}\chi^2(\nu_n,\sigma_n^2)$, conditioning on the previously generated $\mu_{(1)}$
  4. Repeat step 1-3, always conditioning on the most recent draw of the conditioning parameter.

Example Gibbs



Example Gibbs, cont.





Initial value: [1,1]
Initial value: [−1,−1]
Initial value: [−1,1]
Initial value: [1,−1]
x1
x2
6
4
2
0
−2
−4
−4
−2
0
2
4
6

Multinomial model with Dirichlet prior

- *Data*: $y = (y_1, ...y_K)$, where y_k counts the number of observations in the kth category. $\sum_{k=1}^{K} y_k = n$. Example: brand choices.
- Multinomial model:

$$p(y|\theta) \propto \prod_{k=1}^{K} \theta_k^{y_k}, \text{ where } \sum_{k=1}^{K} \theta_j = 1.$$

- *Conjugate prior*: $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$

$$p(\theta) \propto \prod_{k=1}^{K} \theta_j^{\alpha_j - 1}.$$

- Moments of $\theta = (\theta_1, ..., \theta_K)' \sim \mathit{Dirichlet}(\alpha_1, ..., \alpha_K)$

$$\mathrm{E}(\theta_k) = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j}$$

$$\mathrm{V}(\theta_k) = \frac{\mathrm{E}(\theta_k)\left[1 - \mathrm{E}(\theta_k)\right]}{1 + \sum_{k=1}^{K} \alpha_k}$$

- Note that $\sum_{k=1}^{K} \alpha_k$ is the precision (inverse variance).

- 'Non-informative': $\alpha_1 = ... = \alpha_K = 1$ (uniform and proper).
- Simulating from the Dirichlet distribution:
  - Generate $x_1 \sim Gamma(\alpha_1, \beta), ..., x_K \sim Gamma(\alpha_K, \beta)$, independently. Any β will do as long it is the same for all x_i.
  - Compute $y_k = x_k / (\sum_{j=1}^{K} x_j)$.
  - $y = (y_1, ..., y_K)$ is a draw from the $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$ distribution.
- *Prior-to-Posterior updating*:

$$
\begin{aligned}
\textit{Model}: &\quad y = (y_1, ... y_K) \sim \mathrm{Multin}(n; \theta_1, ..., \theta_K) \\
\textit{Prior}: &\quad \theta = (\theta_1, ..., \theta_K) \sim \mathrm{Dirichlet}(\alpha_1, ..., \alpha_K) \\
\textit{Posterior}: &\quad \theta | y \sim \mathrm{Dirichlet}(\alpha_1 + y_1, ..., \alpha_K + y_K).
\end{aligned}
$$