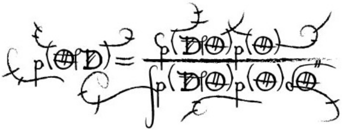

Bayesian Learning 732A46: Lecture 3

Matias Quiroz[1,2]

[1]Division of Statistics and Machine Learning, Linköping University

[2]Research Division, Sveriges Riksbank

April 2016

Lecture overview

- Multiparameter models - direct simulation and marginalization.
- Normal model with unknown variance
- Multinomial model
- Multivariate normal with known covariance matrix

Direct simulation

- Once $p(\theta|y)$ is derived we use it for **posterior analysis**.
- **Direct**: *known distribution* - **Example**: Normal, Beta, Gamma.
- Examples [$\theta \sim p(\theta|y)$ continuous. Replace $\int$ by $\sum$ for discrete θ]

  Expectation: $E(\theta) = \int \theta p(\theta|y) d\theta$

  Variance: $V(\theta) = \int (\theta - E(\theta))^2 p(\theta|y) d\theta$

  Probabilities: $\Pr(\theta \in A) = \int_A p(\theta|y) d\theta$.

  E.g. if $A = \{\theta; \theta \in [0, \infty)\}$ then $\Pr(\theta \leq 2) = \int_0^2 p(\theta|y) d\theta$.
- Note: the function of interest is **averaged over the posterior uncertainty** of the parameters.

Direct simulation, cont.

- Nothing but expectations of a function $h(\theta)$, i.e.

$$E[h(\theta)] = \int h(\theta)p(\theta|y)d\theta.$$

- Expectation: $E(\theta) = \int \theta p(\theta|y)d\theta$. $h(\theta) = \theta$.

  Variance : $V(\theta) = \int (\theta - E(\theta))^2 p(\theta|y)d\theta$. $h(\theta) = (\theta - E(\theta))^2$.

  Probabilities: $\Pr(\theta \in A) = \int_A p(\theta|y)d\theta = \int \mathbb{1}_A(\theta)p(\theta|y)d\theta$. $h(\theta) = \mathbb{1}_A(\theta)$,

$$\mathbb{1}_A(\theta) = \begin{cases} 1, & \text{if } \theta \in A, \\ 0, & \text{if } \theta \notin A, \end{cases}$$

- For **complicated** $h(\theta)$ analytical integration is hard/**impossible**.
- By **simulation** using N draws $\theta^{(i)}$:

$$E[h(\theta)] \approx \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \quad \text{with} \quad \theta^{(i)} \sim p(\theta|y)$$

Direct simulation, cont.

- Expectation: $E(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\theta^{(i)}$.
- Variance : $V(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\theta^{(i)} - \bar{\theta}\right)^2$.
- Probabilities: $\Pr(\theta \in A) \approx \frac{\left\{\#\theta^{(i)} \in A\right\}}{N}$
- Want the **posterior distribution** of $\phi = h(\theta)$, i.e. $p(\phi|y)$?
- **Histogram** (or **Kernel density estimate**) of $h(\theta^{(i)})$ is an approximation.
- Posterior analysis by *direct simulation* is **easy**...
- ... the **difficult** part is to make *direct simulation* **possible**.
- **Note**: *Direct simulation* **requires** that you can **analytically derive** what you "directly simulate"!

Multiparameter models

- **Examples**
  1. Normal model with **both** μ and σ^2 unknown.
  2. Multiple regression models $(\beta_1, \ldots, \beta_p)$.

- Five **invaluable techniques** when working with multiparameters. Generalize easily to $p > 2$ parameters (**try it at home**!)

- **Invaluable technique** $\#1$: Simulation in multiparameter models
  - $p(\theta_1, \theta_2|y)$ - **impossible** with direct simulation
  - $p(\theta_1, \theta_2|y) = p(\theta_1|\theta_2, y)p(\theta_2|y)$ - Each piece **possible** with direct simulation

- **Invaluable technique** $\#2$: How to derive $p(\theta_1|\theta_2, y)$ analytically?
  - Note that θ_2 is **treated as a constant** here!

$$p(\theta_1|\theta_2, y) = \frac{p(\theta_1, \theta_2|y)}{p(\theta_2|y)} \propto p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)p(\theta_1, \theta_2).$$

  - The joy of **ignoring a normalizing constant** applies also for θ_2.

- **Invaluable technique** #3: How to derive $p(\theta_2|y)$ analytically?
  - $p(\theta_2|y) = \int p(\theta_1, \theta_2|y) d\theta_1$ - **can make you cry**
  - **Much** easier to use

$$p(\theta_2|y) = \frac{p(\theta_1, \theta_2|y)}{p(\theta_1|\theta_2, y)} \propto \frac{p(y|\theta_1, \theta_2)p(\theta_1, \theta_2)}{p(\theta_1|\theta_2, y)} \quad (1)$$

    Standard trick:
    LHS of (1) **does not depend** on θ_1 ($\implies$ must cancel on **RHS**). Insert a θ_1 that simplifies (1).

  - Note: Analytical derivations are **not always** possible!

Multiparameter models, cont.

- **Invaluable technique** #4: Are some of your parameters **nuisance** (not of direct interest)? **Example:** I only care about θ_1 (θ_2 nuisance).
  - Computing

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

    analytically can make you cry...
  - ... but computing it by simulation can **can make you smile**

$$\begin{aligned} \theta_2^{(i)} &\sim p(\theta_2|y) \\ \theta_1^{(i)}|\theta_2^{(i)} &\sim p(\theta_1|\theta_2^{(i)}, y) \end{aligned}$$

  - **Histogram** (or **Kernel density estimate**) of $\theta_1^{(i)}$ is an approximation of $p(\theta_1|y)$.
  - This is **marginalization by simulation**.
- **Invaluable technique** #5: Interested in **nasty integrals**, e.g.

$$\Pr(\theta_1 > \theta_2|y) = \int\int_{\theta_1 > \theta_2} p(\theta_1, \theta_2|y)d\theta_1 d\theta_2?$$

Remember **the joy** of simulating!

Normal model with unknown variance - Uniform prior

- **Model**

$$y_1, ..., y_n \overset{iid}{\sim} N(\theta, \sigma^2)$$

- **'Non-informative' Prior**

$$p(\theta, \sigma^2) \propto (\sigma^2)^{-1} \quad \left[\text{uniform in} \quad p(\theta, \log(\sigma^2)) \propto c\right]$$

- **Posterior**. Decompose using technique #1,

$$\begin{aligned} \theta|\sigma^2, y &\sim N\left(\bar{y}, \frac{\sigma^2}{n}\right) & (2) \\ \sigma^2|y &\sim \text{Inv-}\chi^2(\nu_n, s_n^2) \quad , & (3) \end{aligned}$$

  where

$$\nu_n = n - 1 \quad \text{and} \quad s_n^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

  is the usual sample variance.

- (2) - derived in **Lecture 1**. Uses technique #2.
- (3) - **White board**. Uses technique #3.

Normal model with unknown variance - Uniform prior, cont.

- $\sigma_n^2 \sim \text{Inv-}\chi^2(\nu_n, s_n^2)$ if

$$p(\sigma^2) \propto \sigma^{-2(\nu_n/2+1)} \exp\left(-\frac{\nu_n s_n^2}{2\sigma^2}\right).$$

- By technique #3

$$p(\sigma^2|y) \propto \frac{p(y|\theta,\sigma^2)p(\theta,\sigma^2)}{p(\theta|\sigma^2,y)} = \frac{p(y|\theta,\sigma^2)(\sigma^2)^{-1}}{\mathcal{N}(\theta|\bar{y},\sigma^2/n)}$$

- **Important**: **As a function** of σ^2 [**at** $\theta = \bar{y}$]
    1. $\mathcal{N}(\theta|\bar{y},\sigma^2/n) \propto (\sigma^{-2})^{-1/2}$
    2. $p(y|\theta,\sigma^2)(\sigma^2)^{-1} \propto (\sigma^{-2})^{n/2+1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\bar{y})^2\right)$
- 1./2. gives

$$\sigma^{-2(\frac{(n-1)}{2}+1)} \exp\left(-\frac{\overbrace{n-1}^{\nu_n}}{2\sigma^2}\overbrace{\frac{1}{n-1}\sum_{i=1}^n (y_i-\bar{y})^2}^{s_n^2}\right).$$

Normal model with unknown variance - Uniform prior, cont.

- **Simulating** the posterior. Uses technique #1.
  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 θ 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

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

  or plot the histogram of only θ [technique #4] from the simulation above.
- **Homework** (if you want): follow the techniques to derive the posterior when

$$\begin{aligned} p(\mu|\sigma^2) &= \mathcal{N}(\mu_0, \sigma^2/\kappa_0) \\ p(\sigma^2) &= 1/\sigma^2. \end{aligned}$$

Multinomial model with Dirichlet prior

- **Easier** - can simulate from $p(\theta_1, \ldots, \theta_K|y)$ directly. No decomposition needed.
- **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)**: iPhone, Android, Blackberry, other ($K = 4$)
- **Multinomial model**:

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

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

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

Multinomial model with Dirichlet prior

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

$$\mathrm{E}(\theta_k) = \frac{\alpha_k}{\sum_{j=1}^{K} \alpha_j} \quad \text{and} \quad \mathrm{V}(\theta_k) = \frac{\mathrm{E}(\theta_k)\left[1 - \mathrm{E}(\theta_k)\right]}{1 + \sum_{j=1}^{K} \alpha_j}.$$

- Note that $\sum_{j=1}^{K} \alpha_j$ is a **precision** parameter.
- **'Non-informative'**: $\alpha_1 = ... = \alpha_K = 1$ (uniform and proper).
- **Simulating** from the Dirichlet distribution:
  1. Generate $x_1 \sim \mathrm{Gamma}(\alpha_1, 1), ..., x_K \sim \mathrm{Gamma}(\alpha_K, 1)$.
  2. Compute $y_k = x_k / (\sum_{j=1}^{K} x_j)$.
  3. $y = (y_1, ..., y_K)$ is a draw from the $\mathrm{Dirichlet}(\alpha_1, ..., \alpha_K)$ distribution.
- **Prior-to-Posterior updating**:

Model	**Prior**	$\rightarrow$	**Posterior**
Mult	$\theta \sim \mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_K)$	$\rightarrow$	$\theta \mid y \sim \mathrm{Dirichlet}(\alpha_1 + y_1, ..., \alpha_K + y_k)$

Multivariate normal - known Σ

- **Model**

$$y_1, \ldots, y_n \overset{iid}{\sim} \mathcal{N}_p(\mu, \Sigma)$$

where Σ is a **known** covariance matrix.

- **Density**

$$p(y|\mu, \Sigma) = |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right).$$

- **Likelihood**

$$\begin{aligned} p(y_1, \ldots, y_n|\mu, \Sigma) &\propto |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right) \\ &= |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\Sigma^{-1}S_\mu\right)\right), \end{aligned}$$

where $S_\mu = \sum_{i=1}^{n}(y_i-\mu)(y_i-\mu)'$.

Multivariate normal - known Σ. Informative prior on μ

- **Prior**

$$\mu \sim \mathcal{N}_p(\mu_0, \Lambda_0).$$

- **Posterior**

$$\mu|y \sim \mathcal{N}_p(\mu_n, \Lambda_n),$$

where

$$\begin{aligned}
\Lambda_n^{-1} &= \Lambda_0^{-1} + n\Sigma^{-1} \\
\mu_n &= (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{y}).
\end{aligned}$$

- **Prior precision**: Λ_0^{-1}. **Data precision**: $n\Sigma^{-1}$.
- **Note**: the posterior mean is a (matrix) **weighted average** of prior and data information.
- **Noninformative prior**: let the precision go to zero: $\Lambda_0^{-1} \to 0$.