Stat 535 C - Statistical Computing & Monte Carlo Methods

Arnaud Doucet

Email: arnaud@cs.ubc.ca
• Suggested Projects:

www.cs.ubc.ca/~arnaud/projects.html

• First assignment on the web: capture/recapture.

• Additional articles have been posted.
2.1– Outline

- Prior distributions: conjugate, maxent, Jeffrey’s.

- Bayesian variable selection.
3.1– How to Select the Prior Distribution?

- Once the prior distribution is specified, inference using Bayes can be performed almost “mechanically”.

- Omitting computational issues, the most critical and criticized point is the choice of the prior.

- Seldom, the available observation is precise enough to lead to an exact determination of the prior distribution.
3.1– How to Select the Prior Distribution?

- Prior includes subjectivity.

- Subjectivity does not mean being nonscientific: vast amount of scientific information coming from theoretical and physical models is guiding specification of priors.

- In the last decades, a lot of research has focused on un-informative and robust priors.
3.2– Conjugate Priors

• Conjugate priors are the most commonly used priors.

• A family of probability distributions $\mathcal{F}$ on $\Theta$ is said to be conjugate for a likelihood function $f(x|\theta)$ if, for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to $\mathcal{F}$.

• In simpler terms, the posterior remains admits the same functional form as the prior and only its parameters are changed.
3.3– Example: Gaussian with unknown mean

• Assume you have observations \( X_i | \mu \sim \mathcal{N}(\mu, \sigma^2) \) and \( \mu \sim \mathcal{N}(m_0, \sigma_0^2) \) then

\[
\mu | x_1, ..., x_n \sim \mathcal{N}(m_n, \sigma_n^2)
\]

where

\[
\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \Rightarrow \sigma_n^2 = \frac{\sigma_0^2 \sigma^2}{n \sigma_0^2 + \sigma^2},
\]

\[
m_n = \sigma_n^2 \left( \frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{m}{\sigma_0^2} \right) = \sigma_n^2 \left( \frac{\sum_{i=1}^n x_i}{\sigma^2} + \frac{m}{\sigma_0^2} \right).
\]

• One can think of the prior as \( n_0 \) virtual observations with \( n_0 = \frac{\sigma^2}{\sigma_0^2} \) and

\[
m_n = \frac{n \sum_{i=1}^n x_i + n_0 m_0}{n + n_0}.
\]
3.4– Example: Gaussian with unknown mean and variance

- Assume you have observations $X_i \mid (\mu, \sigma^2) \sim \mathcal{N}(\mu, \sigma^2)$ and

$$
\pi(\mu, \sigma^2) = \pi(\sigma^2) \pi(\mu \mid \sigma^2)
$$

$$
= \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\gamma_0}{2}\right) \mathcal{N}(\mu; m_0, \delta^2 \sigma^2)
$$

- We have

$$
\mu, \sigma^2 \mid x_1, \ldots, x_n \sim \mathcal{IG}\left(\frac{\nu_0 + n}{2}, \frac{\gamma_0 + \sum_{i=1}^n x_i^2 - (m_n/\sigma_n)^2}{2}\right)
$$

$$
\times \mathcal{N}(\mu; m_n, \sigma_n^2)
$$

where

$$
m_n = \frac{1}{\delta^{-2} + n} \left( \frac{m_0^2}{\delta^2} + \sum_{i=1}^n x_i \right), \quad \sigma_n^2 = \frac{\sigma^2}{\delta^{-2} + n},
$$
3.4– Example: Gaussian with unknown mean and variance

• Assume you have some counting observations $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{P}(\theta)$; i.e.

$$f(x_i|\theta) = e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

• Assume we adopt a Gamma prior for $\theta$; i.e. $\theta \sim \mathcal{G}(\alpha, \beta)$

$$\pi(\theta) = \mathcal{G}(\theta; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta \theta}.$$

• We have

$$\pi(\theta|x_1, \ldots, x_n) = \mathcal{G}(\theta; \alpha + \sum_{i=1}^{n} x_i, \beta + n).$$

• You can think of the prior as having $\beta$ virtual observations who sum to $\alpha$. 

– Prior Distributions
Many likelihood do not admit conjugate distributions BUT it is feasible when the likelihood is in the exponential family

\[ f(x|\theta) = h(x) \exp(\theta^T x - \Psi(\theta)) \]

and in this case the conjugate distribution is (for the hyperparameters \( \mu, \lambda \))

\[ \pi(\theta) = K(\mu, \lambda) \exp(\theta^T \mu - \lambda \Psi(\theta)). \]

It follows that

\[ \pi(\theta|x) = K(\mu + x, \lambda + 1) \exp(\theta^T (\mu + x) - (\lambda + 1) \Psi(\theta)). \]
3.5– Limitations

- The conjugate prior can have a strange shape or be difficult to handle.

- Consider

\[
\Pr(y = 1 | \theta, x) = \frac{\exp(\theta^T x)}{1 + \exp(\theta^T x)}
\]

then the likelihood for \( n \) observations is exponential conditional upon \( x_i \)'s as

\[
f(y_1, \ldots, y_n | x_1, \ldots, x_n, \theta) = \exp \left( \theta^T \sum_{i=1}^{n} y_i x_i \right) \prod_{i=1}^{n} (1 + \exp(\theta^T x_i))^{-1}
\]

and

\[
\pi(\theta) \propto \exp(\theta^T \mu) \prod_{i=1}^{n} (1 + \exp(\theta^T x_i))^{-\lambda}
\]
• If you have a prior distribution $\pi(\theta)$ which is a mixture of conjugate distributions, then the posterior is in closed form and is a mixture of conjugate distributions; i.e. with

$$\pi(\theta) = \sum_{i=1}^{K} w_i \pi_i(\theta)$$

then

$$\pi(\theta|x) = \frac{\sum_{i=1}^{K} w_i \pi_i(\theta) f(x|\theta)}{\sum_{i=1}^{K} w_i \int \pi_i(\theta) f(x|\theta) d\theta} = \sum_{i=1}^{K} w'_i \pi_i(\theta|x)$$

where

$$w'_i \propto w_i \int \pi_i(\theta) f(x|\theta) d\theta, \quad \sum_{i=1}^{K} w'_i = 1.$$ 

• **Theorem** (Brown, 1986): It is possible to approximate arbitrary closely any prior distribution by a mixture of conjugate distributions.
3.7– Pros and Cons of Conjugate Priors

Pros.

• Very simple to handle, easy to interpret (through imaginary observations).

• Some statisticians argue that they are the least “informative” ones.

Cons.

• Not applicable to all likelihood functions.

• Not flexible at all; what is you have a constraint like $\mu > 0$.

• Approximation by mixtures feasible but very tedious and almost never used in practice.
3.8– Invariant Priors

- If the likelihood is of the form

\[ X | \theta \sim f(x - \theta) \]

then \( f(\cdot) \) is translation invariant and \( \theta \) is a *location parameter*.

- An invariance requirement is that the prior distribution should be translation invariant

\[ \pi(\theta) = \pi(\theta - \theta_0) \]

for every \( \theta_0 \); i.e. \( \pi(\theta) = c \).

- This “flat” prior is improper but the resulting posterior is proper as long as

\[ \int f(x - \theta) \, d\theta < \infty. \]
3.8– Invariant Priors

- If the likelihood is of the form

\[ X | \theta \sim \frac{1}{\theta} f \left( \frac{x}{\theta} \right) \]

then \( f(\cdot) \) is scale invariant and \( \theta \) is a scale parameter.

- An invariance requirement is that the prior distribution should be scale invariant; i.e. got any \( c > 0 \)

\[ \pi(\theta) = \frac{1}{c} \pi \left( \frac{\theta}{c} \right). \]

- This implies that the resulting prior is improper

\[ \pi(\theta) \propto \frac{1}{\theta}. \]
3.9– The Jeffreys Prior

- Consider the Fisher information matrix

\[
I(\theta) = \mathbb{E}_{X|\theta} \left[ \frac{\partial \log f(X|\theta)}{\partial \theta} \frac{\partial \log f(X|\theta)^T}{\partial \theta} \right] = -\mathbb{E}_{X|\theta} \left[ \frac{\partial^2 \log f(X|\theta)}{\partial \theta^2} \right].
\]

- The Jeffrey’s prior is defined as

\[
\pi(\theta) \propto |I(\theta)|^{1/2}
\]

- This prior follows from an invariance principle. Let \( \phi = h(\theta) \) and \( h \) be an invertible function with inverse function \( \theta = g(\phi) \) then

\[
\pi(\phi) = \pi(g(\phi)) \left| \frac{dg(\phi)}{d\phi} \right| = \pi(\theta) \left| \frac{d\theta}{d\phi} \right| \propto |I(\phi)|^{1/2}
\]

as

\[
I(\phi) = -\mathbb{E}_{X|\phi} \left[ \frac{\partial^2 \log f(X|\phi)}{\partial \theta^2} \right] = -\mathbb{E}_{X|\theta} \left[ \frac{\partial^2 \log f(X|\phi)}{\partial \theta^2} \cdot \left| \frac{d\theta}{d\phi} \right|^2 \right] = I(\theta) \left| \frac{d\theta}{d\phi} \right|^2.
\]
3.9– The Jeffreys Prior

• Consider \( X|\theta \sim B(n, \theta) \); i.e.

\[
f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x},
\]

\[
\frac{\partial^2 \log f(x|\theta)}{\partial \theta^2} = \frac{x}{\theta^2} + \frac{n-x}{(1-\theta)^2},
\]

\[
I(\theta) = \frac{n}{\theta(1-\theta)}.
\]

• The Jeffreys prior is

\[
\pi(\theta) \propto [\theta (1-\theta)]^{-1/2} = Be\left(\theta; \frac{1}{2}, \frac{1}{2}\right).
\]
3.9– The Jeffreys Prior

• Consider \( X_i \mid \theta \sim N(\theta, \sigma^2) \); i.e.
\[
f(x_{1:n} \mid \theta) \propto \exp \left( - \frac{(\bar{x} - \theta)^2}{2\sigma^2} \right).
\]

• Since
\[
\frac{\partial^2 \log f(x_{1:n} \mid \theta)}{\partial \theta^2} = -\frac{n}{\sigma^2} \Rightarrow \pi(\theta) \propto 1.
\]

• Consider \( X_i \mid \theta \sim N(\mu, \theta) \); i.e.
\[
f(x_{1:n} \mid \theta) \propto \theta^{n/2} \exp \left( -\frac{s}{2\theta} \right)
\]
where \( s = \sum_{i=1}^{n} (x_i - \mu)^2 \). Then
\[
\frac{\partial^2 \log f(x_{1:n} \mid \theta)}{\partial \theta^2} = \frac{n}{2\theta^2} - \frac{s}{\theta^3} \Rightarrow \pi(\theta) \propto \frac{1}{\theta}.
\]
3.10– Pros and Cons of Jeffreys Prior

- It can lead to incoherences; i.e. the Jeffreys’ prior for Gaussian data and $\theta = (\mu, \sigma)$ unknown is $\pi(\theta) \propto \sigma^{-2}$. However if these parameters are assumed a priori independent then $\pi(\theta) \propto \sigma^{-1}$.

- Automated procedure but cannot incorporate any “physical” information.

- It does NOT satisfy the likelihood principle.
3.11– The MaxEnt Priors

- If some characteristics of the prior distributions (moments, etc.) are known and can be written as $K$ prior expectations
  
  $E_{\pi} [g_k (\theta)] = w_k,$

  a way to select a prior $\pi$ satisfying these constraints is the maximum entropy method.

- In a finite setting, the entropy is defined by
  
  $Ent (\pi) = - \sum_{i=1}^{\pi (\theta_i) \log (\pi (\theta_i))}.$

- The distribution maximizing the entropy is of the form
  
  $\pi (\theta_i) = \frac{\exp \left( \sum_{k=1}^{K} \lambda_k g_k (\theta_i) \right)}{\sum_{j=1}^{\exp \left( \sum_{k=1}^{K} \lambda_k g_k (\theta_j) \right)}}$

  where $\{\lambda_k\}$ are Lagrange multipliers.

- However, the constraints might be incompatible; i.e. $E (\theta^2) \geq E^2 (\theta)$. 
• Assume $\Theta = \{0, 1, 2, ...\}$. Suppose that $E_\pi [\theta] = 5$, then

$$
\pi (\theta) = \frac{e^{\lambda_1 \theta}}{\sum_{\theta=0}^{\infty} e^{\lambda_1 \theta}} = (1 - e^{\lambda_1}) e^{\lambda_1 \theta}.
$$

• Maximizing the entropy we find $e^{\lambda_1} = 1/6$, thus

$$
\pi (\theta) = Geo (1/6)
$$

• What about the continuous case???
3.13– The MaxEnt Prior for Continuous Random Variables

- Jaynes argues that the entropy should be defined as the Kullback-Leibler divergence between $\pi$ and some invariant noninformative prior for the problem $\pi_0$; i.e.

$$
Ent(\pi) = - \int \pi_0(\theta) \log \left( \frac{\pi(\theta)}{\pi_0(\theta)} \right) d\theta.
$$

- The maxent prior is of the form

$$
\pi(\theta) = \frac{\exp \left( \sum_{k=1}^{K} \lambda_k g_k(\theta) \right) \pi_0(\theta)}{\int \exp \left( \sum_{k=1}^{K} \lambda_k g_k(\theta) \right) \pi_0(\theta) d\theta}
$$

- Selecting $\pi_0(\theta)$ is not easy!
• Consider a real parameter $\theta$ and set $E_\pi [\theta] = \mu$. We can select $\pi_0 (d\theta) = d\theta$; i.e. the Lebesgue measure.

• In this case

$$\pi (\theta) \propto e^{\lambda \theta}$$

which is a (bad) improper distribution.

• If additionally $Var_\pi [\theta] = \sigma^2$, then you can establish that

$$\pi (\theta) = \mathcal{N} (\theta; \mu, \sigma^2).$$
3.15– Summary

- In most applications, there is “true” prior.

- Although conjugate priors are limited, they remain the most widely used class of priors for convenience and simple interpretability.

- There is a whole literature on the subject: reference & objective priors.

- Empirical Bayes: the prior is constructed from the data.

- In all cases, you should do a sensitivity analysis!!!
Consider the standard linear regression problem

\[ Y = \sum_{i=1}^{p} \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N} (0, 1) \]

Often you might have too many predictors, so this model will be inefficient.

A standard Bayesian treatment of this problem consists of selecting only a subset of explanatory variables.

This is nothing but a model selection problem with \(2^p\) possible models.
A standard way to write the model is

\[ Y = \sum_{i=1}^{p} \gamma_i \beta_i X_i + \sigma V \text{ where } V \sim \mathcal{N}(0, 1) \]

where \( \gamma_i = 1 \) if \( X_i \) is included or \( \gamma_i = 0 \) otherwise. However this suggests that \( \beta_i \) is defined even when \( \gamma_i = 0 \).

A neater way to write such models is to write

\[ Y = \sum_{\{i: \gamma_i = 1\}} \beta_i X_i + \sigma V = \beta_\gamma^T X_\gamma + \sigma V \]

where, for a vector \( \gamma = (\gamma_1, \ldots, \gamma_p) \), \( \beta_\gamma = \{\beta_i : \gamma_i = 1\} \), \( X_\gamma = \{X_i : \gamma_i = 1\} \) and \( n_\gamma = \sum_{i=1}^{p} \gamma_i \).

Prior distributions

\[ \pi_\gamma (\beta_\gamma, \sigma^2) = \mathcal{N} (\beta_\gamma; 0, \delta^2 \sigma^2 I_{n_\gamma}) \mathcal{IG} \left( \sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2} \right) \]

and \( \pi (\gamma) = \prod_{i=1}^{p} \pi (\gamma_i) = 2^{-p} \).
3.18– Bayesian Variable Selection Example

For a fixed model $\gamma$ and $n$ observations $D = \{x_i, y_i\}_{i=1}^{n}$ then we can determine the marginal likelihood and the posterior analytically

$$
\pi_\gamma (D | \beta_\gamma, \sigma^2) = \Gamma \left( \frac{\nu_0 + n}{2} + 1 \right) \delta^{-n_\gamma} |\Sigma_\gamma|^{1/2} \left( \frac{\gamma_0 + \sum_{i=1}^{n} y_i^2 - \mu_\gamma \Sigma_\gamma^{-1} \mu_\gamma}{2} \right)^{-\left( \frac{\nu_0 + n}{2} + 1 \right)}
$$

and

$$
\pi_\gamma (\beta_\gamma, \sigma^2 | D) = \mathcal{N} (\beta_\gamma; \mu_\gamma, \sigma^2 \Sigma_\gamma)
\times \mathcal{IG} \left( \sigma^2; \frac{\nu_0 + n}{2}, \frac{\gamma_0 + \sum_{i=1}^{n} y_i^2 - \mu_\gamma \Sigma_\gamma^{-1} \mu_\gamma}{2} \right)
$$

where

$$
\mu_\gamma = \Sigma_\gamma \left( \sum_{i=1}^{n} y_i x_{\gamma,i} \right), \quad \Sigma_\gamma^{-1} = \delta^{-2} I_{n_\gamma} + \sum_{i=1}^{n} x_{\gamma,i} x_{\gamma,i}^T.
$$
### 3.18– Bayesian Variable Selection Example

- Popular alternative Bayesian models include
  \[ \gamma_i \sim \mathcal{B}(\lambda) \text{ where } \lambda \sim \mathcal{U}[0, 1], \]
  \[ \gamma_i \sim \mathcal{B}(\lambda_i) \text{ where } \lambda_i \sim \mathcal{B}(\alpha, \beta). \]

- g-prior (Zellner)
  \[
  \beta_\gamma | \sigma^2 \sim \mathcal{N}\left(\beta_\gamma; 0, \delta^2 \sigma^2 \left(X^T_\gamma X_\gamma\right)^{-1}\right).
  \]

- Robust models where additionally one has
  \[
  \delta^2 \sim \mathcal{IG}\left(\frac{a_0}{2}, \frac{b_0}{2}\right).
  \]

- Such variations are very important and can modify dramatically the performance of the Bayesian model.
3.19– Bayesian Variable Selection Example

- Caterpillar dataset: 1973 study to assess the influence of some forest settlement characteristics on the development of caterpillar colonies.

- The response variable is the log of the average number of nests of caterpillars per tree on an area of 500 square meters.

- We have $n = 33$ data and 10 explanatory variables
3.20– Bayesian Variable Selection Example

- \( x_1 \) is the altitude (in meters),
- \( x_2 \) is the slope (in degrees),
- \( x_3 \) is the number of pines in the square,
- \( x_4 \) is the height (in meters) of the tree sampled at the center of the square,
- \( x_5 \) is the diameter of the tree sampled at the center of the square,
- \( x_6 \) is the index of the settlement density,
- \( x_7 \) is the orientation of the square (from 1 if southbound to 2 otherwise),
- \( x_8 \) is the height (in meters) of the dominant tree,
- \( x_9 \) is the number of vegetation strata,
- \( x_{10} \) is the mix settlement index (from 1 if not mixed to 2 if mixed).
3.20– Bayesian Variable Selection Example
### Top five most likely models

| \( \pi(\gamma|\mathbf{x}) \) (Ridge \( \delta^2 = 10 \)) | \( \pi(\gamma|\mathbf{x}) \) (g-p \( \delta^2 = 10 \)) | \( \pi(\gamma|\mathbf{x}) \) (g-p, \( \delta^2 \) estimated) |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|
| 0,1,2,4,5/0.1946                                | 0,1,2,4,5/0.2316                                | 0,1,2,4,5/0.0929                                |
| 0,1,2,4,5,9/0.0321                               | 0,1,2,4,5,9/0.0374                              | 0,1,2,4,5,9/0.0325                              |
| 0,12,4,5,10/0.0327                               | 0,1,9/0.0344                                   | 0,1,2,4,5,10/0.0295                             |
| 0,1,2,4,5,7/0.0306                               | 0,1,2,4,5,10/0.0328                             | 0,1,2,4,5,7/0.0231                              |
| 0,1,2,4,5,8/0.0251                               | 0,1,4,5/0.0306                                 | 0,1,2,4,5,8/0.0228                              |