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 this afternoon: capture/recapture.

• Additional articles have been posted.
2.1– Outline

- Bayesian model selection.
- Bayesian linear model and variable selection.
- Extensions.
3.1– Summary of Last Lecture

• Ones wants to compare two hypothesis: \( H_0 : \theta \sim \pi_0 \)
  versus \( H_1 : \theta \sim \pi_1 \) then the prior is

\[
\pi (\theta) = \pi (H_0) \pi_0 (\theta) + \pi (H_1) \pi_1 (\theta)
\]

where \( \pi (H_0) + \pi (H_1) = 1 \).

• One can have in a coin example: \( \pi_0 (\theta) = \mathcal{U} \left[ \frac{1}{2}, 1 \right] \), \( \pi_1 (\theta) = \mathcal{U} \left[ 0, \frac{1}{2} \right] \)
  or \( \pi_0 (\theta) = \delta_{\theta_0} (\theta) \) and \( \pi_1 (\theta) = \mathcal{U} \left[ 0, \frac{1}{2} \right] \) or \( \pi_0 (\theta) = \mathcal{B} \left( \alpha_0, \beta_0 \right) \) and
  \( \pi_1 (\theta) = \mathcal{B} \left( \alpha_1, \beta_1 \right) \).

• To compare \( H_0 \) versus \( H_1 \), we typically compute the Bayes factor
  which partially eliminated the influence of the prior modelling (i.e. \( \pi (H_i) \))

\[
B_{10}^\pi = \frac{\pi (x | H_1)}{\pi (x | H_0)} = \frac{\int f (x | \theta) \pi_1 (\theta) \, d\theta}{\int f (x | \theta) \pi_0 (\theta) \, d\theta}
\]
3.1– Summary of Last Lecture

• You can also compute the posterior probabilities of \( H_0 \) and \( H_1 \)

\[
\pi(H_0 | x) = \frac{\pi(x | H_0) \pi(H_0)}{\pi(x)}
\]

\[
= \frac{\pi(x | H_0) \pi(H_0)}{\pi(x | H_0) \pi(H_0) + \pi(x | H_1) \pi(H_1)}.
\]

• The posterior probabilities satisfy

\[
\frac{\pi(H_1 | x)}{\pi(H_0 | x)} = \frac{\pi(x | H_1) \pi(H_1)}{\pi(x | H_0) \pi(H_0)} = B_{10}^{\pi} \frac{\pi(H_1)}{\pi(H_0)}.
\]
3.1– Summary of Last Lecture

- Testing hypothesis in a Bayesian way is attractive.... but be careful to vague priors!!!

- Assume you have $X \mid (\mu, \sigma^2) \sim \mathcal{N} \left( \mu, \sigma^2 \right)$ where $\sigma^2$ is assumed known but $\mu$ (the parameter $\theta$) is unknown. We want to test $H_0 : \mu = 0$ vs $H_1 : \mu \sim \mathcal{N} \left( \xi, \tau^2 \right)$ then

$$B_{10}^\pi (x) = \frac{\pi (x \mid H_1)}{\pi (x \mid H_0)} = \frac{\int \mathcal{N} \left( x ; \mu, \sigma^2 \right) \mathcal{N} \left( \mu ; \xi, \tau^2 \right) d\mu}{f (x \mid 0)}$$

$$= \frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} \exp \left( \frac{\tau^2 x^2}{2 \sigma^2 (\sigma^2 + \tau^2)} \right) \xrightarrow{\tau^2 \to \infty} 0$$
3.2– Bayesian Polynomial Regression Example

- In practice, you might have more than 2 potential models/hypothesis for your data.

- Consider the following polynomial regression problem where
  \[ D = \{x_i, y_i\}_{i=1}^n \] where \((x_i, y_i) \in \mathbb{R} \times \mathbb{R}\).

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

  \[
  = \beta_{0:M}^T f_M(X) + \sigma V
  \]

- Here the problem is that if \(M\) is too large then there will be overfitting.
3.2– Bayesian Polynomial Regression Example

As $M$ increases, the model overfits.
3.2– Bayesian Polynomial Regression Example

- Candidate Bayesian models $H_M$ for $M \in \{0, \ldots, M_{\text{max}}\}$.

- For the model $H_M$, we take the prior $\pi_M (\beta_{0:M}, \sigma^2)$

  \[
  \pi_M (\beta_{0:M}, \sigma^2) = \pi_M (\beta_{0:M} | \sigma^2) \pi_M (\sigma^2) = \mathcal{N} (\beta_{0:M}; 0, \delta^2 \sigma^2 I_{M+1}) \mathcal{IG} \left( \sigma^2; \frac{\nu_0}{2}, \frac{\gamma_0}{2} \right). 
  \]

- We have the following Gaussian likelihood

  \[
  f \left( D \mid \beta_{0:M}, \sigma^2 \right) = \prod_{i=1}^{n} \mathcal{N} \left( y_i; \beta_{0:M}^T f_M (x_i), \sigma^2 \right) 
  \]
3.2– Bayesian Polynomial Regression Example

- Standard calculations yield

\[
\pi_M (\beta_{0:M}, \sigma^2 | D) = \mathcal{N} (\beta_{0:M}; \mu_M, \sigma^2 \Sigma_M)
\]

\[
\times \mathcal{IG} \left( \sigma^2; \frac{\nu_0 + n}{2}, \frac{\gamma_0 + \sum_{i=1}^{n} y_i^2 - \mu_M^T \Sigma_M^{-1} \mu_M}{2} \right)
\]

where

\[
\mu_M = \Sigma_M \left( \sum_{i=1}^{n} y_i f_M (x_i) \right), \Sigma_M^{-1} = \delta^{-2} I_{M+1} + \sum_{i=1}^{n} f_M (x_i) f_M^T (x_i)
\]

knowing that

\[
\mathcal{IG} (\sigma^2; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma (\beta)} \frac{1}{(\sigma^2)^{\alpha+1}} \exp \left( -\frac{\beta}{\sigma^2} \right).
\]
3.2– Bayesian Polynomial Regression Example

- The marginal likelihood/evidence is given by

\[
\pi(D|H_M) = \int f(D|\beta_{0:M}, \sigma^2) \pi_M(\beta_{0:M}, \sigma^2) d\beta_{0:M} d\sigma^2 \\
= \Gamma\left(\frac{\nu_0+n}{2} + 1\right) \delta^{-(M+1)} |\Sigma_M|^{1/2} \left(\frac{\gamma_0+\sum_{i=1}^n y_i^2 - \mu_T \Sigma_M^{-1} \mu_M}{2}\right)^{-\left(\frac{\nu_0+n}{2} + 1\right)}
\]

- We can also compute

\[
\pi(H_M|D) = \frac{\pi(D|H_M) \pi(H_M)}{\sum_{i=0}^{M_{\text{max}}} \pi(D|H_i) \pi(H_i)}
\]
3.2– Bayesian Polynomial Regression Example

![Graphs showing different models and their evidence](Image)

Model Evidence

- Bayesian Model Selection
3.2– Bayesian Polynomial Regression Example

- We have assumed here that $\delta^2$ was fixed and set to $\delta^2 = 1$.

- As $\delta^2 \to \infty$, the prior on $\beta_{0:M}$ is getting vague but then

  \[
  \lim_{\delta^2 \to \infty} \pi \left( H_0 \mid D \right) = 1
  \]

  as for $M \geq 1$

  \[
  \frac{\pi \left( D \mid H_0 \right)}{\pi \left( D \mid H_M \right)} = \frac{\delta^{-1} |\Sigma_0|^{1/2} \left( \frac{\gamma_0 + \sum_{i=1}^{n} y_i^2 - \mu_0^T \Sigma_0^{-1} \mu_0}{2} \right)^{-\left(\frac{\nu_0 + n}{2} + 1\right)}}{\delta^{-(M+1)} |\Sigma_M|^{1/2} \left( \frac{\gamma_0 + \sum_{i=1}^{n} y_i^2 - \mu_M^T \Sigma_M^{-1} \mu_M}{2} \right)^{-\left(\frac{\nu_0 + n}{2} + 1\right)}} \to \infty \quad \delta^2 \to \infty
  \]

- Do not use vague priors for model selection!!!

- For a robust model, select a random $\delta^2$ and estimate it from the data. However, numerical methods are then necessary.
3.3– Bayesian Model Choice: Example

- In practice, you might have models of different natures for your data $x = (x_1, ..., x_T)$.

- $M_1$: Gaussian white noise $X_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{WN})$.

- $M_2$: An AR process of order $k_{AR}$, $k_{AR}$ being fixed, excited by white Gaussian noise $V_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{AR})$, 
  \[ X_n = \sum_{i=1}^{k_{AR}} a_i X_{n-i} + V_n. \]

- $M_3$: $k_{\sin}$ sinusoids, $k_{\sin}$ being fixed, embedded in a white Gaussian noise sequence $V_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\sin})$, 
  \[ X_n = \sum_{j=1}^{k_{\sin}} \left( a_{c_j} \cos[\omega_j n] + a_{s_j} \sin[\omega_j n] \right) + V_n. \]
3.4– Bayesian Model for Model Choice

- Generally speaking you have a countable collection of models \( \{ M_i \} \).

- For each model \( M_i \), you have a prior \( \pi_i(\theta_i) \) on \( \Theta_i \) and a likelihood function \( f_i(x|\theta_i) \).

- You attribute a prior probability \( \pi(i) \) to each model \( M_i \).

- The parameter space is \( \Theta = \bigcup_i \{ i \} \times \Theta_i \) and the prior on \( \Theta \) is

\[
\pi(i, \theta_i) = \pi(i) \pi_i(\theta_i).
\]
3.4– Bayesian Model for Model Choice

- In the polynomial regression example

\[ \Theta = \bigcup_{i=0}^{M_{\text{max}}} \{i\} \times \mathbb{R}^{i+1} \times \mathbb{R}^+. \]

- Remark: In all models, you have a noise variance to estimate. This parameter has a different interpretation for each model.
3.4– Bayesian Model for Model Choice

• In the non-nested example $\Theta = \{1\} \times \Theta_1 \cup \{2\} \times \Theta_2 \cup \{3\} \times \Theta_3$ where

\[
\theta_1 = \sigma_{WN}^2 \text{ and } \Theta_1 = \mathbb{R}^+,
\]

\[
\theta_2 = (a_1, \ldots, a_{k_{AR}}, \sigma_{AR}^2) \text{ and } \Theta_2 = \mathbb{R}^{k_{AR}} \times \mathbb{R}^+,
\]

\[
\theta_3 = (a_{c1}, a_{s1}, \omega_1, \ldots, a_{ck_{sin}}, a_{sk_{sin}}, \omega_{k_{sin}}, \sigma_{WN}^2), \Theta_3 = \mathbb{R}^{2k_{sin}} \times [0, \pi]^{k_{sin}} \times \mathbb{R}^+.
\]

• Remark: In all models, you have a noise variance to estimate. This parameter has a different interpretation for each model.

• Be careful, we don’t select here $\Theta = \{1, 2, 3\} \times \Theta_1 \times \Theta_2 \times \Theta_3$. 

3.4– Bayesian Model for Model Choice

- The posterior is given by Bayes’ rule

\[
\pi(k, \theta_k | x) = \frac{\pi(k) \pi_k(\theta_k) f_k(x | \theta_k)}{\sum_k \pi(k) \int_{\Theta_k} \pi_k(\theta_k) f_k(x | \theta_k) d\theta_k}.
\]

- We can obtain the posterior model probabilities through

\[
\pi(k | x) = \int_{\Theta_k} \pi(k, \theta_k | x) d\theta_k.
\]

- Once more, it is conceptually simple but it requires the calculation of many/an infinite number of integrals.
### 3.5– Bayesian Model Averaging

- Assume you’re doing some prediction of say $Y \sim g(y|\theta)$. Then in light of $x$, we have

\[
g(y|x) = \int g(y|\theta) \pi(\theta|x) \, d\theta
\]

\[
= \sum_k \int_{\Theta_k} g_k(y|\theta_k) \pi(k,\theta_k|x) \, d\theta_k
\]

\[
= \sum_k \pi(k|x) \int_{\Theta_k} g_k(y|\theta_k) \pi(\theta_k|x,k) \, d\theta_k
\]

- This is called Bayesian model averaging. All the models are taken into account to perform the prediction.
3.5– Bayesian Model Averaging

- An alternative way to make prediction consists of selecting the best “model”; say the model which has the highest posterior proba.

- The prediction is performed according to

$$\int_{\Theta_{k_{\text{best}}}} g_{k_{\text{best}}} (y | \theta_{k_{\text{best}}}) \pi (\theta_{k_{\text{best}}} | x, k_{\text{best}}) \, d\theta_{k_{\text{best}}}$$

- This is computationally much simpler and cheaper. This can also be very misleading.
3.6– Example

• Consider the previous example: 100 simulated data from a sum of three sinusoids with a very large additive noise.

• Priors were selected for the three models: Inverse-Gamma for $\sigma^2$, normal inverse-Gamma for AR and normal-inverse Gamma plus uniform for sinusoids. We set $\pi(H_1) = \pi(H_2) = \pi(H_3) = \frac{1}{3}$.

• We obtain

$$\pi(H_1|x) = 0.02, \quad \pi(H_2|x) = 0.12 \quad \text{and} \quad \pi(H_3|x) = 0.86.$$ 

• If we start using very vague priors....

$$\pi(H_1|x) \to 1.$$
3.7– Bayesian Variable Selection Example

- 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.
3.8— Bayesian Variable Selection Example

- 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 \( \pi(\gamma) = \prod_{i=1}^{p} \pi(\gamma_i) = 2^{-p} \).
• An alternative way to think of it is to write

\[ Y = \beta^T X + \sigma V \]

but the prior follows

\[ \pi(\beta_1, \ldots, \beta_p) = \prod_{i=1}^{p} \pi(\beta_i) \]

with

\[ \beta_i | \sigma^2 \sim \frac{1}{2} \delta_0 + \frac{1}{2} \mathcal{N}(0, \delta^2 \sigma^2). \]

• The regression coefficients follow a mixture model with a degenerate component.
3.9– 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

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

and

$$
\pi_\gamma (\beta_\gamma, \sigma^2 | D) = \mathcal{N} (\beta_\gamma; \mu_\gamma, \sigma^2 \Sigma_\gamma)
$$

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

where

$$
\mu_\gamma = \Sigma_\gamma \left( \sum_{i=1}^n y_i x_{\gamma,i} \right), \quad \Sigma_\gamma^{-1} = \delta^{-2} I_{n_\gamma} + \sum_{i=1}^n x_{\gamma,i} x_{\gamma,i}^T.
$$
3.10– Conclusion

- Bayesian model selection is a simple and principled way to do model selection.

- Bayesian model selection appears in numerous applications.

- Vague/Improper priors have to be banned in the model selection context!!!!

- Bayesian model selection only allows us to “compare” models. It does not tell you if any of the candidate models makes sense.

- Except for simple problems, it is impossible to perform calculations in closed-form.