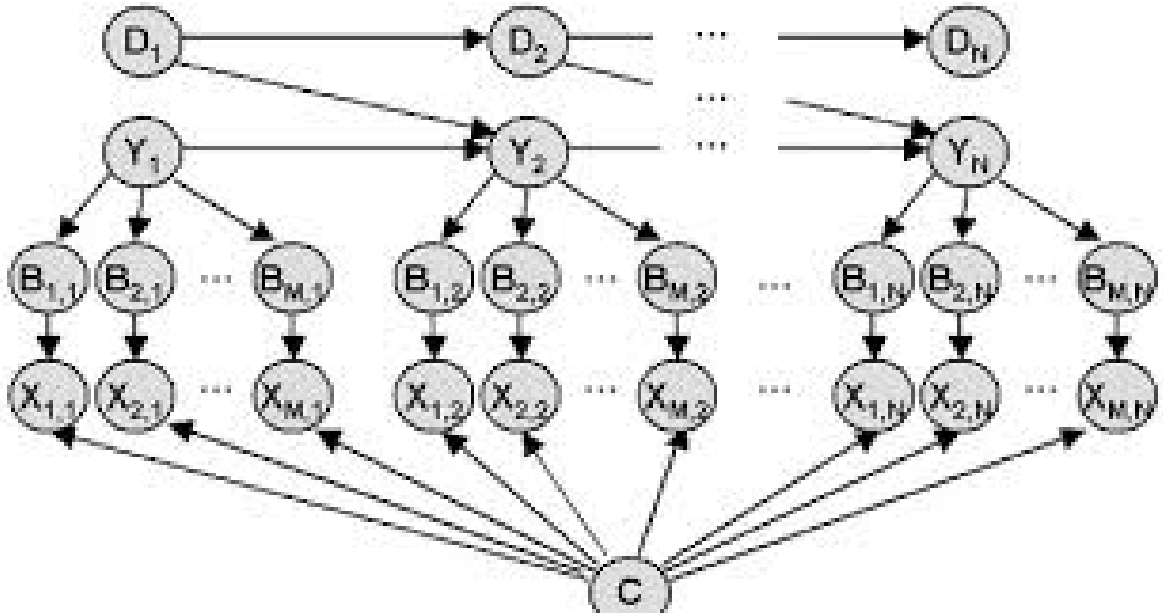
Bayesian Inference

Chapter 5: Model selection

Conchi Ausín and Mike Wiper
Department of Statistics
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative MethodsI
Master in Mathematical Engineering

Objective



In this class we consider problems where we have various contending models and show how these can be compared and also look at the possibilities of model averaging.

Basics

In principle, model selection is easy.

- Suppose that we wish to compare two models, $\mathcal{M}_1$, $\mathcal{M}_2$. Then, define prior probabilities, $P(\mathcal{M}_1)$ and $P(\mathcal{M}_2) = 1 - P(\mathcal{M}_1)$.
- Given data, we can calculate the posterior probabilities via Bayes theorem:

$$P(\mathcal{M}_1|\mathbf{x}) = \frac{f(\mathbf{x}|\mathcal{M}_1)P(\mathcal{M}_1)}{f(\mathbf{x})}$$

  where $f(\mathbf{x}) = f(\mathbf{x}|\mathcal{M}_1)P(\mathcal{M}_1) + f(\mathbf{x}|\mathcal{M}_2)P(\mathcal{M}_2)$.
- Note also that if model $\mathcal{M}_i$ is parametric with parameters θ_i, then

$$f(\mathbf{x}|\mathcal{M}_i) = \int f(\mathbf{x}|\mathcal{M}_i, \theta_i) f(\theta_i|\mathcal{M}_i)\, d\theta_i$$

- Now consider the possible losses (negative utilities) associated with taking a wrong decision.

$$L(\text{select } \mathcal{M}_2|\mathcal{M}_1 \text{ true}) \quad \text{and} \quad L(\text{select } \mathcal{M}_1|\mathcal{M}_2 \text{ true}).$$

  (something like type I and type II errors)
- Take the decision which minimizes the expected loss (Bayes decision).
- Expected loss for choosing $\mathcal{M}_1$ is

$$P(\mathcal{M}_2|\mathbf{x})L(\mathcal{M}_1|\mathcal{M}_2)$$

  and similarly for $\mathcal{M}_2$.
- If the loss functions are equal, then we just select the model with higher probability.
- Setting $L(\mathcal{M}_1|\mathcal{M}_2) = 0.05$ and $L(\mathcal{M}_2|\mathcal{M}_1) = 0.95$ means we select $\mathcal{M}_1$ if $P(\mathcal{M}_1|\mathbf{x}) > 0.05$.

The coin tossing example again

Return to the coin tossing problem of Class I.

- Suppose now that we wished to test the hypothesis $H_0 : \theta = 0.5$ versus the alternative $H_1 : \theta \neq 0.5$.
- Assume that $P(H_0) = P(H_1) = 0.5$ and that $\theta|H_1 \sim \text{Beta}(5, 5)$.

$$
\begin{aligned}
f(\mathbf{x}|H_0) &= \binom{12}{9} 0.5^{12} \\
f(\mathbf{x}|H_1) &= \int_0^1 \binom{12}{9} \theta^9 (1-\theta)^3 \frac{1}{B(5,5)} \theta^{5-1} (1-\theta)^{5-1} \, d\theta \\
&= \binom{12}{9} \frac{B(14,8)}{B(5,5)}
\end{aligned}
$$

which implies that $P(H_0|\mathbf{x}) \approx 0.387$. Therefore, we would reject H_0 under equal loss functions, but would not if we used 0.05 and 0.95 losses for type I and type II errors as previously.

Note that the p-value (under binomial sampling) is 0.0386 so in this case we reject H_0 at a 5% level.

Two sided hypothesis tests: a paradox

In two sided hypothesis tests, Bayesian and classical results can often differ greatly.

- Consider a coin tossing problem where we observe 49,581 heads and 48,870 tails.
- Then a classical 2 tailed test of $H_0 : \theta = 0.5$ vs $H_1 : \theta \neq 0.5$, gives a p-value of 0.0232 and H_0 is clearly rejected.
- However, given the set up of the previous example, the posterior probability that H_0 is true is $P(H_0|\mathbf{x}) \approx 0.89$.
- Given a uniform prior for θ under H_1, this probability increases to 0.95.
- On the contrary, for the one-sided test with $H_1 : \theta > 0.5$, the p-value is 0.0116 and the Bayesian posterior probability of H_0 is $P(H_0|\mathbf{x}) = 0.0117$.

What if we have a lot of possible models?

- In principle we can proceed as earlier but ...
- inference will be sensitive to the selection of the prior probabilities $P(\mathcal{M}_i)$ and ...
- it is often difficult to define these in many contexts (e.g. variable selection in regression models).
- We need a criterion which is less dependent on prior information.

The Bayes factor

The Bayes factor in favour of $\mathcal{M}_i$ and against model $\mathcal{M}_j$ is

$$B_j^i = \frac{P(\mathcal{M}_i|\mathbf{x})}{P(\mathcal{M}_j|\mathbf{x})} \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_i)}.$$

- This is the posterior odds divided by the prior odds.
- How does this get rid of the dependency on the priors?
- $B_j^i = \frac{f(\mathbf{x}|\mathcal{M}_i)}{f(\mathbf{x}|\mathcal{M}_j)}$.
- If the models do not have any parameters, this is the log likelihood ratio.
- Otherwise, recall that the marginal likelihood, $f(\mathbf{x}|\mathcal{M}_i)$ depends on the prior $f(\theta_i|\mathcal{M}_i)$.

Example

In our coin tossing example,

$$
\begin{aligned}
B_1^0 &= \frac{f(\mathbf{x}|H_0)}{f(\mathbf{x}|H_1)} \\
&= \frac{\binom{12}{9} 0.5^{12}}{\binom{12}{9} \frac{B(14,8)}{B(5,5)}} \\
&= 0.6309
\end{aligned}
$$

How do we interpret this number?

Consistency and scales of evidence

- It is clear that $0 \leq B_j^i < \infty$.
- When $B_j^i = 1$, the marginal likelihoods are the same, so the data provide equal evidence for both models.
- If model i is true, then, $B_j^i \to \infty$ and if model j is true, then $B_j^i \to 0$ as $n \to \infty$.

Kass and Raftery (1995) provide the following table for interpreting the Bayes factor.

Bayes factor	Interpretation
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Relationship to classical model selection criteria

The Bayesian information criterion (Schwarz 1978) for evaluating a model, $\mathcal{M}$ is

$$BIC = -2 \log f(\mathbf{x}|\hat{\theta}, \mathcal{M}) + k \log n$$

where $\hat{\theta}$ is the MLE and, the parameters defined under this model, θ, have dimension k.

Then, under certain regularity conditions, as $n \to \infty$, for two models $\mathcal{M}_i$ and $\mathcal{M}_j$, then

$$BIC_i - BIC_j \to -2 \log B^i_j.$$

Problems with the use of Bayes factors I: philosophy

- They require that one of the models is "true".
- There is a true value of the parameter θ under this model.
- There is positive prior mass on this true value under this model.

When the true model is not included, what happens as $n \to \infty$?

Problems with the use of Bayes factors II: calculation

Calculation of the Bayes factor is often tough

- In order to calculate the Bayes factor, we need the marginal likelihoods.
- Outside of conjugate models, these are impossible to evaluate analytically.
- Various alternatives are available in the context of Gibbs sampling and MCMC.

Harmonic mean estimator

Consider an MCMC sample, $\theta^{(1)}, ..., \theta^{(N)}$. Then, we can estimate $f(\mathbf{x}|\mathcal{M})$ or $f(\mathbf{x})$, (dropping dependence on $\mathcal{M}$ for notational convenience) as

$$f(\mathbf{x}) \approx \left(\frac{1}{N} \sum_{i=1}^{n} 1/f(\mathbf{x}|\theta_i) \right)^{-1}.$$

This is consistent as the expectation of what is being averages is

$$\int \frac{1}{f(\mathbf{x}|\theta)} f(\theta|\mathbf{x})\, d\theta = \int \frac{1}{f(\mathbf{x}|\theta)} \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})}\, d\theta = \frac{1}{f(\mathbf{x})} \int f(\theta)\, d\theta = \frac{1}{f(\mathbf{x})}.$$

However, the estimator is highly unstable and can often have infinite variance.

Chib's (1991) approach

Recall Bayes theorem:

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)f(\theta)}{f(\mathbf{x})}$$

so that, taking logarithms and reordering,

$$\log f(\mathbf{x}) = \log f(\mathbf{x}|\theta) + \log f(\theta) - \log f(\theta|\mathbf{x}).$$

Suppose that we have run the Gibbs sampler and have a sensible posterior point estimate, say $\tilde{\theta}$. Then, we can typically calculate $\log f(\tilde{\theta})$ and $\log f(\mathbf{x}|\tilde{\theta})$ analytically.

How do we calculate $\log f(\tilde{\theta}|\mathbf{x})$?

Assume that $\theta = (\theta_1, ..., \theta_k)$. Then, from the law of multiplication,

$$\log f(\tilde{\theta}|\mathbf{x}) = \log f(\tilde{\theta}_1|\mathbf{x}) + \log f(\tilde{\theta}_2|\mathbf{x}, \tilde{\theta}_1) + ... + \log f(\tilde{\theta}_k|\mathbf{x}, \tilde{\theta}_1, ..., \tilde{\theta}_{k-1}).$$

Firstly, if we run the Gibbs sampler again, and generate sample values $\theta^{(1)}, ..., \theta^{(N)}$, we can calculate

$$\log f(\tilde{\theta}_1|\mathbf{x}) \approx \frac{1}{N} \sum_{i=1}^{N} \log f(\tilde{\theta}_1|\mathbf{x}, \theta^{(i)})$$

Now fix $\theta_1 = \tilde{\theta}_1$ and run the Gibbs sampler again, generating a sample $\theta_{-1}^{(i)}$, for $i = 1, ..., N$ to calculate

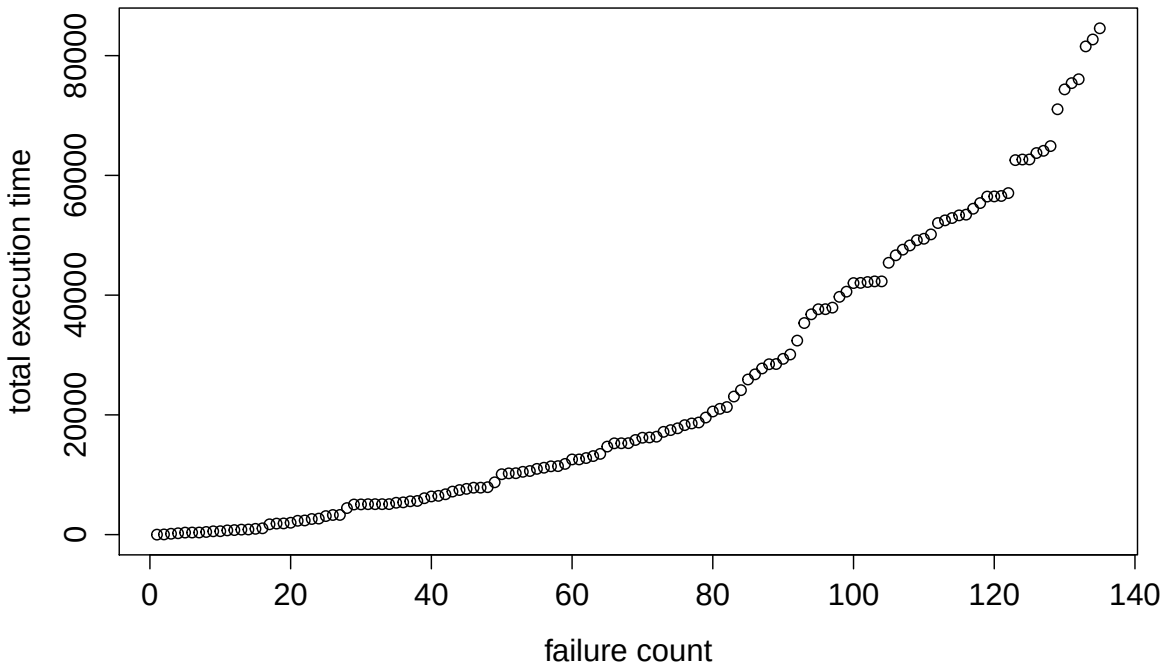
$$\log f(\tilde{\theta}_2|\mathbf{x}, \tilde{\theta}_1) \approx \frac{1}{N} \sum_{i=1}^{N} \log f(\tilde{\theta}_2|\mathbf{x}, \theta_{-1}^{(i)}, \tilde{\theta}_1)$$

Then run the Gibbs sampler again with $\theta_1 = \tilde{\theta}_1$ and $\theta_2 = \tilde{\theta}_2$ and so on.

- Fairly accurate.
- Relies on all conditional distributions being available analytically.
- Extensions to more general MCMC samplers are available.
- Means that we have to run the Gibbs sampler various times which means that it can be very slow.

Software reliability example

Remember the software reliability example. Before we assumed i.i.d. exponential failure times.



Inter-failure times are longer after more faults have been observed (and corrected?)

The Jelinski Moranda model

This model assumes N initial faults, each with rate θ and that after each failure, the fault causing it is perfectly corrected.

$$X_i|\theta, N \sim \textit{exponential}(\theta(N - i + 1)).$$

- The likelihood is

$$f(\mathbf{x}|\theta, N) \propto \frac{N!}{(N-m)!}\theta^m \exp\left(-\theta \sum_{i=1}^{m}(N - i + 1)x_i\right)$$

  where m is the number of observed failures.
- Given an exponential(1) prior for θ, the conditional posterior is gamma.
- Given a Poisson(200) prior for N, the conditional posterior for $N - m$ is Poisson.
- It is easy to run a Gibbs sampler

The estimated marginal likelihood for this model is approx. -983. The marginal likelihood for the i.i.d. model is -1012.

Problems with the use of Bayes factors III: existence

Return to the coin tossing problem and suppose that under H_1, we use Haldane's prior, $f(\theta) \propto \frac{1}{\theta(1-\theta)}$. Then

$$f(\mathbf{x}|H_1) \propto \int_0^1 \binom{12}{9} \theta^9(1-\theta)^3\theta^{-1}(1-\theta)^{-1}\, d\theta$$

but how can we get rid of the proportionality?

When we use improper priors for the parameters within any model, the Bayes factor is not defined!

Possible solution: quasi Bayes factors

- Various quasi Bayes factors have been introduced to get round the problem of improper priors.
- All use some idea of dividing the data into a minimal training set and an evaluation set.
- Idea is to use the (smallest set of) training data to make the improper prior proper ...
- and then use the evaluation data as the sample to calculate a Bayes factor.

Intrinsic Bayes factors

Consider the example of normal data $X|\theta \sim \text{Normal}(\theta, 1)$ and suppose we wish to test $H_0 : \theta \leq 0$ versus $H_1 : \theta > 0$. Assume we define uniform priors for θ conditional on each of these hypotheses. Suppose that we observe a sample of size n.

Note that given a single datum, say x_1, then

$$\theta|x_1, H_i \sim \text{truncated normal}(x_1, 1)$$

where the truncation is onto $\mathbb{Z}^-$ in the case of H_0 and $\mathbb{Z}^+$ in the case of H_1. Then, conditional on x_1, we can define a (partial) Bayes factor

$$\begin{aligned} B_1^0(\mathbf{x}_{-1}|x_1) &= \frac{\int_{-\infty}^{0} f(\mathbf{x}|\theta, x_1) f(\theta|x_1, H_0)\, d\theta}{\int_{0}^{\infty} f(\mathbf{x}|\theta, x_1) f(\theta|x_1, H_1)\, d\theta} \\ &= \frac{1 - \Phi(-x_1)}{\Phi(x_1)} \frac{\Phi(-\sqrt{n}\bar{x})}{1 - \Phi(-\sqrt{n}\bar{x})} \end{aligned}$$

- A problem is that this is clearly sensitive to the training sample chosen.
- If x_1 is an outlier, we could have problems.
- One possibility is to average over all possible training sets of size 1.
  - Geometric IBF:

$$GIBF_1^0 = \sqrt[n]{B_1^0(\mathbf{x}_{-i}|x_i)} = \frac{\Phi(-\sqrt{n}\bar{x})}{1-\Phi(-\sqrt{n}\bar{x})}\left(\prod_{i=1}^{n}\frac{1-\Phi(-x_i)}{\Phi(x_i)}\right)^{1/n}$$

  - Arithmetic IBF

$$AIBF_1^0 = \frac{1}{n}\sum_{i=1}^{n}B_1^0(\mathbf{x}_{-i}|x_i) = \frac{\Phi(-\sqrt{n}\bar{x})}{1-\Phi(-\sqrt{n}\bar{x})}\frac{1}{n}\sum_{i=1}^{n}\frac{1-\Phi(-x_i)}{\Phi(x_i)}$$

- Both methods lose some of the nice properties of the original Bayes factor.

Information criteria

We have seen that the Bayes factor is often very difficult to calculate. In such cases, information criteria may be preferred.

The most well known classical information criteria are

- The Bayesian information criterion

$$BIC = -2 \log f(\mathbf{x}|\hat{\theta}, \mathcal{M}) + k \log n$$

- The Akaike information criterion

$$AIC = -2 \log f(\mathbf{x}|\hat{\theta}, \mathcal{M}) + 2k.$$

Both criteria use the deviance, $D(\theta|\mathbf{x}) = -2 \log f(\mathbf{x}|\theta, \mathcal{M})$ plus a correction term to account for model complexity.

Linear models example

Suppose we have a linear model, $\mathbf{Y} \sim \text{normal}\left(\mathbf{X}\theta, \frac{1}{\tau}\mathbf{I}\right)$. Then recall that the fitted values are

$$E[\mathbf{Y}|\hat{\theta}] = \mathbf{HY} \quad \text{where} \quad \mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

is the hat matrix.

Then the diagonal elements of $\mathbf{H}$ satisfy the restriction $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = k$, the number of linear parameters in the model.

The elements h_{ii} are influence measures. Higher values imply that $\hat{\theta}$ will change more if the i'th datum is removed.

A Bayesian hat matrix

Suppose we introduce a normal prior distribution for θ, $\theta \sim$ normal $\left(\mathbf{m}, \frac{1}{c\tau}\mathbf{V}\right)$.
Then the Bayesian hat matrix is now

$$\mathbf{H} = \mathbf{X}\left(\frac{1}{c}\mathbf{V} + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

and satisfies

$$E_{\mathbf{Y}|\theta}[\mathbf{Y}|\bar{\theta}] = E_{\mathbf{Y}}[\mathbf{Y}] + \mathbf{H}\left(\mathbf{Y} - E_{\mathbf{Y}}[\mathbf{Y}]\right).$$

The effective number of parameters for the linear model is then $p_D = \sum_{i=1}^{n} h_{ii}$.

The deviance information criterion

The effective number of parameters in a model with data $\mathbf{x}$ is defined in Spiegelhalter et al (2002) as

$$p_D = \bar{D} - D(\bar{\theta})$$

where $\bar{\theta} = E[\theta|\mathbf{x}]$ and $\bar{D} = E[D(\theta|\mathbf{x})|\mathbf{x}]$.
Then the deviance information criterion is

$$DIC = \bar{D} + p_D.$$

- $DIC = D(\bar{\theta}) + 2p_D$
- This is like a Bayesian AIC.
- For non-hierarchical, linear models with a non-informative prior on θ, $DIC = AIC$.

Advantages and disadvantages

- Very easy to calculate from MCMC output.
- Doesn't matter if we don't have a true model (unlike Bayes factor or BIC)
- Inconsistent (like AIC)
- Only really useful in nested models.
- p_D can be negative.
- Doesn't work in latent variable models.
- For alternatives, see Celeux et al (2003).

Predictive performance

As well as comparing models in sample, it is useful to assess their ability to predict data. The standard way is to use the log-predictive score:

Divide the data into a training set $\mathbf{x}_T$ and a prediction set, $\mathbf{x}_P$. Then calculate the mean log predictive score

$$-\frac{1}{n_P}\sum_{i:x_i\in\mathbf{x}_P}\log f(x_i|\mathbf{x}_T).$$

- For time varying models, this can show which models predict better at different time periods.
- This can be very computationally intensive.

Model averaging

- Often, various models are available and will give similar fits but very different predictions.
- In such cases, we may wish to use model averaging to make predictions.
- In principle this is straightforward:

  Given a set of models with well defined prior and posterior distributions, we can make predictions as

$$f(y|\mathbf{x}) = \sum_{i=1}^{k} P(\mathcal{M}_i|\mathbf{x}) f(y|\mathbf{x}, \mathcal{M}_i).$$

Practical issues I: computing the sum

If we have lots of possible models, the full computation of the summation may be infeasible.

Two possibilities:

- Occam's window: exclude complex models if data are in favour of simpler models.
  For example, we can consider thresholding type ideas in regression problems.
- MC^3. Use MCMC to directly approximate the summation.
  - We need to construct Metropolis Hastings passes that propose moves through the model space.
  - For models of variable dimension this can be done using e.g. reversible jump moves.

Practical issues II: prior model probabilities

In regression type models with multiple possible regressors, we can consider priors for a model $\mathcal{M}_i$ of form:

$$P(\mathcal{M}_i) = \prod_{j=1}^{k} \pi_j^{\delta_{ji}} (1 - \pi_j)^{1-\delta_{ji}}$$

where there are k possible covariates and $\delta_{ij} = 1$ if covariate j is included in model $\mathcal{M}_i$ and 0 otherwise.

- Setting $\pi_j = 0.5$ gives a uniform prior over model space.
- Setting $\pi_j < 0.5$ penalizes large models.

Summary

In the last chapter we have examined model comparison ideas.

- Bayesian hypothesis testing is similar to classical in one sided tests but often very different in 2 sided tests.
- Theoretical model comparison using Bayes factors is simple but there are many practical complications.
- We can use Bayesian model selection criteria.
- We haven't said anything about goodness of fit.