A Tutorial on the Practice of Bayesian Statistical Inference

David Kaplan

Department of Educational Psychology

THE UNIVERSITY
of
WISCONSIN
MADISON

GESIS, Mannheim, Germany, 8 November, 2017

1. Bayesian computation

2. Bayesian regression

3. Bayesian hierarchical modeling

4. Bayesian factor analysis

- Examples and code are in accompanying folders.

- The key reason for the increased popularity of Bayesian methods in the social and behavioral sciences has been the (re)-discovery of numerical algorithms for estimating the posterior distribution of the model parameters given the data.

- Prior to these developments, it was virtually impossible to analytically derive summary measures of the posterior distribution, particularly for complex models with many parameters.

- Rather than attempting the impossible task of analytically solving for estimates of a complex posterior distribution, we can instead draw samples from $p(\theta|y)$ and summarize the distribution formed by those samples. This is referred to as *Monte Carlo integration*.

- The two most popular methods of MCMC are the Gibbs sampler and the Metropolis-Hastings algorithm.

- Bayesian inference focuses on estimating features of the posterior distribution, such as point estimates (e.g. the EAP or MAP) and posterior probability intervals.

- For complex, high-dimensional problems involving multiple integrals, solving expectations can be virtually impossible.

- Rather than attempting the impossible task of analytically solving high dimensional problems, we can instead draw samples from the posterior distribution and summarize the distribution formed by those samples.

- This is referred to as *Monte Carlo integration*.

Outline
Bayesian
Computation
MCMC
Convergence
Diagnostics
rjags
Bayesian
Model
Building
Multilevel
Modeling
Bayesian
Factor
Analysis

- MCMC is based on first drawing T samples of the parameters of interest $\{\theta_t, t = 1, \ldots, T\}$ from the posterior distribution $p(\theta|y)$ and approximating the expectation by

$$E[p(\theta|y)] \approx \frac{1}{T} \sum_{t=1}^{T} p(\theta_t|y). \tag{1}$$

- Assuming the samples are independent of one another, the law of large numbers ensures that the approximation in equation (1) will be increasingly accurate as T increases.

- Samples do not have to be drawn independently. All that is required, is that the sequence $\{\theta_t, t = 1, \ldots, T\}$ results in samples throughout the support of the distribution

- The "Markov Chain" allows sampling throughout the support of a distribution while also relaxing the assumption of independent sampling.

Outline

Bayesian
Computation
MCMC
Convergence
Diagnostics
rjags

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Factor
Analysis

- Formally, a *Markov chain* is a sequence of dependent random variables $\{\theta^s\}$

$$\theta^0, \theta^1, \ldots, \theta^s, \ldots \qquad (2)$$

such that the conditional probability of θ^s given all of the past variables depends only on θ^{s-1} – that is, only on the immediate past variable.

- This conditional probability is referred to as the *transition kernel* K of the Markov chain.

Outline
Bayesian
Computation
  MCMC
  Convergence
  Diagnostics
  rjags
Bayesian
Model
Building
Multilevel
Modeling
Bayesian
Factor
Analysis

- The Markov chain has a number of very important properties.

- Over a long sequence, the chain will "forget" its initial state θ^0 and converge to its stationary distribution $p(\theta|y)$, which does not depend either on the number of samples T or on the initial state θ^0.

- The number of iterations prior to the stability of the distribution is referred to as the *burn-in* samples.

- Letting m represent the initial number of burn-in samples, we can obtain an *ergodic average* of the posterior distribution $p(\theta|y)$ as

$$\bar{p}(\theta|y) = \frac{1}{T-m} \sum_{t=m+1}^{T} p(\theta_t|y). \qquad (3)$$

- A decision must be made regarding the number of Markov chains to be generated, as well as the number of iterations of the sampler.

- Each chain samples from another location of the posterior distribution based on starting values.

- With multiple chains it may be the case that fewer iterations are required.

- In some cases, the same result can be obtained from one chain, although often requiring a considerably larger number of iterations.

- Once the chain has stabilized, the burn-in samples are discarded. Summary statistics, including the posterior mean, mode, standard deviation, and posterior probability intervals, are calculated on the post-burn-in iterations.

- It is crucial to assess the convergence of MCMC algorithms.

- The difficulty of assessing convergence stems from the very nature of MCMC in that the MCMC algorithm is designed to converge in distribution rather than to a point estimate.

- Because there is not a single adequate assessment of convergence, it is common to inspect several different diagnostics that examine varying aspects of convergence.

Outline

Bayesian
Computation
MCMC
Convergence
Diagnostics
rjags

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Factor
Analysis

1. Trace/history plots

2. ACF plots

3. Geweke diagnostic

4. Gelman-Rubin-Brooks diagnostic

- For this tutorial we will use the R interface of the program "JAGS" (*Just Another Gibbs Sampler* ; Plummer, 2015)

- JAGS is a BUGS like program (*Bayes Under Gibbs Sampling*; Spiegelhalter, Thomas and Best 2000).

- BUGS was the first program (within WinBUGS) that made Bayesian analysis possible.

Outline

Bayesian
Computation
  MCMC
  Convergence
  Diagnostics
  rjags

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Factor
Analysis

- Steps when using rjags:

  1. Bring in data and use R to do any manipulations - e.g. handling missing data, etc.

  2. Write model in jags and save as a .bug file.

  3. Pass the data and model to JAGS using the jags.model command.

  4. Use coda.samples and update to run the Gibbs sample.

  5. Use R to inspect results and get plots.

- See file "normal.rjags.html"

- The frequentist and Bayesian goals of model building are the same.

- The frequentist and Bayesian goals of model building are the same.

  1 Model specification based on prior knowledge

- The frequentist and Bayesian goals of model building are the same.

  1. Model specification based on prior knowledge

  2. Model estimation and fitting

- The frequentist and Bayesian goals of model building are the same.

  1. Model specification based on prior knowledge

  2. Model estimation and fitting

  3. Model evaluation and modification

- The frequentist and Bayesian goals of model building are the same.

  1. Model specification based on prior knowledge

  2. Model estimation and fitting

  3. Model evaluation and modification

  4. Model choice

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

15 / 45

- Despite these similarities there are important differences.

- A major difference between the Bayesian and frequentist goals of model building lie in the model specification stage.

- Because the Bayesian perspective explicitly incorporates uncertainty regarding model parameters using priors, the first phase of modeling building requires the specification of a full probability model for the data and the parameters of the model.

- Model fit implies that the full probability model fits the data. Lack of model fit may be due to incorrect specification of likelihood, the prior distribution, or both.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

16 / 45

- Another difference between the Bayesian and frequentist goals of model building relates to the justification for choosing a particular model among a set of competing models.

- Model building and model choice in the frequentist domain is based primarily on choosing the model that best fits the data.

- This has certainly been the key motivation for model building, respecification, and model choice in the context of structural equation modeling (Kaplan 2009).

- In the Bayesian domain, the choice among a set of competing models is based on which model provides the best posterior predictions.

- That is, the choice among a set of competing models should be based on which model will best predict what actually happened.

- A very natural way of evaluating the quality of a model is to examine how well the model fits the actual data.

- In the context of Bayesian statistics, the approach to examining how well a model predicts the data is based on the notion of *posterior predictive checks*, and the accompanying *posterior predictive p-value*.

- The general idea behind posterior predictive checking is that there should be little, if any, discrepancy between data generated by the model, and the actual data itself.

- Posterior predictive checking is a method for assessing the specification quality of the model. Any deviation between the data generated from the model and the actual data implies model misspecification.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

- Let y^{rep} be data replicated from our current model.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

- Let y^{rep} be data replicated from our current model.

Posterior Predictive Distribution

$$p(y^{rep}|y) = \int p(y^{rep}|\theta)p(\theta|y)d\theta$$

$$= \int p(y^{rep}|\theta)p(y|\theta)p(\theta)d\theta \qquad (4)$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

18 / 45

- In the Bayesian context, the approach to examining model fit and specification utilizes the posterior predictive distribution of replicated data.

- Let y^{rep} be data replicated from our current model.

Posterior Predictive Distribution

$$
\begin{aligned}
p(y^{rep}|y) &= \int p(y^{rep}|\theta)p(\theta|y)d\theta \\
&= \int p(y^{rep}|\theta)p(y|\theta)p(\theta)d\theta
\end{aligned} \tag{4}
$$

- Equation (4) states that the distribution of future observations given the present data, $p(y^{rep}|y)$, is equal to the probability distribution of the future observations given the parameters, $p(y^{rep}|\theta)$, weighted by the posterior distribution of the model parameters.

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

- One approach to model fit in the context of posterior predictive checking is based on Bayesian p-values.

- Denote by $T(y)$ a test statistic (e.g. χ^2) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian p-value is defined to be

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

- One approach to model fit in the context of posterior predictive checking is based on Bayesian p-values.

- Denote by $T(y)$ a test statistic (e.g. χ^2) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian p-value is defined to be

$$p\text{-}value = pr[T(y^{rep}, \theta) \geq T(y, \theta)|y]. \tag{5}$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

19 / 45

- To assess model fit, posterior predictive checking implies that the replicated data should match the observed data quite closely if we are to conclude that the model fits the data.

- One approach to model fit in the context of posterior predictive checking is based on Bayesian p-values.

- Denote by $T(y)$ a test statistic (e.g. χ^2) based on the data, and let $T(y^{rep})$ be the same test statistics for the replicated data (based on MCMC). Then, the Bayesian p-value is defined to be

$$p\text{-}value = pr[T(y^{rep}, \theta) \geq T(y, \theta)|y]. \tag{5}$$

- The p-value is the proportion of replicated test values that equal or exceed the observed test value. High (or low if signs are reversed) values indicate poor model fit.

- A very simple and intuitive approach to model building and model selection uses so-called *Bayes factors* (Kass & Raftery, 1995)

- In essence, the Bayes factor provides a way to quantify the odds that the data favor one hypothesis over another. A key benefit of Bayes factors is that models do not have to be nested.

- Consider two competing models, denoted as M_1 and M_2, that could be nested within a larger space of alternative models. Let θ_1 and θ_2 be the two parameter vectors associated with these two models.

- These could be two regression models with a different number of variables, or two structural equation models specifying very different directions of mediating effects.

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The goal is to develop a quantity that expresses the extent to which the data support M_1 over M_2. One quantity could be the posterior odds of M_1 over M_2, expressed as

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The goal is to develop a quantity that expresses the extent to which the data support M_1 over M_2. One quantity could be the posterior odds of M_1 over M_2, expressed as

Bayes Factors

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \left[\frac{p(M_1)}{p(M_2)} \right]. \tag{6}$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The goal is to develop a quantity that expresses the extent to which the data support M_1 over M_2. One quantity could be the posterior odds of M_1 over M_2, expressed as

Bayes Factors

$$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \left[\frac{p(M_1)}{p(M_2)} \right]. \tag{6}$$

- Notice that the first term on the right hand side of equation (6) is the ratio of two integrated likelihoods.

- This ratio is referred to as the *Bayes factor* for M_1 over M_2, denoted here as B_{12}.

- Our prior opinion regarding the odds of M_1 over M_2, given by $p(M_1)/p(M_2)$ is weighted by our consideration of the data, given by $p(y|M_1)/p(y|M_2)$.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- This weighting gives rise to our updated view of evidence provided by the data for either hypothesis, denoted as $p(M_1|y)/p(M_2|y)$.

- An inspection of equation (6) also suggests that the Bayes factor is the ratio of the posterior odds to the prior odds.

- In practice, there may be no prior preference for one model over the other. In this case, the prior odds are neutral and $p(M_1) = p(M_2) = 1/2$.

- When the prior odds ratio equals 1, then the posterior odds is equal to the Bayes factor.

- A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the *Bayesian information criterion* (BIC), also referred to as the Schwarz criterion.

- Consider two models, M_1 and M_2 with M_2 nested in M_1. Under conditions where there is little prior information, the BIC can be written as

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- A popular measure for model selection used in both frequentist and Bayesian applications is based on an approximation of the Bayes factor and is referred to as the *Bayesian information criterion* (BIC), also referred to as the Schwarz criterion.

- Consider two models, M_1 and M_2 with M_2 nested in M_1. Under conditions where there is little prior information, the BIC can be written as

$$BIC = -2\log(\hat{\theta}|y) + p\log(n) \tag{7}$$

where $-2\log\hat{\theta}|y$ describes model fit while $p\log(n)$ is a penalty for model complexity, where *p* represents the number of variables in the model and n is the sample size.

- The BIC is often used for model comparisons. The difference between two BIC measures comparing, say M_1 to M_2 can be written as

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The BIC is often used for model comparisons. The difference between two BIC measures comparing, say M_1 to M_2 can be written as

$$\Delta(BIC_{12}) = BIC_{(M_1)} - BIC_{(M_2)} \tag{8}$$
$$= \log(\hat{\theta}_1|y) - \log(\hat{\theta}_2|y) - \frac{1}{2}(p_1 - p_2)\log(n)$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The BIC is often used for model comparisons. The difference between two BIC measures comparing, say M_1 to M_2 can be written as

$$
\begin{aligned}
\Delta(BIC_{12}) &= BIC_{(M_1)} - BIC_{(M_2)} \qquad (8) \\
&= \log(\hat{\theta}_1|y) - \log(\hat{\theta}_2|y) - \frac{1}{2}(p_1 - p_2)\log(n)
\end{aligned}
$$

- Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors and the comparison of BIC values from two competing models. Using M_1 as the reference model,

- Rules of thumb have been developed to assess the quality of the evidence favoring one hypothesis over another using Bayes factors and the comparison of BIC values from two competing models. Using M_1 as the reference model,

BIC Difference	Bayes Factor	Evidence against M_2
0 to 2	1 to 3	Weak
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
> 10	> 150	Very strong

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The BIC (ironically) is not fundamentally Bayesian

- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

- Define *Bayesian deviance* as

- The BIC (ironically) is not fundamentally Bayesian

- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

- Define *Bayesian deviance* as

Bayes Deviance

$$D(\theta) = -2\log[p(y|\theta)] + 2\log[h(y)], \tag{9}$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The BIC (ironically) is not fundamentally Bayesian

- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

- Define *Bayesian deviance* as

Bayes Deviance

$$D(\theta) = -2\log[p(y|\theta)] + 2\log[h(y)], \qquad (9)$$

- To make this Bayesian this, we obtain a posterior mean over θ by defining

- The BIC (ironically) is not fundamentally Bayesian

- An explicitly Bayesian approach to model comparison has been developed based on the notion of *Bayesian deviance*.

- Define *Bayesian deviance* as

Bayes Deviance

$$D(\theta) = -2\log[p(y|\theta)] + 2\log[h(y)], \qquad (9)$$

- To make this Bayesian this, we obtain a posterior mean over θ by defining

.

$$\overline{D(\theta)} = E_\theta[-2log[p(y|\theta)|y] + 2log[h(y)]. \qquad (10)$$

- Let $D(\bar{\theta})$ be a posterior estimate of θ. We can define the *effective dimension* of the model as $q_D = \overline{D(\theta)} - D(\bar{\theta})$.

- We then add the model fit term $\overline{D(\theta)}$ to obtain the *deviance information criterion* (DIC) - namely,

- Let $D(\bar{\theta})$ be a posterior estimate of θ. We can define the *effective dimension* of the model as $q_D = \overline{D(\theta)} - D(\bar{\theta})$.

- We then add the model fit term $\overline{D(\theta)}$ to obtain the *deviance information criterion* (DIC) - namely,

Deviance Information Criterion

$$DIC = \overline{D(\theta)} + q_D = 2\overline{D(\theta)} - D(\bar{\theta}). \tag{11}$$

- Let $D(\bar{\theta})$ be a posterior estimate of θ. We can define the *effective dimension* of the model as $q_D = \overline{D(\theta)} - D(\bar{\theta})$.

- We then add the model fit term $\overline{D(\theta)}$ to obtain the *deviance information criterion* (DIC) - namely,

Deviance Information Criterion

$$DIC = \overline{D(\theta)} + q_D = 2\overline{D(\theta)} - D(\bar{\theta}). \qquad (11)$$

- The DIC can be obtained by calculating equation (10) over MCMC samples.

- Models with the lowest DIC values are preferred.

- We will start with the basic multiple regression model.

- Let y be an n-dimensional vector $(y_1, y_2, \ldots, y_n)'$ $(i = 1, 2, \ldots, n)$ of reading scores from n students on the PIRLS reading assessment, and let X be an $n \times k$ matrix containing k background and attitude measures. Then, the normal linear regression model can be written as

- We will start with the basic multiple regression model.

- Let y be an n-dimensional vector $(y_1, y_2, \ldots, y_n)'$ $(i = 1, 2, \ldots, n)$ of reading scores from n students on the PIRLS reading assessment, and let X be an $n \times k$ matrix containing k background and attitude measures. Then, the normal linear regression model can be written as

$$y = X\beta + u, \tag{12}$$

- We will start with the basic multiple regression model.

- Let y be an n-dimensional vector $(y_1, y_2, \ldots, y_n)'$ $(i = 1, 2, \ldots, n)$ of reading scores from n students on the PIRLS reading assessment, and let X be an $n \times k$ matrix containing k background and attitude measures. Then, the normal linear regression model can be written as

$$y = X\beta + u, \qquad (12)$$

- All the usual regression assumptions apply

- We assume that student level PIRLS reading scores scores are generated from a normal distribution.

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

29 / 45

- Variables in this model are:

  1. male (1=male)

  2. ASBG04 = # of books in home

  3. ASBGSBS = Bullying/teasing at school (higher values mean less bullying/teasing)

  4. ASBGSMR = Students motivated to read

  5. ASBGSCR = Students confidence in their reading

  6. ASBR05E = Teacher is easy to understand

  7. ASBR05F = Interested in what teacher says

  8. ASBR05G = Teacher gives interesting things to read

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- See files "multiple.regression.m1.rjags.html" and "multiple.regression.m2.rjags.html"

- The selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Raftery & Madigan (1999) who write

- The selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Raftery & Madigan (1999) who write

  "Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(pg. 382)

- The selection of a particular model from a universe of possible models can also be characterized as a problem of uncertainty. This problem was succinctly stated by Hoeting, Raftery & Madigan (1999) who write

.

"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are."(pg. 382)

- An approach to addressing the problem is the method of *Bayesian model averaging* (BMA). We will show this in the regression example.

- Consider a quantity of interest such as a future observation denoted as Υ.

- Next, consider a set of competing models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior distribution of Υ given data y can be written as

Bayesian Model Averaging

- Consider a quantity of interest such as a future observation denoted as Υ.

- Next, consider a set of competing models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior distribution of Υ given data y can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y). \tag{13}$$

where $p(M_k|y)$ is the posterior probability of model M_k written as

- Consider a quantity of interest such as a future observation denoted as Υ.

- Next, consider a set of competing models M_k, $k = 1, 2, \ldots, K$ that are not necessarily nested.

- The posterior distribution of Υ given data y can be written as

$$p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y). \tag{13}$$

where $p(M_k|y)$ is the posterior probability of model M_k written as

$$p(M_k|y) = \frac{p(y|M_k)p(M_k)}{\sum_{l=1}^{K} p(y|M_l)p(M_l)}, \qquad l \neq k. \tag{14}$$

- The interesting feature of equation (14) is that $p(M_k|y)$ will likely be different for different models.

- The term $p(y|M_k)$ can be expressed as an integrated likelihood

- The interesting feature of equation (14) is that $p(M_k|y)$ will likely be different for different models.

- The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \tag{15}$$

where $p(\theta_k|M_k)$ is the prior distribution of θ_k under model M_k (Raftery et al., 1997).

Outline

Bayesian
Computation

Bayesian
Model
Building
Posterior
Predictive
Checking
Bayes Factors
Bayesian
Linear
Regression
Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- The interesting feature of equation (14) is that $p(M_k|y)$ will likely be different for different models.

- The term $p(y|M_k)$ can be expressed as an integrated likelihood

$$p(y|M_k) = \int p(y|\theta_k, M_k)p(\theta_k|M_k)d\theta_k, \qquad (15)$$

where $p(\theta_k|M_k)$ is the prior distribution of θ_k under model M_k (Raftery et al., 1997).

- BMA provides an approach for combining models specified by researchers.

- Madigan and Raftery (1994) show that BMA provides better predictive performance than that of a single model based on a log-score rule.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- BMA is difficult to implement.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- BMA is difficult to implement.

  1. The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y)$ can be quite large and the corresponding integrals are hard to compute.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- BMA is difficult to implement.

  1. The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y)$ can be quite large and the corresponding integrals are hard to compute.

  2. Eliciting $p(M_k)$ may not be straightforward. The uniform prior $1/M$ is often used.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- BMA is difficult to implement.

  1. The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y)$ can be quite large and the corresponding integrals are hard to compute.

  2. Eliciting $p(M_k)$ may not be straightforward. The uniform prior $1/M$ is often used.

  3. Choosing the class of models to average over is also challenging.

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- BMA is difficult to implement.

  1. The number of terms in $p(\Upsilon|y) = \sum_{k=1}^{K} p(\Upsilon|M_k)p(M_k|y)$ can be quite large and the corresponding integrals are hard to compute.

  2. Eliciting $p(M_k)$ may not be straightforward. The uniform prior $1/M$ is often used.

  3. Choosing the class of models to average over is also challenging.

- The problem of reducing the overall number of models that one could incorporate in the summation has led to solutions based on *Occam's window* and *Markov Chain Monte Carlo Model Composition* – M^3 (Madigan and Raftery, 1994).

Outline

Bayesian
Computation

Bayesian
Model
Building

Posterior
Predictive
Checking

Bayes Factors

Bayesian
Linear
Regression

Bayesian
Model
Averaging

Multilevel
Modeling

Bayesian
Factor
Analysis

- See file "BMA.html"

- A common feature of data collection in the social sciences is that units of analysis (e.g. students or employees) are nested in higher organizational units (e.g. schools or companies, respectively).

- The PIRLS study deliberately samples schools (within a country) and then takes a sample of 4th grade students within sampled schools.

- Such data collection plans are generically referred to as *clustered sampling designs*.

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling
Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- In addition to being able to incorporate priors directly into a multilevel model, the Bayesian conception of multilevel modeling has another advantage – namely it clears up a great deal of confusion in the presentation of multilevel models.

- The literature on multilevel modeling attempts to make a distinction between so-called "fixed-effects" and "random-effects".

- Gelman and Hill have recognized this issue and present five different definitions of fixed and random effects.

- The advantage of the Bayesian approach is that all parameters are assumed to be random. When conceived as a Bayesian hierarchical model, much of the confusion around terminology disappears.

- Perhaps the most basic multilevel model is the random effects analysis of variance model.

- As a simple example consider whether there are differences among G schools $(g = 1, 2, \ldots, G)$ on the outcome of student reading performance y obtained from n students $(i = 1, 2, \ldots, n)$.

- In this example, it is assumed that the G schools are a random sample from a population of schools.

- The model can be written as a two level hierarchical linear model as follows: Let

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- The model can be written as a two level hierarchical linear model as follows: Let

Level - 1

$$y_{ig} = \beta_g + r_{ig}, \qquad (16)$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- The model can be written as a two level hierarchical linear model as follows: Let

Level - 1

$$y_{ig} = \beta_g + r_{ig}, \tag{16}$$

The model for the school random effect can be written as

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- The model can be written as a two level hierarchical linear model as follows: Let

Level - 1

$$y_{ig} = \beta_g + r_{ig}, \qquad (16)$$

The model for the school random effect can be written as

Level - 2

$$\beta_g = \mu + u_g, \qquad (17)$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- The model can be written as a two level hierarchical linear model as follows: Let

Level - 1

$$y_{ig} = \beta_g + r_{ig}, \tag{16}$$

The model for the school random effect can be written as

Level - 2

$$\beta_g = \mu + u_g, \tag{17}$$

Inserting equation (17) into equation (16) yields

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- The model can be written as a two level hierarchical linear model as follows: Let

Level - 1

$$y_{ig} = \beta_g + r_{ig}, \tag{16}$$

The model for the school random effect can be written as

Level - 2

$$\beta_g = \mu + u_g, \tag{17}$$

Inserting equation (17) into equation (16) yields

.

$$y_{ig} = \mu + u_g + r_{ig}. \tag{18}$$

- For equation (18), we first specify the distribution of the reading performance outcome y_{ig} given the school effect u_g and the within school variance σ^2. Specifically,

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- For equation (18), we first specify the distribution of the reading performance outcome y_{ig} given the school effect u_g and the within school variance σ^2. Specifically,

$$y_{ig}|u_g, \sigma^2 \sim N(u_g, \sigma^2). \qquad (19)$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- For equation (18), we first specify the distribution of the reading performance outcome y_{ig} given the school effect u_g and the within school variance σ^2. Specifically,

$$y_{ig}|u_g, \sigma^2 \sim N(u_g, \sigma^2). \tag{19}$$

- Next specify the prior distribution on the remaining model parameters. For this model, we specify conjugate priors

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

40 / 45

- For equation (18), we first specify the distribution of the reading performance outcome y_{ig} given the school effect u_g and the within school variance σ^2. Specifically,

$$y_{ig}|u_g, \sigma^2 \sim N(u_g, \sigma^2). \tag{19}$$

- Next specify the prior distribution on the remaining model parameters. For this model, we specify conjugate priors

$$u_g|\mu, \omega^2 \sim N(0, \omega^2), \tag{20}$$

$$\mu \sim N(b_0, B_0), \tag{21}$$

$$\sigma^2 \sim \text{inverse-Gamma}(a, b), \tag{22}$$

$$\omega^2 \sim \text{inverse-Gamma}(a, b), \tag{23}$$

- We can arrange all of the parameters of the random effects ANOVA model into a vector θ and write the prior distribution as

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- We can arrange all of the parameters of the random effects ANOVA model into a vector $\boldsymbol{\theta}$ and write the prior distribution as

$$p(\boldsymbol{\theta}) = p(u_1, u_2, \ldots, u_G, \mu, \sigma^2, \omega^2), \tag{24}$$

where under the assumption of exchangeability of the school effects u_g we obtain

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- We can arrange all of the parameters of the random effects ANOVA model into a vector $\boldsymbol{\theta}$ and write the prior distribution as

$$p(\boldsymbol{\theta}) = p(u_1, u_2, \ldots, u_G, \mu, \sigma^2, \omega^2), \tag{24}$$

where under the assumption of exchangeability of the school effects u_g we obtain

$$p(\boldsymbol{\theta}) = \prod_{g=1}^{G} p(u_g|\mu, \omega^2) p(\mu) p(\sigma^2) p(\omega^2). \tag{25}$$

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Random
Effects ANOVA

Bayesian
Factor
Analysis

- See files "one.way.anova.html", "multilevel.regression.m1.html" and "multilevel.regression.m2.html".

- We write the confirmatory factor analysis model as

- We write the confirmatory factor analysis model as

CFA Model

$$y = \alpha + \Lambda\eta + \epsilon, \tag{26}$$

- We write the confirmatory factor analysis model as

CFA Model

$$y = \alpha + \Lambda\eta + \epsilon, \qquad (26)$$

- Under conventional assumptions we obtain the model expressed in terms of the population covariance matrix Σ as

- We write the confirmatory factor analysis model as

CFA Model

$$y = \alpha + \Lambda\eta + \epsilon, \qquad (26)$$

- Under conventional assumptions we obtain the model expressed in terms of the population covariance matrix Σ as

$$\Sigma = \Lambda\Phi\Lambda' + \Psi, \qquad (27)$$

- Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

- Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{28}$$

- Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{28}$$

- The uniqueness covariance matrix $\boldsymbol{\Psi}$ is assumed to follow an inverse-Wishart distribution. Specifically,

- Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \tag{28}$$

- The uniqueness covariance matrix $\boldsymbol{\Psi}$ is assumed to follow an inverse-Wishart distribution. Specifically,

$$\boldsymbol{\theta}_{IW} \sim IW(\boldsymbol{R}, \delta), \tag{29}$$

- Let $\boldsymbol{\theta}_{norm} = \{\boldsymbol{\alpha}, \boldsymbol{\Lambda}\}$ be the set of free model parameters that are assumed to follow a normal distribution and let $\boldsymbol{\theta}_{IW} = \{\boldsymbol{\Phi}, \boldsymbol{\Psi}\}$ be the set of free model parameters that are assumed to follow an inverse-Wishart distribution. Thus,

$$\boldsymbol{\theta}_{norm} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega}), \qquad (28)$$

- The uniqueness covariance matrix $\boldsymbol{\Psi}$ is assumed to follow an inverse-Wishart distribution. Specifically,

$$\boldsymbol{\theta}_{IW} \sim IW(\boldsymbol{R}, \delta), \qquad (29)$$

- Different choices for $\boldsymbol{R}$ and δ will yield different degrees of "informativeness" for the inverse-Wishart distribution.

Outline

Bayesian
Computation

Bayesian
Model
Building

Multilevel
Modeling

Bayesian
Factor
Analysis

- See file "CFA.rjags.html"