Bayesian Model Averaging

Journal Club

Andreas Kitsche

Institut fÃ¼r Biostatistik
kitsche@biostat.uni-hannover.de

21. November 2011

Multivariate data

Unit	Variable 1	...	Variable p
1	x_{11}	$\cdots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	...	x_{np}

Observed values are stored in the data matrix $\boldsymbol{X}$:

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Variable Selection and Linear Regression

- given a dependent variable Y
- given a set of candidate predictors $X_1, \ldots, X_k$
- find the best regression model of the form

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_{i_j} X_{i_j} + \epsilon, \qquad \epsilon \sim N(0, \sigma^2 I)$$

  where $X_{i_1}, \ldots, X_{i_p}$ is a subset of $X_1, \ldots, X_k$
- use the model to proceed effect sizes and standard errors
- make predictions

"A typical approach to data analysis is to carry out a model selection exercise leading to a single "best" model and then to make inferences as if the selected model were the true model (the selected model generated the data)." [Raftery et al. (1997)]

Model Uncertainty and Bayesian Model Averaging

Problem: uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are.

"part of the evidence is spent to specify the model"

BMA seeks to average over all possible sets of predictors

Bayesian data analysis

The BayesÂ´rule:

$$\underbrace{p(\theta|D)}_{posterior} = \underbrace{p(D|\theta)}_{likelihood}\,\underbrace{p(\theta)}_{prior} / \underbrace{p(D)}_{evidence}$$

where the evidence (maginal distribution, prior predictive) is

$$p(D) = \int d\theta p(D|\theta)p(\theta)$$

- prior $p(\theta)$ the strength of our belief in θ without the data
- posterior $p(\theta|D)$ the strength of our belief in θ when the Data D have been taken into account
- likelihood $p(D|\theta)$ the probability that the data could be generated by the model with parameter values θ
- evidence $p(D)$ the probability of the data according to the model, determined by summing across all possible parameter values weighted by the strength of belief in those parameter

Bayesian data analysis for model selection

suppose we have two models $M1$ and $M2$, then BayesÂ´ rule is:

$$p(M1|D) = p(D|M1)p(M1)/p(D)$$

$$p(M2|D) = p(D|M2)p(M2)/p(D)$$

The ratio of these is

$$\frac{p(M1|D)}{p(M2|D)} = \underbrace{\frac{p(D|M1)}{p(D|M2)}}_{Bayesfactor} \frac{p(M1)}{p(M2)}$$

Bayesian model averaging

- $M = \{M_1, \ldots, M_K\}$ - the set of all models being considered
- Δ- quantity of interest, i.e. effect size (the parameter estimate divided by its standard error)
- D - data

Posterior distribution of Δ is an average of the posterior distributions under each of the models considered, weighted by their posterior model probabilities:

$$Pr(\Delta|D) = \sum_{k=1}^{K} Pr(\Delta|M_k, D)Pr(M_k|D),$$

where the posterior probability of model M_k is given by

$$Pr(M_k|D) = \frac{Pr(D|M_k)Pr(M_k)}{\sum_{l=1}^{k} Pr(D|M_l)Pr(M_l)}$$

Bayesian model averaging

$$Pr(D|M_k) = \int Pr(D|\theta_k, M_k)Pr(\theta_k|M_k)d\theta_k$$

with:

- $Pr(D|M_k)$ the marginal likelihood of model M_k
- θ_k vector of parameters of model M_k
- $Pr(\theta_k|M_k)$ the prior density of θ_k under model M_k
- $Pr(D|\theta_k, M_k)$ the likelihood
- $Pr(M_k)$ the prior pobability that M_k is the true model

Posterior model probability

Suppose that $(K+1)$ models, $M_0, M_1, \ldots, M_K$ are being considered.
Each of $M_1, \ldots, M_K$ is compared in turn with M_0, yielding Bayes' Factors $B_{10}, \ldots, B_{K0}$.
Then the posterior probability of M_k is:

$$pr(M_k|D) = \alpha_k B_{k0} / \sum_{r=0}^{K} \alpha_r B_{r0}$$

where $\alpha_k = pr(M_k)/pr(M_0)$ is the prior odds for M_k against M_0

Specifying prior model probabilities

A prior probability on model M_i can be specified as:

$$pr(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{1-\delta_{ij}}$$

where $\pi_j \in [0, 1]$ is the prior probability that $\beta_j \neq 0$ in a regression model, and δ^{ij} is an indicator of whether or not variable j is included in model M_i

- $\pi_j = 0$ for all j - uniformed prior across model space
- $\pi_j < 0.5$ for all j - imposes a penalty for large models
- $\pi_j = 1$ - variable j is included in all models

OccamÂ´s Window [Madigan and Raftery(1994)]

Building a subset of models

1. Exclude models not belonging to:

$$A' = \left\{ M_k : \frac{max_l \{Pr(M_l|D)\}}{Pr(M_k|D)} \leq C \right\},$$

   where C is chosen by the data analyst and $max_l \{Pr(M_l|D)\}$ denotes the model with the highest posterior model probability

2. exclude models that receive less support from the data than any of their simpler submodels:

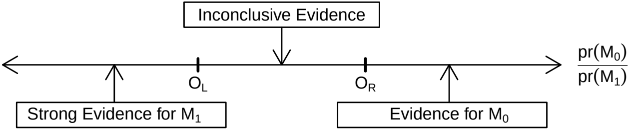
$$B = \left\{ M_k : \exists M_l \in M, M_l \subset M_k, \frac{Pr(M_l|D)}{Pr(M_k|D)} > 1 \right\}$$

Occam‘s Window

M_0 is smaller than M_1

If there is evidence for M_0 then M_1 is rejected, but rejecting M_0 requires strong evidence for the larger model, M_1.

$$\frac{pr(M_0|D)}{pr(M_1|D}$$



Markov Chain Monte Carlo Model Composition (MC^3)

Method developed by [Madigan and YORK(1995)]

- generate a stochastic process which moves through the model space
- Construct a Markov chain $\{M(t), t = 1, 2, \dots, \}$ with state space M and equilibrium distribution $pr(M_i|D)$
- define the function $g(M_i)$ on M and simulate the Markov chain $t = 1, \dots, N$
- $\hat{G} = \frac{1}{N}\sum_{t=1}^{N} g(M(t))$ is a simulation-consistent estimate of $E(g(M))$ as $N \to \infty$
- define $g(M) = pr(\Delta|M, D)$

Markov Chain Monte Carlo Model Composition (MC^3)

- define the neighborhood $nbd(M)$ for each $M \in M$ that consists of the model M itself and the state of models with either one variable more or one variable fewer than M
- define a transition matrix q by setting $q(M \rightarrow M') = 0$ for all $M' \notin nbd(M)$ and $q(M \rightarrow M')$ constant for all $M' \in nbd(M)$
- in state M we proceed by drawing M' from $q(M) \rightarrow M')$
- accept with probability

$$min\left\{1, \frac{Pr(M'|D)}{Pr(M|D)}\right\}$$

R implementation for BMA

R package `BMA`
available functions

- `bis.glm(x,...)` - Bayesian Model Averaging for generalized linear models
- `bic.surv(x,...)` - Bayesian Model Averaging for Cox proportional hazards models for censored survival data
- `bicreg(x,...)` - Bayesian Model Averaging for linear regression models
- `plot(bicreg,...)` - plot of the posterior distribuion of the coefficients produced by model averaging

Limitations: including an ad hoc model selection criterion that may bias posterior estimates

R implementation for BMA

R package `BAS`
For p less than 20-25, BAS can enumerate all models depending on memory availability Bayesian Model Averaging using Bayesian Adaptive Sampling

- `bas.lm(x,...)`
  - `modelprior` - Family of prior distribution on the models
  - `initprobs` - vector of length p with the initial inclusion probabilities

Advantages:

- it can search very large model spaces
- it offers a variety of prior specification options

Limitations: BAS can only estimate ordinary least squares

Predicting Percent Body Fat [Penrose et al. (1985)]

A data frame containing the estimates of the percentage of body fat determined by underwater weighing and various body circumference measurements for 252 men.

- case - case number
- brozek - Percent body fat using Brozek's equation: 457/Density - 414.2
- siri - Percent body fat using Siri's equation: 495/Density - 450
- density - density determined from underwater weighing (gm/cm^3)
- age - Age (years)
- weight - Weight (lbs)
- height - Height (inches)
- neck - Neck circumference (cm)
- chest - Chest circumference (cm)
- abdomen - Abdomen circumference (cm)
- hip - Hip circumference (cm)
- thigh - Thigh circumference (cm)
- knee - Knee circumference (cm)
- ankle -Ankle circumference (cm)
- biceps - Biceps (extended) circumference (cm)
- forearm - Forearm circumference (cm)
- wrist - Wrist circumference (cm)

M. Clyde.
Bayesian model averaging: A tutorial - Comment.
STATISTICAL SCIENCE, 14(4):401–417, 1999.

Anna Genell, Szilard Nemes, Gunnar Steineck, and Paul W. Dickman.
Model selection in Medical Research: A simulation study comparing Bayesian Model Averaging and Stepwise Regression.
BMC MEDICAL RESEARCH METHODOLOGY, 10:–, 2010.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky.
Bayesian model averaging: A tutorial (vol 14, pg 382, 1999).
STATISTICAL SCIENCE, 15(3):193–195, 2000.

D. Madigan and A. E. Raftery.
MODEL SELECTION AND ACCOUNTING FOR MODEL UNCERTAINTY IN GRAPHICAL MODELS USING OCCAMS WINDOW.
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 89(428): 1535–1546, 1994.

D. Madigan and J. YORK.
BAYESIAN GRAPHICAL MODELS FOR DISCRETE-DATA.
INTERNATIONAL STATISTICAL REVIEW, 63(2):215–232, 1995.

K. W. PENROSE, A. G. NELSON, and A. G. FISHER.
GENERALIZED BODY-COMPOSITION PREDICTION EQUATION FOR MEN USING SIMPLE MEASUREMENT TECHNIQUES.
MEDICINE AND SCIENCE IN SPORTS AND EXERCISE, 17(2): 189–189, 1985.

A. E. Raftery, D. Madigan, and J. A. Hoeting.
Bayesian model averaging for linear regression models.
JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 92(437): 179–191, 1997.