Bayesian Model Averaging in Longitudinal Studies using Bayesian Variable Selection Methods

Belay Birlie

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Hasselt University, Diepenbeek, Belgium
&
Department of Statistics, Jimma University, Jimma, Ethiopia

Ethiopian Statistical Association Conference
Addis Ababa, Ethiopia

May 20-22, 2016

Research Teams

- Belay Birlie (Hasselt University and Jimma University)
- Martin Otava (Janssen Pharmaceutical)
- Teshome Degafa (Jimma University)
- Delnesaw Yehwalaw (Jimma University)
- Ziv Shkedy (Hasselt University)

Outline

1. Introduction
   - Model averaging
2. Model Averaging Strategies
   - Frequentist model averaging
   - Bayesian Variable selection
3. Conclusion

Outline

1. Introduction
   - Model averaging
2. Model Averaging Strategies
   - Frequentist model averaging
   - Bayesian Variable selection
3. Conclusion

Motivating Example

Entomological survey on resettled (At risk) and non-resettled (control) villages (Degafa et al, 2015)

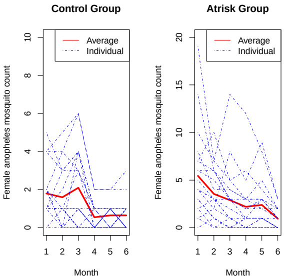
- Female anopheline mosquitoes resting inside human habitations collected monthly from 20 selected houses per village using pyrethrum spray catches
- Six longitudinal measurements per household
- Goal: Quantify the effect of ecological transformation and plan for intervention

Motivating Example

Entomological survey on resettled (At risk) and non-resettled (control) villages (Degafa et al, 2015)

- Female anopheline mosquitoes resting inside human habitations collected monthly from 20 selected houses per village using pyrethrum spray catches
- Six longitudinal measurements per household
- Goal: Quantify the effect of ecological transformation and plan for intervention

- Standard statistical practice
  - Use data-driven search to find best model M^*
  - Check model fit
  - Use M^* to estimate effect size, make predictions



Motivating Example

- Generalized liner Mixed Model

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$
$$\eta_{ij} = log(\lambda_{ij}) = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$$

- Let's restrict attention to linear predictor specification assuming that other structural properties of the model is known

Motivating Example

- Generalized liner Mixed Model

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$
$$\eta_{ij} = log(\lambda_{ij}) = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$$

- Let's restrict attention to linear predictor specification assuming that other structural properties of the model is known

Set of candidate models

- $M_1 : \eta_{ij} = \xi_1 + \xi_3 t_{ij} + b_i$
- $M_2 : \eta_{ij} = \xi_1 + \xi_2 x_i + \xi_3 t_{ij} + b_i$
- $M_3 : \eta_{ij} = \xi_1 + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$
- $M_4 : \eta_{ij} = \xi_1+\xi_2 x_i+(\xi_3+\xi_4 x_i)t_{ij}+b_i$

Motivating Example

- Generalized liner Mixed Model

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$
$$\eta_{ij} = log(\lambda_{ij}) = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$$

- Let's restrict attention to linear predictor specification assuming that other structural properties of the model is known

Set of candidate models

- $M_1 : \eta_{ij} = \xi_1 + \xi_3 t_{ij} + b_i$
- $M_2 : \eta_{ij} = \xi_1 + \xi_2 x_i + \xi_3 t_{ij} + b_i$
- $M_3 : \eta_{ij} = \xi_1 + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$
- $M_4 : \eta_{ij} = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$

- Unsatisfactory approach

Motivating Example

- Generalized liner Mixed Model

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$
$$\eta_{ij} = log(\lambda_{ij}) = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$$

- Let's restrict attention to linear predictor specification assuming that other structural properties of the model is known

Set of candidate models

- $M_1 : \eta_{ij} = \xi_1 + \xi_3 t_{ij} + b_i$
- $M_2 : \eta_{ij} = \xi_1 + \xi_2 x_i + \xi_3 t_{ij} + b_i$
- $M_3 : \eta_{ij} = \xi_1 + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$
- $M_4 : \eta_{ij} = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$

- Unsatisfactory approach
  - What do you do about competing model M^{**}?
  - Too risky to base all of your inferences on M^{*} alone
  - Inference should take in to account uncertainty in the model selection process

Motivating Example

- Generalized liner Mixed Model

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$
$$\eta_{ij} = log(\lambda_{ij}) = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$$

- Let's restrict attention to linear predictor specification assuming that other structural properties of the model is known

Set of candidate models

- $M_1 : \eta_{ij} = \xi_1 + \xi_3 t_{ij} + b_i$
- $M_2 : \eta_{ij} = \xi_1 + \xi_2 x_i + \xi_3 t_{ij} + b_i$
- $M_3 : \eta_{ij} = \xi_1 + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$
- $M_4 : \eta_{ij} = \xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i)t_{ij} + b_i$

- Unsatisfactory approach
  - What do you do about competing model M^{**}?
  - Too risky to base all of your inferences on M^* alone
  - Inference should take in to account uncertainty in the model selection process
- Solution: Model Averaging

Model averaging: Notation

- Consider K models: $\mathcal{M} = \{M_k, k = 1, 2, ..., K\}$
  - Associated with model k are (a vector of) parameters θ_k
- Δ is quantity of interest
  - Effect size
  - Future observation
- D is data
- $P(\theta_k|M_k)$ is prior density of θ_k under M_k
- $P(D|\theta_k, M_k)$ is likelihood of data
- $P(M_k)$ is prior probability that M_k is the true model

Model averaging: Including Model Selection Uncertainty in Estimator

- Model averaged posterior distribution of Δ given data is

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

Model averaging: Including Model Selection Uncertainty in Estimator

- Model averaged posterior distribution of Δ given data is

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

- Let $\hat{\Delta}_k = E[\Delta|D, M_k]$

Model averaging: Including Model Selection Uncertainty in Estimator

- Model averaged posterior distribution of Δ given data is

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

- Let $\hat{\Delta}_k = E[\Delta|D, M_k]$
- The mean and variance of Δ is
  - Mean:

$$E[\Delta|D] = \sum_{k=1}^{K} \hat{\Delta}_k P(M_k|D)$$

  - Variance:

$$Var[\Delta|D] = \sum_{k=1}^{K} (Var[\Delta|D, M_k] + \hat{\Delta}_k^2) P(M_k|D) - E[\Delta|D]^2$$

- This distribution takes into account the model uncertainty
  - i.e. that we do not know the correct model M_k

Model averaging: Posterior Model Probability

- The key ingredient of the model averaged estimates are the posterior model probability

Model averaging: Posterior Model Probability

- The key ingredient of the model averaged estimates are the posterior model probability
- The Posterior probability for model $M_k \in \mathcal{M}$ given data is

$$P(M_k|D) = \frac{P(D|M_k)P(M_k)}{\sum_{s=1}^{K} P(D|M_s)P(M_s)}$$

where

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

Model Averaging: Complication

- Good news: Robust Inference
  - Previous research shows that averaging over all models provides better predictive ability than using single model

Model Averaging: Complication

- Good news: Robust Inference
  - Previous research shows that averaging over all models provides better predictive ability than using single model
- Difficulties in implementation
  - How do you specify prior distribution on M_k and θ_k?
  - How can we compute the marginal likelihoods $P(D|M_k)$ in an economical manner?
  - M can be enormous; what search strategies can be implemented to quickly calculate or approximate $P(D|M_k)$?

Outline

1. Introduction
   - Model averaging
2. Model Averaging Strategies
   - Frequentist model averaging
   - Bayesian Variable selection
3. Conclusion

Model Averaging Strategies

- We will look at two methods
- Frequentist Model Averaging
  - Based on calculating model choice criteria (eg., AIC and BIC)(Burnham and Anderson, 2002; Lin et. al., 2012)

Model Averaging Strategies

- We will look at two methods
- Frequentist Model Averaging
    - Based on calculating model choice criteria (eg., AIC and BIC)(Burnham and Anderson, 2002; Lin et. al., 2012)
- Bayesian Model Averaging (Our main focus)
    - Lots of possible approaches (we will look at one)
        - Bayesian variable selection strategies (BVS)(Kuo and Mallick, 1998; Kasim et al, 2012, Otava, 2014)

Frequentist model averaging

- Using the BIC and AIC approximation
- An alternative expression of the posterior probability is

$$P(M_k|D) = \frac{BF_{kj}P(M_k)}{\sum_{s=1}^{K} BF_{sj}P(M_s)}$$

where $BF_{sj} = P(D|M_s)/P(D|M_j)$

Frequentist model averaging

- Using the BIC and AIC approximation
- An alternative expression of the posterior probability is

$$P(M_k|D) = \frac{BF_{kj}P(M_k)}{\sum_{s=1}^{K} BF_{sj}P(M_s)}$$

where $BF_{sj} = P(D|M_s)/P(D|M_j)$

- Since $-2logBF_{sj} \approx BIC_s - BIC_j)$

$$P(M_k|D) = \frac{\exp\left(-\frac{1}{2}\Delta BIC_k\right)}{\sum_{s=1}^{K} \exp\left(-\frac{1}{2}\Delta BIC_s\right)}$$

with $\Delta BIC_s = BIC_s - BIC_{min}$ and assuming $P(M_k) = 1/K$ for all k

Frequentist model averaging

- Using the BIC and AIC approximation
- An alternative expression of the posterior probability is

$$P(M_k|D) = \frac{BF_{kj}P(M_k)}{\sum_{s=1}^{K} BF_{sj}P(M_s)}$$

where $BF_{sj} = P(D|M_s)/P(D|M_j)$

- Since $-2logBF_{sj} \approx BIC_s - BIC_j)$

$$P(M_k|D) = \frac{\exp\left(-\frac{1}{2}\Delta BIC_k\right)}{\sum_{s=1}^{K} \exp\left(-\frac{1}{2}\Delta BIC_s\right)}$$

with $\Delta BIC_s = BIC_s - BIC_{min}$ and assuming $P(M_k) = 1/K$ for all k

  - Here, one can also use AIC .

Frequentist model averaging

- Steps

Frequentist model averaging

- Steps
  - Develop candidate models based on biological knowledge
  - Fit all candidate models and obtain MLE of parameters, AIC, and BIC of the alternate models
  - Evaluate strength of evidence for alternate models using approximation given above
  - Average MLE of parameters obtained from alternate models by their corresponding posterior model probability

Frequentist model averaging

- Steps
  - Develop candidate models based on biological knowledge
  - Fit all candidate models and obtain MLE of parameters, AIC, and BIC of the alternate models
  - Evaluate strength of evidence for alternate models using approximation given above
  - Average MLE of parameters obtained from alternate models by their corresponding posterior model probability
- Disadvantage
  - Need to fit all candidate models separately
  - Exploration of all K models is not possible for K large

Frequentist model averaging

- Steps
    - Develop candidate models based on biological knowledge
    - Fit all candidate models and obtain MLE of parameters, AIC, and BIC of the alternate models
    - Evaluate strength of evidence for alternate models using approximation given above
    - Average MLE of parameters obtained from alternate models by their corresponding posterior model probability
- Disadvantage
    - Need to fit all candidate models separately
    - Exploration of all K models is not possible for K large
- Solution: Bayesian Variable selection

Bayesian Variable selection (BVS): How to Perform?

- Different models arise from the inclusion/exclusion of ξ_2 and ξ_4

Bayesian Variable selection (BVS): How to Perform?

- Different models arise from the inclusion/exclusion of ξ_2 and ξ_4
- Substitute M by $\delta = (\delta_1, \delta_2)$, a binary indictor variable determining weather or not ξ_2 and/or ξ_4 included in the model where

$$\delta = \begin{cases} (0,0) & \text{if } \xi_2 \& \xi_4 \text{not included,} \\ (1,0) & \text{if } \xi_2 \text{is included,} \\ (0,1) & \text{if } \xi_4 \text{is included,} \\ (1,1) & \text{if } \xi_2 \& \xi_4 \text{is included.} \end{cases}$$

Bayesian Variable selection (BVS): How to Perform?

- Different models arise from the inclusion/exclusion of ξ_2 and ξ_4
- Substitute M by $\delta = (\delta_1, \delta_2)$, a binary indictor variable determining weather or not ξ_2 and/or ξ_4 included in the model where

$$\delta = \begin{cases} (0,0) & \text{if } \xi_2 \& \xi_4 \text{not included}, \\ (1,0) & \text{if } \xi_2 \text{is included}, \\ (0,1) & \text{if } \xi_4 \text{is included}, \\ (1,1) & \text{if } \xi_2 \& \xi_4 \text{is included}. \end{cases}$$

- Use binary system and calculate M using the equation

$$M = 1 + \sum_{l=1}^{L} \delta_l 2^{l-1}, \quad l = 1, ..., L(L = 2here)$$

BVS Model Formulation

Likelihood

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = \xi_1 + \delta_1\xi_2 x_i + (\xi_3 + \delta_2\xi_4 x_i)t_{ij} + b_i$$

Prior specification

$$\begin{aligned}
\xi_k &\sim N(0, \tau_{\xi_h}^{-1}), h = 1, ..., 4 \\
\tau_{\xi_h} &\sim \Gamma(1, 1), \\
\delta_l &\sim B(p_l), l = 1, 2 \\
p_l &\sim U(0, 1), \\
b_i &\sim N(0, \tau_b^{-1}), \\
\tau_b &\sim \Gamma(10^{-3}, 10^{-3})
\end{aligned}$$

BVS Model Formulation

Likelihood

$$Y_{ij}|b_i \sim Poisson(\lambda_{ij})$$

$$log(\lambda_{ij}) = \xi_1 + \delta_1 \xi_2 x_i + (\xi_3 + \delta_2 \xi_4 x_i) t_{ij} + b_i$$

Prior specification

$$\xi_k \sim N(0, \tau_{\xi_h}^{-1}), h = 1, ..., 4$$

$$\tau_{\xi_h} \sim \Gamma(1, 1),$$

$$\delta_l \sim B(p_l), l = 1, 2$$

$$p_l \sim U(0, 1),$$

$$b_i \sim N(0, \tau_b^{-1}),$$

$$\tau_b \sim \Gamma(10^{-3}, 10^{-3})$$

Set of candidate models

One-to-one relation between M and δ

Indictor	Model	Liner predictor
δ	$1 + \sum_l^L \delta_l 2^{L-1}$	$log(\lambda_{ij}))$
(0,0)	1	$\xi_1 + \xi_3 t_{ij} + b_i$
(1,0)	2	$\xi_1 + \xi_2 x_i + \xi_3 t_{ij} + b_i$
(0,1)	3	$\xi_1 + (\xi_3 + \xi_4 x_i) t_{ij} + b_i$
(1,1)	4	$\xi_1 + \xi_2 x_i + (\xi_3 + \xi_4 x_i) t_{ij} + b_i$

BVS: Some Detail

- The BVS model provides a simultaneous framework for estimation and model selection

BVS: Some Detail

- The BVS model provides a simultaneous framework for estimation and model selection
  - Denote $\vartheta_1 = \delta_1 \xi_2$ and $\vartheta_2 = \delta_2 \xi_4$ and let $\theta = (\xi_1, \vartheta_1, \xi_3, \vartheta_2, \sigma_b^2)$
  - Generate a sample $(M^{(b)}, \theta^{(b)}, \delta_l^{(b)}, b = 1, ..., B)$ using an MCMC algorithm

BVS: Some Detail

- The BVS model provides a simultaneous framework for estimation and model selection
  - Denote $\vartheta_1 = \delta_1 \xi_2$ and $\vartheta_2 = \delta_2 \xi_4$ and let $\theta = (\xi_1, \vartheta_1, \xi_3, \vartheta_2, \sigma_b^2)$
  - Generate a sample $(M^{(b)}, \theta^{(b)}, \delta_l^{(b)}, b = 1, ..., B)$ using an MCMC algorithm
  - Estimate posterior inclusion probabilities by

$$\hat{\bar{\delta}}_l = \frac{1}{B} \sum_{b=1}^{B} I(\delta_l^{(b)} = 1), \quad l = 1, 2$$

BVS: Some Detail

- The BVS model provides a simultaneous framework for estimation and model selection
  - Denote $\vartheta_1 = \delta_1 \xi_2$ and $\vartheta_2 = \delta_2 \xi_4$ and let $\theta = (\xi_1, \vartheta_1, \xi_3, \vartheta_2, \sigma_b^2)$
  - Generate a sample $(M^{(b)}, \theta^{(b)}, \delta_l^{(b)}, b = 1, ..., B)$ using an MCMC algorithm
  - Estimate posterior inclusion probabilities by

$$\hat{\bar{\delta}}_l = \frac{1}{B} \sum_{b=1}^{B} I(\delta_l^{(b)} = 1), \quad l = 1, 2$$

  - Estimate posterior model probability by

$$\hat{P}(M_k|D) = \frac{1}{B} \sum_{b=1}^{B} I(M^{(b)} = M_k), \quad k = 1, ..., M$$

BVS: Some Detail

- The BVS model provides a simultaneous framework for estimation and model selection
  - Denote $\vartheta_1 = \delta_1\xi_2$ and $\vartheta_2 = \delta_2\xi_4$ and let $\theta = (\xi_1, \vartheta_1, \xi_3, \vartheta_2, \sigma_b^2)$
  - Generate a sample $(M^{(b)}, \theta^{(b)}, \delta_l^{(b)}, b = 1, ..., B)$ using an MCMC algorithm
  - Estimate posterior inclusion probabilities by

$$\hat{\hat{\delta}}_l = \frac{1}{B}\sum_{b=1}^{B} I(\delta_l^{(b)} = 1), \quad l = 1, 2$$

  - Estimate posterior model probability by

$$\hat{P}(M_k|D) = \frac{1}{B}\sum_{b=1}^{B} I(M^{(b)} = M_k), \quad k = 1, ..., M$$

  - In each iteration b, only one model M is considered, so the estimate of θ is

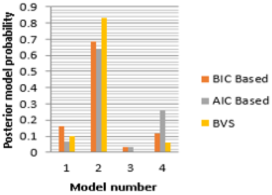
$$\hat{\hat{\theta}} = \frac{1}{B}\sum_{b=1}^{B} n_{M_k}\hat{\theta}_{M_k} = \sum_{k=1}^{K} \hat{P}(M_k|D)\hat{\theta}_{M_k}$$

Application

- We compute posterior model Probability using BVS and also approximate using AIC and BIC for each model
- We fit the full model, the best model and contrast their estimate with model averaged estimates

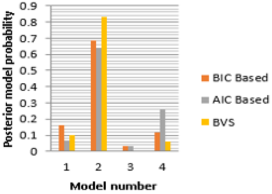
Results: Posterior Model and Inclusion Probability

Model	Parameters in the model				Rank		
	ξ_1	ξ_2	ξ_3	ξ_4	BIC	AIC	BVS
1	1	0	1	0	2	3	2
2	1	1	1	0	1	1	1
3	1	0	1	1	4	4	4
4	1	1	1	1	3	2	3
Inc. Prob.		0.890		0.066			
P-value		0.015		0.645			



Results: Posterior Model and Inclusion Probability

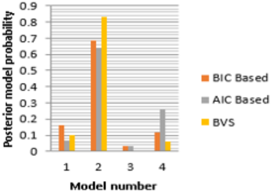
Model	Parameters in the model				Rank		
	ξ_1	ξ_2	ξ_3	ξ_4	BIC	AIC	BVS
1	1	0	1	0	2	3	2
2	1	1	1	0	1	1	1
3	1	0	1	1	4	4	4
4	1	1	1	1	3	2	3
Inc. Prob.		0.890		0.066			
P-value		0.015		0.645			



- All approaches reject the null hypothesis $H_0 : \xi_2 = \xi_4 = 0$
- Clearly, model 2 is indicated as the best by all approaches

Results: Posterior Model and Inclusion Probability

Model	ξ_1	ξ_2	ξ_3	ξ_4	Rank BIC	Rank AIC	Rank BVS
1	1	0	1	0	2	3	2
2	1	1	1	0	1	1	1
3	1	0	1	1	4	4	4
4	1	1	1	1	3	2	3
Inc. Prob.		0.890		0.066			
P-value		0.015		0.645			



- All approaches reject the null hypothesis $H_0 : \xi_2 = \xi_4 = 0$
- Clearly, model 2 is indicated as the best by all approaches
    - This model says that the two groups have a different intercept but identical slope

Results: Parameter Estimates

Par	Full		BM		MA-BIC		MA-AIC		BVS	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
ξ_1	0.690	0.259	0.745	0.229	0.799	0.219	0.767	0.232	0.761	0.242
ξ_2	0.865	0.340	0.786	0.295	0.630	0.238	0.728	0.277	0.661	0.359
ξ_3	-0.245	0.051	-0.265	0.028	-0.259	0.031	-0.260	0.034	-0.262	0.030
ξ_4	-0.028	0.061			-0.002	0.008	-0.006	0.018	-0.001	0.016
σ_b	0.845	0.121	0.846	0.121	0.846	0.120	0.851	0.121	0.905	0.137

Outline

Conclusion

- Post model selection parameter estimation is too risk and may lead to bias
- The use of model averaging is advocated in situations where,
  - The underlying goal of model selection is parameter estimation or prediction
  - No single model is overwhelmingly supported by the data
- The use of frequentist model averaging is limited to situations where we have small number of candidate models
- The BVS method performs simultaneous analyses of all the possible models and provides model averaged parameter estimates

Thank You