On needles and haystacks: Finding answers in complex health data

Rianne Jacobs
Faculty of Science and Engineering
University of Groningen

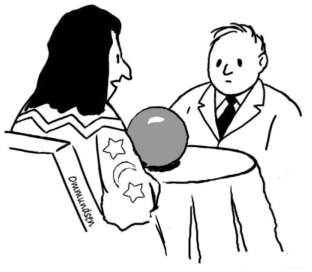
8 November 2019

Table of Contents

1 Introduction

Bayesian variable selection methodology for complex health data

- Large amount of previous knowledge and insights available
- Results have more natural interpretation, e.g. confidence intervals are often interpreted as credible intervals
- Relatively easy to implement complex models



"Is this needed for a Bayesian analysis?"

1 Introduction

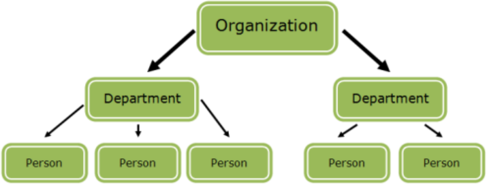
Bayesian variable selection methodology for complex health data

- Prediction
  - Finding best model for good prediction
  - Adjusting the outcome for confounders for fair comparisons
- Explanation
  - Finding (group of) variables that explain some outcome of interest
  - Finding the most important variable that best explains some outcome of interest

1 Introduction

Bayesian variable selection methodology for complex health data

- Structure in the data
  - Multilevel, e.g. longitudinal data, matched data, patients in hospitals, people in neighbourhoods
  - Correlated explanatory variables, possibly multilevel structure
- Many covariates, possibly $p > n$ - need some form of regularization
- Measurement error / misclassification
- Missing values



2 Foodborne disease outbreaks

Identifying the source of food-borne disease outbreaks: An application of Bayesian variable selection (2019) *Statistical Methods in Medical Research*

- Setting
  - Cases and controls fill in extensive food consumption questionnaires
  - Goal: find the food product that best distinguishes cases from controls
  - Logistic regression with variable selection
- Challenges
  - More variables than observations ($p > n$)
  - Misclassification in response
  - Many missing values
  - Cases and controls are matched
  - Covariates may be correlated
  - Small sample estimation problems

3 Bayesian variable selection

When searching for the cause of an outbreak we need a far more sophisticated variable selection procedure

3 Bayesian variable selection

When searching for the cause of an outbreak we need a far more sophisticated variable selection procedure

- We propose Bayesian method
- Can include prior information - may be crucial in beginning of outbreak
- Relatively easy implementation of variable selection, missing value imputation and misclassification correction
- No small sample estimation problems

3 Bayesian variable selection

- Data
  - Salmonella 2012 outbreak data
  - 302 observations
  - 106 exposure covariates
    - 0 - not eaten / not filled in
    - 1 - eaten
    - 2 - maybe (missing in our model)
- Model
  - Main effects model
  - Fixed covariates - age and gender

3 Bayesian variable selection

3.1 Methodology

- Observed response $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)'$
- True response $\boldsymbol{T} = (T_1, T_2, ..., T_n)'$

$$
\begin{aligned}
Y_i &\sim \text{Bernoulli}(\mu_i) \\
\mu_i &= \pi_i(\text{Se}) + (1 - \pi_i)(1 - \text{Sp}) \\
\text{logit}(\pi_i) &= \beta_0 + \boldsymbol{x}_i\boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
P(Y_i = 1|T_i = 1) &= \text{Se} \\
P(Y_i = 0|T_i = 0) &= \text{Sp}
\end{aligned}
$$

- However, no non-infected person entered the dataset as a case
  $P(Y_i = 1|T_i = 0, \boldsymbol{X_i}) = 0 \rightarrow \text{Sp} = P(Y_i = 0|T_i = 0) = 1$

3 Bayesian variable selection

Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993)

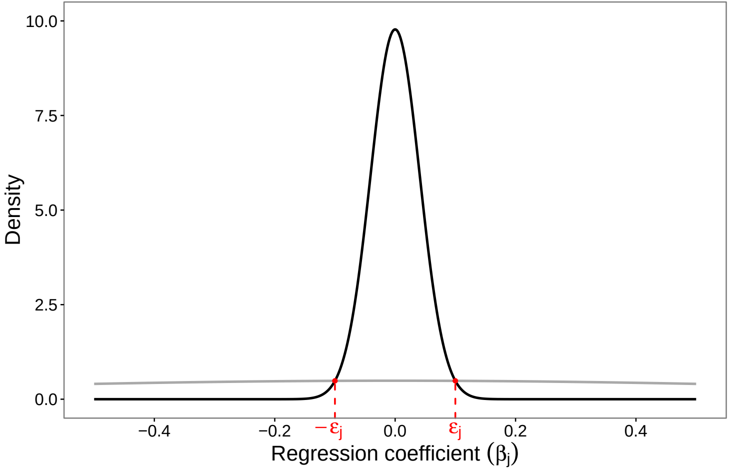
- Mixture prior on β_j, with spike and slab Gaussian components

$$\beta_j|\tau_j^2, c_j^2\tau_j^2, \gamma_j \sim (1-\gamma_j)\,\mathrm{N}(0,\tau_j^2) + \gamma_j\,\mathrm{N}(0, c_j^2\tau_j^2)$$
$$\gamma_j|\omega_j \sim \mathrm{Bernoulli}(\omega_j)$$
$$\omega_j \sim \mathrm{Beta}(a_{0j}, b_{0j})$$

where

- γ_j indicator variable for inclusion of β_j into the model
- ω_j inclusion probability of the j^{th} covariate
- Inclusion probabilities: Beta$(1, 2)$ favours parsimonious models
- Strong informative prior for sensitivity - identifiability

3 Bayesian variable selection



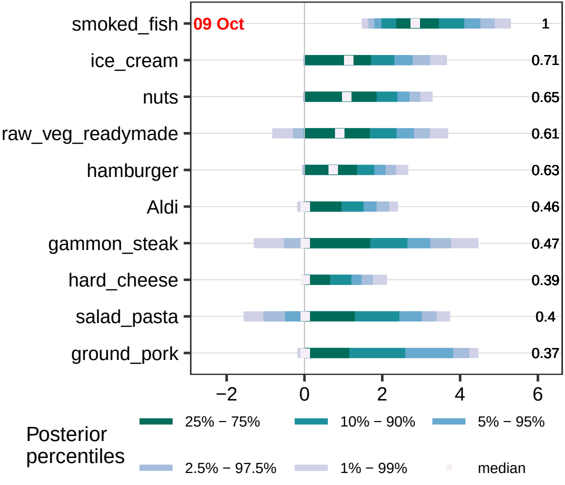
3 Bayesian variable selection

Missing data model

$$p(\boldsymbol{x}_{i,mis}|\boldsymbol{x}_{i,obs}, \boldsymbol{\theta}_X)$$
$$= p(x_{i,mis_1}|\boldsymbol{x}_{i,obs}, \boldsymbol{\theta}_{X_1}) \prod_{j=2}^{p} p(x_{i,mis_j}|x_{i,mis_1}, \ldots, x_{i,mis_{(j-1)}}, \boldsymbol{x}_{i,obs}, \boldsymbol{\theta}_{X_j})$$

- Each conditional distribution is Bernoulli with logit link function
- Possible sparse relationships between variables
- Variable selection in each of the conditional regression models of the covariate probability model
- Mitra and Dunson (2010) developed 2-level SSVS

3.2 Results and Conclusions



- 50% - weak
- 75% - positive
- 95% - strong
- 99% - very strong

3.2 Results and Conclusions

- Cannot fit standard logistic regression model early in outbreak
- Need some regularization or extra information (Lasso or Bayes)
- Compared to standard logistic regression, one week earlier detection
- Lasso diminishes differences in odds ratios between products - more difficult to identify product with large effect
- As evidence in the data increased for smoked fish, Bayes showed a steep increase in odds ratio reflecting this evidence
- Not in Lasso - only slight increase as the odds ratios are kept small due to the shrinkage (strong effects are also shrunk)

4 What is next?

4.1 Structured data

- Incorporating the matched design of the study into the methodology
- Strata with one case and several controls matched on well-known confounders
- Conditional logistic regression

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{k1}}}{e^{\boldsymbol{\beta}'\boldsymbol{x}_{k1}} + e^{\boldsymbol{\beta}'\boldsymbol{x}_{k2}} + \ldots e^{\boldsymbol{\beta}'\boldsymbol{x}_{km}}}$$

where $\boldsymbol{x}_{km}$ denotes covariate values for m^{th} individual in k^{th} stratum.

Main challenge is the Bayesian implementation of this model

- Equivalent to multinomial regression with m categories
- Also possible to rewrite as Poisson regression model
- Dealing with varying strata size

4.2 Other considerations

- Prior specification of sensitivity and inclusion probabilities
  - Extensive literature study
  - Inform priors using data on most likely suspects
- Empirical Bayes - estimate inclusion probabilities from current data and available external data
- Dynamic modelling
  - Posterior of initial analysis becomes prior of subsequent analysis
  - Not trivial to implement in spike-and-slab context

References

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Mitra, R. and Dunson, D. (2010). Two-level stochastic search variable selection in GLMs with missing predictors. *The International Journal of Biostatistics*, 6(1):33.

6 Take home message

- Bayesian variable selection is good alternative to current ad hoc source identification methods
- Relatively easy to implement and great flexibility to adapt model
- Current model still needs some refinements
- Large potential for interesting enhancements and extensions

Thank you

rianne.jacobs@rug.nl