Bayesian Methods for Wisconsin Breast Cancer data

Zhaoliang Zhou

Outline

- Introduction/Motivation
- Data manipulation
- EDA
- Variable selection
- Bayesian logistic regression
- Results
- Discussion/future direction

Introduction/Motivation

- The Wisconsin breast cancer dataset is one of the most popular datasets on Kaggle
  - Many people have applied popular frequentist methods and machine learning algorithms: classification tree, SVM, NN, etc
  - No Bayesian methods have been used for the top votes/hotness
- This project aim to:
  - Provide a Bayesian solution to the problem
  - Bayesian variable selection
  - Bayesian logistic regression for inference
  - Prediction
  - Performance comparison with popular methods on Kaggle

Data manipulation and description

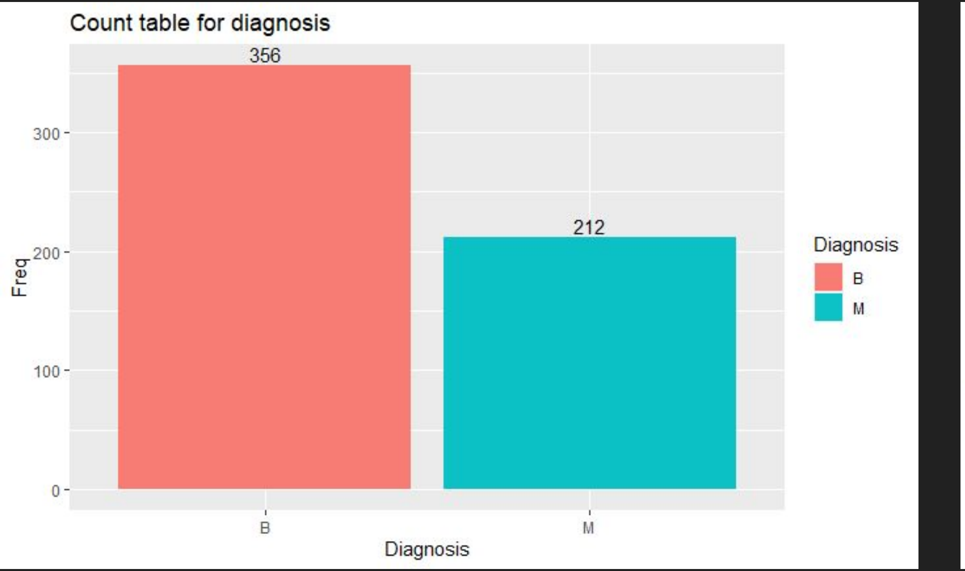
Data description

- Total 30 features - features computed and extracted from digital image of a fine needle aspirate (FNA) of a breast mass. Those features describe the characteristics of the cell nuclei shown in the image.For each cell nucleus, 10 features have been computed, and those features are: radius, texture, perimeter, area,smoothness, compactness, concavity, concave points, symmetry, fractal dimension.
- N = 357

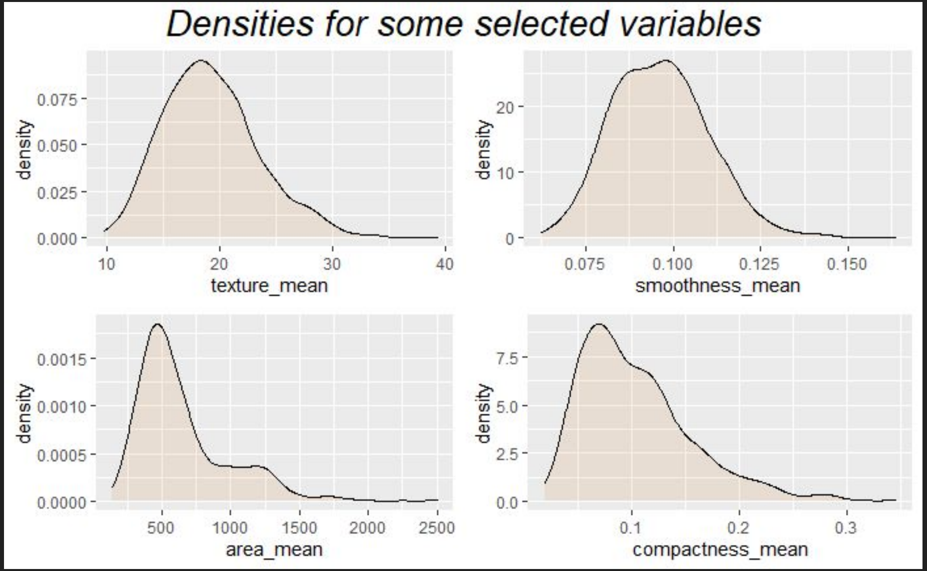
Data manipulation

- Data obtained from Kaggle website
- Removed some unmeaningful variables such as ID
- Recode diagnosis status to 0 and 1

EDA



Count table for diagnosis
356
212
300
200
100
0
Freq
B
M
Diagnosis
Diagnosis
B
M



Densities for some selected variables
density
0.075
0.050
0.025
0.000
10
20
30
40
texture_mean
density
20
10
0
0.075
0.100
0.125
0.150
smoothness_mean
density
0.0015
0.0010
0.0005
0.0000
500
1000
1500
2000
2500
area_mean
density
7.5
5.0
2.5
0.0
0.1
0.2
0.3
compactness_mean

Variable selection

- Many features - consider variable selection
- Traditional methods:
  - Local prior for regression coefficients in the true model - have a prior on the regression coefficient that has a positive prior density function at 0
  - Sometimes difficult in Bayesian framework: difficult to differentiate models with regression coefficients close to 0 and those who do not.
- Johnson & Rossell (2012):
  - Non-local prior densities (NLP)
  - MCMC algorithm to sample from the posterior distribution
  - Show different combinations of the variables with posterior probabilities
  - Select the combination that has the highest posterior probability
- 30 variables -> 15 variables

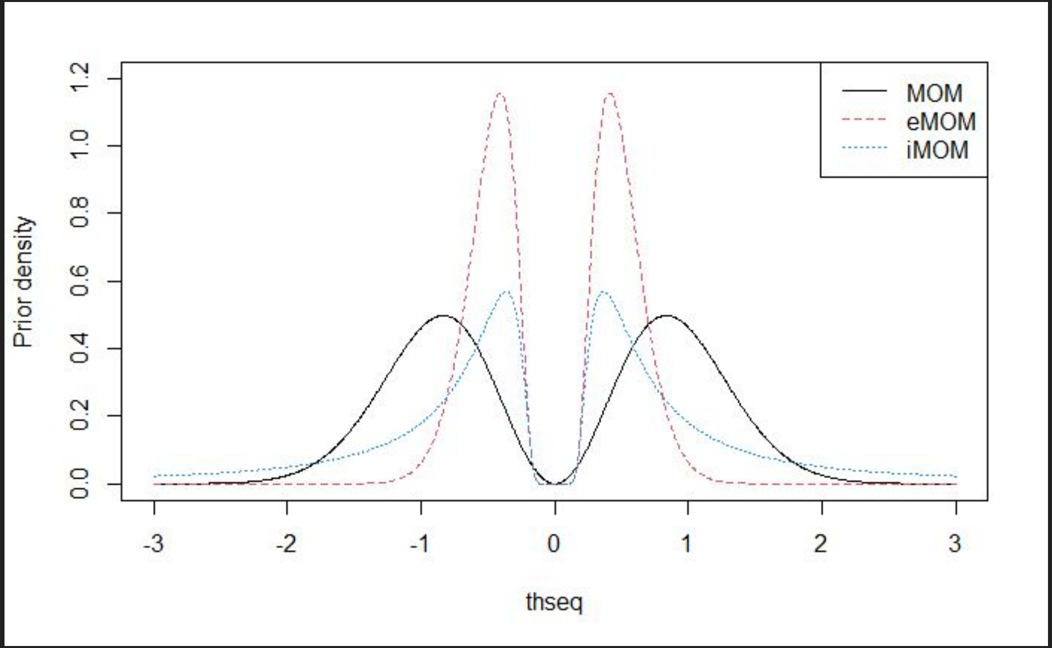
Variable selection cont.

Logistic regression model:

$$y_i|\beta_k \sim Bernoulli(\frac{exp(X_{ik}\beta_k)}{1+exp(X_{ik}\beta)})$$

Product of piMOM densities

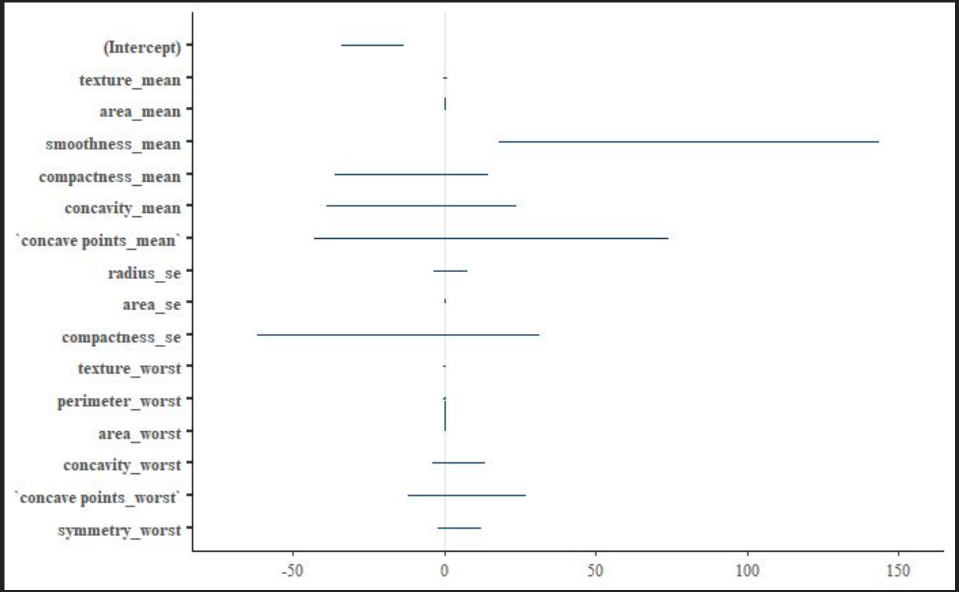
$$\pi(\beta_k|\tau,r) = \frac{\tau^{rk/2}}{\Gamma(r/2)^k}\prod_{i=1}^{k}|\beta_i|^{-(r+1)}exp(-\frac{\tau}{\beta_i^2})$$

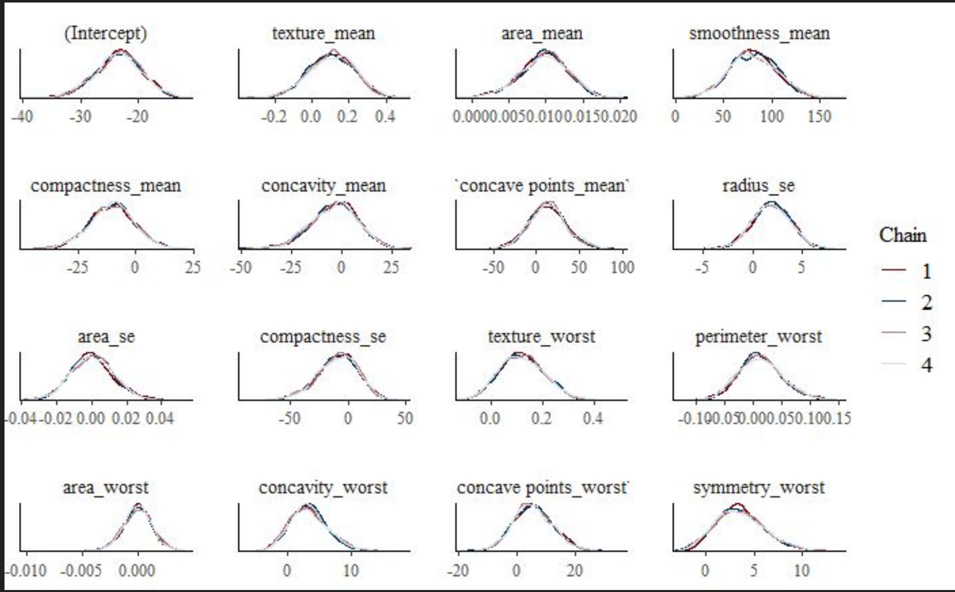


Bayesian GLM

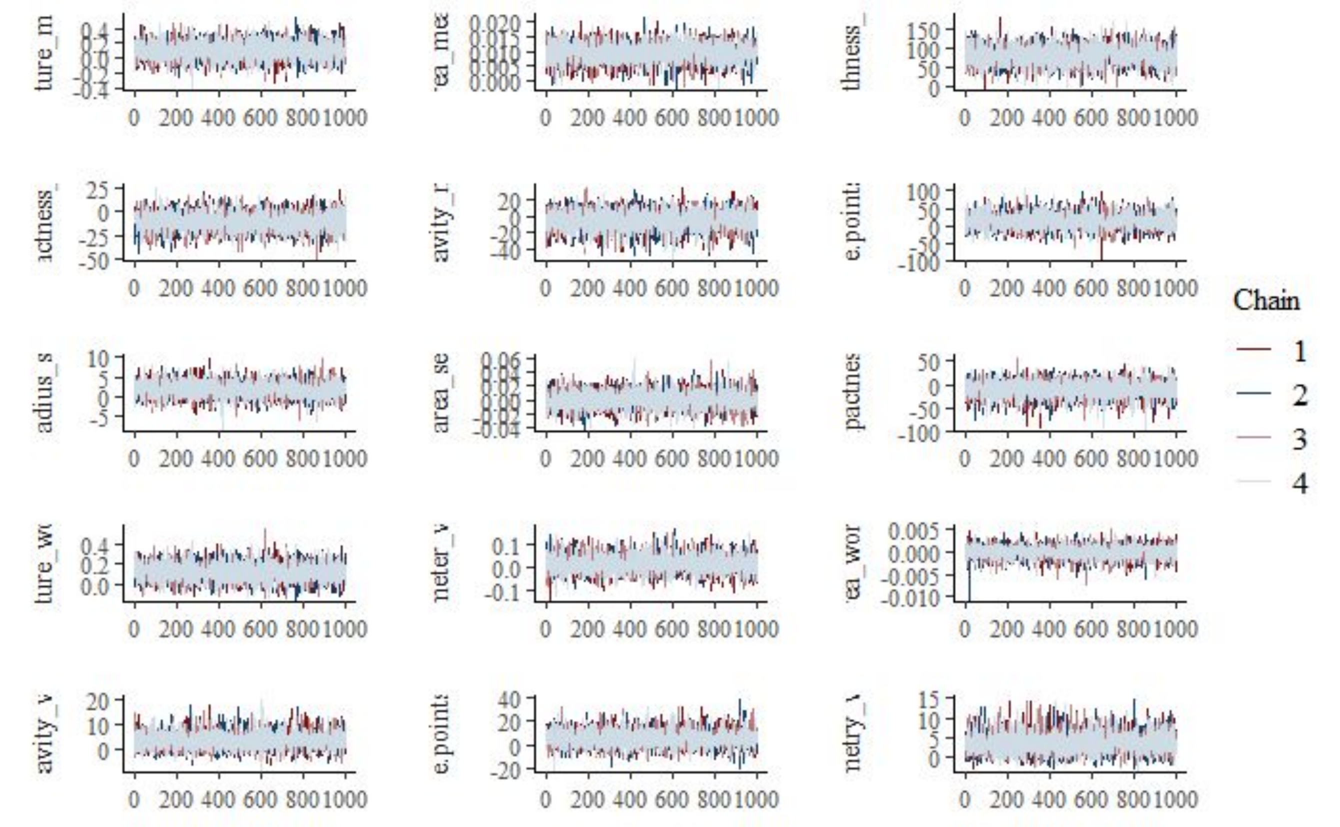
- Fitting Bayesian logistic regression with *stan_glm()*:
  - Intuitive coding and formula - syntax same as all other GLM models in r
  - Provide many tools for summarizing and visualizing posterior densities
- Prior densities on coefficients and intercept:
  - default t-distribution (df = 7, location = 0, scale = 2.5)
  - less prior confidence that the parameters will be close to zero
  - Link = logit
- Syntax:
  - stan_glm(formula, data =,
  - family = binomial(link = "logit"),
  - prior =, prior_intercept =, QR=TRUE,
  - seed =, refresh=0)

Results - posterior distribution for parameters





Results - convergence check



Results - posterior estimates and interval

Parameters	Median (90% CI)
Intercept	-23.22 (-30.21, -16.95)
Texture mean	0.11 (-0.10, 0.30)
Area mean	0.0099 (0.004, 0.015)*
Smoothness mean	79.95 (40.43, 121.27)*
Compactness mean	-10.12 (-26.86, 5.78)
Concavity mean	-3.12 (-24.22, 14.99)
Concave points mean	12.25 (-21.52, 50.12)
Radius se	1.91 (-1.25, 4.97)
Area se	0.00003 (-0.019, 0.021)
Compactness se	-7.86 (40.47, 18.37)*
Texture worst	0.12 (-0.02, 0.27)
Perimeter worst	0.01 (-0.048, 0.081)
Area worst	-0.00007 (-0.0024, 0.0021)
Concavity worst	3.14 (-1.46, 8.99)
Concave points worst	5.33 (-5.20, 17.49)
symmetry_worst	3.43 (-0.45, 8.38)

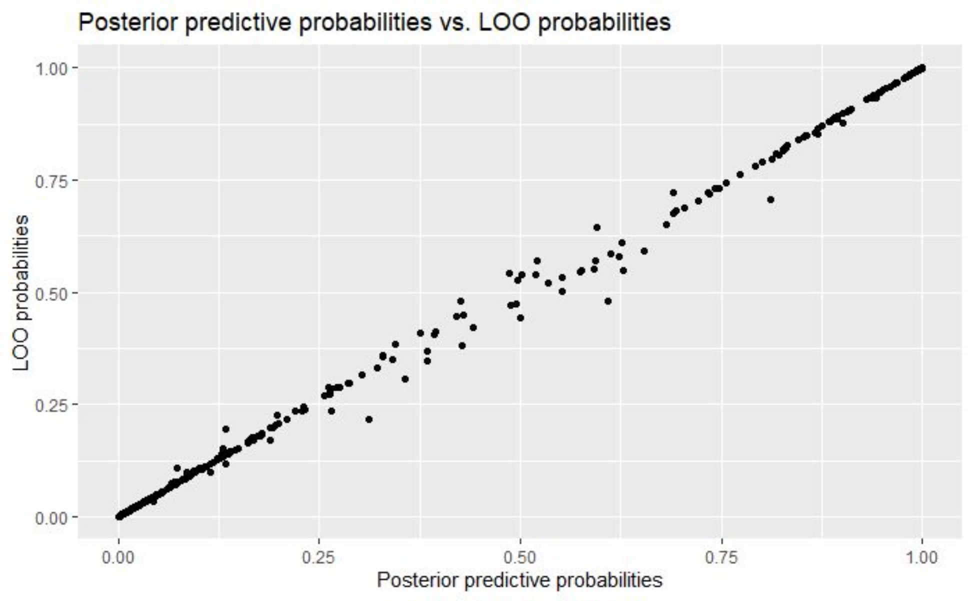
Prediction - baseline model comparison

- Pareto smoothed leave-one-out cross-validation (PSIS-LOO) to compute expected log predictive density (ELPD)
- Compare LOOIC with null model
- Compare ELPD with null model

	LOOIC	ELPD_diff
Alternative model	137.9	
Null model	752.5	-307.3

Prediction - LOO predictive probabilities

- Compute posterior predictive probabilities and then compute LOO classification error
- Posterior classification accuracy: 0.96



Discussion

- Most prediction accuracy on Kaggle is about 0.98+
- Our Bayesian method performed better than NN with single perceptron (0.95 accuracy)
- Most algorithms on Kaggle are “black box” methods
- Bayesian methods have several advantages:
  - Incorporate domain prior knowledge of the data by specifying priori
  - Posterior inferences conditional on the data
  - Interpretable results - posterior estimates and credible interval -> important in clinical setting

Future direction

- Variable selection methods not stable - could repeat many times and select the common variables
- Sensitivity analysis for other NLP during variable selection
- Sensitivity analysis for using different priors on coefficients for model fitting
- Comparison with the model include all the variables

Reference

- Vehtari, Gelman and Gabry (2017a)
- https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
- Amir Nikooienejad, Wenyi Wang, Valen E. Johnson, Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors, Bioinformatics, Volume 32, Issue 9, 1 May 2016, Pages 1338–1345. https://doi.org/10.1093/bioinformatics/btv764
- Johnson V.E. Rossell D. (2012) Bayesian model selection in high-dimensional settings. J. Am. Stat. Assoc.,107, 649–660. https://hannig.cloudapps.unc.edu/STOR757Bayes/handouts/JohnsonRussell2012.pdf
- Vehtari, Aki, et al. "Pareto smoothed importance sampling." arXiv preprint arXiv:1507.02646 (2015).

Questions?