Team PAJ (P.Jordan, A.Jassal, J.Imgrund)
June 25, 2018
Bayesian Statistics/Methods
DATS6450: Yuxiao (James) Huang

Final Project Paper

Title:

What Affects Cancer Diagnosis: A Comparison Using Hierarchy MCMC Model

Introduction:

Cancer is one of the leading causes of death in the United States, it touches us all eventually. According to the Centers for Disease Control and Prevention, breast cancer is the most common cancer in women no matter the race or ethnicity. It is the most common cause of death from cancer in Hispanic women and the second most common cause of death from cancer among white, black, Asian/Pacific Islander, and American Indian/Alaska Native women [1]. Analyzing data pertaining to breast cancer diagnosis can provide insight into what features are most critical to making a correct diagnosis.

Purpose:

The goal of the project was to determine what variables contribute most to a malignant medical test diagnosis. After determining a meaningful variable, a deep-dive analysis into that variable was conducted. Lastly, a comparison between a meaningful and less-meaningful variable was used to highlight extremes in the deep-dive analysis.

Data:

Team PAJ utilized a dataset from the University of Wisconsin located on Kaggle. While the data is part of an open data science competition, the analysis was not associated with the competition. The data format was .csv, which made for an effortless upload into R-Studio. The data consists of 570 rows by 32 columns, each row represented a patient who had a breast tumor and the columns were associated test measurements with one column being a binary flag for benign, not harmful, versus malignant,very virulent, cancer diagnosis (dependent variable). Extra columns were added as necessary to complete the purpose.

Method:

The team utilized two statistical models to achieve the purpose: random forest and hierarchy Markov Chain Monte Carlo (MCMC). The random forest model's sole goal was used for feature selection. Since a single-feature hierarchy MCMC was utilized, the team picked a meaningful variable ('Area_mean') that contributed strongly to a malignant diagnosis [Fig1]. A less-meaningful variable ('Symmetry_mean'), with weak predicting power, was chosen as an extreme opposite of the meaningful variable. Variables were binned into equal-frequency thirds and acted independently of each other. The bins represent **ω** (omega) in MCMC and there were 569 **θs** (theta), or unique patient results [Fig2]. MCMC was run twice on the two separate variables, and the results were analyzed and communicated.

Results:
The team compared six **θs**, unique patient results, in groups of two. The first comparison was between patient-1 and patient-566 [Fig3]. Both patient-1 and patient-566 have large tumor area sizes categorized in the bin3 as well as tumors that were malignant. As expected, the results are similar for the both patient Posterior Probability graphs. The highest density intervals (HDI) defaults to 95%. 95% of data will fall into a distribution between 0.762 to 1.0 and between 0.765 to 0.999 for patient-1 and patient-566, respectively . In addition, difference of **θs** graph is centered around zero, with mode being virtually zero, depicting the difference in the two patients area size in not significant. The red-cross plots to 0 because both patients had a diagnosis of malignant, or 1, and 1 - 1 = 0. In all figures moving forward, the red-cross represents the specific patient's test diagnosis. The scatterplot mimics the other three graph results in that the patients in the highest area bin are more likely to have a malignant medical diagnosis [Fig3].

In the second comparison, patient-21 and patient-10 were consciously selected to highlight the difference between a benign and malignant diagnosis. While bin2, which both patients were categorized, was biased towards a benign diagnosis, it still contained approximately 20% malignant diagnoses [Fig4]. The HDI fell into a distribution between 0.0033 to 0.403 and between 0.0426 to 0.551 for patient-21 and patient-10, respectively [Fig5]. The graph for patient-21 adheres to the probability that a person in bin2 has a higher probability to receive a benign diagnosis; thus the graph is slightly tighter than that for patient-10, who was the opposite. The difference of the **θs** graph is centered near zero, depicting the difference in the two patients tumor area size is not significant.

Patient-42 and patient-492 were selected for the third comparison to highlight 'outliers', whose bins did not match the expected diagnosis. The scatterplot indicates a patient in bin1 has a high likelihood of being benign while a patient in bin3 has a high likelihood of being malignant [Fig6]. Based on all the results presented, a smaller area measurement is more likely to be a benign diagnosis (0), while a larger area measurement is more likely to be malignant (1). The HDI of the data fell into a distribution between 0.0346 to 0.136 and between 0.669 to 0.975 for patient-42 and patient-492, respectively. Notice that the red cross plots outside the HDI of the posterior probability graph for both patients, as their diagnosis run counter to the group probabilities. In addition, the difference of the **θs** graph is centered between -0.943 and -0.599, with mode equal to -0.831 showing that the difference in the two patients tumor area size is significant.

The hierarchical MCMC model was rerun with 'symmetry_mean' as the feature, based on the results of our random forest feature selection we chose this to represent a weak predictor variable. Unlike 'Area_mean', 'Symmetry_mean's' bin2 is more evenly split between benign and malignant, leaning towards the former and bin 3 is nearly split 50/50[Fig7]. As expected in the omega probability graph, bin1 has a higher probability of a benign diagnosis, bin 2 is biased towards benign while bin 3 had a wide distribution with nearly an even probability of either diagnosis[Fig8]. We compared the same six patients, who fell into different bins when using 'symmetry', to have an apples-to-apples comparison. For patient-1, who fell into bin 3 in both tests, the posterior probability was stretched between 0.226 and 1 while patient-566, who was in bin 3 previously was now in bin 2, also had a wider distribution with the symmetry variable. Difference of **θs** graph clearly depicts the difference between the two symmetries is not significant [Fig9]. Patient-21 and patient-10, who both were in bin 2 for the 'area' variable were now in bin 3 for 'symmetry' as we expected bin 3 had a wide spread of probability and yielded a similar non-significant difference between symmetries [Fig10]. Lastly, and one of the most interesting changes, the comparison between patient-42 and patient-492. Recall, we chose these patients because they had a diagnosis which was counter to the posterior probability of the bins they were in. Both moved to different bins in this run, patient-42 from bin 1 to bin 2, but the HDI was still stretched, for patient-492 however the move from bin 3 previously, to bin 1 for symmetry produced a tighter distribution and a more accurate result. Ultimately the difference in symmetries is also not significant [Fig11]. These results are expected because as mentioned before, symmetry did not provide good predictive value for a medical diagnosis.

Conclusion:

The team chose to focus on individual models with the best and worst features to contrast results. Overall the project provided a greater insight into understanding medical datasets. All three purposes were achieved, asserting Area_mean was the best predictor (of the features with the ending in '_mean') for determining a benign or malignant medical diagnosis result. The worst feature (ending in '_mean'), Symmetry_mean was a prime example to contrast the results and highlight extremes in the analysis.

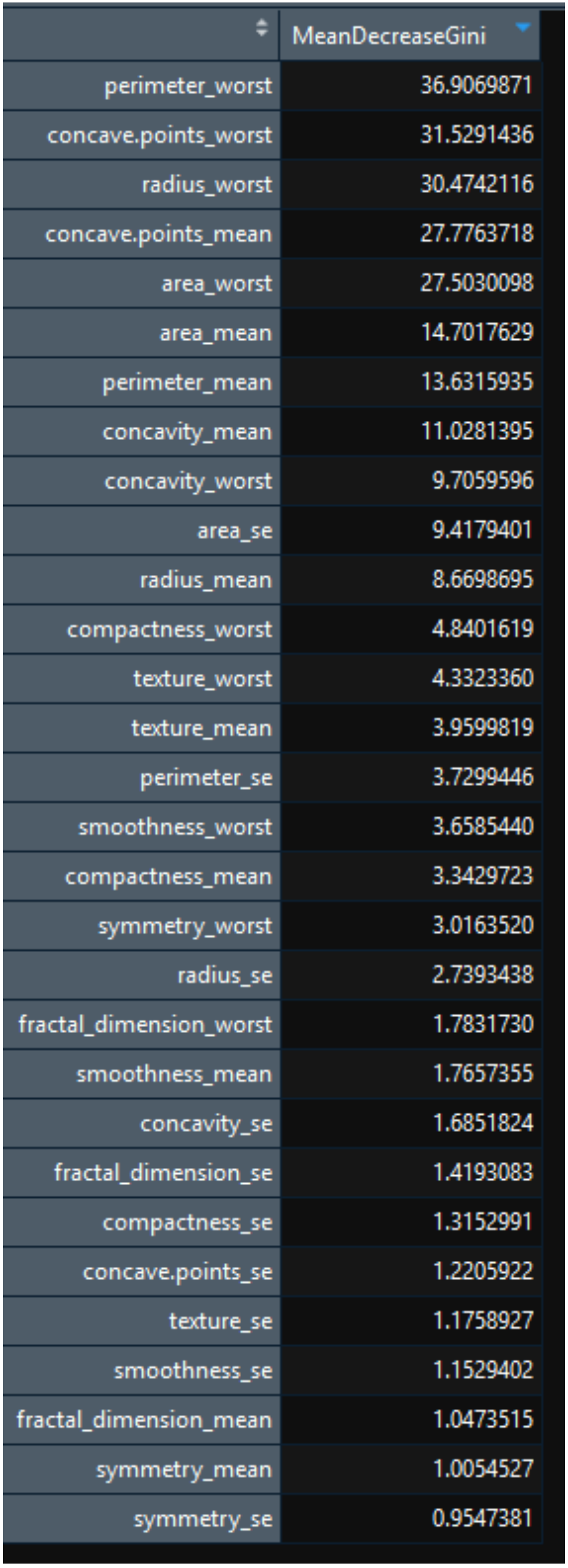
Given the scope of this project, certain models, analysis, and outputs were largely overlooked. For example, a generalized linear model could have been used for feature selection to compliment the random forest model. Additionally, diagnosis and summary statistics of MCMC were produced but were not analysed in detail. If the project continues, these items would be addressed. Having a subject matter expert would help explain correlations in the data on certain features, such as 'Compactness', being derived from other features, which would lead to skewed results due to high multicollinearity. Variations of the MCMC models and associated parameters could also be explored in a future iteration of the project.

Appendix:

[1] "Breast Cancer." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 12 June 2018, www.cdc.gov/cancer/breast/statistics/index.htm.

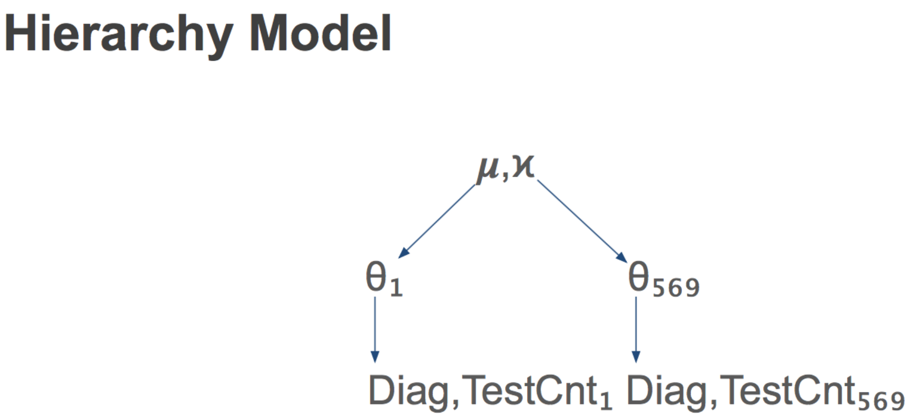
Figures:

[Fig1] Random Forest Feature Importance



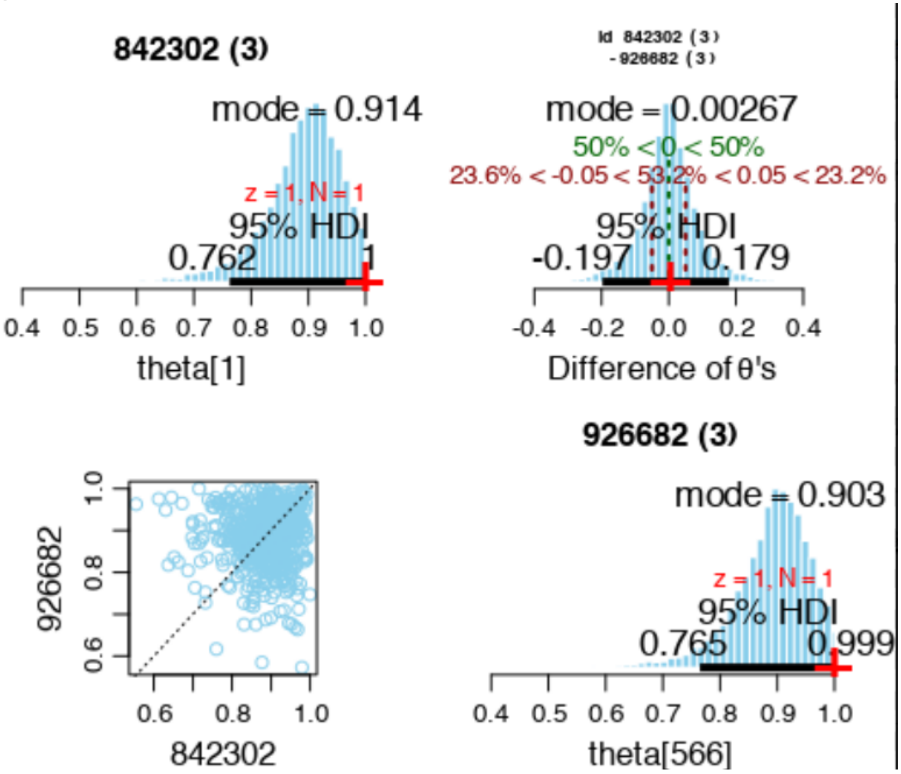
	MeanDecreaseGini
perimeter_worst	36.9069871
concave.points_worst	31.5291436
radius_worst	30.4742116
concave.points_mean	27.7763718
area_worst	27.5030098
area_mean	14.7017629
perimeter_mean	13.6315935
concavity_mean	11.0281395
concavity_worst	9.7059596
area_se	9.4179401
radius_mean	8.6698695
compactness_worst	4.8401619
texture_worst	4.3323360
texture_mean	3.9599819
perimeter_se	3.7299446
smoothness_worst	3.6585440
compactness_mean	3.3429723
symmetry_worst	3.0163520
radius_se	2.7393438
fractal_dimension_worst	1.7831730
smoothness_mean	1.7657355
concavity_se	1.6851824
fractal_dimension_se	1.4193083
compactness_se	1.3152991
concave.points_se	1.2205922
texture_se	1.1758927
smoothness_se	1.1529402
fractal_dimension_mean	1.0473515
symmetry_mean	1.0054527
symmetry_se	0.9547381

[Fig2]Hierarchy Model for the Project

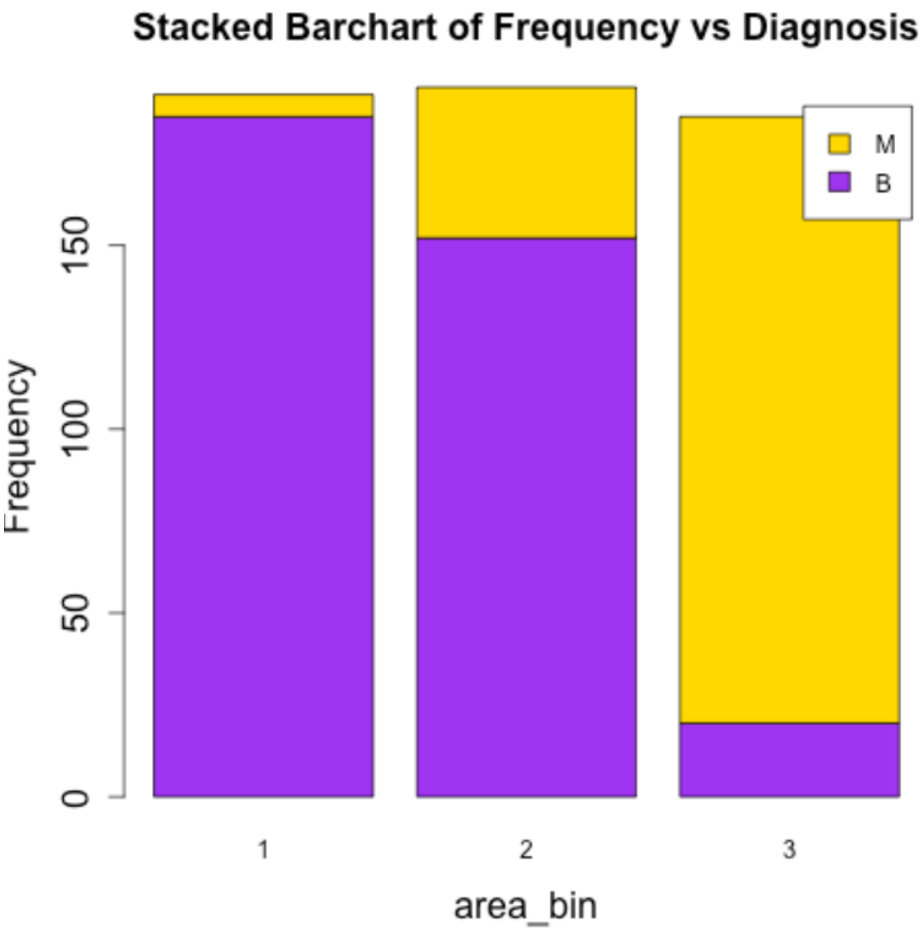


Hierarchy Model

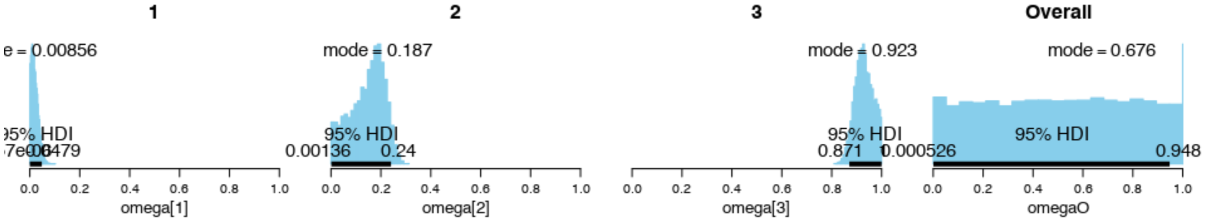
[Fig3] area mean bin 3, difference between theta1 and theta 566



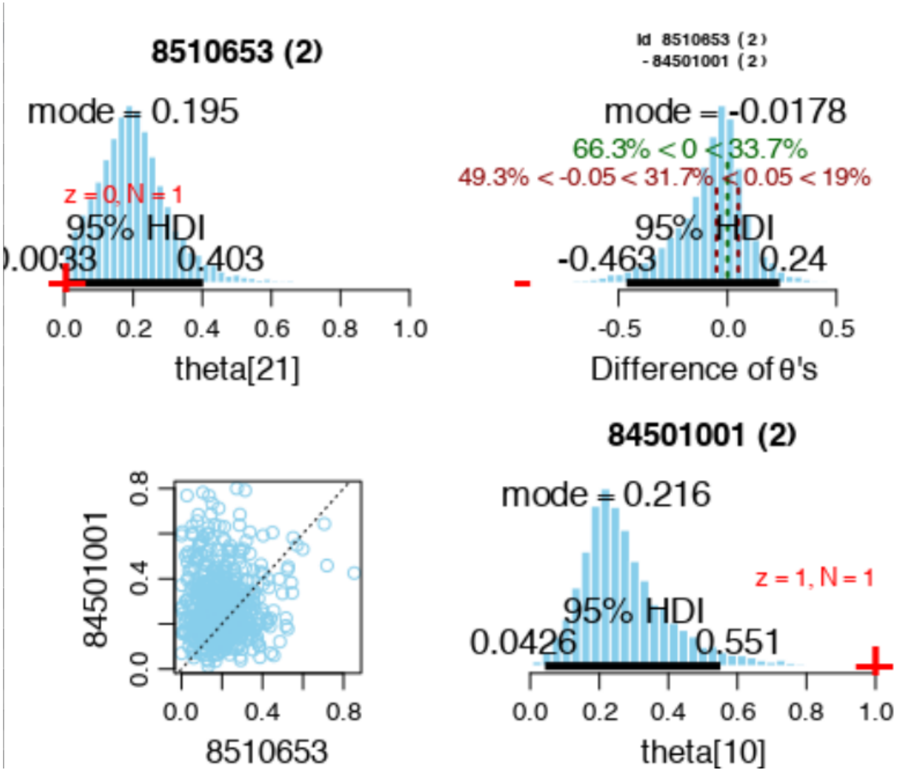
[Fig4] area mean bins stacked histogram



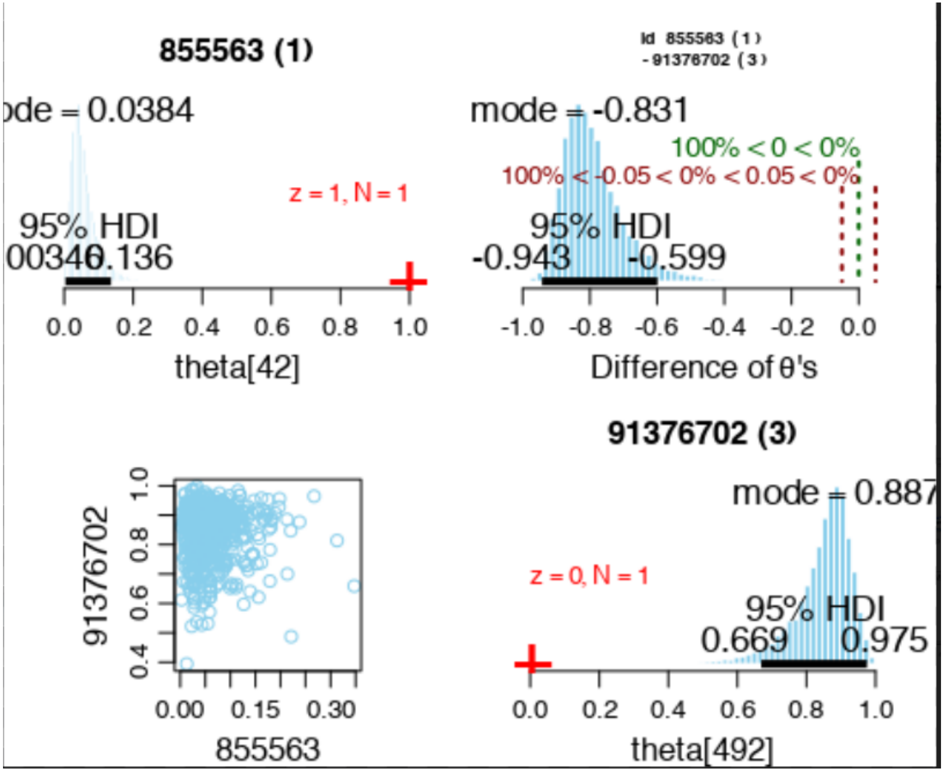
area mean binned distributions for omega



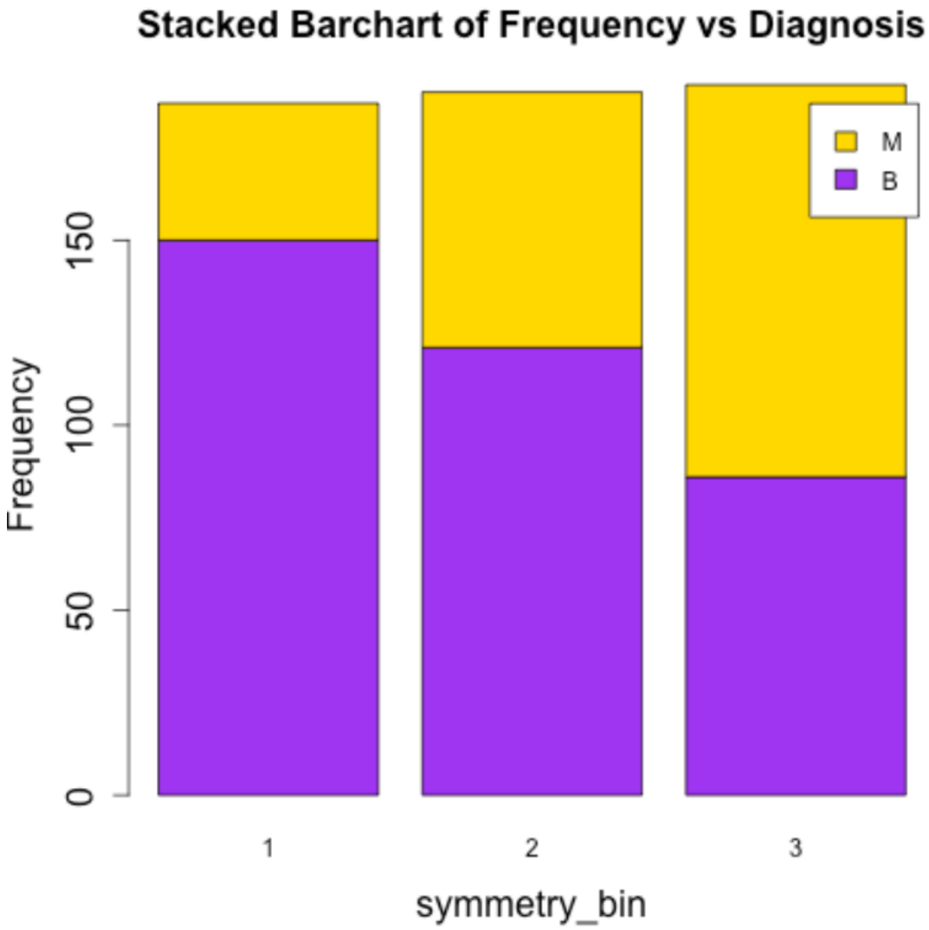
[Fig5] area mean bin 2, difference between theta 21 and theta 10



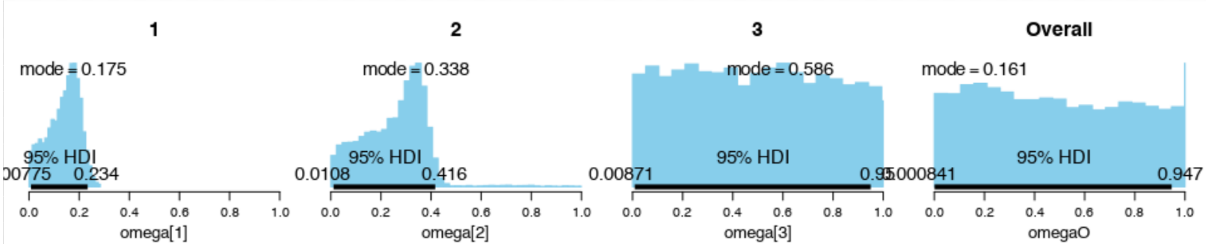
[Fig6] area mean bins 1,3 difference between theta 42 and theta 492



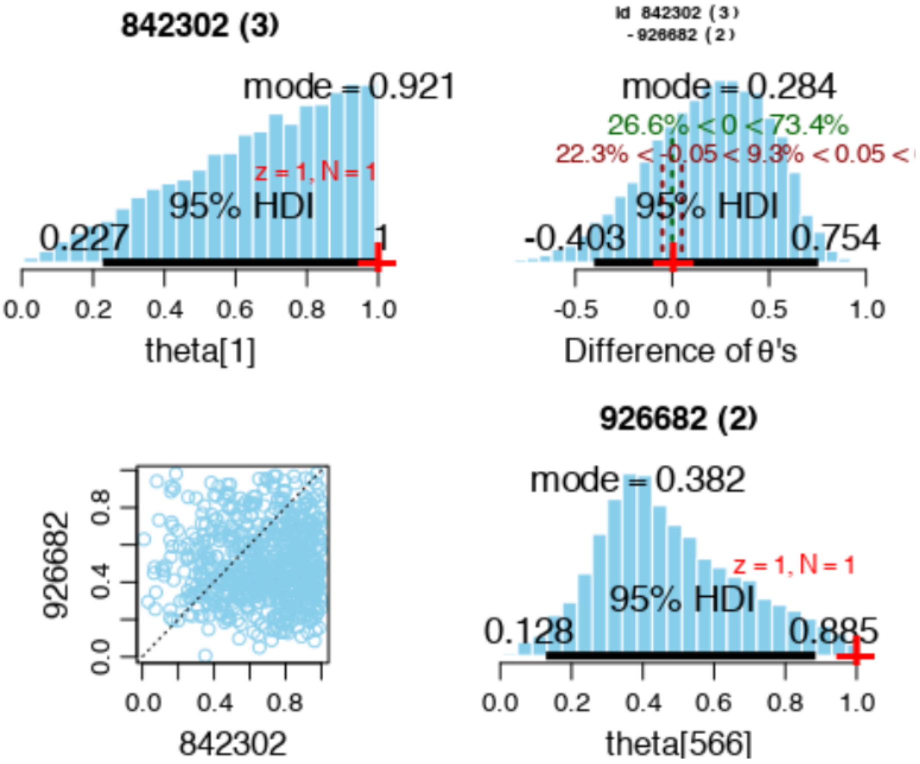
[Fig7] symmetry mean bin stacked histogram



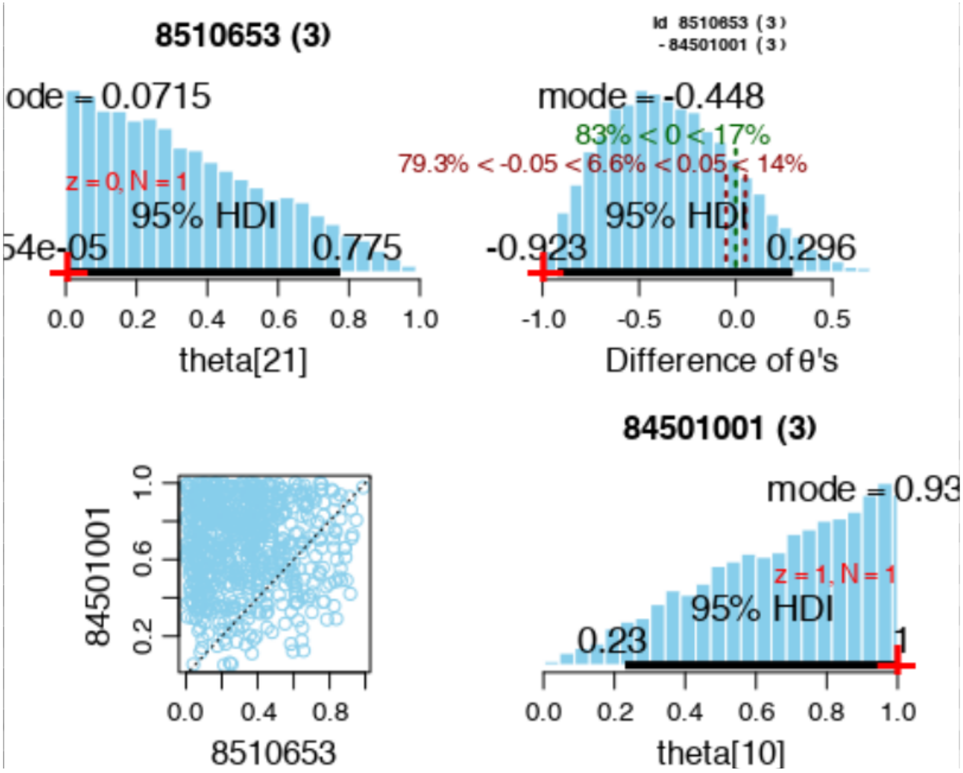
[Fig8] symmetry mean binned omega



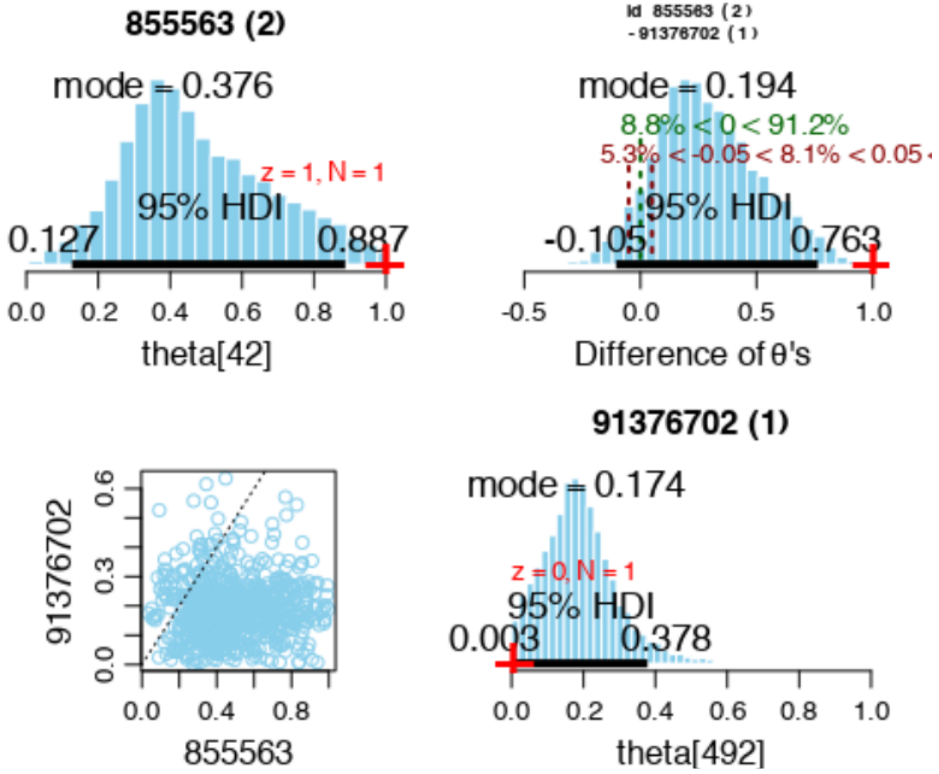
[Fig9] symmetry mean bin 3,2, difference between theta 1 and theta 566



[Fig10] symmetry mean bin3, difference between theta 21 and theta 10



[Fig11] symmetry mean bins 2, 1, difference between theta 42 and theta 492


855563 (2)
mode = 0.376
z = 1, N = 1
95% HDI
theta[42]
Id 855563 (2)
- 91376702 (1)
mode = 0.194
8.8% < 0 < 91.2%
5.3% < -0.05 < 8.1% < 0.05 <
95% HDI
Difference of θ's
91376702 (1)
mode = 0.174
z = 0, N = 1
95% HDI
theta[492]