Bayesian Inference in Breast Cancer Prediction

Yidong Wu

School of Data Science, Fudan University

June 16, 2024

Contents

Introduction

- **Breast cancer:** Affects 2.3 million women annually, over 680,000 deaths each year.
- **Traditional diagnostics:** Subjective, dependent on clinician's experience, variability in diagnosis.
- **Bayesian methods:** Incorporate prior knowledge, continuously update with new data.
- **Advantages:** Quantifies uncertainty, improves prediction accuracy, suitable for medical decision-making.
- **Project goal:** Apply Bayesian logistic regression to predict malignant or benign tumors.

Contents

Data Description

- **Dataset:** Breast Cancer Wisconsin (Diagnostic) dataset
- **Instances:** 569, **Features:** 30
- **Diagnosis:** Malignant (M) as 1, Benign (B) as 0
- **Features computed:** Mean, Standard Error, Worst (mean of three largest values)

Feature	Description
Radius	Mean of distances from center to points on the perimeter
Texture	Standard deviation of gray-scale values
Perimeter	Perimeter of the cell nuclei
Area	Area of the cell nuclei
Smoothness	Local variation in radius lengths
Compactness	(Perimeter^2 / Area - 1.0)
Concavity	Severity of concave portions of the contour
Concave Points	Number of concave portions of the contour
Symmetry	Symmetry of the cell nuclei
Fractal Dimension	'Coastline approximation' - 1

Image Samples

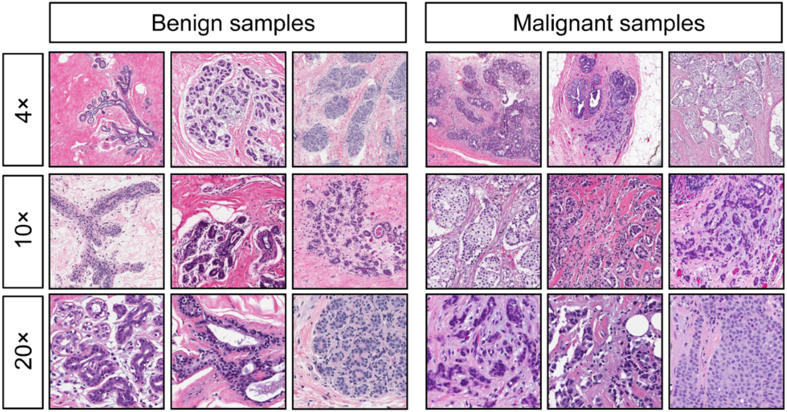


Figure: Image Samples of Breast Cancer Cells

Contents

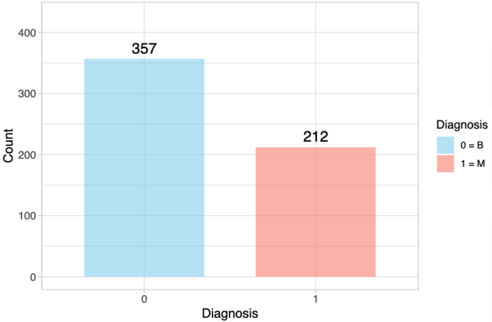
Normalization

Normalization ensures all features contribute equally to the model. This step improves performance and stability by rescaling features to a standard range, enhancing prediction accuracy and reliability.

$$X_j' = \frac{X_j - \mu_j}{\sigma_j}$$

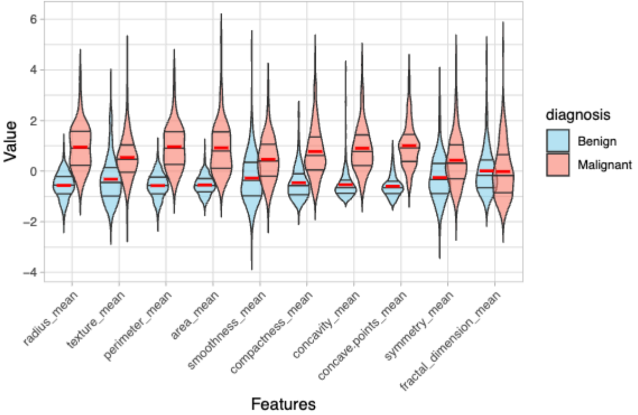
Distribution of Diagnosis

The dataset contains 357 (62.74%) benign (0) and 212 (37.26%) malignant (1) cases, showing an imbalance. Evaluation metrics must account for this imbalance to ensure accurate model performance.



Violin plots of Features by Diagnosis

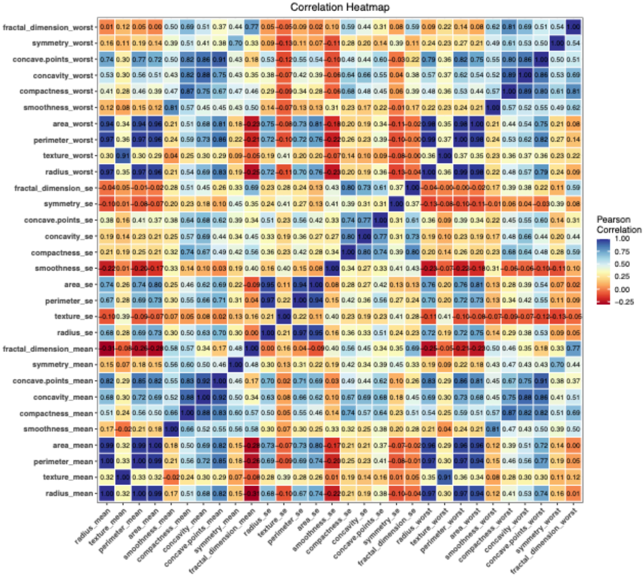
- Red line: mean, black lines: quartiles
- Distinct patterns for several features, indicating importance in distinguishing classes
- Features not skewed enough for transformation



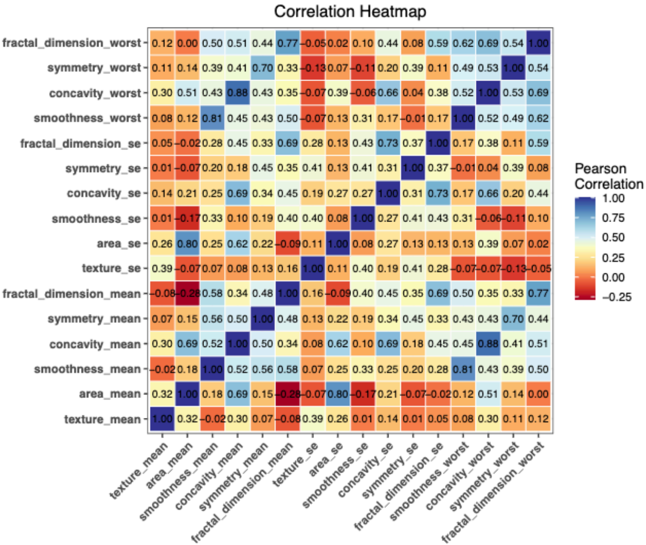
Correlation Heatmap

- Use correlation heatmap to identify and remove highly correlated features
- Reduces multicollinearity, simplifies model
- Retain most informative features, enhance accuracy and interpretability
- Example: `area_mean` highly correlated with `radius_mean` and `perimeter_mean` (both 0.99), we choose `area_mean` for more information
- No absolutely correct answer

Correlation Heatmap



Correlation Heatmap



Contents

Bayesian Logistic Regression

- Model:

$$\Pr(y = 1 \mid X, \beta) = \frac{1}{1 + \exp(-X\beta)}$$

- $\beta = \{\beta_0, \beta_1, \ldots, \beta_p\}$ are coefficients
- Likelihood function:

$$\mathcal{L}(y \mid X, \beta) = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-X_i\beta)} \right)^{y_i} \left(1 - \frac{1}{1 + \exp(-X_i\beta)} \right)^{1-y_i}$$

- Incorporate prior knowledge, obtain full posterior distribution
- Posterior distribution:

$$p(\beta \mid X, y) \propto \mathcal{L}(y \mid X, \beta) \cdot p(\beta)$$

Prior Assumptions

- β_i $(i = 1, \cdots, p)$ are i.i.d.
- $\alpha \sim \mathcal{N}(0, 10^2)$
- Here α is actually β_0
- Our X used in the following code does not have the intercept column

Priors

- Gaussian Prior:

$$\beta_j \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2)$$
$$\theta \mid \sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{4})$$
$$1/\sigma^2 \sim \text{Gamma}(4, 4)$$

- Laplace Prior:

$$\beta_j \mid b \sim \text{Laplace}(0, b)$$
$$b \sim \mathcal{N}(2, 1)$$

- Cauchy Prior:

$$\beta_j \mid \gamma \sim \text{Cauchy}(0, 2)$$
$$\gamma \sim \mathcal{N}(2, 1)$$

Other Machine Learning Methods

- Support Vector Machine
- Random Forest
- These methods are only as a supplement and contrast

Contents

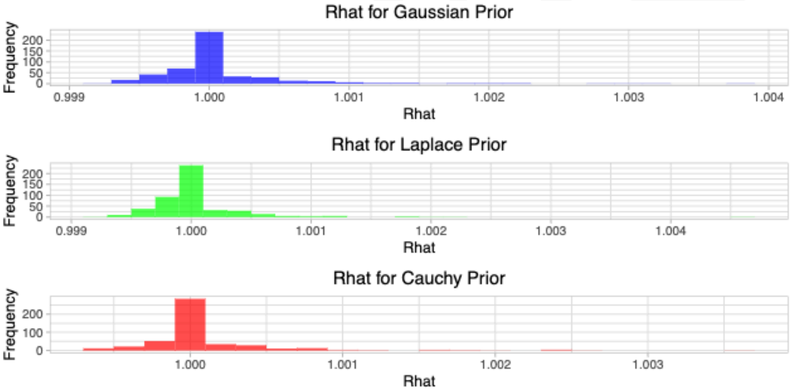
Stan Code

- Stan code too long for main report; see R Markdown file
- Set random seed for reproducibility
- Split dataset into training (80%) and test (20%) sets
- Each model: 5000 iterations across 4 chains

Contents

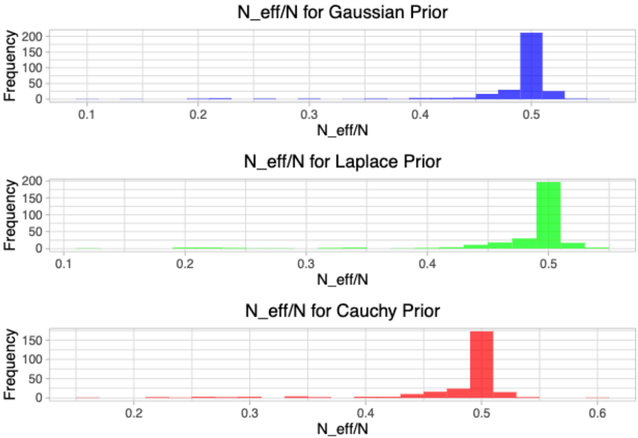
Rhat

- Rhat for all three models are very close to 1
- Chains have converged adequately
- MCMC simulations are reliable for inference.



Effective Sample Size

- Independent samples in the MCMC chain
- A relatively good mixing of the chains and a lower degree of autocorrelation

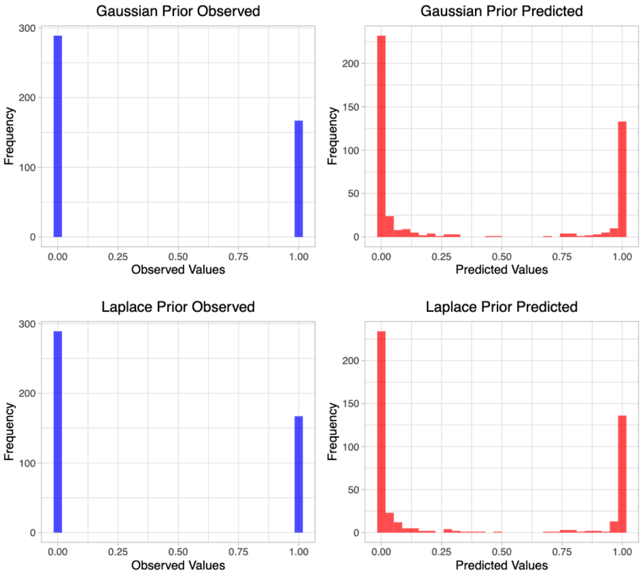


Divergences

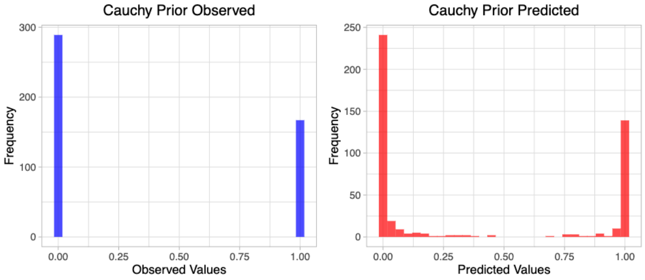
- Zero divergences for all three priors
- MCMC sampling process was stable and reliable across these priors
- No significant issues in exploring the posterior distributions.

Contents

Posterior Predictive Checking



Posterior Predictive Checking



Model	Precision	Recall	Accuracy	F1 Score
Gaussian Prior	0.9551008	0.9473359	0.9718615	0.9332871
Laplace Prior	0.9642164	0.9664749	0.9663301	0.9610915
Cauchy Prior	0.9632764	0.9723962	0.9633485	0.9831463

- Compare observed and predicted values (training set) for models with Gaussian, Laplace, and Cauchy priors
- Histograms show binary outcomes (0 and 1) and intermediate values, which are the results of averaging predictions across samples
- Most predict values close to 0 or 1 and sparse values around 0.5 indicate high confidence in predictions
- High metrics indicate excellent model fit to training data

Contents

Predictive Performance - Posterior Mean

We use the posterior means as the estimates for α and β and run our models on the test set.

Model	Precision	Recall	Accuracy	F1 Score
Gaussian Prior	0.9646018	0.9571429	0.9852941	0.9710145
Laplace Prior	0.9646018	0.9571429	0.9852941	0.9710145
Cauchy Prior	0.9557522	0.9565217	0.9705882	0.9635036

Predictive Performance - MAP

We use the posterior modes as the estimates for α and β and run our models on the test set.

Model	**Precision**	**Recall**	**Accuracy**	**F1 Score**
Gaussian Prior	0.9469027	0.9305556	0.9852941	0.9571429
Laplace Prior	0.9557522	0.9436620	0.9852941	0.9640288
Cauchy Prior	0.9646018	0.9571429	0.9852941	0.9710145

Predictive Performance - Posterior Sampling

We use all the posterior samples of α and β to make predictions on the test set and then averaged the results

Model	Precision	Recall	Accuracy	F1 Score
Gaussian Prior	0.9646018	0.9571429	0.9852941	0.9710145
Laplace Prior	0.9646018	0.9571429	0.9852941	0.9710145
Cauchy Prior	0.9557522	0.9565217	0.9705882	0.9635036

Predictive Performance

- Gaussian and Laplace priors: similar performance, robust choices
- High accuracy, precision, recall, and F1 scores for both priors
- Cauchy prior: competitive results, especially with MAP estimates
- Cauchy provides different regularization, may benefit in certain scenarios
- Consistent results across estimation methods (posterior means, MAP, posterior sampling)

Contents

Sensitivity Analysis

Now, we turn our focus to sensitivity analysis, specifically examining how variations in the parameters of the Gaussian prior influence the results. We will use these four Gaussian priors:

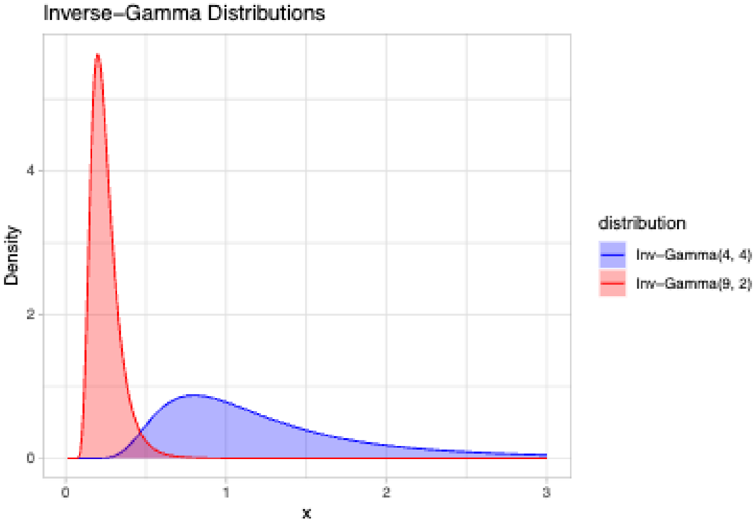
$$\begin{aligned} \text{Prior 1: } & \alpha \sim \mathcal{N}(0, 10^2), \\ & \beta_j \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \\ & \theta \mid \sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{4}), \\ & 1/\sigma^2 \sim \text{Gamma}(9, 2) \end{aligned}$$

$$\begin{aligned} \text{Prior 2: } & \alpha \sim \mathcal{N}(0, 10^2), \\ & \beta_j \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \\ & \theta \mid \sigma^2 \sim \mathcal{N}(0, \sigma^2), \\ & 1/\sigma^2 \sim \text{Gamma}(9, 2) \end{aligned}$$

Sensitivity Analysis

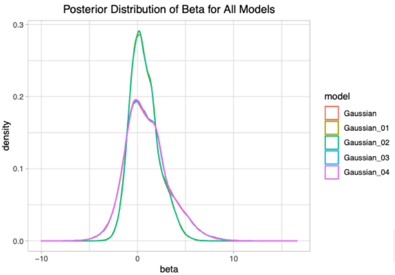
$$
\begin{aligned}
\text{Prior 3: } & \alpha \sim \mathcal{N}(0, 100^2), \\
& \beta_j \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \\
& \theta \mid \sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{4}), \\
& 1/\sigma^2 \sim \mathsf{Gamma}(4, 4) \\
\text{Prior 4: } & \alpha \sim \mathcal{N}(0, 1^2), \\
& \beta_j \mid \theta, \sigma^2 \sim \mathcal{N}(\theta, \sigma^2), \\
& \theta \mid \sigma^2 \sim \mathcal{N}(0, \frac{\sigma^2}{4}), \\
& 1/\sigma^2 \sim \mathsf{Gamma}(4, 4)
\end{aligned}
$$

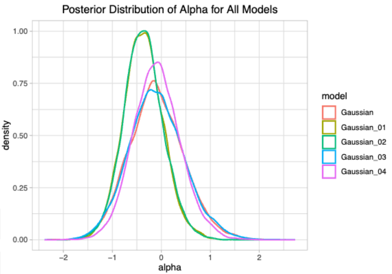
Different Distributions of σ^2



Test Performance

Model	Accuracy	Precision	Recall	F1 Score
Gaussian Prior 0	0.9646018	0.9571429	0.9852941	0.9710145
Gaussian Prior 1	0.9557522	0.9436620	0.9852941	0.9640288
Gaussian Prior 2	0.9557522	0.9436620	0.9852941	0.9640288
Gaussian Prior 3	0.9646018	0.9571429	0.9852941	0.9710145
Gaussian Prior 4	0.9646018	0.9571429	0.9852941	0.9710145





- Overlapping distributions suggest priors do not drastically change β estimates
- Slight differences in tails and peaks indicate regularization effects (Prior 1 and Prior 2)
- Changes in prior for α (Prior 3 and Prior 4) had little effects on β
- β plays a dominant role, influencing α
- α more susceptible to prior modifications
- Model is reliable with different priors

Contents

Model Comparison

- Aim: Achieve better model performance (higher accuracy, other metrics)
- Compare models: Logistic regression (three priors and no prior), SVM, random forest
- Evaluation: k-fold cross-validation ($k = 10$) due to high computational cost of LOO CV

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Gaussian	0.9683584	0.9689284	0.9801832	0.9743269
Laplace	0.9666040	0.9686110	0.9771035	0.9726442
Cauchy	0.9666040	0.9659043	0.9798062	0.9726817
No Prior LR	0.9613095	0.9649735	0.9714437	0.9680479
SVM	0.9630952	0.9659962	0.9744007	0.9699156
Random Forest	0.9630952	0.9629294	0.9779274	0.9701556

Model Comparison

- LR without prior has the worst performance, indicating effectiveness of priors
- Gaussian and Laplace priors are akin to ridge and lasso regularization
- Gaussian prior superior due to complex structure and normality assumption
- SVM and Random Forest did not surpass logistic regression models
- Possible reasons: lack of parameter fine-tuning, simplest linear kernel for SVM
- Random Forest might perform better with more raw features
- **Results are not absolute**

Contents

Discussion and Further Improvements

- **Cross Validation Approach:**
  - Used 10-fold CV instead of LOO CV
  - 10-fold CV balances computational efficiency and performance
  - Parallel computing can accelerate CV
- **Feature Selection Methods:**
  - Current method: correlation heatmap, lacks rigor
  - Future methods: RFE, decision tree-based methods, PCA for dimensionality reduction

Discussion and Further Improvements

- **Prior and Parameter Choices:**
  - Limited exploration of priors, possibly suboptimal
  - Need historical data for better priors
  - β are not i.i.d. in reality
  - Use current dataset as historical data for newer datasets
- **Dataset Limitations and Advanced Models:**
  - Dataset is relatively old
  - Contains redundant features
  - Future improvements: advanced image processing (thus using CNN for better accuracy)

Thanks!