Project Report

ISyE 6420: Fall 2024

Juejing Han, jhan446@gatech.edu

***Abstract*—**This study investigates robust techniques and ensemble methods for mitigating the impact of outliers in regression problems, using a synthetic dataset with a linear trend, Gaussian noise, and significant outliers. Robust Bayesian approaches, including Student's t-distribution and mixture models, and ensemble models like Bayesian Additive Regression Tree (BART) are evaluated. Results show that robust methods and ensemble models effectively handle outlier contamination, with Mixture BART and Mixture Bayesian models achieving superior performance. Notably, utilizing robust methods, BART models with fewer trees outperform the standard BART model with 50 trees. These findings highlight the value of robust techniques and ensemble methods for improving regression models' performance in the presence of outliers.

1 INTRODUCTION

Outliers in observations can significantly impact model performance, particularly for methods that assume normality or homoscedasticity of errors (Montgomery et al., 2012). As real-world datasets often contain anomalies or deviations from expected trends, outlier detection, and noise management are crucial challenges in regression problems. Previous studies have explored various approaches for outlier handling within the Bayesian framework, including standard Bayesian methods (West, 1984; Gelman et al., 2013), robust Bayesian regression using Student-t likelihoods (Gagnon et al., 2020), mixture models (Bishop, 2006), and Bayesian Additive Regression Trees (Chipman et al., 2010; Linero et al., 2018).

This study aims to compare frequentist and Bayesian approaches, as well as parametric and non-parametric models, on a synthetic dataset with a well-defined linear trend, Gaussian noise, and significant outliers. The primary goal is to evaluate the models' performance in capturing the main data trend while effectively mitigating the influence of outliers. Inspired by the PyMC robust regression example[1], which uses Bayesian models with Student-t likelihoods to handle outliers, this project extends the approach by incorporating outlier detection with mixture models, ensemble models, and comparisons across various regression methods.

[1] https://www.pymc.io/projects/examples/en/latest/generalized_linear_models/GLM-robust.html

2 EXPERIMENT DESIGN

2.1 Dataset

A dataset of 110 samples consisting of a linear trend with Gaussian noise and significant outliers is generated for the study (Figure 1). Outliers make up approximately 9% of the dataset. The true regression line is:

$$y = \alpha + \beta x$$

Where $\alpha = 2$ and $\beta = 3$.

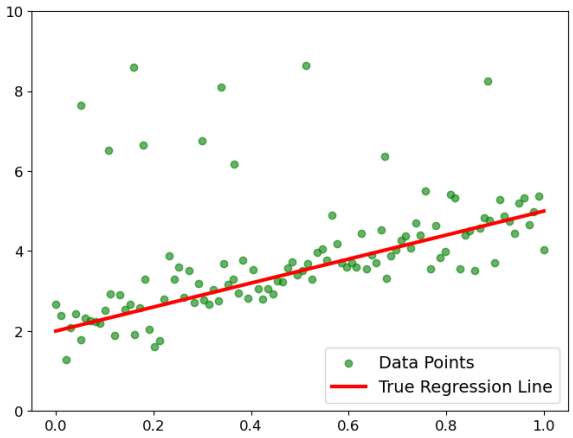

Figure 1—Scatter Plot and True Regression Line

2.2 Models

In this study, two types of regression models are inspected: linear models (frequentist and Bayesian methods) and ensemble models (Bayesian Additive Regression Trees, i.e., BART). For Bayesian models, techniques like mixture models and heavy-tailed distribution (Student's t) are utilized to further evaluate model performance. Models are fitted using PyMC with 2,000 samples per chain after a burn-in period of 1,000 (totaling 8,000 samples and 4,000 burn-in across 4 chains).

2.3 Model Evaluation Methods

Regression checks are used to assess the model's ability to capture the underlying relationship in the data by comparing the predicted regression lines (or posterior means) with the true trend. For Bayesian linear models, 25 posterior predictive samples are drawn and analyzed, while for BART, the posterior mean estimates are examined.

Metric checks are performed to quantitively measure model fit, using statistical metrics such as the Widely Applicable Information Criterion (WAIC) and Leave-One-Out Cross

Validation (LOO), which are effective metrics for evaluating the predictive performance of Bayesian models (Vehtari et al., 2017; Jung et al., 2024). This study uses the default log scale in the Python ArviZ library, where higher WAIC or LOO scores indicate better predictive accuracy. Additionally, Median Absolute Error (MedAE) is included in this study, as it focuses on the median error, which is robust to outliers and highlights the central trend of the data.

3 RESULTS & ANALYSIS

3.1 Linear Models

3.1.1 Frequentist Linear Regression (Frequentist Model)

The Frequentist model yields estimates of $\alpha' = 2.853$ (with a standard deviation of 0.249) and $\beta' = 2.125$ (with a standard deviation of 0.438). Compared to the true values ($\alpha = 2$ and $\beta = 3$), it demonstrates that the outliers, concentrated in the lower range of x (from 0 to 0.5), significantly impact the model. These outliers distort the predicted linear relationship by flattening the regression line relative to the true trend (Figure 2).

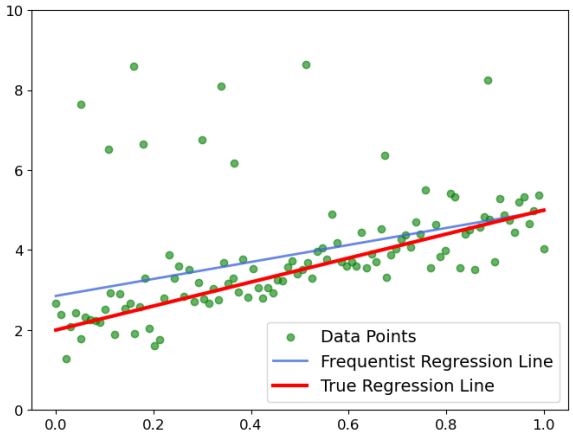

Figure 2—Result Check for Frequentist Model

3.1.2 Standard Bayesian Linear Regression (Standard Bayesian)

The Standard Bayesian model (structure shown in Figure 3, left plot) is defined as follows:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = \alpha + \beta x_i$$

$$\alpha, \beta \sim N(0, 1)$$

$$\sigma \sim HalfNormal(1)$$

Based on the data pattern shown in Figure 1, informative priors are chosen for α and β.

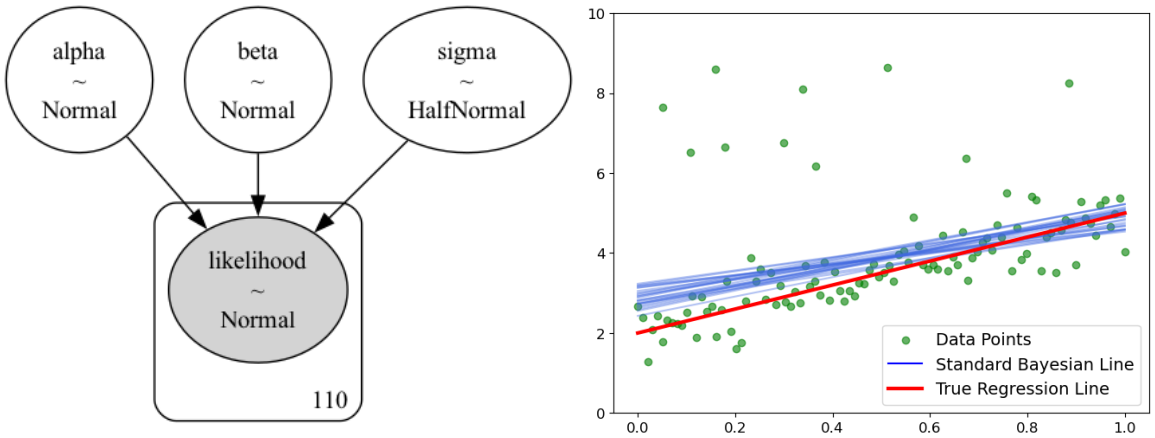

Figure 3—Model Structure (Left) and Result Check (Right) for Standard Bayesian Model

The Standard Bayesian model yields posterior means of $\alpha' = 2.861$ with a 95% HDI (Highest Density Interval) of [2.423, 3.313] and $\beta' = 2.010$ with a 95% HDI of [1.249, 2.779]. The posterior mean values are close to the estimates from the Frequentist model. However, unlike the deterministic point estimates provided by the frequentist approach, Bayesian models offer a broader range of plausible values (as shown in Figure 3, right plot), effectively reflecting the uncertainty around each parameter. Additionally, the Standard Bayesian produces smaller standard deviations (0.227 for α' and 0.393 for β') compared to the Frequentist model.

3.1.3 Robust Bayesian Linear Regression (Robust Bayesian Model)

The Robust Bayesian model (structure shown in Figure 4, left plot) is defined as follows:

$$y_i \sim StudentT(\mu_i, \sigma^2, \nu)$$

$$\mu_i = \alpha + \beta x_i$$

$$\alpha, \beta \sim N(0, 1)$$

$$\sigma \sim HalfNormal(1)$$

$$\nu \sim Exponential(\lambda = 1/30)$$

Where ν represents the degree of freedom for the Student's t-distribution. Smaller ν corresponds to heavier tails, allowing for increased robustness against outliers. Since the data exhibits a well-defined linear trend that most data points follow, with some significant outliers deviating from the trend, the exponential prior with a mean of 30 is chosen for ν. This setting ensures that the Student-t likelihood closely approximates normal

behavior for most data points while retaining heavy tails to effectively handle outliers (Lange et al., 1989; Kruschke, 2014).

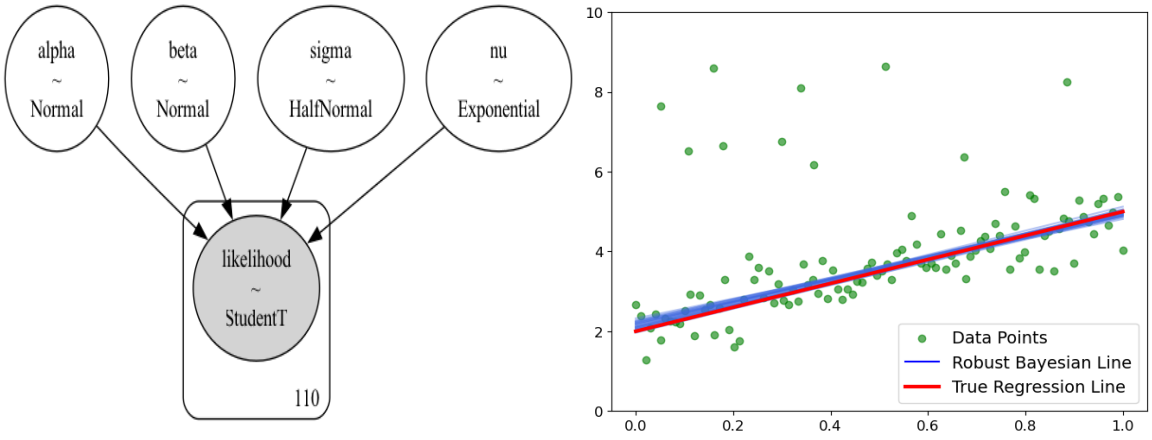

Figure 4—Model Structure (Left) and Result Check (Right) for Robust Bayesian Model

The Robust Bayesian model yields posterior means of $\alpha' = 2.129$ with a 95% HDI of [1.943, 2.315] and $\beta' = 2.829$ with a 95% HDI of [2.505, 3.165]. The posterior means are close to the true values, with the estimated regression lines closely aligning with the true regression line (Figure 4, right plot). Additionally, it produces narrower 95% credible intervals and smaller standard deviations (0.094 for α' and 0.167 for β') compared to the Standard Bayesian model, indicating greater precision in the estimates.

3.1.4 Mixture Bayesian Linear Regression (Mixture Bayesian Model)

The Mixture Bayesian model (structure shown in Figure 5, left plot) is defined as follows:

$$y_i \sim Mixture(1-\omega_i, N(\mu_i, \sigma^2_{inlier}), \omega_i, N(\mu_i, \sigma^2_{outlier}))$$

$$\mu_i = \alpha + \beta x_i$$

$$\alpha, \beta \sim N(0,1)$$

$$\sigma_{inlier} \sim HalfNormal(1)$$

$$\sigma_{outlier} \sim HalfNormal(10)$$

$$\omega_i \sim Bernoulli(p = 0.09)$$

Where ω_i represents the probability of a data point being an outlier. Given that the outliers account for approximately 9% of the total data points, p is set to 0.09. σ_{inlier} and $\sigma_{outlier}$ represent the standard deviations for inlier and outlier data points, respectively. This mixture model assigns higher variance to outliers, effectively mitigating their influence on the overall regression.

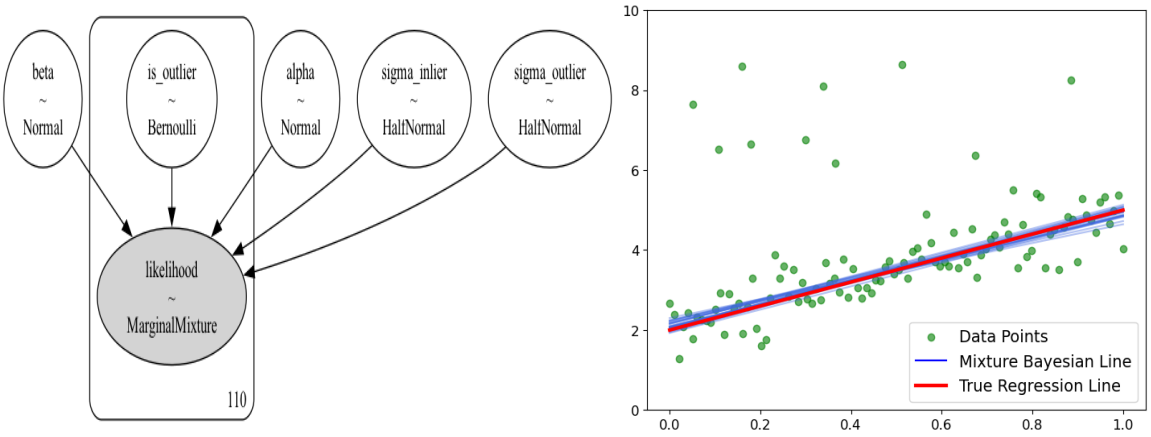

Figure 5—Model Structure (Left) and Result Check (Right) for Mixture Bayesian Model

The Mixture Bayesian model produces results similar to those of the Robust Bayesian model. The posterior means are $\alpha' = 2.136$ with a 95% HDI of [1.948, 2.335] and $\beta' = 2.804$ with a 95% HDI of [2.476, 3.141]. The standard deviations are 0.100 for α' and 0.173 for β'. The posterior means are close to the true values, with the estimated regression lines closely aligning with the true regression line (Figure 5, right plot).

3.2 Ensemble Models

While most data points follow a linear trend, the added noise and significant outliers introduce nonlinearity to the dataset. This is reflected in the Pearson Correlation Coefficient of 0.423, indicating a moderate linear relationship. Real-world datasets often combine linear and non-linear patterns, making the nonparametric BART model a flexible choice since it is well-suited for capturing complexities without a predefined data pattern (Chipman et al., 2010).

3.2.1 Standard BART Regression (Standard BART Model)

The Standard BART model (structure shown in Figure 6, left plot) is defined as follows:

$$y_i \sim N(\mu_i, \sigma^2)$$

$$\mu_i = BART(x_i)$$

$$\sigma \sim HalfNormal(1)$$

Where $BART(x_i)$ is the non-parametric BART function that adaptively models the relationship between x_i and y_i. The number of trees is set to the default value of 50.

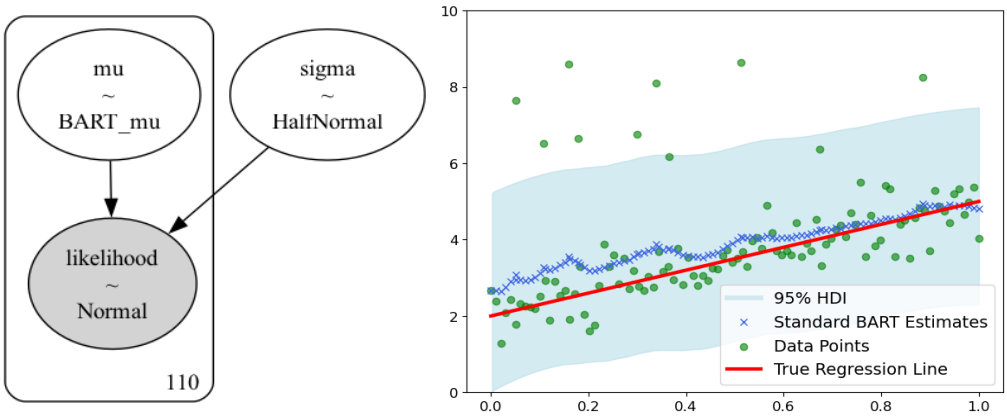

Figure 6—Model Structure (Left) and Result Check (Right) for Standard BART Model

Since BART is nonparametric, the posterior mean estimates are plotted and compared to the true regression line. The results show that the Standard BART model successfully captures the primary trend in the dataset, though it is influenced by the outliers concentrated in the lower range of x (Figure 6, right plot). The 95% HDI band is wide, indicating high uncertainty due to the presence of the outliers.

3.2.2 Robust BART Regression (Robust BART Model)

The Robust BART model (structure shown in Figure 7, left plot) is defined as follows:

$$y_i \sim StudentT(\mu_i, \sigma^2, \nu)$$

$$\mu_i = BART(x_i)$$

$$\sigma \sim HalfNormal(1)$$

$$\nu \sim Exponential(\lambda = 1/30)$$

The number of trees is set to 10.

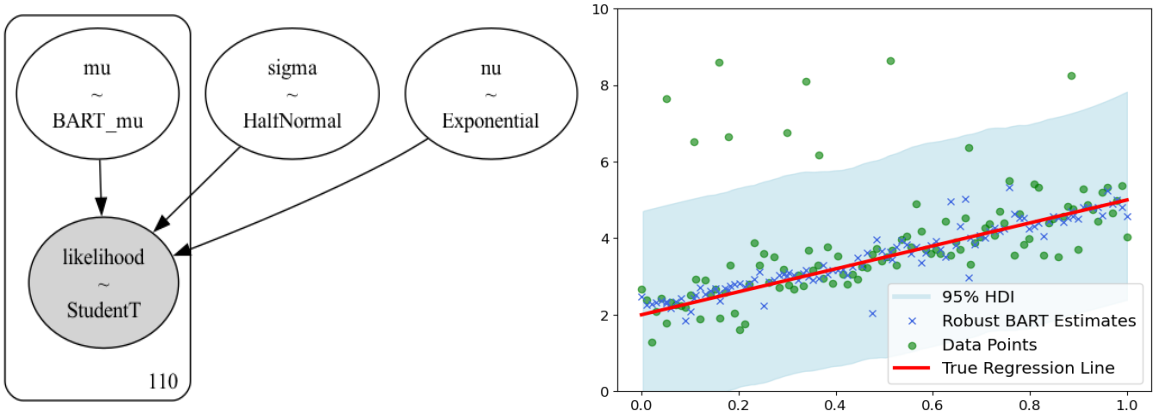

Figure 7—Model Structure (Left) and Result Check (Right) for Robust BART Model

By employing the Student's t-distribution, the Robust BART model successfully mitigates the influence of the outliers in the lower range of x (Figure 7, right plot). However, it also introduces a trade-off, as the model shows deviations from the true regression line in the higher range of x. Furthermore, the 95% HDI band is wide, suggesting a high level of uncertainty in predictions.

3.2.3 Mixture BART Regression (Mixture BART Model)

The Mixture BART model (structure shown in Figure 8, left plot) is defined as follows:

$$y_i \sim Mixture(1-\omega_i, N(\mu_i, \sigma^2_{inlier}), \omega_i, N(\mu_i, \sigma^2_{outlier}))$$

$$\mu_i = BART(x_i)$$

$$\sigma_{inlier} \sim HalfNormal(1)$$

$$\sigma_{outlier} \sim HalfNormal(10)$$

$$\omega_i \sim Bernoulli(p = 0.09)$$

The number of trees is set to 10.

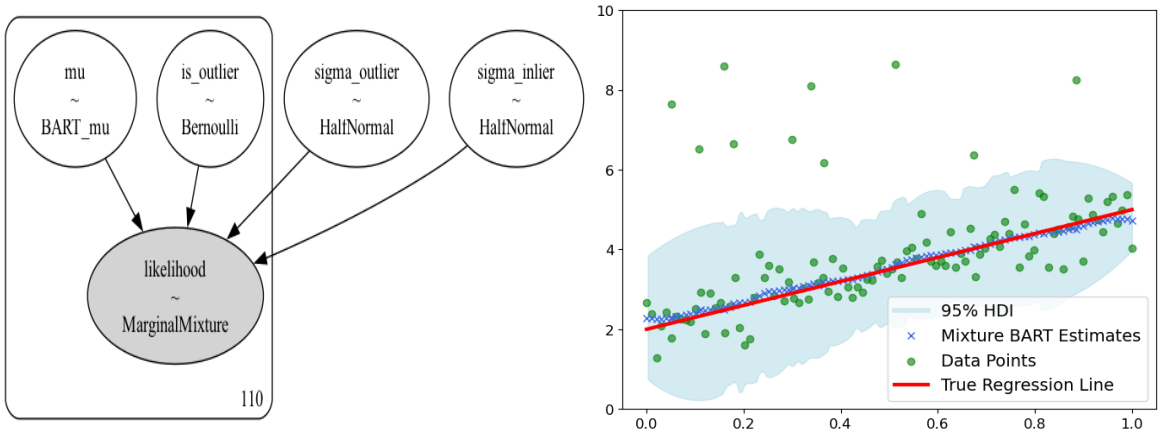

Figure 8—Model Structure (Left) and Result Check (Right) for Mixture BART Model

The Mixture BART model successfully mitigates the influence of the outliers, with its estimates overlapping the true regression line (Figure 8, right plot). This improved performance can be attributed to the model's ability to explicitly account for outliers using a mixture framework, which assigns higher variance to potential outliers, thus reducing their impact on the estimates. Moreover, the model provides a narrower 95% HDI compared to the BART and Robust BART models. This narrower HDI suggests that the Mixture BART model achieves higher confidence in its predictions by successfully balancing robustness to outliers and precision in capturing the underlying trend in the data.

3.3 Metrics Comparison

Mixture BART has the smallest MedAE of 0.313, indicating it is the most effective model at capturing the central trend of the data while being robust to outliers. Additionally, it achieves the second-best LOO score (-114.17) and WAIC score (-112.38), both with the smallest Standard Errors (SEs): LOO SE of 9.80 and WAIC SE of 9.66.

Mixture Bayesian performs similarly, achieving the second smallest MedAE (0.325) and the best LOO score (-107.26) and WAIC score (-107.02), with the second-smallest SEs: LOO SE of 9.95 and WAIC SE of 9.96.

Robust BART has the third smallest MedAE (0.340), closely followed by Robust Bayesian (0.380). Both models achieve comparable LOO and WAIC scores, ranging from -137 to -144. While Robust Bayesian achieves slightly better LOO and WAIC scores, it has marginally larger SEs compared to Robust BART.

Standard BART performs better than Standard Bayesian, with smaller MedAE and better LOO and WAIC scores, accompanied by smaller SEs.

Among all models, the Frequentist model has the largest MedAE, indicating the least robust performance in capturing the central trend.

Table 1—Metrics Comparison.

Model	MedAE	LOO	LOO SE	WAIC	WAIC SE
Frequentist	0.534	-	-	-	-
Standard Bayesian	0.522	-191.49	15.42	-191.44	15.40
Robust Bayesian	0.380	-137.30	15.99	-137.29	15.99
Mixture Bayesian	0.325	-107.26	9.95	-107.02	9.96
Standard BART	0.491	-187.11	14.92	-187.07	14.90
Robust BART	0.340	-144.20	15.55	-144.15	15.55
Mixture BART	0.313	-114.17	9.80	-112.38	9.66

It is worth noting that the mixture technique introduces warnings in the LOO and WAIC calculations, which could be a trade-off between explicit outlier management and diagnostic reliability. Despite the uncertainty surrounding the reliability of LOO and WAIC,

regression checks and MedAE comparisons consistently demonstrate the superiority of the Mixture models.

Overall, the Mixture models stand out as the best-performing models, as evidenced by the smallest MedAE values, supplementary support from the best LOO and WAIC scores, and the smallest SEs (though their reliability needs further examination). The Robust models rank as the second-best, followed by Standard BART, then Standard Bayesian, all of which outperform the Frequentist model. These findings align with the observations from the regression checks.

4 CONCLUSIONS

This study examines the robustness of Student-t and mixture model methods in effectively mitigating the influence of outliers. These approaches utilize their ability to handle heavy-tailed distributions or explicitly model outliers, leading to significant improvements in performance compared to standard methods.

The ensemble models (BART) demonstrate better performance than single regression models in capturing the underlying data trend, particularly in the presence of nonlinearity and significant outliers. Notably, when applying methods like Student-t or mixture modeling, ensemble models with fewer trees (10) outperform the Standard BART model with the default setting of 50 trees.

These findings highlight the importance of utilizing robust techniques and leveraging ensemble methods for datasets with outlier contamination.

5 FUTURE WORK

This study serves as a foundational exploration of robust techniques and ensemble methods using a relatively simple dataset. While the current experiment provides valuable insights into the effectiveness of Student-t distributions, mixture models, and ensemble approaches like BART, future work can expand in several directions.

Future experiments could focus on more complex datasets that incorporate real-world non-linear patterns, diverse noise structures, and larger scales of outlier contamination. Additionally, the exploration could be extended to include state-of-the-art models such as Bayesian deep learning, or hybrid approaches that combine parametric and non-parametric methods. To enhance the robustness of the evaluation, future work could also focus on assessing the reliability of metrics like LOO and WAIC under varying data conditions.

6 REFERENCES

[1] Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Introduction to Linear Regression Analysis (5th ed.). Wiley.

[2] West, M. (1984). Outlier Models and Prior Distributions in Bayesian Linear Regression. Journal of the Royal Statistical Society: Series B (Methodological), 46(3), 431–439.

[3] Gelman, A., Carlin, J. B., Stern, H. S., Dunson D. B., Vehtari A., and Rubin, D. B. (2013). Bayesian Data Analysis (3rd edition). CRC Press.

[4] Gagnon, P., Desgagné, A., and Bédard, M. (2020). A New Bayesian Approach to Robustness Against Outliers in Linear Regression. Bayesian Analysis, 15(2), 389–414.

[5] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

[6] Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. The Annals of Applied Statistics, 4(1), 266–298.

[7] Linero, A. R., & Yang, Y. (2018). Bayesian Regression Tree Ensembles that Adapt to Smoothness and Sparsity. Journal of the Royal Statistical Society (Series B), 80(5), 1087–1110.

[8] Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC. Statistics and Computing, 27(5), 1413–1432.

[9] Jung, A. K., and Templin, J. (2024). Evaluating WAIC and PSIS-LOO for Bayesian Diagnostic Classification Model Selection. arXiv preprint, arXiv:2410.02931.

[10] Lange, K. L., Little, R. J., and Taylor, J. M. (1989). Robust Statistical Modeling Using the t-Distribution. Journal of the American Statistical Association, 84(408), 881–896.

[11] Kruschke, J. K. (2014). Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan (2nd edition). Academic Press.