Comparing VA and MCMC
COM3023, Continuous Assessment

710022436

March 30, 2024

Abstract

This paper presents a comprehensive examination of Bayesian Neural Networks (BNNs), with a focus on exploring the efficacy of Variational Approximation (VA) and Markov Chain Monte Carlo (MCMC) techniques in approximating posterior distributions. Through a comparative study utilizing the abalone age prediction dataset, this research will shed light on the theoretical underpinnings, practical applications, and empirical evidence associated with each approximation method. The investigation reveals the inherent trade-offs between computational efficiency and the accuracy of posterior distribution approximations, highlighting the relevance of these methods in scenarios where uncertainty significantly influences decision-making. This study not only enhances our understanding of Bayesian inference within neural networks but also contributes valuable insights into the applicability of VA and MCMC in real-world data analysis, thereby informing future advancements in the field.

I certify that all material in this document which is not my own work has been identified.

1 Introduction

AI and Machine Learning has improved many sectors ranging from finance to healthcare by providing tools for data analysis, prediction, and decision-making. One of these tools are Bayesian Neural Networks (BNNs). They stand out due to their great ability in managing the uncertainties associated with the parameters of neural networks. This report explores BNNs, focusing particularly on how Bayesian inference is leveraged to navigate through the uncertainties of neural network parameters. Bayesian inference offers a probabilistic approach to learning and decision-making, this makes BNNs especially valuable in scenarios where uncertainty is a critical factor like important financial or health decisions. However, the complexity of BNNs lies in the challenge of computing the posterior distribution. This problem doesn't have a closed-form solution. Due to this, the use of approximation methods to estimate the posterior distribution become more desirable. In this field, Variational Approximation (VA) and Markov Chain Monte Carlo (MCMC) are notorious methods. This research aims to compare these two approximation techniques in terms of their effectiveness, efficiency, and applicability to real-world data. This investigation aims to deepen the understanding of Bayesian inference within the context of neural networks and enhance practical skills in applying these sophisticated approximation methods.

2 Bayesian Neural Networks

BNNs use the principles of Bayesian inference to estimate the uncertainty in the predictions made by neural networks. They infer a posterior distribution over the network's parameters, given the observed data. This posterior distribution captures the uncertainty in the model parameters, allowing the BNNs to provide not just predictions but also measures of uncertainty.[5] This feature is particularly useful in critical applications such as healthcare, autonomous driving, and financial forecasting, where understanding the confidence level of predictions is as important as the predictions themselves.

One of the significant benefits of BNNs is their ability to achieve substantial performance gains over standard neural network training and deep ensembles.[5] This is mostly because BNNs can explore a wide range of model configurations through the posterior distribution over parameters, effectively incorporating model uncertainty into predictions. This capability enables them to avoid overfitting and generalize better to unseen data, thereby improving prediction reliability, especially in scenarios where data is scarce or noisy.

Moreover, BNNs support a principled approach to model comparison and model averaging, where predictions from multiple models are combined based on their posterior probabilities. This approach naturally supports advanced machine learning concepts such as active learning, continual learning, and decision-making under uncertainty, making BNNs a powerful tool for developing sophisticated AI systems that require a high degree of reliability and interpretability in their predictions[6].

3 Comparing Methods

3.1 Theoretical Foundations

Variational Approximation (VA) is a technique that tries to approximate the true posterior distribution over neural network parameters by finding a simpler, parameterized distribution that minimizes the divergence from the true posterior. This approach uses the Kullback-Leibler (KL) divergence or the Evidence Lower Bound (ELBO) to quantify the difference between the true and the approximated distributions, aiming to minimize this divergence. Specifically, VA seeks to maximize the ELBO, which is defined as the expectation of the log likelihood minus the KL

divergence between the approximated and the true posterior distributions. This process ensures that the variational distribution closely approximates the true posterior, leveraging the trade-off between the tractability of the approximation and the fidelity to the true distribution.[2]

Markov Chain Monte Carlo (MCMC), on the other hand, is a class of algorithms that generate samples from the posterior distribution without requiring a closed-form expression of the distribution itself. MCMC evaluates $E[f(X)]$. By drawing samples $\{X_t, t = 1, \ldots, n\}$ from $\pi(.)$ and approximating

$$E[f(X)] \approx \frac{1}{n} \sum_{t=1}^{n} f(X_t).$$

The population mean gets estimated by the sample mean and it can be made more or less accurate by changing the sample size.[1]

3.2 Practical Applications and Flexibility

VA is often praised for its computational efficiency and scalability, making it suitable for large-scale applications. It is particularly advantageous in scenarios where the primary goal is to perform inference quickly and where the exact posterior distribution is less critical.

MCMC methods are renowned for their theoretical guarantee to eventually converge to the exact posterior distribution, given infinite computation time. This makes MCMC highly reliable for applications where the accuracy of the posterior approximation is paramount, albeit at a significant computational cost.

3.3 Advantages

The main advantage of VA, as demonstrated in practical applications like legislative voting blocs analysis[2], is its computational efficiency and scalability. VA transforms the problem of posterior approximation into an optimization problem, allowing for quick and reasonably accurate estimates of the posterior distribution. This characteristic is especially beneficial in large-scale applications and scenarios with limited computational resources, evidenced by its application to over 64,000 Senate press releases, achieving convergence in a notably short time frame.[2]

Conversely, the key advantage of MCMC methods lies in their ability to approximate the true posterior distribution without significant assumptions about its form, offering high versatility and accuracy. This makes MCMC particularly suitable for complex models where simpler approximation methods might not suffice. The flexibility of MCMC to adapt to various model complexities without compromising on the quality of inference underlines its continued relevance in statistical modeling, despite the computational advantages offered by methods like VA.

3.4 Limitations

The primary limitation of VA arises from its reliance on the assumption that the posterior distribution can be approximated well by a simpler, parameterized distribution. This assumption may not be valid in complex scenarios, which could potentially lead to inaccurate or biased approximations.

Although MCMC methods are theoretically powerful, they are computationally intensive and can be impractical for large datasets or models many parameters. The convergence of MCMC algorithms can take a long time, and diagnosing convergence issues can be challenging.

3.5 Empirical Evidence

Empirical studies comparing VA and MCMC often highlight the trade-off between computational efficiency and approximation accuracy. VA methods tend to be faster but might introduce

bias or underestimation of uncertainty. MCMC methods, though slower, are typically more accurate and provide a more comprehensive view of the uncertainty in model parameters.

Both VA and MCMC have their place in the toolbox of methods for approximating posterior distributions in BNNs. The choice between VA and MCMC should be informed by the specific requirements of the application, including the need for computational efficiency, the complexity of the model, and the importance of accurately capturing the posterior distribution's uncertainty. Future research and development may focus on hybrid approaches that aim to combine the strengths of both methods, potentially offering both computational efficiency and high approximation accuracy.

4 The Data

The dataset chosen for this comparative study on Bayesian Neural Networks (BNNs) is the abalone age prediction dataset, a multifaceted collection of abalone data that is particularly suited for examining the intricacies of age predictors. The robustness of this dataset lies in its inclusion of a wide variety of features that are believed to be influence by the age of an abalone. Variables such as height, sex, shell weight, diameter, and length are among the diverse array of attributes recorded. With 4178 data entries, this dataset offers a broad scope for analysis[7].

This dataset is particularly well-suited for the task at hand due to its comprehensive nature, which allows for the modelling of complex relationships between input variables and the outcome. The richness and variety of the data provide an excellent opportunity to deploy and compare the effectiveness of Variational Approximation (VA) and Markov Chain Monte Carlo (MCMC) techniques in handling uncertainties and predicting outcomes within BNN frameworks. The empirical portion of this study will involve applying these techniques to the dataset to determine which method yields the most accurate and computationally efficient predictions, potentially offering valuable insights for professionals in the ecological field.

5 Experimental Design

When comparing the performance of Variational Approximation (VA) and Markov Chain Monte Carlo (MCMC) in approximating posterior distributions within Bayesian Neural Networks. The primary objectives are to evaluate:

- Accuracy of posterior distribution approximation.
- Computational efficiency and scalability.
- Capability in handling uncertainties in predictions.

5.0.1 Dataset and Preprocessing

The abalone Prediction Dataset consists of 4178 entries with the features diameter, height, shell weight and length being selected. After initial data cleaning to remove missing values, standardising the data and putting it into training and testing sets, the model will be prepared to predict age.

5.0.2 Model Specification

- Initializes a linear layer with weights and bias as Pyro samples from normal distributions. This introduces uncertainty into the parameters, distinguishing it from a standard linear regression model.

- In the forward pass, samples a sigma value representing the noise in the data, then defines a likelihood distribution for observations given the mean (from the linear transformation of inputs) and sigma. The model either observes the actual data or makes predictions based on the mean and sigma.

5.0.3 Variational Approximation (VA)

- Clears any previous parameters stored in Pyro.
- Instantiates the model and an automatic guide for approximate posterior inference.
- An optimizer (Adam) is set with a learning rate of 0.05
- Trains the model using SVI, minimizing the negative ELBO between the model's posterior and the observed data. SVI was chosen as it can handle large datasets.[3]
- Prints out the loss periodically.

5.0.4 Markov Chain Monte Carlo (MCMC)

- The No-U-Turn Sampler (NUTS), a form of Hamiltonian Monte Carlo (HMC) that automatically tunes its parameters, is used for inference.
- MCMC with NUTS is executed to sample from the posterior distribution of the model parameters given the data. This is a way to approximate the Bayesian inference without needing explicit formulae for posterior distributions.[4]
- After running MCMC, samples from the posterior are collected. These samples represent the uncertainty in the model parameters given the observed data.

5.0.5 Metrics for Comparison

- Accuracy: The predictive accuracy of both methods will be compared using the mean squared error (MSE) between the predicted and actual ages.
- Computational Efficiency: The time taken to reach convergence or an acceptable loss level for VA, and the total sampling time for MCMC, will be compared.
- Uncertainty Quantification: The ability of each method to quantify uncertainty in predictions will be assessed by examining the width of the 95% credible intervals for the predicted ages and how these intervals capture the observed data.

6 Results

Table 1: Performance Metrics Comparison

Metric	VA Model	MCMC Model
MSE	10.268	6.782
R-squared	82.772	91.224
Run Time (s)	15	120

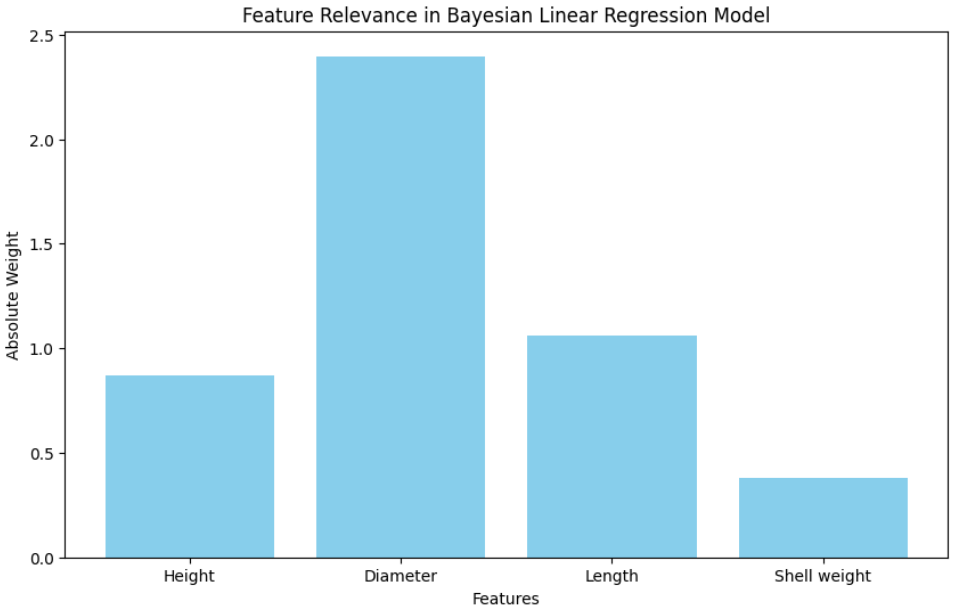

Figure 1: Feature relevance plot

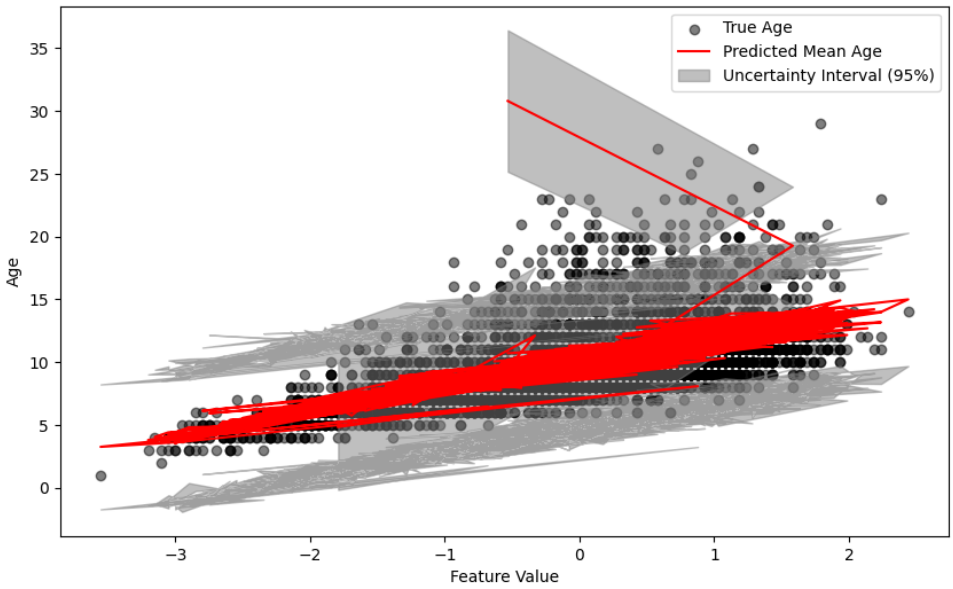

Figure 2: Predictive distribution of VA with 95% CI

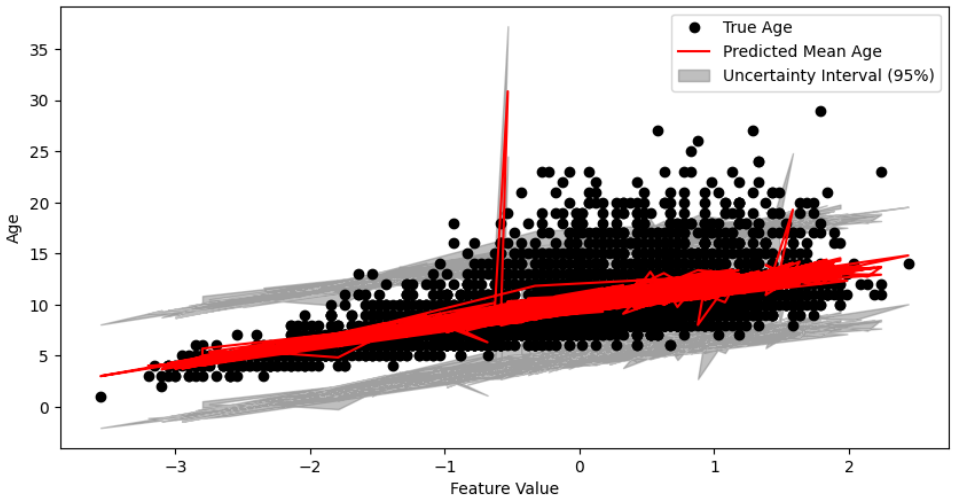

Figure 3: Predictive distribution of MCMC with 95% CI

7 Results Analysis

7.1 Initial Results

In the initial stages of this research, the objective was to construct a Bayesian Neural Network (BNN) that was capable of accurately predicting the age of abalones based on a set of physical features. The initial model architecture incorporated two hidden layers. The preliminary results using Variational Approximation (VA) were promising, as evidenced by a Mean Squared Error (MSE) of 6.423, indicating a high degree of accuracy.

However, the implementation of MCMC was met with substantial computational challenges. Unlike the VA method, which optimizes a lower bound on the model evidence, MCMC requires extensive sampling to approximate the posterior distribution, substantially increasing the computational burden. The two-hidden-layer model proved to be too complex, consuming all available RAM and rendering my local Jupyter Notebook environment incapable of processing the algorithm.

In an attempt to circumvent these limitations. the workspace was moved to Google Colab, which boasts a higher allocation of computational resources, including RAM. Yet, the problem persisted; the MCMC method continued to deplete all available memory.

The persistent computational difficulties necessitated a re-evaluation of the model architecture. Both models were reduced to a single hidden layer. This decision was not without consequence; the reduction in model complexity invariably led to a loss in the capacity to model the complex dependencies within the data. Notably, the two-layer VA model had already demonstrated a higher level of predictive precision, as indicated by the lower MSE score.

Nonetheless, the simplification to a single hidden layer was essential to facilitate a fair comparative study between the VA and MCMC methods. This adjustment allowed the MCMC model to function within the computational confines provided by Google Colab. The MCMC's demand for computational resources is particularly pronounced when dealing with models of high complexity, and the reduction in complexity was the only viable solution to enable the model to run to completion.

7.2 Feature Analysis

The 'Diameter' feature has the highest absolute weight, significantly surpassing the others, which indicates that it has the greatest influence on the age prediction in the dataset. The

large weight suggests that changes in the diameter of an abalone are strongly associated with changes in its age.

'Height' has the next highest weight, albeit it is less than half of the weight of 'Diameter', suggesting a moderate relevance to the age prediction. 'Length' follows, with a lower weight than 'Height', indicating it has a smaller yet non-negligible effect. 'Shell weight' has the least weight of the four, suggesting that it has the smallest direct influence on the age prediction according to this model.

These weights are crucial in understanding the model's behavior and in the context of feature engineering. For instance, they suggest that 'Diameter' might be the most informative single feature for age prediction and should be prioritized in model training and interpretation. Moreover, it suggests that, in the context of Bayesian model selection or regularization, 'Shell weight' might contribute less to model performance and could be a candidate for exclusion in simpler models to prevent overfitting.

The model's reliance on 'Diameter' aligns with biological knowledge since larger diameters generally indicate older abalones. This is a useful validation of the model, suggesting that it is learning meaningful patterns in the data rather than spurious correlations. For researchers or practitioners, such insights into feature relevance are valuable for refining data collection strategies, prioritizing certain measurements, and improving the interpretability of the model's outputs.

7.3 Metrics Analysis

The Table and the plots show a performance comparison between two Bayesian Neural Network (BNN) models applied to the abalone dataset, with plots depicting the true versus predicted age, as well as Table 1 which summarizes key performance metrics.

Starting with Table 1, we observe a significant difference in the Mean Squared Error (MSE) with the MCMC model achieving a lower value of 6.782 compared to 10.268 of the VA model. This suggests that the MCMC has a better fit to the data, possibly capturing the underlying distribution more accurately. This is further supported by the R-squared values, where the MCMC model again leads with 91.224% over the 82.772% of the VA model, indicating a higher proportion of variance explained by the MCMC model. However, this superior performance comes at the cost of computational efficiency, with the MCMC model taking 120 seconds to run, significantly longer than the 15 seconds for the VA model.

Examining Figure 2, the VA model shows a clear trend of the predicted mean age increasing with the feature value. The uncertainty interval depicted in gray suggests that the model is reasonably confident about its predictions for most of the feature value range but shows increasing uncertainty for higher feature values. The red line, indicating the mean predicted age, generally follows the true age points (black dots), but the spread indicates variance in predictions.

In Figure 3, the MCMC model demonstrates a more confident prediction trend, with a narrower uncertainty interval for a significant portion of the data, particularly around the center of the feature value range. It's noteworthy that there is a spike in uncertainty at a feature value around 1.5, which might indicate a potential outlier or an area where the model is less certain about its predictions.

From a comparative perspective, the MCMC model appears to provide a more accurate representation of the dataset with its narrower uncertainty intervals and higher R-squared value, pointing to its ability in capturing the nuances in the data. The VA model, while less computationally demanding, seems to offer a wider range of prediction intervals, which could suggest a less precise understanding of the data dynamics.

Considering the trade-offs, the choice between VA and MCMC hinges on the specific requirements of the task at hand. For applications where computational resources or time are limited, the VA method offers a viable alternative with acceptable accuracy. However, for tasks

where the utmost precision is necessary and computational resources are abundant, the MCMC model clearly stands out as the superior choice.

The MCMC model's enhanced ability to characterize the uncertainty in the abalone dataset is evident in both the numerical metrics and visual analysis of the plots. Despite its computational intensity, the precise estimation of the posterior distribution through MCMC could be critical in high-stakes scenarios where accurate predictions are paramount. Conversely, the VA model's efficiency may be preferred for quick assessments or when resources are constrained, albeit at the expense of some accuracy and a broader confidence interval for predictions.

8 Conclusion

In conclusion, this research intricately analyzes the performance of VA and MCMC methods within the realm of Bayesian Neural Networks, utilizing the abalone dataset for empirical validation. The findings indicate that MCMC provides a more precise approximation of the posterior distribution, albeit with higher computational demands, while VA offers a balance between speed and accuracy, making it suitable for applications with constrained computational resources. The comparative analysis underscores the significance of selecting an appropriate approximation technique based on the specific needs of the task, whether it prioritizes computational efficiency or the accuracy of uncertainty estimation. Future endeavors in this domain may explore hybrid models that amalgamate the strengths of both VA and MCMC, potentially introducing a new era of computational efficiency coupled with high accuracy posterior approximation.

References

[1] Walter R Gilks, Sylvia Richardson, and David J Spiegelhalter. *Markov Chain Monte Carlo In Practice*. 1995.

[2] Justin Grimmer. An Introduction to Bayesian Inference via Variational Approximations. `https://stanford.edu/~jgrimmer/VariationalFinal.pdf`, 2010.

[3] Matt Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. `https://arxiv.org/abs/1206.7051`, 2013.

[4] Matt Hoffman and Andrew Gelman. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. `https://www.jmlr.org/papers/volume15/hoffman14a/hoffman14a.pdf`, 2014.

[5] Pavel Izmailov, Sharad Vikram, Matthew D. Hoffman, and Andrew Gordon Wilson. What Are Bayesian Neural Network Posteriors Really Like? `https://proceedings.mlr.press/v139/izmailov21a.html`, 2021.

[6] Kononenko. I. Bayesian neural networks. Biol. Cybern. 61. `https://doi.org/10.100/BF00200801`, 1989.

[7] Rodolfo Mendes. Abalone Dataset. `https://www.kaggle.com/datasets/rodolfomendes/abalone-dataset`, 2018.