ML–DS I: Project VI Report

Sebastijan Trojer
ML-DS I 2024/25 , FRI, UL
st5804@student.uni-lj.si

I. Introduction

In this assignment our task was to leverage Bayesian inference to gain insight into the relationship between the explanatory and target variables the latter being the number of goals scored. We utilized the Poisson GLM and used Markov Chain Monte Carlo (MCMC) diagnostics, such as autocorrelation and traceplot to verify the computation. Lastly we derived the Poisson posterior and its derivatives and applied them to compute the Laplace approximation of the parameters.

II. Methodology and results

Our initial task was to construct a Bayesian Poisson model on the data we were provided with. The data consisted of 8 explanatory variables that consisted of pairs, specifically, we had information about each team's score rate, concede rate, corner, and foul ratio, and each data point represented a football match.

A. Building the model

We built the model using the library *pymc* that supports MCMC inference. First, the data was standardized and an intercept term was added. Then we built the model. We placed a normal prior on the coefficients, specifically $\beta \sim N(0, 5)$. The likelihood assumed that the target variable follows a Poisson distribution, specifically $y \sim Poisson(\lambda); \lambda = e^{X\beta}$ where y is the target variable (number of goals), X the feature matrix, and β the model parameters.

B. Exploring the model

After the model was built, we perfromed Bayesian inference using MCMC. We ran five independent chains, each drawing 1000 posterior samples, following a tuning phase of 1000 iterations per chain, which is also reflected by the chain stability in the traceplot (Figure 2) and the autocorrelation plots (Figure 1) which we used to better understand the model.

Specifically, if we inspect the trace plot and the posterior distributions of the parameters, shown in Figure 2, further, we notice that the trace plots indicate that the model had converged. The chains on the trace plot appear well-mixed and stationary, and the posterior distributions on the left plot seem well separated and uni-modal. The posterior distributions are shown per chain and their close alignment

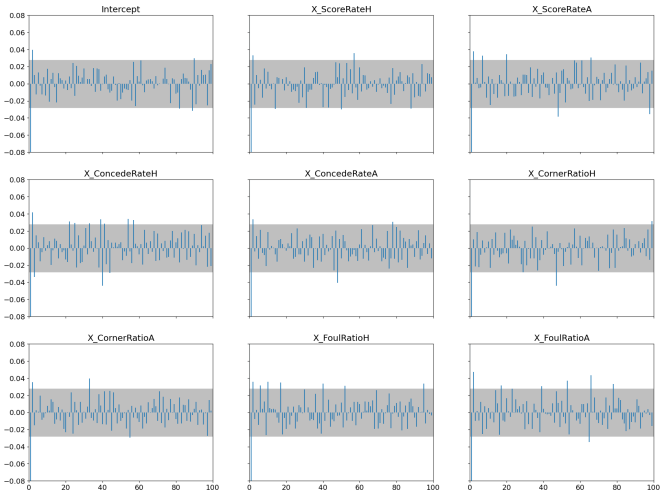

Fig. 1. Autocorrelation of samples corresponding to each feature.

suggests that the sampler consistently explores the same region of the posterior.

The posterior distributions also allow us to draw conclusions about the features and their relationship to the target variable. For example, the parameter corresponding to the score rate of the home team is the most positive among all of them, even in absolute value, indicating that it has the most impact on the number of goals that were scored in a match, and since it is positive, it implies that this relationship is positive, meaning that the higher the score rate of the home team, the higher the number of goals scored. Another example is the corner rate of the away team, which seems to have minimal impact on the number of goals. We can draw similar insights from the other distributions.

Parameter	ESS	$\hat{R}$
ScoreRateH	6703	1.00
ScoreRateA	6879	1.00
ConcedeRateH	8157	1.00
ConcedeRateA	7779	1.00
CornerRatioH	7103	1.00
CornerRatioA	6733	1.00
FoulRatioH	7553	1.00
FoulRatioA	7826	1.00

TABLE I
Mean ESS scores and $\hat{R}$ values for all variables.

We also evaluated the effective sample size (ESS) and the $\hat{R}$ statistic for each parameter, as summarized in Table I. The

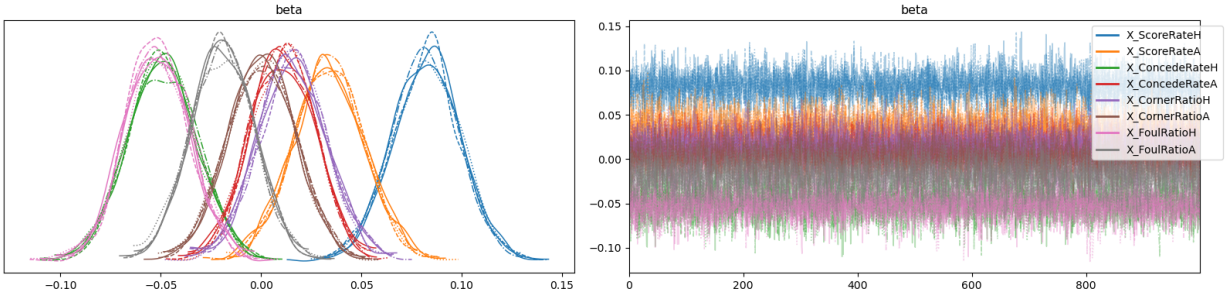

Fig. 2. Posterior distributions (left) and trace plot (right). The posteriors are shown for each chain separately.

$\hat{R}$ values are all equal to 1.00, indicating great convergence across chains. The ESS values are high, suggesting a large number of effectively independent samples. Overall the analysis indicates that the posterior is credible and do not need to refine the model.

C. Laplace Approximation

In the second part of the assignment we focused on fitting the model using Laplace approximation, which relies on approximating the posterior by a Gaussian centered at the mode (MAP), so we don't have to rely on MCMC.

First, we derived the posterior, its gradient and Hessian matrix (see Appendix). Then we used an optimizer to find the mode of the posterior and approximate with a Gaussian. The covariance of the Gaussian is determined by the inverse of the Hessian that we derived. That gives us both parameters that we require to approximate the posterior.

1) MCMC vs. Laplace approximation – comparison: Figure 3 shows the posteriors from MCMC and Laplace approximation on a single plot. Notice that the distributions match closely, which confirms that the approximation works well, the key difference being that the latter are perfect Gaussians unlike MCMC distributions.

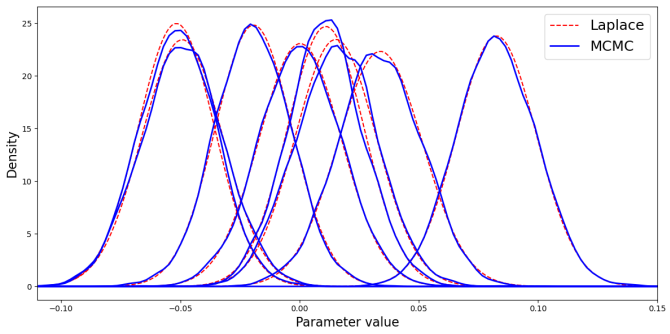

Fig. 3. Posterior densities from MCMC (blue) and Laplace approximation (red).

2) Predicting on test data: Lastly, we used the Laplace-approximated model to generate predictions on the test set. Specifically, we drew 1000 samples from the posterior distribution of each parameter and used these to make predictions for each data point, resulting in 1000 predicted values per instance. We then aggregated these predictions in three different ways to simulate three evaluation metrics: squared error, absolute error, and accuracy. These correspond to the mean, median, and mode of the sampled predictions, respectively.

Figure 4 shows two distributions. The left panel presents the target distribution, computed by sampling from the MCMC posterior and generating predictions accordingly. It includes both the observed targets and the predicted values, and spans a wider range. The right panel shows the distribution of predictions from the Laplace approximation. Note that for the accuracy metric, predictions are binned into whole numbers, since accuracy requires exact match and the target variable is an integer. In contrast, predictions evaluated using squared and absolute error are continuous and closely aligned. In practice, the strategy that will yield best results depends on the use case, so it's hard to comment on that further.

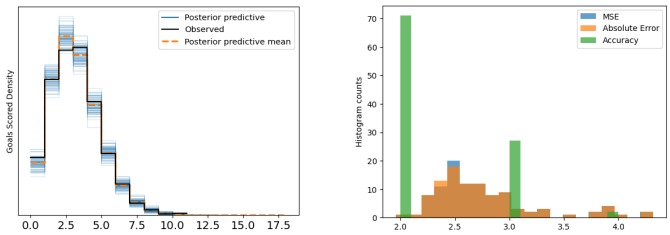

Fig. 4. Observed and predicted target distribution with MCMC model (left) and predicted target distributions using three distinct loss functions using the Laplace approximated model (right).

III. Conclusion

In this assignment, we built two Bayesian predictive models and thoroughly analyzed them. The first one, MCMC based, allowed us to verify model reliability by analyzing standard MCMC diagnostics, such as autocorrelation and ESS, while the Laplace approximated model demonstrated that such approximations can work and that we can approximate the parameters relatively well, if certain assumptions hold. Lastly, we demonstrated how the model can be used to make predictions using several different loss functions without the need for retraining, unlike in the MLE paradigm.

IV. Appendix

$p(y|x) = p(x|y)\,p(y)$ — Poisson lik. $\sim$ Normal $(0, \alpha I)$

Log prior: $f(\beta) = \frac{1}{(2\pi)^{\frac{k}{2}} (\det(\alpha I))^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\beta^T(\alpha I)^{-1}\beta\right)$

$l(\beta|x,y) = \prod^n f(\beta)$

$\log l(x|y) = \log \frac{1}{(2\pi)^{k/2}(\det(\alpha I))^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2}\beta^T(\alpha I)^{-1}\beta\right)$

$= \log\left(\frac{1}{(2\pi^{k/2})\cdot(\det(\alpha I))^{\frac{1}{2}}}\right) + \left(-\frac{1}{2}\beta^T(\alpha I)^{-1}\beta\right)$

$= \log(\text{—11—}) - \frac{1}{2}\beta^T(\alpha I)^{-1}\beta$

$= \log(\text{—11—}) - \frac{1}{2}\beta^T\frac{1}{\alpha}I\beta = \log(\text{—11—}) - \frac{1}{2\alpha}\beta^T\beta$

Log likelihood:

$p(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \qquad \lambda = e^{x^T\beta}$

$p(y|\beta,x) = \frac{1}{y!} e^{x^T\beta y} \; e^{-e^{x^T\beta}}$

$L(y_i|\beta,x_i) = \prod^n \left(\frac{1}{y_i!} e^{x_i^T\beta y_i} e^{-e^{x^T\beta}}\right)$

$l(y_i|\beta,x_i) = \sum^n \log\left(\frac{1}{y_i!} e^{x_i^T\beta y_i} e^{-e^{x^T\beta}}\right) = \sum^n \log\frac{1}{y_i!} + \sum^n x_i^T\beta y_i - \sum e^{x^T\beta}$

Posterior

$\log p(\beta|x_i,y_i) \propto \log p(y_i|\beta,x_i) + \log f(\beta) =$

$= \sum^n \log\frac{1}{y_i!} + \sum^n x_i^T\beta y_i - \sum e^{x^T\beta} + \log(\text{—11—}) - \frac{1}{2\alpha}\beta^T\beta$

$= \sum^n x_i^T\beta y_i - \sum^n e^{x_i^T\beta} - \frac{1}{2\alpha}\beta^T\beta = \sum^n \left(y_i^T x_i\beta - e^{x_i^T\beta}\right) - \frac{1}{2\alpha}\beta^T\beta$

$\Rightarrow \log p(\beta|x,y) \propto y^T(x\beta) - \sum^n e^{x_i^T\beta} - \frac{1}{2\alpha}\beta^T\beta$

$\frac{\partial l}{\partial \beta} = x^T y - x^T e^{x^T\beta} - \frac{1}{\alpha}\beta$

$\frac{\partial}{\partial\beta} a^T\beta = a$

$\frac{\partial}{\partial\beta}(y^Tx)\beta = (y^Tx)^T = x^Ty$

$\frac{\partial^2 l}{\partial^2\beta} = -\sum^n x_i^T e^{x_i^T\beta} x_i - \frac{1}{\alpha}I = -x^T \operatorname{diag}(e^{x^T\beta})\,x - \frac{1}{\alpha}I$