Bayesian Inference

Marko Medved

I. Implementation of the Poisson GLM with Bayesian inference using MCMC

A. Implementation details

For this part of the assignment, we implemented a Poisson Generalized Linear Model (GLM) using Bayesian inference. As a preprocessing step, we standardized all predictor variables. For the prior distribution of the regression coefficients (β), we used a normal distribution with zero mean and a large variance ($\alpha = 10$). This choice reflects a weakly informative prior—wide enough to avoid imposing strong constraints on the coefficients, allowing the data to primarily inform the posterior.

To perform inference, we used the Markov Chain Monte Carlo (MCMC) method to sample from the posterior distribution. To assess convergence reliably, we ran 4 independent chains. Each chain generated 1000 posterior samples following a burn-in (warm-up) phase of 1000 iterations, which were discarded.

B. Relationship between the explanatory and GoalsScored variables

To describe the relationship between features and the target variable, we plot the posterior densities of the regression coefficients in Figure 1. The Score Rate of the home team stands out as the most important feature, showing a clear positive effect on the expected number of goals (through the exponential link). In contrast, the home team's foul ratio (somewhat unexpectedly) and concede rate (as expected) have the strongest negative effects, reducing the predicted goal count. Other coefficients are close to zero, indicating little effect since their multiplicative impact on the mean is near one after applying the link function. The intercept was excluded from this visualization for clarity but shows a similar distribution centered around one.

The similar widths of the coefficient distributions suggest comparable uncertainty across all features.

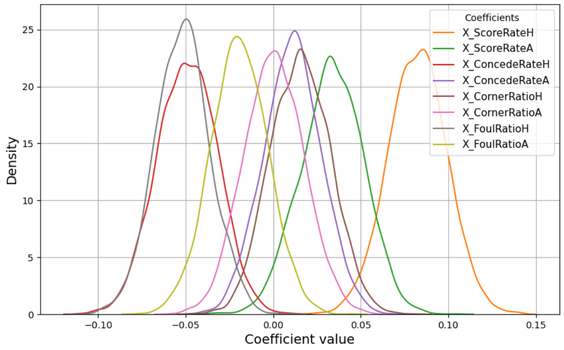

Figure 1. Densities of coefficients for the features with the exclusion of the intercept density

C. MCMC diagnostics

We use standard MCMC diagnostics to assess convergence. Figure 2 shows the trace plots of all regression coefficients across the 4 chains. The samples for each parameter appear stable and well-mixed, with no visible trends, and the chains overlap closely—indicating good convergence.

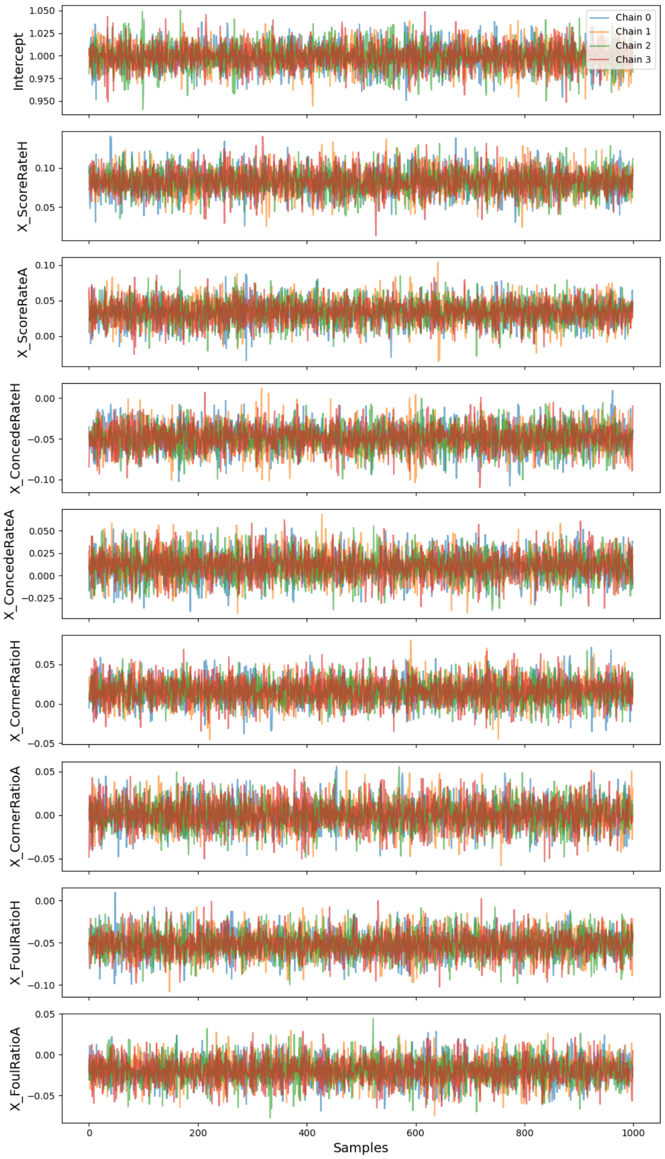

Figure 2. Trace plots for each of the regression coefficients for one of the chains

In Figure 3, we present the autocorrelation plots for all the regression coefficients (only for one chain to save space). The autocorrelations across most coefficients and lags are close to zero, indicating that the samples are really nicely uncorrelated.

In Table I, we report the ESS and R-hat statistics for all regression coefficients. The ESS values are consistently high, suggesting efficient sampling. In fact, for some coefficients, the ESS exceeds the total number of samples due to the presence of negative autocorrelation, which can reduce variance in the

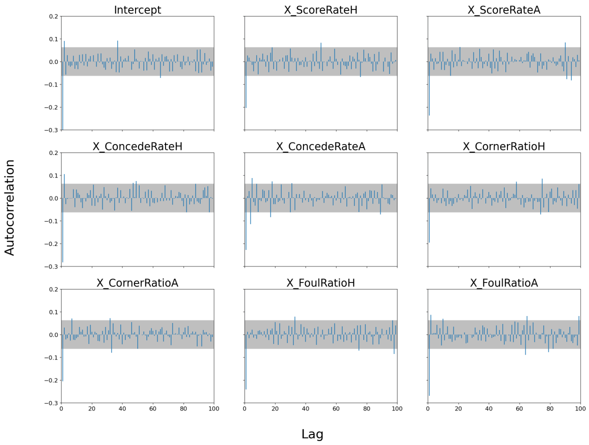

Figure 3. Autocorrelation plots for one chain

sample mean. Additionally, all R-hat values are equal to 1.00, indicating that the between-chain and within-chain variances are nearly identical, confirming good convergence of the Markov chains.

Regression coefficients	**ESS**	**R-hat**
Intercept	6551	1.00
X_ScoreRateH	5878	1.00
X_ScoreRateA	6180	1.00
X_ConcedeRateH	6050	1.00
X_ConcedeRateA	6963	1.00
X_CornerRatioH	6529	1.00
X_CornerRatioA	6416	1.00
X_FoulRatioH	6123	1.00
X_FoulRatioA	6782	1.00

Table I
ESS and R-hat diagnostics for regression coefficients

II. Laplace approximation

A. Implementation

To implement the Laplace approximation, we first derive the negative log-posterior, its gradient, and Hessian with respect to the regression coefficients $\boldsymbol{\beta}$. Note that since we are using a minimizer, we provide their negative counterparts.

Negative log-posterior:

$$-\log p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}) = -\left(\mathbf{y}^\top \mathbf{X}\boldsymbol{\beta} - \mathbf{1}^\top \exp(\mathbf{X}\boldsymbol{\beta}) - \frac{1}{2\alpha}\boldsymbol{\beta}^\top \boldsymbol{\beta}\right) \quad (1)$$

where α is the prior variance parameter for a Gaussian prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$.

The gradient of the negative log-posterior is

$$\nabla_{\boldsymbol{\beta}} \left(-\log p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y})\right) = -\left(\mathbf{X}^\top \mathbf{y} - \mathbf{X}^\top \exp(\mathbf{X}\boldsymbol{\beta}) - \frac{1}{\alpha}\boldsymbol{\beta}\right) \quad (2)$$

The Hessian matrix is given by

$$\mathbf{H} = \mathbf{X}^\top \operatorname{diag}\left(\exp(\mathbf{X}\boldsymbol{\beta})\right) \mathbf{X} + \frac{1}{\alpha}\mathbf{I}$$

These expressions are used to find the mode of the posterior and approximate it with a Gaussian distribution centered at the mode with covariance $\mathbf{H}^{-1}$.

B. Comparison with MCMC

To compare the Laplace approximation with the MCMC results, we plot the marginal distributions obtained from the approximation in Figure 4. These densities closely resemble those from the MCMC sampling, demonstrating that the Laplace method captures the posterior distributions well. However, unlike the MCMC samples, the approximated distributions are obviously exact Normal distributions.

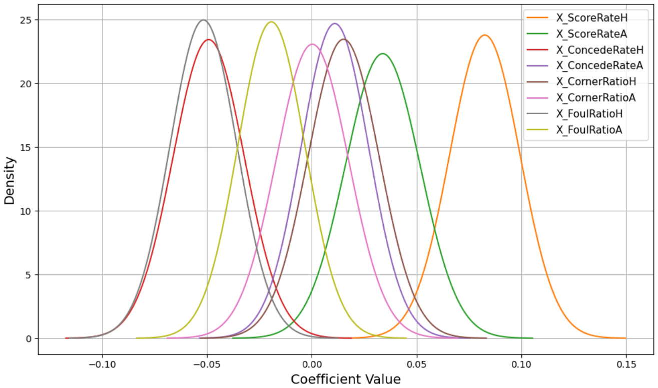

Figure 4.

C. Predictions for test cases

Lastly, we use our model to make predictions on the test dataset. To obtain the predictive distributions, we first draw 1000 samples from the multivariate normal distribution defined by the estimated regression coefficients' means and the covariance matrix (the inverse of the Hessian at the calculated means of β-s). For each sample of coefficients, we multiply by the feature vectors and apply the Poisson GLM's link function (exponential) to generate the corresponding predicted counts.

For point predictions, we consider three loss functions and select the summary statistic of the predictive distribution accordingly:

- **Squared error loss:** use the mean
- **Absolute error loss:** use the median
- **Accuracy:** use the mode

Since we do not posses the actual target variables for test cases, we don't compute the metrics.

III. Appendix