# Out-Sample Performance Check

## WC 2022-Group stages

In this section, we considered to evaluate the predictive performance of some proposed models focused on the goals scored by both competing teams during the group stage period of the tournament. The models which were compared among each other:

1. Diagonal Inflated Bivariate Poisson
2. Bivariate Poisson
3. Double Poisson

For the fitting of all these models, we used the same data for their training as well as the same priors apart from both formulations of bivariate Poisson (models 1 and 2) where we had the additional parameters of π, ρ for model 1 and the one of parameter ρ for model 2, respectively. The dataset for the performance evaluation consists of 48 matches.

**Results**

First, we calculated the Bayesian Information Criterion of LOOIC (Vehtari et al., 2017) across Models 1, 2, 3 in every matchday of the group-stages. Briefly, we report that the LOOIC of the Model 1 was the worst one in every match day while the best one was the third one but with much slight difference compared to the model 2.

As a next step, we considered to compare our fitted models’ predictive performance, using the posterior predictive density of goal difference, with the observed goal differences. More specifically, we proceeded to the comparison of our posterior predictive distributions (of fitted models) with the observed goal differences graphically through the posterior predictive checking (PPC) bar plots. These plots are useful since we can depict how close are the 95% posterior intervals of the predicted frequencies of goal differences to the corresponding observed ones. Also, the plots will include the metric of “Mean Absolute Error” calculated within Bayesian framework to quantify how close are our model-based predictions with the observed results across the group stage matches. Based on Figure 1, we observe that the Model 1 (diag. infl. Biv. Poisson) has almost identical error with the simple Bivariate Poisson in favor of the diagonal inflated version. The similarity between predictive performances of both Models 1 and 2 is obvious from the corresponding PPC graphs where in both Models 1 and 2 the posterior medians (dark points) are close to the observed frequencies. An obvious difference between Models 1 and 2 is that the former one’s 95% posterior intervals “capture” better the observed goal difference value of -2.

Timeline

Description automatically generated

**Figure 1:** 95% Posterior intervals of predicted frequencies of goal differences; yrep and y are the generated (median) and observed quantities, respectively. The MAE metric is the posterior mean of the MAE; within brackets the posterior standard deviation of this metric.

In general, the results from Figure 1 indicate that any model among them provides similar predictions with a little advantage in favor of diag. infl. Bivariate Poisson. Thus, our predictions for the next phase of 16 in World Cup will be again based on this model formulation.

Nevertheless, an external reader can obtain similar results in terms of error metrics by fitting any choice among those proposed ones to predict the goals scored by two competing teams as well as the goal difference in each match. Another advantage in favor of Models 2, 3 is that they have simpler structure and less parameters during their fitting. Also, those models are better in terms of LOO.

By the end of tournament, as we will obtain more and more matches to our sample, we are going to proceed a more detailed evaluation of the predictive performance of fitted models by using additional error metrics such as log-loss metrics, another classification metrics, etc… to understand at great extent which model would perform better to such tournament based on the training data that we use for.

# Bibliography

Grandini, M., Bagli, E., & Visani, G. (2020). Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*.

VEHTARI, A., GELMAN, A., & GABRY, J. (2017) Practical Bayesian model evaluation using leave one-out cross-validation and WAIC, Statistics and Computing, 27, 1413–1432.