Improved Bayesian Hierarchical Model: Predicting Scoring Outcomes in EPL Soccer

Ketkar, Akhil – akhilketkar@g.harvard.edu
Prunskis, Owen – prunskis@college.harvard.edu

Advanced Scientific Computing: Stochastic Optimization Methods
Spring 2015

Introduction

Soccer is the most popular sport in the world, particularly in Europe. The plethora of in-depth records makes the sport a natural target for statisticians and mathematicians to flex their intellectual muscle in predicting match outcomes, scoring differentials, season champions, and far more esoteric metrics.

Baio-Blangiardo proposes a Bayesian hierarchical model for predicting the number of goals scored per game drawing from a bivariate Poisson distribution, which considers attack and defense strength and home field effects. The Weitzenfield expansion notably adds an intercept term to capture average goals score by the away team. We implement and improve upon these models.

Abstract: We expand on the the Baio-Blangiardo hierarchical Bayesian soccer model to predict match outcomes, scores, and season standings. The model incorporates performance-based stratifications for mean goal scoring, considers home-field advantage effects for each team, and includes game importance and schedule density parameters. We apply our model to English Premier League data, which yields improved results.

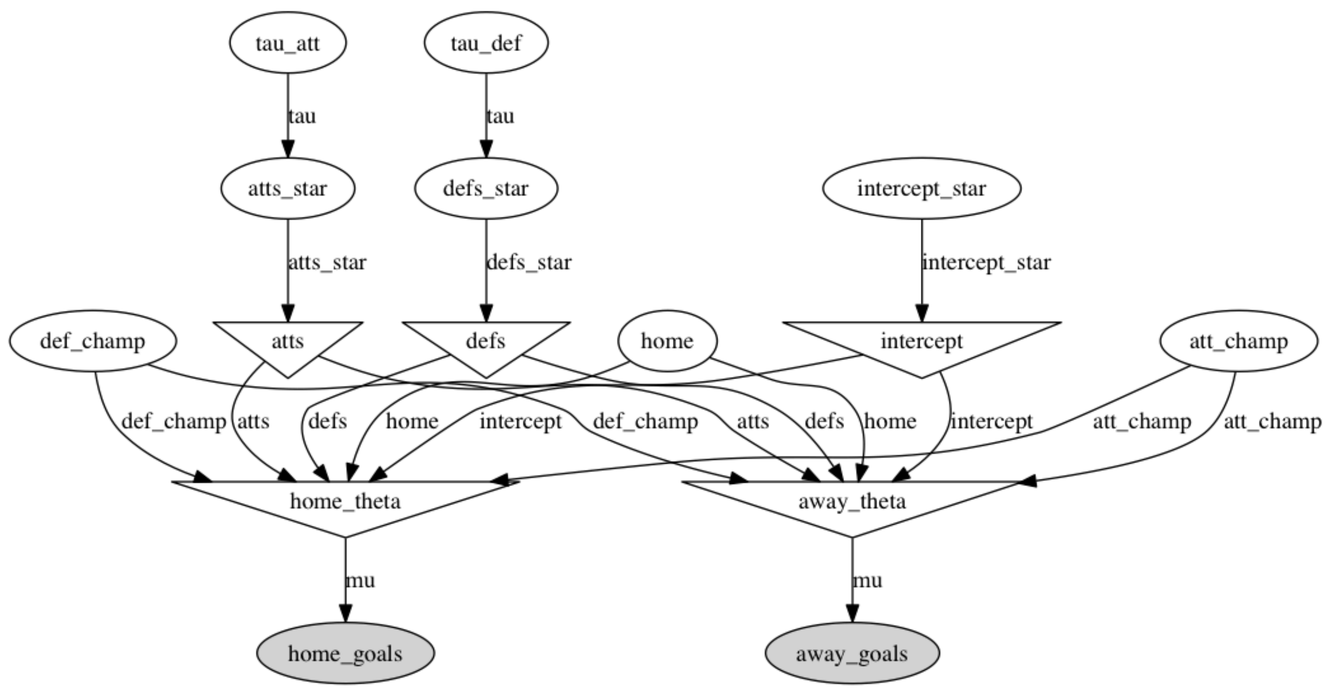


Fig 1. DAG representation of the hierarchical model

Approach

Building from the Weitzenfield expansion of the Baio-Blangiardo Bayesian hierarchical model, we add the following considerations:

- Stratification of teams by ability for separate intercept parameters,
- team-specific home field effects,
- game importance, defined by risk of relegation or possibility of winning championship, and
- schedule density, considering non-league games (not shown in Fig.1)

Results

Our model represents a signifcant improvement to both the Baio-Blangiardo and Weitzenfield improvements. Figure 2 shows the comparison between observed cumulative goals (vertical axis) and our model predicted cumulative goal totals (horizontal axis). We include the red 45-degree line to assist in visualization. Our result represents a significant improvement to the base model

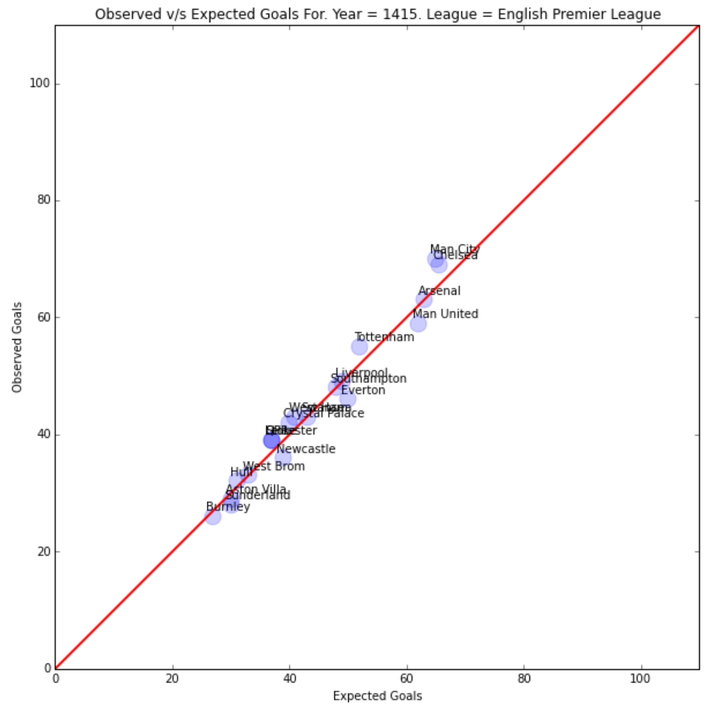
Conclusions

We propose and successfully implement several improvements to the Baio-Blangiardo Bayesian hierarchical model. Given the success of the modesl in predicting stochastic outcomes in soccer, we apply the model to data from both the National Basketball Association and Major League Baseball and find that our model is fairly robust across athletic disciplines, the latter of which performs slightly better than the former due to similarities in EPL and MLB scoring distributions.

Data

The original paper by Baio and Blangiardo uses Italian Serie A match data for the 2007-2008 season.

Our primary model pulls English Premier League match data for the 2013-2014 and compare our simulated 2014-2015 season to actual results. In additional expansions, we consider NBA basketball and MLB baseball data from the previous 2 seasons to assess our model's cross-disciplinary application.



Citations

Baio, G. and Blangiardo, M.A. 2010. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistic,* vol. *37.2;* 253-264.

Dixon, J. D. and Coles, S.G. 1997. Modelling Association Football Scores and Inefficiencies in the Football Betting Market. *Applied Statistics,* vol. 46.2; 265-280.

Weitzenfeld, D. 2014. A hierarchical bayesian model of the premier league. *Pass the ROC*.