Modeling Soccer Scores by Integrated Nested Laplace Approximation

RSCAM Group Project

SHEN Dongrui

Joint work with LEROY Alix, JIANG Chris and XIAO Shawn

University of Edinburgh

September 5, 2022

Score Prediction and Bayesian Inference

Figure: There is a growing interest in soccer score predictions due to the availability of a large amount of data, the development of betting internet sites and the large media coverage of sporting events. [1]

In this study, we explore a Bayesian regression model building on the INLA framework, for the number of goals scored by the two teams in each match.

[1] How much money is being bet on sports every year? April 3, 2019. https://afootballreport.com

Outline

Introduction to INLA

How to model outcomes?

Experiments and Results

Conclusion

Section 1

Introduction to INLA

Bayesian Inference

- In a Bayesian model, we generally want posterior distributions for our models.

Definition (The posterior distribution)

The posterior distribution is equal to the data likelihood multiplied by the priors over the normalizing constant (so the posterior integrates to one).

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)\mathrm{d}\theta} \propto p(y|\theta)p(\theta), \tag{1}$$

In a Bayesian analysis we obtain a posterior distribution for the parameter, for which we can provide summary statistics (median, median, or mode) and quantiles to directly obtain credible intervals.

Latent Gaussian Models

- Let θ be the vector of hyperparameters and x be the vector of latent effects. Generally we can partition $\theta = (\theta^{(1)}, \theta^{(2)})$.

Definition (The general form for an LGM)

$$y|x, \theta^{(2)} \sim \prod_i p(y_i|\eta_i, \theta^{(2)}) \qquad \text{(Likelihood)}$$

$$x|\theta^{(1)} \sim p(x|\theta^{(1)}) = \mathcal{N}(0, \Sigma) \qquad \text{(Latent Field)}$$

$$\theta = [(\theta^{(1)}, \theta^{(2)})]^T \sim p(\theta) \qquad \text{(Hyperpriors)}$$

where y is an observed data set, x are not covariates but rather is the joint distribution of all parameters in the linear predictor (including itself), and θ are the hyperparameters of the latent field that are not Gaussian.

Latent Gaussian Models Cont.

- Let $\mu_i = \mathbb{E}(y_i|x, \theta)$ be the mean of the observation i given the model parameters, for $1 \leq i \leq n$. This mean is connected to another random variable η_i called the linear predictor of observation i by an invertible link function g, i.e.

$$\eta_i = g\left(\mu_i\right), \quad \eta_i = \beta_0 + \sum_{j=1}^{n_\beta} \beta_j z_{ji} + \sum_{k=1}^{n_f} f^{(k)}\left(u_{ki}\right) + \epsilon_i, \quad (2)$$

  where ϵ_i are Gaussian error terms, $\boldsymbol{\beta}$ is the regression coefficient, $\boldsymbol{f}$ is a set of random functions, $\boldsymbol{z}$ and $\boldsymbol{u}$ are some covariates.
- These parts form the vector of latent effects $\boldsymbol{x} = (\eta, \boldsymbol{\beta}, \boldsymbol{f})$.
- By varying the form of the functions $\boldsymbol{f}(k)$, this framework can accommodate a wide range of models such as standard regression models, hierarchical models and etc.

Laplace Approximation

Definition (Main Idea)

For some unnormalized probability density $q(x)$, we approximate log q(x) by a quadratic function centered at the mode $\hat{x}$.
By Taylor series expansion of order 2 around the mode $\hat{x}$, as $(logq)'(\hat{x}) = 0$ we have

$$\begin{aligned} \log q(x) &\approx \log q(\hat{x}) + \frac{1}{2}(\log q)''(\hat{x})(x - \hat{x})^2 \\ &\approx \log q(\hat{x}) - \frac{1}{2\hat{\sigma}^2}(x - \hat{x})^2 \end{aligned} \tag{3}$$

by setting $\hat{\sigma}^2 := -((\log q)''(\hat{x}))^{-1}$. Hence $q \approx N(\hat{x}, \hat{\sigma}^2)$.

- Laplace approximation is also applicable in higher dimensions, with $\hat{\sigma}^2$ replaced by the covariance matrix $\hat{\Sigma} = -\left(\nabla^2(\log q)(\hat{x})\right)^{-1}$, and $\int_x q(x)\mathrm{d}x \approx q(\hat{x})(2\pi)^{d/2}(\det \hat{\Sigma})^{1/2}$.

Integrated Nested Laplace Approximation(INLA)

The general idea of INLA is to use the approximation

$$p(z|w) \approx \frac{p(x, z|w)}{\tilde{p}(x|z, w)}, \tag{4}$$

where $\tilde{p}(x|z, w)$ is the Laplace approximation to the conditional density $p(x|z, w)$.

Brief Summary of INLA

- The basic idea of INLA is the application of Laplace approximation repeatedly in a nested manner, i.e. not directly on the posterior, but on some of the conditional distributions that arise in calculations.
- Take advantage of the structure of the Latent Gaussian Model (such as the sparsity of the GMRF prior) to speed up calculations.
- Use numerical integration over the hyperparameter space. The grid can be further refined if the precision is not yet sufficient (by increasing the number of gridpoints).

Section 2

How to model outcomes?

Main Ideas

How to model outcomes?[2]

(a) Model outcomes (win, draw, lose)

(b) Model the number of goals scored

Bradley Terry Models vs Poisson Log-linear Model

[2]Alkeos Tsokos et al., 2019. Modeling outcomes of soccer matches. Machine learning, 108(1), pp.77–95. https://doi.org/

Bradley-Terry Model

Bradley-Terry models could be considered as the extension of linear regression models.

Definition (Bradley–Terry)

The extended Bradley–Terry model assumes that

$$p(Y_{ijt} = y) = f_y(\lambda_{it}, \lambda_{jt}), \tag{5}$$

where $y = 0, 1, 2$ represent lose, draw and win, the predictor λ_{it} is the "strength" of team i at match t, and we use $f_y(\cdot)$ to approximate the probabilities of getting outcome y.

- Two different way to define $f_y(\cdot, \cdot)$: Ordinal, Davidson
- Three different way to define "strength": BL, CS, LF
- Use MLE

Bradley-Terry Model Cont.

BL: baseline

The simplest specification of all only considers the home-field advantage:

$$\lambda_{it} = \beta h_{it},$$

where $h_{it} = 1$ if team i play at home at time t, and $h_{it} = 0$ otherwise.

CS: constant strength

This specification takes account of the constant strength:

$$\lambda_{it} = \alpha_i + \beta h_{it},$$

where α_i represents the time-invariant strength of the ith team.

LF: linear with features

This specification assumes the team strengths are a linear combination of match- and team-specific features:

$$\lambda_{it} = \sum_{k=1}^{p} \beta_k \mathrm{z}_{itk},$$

where z_{itk} is the kth element of the feature vector z_{it}.

Poisson Log-linear Model

Poisson Log-linear Model fits in the LGM framework.

Definition (Poisson Log-linear Model)

The Poisson log-linear model assumes the Poisson likelihood,

$$y_{gj}|\lambda_{gj} \sim \text{Poisson}(\lambda_{gj}). \tag{6}$$

$y_{gj}(j = 1, 2)$ is the observed goal counts in the match g for the home or away team. The parameters λ_{gj} represents the corresponding scoring intensity.

- These rates λ_{gj} are linked to some linear predictors η_{gj} by the log link function: $\eta_{gj} = \log \lambda_{gj}$
- Two ways to define the linear predictor: PL baseline, PL features
- Use INLA

Poisson Log-linear Model Cont.

PL baseline
A basic model that only considers team strengths and the home-field advantage:

$$\eta_{gj} = \log \lambda_{gj} = \beta h_{gj} + \alpha_{h_g} + \xi_{a_g}.$$

PL features
This specification takes account of more features:

$$\eta_{gj} = \log \lambda_{gj} = \sum_{k=1}^{p} \beta_k z_{gjk} + \alpha_{h_g} + \xi_{a_g},$$

where z_{gjk} is the kth element of the feature vector z_{gj}. α_t and ξ_t denote the attacking strength and defense strength of team t.

Section 3

Experiments and Results

Experiments Workflow

The main purpose is to compare the two distinct modeling approaches:

- Bradley-Terry Model by MLE
- Poisson Log-linear Model by INLA

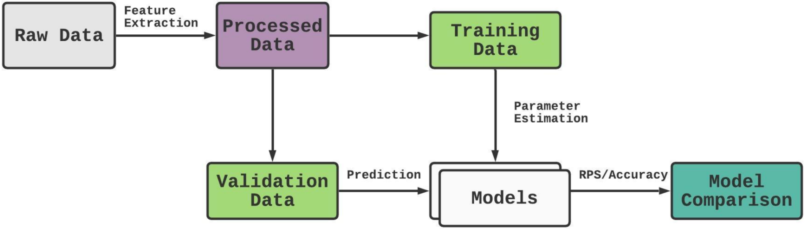

Figure: The flowchart of the experiments for model comparison.

Data Preparation

The following experiments will focus on the data for the 2016-17 Bundesliga.[3]
There are 18 teams in the Bundesliga. Every team in the league plays twice against each other, so in total, there are 306 games and 34 rounds per season.

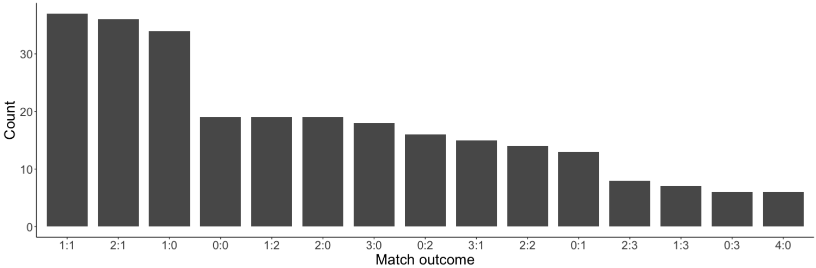

Figure: The 10 most frequent match scorelines in the 2016-17 Bundesliga.

[3]The data set used in the project is obtained from the Open International Soccer Database, 2017. https://osf.io/kqcye/

Data Preparation Cont.

Table: Example of Feature values: the first 3 matches of 16-17 season for Bayern Munich.

	Match 1	**Match 2**	**Match 3**
League	Bundesliga	Bundesliga	Bundesliga
Date	2016-08-26	2016-09-09	2016-09-17
Round	1	2	3
Attack	Bayern Munich	Bayern Munich	Bayern Munich
Defense	Werder Bremen	Schalke 04	Ingolstadt
WDL	W	W	W
Goals	6	2	3
Goals conceded	0	0	1
Home	1	0	1
Lag time	-	14	8
Matches played	0	1	2
Form	-	-	1.000
Points tally	0	3	6
Goal difference	0	6	8
Newly promotes	0	0	0
Previous round rankings	-	1	1
Season	16-17	16-17	16-17
Season window	Aug-May	Aug-May	Aug-May
Quarter	3	3	3

Validation Framework

Temporal Validation Scheme

- Assessed the performance of a predictive model using data collected after the model was developed.

Validation Criteria

- Accuracy

$$\text{Accuracy} = \begin{cases} 1, & \text{if predicted outcome} = \text{actual outcome}, \\ 0, & \text{if predicted outcome} \neq \text{actual outcome}. \end{cases} \tag{7}$$

  Accuracy is not a great metric in scenarios with 3 or more possible outcomes.

- RPS: Ranked probability score

$$\text{RPS} = \frac{1}{r-1} \sum_{i=1}^{r-1} \left\{ \sum_{j=1}^{i} (p_j - a_j) \right\}^2, \tag{8}$$

  where $r = 3$ is the number of possible outcomes, $(p_j)_{j=1}^{r}$ is the r-vector of predicted probabilities. $(a_j)_{j=1}^{r}$ is the actual probabilities.

Predictive performance comparison

Model	Draws	Ranked Probability Score	Accuracy	Running Time
BL	Davidson	0.2279436	0.4477124	1.327
BL	Ordinal	0.2279438	0.4477124	2.009
CS	Davidson	0.2279427	0.4832215	3.944
CS	Ordinal	0.2279437	0.4832215	4.347
LF	Davidson	0.2234490	0.4941634	7.299
LF	Ordinal	0.2235946	0.4941634	9.211
PL baseline		0.2150916	0.4863813	11.634
PL features		0.2145319	0.4785992	11.63362

Table: The ranked probability score and classification accuracy for the models, as estimated from the temporal validation framework. The model with low RPS and high accuracy is considered as a good model.

Section 4

Conclusion

Conclusion

- Introduction of INLA, which allows for computationally efficient Bayesian inference for a large class of LGMs.
- The INLA framework gives promising results for the problem of modeling outcomes of soccer matches.
- Further models with INLA could be used, using different features and test for more accurate predictions.
- A challenging aspect of modeling soccer outcomes is devising ways to consider which season matches belong to but also to borrow information across different seasons. Further improvements can be achieved by using a hierarchical model.

Reference

[1] *How much money is being bet on sports every year?*. `https://afootballreport.com/`, 2019.

[2] Alkeos Tsokos et al. *Modeling outcomes of soccer matches. Machine Learning*, 108(1): 77–95, `https://doi.org/`, 2019.

[3] Werner Dubitzky et al. *The open international soccer database for machine learning*. *Machine Learning*, 108(1): 9–28, `https://osf.io/kqcye/`, 2019.

[4] Brian J.Reich, Sujit K.Ghosh. *Bayesian Statistical Methods*. CRC Press, 2019.