Modeling Football League Results (A study on 2000-2001 Serie-A Data)

Introduction:

The surge in interest surrounding football outcome modeling lacks a comprehensive computational tool to seamlessly fit various football models. Addressing this gap, the **footBayes** package aims to provide a unified solution enabling the fitting, interpretation, and visual exploration of diverse goal-based Bayesian football models. Leveraging the Stan environment **(Stan Development Team (2020)),** this tool facilitates the seamless execution of goal-based Bayesian models within a singular framework.

I have downloaded the package from the following link:
https://github.com/LeoEgidi/footBayes

	Date	Season	home	visitor	FT	hgoal	vgoal	tier
23421	2000-10-01	2000	Udinese Calcio	Brescia Calcio	4-2	4	2	1
23422	2000-10-01	2000	AS Roma	Bologna FC	2-0	2	0	1
23423	2000-10-01	2000	AC Perugia	US Lecce	1-1	1	1	1
23424	2000-10-01	2000	Reggina Calcio	Inter	2-1	2	1	1
23425	2000-10-01	2000	SSC Napoli	Juventus	1-2	1	2	1
23426	2000-10-01	2000	AC Milan	Vicenza Calcio	2-0	2	0	1
23427	2000-10-01	2000	Atalanta	Lazio Roma	2-2	2	2	1
23428	2000-10-01	2000	AS Bari	Hellas Verona	1-1	1	1	1
23429	2000-10-01	2000	Parma AC	ACF Fiorentina	2-2	2	2	1
23430	2000-10-15	2000	Inter	SSC Napoli	3-1	3	1	1
23431	2000-10-15	2000	Vicenza Calcio	Atalanta	1-2	1	2	1
23432	2000-10-15	2000	Hellas Verona	Udinese Calcio	1-1	1	1	1
23433	2000-10-15	2000	US Lecce	AS Roma	0-4	0	4	1
23434	2000-10-15	2000	Juventus	AS Bari	2-0	2	0	1
23435	2000-10-15	2000	ACF Fiorentina	Reggina Calcio	2-1	2	1	1
23436	2000-10-15	2000	Brescia Calcio	Parma AC	0-0	0	0	1
23437	2000-10-15	2000	Bologna FC	AC Milan	2-1	2	1	1
23438	2000-10-15	2000	Lazio Roma	AC Perugia	3-0	3	0	1
23439	2000-10-22	2000	AS Roma	Vicenza Calcio	3-1	3	1	1
23440	2000-10-22	2000	SSC Napoli	Bologna FC	1-5	1	5	1
23441	2000-10-22	2000	Hellas Verona	Lazio Roma	2-0	2	0	1

Showing 23,420 to 23,441 of 27,684 entries, 8 total columns

	Season	home	visitor	hgoal	vgoal
23421	2000	Udinese Calcio	Brescia Calcio	4	2
23422	2000	AS Roma	Bologna FC	2	0
23423	2000	AC Perugia	US Lecce	1	1
23424	2000	Reggina Calcio	Inter	2	1
23425	2000	SSC Napoli	Juventus	1	2
23426	2000	AC Milan	Vicenza Calcio	2	0
23427	2000	Atalanta	Lazio Roma	2	2
23428	2000	AS Bari	Hellas Verona	1	1
23429	2000	Parma AC	ACF Fiorentina	2	2
23430	2000	Inter	SSC Napoli	3	1
23431	2000	Vicenza Calcio	Atalanta	1	2
23432	2000	Hellas Verona	Udinese Calcio	1	1
23433	2000	US Lecce	AS Roma	0	4
23434	2000	Juventus	AS Bari	2	0
23435	2000	ACF Fiorentina	Reggina Calcio	2	1
23436	2000	Brescia Calcio	Parma AC	0	0
23437	2000	Bologna FC	AC Milan	2	1
23438	2000	Lazio Roma	AC Perugia	3	0
23439	2000	AS Roma	Vicenza Calcio	3	1
23440	2000	SSC Napoli	Bologna FC	1	5
23441	2000	Hellas Verona	Lazio Roma	2	0

Showing 1 to 21 of 306 entries, 5 total columns

Meta Dataset (1934-2021 Seasons):
27684 observations of 8 variables

2000-2001 Season Data:
306 observations of 5 variables

I propose to perform the following fits and checks for analysis and posterior predictive checks:

I. Fitting static maximum likelihood and Bayesian models.
II. Changing the prior distributions and perform some sensitivity tests.
III. Fitting dynamic Bayesian models.
IV. Model Checks
V. Predictions and Predictive Accuracy
VI. Comparing models.

Distributions for Modelling

We are considering 2 distributions. One being the **Double Poisson** and the other one being **Bivariate Poisson**. One major concern with the Double Poisson model is that the goals scored during a match by two competing teams are conditionally independent. But, in team sports, such as football, handball, hockey, and basketball it is reasonable to assume that the two outcome variables are **correlated** since the **two teams interact** during the game. Lets say, in the practical football match case of the home team leading with 1-0 or 2-0, when only ten minutes are left to play. The away team can then become more determined and can take more risk in an effort to score and achieve the draw within the end of the match. Or, even when one of the two teams is leading say with 3-0, or 4-0, it is likely that team will be relaxing a bit, and the opposing team could score at least one goal quite easily. For this, goals' correlation due to a change in the performance of the team or both teams could be captured by a dependence parameter, accounting for positive correlation. Positive parametric goals' dependence is made possible by using a **bivariate Poisson distribution**.

The Joint Probability Function $\mathbf{P}_{X,Y}{}^{(X=x,Y=y)}$ =

$$\sum_{k=0}^{\min(x,y]} ({}^{x}c_k)({}^{y}c_k) * k! * (\lambda 3/(\lambda 1\lambda 2))k.$$

Marginally, each random variable follows a Poisson distribution with E(X)=λ1+λ3, E(Y)=λ2+λ3, and cov(X,Y)=λ3 ; λ3 acts as a measure of dependence between the goals scored by the two competing teams. If λ3=0, then the two variables are conditionally independent and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions, the double Poisson case.

Let (xn,yn) denote the observed number of goals scored by the home and the away team in the n-th game, respectively. A general bivariate Poisson model allowing for goals' correlation, as in **Karlis and Ntzoufras (2003)** is the following:

$$\begin{gathered} Xn,Yn[\lambda 1n,\lambda 2n,\lambda 3n]=BivPoisson(\lambda 1n,\lambda 2n,\lambda 3n) \\ Log(\lambda 1n] = \mu+home+atthn+defan \\ Log(\lambda 2n] = \mu+attan+defhn \\ Log(\lambda 3n] = \beta 0+\gamma 1\beta homehn+\gamma 2\beta awayan+\gamma 3\boldsymbol{\beta} wn , \end{gathered}$$

where λ1n,λ2n represent the **scoring rates** for the home and the away team, respectively; μ represents the **constant intercept**; home represents the **home-effect**, i.e. the well-known advantage of the team hosting the game; attt and deft represent the **attack** and the **defence** abilities, respectively, for each team t, t=1,…,T, the nested indexes hn ,an=1,…,T denote the home and the away team playing in the n-th game, respectively; β0 is a constant parameter; βhomehn and βawayan are parameters that depend on the home and away team respectively, wn is a vector of covariates for the n-th match used to model the covariance term and ***β*** is the corresponding vector of regression coefficients. The parameters γ1,γ2, and γ3 are dummy binary indicators taking values 0 or 1 which may activate distinct sources of the linear predictor. Hence when γ1=γ2=γ3=0, we consider constant covariance as in **Egidi and Torelli (2020)** , whereas when (γ1,γ2,γ3)=(1,1,0) we assume that the covariance depends on the teams' parameters only but not on further match covariates, and so on.

The case λ3n=0 (the scores' correlation parameter equals zero) reduces to the double Poisson model, as in **Baio and Blangiardo (2010)** .

To achieve model's identifiability, attack/defence parameters are imposed a **sum-to-zero** constraint:

$$\sum t=1 \text{ to } T \ attt=0 \ , \sum t=1 \text{ to } T \ deft=0$$

Another identifiability constraint, largely proposed in the football literature, is the **corner**-constraint, which assumes the abilities for the T-th team are equal to the negative sum of the others, and then achieves a sum-to-zero as well:

$att_T = -(\sum_{t=1}^{T-1} att_t)$, $def_T = -(\sum_{t=1}^{T-1} def_t)$

Fitting Static Maximum Likelihood and Bayesian Models

To fit and interpret the models, and we'll mainly focus on the bivariate Poisson case.
Classical estimates for BP models are provided, among the others, by **Karlis and Ntzoufras (2003)** (MLE through an EM algorithm) and **Koopman and Lit (2015).**

Likelihood approach:

Given the parameter-vector $\boldsymbol{\vartheta}=(\{att_t, def_t, t=1,\ldots,T\}, \mu, home, \beta_{home_{h_n}}, \beta_{away_{a_n}}, \beta_0, \boldsymbol{\theta})$, the likelihood function of the bivariate Poisson model above takes the following form:

$$L(\boldsymbol{\vartheta})=\prod_{n=1}^{N} \exp\{-(\lambda_{1n}+\lambda_{2n}+\lambda_{3n})\}*((\lambda_{1n}^{x_n}/x_n!)*(\lambda_{2n}^{y_n}/y_n!) \times \sum_{k=0}^{\min(x_n,y_n)} (x_n \text{ C } k)(y_n \text{ C } k)*k!*(\lambda_{3n}/\lambda_{1n}\lambda_{2n})^k.$$

Maximum-likelihood parameters estimation can be performed by searching the MLE $\boldsymbol{\vartheta}$ such that:

$$\boldsymbol{\vartheta} = \operatorname{argmax} L(\boldsymbol{\vartheta}), \text{ where } \theta \in \Theta$$

by imposing the following system of partial (log)-likelihood equations:

$$l'(\boldsymbol{\vartheta})=0$$

Wald and deviance-confidence intervals may be constructed for the MLE $\boldsymbol{\vartheta}$. A 95% Wald-type interval satisfies:

$$\boldsymbol{\vartheta} \pm 1.96 \, se(\boldsymbol{\vartheta})$$

As we'll see, the footBayes package allows the MLE computational approach (along with Wald-type and profile-likelihood confidence intervals) for static models only, i.e. when the model complexity is considered acceptable. As the parameters' space grows—as it commonly happens when adding dynamic patterns—MLE becomes computationally expensive and less reliable.

Bayesian approach:

The goal of the Bayesian analysis is to carry out inferential conclusions from the joint posterior distribution $\pi(\boldsymbol{\vartheta}|⍰)$, where $⍰=(x_n,y_n)$ $n=1,\ldots,N$ denotes the set of observed data for the N matches. The joint posterior satisfies:

$$\pi(\boldsymbol{\vartheta}|⍰)= p(⍰|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}) / (p(⍰)) \propto p(⍰|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta}),$$

where $p(⍰|\boldsymbol{\vartheta})$ is the model sampling distribution (proportional to the likelihood function), $\pi(\boldsymbol{\vartheta})$ is the joint prior distribution for $\boldsymbol{\vartheta}$, and $p(⍰) = \int_\Theta p(⍰|\boldsymbol{\vartheta})\pi(\boldsymbol{\vartheta})d\theta$ is the marginal likelihood that does not depend on θ.

In the majority of the cases, $\pi(\boldsymbol{\vartheta}|⍰)$ does not have a closed form and for such reason we need to approximate it by simulation. The most popular class of algorithms designed to achieve this task is named **Markov Chain Monte Carlo Methods** (see **Robert and Casella (2013)** for a deep theoretical overview). These methods allow to sample weak correlated samples from some Markov chains whose stationary and limiting distribution coincide with the posterior distribution that we wish to approximate and sample from.

The footBayes package relies on a sophisticated MCMC enginery, namely the **Hamiltonian Monte Carlo** performed by the Stan software: the HMC borrow its name from the Hamiltonian dynamics of physics and is aimed at suppressing random-walk and wasteful behaviours in the exploration of the posterior distribution which typically arise when using the Gibbs sampling and the Metropolis-Hastings algorithm. For a deep and great summary about HMC, you may read the paper **Betancourt (2017).**

In terms of inferential conclusions, we are usually interested in summaries from the marginal posterior distributions of the single parameters: posterior means, medians, credibility intervals, etc.. We can write out the formula for the posterior distribution of the bivariate Poisson model above as:

$$\pi(\boldsymbol{\vartheta}|\square) \propto \pi(\boldsymbol{\vartheta})* \prod_{n=1}^{N} \text{BivPoisson}(\lambda 1n,\lambda 2n,\lambda 3n)$$

where $\pi(\boldsymbol{\vartheta})=\pi(att)\pi(def)\pi(\mu)\pi(home)\pi(\beta homehn)\pi(\beta awayan)\pi(\beta 0)\pi(\boldsymbol{\beta})$ is the joint prior distribution under the assumption of a-prior independent parameters' components.

The standard approach is to assign some **weakly-informative** prior distributions to the team-specific abilities. These parameters are considered **exchangeable** from two common (prior) distributions:

$$attt \sim \square(\mu att,\sigma att)$$
$$deft \sim \square(\mu def,\sigma def),\ \ t= 1,\ldots,T$$

with **hyperparameters** $\mu att,\sigma att,\mu def,\sigma def$. The model formulation is completed by assigning some weakly-informative priors to the remaining parameters. In what follows, some priors' options will be handled directly by the user.

STATIC FIT:

We are currently ignoring any time-dependence in our parameters, considering them to be **static** across distinct match-times. Initially, we use a static Bivariate-Poisson distribution along with Monte-Carlo Sampling. A positive influence was observed for the home, sigma_att, sigma_def.

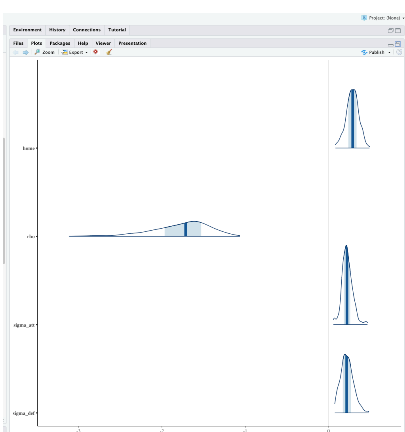


Fig.1_ Marginal Posterior with Bayesplot

The Likelihood Approach: the mle_foot function returns the MLE estimates along with 95% profile-likelihood deviance confidence intervals (by default) and Wald-type confidence

The Bayesian approach: the stan_foot function produces an Hamiltonian Monte Carlo posterior sampling by using the underlying rstan ecosystem. The user can choose the number of iterations (iter), the number of Markov chains (chains), and other optional arguments values.

As we could expect, there is a positive effect from the home-effect (posterior mean about 0.3), and this implies that if two teams are equally good (meaning that their attack and defence abilities almost coincide), assuming that the constant intercept $\mu \approx 0$, we get that the average number of goals for the home-team will be $\lambda 1=\exp\{0.3\} \approx 1.35$, against $\lambda 2=\exp\{0\}=1$

We fit the same model under the MLE approach with Wald-type confidence intervals. We can then print the MLE estimates.

```
> fit1_mle$home
    2.5% mle 97.5%
[1,]  0.2 0.3  0.39
```

We got a very similar estimate to the Bayesian model for the home-effect.

MODEL CHECKS

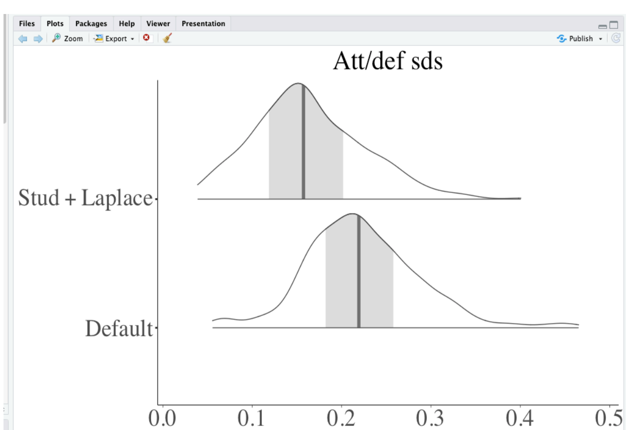
The default priors for the team-specific abilities and their related team-level standard deviations are:

$$\begin{aligned} att_t &\sim \square(\mu_{att}, \sigma_{att}), \\ def_t &\sim \square(\mu_{def}, \sigma_{def}), \\ \sigma_{att}, \sigma_{def} &\sim \text{Cauchy}^+(0, 5), \end{aligned}$$

where Cauchy$^+$ denotes the half-Cauchy distribution with support $[0,+\infty)$. However, we are free to elicit some different priors, possibly choosing one among the following distributions: Gaussian (normal), student-t (student_t), Cauchy (cauchy) and Laplace (laplace). The prior optional argument allows to specify the priors for the team-specific parameters att and def, whereas the optional argument prior_sd allows to assign a prior to the group-level standard deviations $\sigma att, \sigma def$. For instance, for each team t, $t=1,\ldots,T$, we could consider:

$$\begin{aligned} att_t &\sim t(4, \mu_{att}, \sigma_{att}), \\ def_t &\sim t(4, \mu_{def}, \sigma_{def}), \\ \sigma_{att}, \sigma_{def} &\sim \text{Laplace}^+(0, 1), \end{aligned}$$

where $t(df,\mu,\sigma)$ denotes a student-t distribution with df degrees of freedom, location μ and scale σ, whereas Laplace$^+$ denotes a half-Laplace distribution.



The student+laplace prior induces a lower amount of group-variability in the σatt marginal posterior distribution (then, a larger shrinkage towards the grand mean μatt.

Fig.2_ Comparing posteriors of Student-t + Laplace and Default Cauchy

Fitting Dynamic Bayesian Models

Teams' performance tend to be *dynamic* and change across different years, if not different weeks. Many factors contribute to this football aspect:

1. Teams act during the summer/winter players' transfermarket, by dramatically changing their rosters.
2. Some teams' players could be injured in some periods, by affecting the global quality of the team in some matches.
3. Coaches could be dismissed from their teams due to some non satisfactory results.
4. Some teams could improve/worsen their attitudes due to the so-called turnover.

We use the dynamic_type argument in the stan_foot function, with possible options 'seasonal' or 'weekly' in order to consider more seasons or more week-times within a single season, respectively. Let's fit a weekly-dynamic parameters model on the Serie A 2000/2001 season.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
home	0.26	0.00	0.07	0.13	0.22	0.27	0.31	0.39	262	1.02
rho	-1.74	0.02	0.34	-2.48	-1.93	-1.69	-1.51	-1.21	378	1.00
sigma_att	0.06	0.01	0.01	0.04	0.05	0.06	0.07	0.09	4	1.90
sigma_def	0.07	0.01	0.02	0.04	0.05	0.06	0.09	0.11	4	2.29

From the printed summary, we may note that the degree of goals' correlation seems to be again very small here. Moreover, the Gelman-Rubin statistic for σatt is relatively high, whereas the effective sample sizes for σatt and σdef are quite low. This is suggesting possible inefficiencies during the HMC sampling and that a model-reparameterization could be suited and effective at this stage. Another option is to play a bit with the prior specification for σatt and σdef, for instance by specifying a prior inducing less shrinkage in the team-specific abilities. To deal with these issues and broaden the set of candidate models, let's fit also a dynamic double-Poisson model with the `double_pois` option for the argument `model`:

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
home	0.39	0.00	0.04	0.31	0.36	0.39	0.42	0.47	148	1.03
sigma_att	0.05	0.01	0.01	0.04	0.04	0.05	0.06	0.09	4	1.92
sigma_def	0.06	0.01	0.01	0.04	0.05	0.06	0.07	0.09	3	1.83

The fitting problems mentioned above remain also for the double Poisson mode. Thus, it's time to play a little bit with the prior distributions. Also, in the dynamic approach we can change the default priors for the team-specific abilities and their standard deviations, respectively, through the optional arguments `prior` and `prior_sd`. The specification follows almost analogously the static case: with the first argument we may specify the prior's family for the team-specific abilities and the specific priors for attt,1, deft,1 along with the hyper-prior location μatt, μdef
whereas σatt and σdef need to be assigned some proper prior distribution. Assume to fit the same double Poisson model, but here we suppose student-t distributed team-specific abilities with 4 degrees of freedom to eventually capture more extreme team-specific abilities (more variability, i.e. less shrinkage), along with a Cauchy+(0,25) for their standard deviations (to better capture a possible larger evolution variability):

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
home	0.41	0.00	0.05	0.31	0.38	0.41	0.44	0.49	203	1.03
sigma_att	0.05	0.01	0.02	0.03	0.04	0.04	0.05	0.10	4	2.63
sigma_def	0.06	0.01	0.01	0.04	0.05	0.06	0.07	0.09	3	2.24

By doing so, it was observed that
The situation has been only slightly improved.

Predictive Intervals for team-specific Football Abilities:

The function foot_abilities allows to depict posterior/confidence intervals for global attack and defense abilities on the considered data (attack abilities are plotted in red, whereas defense abilities in blue colors). The higher the attack and the lower the defence for a given team, and the better is the overall team's strength.

AS Roma, the team winning the Serie A 2000/2001, is associated with the highest attack ability and the lowest defence ability according to both the models. In general, the models seem to well capture the static abilities: AS Roma, Lazio Roma and Juventus (1st, 3rd and 2nd at the end of that season, respectively) are rated as the best teams in terms of their abilities, whereas AS Bari, SSC Napoli and Vicenza Calcio (all relegated at the end of the season) have the worst abilities.

We can also depict the team-specific dynamic plots for the dynamic models:

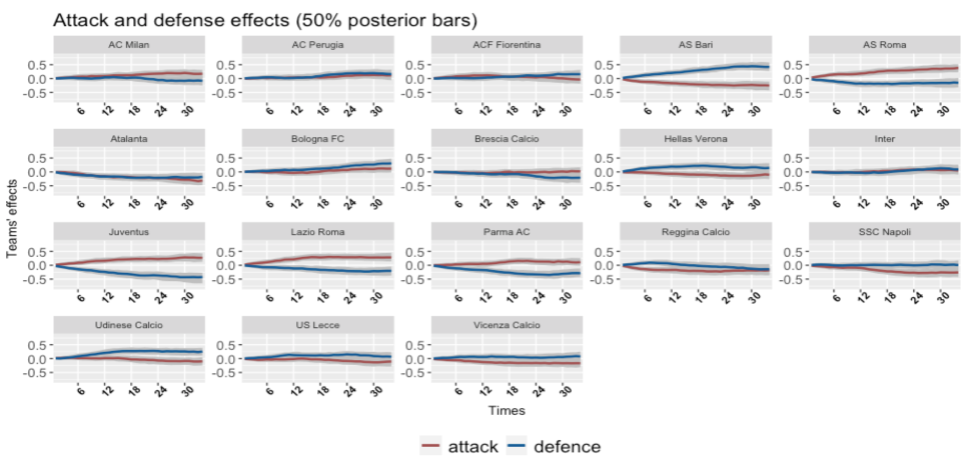
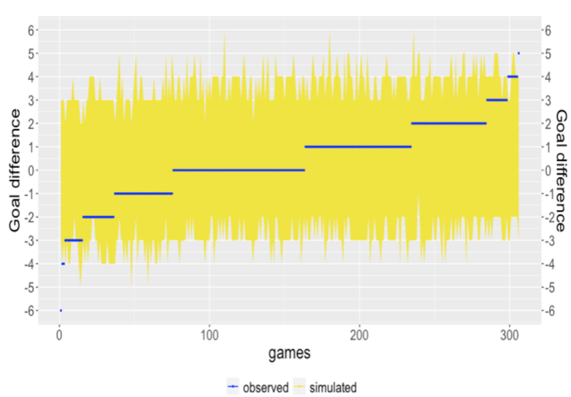


Fig.3_foot_abilities(fit2_stan, italy_2000)

The function pp_foot allows to obtain:

- an aggregated plot depicting the observed frequencies of the observed goal differences Zn=Xn−Yn, n=1,…,N plotted against the replicated ones
- a visualization of the match-ordered goal differences(GD) along with their 50% and 95% credible intervals



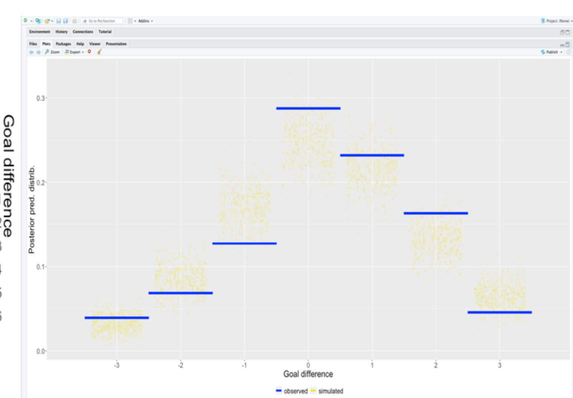


Fig.4_Aggregated GD plot.

Fig.5_Ordered GD plot.

The aggregated goal difference frequencies seem to be decently captured by the model's replications: in the first plot, blue horizontal lines denote the observed goal differences frequencies registered in the dataset, whereas yellow jittered points denote the correspondent replications. Goal-difference of 0, corresponding to the draws occurrences, is only slightly underestimated by the model. However, in general there are no particular clues of model's misfit.In the second plot, the ordered observed goal differences are plotted against their replications (50% and 95% credible intervals), and from this plot also we do not have particular signs of model's misfits.

Predictions and Predictive Accuracy

The hottest feature in sports analytics is to obtain future predictions. By considering the **posterior predictive distribution** for future and observable data ▯ , we acknowledge the whole model's prediction uncertainty (which propagates from the posterior model's uncertainty) and we can then generate observable values D conditioned on the posterior model's parameters estimates.
We may predict test matches by using the argument predict of the stan_foot function, for instance considering the last four weeks of the 2000/2001 season as the test set, and then computing posterior-results probabilities using the function foot_prob for two teams belonging to the test set, such as Reggina Calcio and AC Milan:
Darker regions are associated with higher posterior probabilities, whereas the red square corresponds to the actual observed result, 2-1 for Reggina Calcio. This final observed result had about a 5% probability to happen! While other results had good predictions, this match didn't get the observed result.

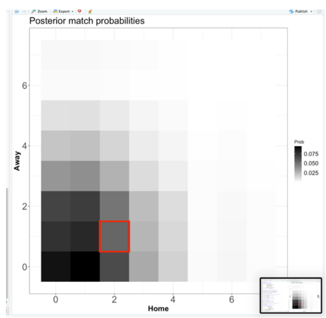


Fig.5_ Posterior Match Probability Plot

Home Win Probabilities / Rank-League Reconstruction

We can also use the out-of-sample posterior-results probabilities to compute some aggregated **home/draw/loss** probabilities (based then on the S draws from the MCMC method) for a given match:

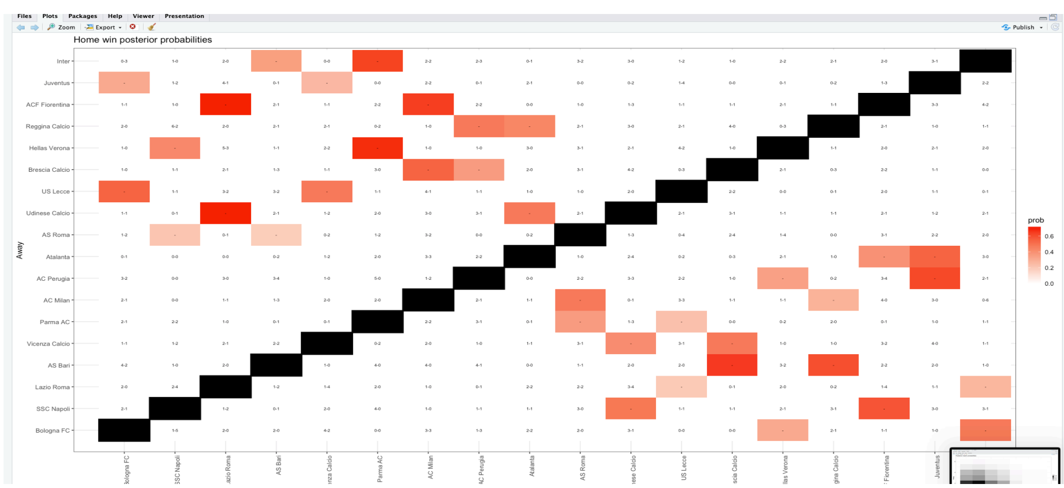


Fig.6_ Home Win Probability Plot

Red cells denote more likely home-wins (close to 0.6), such as: Lazio Roma - Fiorentina (observed result: 3-0, home win), Lazio Roma - Udinese (observed result: 3-1, home win), Juventus - AC Perugia (observed result: 1-0, home win), Brescia Calcio - AS Bari (observed result: 3-1, home win). Conversely, lighter cells denote more likely away wins (close to 0.6), such as: AS Bari - AS Roma (observed result: 1-4, away win), AS Bari - Inter (observed result: 1-2, away win).

However, predicting the final rank position (along with the teams' points) is often assimilated to an *oracle*, rather than a coherent statistical procedure:
rank-league reconstruction for the first model fit1_stan by using the **in-sample** replications (yellow ribbons for the credible intervals, solid blue lines for the observed cumulated points).

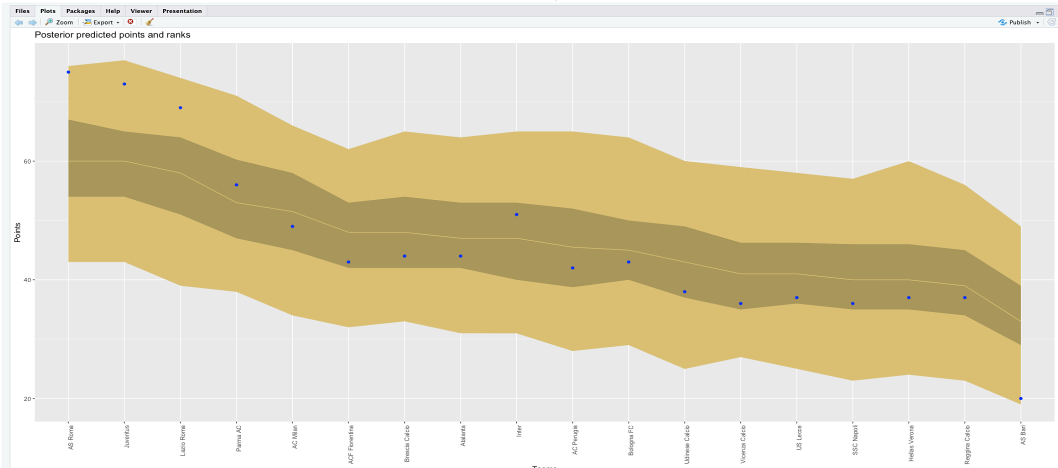


Fig.7_ Rank League Reconstructions with the Fit1_stan model

Rank-league prediction (aggregated or at team-level) for the fourth model fit4_stan by using the **out-of-sample** replications.

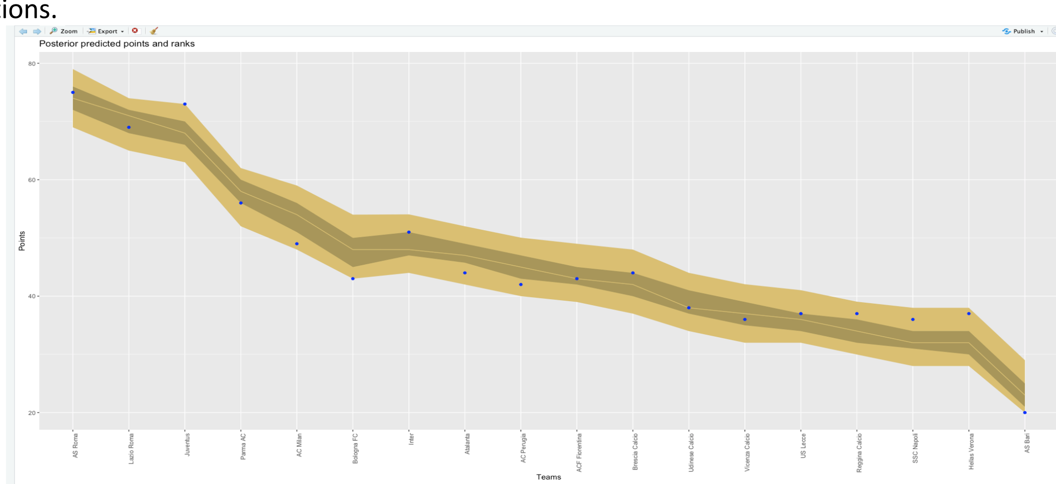


Fig.7_ Rank League Reconstructions with the Fit4 _stan model

We can clearly observe that, **Fit4_stan** has better predictive ability than the **Fit1_stan** model

MODEL COMPARISONS

```
> loo_compare(loo1, loo1_t, loo2, loo3_t)
       elpd_diff se_diff
model1  0.0       0.0
model2 -0.4       0.5
model3 -4.5       2.6
model4 -8.8       4.2
```

According to the above model LOOIC comparisons, the **weekly-dynamic double Poisson** models attain the lowest LOOIC values and are then the favored models in terms of predictive accuracy.

CONCLUSION

- The static model's fit1_stan final looic is suggesting that the assumption of static team-specific parameters is too restrictive and oversimplified to capture teams' skills over time and make reliable predictions. Anyway, from model checking we have the suggestion that even the static model has a reliable goodness of fit and could be used for some simplified analysis not requiring complex dynamic patterns.
- The dynamic model's fit4-stan, that is, the model 4 has the best predictive capability and its weekly-dynamic fit allows it to do so. It is based out of the double-poisson model and uses Monte Carlo Sampling.

THANK YOU