Bayesian hierarchical models for predicting individual performance in football (soccer)

L. Egidi* and J. S. Gabry**

*Dipartimento di Scienze Statistiche, Università degli Studi di Padova, email address: egidi@stat.unipd.it

** Department of Statistics, Columbia University, New York, email address: jgabry@gmail.com

Abstract

The task of predicting the performance of football (soccer) players is gaining increasing attention in the sports and statistical communities. We discuss the merits and flaws of a variety of hierarchical Bayesian models for detecting factors relevant to player performance in the presence of noisy data, and we compare the models on their predictive accuracy on hold-out data. We apply our analyses to the 2015–2016 season in the top Italian league, Serie A, and use the player ratings provided by a popular Italian fantasy football game as a motivating example. Our central goals are to explore what can be accomplished with a simple freely available dataset and to focus on a small number of interesting modeling and prediction questions that arise. We validate our models through graphical posterior predictive checks and we provide out-of-sample predictions for the second half of the season, using the first one as training set.

1 Introduction

In most of the published statistical research on football — Baio and Blangiardo (2010), Dixon and Coles (1997), Karlis and Ntzoufras (2009) — the authors primarily focus on modeling some aspect of the global result of a match between opposing teams (e.g., goal differential), or on predicting the order of the league table at the end of a season, and rarely on the performance of individual players over the course of a season. One reason for not focusing on predictions at the individual player level is that the performance of individual football players is noisy and hard to predict. The dimensions of the pitch combined with the number of players, the difficulty of controlling the ball without the use of hands, and many other factors all contribute to the predictive challenge. In fact, as far as well can tell from reviewing the current literature, there have been no published attempts to use a hierarchical Bayesian framework to address the challenges of modeling this kind of data.

Nevertheless, we suspect that even in football —in fantasy football at least (Bonomo et al., 2014)— a prediction task for individual performance could be well posed. In this paper we present and critique several Bayesian hierarchical models (Gelman et al., 2013, Gelman and Hill, 2006) designed to predict the results of an Italian fantasy football game with players nested within position and team. All models are estimated via Markov chain Monte Carlo using RStan, the R (R Core Team, 2016) interface to the Stan C++ library (Stan Development Team, 2016*a*).

The outcome of interest is the fantasy rating of each player in Italy's top league, Serie A, for each match of the 2015–2016 season. In some sense, we are using these data with a dual purpose: we would like to provide estimates and predictions both for the fantasy game and for the sport itself. That is, we use the fantasy

ratings as both an outcome of interest and also as a (crude) proxy for the quality of a player's performance. Although we take Fantacalcio, an Italian fantasy football product, as our example, the process of developing these models and comparing them on predictive performance does not depend on the idiosyncrasies of this particular fantasy system and is applicable more broadly.

Our central goals are to explore what can be accomplished with a simple freely available dataset (comprising only a few variables) and to focus on a small number of interesting modeling and prediction questions that arise. For this reason we also gloss over many issues that we believe should be of interest in subsequent research, for instance variable selection, additional temporal correlation structures, and the possibility of constructing more informative prior distributions.

The rest of the paper is structured as follows. In Section 2 we briefly introduce the Italian fantasy football game Fantacalcio. We then describe our dataset and present the models we fit in Section 3, where a mixture model (Section 3.3) is explained in detail and the other models derived as consequence. Preliminary results are presented in Section 4, along with a variety of posterior predictive checks as well as out-of-sample prediction tasks. Section 5 concludes.

2 Overview of the game

Fantasy sports games typically involve roster selection and match-by-match challenges against other participants with the results determined by the collective performance of the players on the fantasy rosters. In Italy, fantasy football was popularized by the brand Fantacalcio edited by Riccardo Albini in the 1990s (see `http://www.fantacalcio.it` for further details) and in the rest of the paper we use the original denomination for referring at the Italian game.

At the beginning of the season, the virtual managers are allocated a limited amount of virtual money with which to buy the players that will comprise their roster. Each player in the Italian Serie A league has an associated price determined by various factors including past performance and forecasts for the upcoming season. After every match in Serie A, the prominent Italian sports periodicals assign each player a rating, a so-called *raw score*, on a scale from one to ten. In practice there is not much variability in these scores; they typically range from four to eight, with the majority between five and seven. These raw scores are very general and largely subjective performance ratings that do not account for significant individual events (goals, assists, yellow and red cards, etc.) in a consistent way.

As a means of systematically including specific in-game events in the ratings, Fantacalcio provides the so-called *point scoring* system. Points are added or deducted from a player's initial raw score for specific positive or negative events during the match. The point scores are more variable than the raw scores, especially across positions (e.g., when comparing defending and attacking players). Goalkeepers suffer the most from the point scoring system, as they are deducted a point for every goal conceded. On the other extreme, forwards (attacking players) typically receive the highest point scores because every goal scored is worth three points.

For player i in match t the total rating y_{it} is

$$y_{it} = \mathsf{R}_{it} + \mathsf{P}_{it}, \tag{1}$$

where R is the raw score and P is the point score. Table 1 lists the game features that contribute to a player's point score P_{it} for a given match.

Event	Points
Goal	+3
Assist	+1
Penality saved*	+3
Yellow card	−0.5
Red Card	−1
Goal conceded*	−1
Own Goal	−2
Missed penality	−3

Table 1: Bonus/Malus points in Fantacalcio. The symbol * denotes an event only applicable to goalkeepers.

Importantly, there are two general ways we observe an outcome of $y_{it} = 0$. First, player i's rating for match t will be zero if the player does not play in the match — because of injury, disqualification, coach's decision, or some other reason — or he does not participate in the match for long enough for their impact to be judged by those tasked with assigning the subjective raw score ($\mathsf{R}_{it} = 0$). We will refer to this first type of zero as a *missing* observation because the player did not enter the match. Second, due to the nature of the Fantacalcio scoring system, a player can also receive a score of zero even if he does play in the match. For example, a goalkeeper who receives a raw score of four and concedes four goals will have a score of zero for the match. We will refer to this second type of zero — quite uncommon — as an *observed* zero.

One of the main aims of this paper is the attempt to model the missing values which naturally arise over the season.

3 Data and models

3.1 Data

All data for this paper are from the 2015–2016 season of the Italian Serie A and were collected from the Italian publication La Gazzetta dello Sport (`http://www.gazzetta.it`). We decided to select those players which participated in at least a third of matches during the *andata* (the first half of the season); this results in a dataset containing ratings for 237 players (18 goalkeepers, 90 defenders, 78 midfielders, and 51 forwards). For illustration purposes of the data at hand, Figure 1 displays the average ratings for the players of our dataset plotted against the initial standardized prices for each player, discussed in Section 2. For a wider overview on the data we used, see `http://www.gazzetta.it/calcio/fantanews/statistiche/serie-a-2015-16/`.

There are $N = 237$ players and $T = 38$ matches in the dataset. When fitting our models we use only the $T_1 = 19$ matches from the first half of the 2015–2016 Serie A season. The remaining matches are used later for predictive checks. The players are grouped into $J = 4$ positions (forward, midfielder, defender, goalkeeper) and $K = 5$ team clusters. The five clusters (not listed here) were determined using the official Serie A rankings at the midpoint of the season. The purpose of the team clustering is both to use a grouping structure that has some practical meaning in this context and also to reduce the computational burden somewhat by

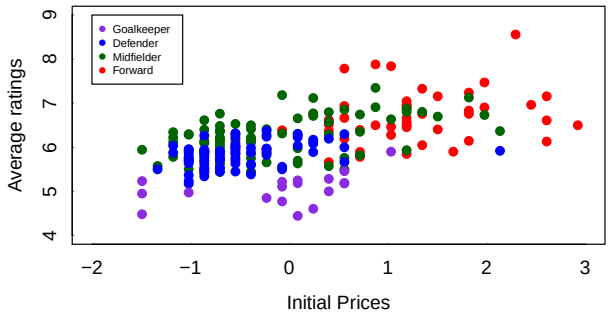


Figure 1: Average ratings plotted against the initial standardized prices for each of the 237 players of the dataset, taking into account the four different positions.

including cluster-specific parameters rather than team-specific parameters.

3.2 General framework and notation

The notation we use for data and parameters is similar to the convention adopted by Gelman and Hill (2006) for multilevel models. For match $t \in \{1,\dots,T\}$, let y_{ijkt} denote the value of the total rating for player $i \in \{1,\dots,N\}$, with position (role on the team) $j \in \{1,\dots,J\}$, on a team in team-cluster $k \in \{1,\dots,K\}$. To ease the notational burden, throughout the rest of the paper the subscripts j and k will often be implicit and we will use y_{it} in place of y_{ijkt}.We denote by $\mathbf{Z}$ the $N \times T$ binary matrix in which each element z_{it} is 1 if player i's team plays match t at its home stadium and 0 otherwise. And let q_i denote the initial standardized price for player i. These values are assigned by experts and journalists at the beginning of the season based on their personal judgement and then updated throughout the season to reflect each player's performance.

Let α_i denote the individual intercept for each player, with $i = 1,...,N$. We denote with $\gamma_{k[i]}$ the team-cluster intercept and with $\beta_{k[i],t}$ the team-cluster of the opponent in match t, with $k = 1,...,K$. In our simplified framework we set the number of team-clusters $K = 5$. $\rho_{j[i]}$ is the position intercept, with $j = 1,...,J$ and $J = 4$. The standardized prices are multiplied by a coefficient $\delta_{j[i]}$, which also varies over the J positions. Because we are interested in detecting trends in player ratings, we also incorporate the average rating up to the game $t-1$, $s_{i,t-1}$, multiplied by a factor $\lambda_{j[i]}$ estimated from the data. For the mixture model in Section 3.3, the same average rating $s_{i,t-1}$ is also multiplied by a coefficient $\zeta_{j[i]}$ in order to model the probability of participating in the match t.

For illustration purpose, here we present in detail the mixture model (hereafter, MIX), and we gloss over the technical details for the other two models we fit, which may be conceptually derived from the first one: the hierarchical autoregressive model (HAr), whose estimates are carried out by replacing all the missing values (see Section 2) with some zeros; and the hierarchical autoregressive missing model (HAr-mis), which actually treats the unobserved ratings as modeled parameters — we wrote a simple Stan program implementing the joint model for the observed and missing observations —. It is worth noticing that the MIX and the HAr-mis model are actual attempts for modeling the missingness in our dataset.

3.3 Mixture model (MIX)

Even if we found that some players have a tendency to be ejected from matches due to red cards, for instance, or tend to suffer injuries at a high rate, it would still be very challenging to arrive at sufficiently informative probability distributions for these events. Even with detailed player histories over many seasons, it would be hard to predict the number of missing matches in the current season. Nevertheless, we can try to incorporate the *missingness* behavior intrinsic to the game into our models. Assuming that it is very rare for a player to play in every match during a season, we can try to model the overall propensity for missingness. A general way of doing this entails introducing a latent variable, which we denote V_{it} and define as

$$V_{it} = \begin{cases} 1, & \text{if player } i \text{ participates in match } t, \\ 0, & \text{otherwise.} \end{cases}$$

If for each player i we let $\pi_{it} = Pr(V_{it} = 1)$, then we can specify a mixture of a Gaussian distribution and a point mass at 0 (Gottardo and Raftery, 2008)

$$p\left(y_{it} \,|\, \eta_{it}, \sigma_y^2\right) = \pi_{it}\, \mathsf{Normal}\left(y_{it} \,|\, \eta_{it}, \sigma_y^2\right) + (1 - \pi_{it})\, \delta_0, \tag{2}$$

where δ_0 is the Dirac mass at zero, σ_y^2 is the variance of the error in predicting the outcome and η_{it} is the linear predictor

$$\eta_{it} = \alpha_i + \beta_{k[i],t} + \gamma_{k[i]} + \rho_{j[i]} + \delta_{j[i]} q_i + \theta z_{it} + \lambda_{j[i]} s_{i,t-1}. \tag{3}$$

The probability π_{it} is modeled using a logit regression,

$$\pi_{it} = \text{logit}^{-1}\left(p_0 + \zeta_{j[i]} s_{i,t-1}\right), \tag{4}$$

which takes into account predictors that are likely to correlate with player participation. $s_{i,t-1}$ is the average rating for player i up to match $t-1$ and p_0 is the intercept for the logit model.

For the new parameters introduced in (4) we use the weakly informative priors

$$(p_0, \zeta) \overset{iid}{\sim} \mathsf{Normal}(0, 5^2).$$

The models for the group-level and individual parameters are

$$\alpha_i \sim \mathsf{Normal}(\mu_\alpha, \sigma_\alpha^2), \quad i = 1, \ldots, N \tag{5}$$

$$\gamma_k \sim \mathsf{Normal}(0, \sigma_\gamma^2), \quad k = 1, \ldots, K \tag{6}$$

$$\beta_k \sim \mathsf{Normal}(0, \sigma_\beta^2), \quad k = 1, \ldots, K \tag{7}$$

$$\rho_j \sim \mathsf{Normal}(\mu_\rho, \sigma_\rho^2), \quad j = 1, \ldots, J \tag{8}$$

with weakly informative prior distributions for the remaining parameters and hyperparameters.

In this formulation, the parameters μ_α and μ_ρ are the prior means of the individual intercepts and of the position-specific intercepts.

The HAr and the HAr-mis models — which differ only concerning how they use and code the missing values — may be easily defined through the distribution $\mathsf{Normal}\left(y_{it} \,|\, \eta_{it}, \sigma_y^2\right)$, with the same η_{it} as in (3).

4 Preliminary results, posterior predictive checks and predictions

4.1 Results

We fit the models via Markov chain Monte Carlo using RStan, the R interface to the Stan C++ library (Stan Development Team, 2016*a*), and monitored convergence as recommended in Stan Development Team (2016*b*). Figure 2 shows the parameter estimates for the HAr, the HAr-mis and the MIX model. At a first glance, the magnitude and the sign of the parameters for the MIX model and the HAr-mis are quite close. According to all the models, the beta's, gamma's and delta's coefficients are almost all shrunk towards their grand mean 0, with a low variability.

As it is evident, the largest source of variation for the three models is represented by the position. For what concerns the lambda's, the estimates obtained through the HAr model are greater than those obtained under the HAr-mis and the MIX model. We recall that, for every t, these coefficients are multiplied by the lagged average rating $s_{i,t-1}$; then, we strongly believe that the greater HAr values are mainly due to coding the missing values as zeros, instead of modeling as parameters, as for the HAr-mis model. All the models recognize a slight advantage due to playing at home ($\theta > 0$).

4.2 Posterior predictive checks

Now that we have estimated all of the models, we turn our attention to evaluating the fit of the models to the observed data. We use the 19 match days comprising the first half of the Serie A season — the *andata* — as training data, and for every player we make in-sample predictions for those 19 matches.

Figure 3 shows an example of a graphical posterior predictive check focusing on the *cumulative* ratings for each player over the matches in the training data. For illustration purposes, here we only show the results for one team, Napoli: the dashed black lines represent the observed values, while the red, green, blue lines represent predictions from the HAr, MIX and HAr-mis models, respectively. HAr and MIX models make predictions quite close to the observed values for many of the players. In correspondence of players with a non-trivial amount of missing (here zero) values, these models result to be preferable to the HAr-mis (see the plots for El Kaddouri, for instance).

We are also interested in the calibration of the model. In Figure 4 we display the median predictions and 50% posterior predictive intervals under the MIX for our selected team Napoli, overlaying the observed data points. In a well-calibrated model we expect half of the observed values to lie outside the corresponding 50% intervals. By this measure the MIX model has decent but not excellent calibration, since for most of the players — especially for the goalkeeper and the defenders— the 50% intervals cover more than 50% of the observed (blue) points. Conversely, for the volatile superstar Higuaín (an outlier even among forwards) a few points fall inside the intervals.

4.3 Out of sample predictions

As usual in a Bayesian framework, the prediction for a new dataset may be directly performed via the posterior predictive distribution for our unknown set of observable values. Following the same notation of Gelman et al. (2013), let us denote with $\tilde{y}$ a generic unknown observable. Its distribution is then conditional on the observed y,

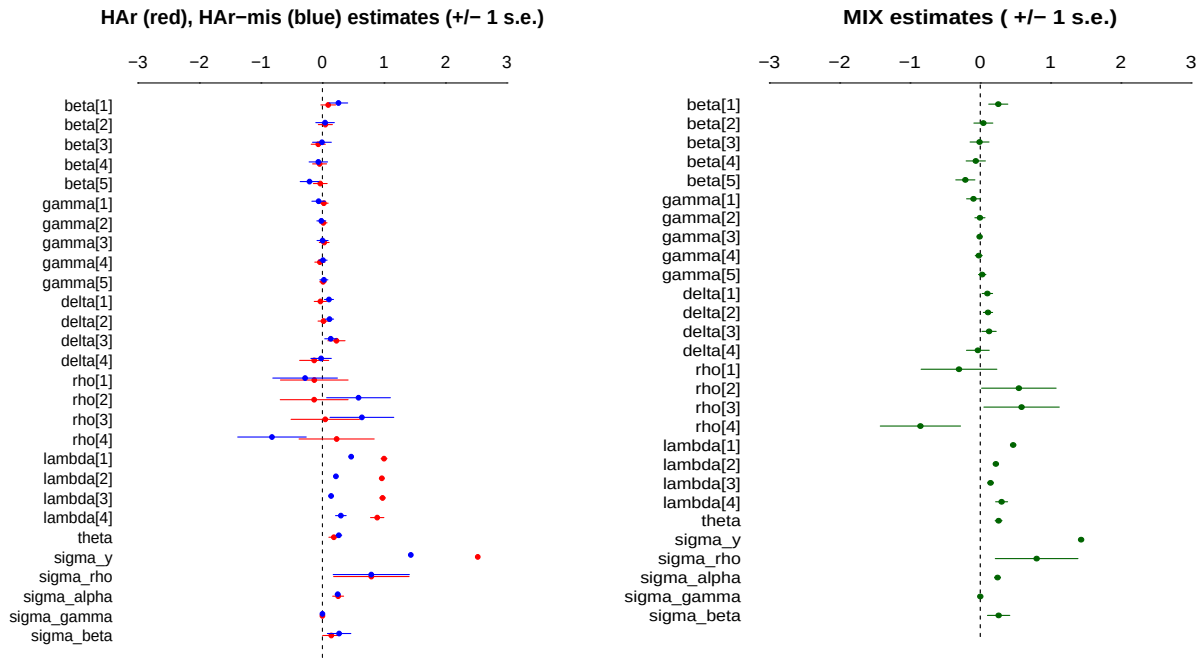


Figure 2: Posterior summary statistics for the HAr, HAR-mis and MIX model. $\beta_k,\ k = 1,...,5$ are the coefficients for the clusters opponent team (5=good, 4 = quite good, 3= medium, 2=low, 1=very low); $\gamma_k,\ k = 1,...,5$ are the coefficients for the clusters own team, same classification as before; $\delta_j,\ j = 1,...,J$ are the coefficients for the initial prices of the players; $\lambda_j,\ j = 1,...,J$ are the coefficients of the lagged observed average rating; $\rho_j,\ j = 1,...,J$ are the positions parameters (1 = Forward, 2=Midfield, 3=Defender, 4=Goalkeeper); θ is the coefficient for the home/away predictor; σ_y is the individual standard deviation; σ_α is the standard deviation for the individual intercepts $\alpha_i,\ i = 1,...,N$; σ_ρ is the position's parameters standard deviation; σ_γ is the clusters own teams standard deviation; σ_β is the clusters opponent teams standard deviation. The further set of parameters for the MIX model, represented by $\zeta_j,\ j = 1,...,J$ and p_0, is not shown here.

$$p(\tilde{y}|y) = \int_\Theta p(\tilde{y},\theta|y)d\theta = \int_\Theta p(\theta|y)p(\tilde{y}|\theta)d\theta$$

where the conditional independence of y and $\tilde{y}$ given θ is assumed. We fit the models over the $T = 19$ matches in the first half of the season and then generate predictions for the $T^\star = 19$ matches in the second half of the season.

Based on average predicted ratings for the held-out data from the second half of the 2015–2016 Serie A season, Figure 5 displays the best teams of eleven players that can be assembled from the available players according to each of the models. Also shown is the best team assembled using the observed ratings from the same set of matches. As is evident at a first glance, the predictions obtained through the HAr model are quite inefficient: this model tends to overestimate the players' rating, which are quite far from the observed ratings of the second part of the season. The team created based on the predictions from the HAr-mis and

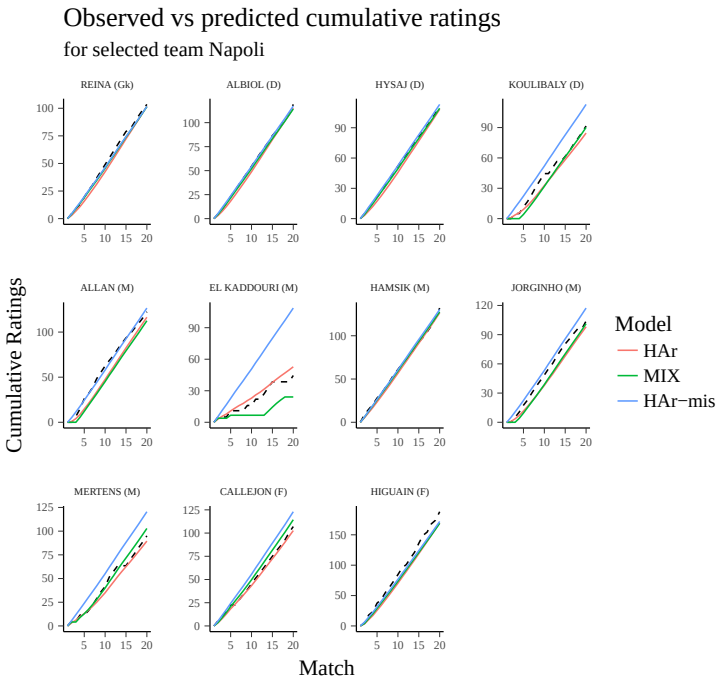


Figure 3: Posterior predictive validation of the HAr model against MIX and HAr-mis models for selected team Napoli, throughout the first half of the 2015–2016 Serie A season. The dashed black line represents the observed cumulative ratings, while the red, green, and blue lines show the medians of the predictions from the HAr, MIX and HAr-mis models, respectively.

the MIX model include four of the eleven players (Acerbi, Pogba, Hamsik, Higuaín) from the team based on the actual ratings. Dybala, who is the third best forward according to these models, is also rated highly (fifth best forward) according the observed ratings. And Rudiger, the second best defender according to the models, is also rated highly (eighth best defender).

Informally, the teams selected by the MIX and the HAr-mis models appear to be quite competitive: from this section, it is evident that modeling the missingness allows to obtain better predictions.

5 Discussion

The recent successes of so-called football (soccer) analytics are due in large part to the increasing number of available metrics for analyzing and describing the game. According to our current knowledge, the only attempt to using these and many other metrics for measuring player performance is the OPTA index. Compared to attempts like the OPTA index, our ratings may seem like very crude approximations to player performance —and they are— since they gloss over many games events. But the formulation of an index based on as many variables as possible has not been the aim of this paper. The attractiveness of our general approach is that it is based on a coherent statistical framework: we have an outcome variable y (the player rating) that is actually available, probability models relating the outcome to predictors, the ability to add

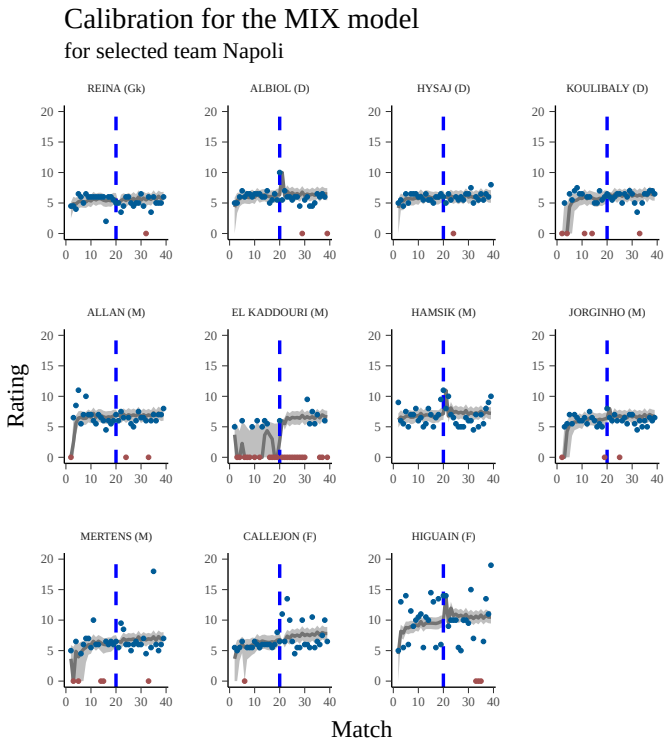
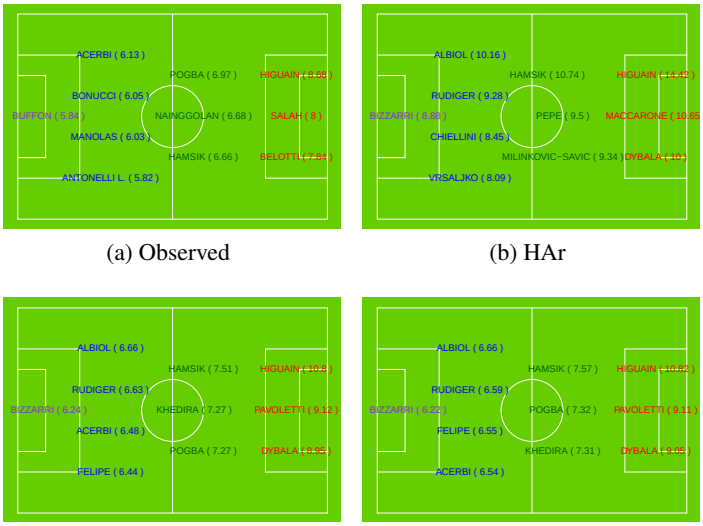


Figure 4: Calibration check for the MIX model for selected team Napoli . Blue points are observed values, red points are the zeros (missing values). The light gray ribbons represent 50% posterior predictive intervals and the dark gray lines are the median predictions. The dashed vertical blue line delimits the in-sample predictions from the out-of sample predictions.

prior information into an analysis in a principled way, and the ability to propagate our uncertainty into the predictions by drawing from the posterior predictive distribution.

We proposed some hierarchical models for predicting player ratings, taking care of the missingness as a part of the models. As expected, we preliminarily found that a player's position is, in most cases, an important factor for predicting performance (as measured by the Fantacalcio ratings). However, it is somewhat counterintuitive that the inferences from these models suggest that the quality of a player's team and the opposing team and the initial price of the players do not account for much of the variation in player ratings. It is also notable that the association between the current and lagged performance ratings —expressed by the average lagged rating—- is slightly different from zero after accounting for the other inputs into the models. Future research should consider whether other functional forms for describing associations over time are more appropriate, to what extent the inclusion of other variables in the models could improve the predictive performance, and if more informative priors can be developed at the position and team levels of the models. Another future issue should concern the choice of the training and the test set: for simplicity, in this paper we considered only the fist part of the season as training set and the second one as test set; however, we strongly believe that our models may be used in a dynamic way, using data at match day t for predicting the players' performances at match day $t+1$.



(a) Observed (b) HAr

(c) MIX (d) HAr-mis

Figure 5: Best teams according to out-of-sample prediction of average player ratings for the HAr, MIX and HAr-mis model compared to the observed best team for the second part of the season. The averaged ratings are computed for those players who played at least 15 matches in the second half of the season.

References

Baio, G. and Blangiardo, M. (2010), 'Bayesian hierarchical model for the prediction of football results', *Journal of Applied Statistics* **37**(2), 253–264.

Bonomo, F., Durán, G. and Marenco, J. (2014), 'Mathematical programming as a tool for virtual soccer coaches: a case study of a fantasy sport game', *International Transactions in Operational Research* **21**(3), 399–414.

Dixon, M. J. and Coles, S. G. (1997), 'Modelling association football scores and inefficiencies in the football betting market', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **46**(2), 265–280.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B. and andDonald B. Rubin, A. V. (2013), *Bayesian Data Analysis*, third edn, Chapman & Hall/CRC.

Gelman, A. and Hill, J. (2006), *Data analysis using regression and multilevel/hierarchical models*, Cambridge University Press.

Gottardo, R. and Raftery, A. E. (2008), 'Markov chain monte carlo with mixtures of mutually singular distributions', *Journal of Computational and Graphical Statistics* **17**(4), 949–975.

Karlis, D. and Ntzoufras, I. (2009), 'Bayesian modelling of football outcomes: using the skellam's distribution for the goal difference', *IMA Journal of Management Mathematics* **20**(2), 133–145.

R Core Team (2016), 'R: A language and environment for statistical computing'.
URL: *https://www.R-project.org/*

Stan Development Team (2016*a*), 'The Stan C++ library, version 2.14.0'.
URL: *http://mc-stan.org*

Stan Development Team (2016*b*), *Stan Modeling Language User's Guide and Reference Manual, Version 2.14.0*. `http://mc-stan.org/`.