MathSport International 2017 Conference

Padova, 26-28 June 2017

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

Leonardo Egidi[a] Jonah Sol Gabry[b]

[a]Dipartimento di Scienze Statistiche, Università degli Studi di Padova
egidi@stat.unipd.it

[b]Department of Statistics, Columbia University, New York
jgabry@gmail.com

Fantasy football has become a cornerstone among football fans and statistical amateurs. Generally, fantasy games involve

- roster selection at the beginning of the season;
- match-by-match challenges against other participants, with the results determined by the collective performance of the players on the fantasy rosters;
- a lot of free and available data, which allows for statistical analysis.

So far, there is no statistical literature referring to fantasy football models: *we try to fill this gap*, by using hierarchical Bayesian models (Gelman and Hill, 2006) for predicting the players' performances.

For player i in match t the total *fantasy rating* y_{it} is given by

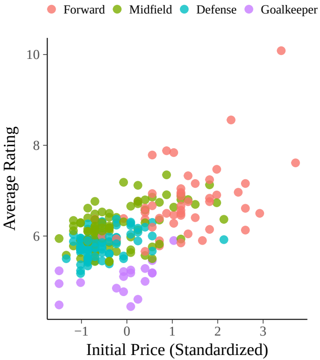
$$y_{it} = \mathsf{R}_{it} + \mathsf{P}_{it}, \tag{1}$$

where R is the **raw subjective score** on a scale from one to ten assigned by some prominent newspaper, and P is the **point score**, that takes care of specific in-game events.

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: *Point scores. * = events only applicable to goalkeepers.*

We refer to the Italian fantasy football version *Fantacalcio*. At the beginning of the season, Fantacalcio managers are allocated a limited amount of virtual money with which to buy the players that will comprise their roster.

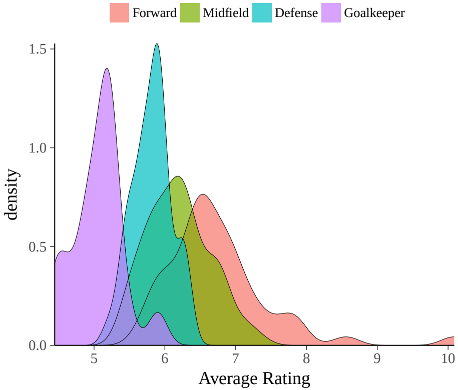


Main challenge There may be **missing values**: in fact, y_{it} will be missing if the player

- does not play in the match;
- does not participate in the match for long enough for being judged by the subjective raw score.

A natural question is: how modeling the missingness?

Data All data are from the 2015–2016 season of the Italian Serie A and were collected from the Italian publication La Gazzetta dello Sport.[1]



- $N = 237$ players, grouped into
- $J = 4$ positions (18 goalkeepers, 90 defenders, 78 midfielders, and 51 forwards), and $K = 5$ team clusters;
- $T = 38$ matches.

[1] http://www.gazzetta.it.

Predictors and models notation

- h_{it}: home/away predictor. $h_{it} = 1$ if player i's team plays match t at its home stadium and $h_{it} = 0$ if the match is played at the opponent's stadium;
- q_i: initial standardized price for player i;
- α_i: individual intercepts corresponding to each player $i = 1, ..., N$;
- $\gamma_{k[i]}$ and $\beta_{k[i],t}$: intercepts for the team-cluster of player i and the team-cluster of the team opposing player i in match t, respectively, with $k = 1, ..., K$;
- $\rho_{j[i]}$: the position-specific intercept, with $j = 1, ..., J$;
- $\delta_{j[i]}$: coefficient for the prices;
- $\lambda_{j[i]}\bar{y}_{i,t-1}$: autoregressive term;
- $\zeta_{j[i]}\bar{y}_{i,t-1}$: autoregressive term in the mixture model.

Mixture (MIX) model for the ratings

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*. Let V_{it} denote a latent variable

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

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

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

where δ_0 is the Dirac mass at zero and η_{it} is the linear predictor:

$$\eta_{it} = \alpha_0 + \alpha_i + \beta_{k[i],t} + \gamma_{k[i]} + \rho_{j[i]} + \delta_{j[i]} q_i + \lambda_{j[i]} \bar{y}_{i,t-1} + \theta h_{it}, \quad (3)$$

and σ_y is the standard deviation of the error in predicting the outcome.

Mixture (MIX) model for the ratings

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

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

which takes into account $\bar{y}_{i,t-1}$, the average rating for player i up to match $t-1$; p_0 is an intercept for the logit model. The individual-level, position-level, and team-cluster-level parameters are given hierarchical normal priors,

$$\alpha_i \sim \text{Normal}(0,\, \sigma_\alpha), \quad i = 1, \ldots, N \tag{5}$$
$$\gamma_k \sim \text{Normal}(0,\, \sigma_\gamma), \quad k = 1, \ldots, K \tag{6}$$
$$\beta_k \sim \text{Normal}(0,\, \sigma_\beta), \quad k = 1, \ldots, K \tag{7}$$
$$\rho_j \sim \text{Normal}(0\, \sigma_\rho), \quad j = 1, \ldots, J \tag{8}$$

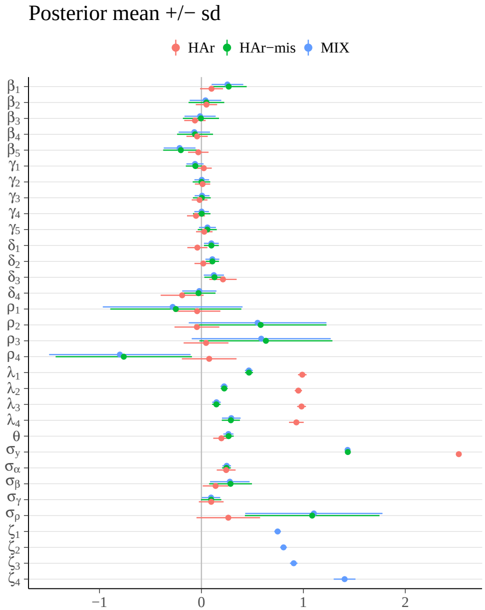
with weakly informative prior distributions for the remaining parameters and hyperparameters.

Our mixture specification allows for some natural other models extensions

- $\pi_{it} \sim logit^{-1} \rightarrow$ MIX
- $\pi_{it} = 1$, fixed
    - missing $y_{it} = 0 \rightarrow$ Hierarchical autoregressive model (HAr);
    - missing $y_{it} \sim f \rightarrow$ Hierarchical autoregressive model with missing model (HAr-Mis);

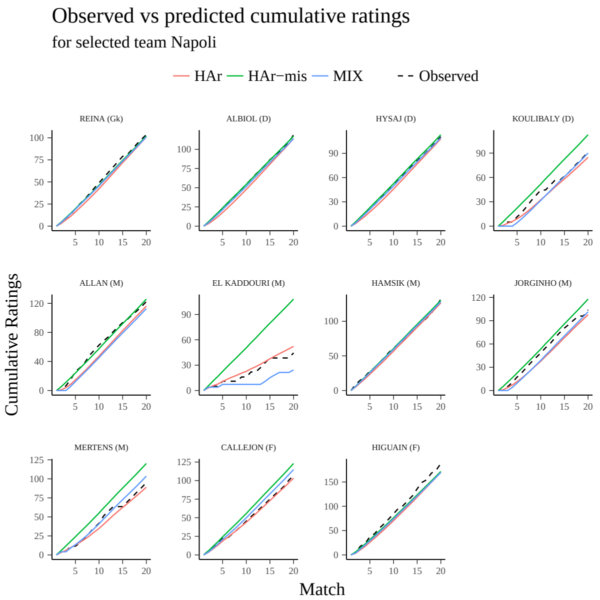
Remark We want to estimate our models and predict the fantasy rating on a test set. Some interesting issue arise: missingness, model calibration, posterior predictive checks, out-of-sample predictions...

Setup We use the first half of the season as training set and the second half as test set.



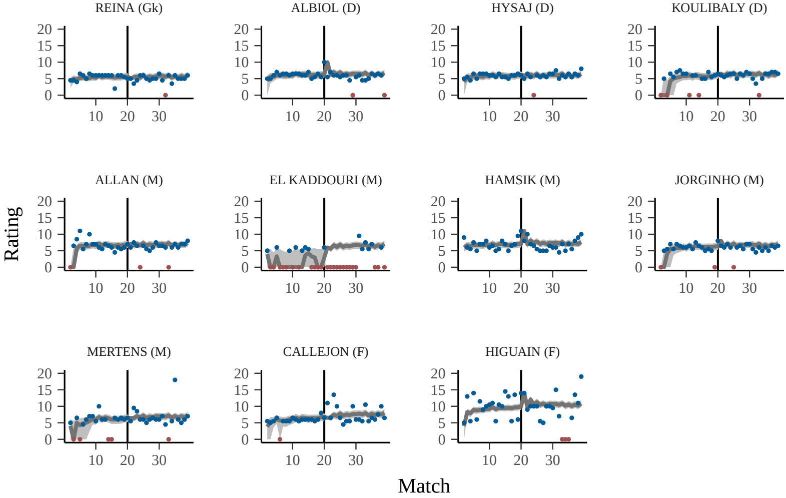
MIX and HAr-Mis, that take care of the missingness, produce similar result. (Models fitted via Markov chain Monte Carlo (3000 iter., burn-in=1000)) using RStan Stan Development Team (2016a) and monitored convergence as recommended in Stan Development Team (2016b)).

Posterior predictive checks

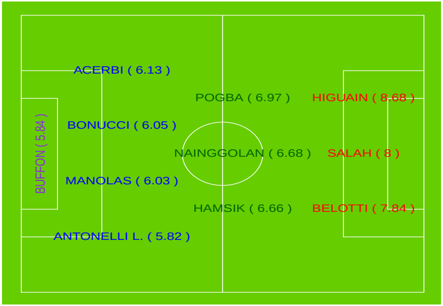


Calibration for the MIX model

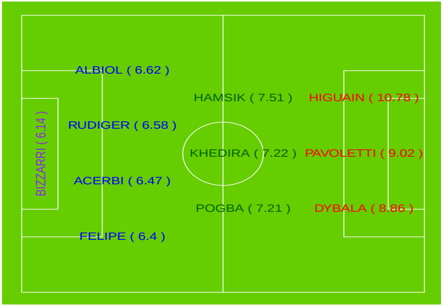
for selected team Napoli



Final aim **Select the best roster.** According to our posterior predictions for the second part of the season, we can create the best roster.



(a) Observed team



(b) MIX team

Let us note that the MIX is quite competitive; moreover Rudiger (defender, Roma) and Khedira (midfield, Juventus) performed pretty well in the 2016-2017 Serie A season.

Discussion and further work

- We proposed a class of hierarchical Bayesian models for predicting player ratings, in the presence of noisy fantasy football (soccer) data;
- these fantasy ratings may be seen as a crude proxy for players' performances;
- we took care of the missingness issue;
- after controlling for missingness, the out-of sample predictive fit is good (the selected team appears to be competitive). Still checking for calibration.
- Further work
  - Dynamic prediction (match after match), adding data for more seasons, adding predictors;
  - **app** for fantasy football managers (working on).

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

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

Stan Development Team (2016a). RStan: the R interface to Stan, version 2.14.1.

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