A Bayesian inference approach for determining player abilities in football

G.A. Whitaker*†, R. Silva* and D. Edwards†

* Department of Statistical Science, University College London, UK
† Stratagem Technologies, UK

gavin.whitaker@ucl.ac.uk (https://www.ucl.ac.uk/statistics/people/gavin-whitaker)

The Problem

- In football it is natural to ask "How good is a player at a specific event/skill?" Or "Can we learn anything about the present to help predict future results?"
- We consider the task of determining a football player's ability for a given event—goal scoring, shot taking and being involved in creating a chance (along with the defensive counterparts: stopping a goal, stopping a shot and disrupting play).
- We then use these inferred abilities to help predict future matches, extending the Bayesian hierarchical model of Baio & Blangiardo (2010).
- A large dataset is available to us, which gives counts for each player (740), in each event (39 events) over 2 full seasons of the English Premier League (2013/14 & 14/15).
- Given the large dataset (and number of parameters) we appeal to variational inference (VI) methods to fit the model, and to allow a computationally efficient approach.
- These techniques give a method to access the abilities of players, whilst quantifying the uncertainty around any given player.

Player Ability Model

- For a given ability/event, we have K matches, numbered $k = 1, \ldots, K$. The teams in fixture k are $T_k = \{T_k^H, T_k^A\}$ (H: home team, A: away team).
- Take P to be the set of all players who feature in the dataset, and $P_k^j \in P$ to be the players who play for team j in fixture k.
- We model event e_1 against event e_2, such that $E = \{e_1, e_2\}$. An example being Goal against GoalStop.
- Let the number of occurrences of an event in a match, for a player $\left(X_{i,k}^e\right)$, follow a Poisson distribution, that is $X_{i,k}^e \sim Pois\left(\eta_{i,k}^e \tau_{i,k}\right)$, where

$$\eta_{i,k}^e = \exp\left\{\Delta_i^e + \tau_{i,k}\left(\lambda_1^e \sum_{i' \in P_k^j} \Delta_{i'}^e - \lambda_2^e \sum_{i' \in P_k^{T_k \setminus j}} \Delta_{i'}^{E \setminus e}\right) + \left(\delta_{T_k^H, j}\right)\gamma^e\right\},$$

$\delta_{r,s}$ is the Kronecker delta, $\tau_{i,k}$ is the fraction of time player i spent on the pitch, γ^e is the home effect and Δ_i^e represents the (latent) ability of each player for a specific event. The impact of a player's own team is captured through λ_1^e, with λ_2^e describing the opposition's ability to stop the player in that event.
- We take the (reasonably uninformative) prior, $\pi\left(\Delta_i^e\right) \sim N\left(-2, 2^2\right)$.
- The model is formed using *standard* VI methods, where we make the *mean-field* assumption, in which the latent variables are assumed to be mutually independent.
- The model is fit by maximising the ELBO which is available in closed-form here.

Hierarchical Bayesian Model

- To predict future matches we extend the model of Baio & Blangiardo (2010) (who present the model of Karlis & Ntzoufras (2003) in a Bayesian framework), to include the inferred player abilities (Δ).
- The model is a Poisson-log normal model. For ease, we present the baseline model for a single fixture (the extension is trivial).
- We let y_t, $t \in \{H, A\}$ be the total number of goals scored for a team, where $y_t | \theta_t \overset{indep}{\sim} Pois\left(\theta_t\right)$, with

$$\log\left(\theta_H\right) = \text{home} + \text{att}_H + \text{def}_A \quad \text{and} \quad \log\left(\theta_A\right) = \text{att}_A + \text{def}_H.$$

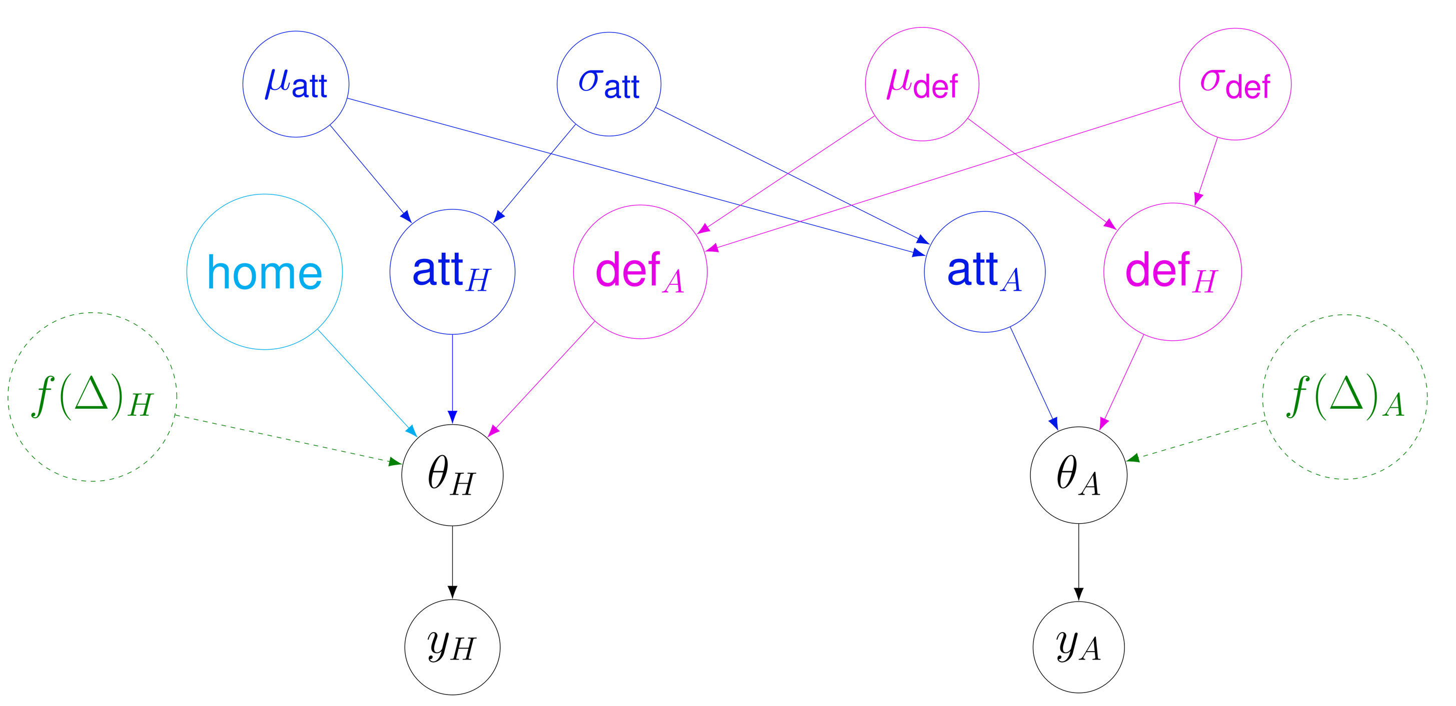
- Each team has their own team-specific attack and defence ability. A constant home effect is also included in the rate of the home team's goals. The attack and defence parameters for each team are seen to be draws from a common distribution

$$\text{att}_t \sim N\left(\mu_{\text{att}}, \sigma^2_{\text{att}}\right) \quad \text{and} \quad \text{def}_t \sim N\left(\mu_{\text{def}}, \sigma^2_{\text{def}}\right).$$

For identifiability, we impose sum-to-zero constraints on the attack and defence parameters.
- We follow Baio & Blangiardo (2010) and assume the priors

$$\mu_{\text{att}} \sim N\left(0, 100^2\right), \quad \mu_{\text{def}} \sim N\left(0, 100^2\right), \quad \text{home} \sim N\left(0, 100^2\right)$$
$$\sigma_{\text{att}} \sim Inv\text{-}Gamma(0.1, 0.1), \quad \sigma_{\text{def}} \sim Inv\text{-}Gamma(0.1, 0.1).$$

- Graphically the model is



Dashed-green indicates our extension.

- Our extension includes the latent Δs of the Player Ability model in the scoring intensities, through $f(\Delta)_*$.
- For a single pair of events, a suitable choice could be

$$f(\Delta)_H = \sum_{i \in I^{T^H}} \mu_{\Delta_i^e} - \sum_{i \in I^{T^A}} \mu_{\Delta_i^{E \setminus e}} \quad \text{and} \quad f(\Delta)_A = \sum_{i \in I^{T^A}} \mu_{\Delta_i^e} - \sum_{i \in I^{T^H}} \mu_{\Delta_i^{E \setminus e}},$$

where I^j is the initial eleven players who start a fixture for team j and μ_Δ is the mean of the marginal posterior variational densities.
- We fit the model using `PyStan` (Stan Development Team 2016).
- Full details of both models can be found in Whitaker et al. (2017).

Applications

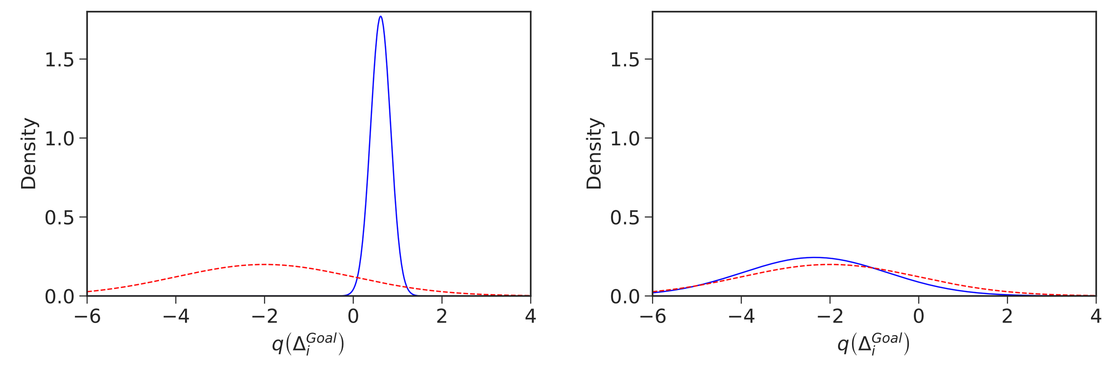
Determining a player's ability

- We look to create an ordering of players abilities, considering occurrences of Goal against GoalStop (GoalStop is an event type of our own creation aiming to represent all the things a team can do to stop the other team from scoring a goal).
- The top 10 goal scorers in the 2013/2014 English Premier League based on the 2.5% quantile of the marginal posterior variational density for each player, $q(\Delta_i^{\text{Goal}})$ are

Goal - top 10

Rank	Player	2.5% quantile	Mean	Standard deviation	Observed	Observed rank	Rank difference	Time played
1	Suarez	0.508	0.869	0.184	31	1	0	3185
2	Sturridge	0.176	0.617	0.225	21	2	0	2414
3	Aguero	0.147	0.636	0.250	17	4	+1	1616
4	Y. Toure	-0.043	0.395	0.224	20	3	-1	3113
5	Rooney	-0.056	0.421	0.243	17	5	0	2625
6	Dzeko	-0.065	0.424	0.249	16	8	+2	2128
7	van Persie	-0.136	0.430	0.289	12	15	+8	1690
8	Remy	-0.230	0.302	0.271	14	11	+3	2274
9	Bony	-0.257	0.238	0.252	16	7	-2	2644
10	Rodriguez	-0.354	0.161	0.263	15	10	0	2758

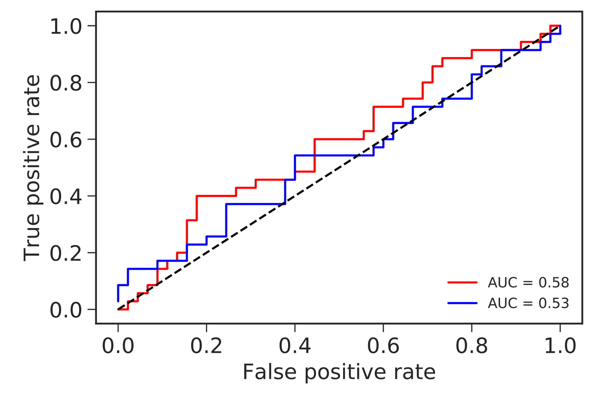
- The ranking shown appears sensible, and is very close to that obtained by ranking players on the total number of goals scored over the season.
- Through the marginal posterior variational densities we can see the differences between different players in the 2013/2014 English Premier League, especially those that play a lot or not.



Left: Sturridge (29 appearances), *Right*: Reed (4 appearances, totalling 8 mins). Prior, posterior.

Prediction

- We look to predict the goals in future matches to inform decisions for the over/under (OU) and Asian (AH) handicap markets.
  - OU: whether a certain number of goals will be scored (over) or not (under).
  - AH: will a team win given a certain handicap to their score.
- We fit the model to the past to predict the future using incremental blocks. There are 5 blocks over a season.
- The latent player abilities (Δs) are included for the event types Goal, Shots and being involved in creating a chance; along with the defensive counterparts.
- We compare our extension (Δ model) against the baseline of Baio & Blangiardo (2010).
- Averaging probabilities across the posterior sample we can construct ROC curves.

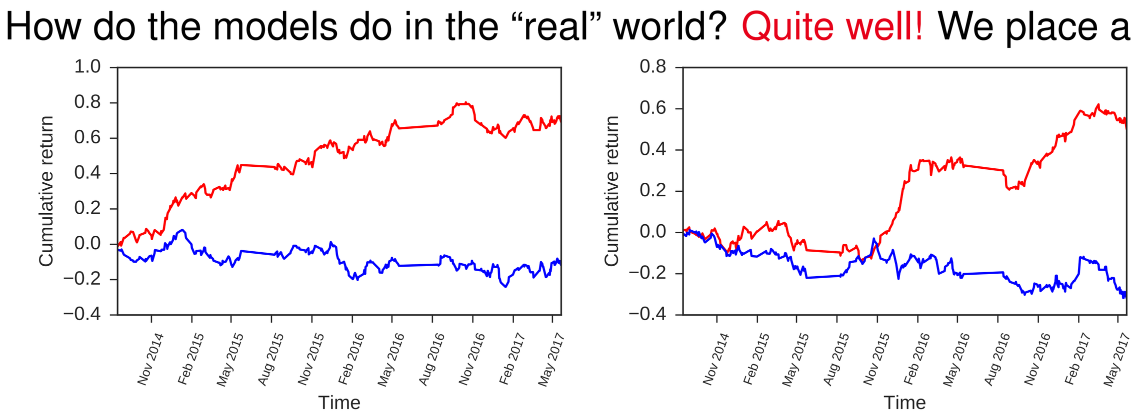


Area under the curve values

Model	Block 1	2	3	4	5
Δ model	0.54	0.65	0.58	0.68	0.62
Baseline	0.47	0.60	0.53	0.55	0.61

Including the latent player abilities in the model leads to a better predictive performance.

- How do the models do in the "real" world? Quite well! We place a flat stake on each bet of $£100$.



Left: OU, Δ model: $£6914.09$, baseline: -$£1189.51$. *Right*: AH, Δ model: $£5016.87$, baseline: -$£3193.15$.

- See Whitaker et al. (2017) for a full analysis of the dataset.

Summary and References

- We have provided a framework to establish player abilities in a Bayesian inference setting. Our approach is computationally efficient and centres on variational inference methods.
- We have shown that inferences for player's abilities are reasonably accurate and have close ties to reality.
- By extending the Bayesian hierarchical model of Baio & Blangiardo (2010) to include these latent player abilities, we can gain reasonable predictions of future matches.
- We observed an improvement in performance over the baseline model, and a profitable strategy when considering the betting market.

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

Karlis, D. & Ntzoufras, I. (2003), 'Analysis of sports data by using bivariate Poisson models', *Journal of the Royal Statistical Society: Series D (The Statistician)* **52**(3), 381–393.

Stan Development Team (2016), 'PyStan: the Python interface to Stan, version 2.15.0.0'. **URL:** *http://mc-stan.org*

Whitaker, G. A., Silva, R. & Edwards, D. (2017), 'A Bayesian inference approach for determining player abilities in soccer', *arXiv preprint arXiv:1710.00001* .