Bayesian hierarchical model for the prediction of football results

Tianyu Cao (u5687658), Chen Chen (u6111797), Zheyan Chen (u6324634) & Rui Qiu (u6139152)

2018-05-27

background

- paper information:
  - *Bayesian hierarchical model for the prediction of football results*, Gianluca Baio & Marta Blangiardo, Journal of Applied Statistics, 21 Jan 2010.
- statistical question:
  - estimate the characteristics that bring a team to lose or win a game, and to predict the score of a particular match
- why hierarchical?
  - "Hierarchical models are widely used in many different fields as they are natural way of taking into account relations between variables, by assuming a common distribution for a set of relevant parameters thought to underlie the outcomes of interest."

data - Serie A (1991-1992)

g	Home Team	Visiting Team	$h(g)$	$a(g)$	y_{g1}	y_{g2}
1	Verona	Roma	18	15	0	1
2	Napoli	Atalanta	13	2	1	0
3	Lazio	Parma	11	14	1	1
4	Cagliari	Sampdoria	4	16	3	2
...	...	...	...	...	...	...
303	Sampdoria	Cremonese	16	5	2	2
304	Roma	Bari	15	3	2	0
305	Inter	Atalanta	9	2	0	0
306	Torino	Ascoli	17	1	5	2

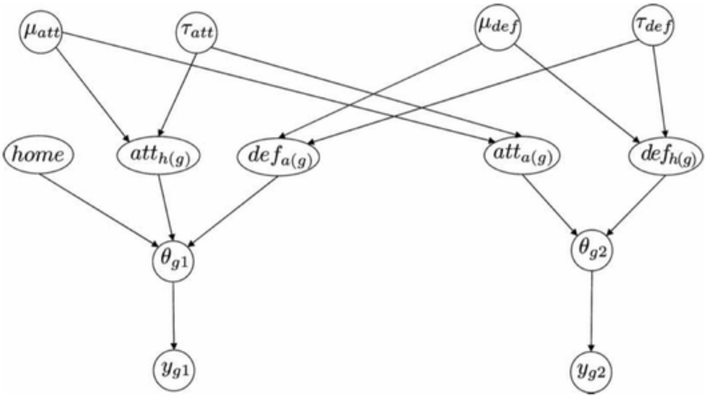
base model - parameters and variables

- $T = 18$, $G = 306$
- $y_{gj} \mid \theta_{gj} \sim Poisson(\theta_{gj})$, conditional on parameters $\theta = (\theta_{g1}, \theta_{g2})$ (scoring intensity in the g-th game)
- $\log \theta_{g1} = home + att_{h(g)} + def_{a(g)}$
- $\log \theta_{g2} = att_{a(g)+def_{a(g)}}$
  - $home$
  - att, def

base model - priors

- *home*: fixed effect, uninformative:
    - $home \sim N(0, 0.0001)$
- att, def: team-specific
    - $att_t \sim N(\mu_{att}, \tau_{att}), def_t \sim N(\mu_{def}, \tau_{def})$
    - zero-sum contraint: $\sum_{t=1}^{T} att_t = 0, \sum_{t=1}^{T} def_t = 0$
- hyperpriors of att, def:
    - uninformative
    - $\mu_{att} \sim N(0, 0.0001), \mu_{def} \sim N(0, 0.0001)$
    - $\tau_{att} \sim Gamma(0.1, 0.1), \tau_{def} \sim Gamma(0.1, 0.1)$

base model - priors



- unobservable hyper-parameters $\eta = (\mu_{att}, \mu_{def}, \tau_{att}, \tau_{def})$

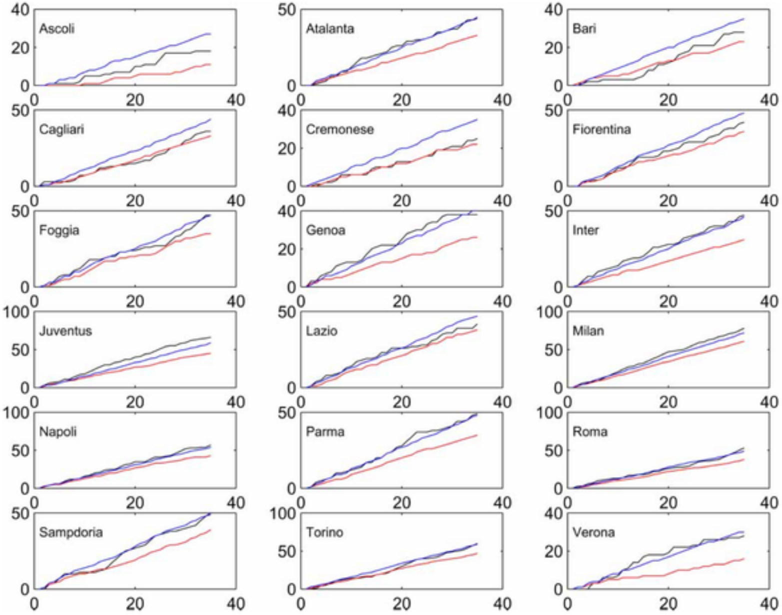
base model - estimate

estimate the value of the main effects that we used to explain the scoring rates (by entering the evidence provided by the observed results, **y** vector and updating the prior distributions by means of the Baye's thm using a MCMC)

base model - estimate

Teams	Attack effect				Defense effect			
	Mean	2.5%	Median	97.5%	Mean	2.5%	Median	97.5%
Ascoli	−0.2238	−0.5232	−0.2165	0.0595	0.4776	0.2344	0.4804	0.6987
Atalanta	−0.1288	−0.4050	−0.1232	0.1321	−0.0849	−0.3392	−0.0841	0.1743
Bari	−0.2199	−0.5098	−0.2213	0.0646	0.1719	−0.0823	0.1741	0.4168
Cagliari	−0.1468	−0.4246	−0.1453	0.1255	−0.0656	−0.3716	−0.0645	0.2109
Cremonese	−0.1974	−0.4915	−0.1983	0.0678	0.1915	−0.0758	0.1894	0.4557
Fiorentina	0.1173	−0.1397	0.1255	0.3451	0.0672	−0.1957	0.0656	0.3372
Foggia	0.3464	0.1077	0.3453	0.5811	0.3701	0.1207	0.3686	0.6186
Genoa	−0.0435	−0.3108	−0.0464	0.2149	0.1700	−0.0811	0.1685	0.4382
Inter	−0.2077	−0.4963	−0.2046	0.0980	−0.2061	−0.5041	−0.2049	0.0576
Juventus	0.1214	−0.1210	0.1205	0.3745	−0.3348	−0.6477	−0.3319	−0.0514
Lazio	0.0855	−0.1626	0.0826	0.3354	0.0722	−0.1991	0.0742	0.3145
Milan	0.5226	0.2765	0.5206	0.7466	−0.3349	−0.6788	−0.3300	−0.0280
Napoli	0.2982	0.0662	0.2956	0.5267	0.0668	−0.2125	0.0667	0.3283
Parma	−0.1208	−0.3975	−0.1200	0.1338	−0.2038	−0.5136	−0.2031	0.0859
Roma	−0.0224	−0.2999	−0.0182	0.2345	−0.1358	−0.4385	−0.1300	0.1253
Sampdoria	−0.0096	−0.2716	−0.0076	0.2436	−0.1333	−0.4484	−0.1317	0.1346
Torino	0.0824	−0.1821	0.0837	0.3408	−0.4141	−0.7886	−0.4043	−0.1181
Verona	−0.2532	−0.5601	−0.2459	0.0206	0.3259	0.1026	0.3254	0.5621
Home	0.2124	0.1056	0.2128	0.3213				

base model - prediction



why not base model

- **overshrinkage**: some of extreme occurrences tend to be pulled towards the grand mean of the observations.
- it could be possible for our case since the performance of different teams in a league in a season can diverge, some are really good, some are really bad.
- previous model the hyper-parameters assume all attack/defense intensity are drawn by a common process, which is not sufficient to capture the different skill levels of each team, therefore, shrinkage, penalizing extremely good teams and overestimating the bad teams

hierarchical model

- stratify the teams into 3 levels: **top, mid** and **bottom**
- model the *att* and *def* parameters using a non central t distribution on $\nu = 4$ degrees of freedom, instead 2 from normal
- the observable variables, the prior specificaition for θ_{gj}, the hyper-parameter home is unchanged, other hypers are modeled as
  - each team t has two latent (unobservable) variables $grp^{att}(t), grp^{def}(t)$ taking on the value of $1, 2, 3$ representing **bottom, mid, top** level. These are given suitable categorical dist each depending on a vector of prior probability $\pi^{att} = (\pi^{att}_{1t}, \pi^{att}_{2t}, \pi^{att}_{3t})$ and $\pi^{def} = (\pi^{def}_{1t}, \pi^{def}_{2t}, \pi^{def}_{3t})$
  - $\pi \sim Dirichlet(1, 1, 1)$

hierarchical model

- $att_t \sim nct(\mu^{att}_{grp(t)}, \tau^{att}_{grp(t)}, \nu)$
- $def_t \sim nct(\mu^{def}_{grp(t)}, \tau^{def}_{grp(t)}, \nu)$
- $grp^{att}(t), grp^{def}(t) \implies$
  - $att_t = \sum_{k=1}^{3} \pi^{att}_{kt} \times nct(\mu^{att}_{grp(t)}, \tau^{att}_{grp(t)}, \nu)$
  - $def_t = \sum_{k=1}^{3} \pi^{def}_{kt} \times nct(\mu^{def}_{grp(t)}, \tau^{def}_{grp(t)}, \nu)$
- $\mu^{att}_1 \sim truncN(0, 0.001, -3, 0), \mu^{def}_1 \sim truncN(0, 0.001, 0, 3)$
- $\mu^{att}_3 \sim truncN(0, 0.001, 0, 3), \mu^{def}_3 \sim truncN(0, 0.001, -3, 0)$
- $\mu^{att}_2 \sim N(0, \tau^{att}_2), \mu^{def}_2 \sim N(0, \tau^{def}_2)$
- $\tau^{att}_k \sim Gamma(0.01, 0.01), \tau^{def}_k \sim Gamma(0.01, 0.01)$

hierarchical model - updated results (07-08)

- context changed: winning points, league expansion
- bigger gap between two extremes

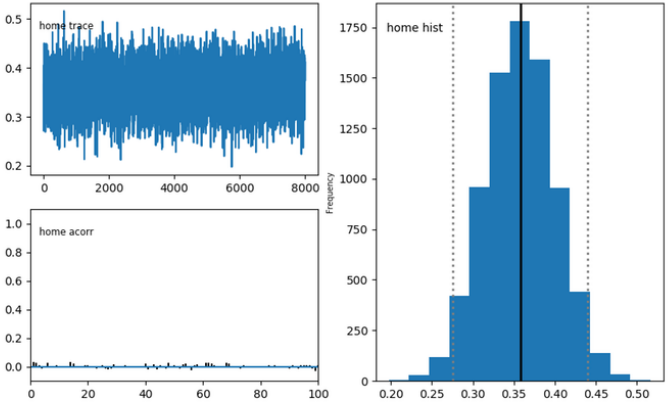
Table 3. Posterior predictive validation of the model. Observed and estimated league table (2007/2008).

Teams	Observed results			Basic model (medians)			Mixture model (medians)		
	Points	Scored	Conc'd	Points	Scored	Conc'd	Points	Scored	Conc'd
Inter	85	69	26	69	62	38	76	65	30
Roma	82	72	37	67	64	42	70	68	40
Juventus	72	72	37	68	65	42	69	67	41
Fiorentina	66	55	39	59	52	43	58	51	43
Milan	64	66	38	64	60	42	66	62	42
Sampdoria	60	56	46	57	53	47	57	54	47
Udinese	57	48	53	50	47	50	50	45	50
Napoli	50	50	53	52	49	49	50	47	51
Genoa	48	44	52	52	50	51	48	43	50
Atalanta	48	52	56	49	45	49	52	50	52
Palermo	47	47	57	49	47	52	47	43	52
Lazio	46	47	51	50	46	49	49	44	50
Siena	44	40	45	49	42	46	48	41	48
Cagliari	42	40	56	46	41	51	45	41	52
Torino	40	36	49	44	39	51	45	39	49
Reggina	40	37	56	45	38	48	44	40	52
Catania	37	33	45	45	37	46	45	37	48
Empoli	36	29	52	41	34	50	40	34	51
Parma	34	42	62	45	42	54	44	42	55
Livorno	30	35	60	40	38	53	42	38	54

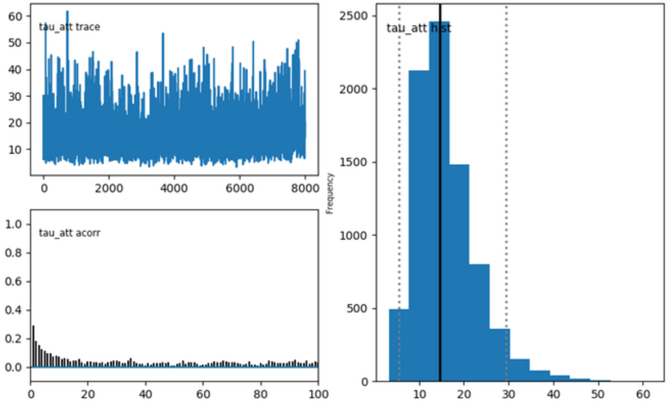
replication - 07-08 Serie A data

- data: jokecamp/FootballData (raw data), wikipedia page (result table)
- tool: pymc
- setup for mcmc: `iter=200,000, burnin=40,000, thin=20`
- Note: simplification $att_t \sim N(0, \tau_{att}), def_t \sim N(0, \tau_{def})$

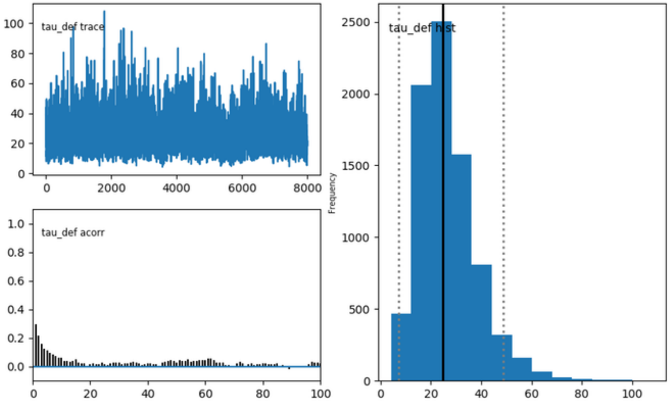
replication - diagnostics



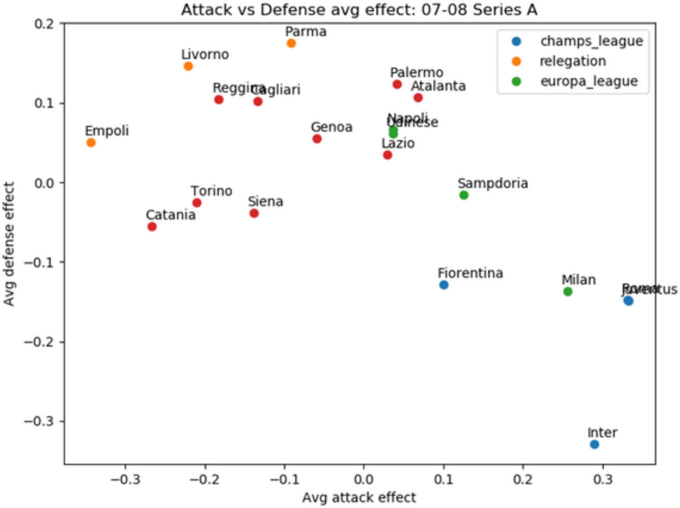
replication - diagnostics



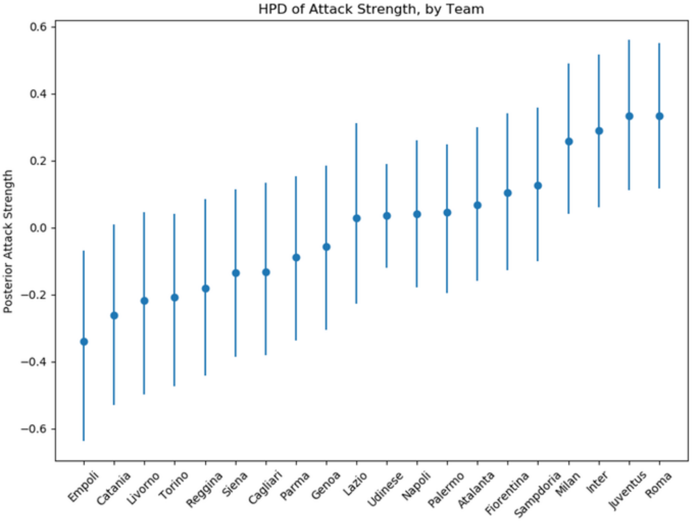
replication - diagnostics



replication - mean att vs mean def



replication - HPD



replication - simulation

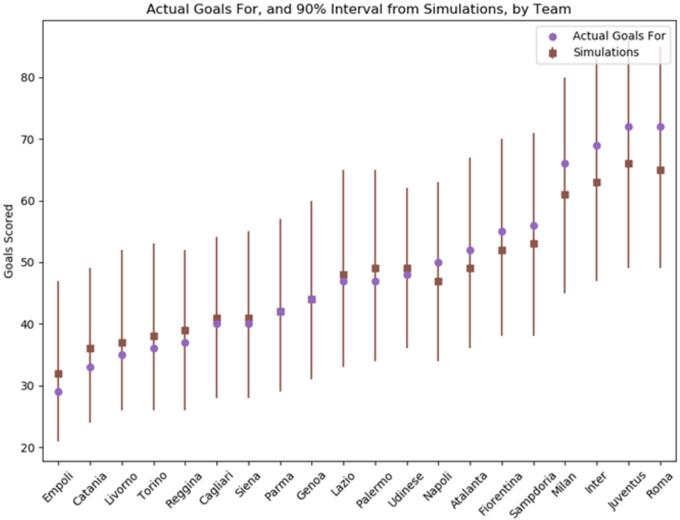
simulated

- top 4: Juventus (70/60), Roma (68/52), Inter (67/50), Milan (66/66)
- bottom 3: Udinese (38/33) , Empoli (37/34), Torino (29/35)

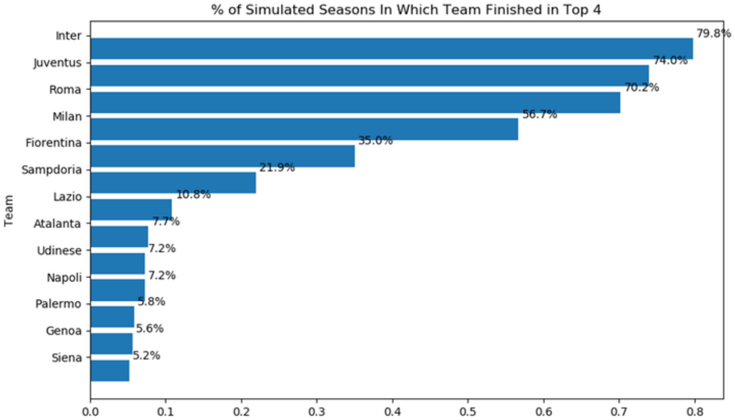
real

- top 4: Inter (85/69), Roma (82/72), Juventus (72/72), Fiorentina (66/55)
- bottom 3: Empoli (36/29), Parama (34/42), Livorno (30/35)

replication - 90% credible interval



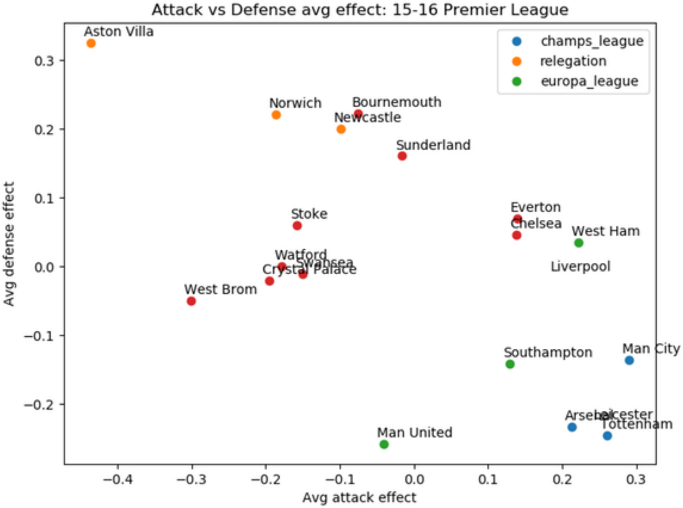
replication - top 4 summary



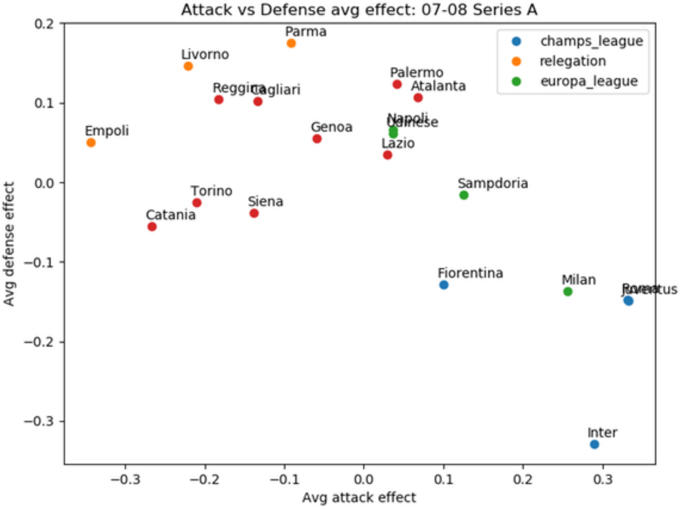
application - thoughts

- wait, but what can we **really** learn from these data?
- we are using history to "predict" history, seems not very meaningful
  - maybe, use past seasons/years to predict this year? (under the circumstance that no transfer, nor relegation/promotion happens!)
- but at least, we can learn some *features* about a league

application - 15-16 Premier League data



application - comparison



other attempts

- existing other attempts
  - Sabermetrics
    - search for objective knowledge about baseball
    - the mathematical and statistical analysis of baseball record
  - ELO ratings
    - a measure of the team's current strength
  - models based on machine learning
    - one of the intelligent methodologies
- the fields can be applied to
  - gambling
  - coaching improvements
  - journalism

conclusion

- base model
  - extreme occurences – overshrinkage
- Bayesian hierarchical model
  - bettering performance in prediction
  - easily implemented by using standard MCMC algorithms
  - easily extended to include a mixture structure

conclusion

- limitation and improvement
  - predictions are obtained in one batch
    - using the observed results to estimate the parameters
  - include more variables

references

- Baio, G. and Blangiardo, M., 2010. Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics*, Volume 37 Issue 2.
- Milad Kharratzadeh, *A Hierarchical Bayesian Model for Predicting Soccer Results*
- Gavin A. Whitaker, Ricardo Silva, Daniel Edwards, A Bayesian inference approach for determining player abilities in soccer. *arXiv.org*.
- Rasmus Baath, *Modeling Match Result in Soccer using a Hierarchical Bayesian Poisson Model*, part 1 & part 2
- Albert, J., 1997. An introduction to sabermetrics. *Bowling Green State University (http://www-math.bgsu.edu/~albert/papers/saber.html).*
- Bunker, R.P. and Thabtah, F., 2017. A machine learning framework for sport result prediction. *Applied Computing and Informatics.*

references

- Costa, G.B., Huber, M.R. and Saccoman, J.T., 2007. *Understanding sabermetrics: An introduction to the science of baseball statistics.* McFarland.
- Hvattum, L.M. and Arntzen, H., 2010. Using ELO ratings for match result prediction in association football. *International Journal of forecasting,* 26(3), pp.460-470.
- Ulmer, B., Fernandez, M. and Peterson, M., 2013. *Predicting Soccer Match Results in the English Premier League* (Doctoral dissertation, Doctoral dissertation, Ph. D. dissertation, Stanford).