Using player abilities to predict football

Gavin Whitaker

ISBIS: Data Science in Industry, July 2018

Who are Stratagem?

- A data science and trading company - focus on sports
- Football/tennis/basketball
- Combine state-of-the-art analytics, data and expert analysis to find cutting edge solutions
- Aim to utilise the full power of the modern AI toolkit
- Offer:
  - Betting insights
  - Modelling
  - Bet prices
  - Trading services

The problem

Aim

- Wish to determine the ability of a given player in a specific event, e.g. passing, scoring a goal etc
- Use these abilities to help predictions, Over/Under or Asian handicap markets

Possible questions

- Can we rate/rank players on their ability in an event?
- How important is a player to a team?
- What happens if that player is missing from the team?
- What happens if you add a player to a team?
- What happens if you swap player x for player y?

Example fixture

Japan vs Poland - 28/6/18

Outcome	Market	Model
Home win		
Away win		
Draw		

Example fixture

Japan vs Poland - 28/6/18

Outcome	Market	Model
Home win		25.7%
Away win		43.7%
Draw		30.6%

Japan vs Poland - 28/6/18

Outcome	Market	Model
Home win	37.5%	25.7%
Away win	32.0%	43.7%
Draw	30.5%	30.6%

Example fixture

Japan vs Poland - 28/6/18

Outcome	Market	Model
Home win	37.5%	25.7%
Away win	32.0%	43.7%
Draw	30.5%	30.6%

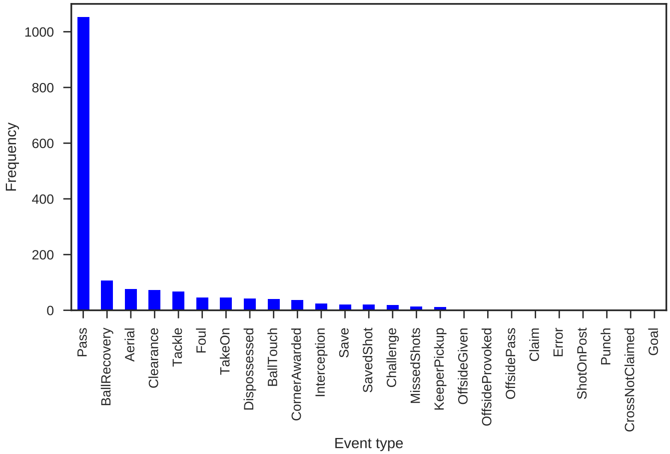
Final score: Japan 0 - 1 Poland

The data

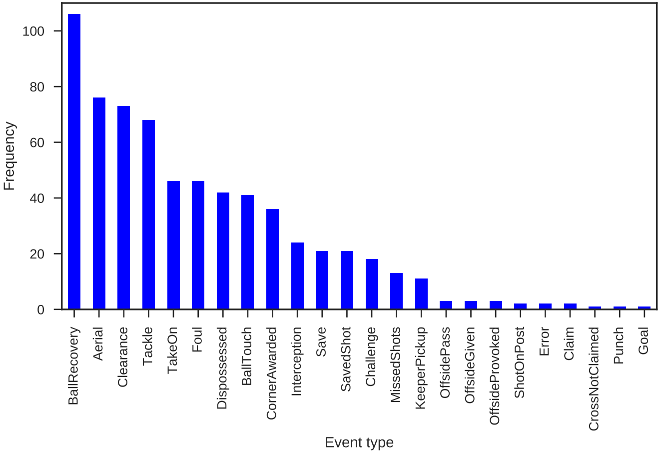
	expanded_minute	minute	second	period	team_id	player_id	type	outcome	x	y	end_x	end_y
2	0	0	1	FirstHalf	663	91242	Pass	Successful	50.1	51.0	53.1	48.7
3	0	0	2	FirstHalf	663	23736	Pass	Successful	53.1	48.7	46.2	54.5
4	0	0	3	FirstHalf	663	17	Pass	Successful	46.2	54.5	32.2	84.4
5	0	0	4	FirstHalf	663	14230	Pass	Successful	32.2	84.4	22.6	61.1
6	0	0	5	FirstHalf	663	7398	Pass	Successful	22.6	61.1	32.0	61.1
7	0	0	6	FirstHalf	663	31451	Pass	Successful	32.2	61.8	22.2	66.0
8	0	0	9	FirstHalf	663	7398	Pass	Successful	22.8	68.1	40.7	73.8
9	0	0	10	FirstHalf	690	38772	Tackle	Successful	60.4	22.8	60.4	22.8
10	0	0	10	FirstHalf	663	80767	Dispossessed	Successful	39.6	77.2	39.6	77.2
11	0	0	12	FirstHalf	690	8505	Pass	Successful	68.8	20.9	73.6	24.7

- Touch data
- 2013/2014 - 2016/2017 Premier league seasons
- Roughly 2.4 million events in total
- $\approx$ 1600 events per game

Liverpool vs Stoke, 17th August 2013



Liverpool vs Stoke, 17th August 2013 (pass removed)



Stop	Control	Disruption	Questionable
Card	Aerial	BlockedPass	CornerAwarded
End	BallRecovery	Challenge	CrossNotClaimed
FormationChange	BallTouch	Claim	KeeperSweeper
FormationSet	ChanceMissed	Clearance	ShieldBallOpp
OffsideGiven	Dispossessed	Interception	
PenaltyFaced	Error	KeeperPickup	
Start	Foul	OffsideProvoked	
SubstitutionOff	Goal	Punch	
SubstitutionOn	GoodSkill	Save	
	MissedShots	Smother	
	OffsidePass	Tackle	
	Pass		
	SavedShot		
	ShotOnPost		
	TakeOn		

The model

- K matches, numbered $k = 1, \ldots, K$
- Set of teams in fixture k is T_k, with T_k^H and T_k^A, $T_k = \{T_k^H, T_k^A\}$
- P is the set of all players, $P_k^j \in P$ is the subset of players who play for team j in fixture k
- Consider how players' abilities over different events interact, group events to create meaningful interactions
- For simplicity lets consider 2 events, e.g. "Pass" and "AntiPass"
- Denote the set of events as $E = \{e_1, e_2\}$

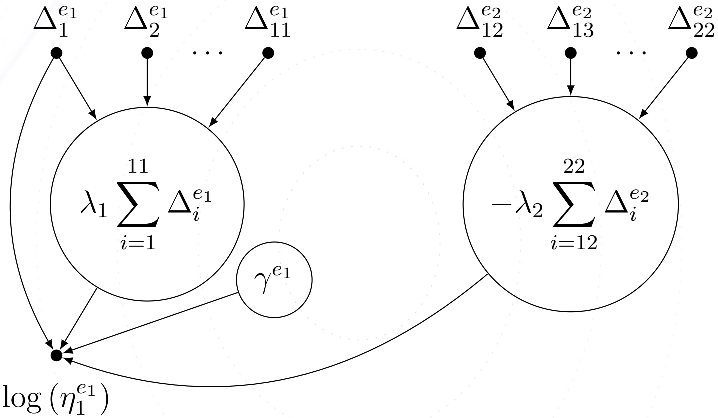
- $X^e_{i,k}$ as the number of occurrences of event e, by player i (for team j), in match k
- We construct a Poisson process

$$X^e_{i,k} \sim Pois\left(\eta^e_{i,k}\tau_{i,k}\right)$$

where

$$\eta^e_{i,k} = \exp\left\{\Delta^e_i + \tau_{i,k}\left(\lambda^e_1 \sum_{i' \in P^j_k} \Delta^e_{i'} - \lambda^e_2 \sum_{i' \in P^{T_k \setminus j}_k} \Delta^{E \setminus e}_{i'}\right) + \left(\delta_{T^H_k, j}\right)\gamma^e\right\}$$

- Δ_i^e is the latent ability for player i for event e
- $\delta_{r,s}$ is the Kronecker delta
- $\tau_{i,k}$ is the fraction of time player i spent on the pitch in match k, $\tau_{i,k} \in [0, 1]$
- γ^e is the home effect for event e
- λ_1^e is the impact of a player's own team
- λ_2^e describing the opposition's ability to stop the player/other team
- We impose the constraint that the λs must be positive



For simplicity we assume only 11 players on each team and drop the time dependence

The log-likelihood is

$$\ell = \sum_{e \in E} \sum_{k=1}^{K} \sum_{j \in T_k} \sum_{i \in P_k^j} X_{i,k}^e \log\left(\eta_{i,k}^e \tau_{i,k}\right) - \eta_{i,k}^e \tau_{i,k} - \log\left(X_{i,k}^e\,!\right)$$

- Large number of players $\rightarrow$ large number of parameters (MCMC not feasible)
- Appeal to variational inference techniques combined with automatic differentiation
- Can utilise a prior for those players with few data points/minutes played

Variational inference

- ▷ See Blei et al. (2017), Kucukelbir et al. (2016), Duvenaud and Adams (2015) and Chapter 19 of Goodfellow et al. (2016)
- ▷ Specify a variational family of densities over the latent variables
- ▷ Latent variables - ν
- ▷ Aim to find best candidate approximation $q(\nu)$
- ▷ Do this by maximising the evidence lower bound (ELBO)
- ▷ ELBO - close relations to the Kullback-Leibler divergence

$$ELBO\left(\nu\right) = E_{\nu}\left[\log\left\{\pi\left(\nu, x\right)\right\}\right] - E_{\nu}\left[\log\left\{q\left(\nu\right)\right\}\right]$$

- ▷ We consider the *mean-field variational family*
- ▷ The latent variables are assumed to be mutually independent

- Let $\nu = \Delta$, and set

$$q\left(\Delta_i^e\right) \sim N\left(\mu_{\Delta_i^e}, \sigma^2_{\Delta_i^e}\right),$$

where

$$q\left(\Delta\right) = \prod_{e \in E} \prod_{j \in T_k} \prod_{i \in P_k^j} q\left(\Delta_i^e\right)$$

- Aim - find suitable candidate values for $\mu_{\Delta_i^e}$ and $\sigma_{\Delta_i^e}$, $\forall i, \forall e$. These are the variational parameters
- Take $(\lambda_1^e, \lambda_2^e, \gamma^e)^T$ to be fixed parameters
- Prior - $\pi(\Delta_i^e) \sim N(-2, 2^2)$

- Fit using automatic differentiation - `Python` package `autograd` (Maclaurin et al., 2015)
- Minimise (negative ELBO) using ADAM

Application - 2013/2014 season

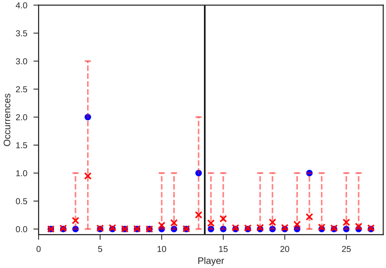
- Initially look at the 2013/2014 English Premier League
  - $k = 1, \ldots, 380$
  - $j \in T_k$ where T_k consists of a subset of $\{1, \ldots, 20\}$
  - $i \in P_k^j$ where P_k^j is a subset of $P = \{1, \ldots, 544\}$
- The full model has 2182 parameters
- Compare "Goal" and "GoalStop"

English Premier League 2013/2014

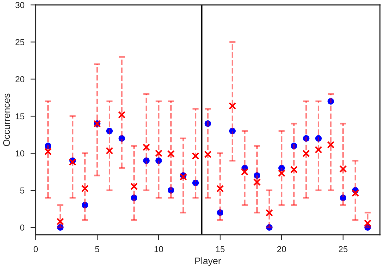
#	Team	Pl	W	D	L	F	A	GD	Pts
1	Manchester City	38	27	5	6	102	37	65	86
2	Liverpool	38	26	6	6	101	50	51	84
3	Chelsea	38	25	7	6	71	27	44	82
4	Arsenal	38	24	7	7	68	41	27	79
5	Everton	38	21	9	8	61	39	22	72
6	Tottenham Hotspur	38	21	6	11	55	51	4	69
7	Manchester United	38	19	7	12	64	43	21	64
8	Southampton	38	15	11	12	54	46	8	56
9	Stoke City	38	13	11	14	45	52	-7	50
10	Newcastle United	38	15	4	19	43	59	-16	49
11	Crystal Palace	38	13	6	19	33	48	-15	45
12	Swansea City	38	11	9	18	54	54	0	42
13	West Ham United	38	11	7	20	40	51	-11	40
14	Sunderland	38	10	8	20	41	60	-19	38
15	Aston Villa	38	10	8	20	39	61	-22	38
16	Hull City	38	10	7	21	38	53	-15	37
17	West Bromwich Albion	38	7	15	16	43	59	-16	36
18	Norwich City	38	8	9	21	28	62	-34	33
19	Fulham	38	9	5	24	40	85	-45	32
20	Cardiff City	38	7	9	22	32	74	-42	30

Within sample predictive distributions

Goal



GoalStop

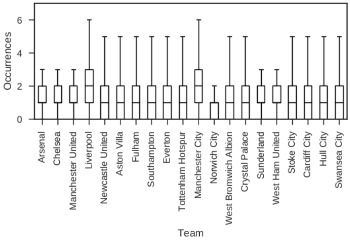


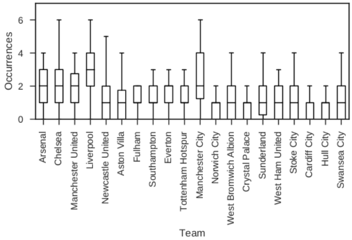
Red: model combinations of $\eta^e_{i,k}$. Blue: observed. The red dotted bars show the 95% prediction interval for each $\eta^e_{i,k}$. The black line separates the players from the two teams

Model

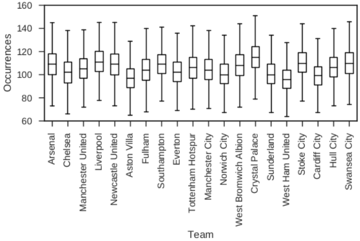
Observed

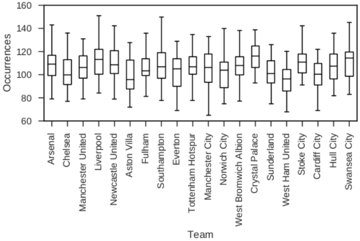
Goal





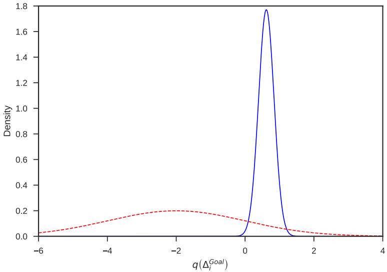
Goal Stop



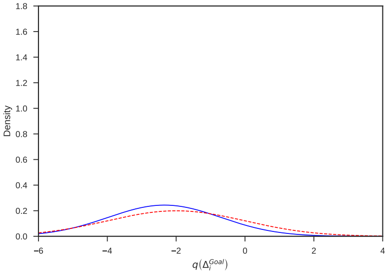


Goal - marginal posterior variational densities

Sturridge



Reed



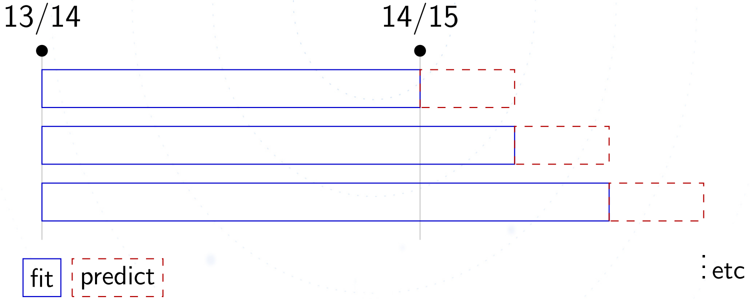
Red-dotted: prior. Blue-solid: posterior

Goal - top 10						
Rank	Player	2.5% quantile	Standard deviation	Observed	Observed rank	Rank difference
1	Suarez	0.508	0.184	31	1	0
2	Sturridge	0.176	0.225	21	2	0
3	Aguero	0.147	0.250	17	4	+1
4	Y. Toure	-0.043	0.224	20	3	-1
5	Rooney	-0.056	0.243	17	5	0
6	Dzeko	-0.065	0.249	16	8	+2
7	van Persie	-0.136	0.289	12	15	+8
8	Remy	-0.230	0.271	14	11	+3
9	Bony	-0.257	0.252	16	7	-2
10	Rodriguez	-0.354	0.263	15	10	0

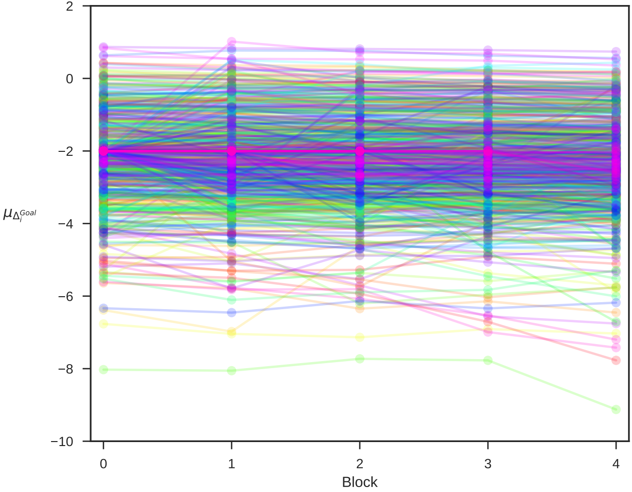
GoalStop - top 10						
Rank	Player	2.5% quantile	Standard deviation	Observed	Observed rank	Rank difference
1	Mulumbu	2.575	0.040	631	1	0
2	Kallstrom	2.553	0.177	33	405	+403
3	Mannone	2.528	0.044	508	12	+9
4	Yacob	2.510	0.053	359	43	+39
5	Tiote	2.474	0.044	517	8	+3
6	Lewis	2.446	0.213	23	436	+430
7	Palacios	2.441	0.101	100	286	+279
8	Jedinak	2.420	0.041	603	2	-6
9	Ruddy	2.411	0.041	600	3	-6
10	Arteta	2.409	0.048	431	21	+11

Application - past/future

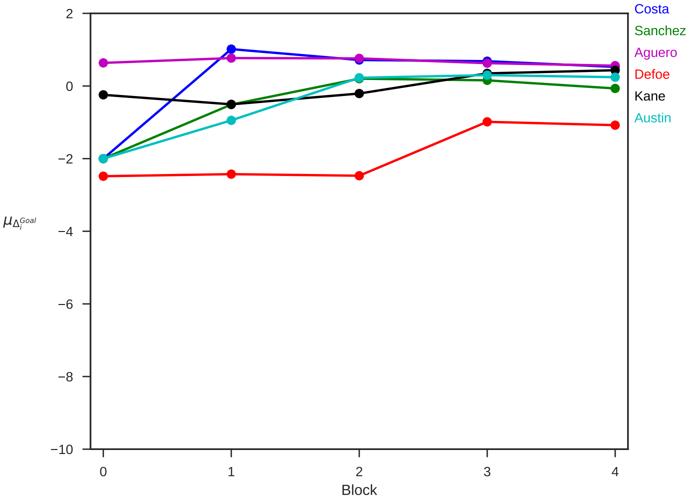
- Now look at 2013/2014 and 2014/2015 English Premier League seasons
- Initial train using complete 2013/2014 season
- Introduce games in blocks of 80 games
- Fit on all the past to predict the future



Goal: 13/14 - 14/15

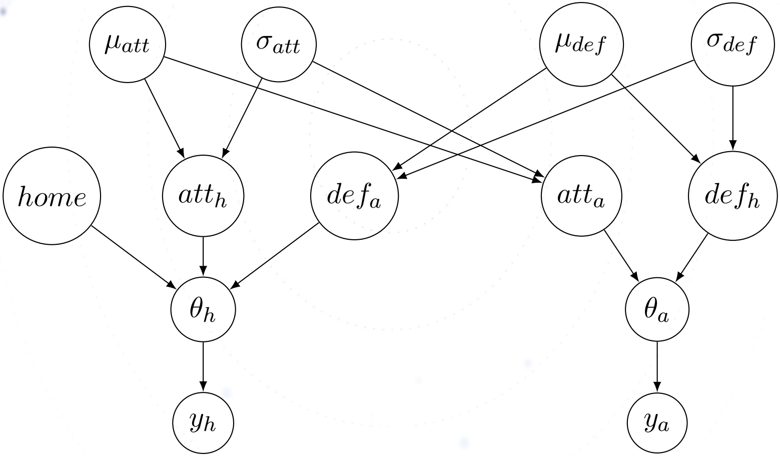


Goal: 13/14 - 14/15

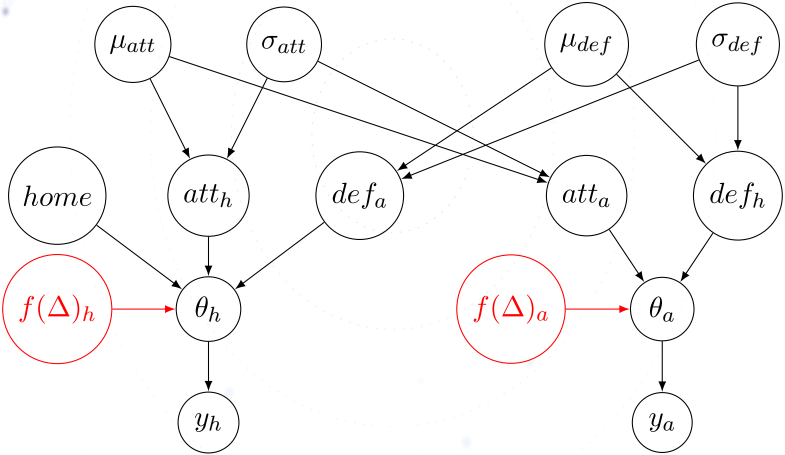


Prediction

- Use these Δs as covariates in a hierarchical Bayesian model, comparing against a baseline model found in Baio and Blangiardo (2010)



- Use these Δs as covariates in a hierarchical Bayesian model, comparing against a baseline model found in Baio and Blangiardo (2010)



- $y \equiv (y_h, y_a) = (\text{home goals, away goals})$

$$y_h|\theta_h \sim Pois\left(\theta_h\right),$$
$$y_a|\theta_a \sim Pois\left(\theta_a\right),$$

$$\log\left(\theta_h\right) = home + att_h + def_a,$$
$$\log\left(\theta_a\right) = att_a + def_h$$

- $att_* \sim N\left(\mu_a, \sigma_a^2\right)$ and $def_* \sim N\left(\mu_d, \sigma_d^2\right)$
- Priors

$$(\mu_a, \mu_d) \sim N\left(0, 10^2\right), \qquad \text{independently}$$
$$(\sigma_a, \sigma_d) \sim Inv\text{-}Gamma(0.1, 0.1), \qquad \text{independently}$$
$$home \sim N\left(0, 10^2\right)$$

- Let p^* be the players which start a game (predicted line up)

$$f(\Delta)_h = \sum_{i' \in p_h^*} \mu_{\Delta_{i'}^{e}} - \sum_{i' \in p_a^*} \mu_{\Delta_{i'}^{E \setminus e}}$$

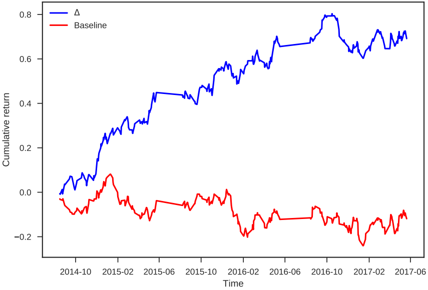
$$f(\Delta)_a = \sum_{i' \in p_a^*} \mu_{\Delta_{i'}^{e}} - \sum_{i' \in p_h^*} \mu_{\Delta_{i'}^{E \setminus e}}$$

  where p_h^* is the home team and p_a^* the away team
- Fit the model using STAN (HMC)
- Fit the model on the past, before predicting on the next set of fixtures
- Use output from hierarchical Bayesian model (θ) to form predictions, e.g. out-of-sample $\Pr(\text{goals} > 2.5)$

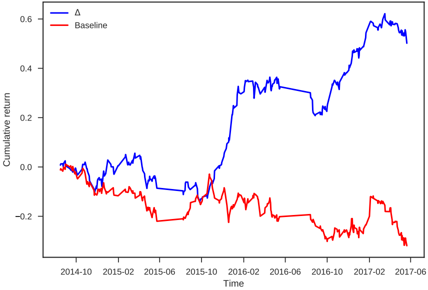
- Use Δs for the abilities
  - Goal
  - Shots
  - Chained Event (our own devising)
- Use 2013/2014 for training only and predict from 2014/2015 onwards using block structure (to end of 2016/2017 season)
- Only predict matches where we have already observed both teams involved (only affects 1st block of each season)
- Bet £100 stake on each game
- Predict
  - Over/under (OU)
  - Asian handicap (AH)

England - Premier League

OU



AH

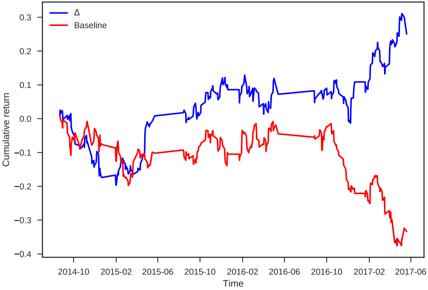


OU, Δ: £6914.09, Baseline: £-1189.51

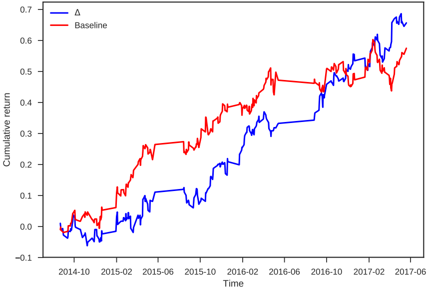
AH, Δ: £5016.87, Baseline: £-3193.15

Germany - Bundesliga

OU



AH

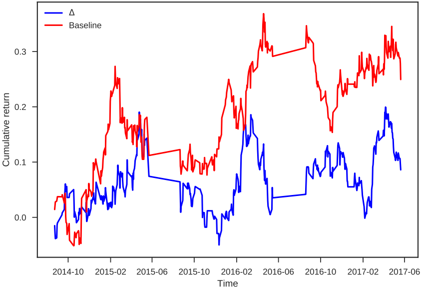


OU, Δ: £2502.33, Baseline: £-3335.97

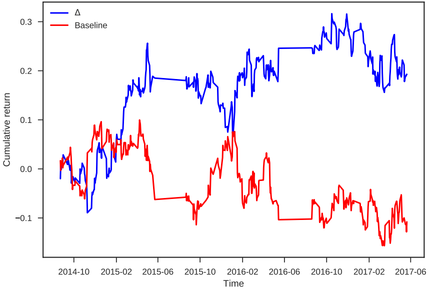
AH, Δ: £6565.97, Baseline: £5749.04

Spain - La Liga

OU



AH

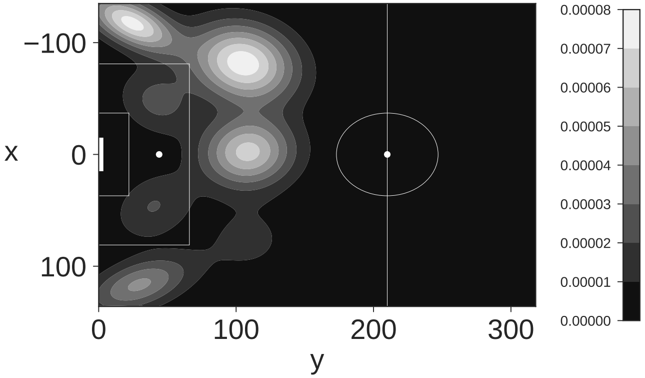


OU, Δ: £859.31, Baseline: £2493.50
AH, Δ: £1927.65, Baseline: £-1079.97

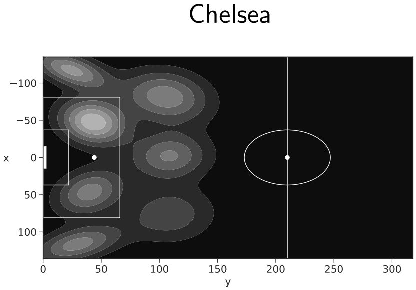
Future work

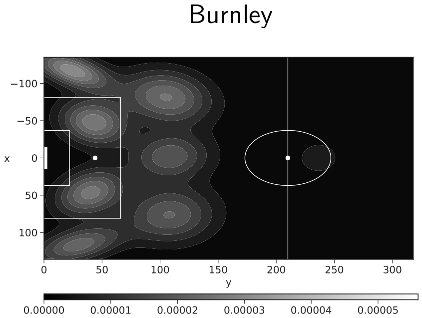
- See if we really are capturing the style of play in each league using Goal, Shots and Chained Event
- What can we introduce to more accurately capture the leagues, e.g. other abilities?
- Look at down-weighting the past
- Possibilities towards a spatial model
- Hope to combine - finding the links between players and how the abilities interact over the links

Eriksen assist locations under a Gaussian mixture model in the 2016/2017 English Premier League, 1^{st} 15 minutes of games



Assist location maps for 2 teams using data until 1^{st} March in 2016/2017 season





References

- Baio, G. and Blangiardo, M. *Bayesian hierarchical model for the prediction of football results.* Journal of Applied Statistics, 37 (2) 253-264,2010
- Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. *Variational inference: A review for statisticians.* Journal of the American Statistical Association, 2017
- Duvenaud, D. and Adams, R. P. *Black-box stochastic variational inference in five lines of python.* NIPS Workshop on Black-box Learning and Inference, 2015
- Franks, A., Miller, A., Bornn, L. and Goldsberry, K. *Characterizing the spatial structure of defensive skill in professional basketball.* The Annals of Applied Statistics, 9 (1) 94-121, 2015
- Goodfellow, I., Bengio, Y. and Courville, A. *Deep Learning.* MIT Press, 2016.
- Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. and Blei, D. M. *Automatic differentiation variational inference.* Journal of Machine Learning Research, 18 (14) 1-45, 2017

- Maclaurin, D., Duvenaud, D. and Adams, R. P. *Autograd: Effortless gradients in numpy.* ICML 2015 AutoML Workshop, 2015
- Miller, A., Bornn, L., Adams, R. and Goldsberry, K. *Factorized point process intensities: A spatial analysis of professional basketball.* Proceedings of the 31st International Conference on Machine Learning - Volume 32, 235-243, 2014
- Saul, L. and Jordan, M. I. *Exploiting tractable substructures in intractable networks.* Advances in Neural Information Processing Systems 8, MIT Press, 486-492, 1996
- Whitaker, G. A., Silva, R., Edwards, D. *A Bayesian inference approach for determining player abilities in soccer.* arXiv preprint, https://arxiv.org/abs/1710.00001.pdf, 2017
- Whitaker, G. A., Silva, R., Edwards, D. *Modeling goal chances in soccer: a Bayesian inference approach.* arXiv preprint, https://arxiv.org/abs/1802.08664.pdf, 2018