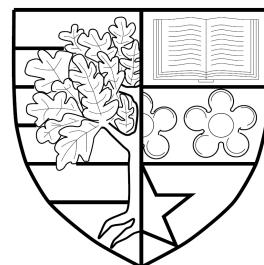


**STATISTICAL MODELLING AND BAYESIAN
INFERENCE FOR MATCH OUTCOMES AND TEAM
BEHAVIOUR IN ASSOCIATION FOOTBALL**

by
Jeffrey Pollock



Submitted for the degree of
Doctor of Philosophy

DEPARTMENT OF ACTUARIAL MATHEMATICS & STATISTICS
SCHOOL OF MATHEMATICAL AND COMPUTER SCIENCES
HERIOT-WATT UNIVERSITY

September 2015

The copyright in this thesis is owned by the author. Any quotation from the report or use of any of the information contained in it must acknowledge this report as the source of the quotation or information.

Abstract

This thesis presents advances in modelling and inference for match outcomes in the association football English Premier League. We firstly extend earlier models by introducing a behavioural aspect which can be used to investigate how teams react to the state of play in a match. We show that the model, in its simplest form, outperforms existing models and is able to select a portfolio of profitable bets against a bookmaker. Secondly, we introduce a dynamic component to the model by allowing team ability parameters to vary stochastically in time. We employ particle filtering methods to cope with a mixture of static and dynamic parameters and find that the updating of posterior distributions is particularly fast, a necessary attribute should we wish to update parameter estimates while matches are in-play. Furthermore, it is shown that the methods are able to recover model parameters based on simulated league data. Finally, we propose an extension to the model so that we are able to investigate how a team modifies its behaviour based on their league situation. We consider league positions that are closely attainable and suggest that since teams modify their behaviour based on their current league position, outcomes of different matches are not necessarily independent.

Acknowledgements

I'd like to thank my supervisors, Professor Gavin Gibson and Doctor George Streftaris, from whom I have learned much during my PhD.

On a more personal level, I'd like to thank my mum Margaret, my sister Emily, and my partner Lavinia, for pretending to listen to me talk about a mixture of football and statistics for the past three years.

Contents

1	Introduction	1
1.1	An overview of gambling in sport	1
1.2	Gambling in association football	3
1.3	Bookmaking	5
1.4	The English Premier League	6
1.5	Statistical models in association football	8
1.6	Data	10
1.7	Aim and structure of the thesis	12
2	Bayesian computational methods	15
2.1	Introduction	15
2.2	The Bayesian approach to inference	15
2.2.1	Bayesian versus Frequentist	17
2.2.2	The non-informative prior	18
2.3	MCMC	19
2.3.1	The Metropolis-Hastings algorithm	22
2.3.2	The Gibbs sampler	24
2.3.3	Burn-in, thinning, and convergence diagnostics	26
2.4	RJMCMC	29
2.5	Particle filtering methods	30
2.6	Bayesian model choice	32
2.6.1	Bayesian model averaging	33
2.6.2	The Bayes factor	33
2.6.3	The Jeffreys-Lindley paradox	34
2.6.4	The Deviance information criterion	36
2.6.5	The posterior predictive distribution and scoring rules	37
3	An adaptive behaviour model for association football using rankings as prior information	40
3.1	Introduction	40
3.2	Association football models	41
3.2.1	The model of Dixon and Robinson	41

3.2.2	The Bradley-Terry model	42
3.2.3	A new non-homogeneous Poisson process model	42
3.3	Bayesian inference for parameter estimation	44
3.3.1	The log-likelihood	44
3.3.2	Simulation of goal times	46
3.3.3	Prior choice	48
3.3.4	A prior for R using ranking information	49
3.3.5	Inference results	54
3.3.6	Goodness of fit	64
3.4	Model comparisons	66
3.4.1	Comparison using a scoring rule	67
3.4.2	A Hosmer-Lemeshow type test	70
3.4.3	Use of models to inform betting strategies	72
3.5	Concluding remarks	75
4	Fast updating of dynamic and static parameters using particle filters	76
4.1	Introduction	76
4.2	Particle filtering methods	77
4.2.1	The bootstrap filter	77
4.2.2	General particle filter	78
4.2.3	Auxiliary particle filter	79
4.2.4	Practical implementation	80
4.2.5	Resampling methods	80
4.3	A mixture of dynamic and static parameters	81
4.3.1	Artificial evolution	83
4.3.2	Kernel smoothing	83
4.3.3	Kernel smoothing methods for variance reduction in artificial evolution	85
4.3.4	An example of variance reduction in artificial evolution	89
4.4	An updated model	89
4.4.1	Updated model inference	91
4.4.2	Finding the optimal variance	92
4.4.3	Inference results	95
4.4.4	Model and inference performance	100
4.4.5	Combining MH and particle filtering methods	104
4.4.6	Inference using simulated data	104
4.5	Concluding remarks	111
5	Incorporating league position into team behaviour	113
5.1	Introduction	113

5.2	Association football models using league information	114
5.3	A non-homogeneous Poisson process model using a concept of utility	116
5.4	Data	119
5.5	Four competing models	120
5.5.1	Prior choice	121
5.5.2	Results of model fitting	123
5.5.3	Model choice using DIC	128
5.6	Model inference using RJMCMC	129
5.6.1	Knot birth	130
5.6.2	Knot death	131
5.6.3	Prior choice	133
5.6.4	Model inference results	135
5.7	An exemplar match	138
5.8	Concluding remarks	141
6	Conclusion	143
6.1	Results and discussion	143
6.2	Future work	145
Bibliography		148

List of Tables

1.1	The league table at the end of the EPL 2011/2012 season	7
2.1	An interpretation of Bayes factors from Kass and Raftery (1995)	34
3.1	A summary of posterior estimates for the 10 non-resource model parameters	61
3.2	Potential scale reduction factors from Gelman-Rubin's convergence diagnostic obtained from three chains	63
3.3	A comparison of the four competing models in terms of the sum of the logarithmic scoring rule and the geometric mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed, for weeks 6 to 38	67
3.4	The results of the hypothesis test H_0 : the bookmaker's estimated probabilities are correct, against H_1 : the use of model probabilities gives the bettor an advantage over the bookmaker	73
4.1	A comparison of the sum of the logarithmic scoring rule, the geometric mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed, and the betting profit, for models M and DM for weeks 6 to 38	100
5.1	A summary of the different knot positions in each of the four models	120
5.2	The means of the Γ prior density for the utility value at the knot positions p used within models m_0 , m_1 , m_2 , and m_3	121
5.3	p_D and DIC for the four competing models	128
5.4	The means of the Γ prior density for the utility value at each possible knot position	133

5.5 The league table before the last match of the EPL 2011/2012 season. ‘p’ denotes the league position, ‘pld’ the number of matches played (note this is one less than the week of the season), ‘w’ the number of matches won, ‘d’ the number of matches drawn, ‘l’ the number of matches lost, ‘gf’ the number of goals for (scored), ‘ga’ the number of goals against (conceded), ‘gd’ the goal difference (‘gf’ - ‘ga’), and ‘pts’ the league points	140
--	-----

List of Figures

1.1	A screenshot of the Betfair exchange showing available back and lay odds for the runners of the 14:30 at Royal Ascot	2
1.2	A screenshot from the website of one of the top UK bookmakers William Hill, showing the odds for some of the markets available on an international match between Peru and Venezuela	4
1.3	A bar plot of goal times (in minutes) for the EPL season 2011/2012 .	11
1.4	A stacked bar chart showing the value of $1/d_{m,H} + 1/d_{m,D} + 1/d_{m,A}$ from the Bet365 1X2 betting market for all 380 matches (m) in the 2011/2012 season. ■ $1/d_{m,H}$, ■ $1/d_{m,D}$, ■ $1/d_{m,A}$	13
3.1	Marginal plots of samples from R_k for Manchester United, Arsenal, Everton, and Wolverhampton Wanderers	52
3.2	Bivariate plots of samples from \mathbf{R} plotted using hexagonal bins. The darker red hexagons represent a higher count of samples in that bin .	53
3.3	Density histograms of the posterior samples for $e_i - 0.5(c_i + d_i)$ for $i = -1$ (top), $i = 0$ (middle), and $i = 1$ (bottom)	55
3.4	Transparent plots of 1,500 posterior samples from the functions $\alpha_k(t)$ (top), $\alpha_k^{(1)}(t)$ (middle), and $\alpha_k^{(2)}(t)$ (bottom) for the states losing, drawing, and winning. — the posterior mean	56
3.5	A violin plot of the marginal posterior distribution of the teams' resources (R_k for team k). The teams have been ordered by the posterior mean of their resource	57
3.6	A plot displaying the posterior distribution of a ranking of the teams based on the parameter vector \mathbf{R} . Darker regions represent areas of higher posterior probability and again, the teams have been ordered by the posterior mean of their resource. • denote the final league position of the teams in the 2011/2012 season	58
3.7	Trace plots and density histograms of the posterior samples for parameters h , a , c_{-1} , c_0 , and c_1	59
3.8	Trace plots and density histograms of the posterior samples for parameters d_{-1} , d_0 , d_1 , ρ_1 , and ρ_2	60

3.9	A plot of $\lambda_m(t)$ (—) and $\mu_m(t)$ (---) for match m where Manchester City (i) played Queens Park Rangers (j). — denotes goal times, the resulting score in the form $i - j$ is annotated	62
3.10	A density histogram of the simulated χ_i^2 statistics. — the overall χ^2 statistic	65
3.11	A density histogram of $\sum_{w=6}^{38} LSR_w$ under H_0 : the distribution of LSR_w given D_{w-1} is approximately equal for all four models each week w . Individual model estimates of $\sum_{w=6}^{38} LSR_w$ are denoted by — M, — DR, — BTC, — BTB	68
3.12	A density histogram of $\sum_{w=6}^{38} LSR_w$ under H_0 : the distribution of LSR_w given D_{w-1} is approximately equal for the three best performing models each week w . Individual model estimates of $\sum_{w=6}^{38} LSR_w$ are denoted by — M, — DR, — BTB	69
3.13	A plot of $O_{\mathcal{M}_{I,E}}$ and a 95% prediction interval around $\mathbb{E}(X_{\mathcal{M}_{I,E}})$ for the events H (top), D (middle), and A (bottom). In each probability interval I , $O_{\mathcal{M}_{I,E}}$ is denoted by (•), (•), (•), and (•) for models M, DR, BTC, and BTB respectively, the 95% prediction interval follows similarly	71
3.14	A plot of the observed profit $O_{P,r}$ for varying levels of r and when using estimated probabilities from each of the four models. — M, — DR, — BTC, — BTB	74
4.1	An example of residual resampling to gain a sample from a $N(0, 1)$ distribution. ■ particles which counts have been deterministic from the floor of the particle count expectation, ■ particles which have been included from the multinomial resampling step	82
4.2	Smooth kernel density estimates when $h = 0.1$ (top), $h = 0.5$ (middle), and $h = 0.9$ (bottom). — the kernel density estimate with no shrinkage, --- the kernel density estimate with shrinkage	86
4.3	Example of the Auxiliary Particle Filter at time 1 (top), 10 (middle), 20 (bottom). The histogram represents the particle posterior approximation of the static parameter μ . — the prior density, — the theoretical posterior	90
4.4	Values of the metric $GM_{1,38}$ for different values of σ^2 in the range 0 to 0.5 (top) and 0 to 0.02 (bottom). — the calculated values, --- a smooth LOESS estimate	94
4.5	Time series plot of parameters h, a, c_{-1}, c_0, c_1 , and d_{-1} . — posterior mean, --- 95% BCI	95
4.6	Time series plot of parameters d_1, d_0, ρ_1, ρ_2 and the log-resource team parameters. — posterior mean, --- 95% BCI	96

4.7	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI	97
4.8	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI	98
4.9	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI	99
4.10	Scatter plot of the home win (top), draw (middle), and away win (bottom) probabilities predicted by the models M and DM for 330 matches from weeks 6 to 380. -··- the line $y = x$, — a smooth LOESS estimate	101
4.11	The cumulative time to take 100,000 samples from the posteriors $p(\boldsymbol{\theta}_M \mathbf{D}_w)$ (—) and $p(\boldsymbol{\theta}_{DM} \mathbf{D}_w)$ (—) for weeks $w = 1, \dots, 38$	103
4.12	Values of $GM_{1,38}$ corresponding to different values of σ^2	105
4.13	Time series plot of parameters h , a , c_{-1} , c_0 , c_1 , and d_{-1} . — posterior mean, -·- 95% BCI, — the true parameter value	106
4.14	Time series plot of parameters d_0 , d_{-1} , ρ_1 , ρ_2 , and the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value	107
4.15	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value	108
4.16	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value	109
4.17	Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value	110
5.1	An illustration of how a utility function which values each league position might look. Knot locations are denoted by •	117
5.2	A plot of the prior density of the utility value for the possible knot locations. — position 1, — position 2, — position 4, — position 5, — position 6, — position 17, — position 18	122
5.3	A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_0 . — the posterior mean	124
5.4	A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_1 . — the posterior mean	125
5.5	A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_2 . — the posterior mean	126

5.6	A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_3 . — the posterior mean	127
5.7	The utility function given by model m (top), the proposal of a new knot in position 10 (middle), the resulting utility function of model m' (bottom). • utility function knot locations, --- the superimposed proposal density	132
5.8	The utility function given by model m' (top), the proposal of the death of the knot in position 3 (middle), the resulting utility function of model m (bottom). • utility function knot locations	134
5.9	A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) using RJMCMC. — the posterior mean	136
5.10	A bar plot of the prior and posterior probabilities for the total number of knots (top). A bar plot of the posterior probability of a knot in each position (bottom). ■ the prior probabilities, ■ the posterior probabilities	137
5.11	A plot of the rates of scoring ($\lambda_m(t)$ and $\mu_m(t)$) (top) and the resource allocation $\alpha_k(t, w)$ (bottom) for match m where Manchester City played at home to Queens Park Rangers. — denotes Manchester City, -·- denotes Queens Park Rangers, — goals scored in this match, --- goals scored in other concurrent matches	139

Glossary

AIC Akaike Information Criterion.

BCI Bayesian Credible Interval.

BIC Bayesian Information Criterion.

DIC Deviance Information Criterion.

EPL English Premier League.

IID Independent Identically Distributed.

LOESS Local Regression.

MCMC Markov Chain Monte Carlo.

MH Metropolis-Hastings.

MLE Maximum Likelihood Estimate.

PDF Probability Density Function.

RJMCMC Reversible Jump Markov Chain Monte Carlo.

SIR Sequential Importance Resampling.

SIS Sequential Importance Sampling.

SMC Sequential Monte Carlo.

SPL Scottish Premier League.

WAIC Widely Applicable Information Criterion.

Chapter 1

Introduction

1.1 An overview of gambling in sport

Gambling has been a hugely popular activity throughout human life, and has recorded mentions in ancient Roman and Greek history. While the practice of gambling has been frowned upon or even illegal in certain periods of time or locations over the globe, in recent times in the UK gambling (under regulation) has been widely accepted by the public. Most UK high streets now contain at least one bookmaking office (shop) and even more gambling is done online. Horse racing and association football (commonly referred to as *soccer* in the United States and football elsewhere) are the most gambled upon sports in the UK, with sports like tennis and greyhound racing also being popular.

After a government go-ahead in 1960, UK bookmaking offices began to open in 1961 with the aim of ending unregulated, illegal gambling. Before then, bets could only be legally placed at racing tracks. The number of bookmaking offices quickly grew, and in 2013 was estimated at around 8,700 (Barford and Judah (2013)). It is however thought that the number of bookmaking offices is declining due to the popularity of on-line gambling.

Founded in 2000, Betfair (<https://www.betfair.com/exchange>) offered the first on-line sports betting exchange. It allowed bettors to act either as a traditional bettor, ‘backing’ an event, or act in the traditional role of the bookmaker, ‘laying’ an event. This allowed bettors to bet against each other and thus offer better odds than a bookmaker. Bookmakers typically worked an expected profit of around 15% into their odds, whereas Betfair would only take a small percentage (around 5%) of any winnings. An example of the Betfair exchange can be seen in Figure 1.1, which shows the market of a race at the hugely popular Royal Ascot. The Figure shows the odds which are available to back (in bold under ‘Back all’) and the amount of money which is in the exchange at those odds (directly underneath the odds).

14:30 Royal Ascot
Fri 19 Jun | 6f Grp 3

Live Stream Radio
Form & Results
Tote Multiples

Win	Each Way	Place	4 Places	More ▾			
<input checked="" type="checkbox"/> Going In-Play <input type="checkbox"/>	<input checked="" type="checkbox"/> Rules <input type="checkbox"/>				Matched: GBP 575,652	<input type="button" value="Refresh"/>	
<input type="checkbox"/> Betfair SP [?]	<input type="checkbox"/> Timeform						
18 selections							
					Back all	Lay all	
9 (13)	Illuminate Richard Hughes	5.2 £2403	5.3 £2549	5.4 £1293	5.5 £657	5.6 £1161	5.7 £1706
11 (5)	Laxfield Road Frankie Dettori	7.6 £1002	7.8 £1337	8 £673	8.2 £787	8.4 £224	8.6 £228
3 (12)	Back At The ... Joel Rosario	8 £3432	8.2 £1360	8.4 £451	8.6 £760	8.8 £352	9 £311
16 (8)	Spanish Ro... Cristian Demuro	10 £623	10.5 £602	11 £2709	11.5 £548	12 £808	12.5 £254
18 (19)	Tutu Nguru Pat Cosgrave	11.5 £304	12 £336	12.5 £729	13 £614	13.5 £1144	14 £125
14 (2)	Our Joy Oisin Murphy	11.5 £223	12 £568	12.5 £634	13 £566	13.5 £697	14 £792
1 (18)	Ashadihan Jamie Spencer	14.5 £147	15 £683	15.5 £12	16 £196	16.5 £227	17 £165
10 (17)	Jersey Breeze Charles Bishop	16.5 £151	17 £452	17.5 £116	18 £216	18.5 £103	19 £75
2 (16)	Azhar James Doyle	21 £161	22 £311	23 £169	24 £170	25 £154	26 £95
4 (10)	Elegant Supe... Antoine Hamelin	29 £162	30 £238	32 £192	34 £117	36 £111	38 £30
5 (7)	Fireglow William Buick	28 £54	29 £25	30 £78	32 £43	34 £86	36 £55
15 (14)	Palenville Pat Dobbs	32 £51	34 £242	36 £152	38 £157	40 £83	42 £44
6 (11)	First Party Joe Fanning	42 £67	44 £83	46 £27	50 £27	55 £12	70 £19
13 (4)	Miss Moneyp... Richard Kings...	34 £77	36 £47	38 £163	40 £43	42 £117	44 £33
.. 10	Vallance Road	55	60	65	70	75	85

Figure 1.1: A screenshot of the Betfair exchange showing available back and lay odds for the runners of the 14:30 at Royal Ascot

Similarly, the odds which you can lay are shown. The screenshot was taken about an hour before the start of the race, and already the exchange had matched bets up to a total value of £575,652. It is not uncommon to see several millions matched on big horse racing or association football markets.

Somewhat in response to Betfair, high street bookmakers had to invest in websites which easily allowed on-line gambling (as we will see in the forthcoming section) and also operated at a lower expected profit, in order to offer competitive odds. This led to a massive increase in turnover for bookmakers as many people are not comfortable with the Betfair exchange and prefer a simpler betting approach with a bookmaker. This also suggests an additional opportunity for bettors, betting against high street bookmakers who now offer better odds.

1.2 Gambling in association football

The first popular association football bet in the UK was known as the ‘football pools’. Littlewoods football pools was the first of its kind, beginning in 1923 when the football pools coupons were offered outside Manchester United’s ground Old Trafford. The bet quickly spread across the whole of the UK, probably because, for a small stake the bet offered the chance of a share of a massive jackpot (pool of money). The aim of the bet was to select the outcome of several matches which at the time, were all played concurrently at 3pm on a Saturday. More recently, since association matches are televised, matches occur at different times during the week, although the bulk are played in the traditional 3pm Saturday slot. At its peak of popularity, it is estimated around that the pools had around 10 million players in the UK, this figure however severely plummeted following the introduction of the UK national lottery in 1994, which offered even bigger jackpots. It is also likely that bettors now favour association football bets with bookmakers, who can offer odds on a much larger selection of events.

Bookmakers now offer a large number of association football betting markets to customers. This is no real surprise since to quote Constantinou et al. (2012), ‘[association football] is the world’s most popular sport … and constitutes the fastest growing gambling market’. Some of the most popular betting markets are: ‘1X2’ (also called match betting), in which the bettor chooses the final outcome of the match (home team win, draw, or away team win), ‘total goals under/over’, in which the bettor chooses if the total number of match goals will be under or over a certain line (usually 2.5), and ‘correct score’, in which the bettor chooses the exact final score in the form x-y. There are also more obscure betting opportunities, for example betting on the number of corners in the last 15 minutes of a match. In fact, as can be seen in Figure 1.2, bookmakers such as William Hill

(<http://sports.williamhill.com/bet/en-gb>) may offer in hundreds of betting markets (218 for this particular example) on an association football match.

+ My Markets (0)	All Markets (218)	5 Minutes (7)	Scorer and Corners (37)	
Goals (56)	Goal Time (26)	Cards (34)	Other Markets (17)	
▼ Match Betting				
Peru	2.80	Draw	3.10	
▼ Match Result and 4 or More Goals in The Match				
Peru	11.00	Draw	9.00	
Venezuela		Venezuela	11.00	
▼ Match Result and Both Teams To Score				
Peru	6.00	Draw	4.00	
Venezuela		Venezuela	5.50	
▼ Headline Offers				
Jose Paolo Guerrero To Score And Peru To Win	5.50	Salomon Rondon To Score And Venezuela To Win	7.00	
Bet Void If Player Doesn't Take Part				
▼ Correct Score				
Peru 1-0	8.00	Draw 0-0	8.50	
Venezuela 1-0		Venezuela 1-0	7.50	
Peru 2-0	13.00	Draw 1-1	6.50	
Venezuela 2-0		Venezuela 2-0	13.00	
Peru 2-1	11.00	Draw 2-2	17.00	
Venezuela 2-1		Venezuela 2-1	10.00	
Peru 3-0	29.00	Draw 3-3	67.00	
Venezuela 3-0		Venezuela 3-0	26.00	
Peru 3-1	26.00	Draw 4-4	201.00	
Venezuela 3-1		Venezuela 3-1	23.00	
Peru 3-2	41.00		Venezuela 3-2	41.00
Peru 4-0	81.00		Venezuela 4-0	67.00
Peru 4-1	67.00		Venezuela 4-1	67.00
Peru 4-2	101.00		Venezuela 4-2	81.00
Peru 4-3	151.00		Venezuela 4-3	126.00
Peru 5-0	251.00		Venezuela 5-0	201.00
Peru 5-1	201.00		Venezuela 5-1	201.00
Peru 5-2	251.00		Venezuela 5-2	251.00
Peru 6-0	501.00		Venezuela 6-0	501.00
Peru 6-1	501.00		Venezuela 6-1	501.00
Peru 6-2	501.00		Venezuela 6-2	501.00
			Venezuela 7-1	501.00

Figure 1.2: A screenshot from the website of one of the top UK bookmakers William Hill, showing the odds for some of the markets available on an international match between Peru and Venezuela

There has also been a recent increase in what is known as ‘in-play’ betting, where bets can be placed during a match. Previously, as soon as a match had started, the betting markets would suspend, however with in-play betting this is no longer the case. In-play betting markets see the odds change continuously throughout time in a match, and react to events such as goals or player dismissals.

1.3 Bookmaking

Bookmakers typically estimate the probability of an event (for example a score of 3-0 or a home team win), and then add what is called ‘over-round’ in order to ensure an expected net profit from their published odds - effectively worsening the fair odds. Odds may be offered in various forms (for example fractional or American) but we consider the odds in decimal form, which are the most natural to work with mathematically. Decimal odds are given by:

$$d_{m,E} = \frac{1}{p_{m,E} + o_{m,E}} \quad (1.1)$$

where $d_{m,E}$ is the bookmaker’s decimal odds, $p_{m,E}$ is the bookmaker’s estimated probability, and $o_{m,E}$ is the added over-round, for event E in match m . From a bettor’s point of view, a 1-unit bet at decimal odds d results in either a loss of 1 unit, or a profit of $d - 1$ units.

In the past, bookmakers would devise odds based on their expert knowledge of sport and any bets they had already taken. For example they might begin with their best guess of the statistically profitable odds (which include over-round), then as bets are placed, shorten (decrease) the odds on events which have been heavily bet (discouraging future bets) and lengthen (increase) the odds on events which have not been bet so much (encouraging future bets). This strategy aims to keep a ‘balanced book’, that is, a low-risk spread of bets taken over all events which ideally results in arbitrage for the bookmaker, and largely no cases where the bookmaker has a large sum to pay should a particular event occur. The strategy is in a sense Bayesian in flavour since the Bookmaker first assigns some prior odds based on expert knowledge, and then updates the odds based on new information (bets) becoming available.

We note two points from the above bookmaking strategy, firstly, it is very labour intensive, requiring expert knowledge in both the prior setting of the odds and any movement of the odds as bets are placed. Secondly, the bookmaker moves odds based on his current book of bets and not necessarily on the true underlying probability of an event. Therefore, should a bookmaker take a lot of bets on a particular event, for example a UK bookmaker taking bets on England to win the association football World Cup, then the odds on that particular event may be particularly bad value and there may be scope for good-value, long-term profitable betting strategies focused on less fancied teams within the betting market.

The first point mentioned above has been a real problem for bookmakers, and has been remedied in recent years by the use of statistical models which can calculate odds for hundreds of betting market events. They are most prevalent for in-play betting markets, since theoretically the odds should change continuously throughout

time and a bookmaker who was manually trading may not be able to keep up with the ever changing market. Furthermore, if the bookmaker believes they have a good statistical model, they may choose not to change the odds based on incoming bets, instead opting for a high-risk, high-rate-of-return strategy.

The second point mentioned above is mainly of interest to bettors, and suggests the use of statistical models to bet against the bookmaker, of which we present an example in Chapter 3 Section 3.4.3.

In practice, it is likely that different bookmakers opt for different strategies with regards to the use of statistical models and movement of odds based on bets. However, statistical models have considerably reduced the amount of manual labour required for a bookmaker to offer a large number of betting markets - as we will discuss further in Section 1.5.

1.4 The English Premier League

The English Premier League (EPL) is one of the top association football leagues and attracts interest internationally. At the end of the 2011/2012 season over £1.1bn was paid out to the 20 participating teams in broadcast money alone (www.premierleague.com). We thus choose this league as a test system for our subsequent statistical models.

Each EPL season involves 20 teams playing against each other twice, once at home and once away, giving a total of 380 matches which are played over 38 weeks (10 matches per week). Teams are awarded three points for a win, one point for a draw, and zero points for a loss, goals scored, conceded, and difference are also recorded for calculation of the league standings. Teams are first ranked on points, then goal difference, and then goals scored. Ending the season in positions 1-4 grants qualification to what is known as the Champions League, a short tournament which contains the top teams in Europe, similarly, position 5 grants qualification to the Europa League. At the other end of the table, teams in positions 18-20 are relegated to the league below (known as the Championship), while the top 3 teams from the Championship are promoted into the EPL. The league table at the end of season 2011/2012 is shown in Table 1.1 where ‘p’ denotes the league position, ‘pld’ the number of matches played, ‘w’ the number of matches won, ‘d’ the number of matches drawn, ‘l’ the number of matches lost, ‘gf’ the number of goals for (scored), ‘ga’ the number of goals against (conceded), ‘gd’ the goal difference (‘gf’ - ‘ga’), and ‘pts’ the league points. At the end of the 2010/2011 season (the previous season) Norwich City, Swansea City, and Queens Park Rangers were promoted into the EPL, replacing Birmingham City, Blackpool, and West Ham United.

p	team	pld	w	d	l	gf	ga	gd	pts
1	Manchester City	38	28	5	5	93	29	+64	89
2	Manchester United	38	28	5	5	89	33	+56	89
3	Arsenal	38	21	7	10	74	49	+25	70
4	Tottenham Hotspur	38	20	9	9	66	41	+25	69
5	Newcastle United	38	19	8	11	56	51	+5	65
6	Chelsea	38	18	10	10	65	46	+19	64
7	Everton	38	15	11	12	50	40	+10	56
8	Liverpool	38	14	10	14	47	40	+7	52
9	Fulham	38	14	10	14	48	51	-3	52
10	West Bromwich Albion	38	13	8	17	45	52	-7	47
11	Swansea City	38	12	11	15	44	51	-7	47
12	Norwich City	38	12	11	15	52	66	-14	47
13	Sunderland	38	11	12	15	45	46	-1	45
14	Stoke City	38	11	12	15	36	53	-17	45
15	Wigan Athletic	38	11	10	17	42	62	-20	43
16	Aston Villa	38	7	17	14	37	53	-16	38
17	Queens Park Rangers	38	10	7	21	43	66	-23	37
18	Bolton Wanderers	38	10	6	22	46	77	31	36
19	Blackburn Rovers	38	8	7	23	48	78	-30	31
20	Wolverhampton Wanderers	38	5	10	23	40	82	-42	25

Table 1.1: The league table at the end of the EPL 2011/2012 season

Each match in the EPL consists of two 45-minute halves. However, at the end of each half the referee may allow the match to continue for longer than the allocated 45 minutes in what is known as ‘injury time’. Injury time typically adds around 2 minutes to the first half, and 3.5 minutes to the second half.

As touched upon in Section 1.2, each week the 10 EPL matches were traditionally all played at the same time of 3pm on a Saturday. This is however no longer the case, with television companies demanding that matches be played at different times so more can be televised. The tradition however holds on the last set of 10 matches in week 38, which often sees teams concurrently playing to avoid league relegation, or at the other end of the league table, for league victory. Furthermore, as information now flows quickly between matches, at all times teams are aware of the state of other concurrent games which may affect them.

1.5 Statistical models in association football

Coinciding with the rise in on-line and in-play betting has been a demand for statistical models which can capture the details of sporting contests. As seen in Section 1.3, bookmakers offer hundreds of markets and thus require a statistical model which is capable of predicting the probability of each event in each market, both before and during a match. Statistical models may also be used by bettors, to inform their own betting strategy against bookmakers, or other bettors on an exchange like Betfair.

Much of the published work on this area typically extends a model proposed by Maher (1982) who suggested modelling the numbers of goals scored by the home and away teams goals as independent Poisson random variables, where the mean of each distribution depends on the strength of the home and away team’s attacking and defensive capabilities. Dixon and Coles (1997) proposed an adjustment to the probability of certain match scores in the independent Poisson model and also suggested that a model’s forecasting ability improved if the likelihood incorporated weightings that decreased exponentially with the time elapsed since the observation of an event. Karlis and Ntzoufras (2003) built on these ideas with a diagonally inflated bivariate Poisson model for modelling scores. McHale and Scarf (2011) also presented a method of modelling the number of home and away team goals, by assuming the number of goals each team scored had a marginal negative binomial distribution, and considered the joint distribution of home and away team goals using copula methods (see Nelsen (2007)).

Dixon and Robinson (1998) introduced a richer model which modelled match goal times as opposed to only the final score. They demonstrated that when goal times were modelled as a non-homogeneous Poisson process, the rate of scoring of teams

changed depending on the current score. Volf (2009) later presented a similar (albeit much less cited) model which he applied to international matches. While not applied to modelling association football goal times, Baker and McHale (2013) developed a 10-dimensional point-process model for the different methods of scoring in American football, and is also an excellent reference for the type of models mentioned here.

Owen (2011) presented a dynamic generalized linear model which extended the model of Maher (1982) to the dynamic spectrum where each team's attack and defense parameter followed a random walk throughout the football season. It was shown that, when applied to Scottish Premier League data, it was beneficial in terms of predictive performance to allow a dynamic model component. Koopman and Lit (2015) then modelled each team's attack and defense parameter as an auto-regressive process which fed into a bivariate Poisson model for the number of home and away team goals. They used a mixture of Monte Carlo and maximum likelihood methods for efficient estimation of all model parameters, and showed some evidence that the model produced a significant profit when compared to bookmaker's odds.

All of the above mentioned models describe the strength of each team with two parameters, one which represents the team's attacking capability, and one which represents their defensive capability. The EPL involves 20 teams competing throughout a season and thus these models use 39 parameters to describe the team strengths (one parameter must be constrained for model identifiability). Furthermore, ranking of the teams based on these parameters is not straightforward.

A further class of models is only concerned with the final match outcome (home team win, draw, or away team win). Fahrmeir and Tutz (1994); Knorr-Held (2000); Cattelan et al. (2013) modelled this outcome using Bradley-Terry type models where parameters representing a team's strength could vary dynamically with time. However these models have the disadvantage of not being able to process all the information contained in a match's score, as for example, a score of 5-0 is treated the same as a score of 3-2.

Scarf and Shi (2008) also employed a Bradley-Terry type model, which they used in order to estimate an idea of 'match importance' by considering the difference in probability of events like winning the league, when a team won or lost a match. These ideas fit in with the concept that teams may change their behaviour if a match is deemed important. Somewhat similarly, the Bradley-Terry type model of Goddard and Asimakopoulos (2003) used covariates related to if matches had league promotion or relegation implications, in a thought that teams may try harder and thus have greater chance of winning in these matches.

From the view of a bookmaker, Bradley-Terry type models are limited in that they are only able to predict probabilities (and thus provide odds) for the 1x2 market

before a match begins. The Poisson models which extend the work of Maher (1982) are a little more useful, in that they can predict probabilities for any correct score $x-y$ and thus are able to provide odds for markets such as total goals under/over, but again, only before a match. The non-homogeneous Poisson process model of Dixon and Robinson (1998) however models goal times, and thus is able to provide odds for the most markets both before the match and in-play. This type of model is thus the most useful to bookmakers who wish to use models for the purposes of making a betting market, and bettors who wish to bet both before a match and in-play on a variety of betting markets.

It should also be mentioned that forecasting match outcomes is not the only use of statistical models in association football. They have become increasingly popular in advising managers on match tactics and player transfers. Data is now collected for each intricate movement a player makes, from missing a tackle to scoring a goal from 30 metres out, and teams are becoming aware of the benefits of analysis of this data. Many teams now house data analysis departments whose job is to value players in order to advise which player transfers represent good value, and to provide feedback on each individual player's performance. Statistics of this kind have been commonly used in other sports (particularly Baseball, which even has a book and film about this very topic, 'Moneyball', Lewis (2004)) but have only recently become utilised in association football. In fact, in 2012, one of the top EPL teams, Manchester City, made player data publicly available in order to promote analysis of player performance.

1.6 Data

We use data that record for each match the particular minute(s) during which a goal is scored and the team scoring the goal. Our analyses in Chapter 3 and Chapter 4 concern all matches in the 2011/2012 EPL season, with specification of prior distributions based on the previous season (2010/2011). While analyses in Chapter 5 concern all matches from five seasons of EPL data, from seasons 2007/2008 to 2011/2012. The goal-time data are available at www.scoresway.com.

As mentioned in Section 1.4 at the end of each 45 minute half, the referee allows extra injury time to be played. Goals scored in first half injury time are simply recorded in the data as having been scored at minute 45, and similarly for goals in second half injury time at minute 90. Hence, the histogram shown in Figure 1.3 displays visible spikes at these times. We also note here that the vast majority of the figures in this thesis have been created via the use of R and ggplot2 (R Core Team (2015); Wickham (2009)).

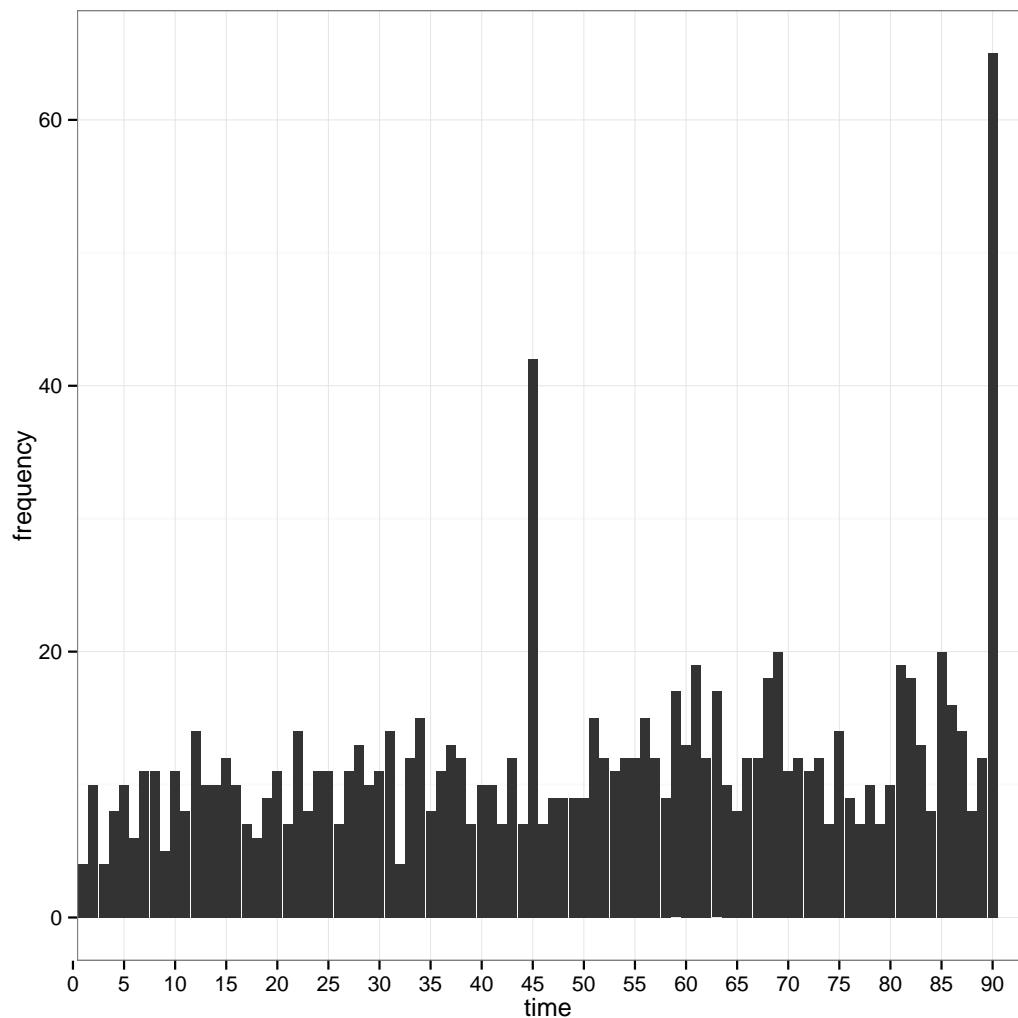


Figure 1.3: A bar plot of goal times (in minutes) for the EPL season 2011/2012

We also use historical data on bookmaker's odds from the 2011/2012 EPL season which are available as a direct download on the website www.football-data.co.uk. These data contain pre-match odds from the UK bookmaker Bet365 for the 1X2 betting market. In the notation of Section 1.3, the data comprise $d_{m,H}$, $d_{m,D}$, and $d_{m,A}$ for matches $m = 1, \dots, 380$. For any match m , $p_{m,H} + p_{m,D} + p_{m,A} = 1$, but calculation of $1/d_{m,H} + 1/d_{m,D} + 1/d_{m,A}$ for a given match m typically yields a value in the region of 1.045 to 1.06, which highlights the use of over-round, $o_{m,H}$, $o_{m,D}$, and $o_{m,A}$ in Bet365's published odds. Values of $1/d_{m,H}$, $1/d_{m,D}$, and $1/d_{m,A}$ and their sum are displayed in Figure 1.4 via a stacked bar chart. In Chapter 3 Section 3.4.3 we make the assumption that $o_{m,H} = o_{m,D} = o_{m,A}$ in order to estimate the bookmaker's assumed probabilities $p_{m,H}$, $p_{m,D}$, and $p_{m,A}$.

1.7 Aim and structure of the thesis

This thesis aims to achieve the following points in the noted chapter:

1. *The creation of models which are widely applicable, have improved predictive power, and are more parsimonious than models currently in literature.* As mentioned throughout this introductory chapter, bookmakers need to offer hundreds of markets for association football matches both before the match and in-play. We thus seek to create a model which is applicable to a wide selection of markets, that is, a model of the goal times in a similar vein to Dixon and Robinson (1998). We consider model parsimony important, and seek to create a more parsimonious, but still rich, model. A model of this nature is introduced in Chapter 3
2. *Inference for models in a Bayesian framework.* The use of Bayesian methods is relatively rare in the literature concerning association football models. We propose reasons for why the Bayesian framework is preferable for this particular application in Chapter 2, and then throughout the remainder of the thesis, use Bayesian methods for all model inference
3. *Methods of inference which are quick to compute.* We also view speed of computation as important, since for in-play markets, updated predictions are needed continuously throughout time, in particular after goals. We thus explore 'particle filtering' methods in Chapter 4 which also naturally present an opportunity to add a dynamic element to the model, in a similar vein to Owen (2011)
4. *The creation of models which can capture behavioural aspects of teams.* Finally, we add another layer to the model proposed in Chapter 3, which represents how teams change their behaviour in relation to their league situation. This research is able to investigate if a notion of 'value' can be placed on the dif-

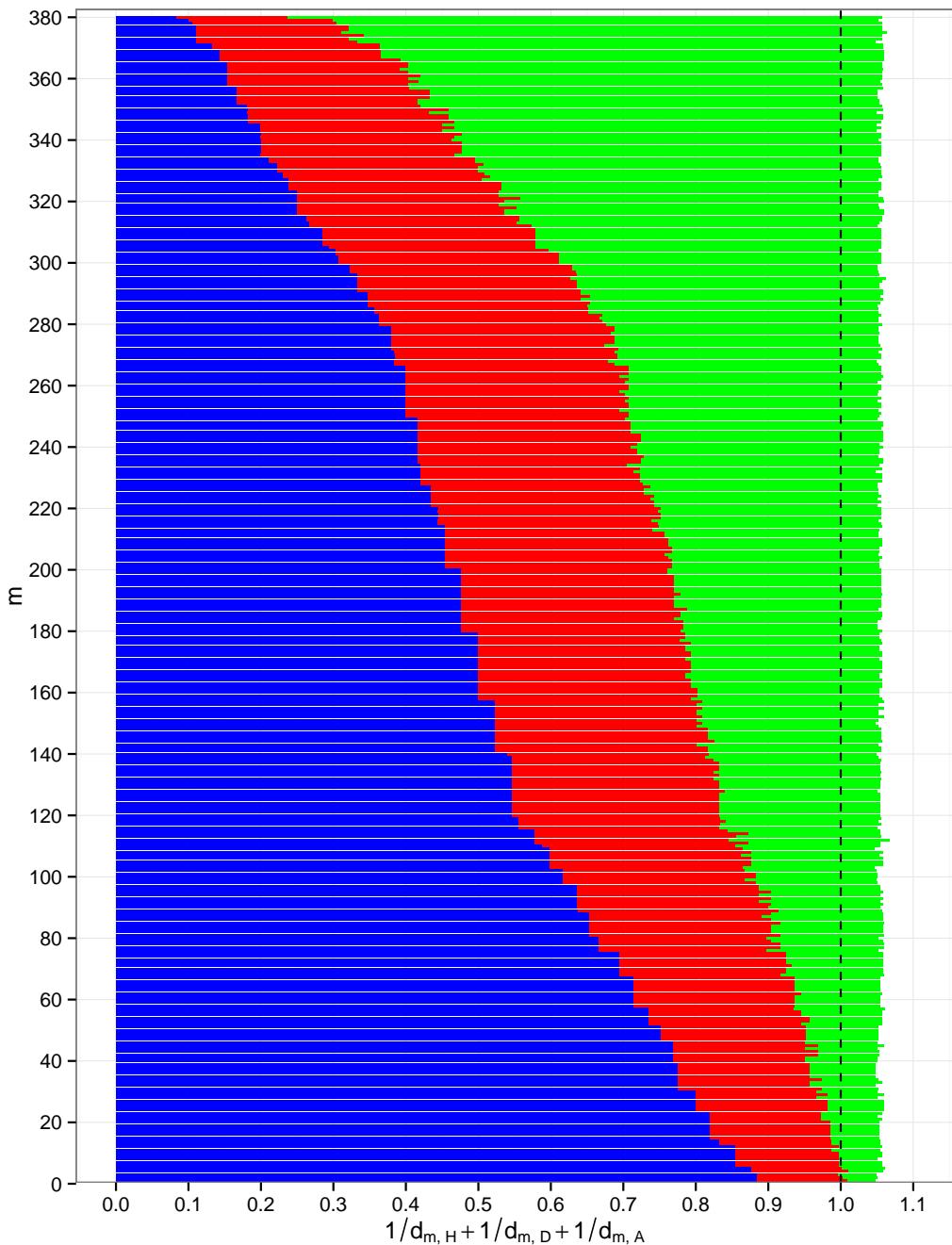


Figure 1.4: A stacked bar chart showing the value of $1/d_{m,H} + 1/d_{m,D} + 1/d_{m,A}$ from the Bet365 1X2 betting market for all 380 matches (m) in the 2011/2012 season. ■
 1/ $d_{m,H}$, ■ 1/ $d_{m,D}$, ■ 1/ $d_{m,A}$

ferent EPL positions, and if there is evidence that concurrent matches are not necessarily independent. This work is presented in Chapter 5

We also introduce the Bayesian methods used throughout the thesis in Chapter 2, and present a conclusion to the thesis in Chapter 6.

Chapter 2

Bayesian computational methods

2.1 Introduction

Here we review some of the Bayesian methods used throughout this thesis. This chapter focuses on introducing methodology which is further discussed and applied in later chapters. We use the notation $\mathbb{P}(.)$ to denote value of probability and $p(.)$ to denote a value of a Probability Density Function (PDF).

2.2 The Bayesian approach to inference

There are currently two main schools of thought with statistics, each with their own approach to parameter inference. The Bayesian approach to inference involves assuming unknown model parameters are random variables. The Frequentist approach however, involves assuming that model parameters are fixed but unknown, and only an estimate of the model parameters, typically the Maximum Likelihood Estimate (MLE), is a random variable. For a Frequentist, probability statements only make sense for repeatable experiments. For example a 95% confidence interval of a model parameter means that when repeating the experiment and each time calculating the confidence interval, 95% of the time the interval should contain the true parameter value. The interpretation of the comparable interval in the Bayesian framework, the Bayesian Credible Interval (BCI), is much clearer - there is a 95% probability the true parameter value is in said interval given the observed data.

Bayes theorem (and in turn Bayesian inference) is named after the Reverend Thomas Bayes (1702-1761) whose work, *An essay towards solving a problem in the doctrine of chances*, which contained the theorem, was published by Richard Price (Bayes and Price (1763)). Bayes' theorem provides the conditional probability of one event given another. For example for events A and B , the probability of observing event

A given we have already observed event B is:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \quad (2.1)$$

What can be slightly confusing is that Bayes' theorem is not inherently Bayesian - it is unarguable and can be easily derived from the definition of conditional probability. A quick derivation is given by noting the definition of conditional probability, $\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B)$, and then by noting that $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$ due to the commutativity of the \cap operator. What is Bayesian however are the methods of inference presented in this chapter, which all rely on Bayes' theorem.

Firstly let us consider a more general formulation of Bayes' theorem relating to partitions of a sample space S . If A_1, \dots, A_n denote a partition of S , that is, they are mutually exclusive and their union is S , we have for any $j \in \{1, \dots, n\}$:

$$\mathbb{P}(A_j|B) = \frac{\mathbb{P}(B|A_j)\mathbb{P}(A_j)}{\sum_{i=1}^n \mathbb{P}(B|A_i)\mathbb{P}(A_i)}. \quad (2.2)$$

Now, the Bayesian approach to inference is to treat the unknown model parameter vector $\boldsymbol{\theta}$ as a random variable, and so inference is concerned with determining the distribution of $\boldsymbol{\theta}$, given we have observed some data \mathbf{D} . This means we consider what is known as the posterior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}|\mathbf{D})$. In a similar vein to Equation (2.2), the posterior distribution may be written as:

$$p(\boldsymbol{\theta}|\mathbf{D}) = \frac{p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{D}|\boldsymbol{\theta}')p(\boldsymbol{\theta}') d\boldsymbol{\theta}'} \quad (2.3)$$

where $p(\mathbf{D}|\boldsymbol{\theta})$ is the likelihood function (which will be determined by the choice of model employed) and $p(\boldsymbol{\theta})$ is the prior density function. The prior $p(\boldsymbol{\theta})$ is chosen to reflect any information regarding the model parameters $\boldsymbol{\theta}$, and the ability to incorporate such information is one of the benefits (but also main criticisms) of the Bayesian approach to inference. Note that Equation (2.3) includes a constant normalising factor, $\int p(\mathbf{D}|\boldsymbol{\theta}')p(\boldsymbol{\theta}') d\boldsymbol{\theta}'$. In many of the Bayesian methods, we need only be concerned with the posterior distribution up to a constant of proportionality, and thus we use:

$$p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (2.4)$$

which is often referred to as Bayes' rule (as opposed to Bayes' theorem).

2.2.1 Bayesian versus Frequentist

The Bayesian framework naturally allows us to make probability statements around model parameters, for example we can compute $\mathbb{P}(\theta_1 > \theta_2)$ by considering posterior probabilities. The equivalent Frequentist approach would be to use a hypothesis test of $H_0 : \theta_1 = \theta_2$ against $H_1 : \theta_1 > \theta_2$ and calculate a p-value, rejecting H_0 in favour of H_1 if the p-value is less than the chosen significance level, usually a somewhat arbitrary 0.05. In fact, in 2015 the academic journal ‘Basic and Applied Social Psychology’ banned the use of p-values stating ‘We believe that the $p < 0.05$ bar is too easy to pass and sometimes serves as an excuse for lower quality research’ (Trafimow and Marks (2015)). Others have also been critical of the use of p-values for example Colquhoun (2014) who made the rather bold statement ‘if you use $p = 0.05$ to suggest that you have made a discovery, you will be wrong at least 30% of the time’. The false positive rate of at least 30% is however somewhat misleading - it is based on the assumptions of real effects being present in only 10% of experiments, and an underlying statistical power of 80%. If 100% of experiments contained real effects, the false discovery rate would be 0%, and similarly, the false discovery rate would decrease as the statistical power increased - so the results are highly sensitive to the assumptions made. There are numerous discussions of p-values, both old and new (Casella and Berger (1987); Ioannidis (2005); Senn (2015)), what is important to take from all this debate however is that p-values should not represent the end of a statistical analysis, but rather the beginning.

The Bayesian approach also very naturally propagates uncertainty surrounding estimates of model parameters, which is contained in the posterior distribution. Frequentists typically rely on an asymptotic normality assumption of the MLE in order to describe uncertainty of model parameter estimates. There are no such assumptions in the Bayesian approach and furthermore, a Bayesian analysis can provide credible intervals for parameters or any function of the parameters which are more easily interpreted than the concept of confidence interval in a Frequentist setting.

Finally (in terms of pro Bayesian arguments), in some areas inference has only been possible with a Bayesian framework. For example, when there are cases of missing data which prevent the calculation of the likelihood function, authors like Streftaris and Gibson (2004) have successfully managed to infer model parameters in a Bayesian framework using data augmentation to infer the missing data.

The main criticism of the Bayesian approach regards the choice of prior distributions, in that the choice is completely subjective. For example two Bayesian statisticians who are very confident in their own, differing, opinion of the likely values of model parameters may choose to use quite different prior distributions. The two statisticians may then come to different conclusions regarding model parameters (in

particular if the data sample is small then the posterior distribution can be quite sensitive to the choice of prior). In this case the difference has not come from the data - it has emerged based on the statistician's opinion. Frequentists typically believe that conclusions must be based solely on the data and Frequentists should always come to the same conclusions regarding model parameters.

Many however see the prior distribution as an asset to the Bayesian statistician's toolbox. It offers the chance to include further information and experience into the belief surrounding model parameters, and extra information should always be appropriately used when available. Sports modelling presents an ideal example for when informative priors can and should be used. For example we have knowledge of the likely values of model parameters based on our experiences of years of watching association football and the vast amount of historical data on the game. Informative priors are used for model parameters throughout this thesis.

In response to the criticism of the use of subjective prior distributions, much research has been discussed on 'non-informative' priors. We discuss the non-informative prior in the following section.

2.2.2 The non-informative prior

An obvious choice of a non-informative prior is the uniform distribution over an allowable/sensible parameter range, or to follow Laplace and choose $p(\boldsymbol{\theta}) = 1$ (Berger and Bernardo (1992)), however, this approach may not be as non-informative as one might first believe. For example if $\psi \in [0, 1]$ is the model parameter of interest, one may elicit the 'non-informative' prior $\psi \sim U(0, 1)$. This however implies $\psi^2 \sim B(\frac{1}{2}, 1)$, suggesting ψ^2 is closer to 0 than 1, so the prior is non-informative regarding ψ but not ψ^2 . That is, the approach of the non-informative uniform prior is not invariant under transformations of the model parameters - different model parameterisations may lead to different posterior distributions.

This lack of invariance prompted the development of a non-informative prior which was also invariant under parameter transformations. Jeffreys (1946) thus proposed what is known as the *Jeffreys' prior*:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(I(\boldsymbol{\theta}))} \quad (2.5)$$

where:

$$I(\boldsymbol{\theta}) = -\mathbb{E} \left(\frac{d \log(p(\mathbf{D}|\boldsymbol{\theta}))}{d\boldsymbol{\theta}^2} \right) \quad (2.6)$$

is the *Fisher information*. The prior is most often improper, that is, it does not inte-

grate to a finite value, and furthermore, the ‘non-informative’ label may be somewhat misleading - the prior requires input from the statistician via the statistical model employed.

Objective Bayesian methods (in which the analysis is data-driven) have been most notably furthered by Bernardo (1979); Berger and Bernardo (1992); Berger et al. (2009) who discuss a ‘reference prior’. The resulting ‘reference posterior’ provides a standard to which posterior distributions, formed from using different priors based on subjective or objective items of information, may be compared. The intuition behind the reference prior is that it is the prior which maximises the distance or divergence between the posterior and prior distribution so that the data have maximum effect on the posterior distribution. A sensible question is ‘how can you choose a prior to maximise the divergence between the prior and posterior distributions before you have seen any data?’ The answer is that reference priors deal with the expectation of the divergence, given a statistical model.

One commonly used divergence measure is the *Kullback-Leibler divergence* introduced by Kullback and Leibler (1951). If we consider a sufficient statistic $T = T(\mathbf{D})$, the Kullback-Leibler divergence for the prior and posterior distributions is:

$$D_{KL} = \int p(\boldsymbol{\theta}|t) \log \left(\frac{p(\boldsymbol{\theta}|t)}{p(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} \quad (2.7)$$

where t is an observation of T . The expectation of D_{KL} over the distribution of T is:

$$\mathbb{E}(D_{KL}) = \int p(t) \int p(\boldsymbol{\theta}|t) \log \left(\frac{p(\boldsymbol{\theta}|t)}{p(\boldsymbol{\theta})} \right) d\boldsymbol{\theta} dt \quad (2.8)$$

$$= \int \int p(\boldsymbol{\theta}, t) \log \left(\frac{p(\boldsymbol{\theta}, t)}{p(\boldsymbol{\theta})p(t)} \right) d\boldsymbol{\theta} dt \quad (2.9)$$

which may be recognised as the mutual information between $\boldsymbol{\theta}$ and t . Thus, the reference prior involves finding $p(\boldsymbol{\theta})$ which maximises this mutual information.

As noted by Berger and Bernardo (1992), this method is typically very hard to implement. It is however consistent with the Jeffreys’ prior for the single parameter case, and it is generally much more recommended in the multi-parameter case due to the shortcomings of the Jeffreys’ prior in a multivariate setting - as acknowledged by Jeffreys himself (Jeffreys (1946)).

2.3 MCMC

In most cases, the form of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$ is analytically intractable. For example, for the chosen prior distributions and model specifications

employed in Chapter 3, $p(\boldsymbol{\theta}|\mathbf{D}) \propto p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ does not conform to a recognisable distribution. As a result, many Bayesian analyses use sampling methods to approximate the distribution of the posterior $p(\boldsymbol{\theta}|\mathbf{D})$, of which Markov Chain Monte Carlo (MCMC) is the most popular. MCMC methods are inherently very computationally demanding, and so have only been possible with advances in computing power.

The basic Monte Carlo principle is to draw an independent and identically distributed set of samples $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n\}$ from our target distribution (the posterior $p(\boldsymbol{\theta}|\mathbf{D})$) which may be used directly for parameter inference and prediction. The n samples can be used to approximate the expectation:

$$\begin{aligned}\mathbb{E}(f(\boldsymbol{\theta}|\mathbf{D})) &= \int f(\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \\ &\approx \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{\theta}_i)\end{aligned}\tag{2.10}$$

for some function f which must be Lebesgue integrable with respect to $p(\boldsymbol{\theta}|\mathbf{D})$ in order for the expectation to exist. The strong law of large numbers then holds and the approximation becomes more accurate as $n \rightarrow \infty$. It is thus possible to approximate features such as the posterior mean using $f(\boldsymbol{\theta}) = \boldsymbol{\theta}$, or posterior probabilities like $\mathbb{P}(\boldsymbol{\theta} < \mathbf{L}|\mathbf{D})$ using $f(\boldsymbol{\theta}) = I(\boldsymbol{\theta} < \mathbf{L})$ (the indicator function which is 1 if the comparison is true and 0 otherwise). We can quickly see that having a Monte Carlo sample from the posterior distribution of $\boldsymbol{\theta}$ is sufficient to enable inference to be made.

In order to generate the Monte Carlo sample, we seek to create a Markov chain where the stationary distribution corresponds to our posterior distribution. We firstly consider a discrete state-space $S = \{s_1, \dots, s_m\}$ for the purposes of illustration of the ideas of Markov chains, before describing methods relating to continuous state-spaces in Section 2.3.1. The extension to continuous state-spaces do not require any new concepts (Roberts (1996); Tierney (1996)).

A Markov chain is a stochastic process which takes values in S at discrete time points t while satisfying the following condition: if the state of the chain at time t is X_t , then the distribution of X_t conditional on X_1, \dots, X_{t-1} is the same as the distribution of X_t conditional only on X_{t-1} . That is:

$$\mathbb{P}(X_t|X_1, \dots, X_{t-1}) = \mathbb{P}(X_t|X_{t-1})\tag{2.11}$$

in words, the future states which the process takes depend only on the current state and not the past. This finite state Markov Chain is defined by a $m \times m$ transition matrix \mathbf{T} where entry $T_{i,j}$ contains $\mathbb{P}(X_{t+1} = s_j | X_t = s_i)$ for all integer values of $t \geq 0$. Furthermore, if \mathbf{q} is a row vector of length m (like all the distributions we will be concerned with here) containing the distribution of the Markov chain state

at time 0 so $q_i = \mathbb{P}(X_0 = s_i)$, then the distribution of the Markov chain state at time 1 is $\mathbf{q}\mathbf{T}$. It follows by repeated application of the matrix multiplication that $\mathbf{q}\mathbf{T}^k$ is the the distribution of the Markov chain state at time k .

Roberts (1996) succinctly describes the three properties which \mathbf{T} must satisfy in order for the corresponding Markov chain to converge to stationary distribution $\boldsymbol{\pi}$:

1. *Irreducible.* That it is possible for the Markov process to reach any state s_i from any state s_j within a finite number of time steps. That is, the Markov chain transition graph is connected
2. *Aperiodic.* The chain does not oscillate between different sets of states in a regular periodic movement. In other words, the Markov chain transition graph has period 1
3. *Positive recurrent.* If an initial value x_0 is sampled from $\boldsymbol{\pi}$ then x_1 and all subsequent iterates will also be distributed according to $\boldsymbol{\pi}$. In terms of the stationary distribution $\boldsymbol{\pi}$, this implies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}$

The positive recurrent property can sometimes be shown by demonstrating that a stronger property holds, known as *detailed balance*:

$$\pi_i T_{i,j} = \pi_j T_{j,i} \quad \text{for all } i, j. \quad (2.12)$$

Furthermore, summing both sides over j yields:

$$\pi_i = \sum_{j=1}^m \pi_j T_{j,i} \quad (2.13)$$

since $\sum_{j=1}^m \pi_i T_{i,j} = \pi_i \sum_{j=1}^m T_{i,j} = \pi_i$, it then follows that $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}$. Detailed balance can also be seen as a reversibility condition. In words, it means that the probability of the Markov chain being in state s_i and moving from s_i to s_j must be the equal to the probability of being in state s_j and moving from s_j to s_i .

Thus for any initial distribution \mathbf{q} and a transition matrix which is *irreducible*, *aperiodic*, and satisfies $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{T}$, we have that $\mathbf{q}\mathbf{T}^k$ tends to $\boldsymbol{\pi}$ as $k \rightarrow \infty$. That is, that no matter what the initial distribution of the chain is, after a sufficient number of steps the distribution of the state of the chain will approach $\boldsymbol{\pi}$ and will be independent of the initial distribution \mathbf{q} . Thus if we simulate the process starting from any of the states in S , after a sufficient amount of time steps, we will be able to take samples from the stationary distribution $\boldsymbol{\pi}$.

2.3.1 The Metropolis-Hastings algorithm

In Section 2.3 we reviewed some of the theory of how a Markov chain can be defined by its transition matrix \mathbf{T} in order to have some stationary distribution over a discrete state-space S . Here, we review the Metropolis-Hastings (MH) algorithm (Metropolis et al. (1953), Hastings (1970)) which ensures the stationary distribution of the Markov chain is our posterior distribution of interest over a continuous state-space.

It is intuitive that Markov chains can be used in cases where the state-space is discrete, however the theory of Markov chains in a continuous state-space is nearly identical after replacing the discrete stationary distribution $\boldsymbol{\pi}$ with a continuous posterior density function $p(\boldsymbol{\theta}|\mathbf{D})$, discrete sums with continuous integrals, and the discrete support transition matrix \mathbf{T} with a continuous support transition kernel K . For example, in a continuous state-space setting we may write Equation (2.13) as:

$$p(\boldsymbol{\theta}_{t+1}|\mathbf{D}) = \int p(\boldsymbol{\theta}_t|\mathbf{D})K(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) d\boldsymbol{\theta}_t \quad (2.14)$$

so the kernel K is the conditional density of $\boldsymbol{\theta}_{t+1}$ given $\boldsymbol{\theta}_t$, where $\boldsymbol{\theta}_t$ denotes the state of the chain at time t .

The MH algorithm is commonly referred to as the most popular MCMC method (Hitchcock (2003)) and is the base for many other algorithms which are special cases or extensions, for example the Gibbs sampler (Geman and Geman (1984)) which we review in Section 2.3.2. An outline of the algorithm is as follows:

Let the state of the chain at time t be $\boldsymbol{\theta}_t$ where $\boldsymbol{\theta}_0$ is some user-specified value. Then with a user-specified proposal distribution $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)$, repeat the following steps for $t = 0, \dots, n$:

1. Sample $\boldsymbol{\theta}'$ from the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)$
2. Calculate the acceptance probability $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}_t)$
3. With probability $\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}_t)$ set $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}'$, otherwise $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$

The acceptance probability is given by:

$$\begin{aligned} \alpha(\boldsymbol{\theta}', \boldsymbol{\theta}_t) &= \min \left(1, \frac{p(\boldsymbol{\theta}'|\mathbf{D})}{p(\boldsymbol{\theta}_t|\mathbf{D})} \times \frac{q(\boldsymbol{\theta}_t|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)} \right) \\ &= \min \left(1, \frac{p(\mathbf{D}|\boldsymbol{\theta}')}{p(\mathbf{D}|\boldsymbol{\theta}_t)} \times \frac{p(\boldsymbol{\theta}')}{p(\boldsymbol{\theta}_t)} \times \frac{q(\boldsymbol{\theta}_t|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)} \right). \end{aligned} \quad (2.15)$$

Note that the appearance of the ratio in the acceptance probability means that the posterior normalising factor (seen in Equation (2.3)) cancels. The transition kernel

for the MH algorithm from point current state $\boldsymbol{\theta}_t$ to the next state $\boldsymbol{\theta}_{t+1}$ is:

$$K(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) = q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)\alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) + \delta_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1}) \int q(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t)(1 - \alpha(\boldsymbol{\theta}^*, \boldsymbol{\theta}_t)) d\boldsymbol{\theta}^* \quad (2.16)$$

where $\delta_{\boldsymbol{\theta}_t}(\boldsymbol{\theta}_{t+1})$ is the Dirac delta measure located at $\boldsymbol{\theta}_t$. The first term on the right hand side of Equation (2.16) corresponds to an accepted proposal, and the second a rejected proposal. For further details on the transition kernel of the MH algorithm, see for example Tierney (1998).

It can be proved that the Markov chain produced by the MH algorithm converges and has the desired stationary distribution $p(\boldsymbol{\theta}|\mathbf{D})$ by considering the three properties described in Section 2.3. Firstly, the chain is irreducible so long as the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)$ allows the chain to cover the support of the posterior $p(\boldsymbol{\theta}|\mathbf{D})$. Secondly, the chain is aperiodic since a rejected proposal implies $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$. Finally, we show the chain is positive recurrent by demonstrating that detailed balance holds, that is:

$$K(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathbf{D}) = K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})p(\boldsymbol{\theta}_{t+1}|\mathbf{D}). \quad (2.17)$$

In the case of a rejected proposal we have $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$ and so Equation (2.17) clearly holds. In the case of an accepted proposal we have:

$$\begin{aligned} K(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t) &= q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)\alpha(\boldsymbol{\theta}_{t+1}, \boldsymbol{\theta}_t) \\ &= \min \left(q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t), \frac{p(\boldsymbol{\theta}_{t+1}|\mathbf{D})q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})}{p(\boldsymbol{\theta}_t|\mathbf{D})} \right) \end{aligned} \quad (2.18)$$

and therefore:

$$\begin{aligned} K(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t)p(\boldsymbol{\theta}_t|\mathbf{D}) &= \min(p(\boldsymbol{\theta}_t|\mathbf{D})q(\boldsymbol{\theta}_{t+1}|\boldsymbol{\theta}_t), p(\boldsymbol{\theta}_{t+1}|\mathbf{D})q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})) \\ &= K(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t+1})p(\boldsymbol{\theta}_{t+1}|\mathbf{D}) \end{aligned} \quad (2.19)$$

by noting that $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t+1}$ are symmetric inside the min function. We now have proof of Equation (2.17), that is, the MH algorithm produces a Markov Chain which converges to a stationary distribution of $p(\boldsymbol{\theta}|\mathbf{D})$.

Much research has been undertaken regarding the optimal acceptance rate within the MH algorithm, that is, the proportion of proposals $\boldsymbol{\theta}'$ that are accepted. Intuitively, the proposal density $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t)$ should be chosen so that the acceptance rate is not close to 0 (the proposals are never accepted, for example when the proposal distribution has very high variance, and the chain is ‘stuck’) and not close to 1 (the proposals are always accepted, for example when the proposal distribution has very small variance, and the chain moves very slowly).

A common choice for the proposal density is $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t) = N(\boldsymbol{\theta}_t, \sigma^2)$ where the user may tune σ^2 to achieve a certain acceptance rate, the resulting algorithm is known as *random-walk MH*. Also of note is $q(\boldsymbol{\theta}'|\boldsymbol{\theta}_t) = q(\boldsymbol{\theta}')$ which results in the *independence sampler* algorithm.

Roberts et al. (1997) (the original work was published in 1994) were the first authors to publish theoretical results regarding the optimal acceptance rate, and suggested a commonly cited value of 0.234. This value was obtained however under the assumption of a high-dimensional target distribution made of Independent Identically Distributed (IID) components. Gelman et al. (1996) provided some theoretical justification for acceptance rates in the range of 0.15 to 0.5, but note that these values do not guarantee efficiency of the algorithm, in particular when sampling from a highly multi-modal distribution. More recently, Bedard (2008) warned statisticians of the problems of blindly tuning a MH algorithm to the 0.234 rule, and showed that when the assumption of IID components was relaxed, the optimal acceptance rate may be far from 0.234. For MH algorithms implemented in this thesis we follow Gelman et al. (1996) and are satisfied with acceptance rates between 0.15 and 0.5.

Algorithm 1 provides an implementation of the MH algorithm which takes n samples for d parameters, and stores the sample from each iteration in the rows of a matrix denoted by \mathbf{S} . The algorithm performs updates to each of the d parameters separately.

2.3.2 The Gibbs sampler

We describe the Gibbs sampler for a d dimensional posterior $p(\boldsymbol{\theta}|\mathbf{D})$ with $\boldsymbol{\theta} = (\theta^{(1)}, \dots, \theta^{(d)})$. The Gibbs sampler relies on simulation from the conditional posterior density of $\theta^{(i)}$ given all other parameters in $\boldsymbol{\theta}$ for all $i = 1, \dots, d$. For a user-specified starting value $\boldsymbol{\theta}_0$, the Gibbs sampler algorithm repeats the following steps for $t = 0, \dots, n$:

1. sample $\theta_{t+1}^{(1)}$ from conditional density $p(\theta^{(1)}|\mathbf{D}, \theta_t^{(2)}, \dots, \theta_t^{(d)})$
2. sample $\theta_{t+1}^{(2)}$ from conditional density $p(\theta^{(2)}|\mathbf{D}, \theta_{t+1}^{(1)}, \theta_t^{(3)}, \dots, \theta_t^{(d)})$
- ⋮
- k. sample $\theta_{t+1}^{(k)}$ from conditional density $p(\theta^{(k)}|\mathbf{D}, \theta_{t+1}^{(1)}, \dots, \theta_{t+1}^{(k-1)}, \theta_t^{(k+1)}, \dots, \theta_t^{(d)})$
- ⋮
- d. sample $\theta_{t+1}^{(d)}$ from conditional density $p(\theta^{(d)}|\mathbf{D}, \theta_{t+1}^{(1)}, \dots, \theta_{t+1}^{(d-1)})$

The key note from this algorithm is that after having updated a component, the updated value must be used when we condition on it in subsequent samples.

Algorithm 1 MH algorithm

```

1: procedure MH( $n, m$ )
2:    $\boldsymbol{\theta} = (\theta_0^{(1)}, \dots, \theta_0^{(d)})$ 
3:    $\mathbf{S} = \text{Matrix}(n, d)$ 
4:   for  $t = 1$  to  $n$  do
5:     for  $p = 1$  to  $d$  do
6:        $\boldsymbol{\theta}' = \boldsymbol{\theta}$ 
7:        $\theta'_p = q(\theta_p)$ 
8:        $\log D = \log(p(\mathbf{D}|\boldsymbol{\theta})) + \log(p(\boldsymbol{\theta})) + \log(q(\theta'_p|\theta_p))$ 
9:        $\log N = \log(p(\mathbf{D}|\boldsymbol{\theta}')) + \log(p(\boldsymbol{\theta}')) + \log(q(\theta_p|\theta'_p))$ 
10:      if  $u() < \exp(\log N - \log D)$  then
11:         $\theta_p = \theta'_p$ 
12:      end if
13:       $S_{t,p} = \theta_p$ 
14:    end for
15:  end for
16: end procedure

17: function  $q(x)$ 
18:   return random draw from the proposal density  $q(x'|x)$ 
19: end function

20: function  $u()$ 
21:   return random draw from  $U(0, 1)$ 
22: end function

```

It can be shown that the Gibbs sampler is a composition of MH moves where the proposed value is always accepted. To illustrate this we consider the simple two dimensional case, the extension to higher dimensions follows similarly. Consider $\boldsymbol{\theta} = (\theta^{(1)}, \theta^{(2)})$ and a proposal to update component $\theta^{(1)}$ of θ' at current time t where the state is $\boldsymbol{\theta}_t = (\theta_t^{(1)}, \theta_t^{(2)})$. The MH acceptance probability is given by:

$$\begin{aligned}\alpha &= \min \left(1, \frac{p(\theta', \theta_t^{(2)} | \mathbf{D})}{p(\theta_t^{(1)}, \theta_t^{(2)} | \mathbf{D})} \times \frac{p(\theta_t^{(1)} | \mathbf{D}, \theta_t^{(2)})}{p(\theta' | \mathbf{D}, \theta_t^{(2)})} \right) \\ &= \min \left(1, \frac{p(\theta', \theta_t^{(2)} | \mathbf{D})}{p(\theta_t^{(1)}, \theta_t^{(2)} | \mathbf{D})} \times \frac{p(\theta_t^{(1)}, \theta_t^{(2)} | \mathbf{D})}{p(\theta_t^{(2)})} \times \frac{p(\theta_t^{(2)})}{p(\theta', \theta_t^{(2)} | \mathbf{D})} \right) \\ &= 1.\end{aligned}\tag{2.20}$$

The update step of $\theta^{(2)}$ follows similarly when conditioning on $\theta_{t+1}^{(1)} = \theta'$.

Since a Gibbs move is a special case of a MH move, it follows that each individual Gibbs component update satisfies detailed balance as described in Section 2.3.1. It is also clear that the chain will be irreducible since the support of the conditional posterior densities covers the same support as the full posterior distribution. It may not however be obvious that the chain is aperiodic since the chain does not allow for rejections. However it is difficult to imagine a sequence of conditional posterior distributions that would ‘force’ the chain away from the current value. We refer the reader to Roberts and Polson (1994) for more technical details regarding the convergence of the Gibbs sampler and a proof of the aperiodicity of the Gibbs sampler.

The Gibbs sampler is in some ways simpler than the MH algorithm since it does not require tuning of the proposal density in order to acquire satisfactory mixing of the Markov chain. It does however require simulation from particular conditional distributions which may not always be straightforward, as is the case in Chapter 3 and onwards. It should also be noted that it is perfectly acceptable to mix the Gibbs sampling and MH algorithm sampling steps, with Gibbs updating steps for some components and MH updating steps for others.

Algorithm 2 provides an implementation of the Gibbs sampling algorithm which takes n samples for d parameters, and stores the sample from each iteration in the rows of a matrix denoted by \mathbf{S} . The algorithm performs updates to each of the d parameters separately.

2.3.3 Burn-in, thinning, and convergence diagnostics

Although algorithms like MH are guaranteed to produce Markov chains which converge, there is no guarantee on how many time steps this may take. In particular,

Algorithm 2 Gibbs sampler algorithm

```

1: procedure GIBBS( $n, d$ )
2:    $\boldsymbol{\theta} = (\theta_0^{(1)}, \dots, \theta_0^{(d)})$ 
3:    $\mathbf{S} = \text{Matrix}(n, d)$ 
4:   for  $t = 1$  to  $n$  do
5:     for  $p = 1$  to  $d$  do
6:        $\theta_p = q(\boldsymbol{\theta}, p)$ 
7:        $S_{t,p} = \theta_p$ 
8:     end for
9:   end for
10:  end procedure

11: function  $q(\boldsymbol{\theta}, p)$ 
12:   return random draw from the conditional density
         $p(\theta^{(p)} | \mathbf{D}, \theta^{(1)}, \dots, \theta^{(p-1)}, \theta^{(p+1)}, \dots, \theta^{(d)})$ 
13: end function

```

time to converge may be long if the chain is initialised in a position of low posterior probability. Furthermore, after a sufficient number of steps have been taken and the chain has converged, adjacent samples may exhibit high dependence (positive auto-correlation), breaking the independence assumption of a Monte Carlo sample.

One solution to the two above mentioned problems is the use of burn-in and thinning. To burn samples means to exclude the first B samples from analysis, B is typically determined by visual inspection of trace plots to see how long it takes for the sampled value to ‘settle’. To thin samples means to only use every T -th sample, with T usually determined by visual inspection of auto-correlation plots to see what value of T would ensure that adjacent samples are sufficiently independent.

The strategies of burn-in and thinning are however very wasteful. They involve discarding information, which should be a statistical sin. Authors like Geyer (2011) suggest that instead of burn, ‘any point you don’t mind having in a sample is a good starting point’ and suggest practical ways of initialising the Markov chain so no burn-in is required. For example practising statisticians may run MCMC procedures multiple times, in particular if using the MH algorithm which requires the tuning of the proposal distribution based on the acceptance rate. Thus a good rule to follow is to initialise each Markov chain where the previous one ended. Alternatively, it may be possible to initialise the Markov chain at the posterior mode which can sometimes be found via optimisation methods, or the MLE may be a good starting point if there is thought to be little influence of the prior on the posterior distribution.

The smallest amount of thinning involves discarding every 2nd sample, so at best,

thinning involves immediately halving the sample size. Link and Eaton (2012) show results from simulations suggesting that estimation of posterior features is more precise when based on un-thinned chains, which is intuitive since they use more information. Furthermore, for unbiased estimation of a posterior expectation using Equation (2.10), there is no requirement for independence within the MCMC samples.

Estimation of standard errors of estimates (MCMC error) however does require independent samples. A useful reference for this point (and many others related to MCMC) is the discussion paper by Kass et al. (1998). For example Carlin describes how it is perfectly fine to estimate the posterior mean with the the sample mean $\bar{\theta} = \frac{1}{n} \sum_{i=1}^n \theta_i$, since the estimate is unbiased even if the samples are not independent. However calculation the sample variance $s^2 = \frac{1}{n-1} \sum_{i=1}^n (\theta_i - \bar{\theta})^2$ and using s^2/n as an estimate of the MCMC error of $\bar{\theta}$ would very likely be an underestimate due to the positive auto-correlation in the samples.

Numerous methods for calculating MCMC error are discussed in Kass et al. (1998), including using thinned chains (which is deemed sub-optimal). Neal suggests (and further explains in Neal (1993)) the calculation of an *effective sample size*, n_{eff} , which accounts for the positive auto-correlation within the samples:

$$n_{eff} = \frac{n}{1 + 2 \sum_{k=1}^{\infty} acf(k)} \quad (2.21)$$

where $acf(k)$ is the estimated auto-correlation at lag k . The sum must then be truncated at some reasonable value and MCMC error estimates may be based on s^2/n_{eff} . Calculation of n_{eff} may also be useful in order to gain an idea how many samples, n , to take. Other methods like *non-overlapping batch means* are described in Geyer (2011).

Throughout this thesis we only burn samples if we are unsure of the validity of the chain starting points, and no thinning is used.

Formal diagnostics are available for checking the convergence of MCMC algorithms. The most popular of these are the Gelman-Rubin diagnostics (see Gelman and Rubin (1992); Brooks and Gelman (1998)) which involve running multiple chains from over-dispersed starting points and comparing the between chain and within chain variability, and the Geweke diagnostic (see Geweke et al. (1991)), which produces z-scores by comparing the means of the first 10% and last 50% (by default for most computer packages) of samples.

2.4 RJMCMC

Reversible Jump Markov Chain Monte Carlo (RJMCMC) is a framework in which reversible Markov chain samplers are able to ‘jump’ between parameter state spaces of differing dimensionality. The RJMCMC algorithm was first described in Green (1995) and is an extension of the MH algorithm. It has since been used to tackle numerous problems, mainly concerning Bayesian model choice, for example in Hastie and Green (2012); Punska et al. (1999), but may also be used for data imputation. For example if we are concerned with modelling goal times in association football, and obtain data which records goal times for several matches, but also data which only records the match outcome (home team win, draw, or away team win) for several matches, we could use RJMCMC methods to impute a varying dimension vector of goal times for the matches for which goal times were not recorded. For an example of the use of RJMCMC to impute unobserved data, see Gibson and Renshaw (1998).

RJMCMC allows the MCMC routine to explore different models m , each with their own parameter space $\boldsymbol{\theta}_m$ and aims to generate samples from the posterior distribution $p(m, \boldsymbol{\theta}_m | \mathbf{D})$. The dimension of $\boldsymbol{\theta}_m$, $\dim(\boldsymbol{\theta}_m)$, need not be the same for each m . An outline of the algorithm for a between-models move is as follows:

Let the current state of the chain be $(m, \boldsymbol{\theta}_m)$, then with a user-specified model proposal density $j(m'|m)$, parameter proposal density $q(\boldsymbol{\mu}|\boldsymbol{\theta}_m, m, m')$, and invertible parameter transformation function $g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{\mu})$:

1. Sample the proposal model m' from $j(m'|m)$
2. Sample $\boldsymbol{\mu}$ from $q(\boldsymbol{\mu}|\boldsymbol{\theta}_m, m, m')$
3. Set $(\boldsymbol{\theta}'_{m'}, \boldsymbol{\mu}') = g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{\mu})$
4. With probability α , set the new state of the chain to $(m', \boldsymbol{\theta}'_{m'})$

Here, the acceptance probability is given by:

$$\alpha = \min \left(1, \frac{p(\mathbf{D}|\boldsymbol{\theta}')}{p(\mathbf{D}|\boldsymbol{\theta})} \times \frac{p(\boldsymbol{\theta}'|m')}{p(\boldsymbol{\theta}|m)} \times \frac{p(m')}{p(m)} \times \frac{j(m|m')}{j(m'|m)} \times \frac{q(\boldsymbol{\mu}'|\boldsymbol{\theta}'_{m'}, m', m)}{q(\boldsymbol{\mu}|\boldsymbol{\theta}_m, m, m')} \times \left| \frac{\partial g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{\mu})}{\partial (\boldsymbol{\theta}_m, \boldsymbol{\mu})} \right| \right) \quad (2.22)$$

The algorithm is essentially the same as MH but formulated to allow for sampling from a union of models with parameter spaces of differing dimension, and hence comments on the MH algorithm are also applicable here.

When moving across model spaces it is likely that there is a natural way to propose the new model parameters $\boldsymbol{\theta}'_{m'}$ based on the current model parameters $\boldsymbol{\theta}_m$ and the function $g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{\mu})$ should be chosen accordingly. The variates $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ provide

dimension matching between spaces such that $\dim(\boldsymbol{\theta}_m) + \dim(\boldsymbol{\mu}) = \dim(\boldsymbol{\theta}'_{m'}) + \dim(\boldsymbol{\mu}')$ (note $\dim(\boldsymbol{\mu}')$ or $\dim(\boldsymbol{\mu})$ can be 0), a necessary condition for the Jacobian term to always be calculable, while the Jacobian term appears in α to account for the deterministic transformation of the parameter space when switching models. Typically when proposing a move from model m to model m' where m is nested in m' , $\boldsymbol{\mu}$ when passed into $g_{m,m'}(\boldsymbol{\theta}_m, \boldsymbol{\mu})$ will be transformed (or given directly) into the additional parameters in $\boldsymbol{\theta}'_{m'}$ and $\dim(\boldsymbol{\mu}') = 0$. Oppositely, when proposing a move from model m to model m' where m' is nested in m , that is, the removal of parameters, $\dim(\boldsymbol{\mu}) = 0$ and $\boldsymbol{\mu}'$ will account for the removed parameters.

Furthermore, often the Jacobian term will be 1 (which it is in a standard MH procedure) for example when all the parameters of the proposed model are generated directly from the proposal distribution (essentially an independence sampler). A detailed example of an implementation of RJMCMC is shown in Chapter 5 Section 5.6. For an example of when the Jacobian term is not equal to 1; see the ‘Poisson versus negative binomial’ section in Hastie and Green (2012).

2.5 Particle filtering methods

Particle filtering methods have been applied in a wide range of fields such as robotics (Montemerlo et al. (2002); Rekleitis (2004)), navigation (Gustafsson et al. (2002)), and image processing (Nummiaro et al. (2003)), where they are used to quickly update a posterior belief around a dynamic system as new data arrives throughout time. The particle filter does not limit the dynamic system to be linear or for the system and observational noise to be Gaussian (for which the Kalman filter (Kalman (1960)) is a popular approach). In addition, particle filtering methods can be surprisingly straightforward to implement. Practically, the implementation of a particle filter can readily take advantage of parallel processing which is not always straightforward in most MCMC methods (for example in MH).

The methods are typically used for what is known as *on-line learning*, that is, updating of posterior belief as data arrive sequentially through time. This differs the more common case of *off-line learning* where the data are fixed. For example, updating posterior belief regarding model parameters every few seconds of an association football match would clearly constitute on-line learning, and would naturally require computationally fast methods to be useful. Inferring model parameters for a fixed amount of past data would be off-line learning, and is typically not be so time restricted.

Also known as Sequential Importance Resampling (SIR) or Sequential Monte Carlo (SMC) (Sanjeev Arulampalam et al. (2002)), particle filtering methods have been

designed for filtering of hidden Markov models. That is, at each time t we sequentially observe data \mathbf{y}_t and wish to infer the distribution of the unobserved (hidden) underlying Markov process \mathbf{x}_t where \mathbf{y}_t is observed with noise, for example $\mathbf{y}_t = f(\mathbf{x}_t) + \boldsymbol{\epsilon}_t$. The posterior distribution of the hidden state is thus $p(\mathbf{x}_t | \mathbf{D}_t)$ where $\mathbf{D}_t = \{\mathbf{D}_{t-1}, \mathbf{y}_t\}$ (the total data observed up to time t).

While not used in this thesis, it should also be noted that these methods may be used to sample from a target distribution $p(\mathbf{x}|\mathbf{D})$ in non-sequential problems. For example if $p(\mathbf{x}|\mathbf{D})$ is highly multi-modal and difficult to sample from, it may be beneficial to begin sampling from an easy-to-sample distribution and move the sample through an artificial sequence of distributions which ultimately ends with $p(\mathbf{x}|\mathbf{D})$. The idea is that at each time step, the sample distribution approaches closer to the target distribution $p(\mathbf{x}|\mathbf{D})$. Such methods have been discussed by Neal (2001); Chopin (2002); Del Moral et al. (2006) and in some cases outperform traditional MCMC methods. For example Del Moral et al. (2006) considers the artificial sequence of distributions:

$$p_t(\mathbf{x}|\mathbf{D}) \propto p(\mathbf{D}|\mathbf{x})^{\phi_t} p(\mathbf{x}) \quad (2.23)$$

where $0 \leq \phi_1 < \dots < \phi_T = 1$. At $t = 1$ the sample is very close to a sample from the prior (or exactly if $\phi_1 = 0$), and at each time point t , the distribution of the samples moves closer to the target distribution, a sample from the target distribution is ultimately obtained at $t = T$.

We now discuss the underlying ideas of the particle filtering algorithm. Suppose we have a size n sample from the posterior distribution $p(\mathbf{x}_t | \mathbf{D}_t)$ at time t . The sample is defined by a set of points $\{\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(n)}\}$ with corresponding importance weights $\{\omega_t^{(1)}, \dots, \omega_t^{(n)}\}$ and will be denoted as \mathbf{P}_t , the particle approximation of $p(\mathbf{x}_t | \mathbf{D}_t)$. In the case where all the weights are equal ($\omega_t^{(i)} = 1/n$ for all i), \mathbf{P}_t is a direct Monte Carlo sample - as will be the case for the methods described in this thesis.

We assume that the model transition density of the dynamic parameter \mathbf{x}_t is available and is $p(\mathbf{x}_{t+1} | \mathbf{x}_t)$. We further assume the likelihood function for the data observed at time t , \mathbf{y}_t , is available and is $p(\mathbf{y}_t | \mathbf{x}_t)$.

After obtaining data \mathbf{y}_{t+1} we wish to sample from the posterior distribution $p(\mathbf{x}_{t+1} | \mathbf{D}_{t+1})$. Using Bayes rule:

$$\begin{aligned} p(\mathbf{x}_{t+1} | \mathbf{D}_{t+1}) &\propto p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}) p(\mathbf{x}_{t+1} | \mathbf{D}_t) \\ &= p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}) \int_{-\infty}^{\infty} p(\mathbf{x}_{t+1} | \mathbf{x}_t) p(\mathbf{x}_t | \mathbf{D}_t) d\mathbf{x}_t \end{aligned} \quad (2.24)$$

where $p(\mathbf{x}_{t+1} | \mathbf{D}_t)$ is the prior density of \mathbf{x}_{t+1} at time t . At time $t = 0$, a known prior $p(\mathbf{x}_1 | \mathbf{D}_0) = p(\mathbf{x}_1)$ is used. Now, we have our discrete particle representation

of $p(\mathbf{x}_t|\mathbf{D}_t)$, \mathbf{P}_t , and so the integral in Equation (2.24) is replaced with a weighted summation where each particle i is sampled from the transition density, $p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(i)})$. The posterior distribution $p(\mathbf{x}_{t+1}|\mathbf{D}_{t+1})$ is then approximated by:

$$\hat{p}(\mathbf{x}_{t+1}|\mathbf{D}_{t+1}) \propto p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}) \sum_{i=1}^n \omega_t^{(i)} p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(i)}). \quad (2.25)$$

So in theory a set of particles \mathbf{P}_t , should be able to represent a posterior distribution throughout time t with some movement in the particles, followed by a weighting step based on the likelihood of the data at that point. We continue the discussion of particle filtering methods in Chapter 4 where we apply them to estimate parameters in the models we develop.

2.6 Bayesian model choice

In a Bayesian framework it is only natural to represent all uncertainty by probability distributions. Typically this is done by displaying the belief of a parameter by its posterior distribution, but can also extend to posterior probabilities for a collection of models. The discussion paper by Draper (1995) promotes consideration of model uncertainty, as opposed to finding the ‘best’ model from a collection of competing models and basing all inference on that single model, which may lead to over-confidence in uncertainty of predictions.

For a collection of models m_1, \dots, m_K , the posterior probability of model m_k is:

$$p(m_k|\mathbf{D}) = \frac{p(m_k)p(\mathbf{D}|m_k)}{\sum_{i=1}^K p(m_i)p(\mathbf{D}|m_i)} \quad (2.26)$$

where $p(m_k)$ is the prior model probability of model m_k and:

$$p(\mathbf{D}|m_k) = \int p(\mathbf{D}|m_k, \boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k|m_k) d\boldsymbol{\theta}_k \quad (2.27)$$

where $\boldsymbol{\theta}_k$ is the parameter vector for model m_k , $p(\boldsymbol{\theta}_k|m_k)$ is the prior density of $\boldsymbol{\theta}_k$ under model m_k and $p(\mathbf{D}|m_k, \boldsymbol{\theta}_k)$ is the likelihood function. That is, $p(\mathbf{D}|m_k)$ is the marginal likelihood for model m_k obtained by integrating the likelihood function with respect to the prior distribution of the parameters.

In practice the integration in Equation (2.27) is often not analytically tractable and thus approximations are needed, for example the methods described in Chib and Jeliazkov (2001). We note that the method of approximating the marginal likelihood by computing the harmonic mean of the likelihood with respect to posterior samples is generally frowned upon by the statistical community, and was even branded as

‘The Harmonic Mean of the Likelihood: Worst Monte Carlo Method Ever’ (Neal (2008)). In addition, the posterior model probabilities are by construction quite sensitive to the choice of prior $p(\boldsymbol{\theta}_k|m_k)$.

Model choice could be based on the posterior model probabilities by simply choosing model m_k with the highest $p(m_k|\mathbf{D})$. However, in a Bayesian framework it is not necessary to limit ourselves to only one model (which may even have a posterior probability of being the true model much less than 1) and hence Bayesian model averaging can be used.

2.6.1 Bayesian model averaging

In order to perform inference or make predictions using a combination/ensemble of models, one may use a method such as *Bayesian model averaging*. The basic concept is to calculate a weighted average of the output from each model, where the weights are the posterior model probabilities.

If a quantity of interest is Δ which models m_1, \dots, m_K aim to shed light on (such as the outcome of an association football match, or a particular parameter common to all models), then the posterior distribution of Δ using Bayesian model averaging is:

$$p(\Delta|\mathbf{D}) = \sum_{i=1}^K p(\Delta|m_i, \mathbf{D})p(m_i|\mathbf{D}). \quad (2.28)$$

It has been suggested by Hoeting et al. (1999); Madigan and Raftery (1994) that averaging over all models in this fashion provides a better average predictive ability than any single model. An example of Bayesian model averaging can be found in Chapter 5 Section 5.6 which uses RJMCMC to sample posterior model probabilities from a large collection of models in order to estimate a particular utility function.

2.6.2 The Bayes factor

The actual values of posterior model probabilities can however be somewhat misleading. They do not necessarily imply the probability that a particular model is the true model, but the probability of one model in relation to another. For example when the number of models under consideration increases, under a uniform prior on the models, the posterior model probability of any given model will decrease. Thus, it is intuitive to consider ratios of evidence in favour of models, known as the Bayes factor, which do not change with the number of models under consideration.

The Bayes factor appeared in 1939 in the first text to develop a fundamental theory

$B_{x,y}$	Interpretation
1 to 3	Not worth more than a bare mention
3 to 20	Positive
20 to 150	Strong
> 150	Very strong

Table 2.1: An interpretation of Bayes factors from Kass and Raftery (1995)

of scientific inference based on Bayesian statistics, *Theory of Probability* (see Jeffreys (1939)). The Bayes factor for comparison of models m_1 and m_2 is:

$$BF_{1,2} = \frac{p(\mathbf{D}|m_1)}{p(\mathbf{D}|m_2)} \quad (2.29)$$

where $p(\mathbf{D}|m_1)$ and $p(\mathbf{D}|m_2)$ are calculated as in Equation (2.27). The Bayes factor is thus linked to the ratio of posterior model probabilities via:

$$\frac{p(m_1|\mathbf{D})}{p(m_2|\mathbf{D})} = BF_{1,2} \times \frac{p(m_1)}{p(m_2)}. \quad (2.30)$$

Literature often cites (somewhat arbitrary) tables which show how much evidence the Bayes factor provides in favour of model m_1 against model m_2 , the most popular being by Kass and Raftery (1995) which is shown in Table 2.1.

2.6.3 The Jeffreys-Lindley paradox

There are however inherent problems with the above methods which use posterior model probabilities. Typically, the marginal likelihood (Equation (2.27)) is very sensitive to the choice of prior distribution $p(\boldsymbol{\theta}_k|m_k)$. Choosing a very non-informative prior (which is often used in Bayesian analysis) can mean that when comparing two nested models the Bayes factor may favour the more parsimonious model even when a classical hypothesis test clearly rejects it. This clear contradiction is often referred to as the *Jeffreys-Lindley paradox* after Lindley (1957) further discussed the paradox based on findings in Jeffreys (1939). The intuition behind the paradox is that even though a more complex model may have a much higher posterior mode, it is likely very peaked, and the additional dimensionality means that the integration (seen in Equation (2.27)) of the likelihood with respect to the prior spans over a larger range of low density.

The Jeffreys-Lindley paradox is probably best illustrated with an example (taken from Wikipedia (2014)). Consider a certain city where $m' = 49,581$ males are born from a total of $n = 98,451$ births. We assume the number of male births M is a binomial random variable such that $M \sim Bin(n, p)$. We are interested in testing

whether p is 0.5 or some other value. That is, we test the hypothesis $H_0: p = 0.5$ against $H_1: p \neq 0.5$.

The classical approach to the hypothesis test is to calculate a p-value (in this case 2-sided), $2 \times \mathbb{P}(M > m')$ assuming H_0 is true. Using a Normal approximation the p-value is 0.0235, which usually allows for the rejection of H_0 in favour of H_1 (since the p-value is below the magic cut-off value, 0.05).

The Bayesian approach is to compute the posterior probabilities of H_0 and H_1 . We assign equal prior probabilities, $\mathbb{P}(H_0) = \mathbb{P}(H_1) = 0.5$, and then calculate the posterior model probabilities in a similar fashion to as described in Equation (2.26):

$$\mathbb{P}(H_0|m') = \frac{\mathbb{P}(m'|H_0)}{\mathbb{P}(m'|H_0) + \mathbb{P}(m'|H_1)}. \quad (2.31)$$

$\mathbb{P}(m'|H_0)$ simply refers to the probability of observing $M_0 = m'$ where $M_0 \sim \text{Bin}(n, 0.5)$ (determined by H_0) and $\mathbb{P}(m'|H_1)$ is the probability of observing $M_1 = m'$ where $M_1 \sim \text{Bin}(n, p_1)$ and $p_1 \sim U(0, 1)$ (determined by H_1). Therefore $\mathbb{P}(m'|H_0) = 0.0001951$ and:

$$\begin{aligned} \mathbb{P}(m'|H_1) &= \int_0^1 \binom{n}{m'} p^{m'} (1-p)^{n-m'} dp \\ &= \binom{n}{m'} B(m' + 1, n - m' + 1) \\ &= 0.00001016 \end{aligned} \quad (2.32)$$

by rearranging the terms in the integral into a beta PDF which integrates to 1 and calculating the beta function B and binomial coefficient on a log scale (to prevent numerical issues). The posterior probabilities are thus $\mathbb{P}(H_0|m') = 0.9505$ and $\mathbb{P}(H_1|m') = 0.04948$. Which is strongly in favour of H_0 .

The root of the disagreement in the two methods is that the classical test looks for evidence against H_0 without making any real reference to H_1 , whereas the Bayesian test directly compares H_0 and H_1 . The tests do not *exactly* contradict, either, one says there is evidence against $p = 0.5$, the other tells that $p = 0.5$ explains the data better than $p \sim U(0, 1)$.

In this thesis we mainly consider Bayesian methods and the above example warns us of the problems of Jeffreys-Lindley paradox when comparing models with posterior model probabilities. The prior of p under H_0 is very informative, $p = 0.5$, and under H_1 it is very vague, $p \sim U(0, 1)$. Although H_1 contains the MLE $\hat{p} = m'/n = 0.5036$, it largely contains values of p which are not in anyway consistent with the data, for example $p < 0.4$ or $p > 0.6$, and hence it is rejected. The same ideas apply to a more complex model having a higher dimension of parameter space and thus scope to contain the best model for explaining the data, but also scope for more space of

low posterior probability, in particular when vague priors are used. As is stated in Robert (2014) one of the main problems with the paradox is that it persists even when the sample size grows to infinity, and is hence why it is still a topic of discussion amongst statisticians, for example in the aptly named ‘Who should be afraid of the Jeffreys-Lindley paradox?’ Spanos (2013).

2.6.4 The Deviance information criterion

The Deviance Information Criterion (DIC) was first proposed by Spiegelhalter et al. (2002) in a lengthy discussion paper and is often used as a model comparison measure in Bayesian analysis. It is simple to calculate and is a readily available statistic from the widely used program WinBUGS (Lunn et al. (2000)), which may have aided in its popularity. Methods like DIC have been proposed due to a desire for model comparison techniques which are not sensitive to the choice of prior distribution or potentially susceptible to the Jeffreys-Lindley paradox (like many methods which make use of the marginal likelihood, for example the Bayes factor). In a similar vein was the proposal of the *posterior Bayes factor* in a discussion paper by Aitkin (1991) which (unfortunately for the author) was not so well received by the statistical community. For example in the discussion Lindley states ‘the method of posterior Bayes factors is seriously flawed and cannot be recommended’.

In a Bayesian framework it can often be difficult to calculate the effective number of parameters in a model. For example when there are several levels of parameterisation in a hierarchical model. Spiegelhalter et al. (2002) provide a method for determining the effective number of model parameters, which they define as:

$$p_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) \quad (2.33)$$

where:

$$D(\boldsymbol{\theta}) = -2 \log(p(\mathbf{D}|\boldsymbol{\theta})) + 2 \log(f(\mathbf{D})). \quad (2.34)$$

$D(\boldsymbol{\theta})$ is termed as the ‘Bayesian Deviance’ which depends on the log-likelihood of the data \mathbf{D} for parameter $\boldsymbol{\theta}$, that is, $\log(p(\mathbf{D}|\boldsymbol{\theta}))$, and a standardising term, $2 \log(f(\mathbf{D}))$, where $f(\mathbf{D})$ is the likelihood of the data under a saturated model. The DIC (Spiegelhalter et al. (2002)) is then proposed as:

$$DIC = D(\bar{\boldsymbol{\theta}}) + 2p_D \quad (2.35)$$

that is, a Bayesian measure of model fit, $D(\bar{\boldsymbol{\theta}})$, penalised by a measure of model complexity, p_D .

When using DIC for model comparison purposes (as we do in Chapter 5 Section 5.5.3) it is sufficient to use $2\log(f(\mathbf{D})) = 0$, since this term will be identical (and thus cancel) over models. DIC is analogous to other model selection information criteria (for example Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), see Burnham and Anderson (2004)) in that it considers a measure of fit to the data, and a measure of model complexity. In a similar fashion, smaller values of DIC correspond to a preferable model. In addition, for models using negligible prior information, DIC is approximately equivalent to AIC (Spiegelhalter et al. (2002)).

In practice, for model comparisons, $\overline{D(\boldsymbol{\theta})}$ is calculated by storing the value of $\log(p(\mathbf{D}|\boldsymbol{\theta}_i))$ for each sample $i = 1, \dots, n$ taken in the MCMC procedure. Then the estimate is:

$$\overline{D(\boldsymbol{\theta})} = \frac{-2}{n} \sum_{i=1}^n \log(p(\mathbf{D}|\boldsymbol{\theta}_i)). \quad (2.36)$$

Similarly, $D(\bar{\boldsymbol{\theta}}) = -2 \log(p(\mathbf{D}|\bar{\boldsymbol{\theta}}))$ where $\bar{\boldsymbol{\theta}}$ is the posterior sample mean. An example of model choice based on DIC is shown in Chapter 5 Section 5.5.3.

2.6.5 The posterior predictive distribution and scoring rules

In this thesis we are naturally concerned with the one-week-ahead forecasting abilities of our chosen models. It is thus only sensible to evaluate models based on how accurately they are able to forecast results. Given a posterior distribution $p(\boldsymbol{\theta}|\mathbf{D})$ and a random variable of interest Δ , the posterior predictive distribution of Δ has probability function:

$$p(\Delta|\mathbf{D}) = \int p(\Delta|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \quad (2.37)$$

which provides an indication of the uncertainty in the prediction of Δ via the uncertainty present in $p(\boldsymbol{\theta}|\mathbf{D})$. Equation (2.37) shows the expectation of $p(\Delta|\boldsymbol{\theta})$ with respect to the posterior $p(\boldsymbol{\theta}|\mathbf{D})$, and so we may approximate $p(\Delta|\mathbf{D})$ with:

$$\hat{p}(\Delta|\mathbf{D}) = \frac{1}{n} \sum_{i=1}^n p(\Delta|\boldsymbol{\theta}_i) \quad (2.38)$$

following from the methodology presented in Equation (2.10).

If we consider the weekly arrival of data and denote \mathbf{D}_w to be the data collected up to week w , then we may calculate the one-week-ahead posterior predictive distribution $\hat{p}(\Delta_{w,i}|\mathbf{D}_{w-1})$ where $\Delta_{w,i}$ is a random variable which denotes the outcome of event i in week w . This implies use of the week $w-1$ posterior distribution, $p(\boldsymbol{\theta}|\mathbf{D}_{w-1})$.

We make use of $\hat{p}(\Delta_{w,i} | \mathbf{D}_{w-1})$ in Chapters 3 and 4 in order to calculate a *scoring rule*. A scoring rule is described by Dawid et al. (2012) as a function measuring the quality of a quoted probability distribution Q for a random variable Δ , in the light of the realised outcome O of Δ . For example if we consider a random variable which denotes the end of match result in association football (a home team win, a draw, or an away team win), the scoring rule measures the quality of our prediction of the result based on the observed result. Furthermore, a scoring rule is known as *strictly proper* if it is uniquely optimised when Q is the true probability distribution of Δ .

We make use of the strictly proper *logarithmic scoring rule* which we calculate as:

$$LSR_w = \sum_{i=1}^{n_w} \log(\hat{\mathbb{P}}(O_{w,i} | \mathbf{D}_{w-1})) \quad (2.39)$$

where n_w is the number of considered events in week w and $O_{w,i}$ is the observed outcome of the random variable $\Delta_{w,i}$ in week w . Larger values of LSR_w then imply high quoted probabilities for the events which were actually realised in week w , an indication of how well a model's probabilities forecast the observed outcomes when using the data up to and including week $w - 1$. Furthermore, we may consider $\sum_w LSR_w$ as an indication of the forecasting ability throughout all weeks w (Gneiting and Raftery (2007)).

One common method for model assessment and selection is *cross-validation* (an excellent reference is Hastie et al. (2009)) where model parameters are inferred using a partition of the available data (training data) and then performance is tested on the remaining partition of data (testing data). The process is typically repeated k times (and the resulting method *k -fold-cross-validation*) so that k non-overlapping testing data partitions comprise the entire data, and the results are averaged. That is, all the data has been used for testing. The idea of methods like cross-validation is to assess how well a model generalises to unseen observations.

Calculation of LSR_w is somewhat similar to cross-validation methods in that the model parameters are inferred on a particular partition of the data and then model performance is assessed on the remaining data. The method proposed for calculation of LSR_w however respects the sequential nature in which the data we consider in this thesis arrives, and so it is a performance metric which we use for the tuning of model parameters and model selection.

We also note that the approximation of the posterior predictive distribution, $\hat{p}(\Delta | \mathbf{D})$, may be used for what is known as *posterior predictive checks*. The notion is to 'check' where an observed outcome O of Δ sits within the distribution $\hat{p}(\Delta | \mathbf{D})$, which in turn suggests the *posterior predictive p-value* (see Meng (1994); Gelman (2013)). Again, we are naturally concerned with forecasting ability of models, since this has a direct relation to bookmakers odds, and so we prefer to evaluate models based

on a metric like $\sum_w LSR_w$. We do however consider a posterior predictive check in Chapter 3 Section 3.3.6. These checks are typically useful to understand where the model is not fitting the data and thus offer insight as to how a model may be improved.

Chapter 3

An adaptive behaviour model for association football using rankings as prior information

3.1 Introduction

We propose a model in which teams are defined by their overall capacity (a single parameter) which they then partition into attacking or defending according to the current time and state of the match. The model is most closely related to that of Dixon and Robinson (1998) through its explicit modelling of goals times. The use of a single parameter to represent a team's strength (as opposed to two parameters) offers parsimony in comparison to other similar models in literature, and also means that a ranking of the teams (based solely on the value of the single parameter) is readily available. We are thus also able to place an informative prior upon the team strength parameters jointly that reflects belief about the relative ranking of the teams. Furthermore, the model developed here offers insight into how teams adapt their behaviour in response to the time and state of a match. Also, models such as the one presented here, which use the extra information of goal times, are relatively uncommon in the literature when compared to models which only use the final count of goals or the final match outcome. We hope to highlight the merits of models which use more detailed data which, in essence, should lead to higher predictive ability, as will be seen in Section 3.4.

The chapter is organised as follows: Section 3.2 presents models in the current literature and introduces a new model for association football. Section 3.3 describes Bayesian computational methods for parameter inference. Section 3.4 compares the predictive performance of four models and compares the probabilities predicted by the models with those used by a major UK bookmaker, Bet365. Lastly, concluding remarks are presented in Section 3.5.

3.2 Association football models

3.2.1 The model of Dixon and Robinson

Dixon and Robinson (1998) proposed a non-homogeneous Poisson process model considering goal times for each of the competing teams. They denoted their best model ‘model VI’ which has the following specification for the instantaneous rates of scoring in match m where team i plays at home against team j :

$$\lambda_m^{DR}(t) = \gamma_h \alpha_i \beta_j \lambda_{xy} \rho(t) + \epsilon_1 t \quad (3.1)$$

$$\mu_m^{DR}(t) = \alpha_j \beta_i \mu_{xy} \rho(t) + \epsilon_2 t \quad (3.2)$$

where $\lambda_m^{DR}(t)$ and $\mu_m^{DR}(t)$ are the instantaneous rates of scoring for team i and team j respectively, γ_h is a constant parameter to represent the home advantage, α_k and β_k are constant parameters representing the attacking and defensive capability of team k respectively, ϵ_1 and ϵ_2 are parameters designed to account for the increase in rate of goals throughout a match for the home and away teams respectively, λ_{xy} and μ_{xy} are parameters which adjust the rates of scoring for the home and away teams respectively based on the current score being $x-y$ (x for the home team and y for the away team), and t is the time elapsed in the current match, $t \in [0, 1]$. Note that time is measured on a scale for which 1 unit is equivalent to 90 minutes and thus observed injury time goals will be assigned to $t = 0.5$ or $t = 1.0$. The parameter $\rho(t)$ is included to account for recording of injury-time goals as occurring in the last minute of the respective half of the match and is specified as:

$$\rho(t) = \begin{cases} \rho_1 & \text{if } t \in (44/90, 45/90] \\ \rho_2 & \text{if } t \in (89/90, 90/90] \\ 1 & \text{otherwise.} \end{cases} \quad (3.3)$$

The parameter λ_{xy} is defined as:

$$\lambda_{xy} = \begin{cases} \lambda_{10} & \text{if the current score is 1-0} \\ \lambda_{01} & \text{if the current score is 0-1} \\ \lambda_{21} & \text{if the home team is winning and the score is not 1-0} \\ \lambda_{12} & \text{if the away team is winning and the score is not 0-1} \\ 1 & \text{otherwise} \end{cases} \quad (3.4)$$

and the parameter μ_{xy} is defined similarly. To provide model identifiability, the constraint $\frac{1}{20} \sum_{k=1}^{20} \alpha_k = 1$ is necessary.

With its use of separate parameters α_k , $1 \leq k \leq 19$, and β_l , $1 \leq l \leq 20$ the model has a high number of parameters. However it is natural to assume (and this can be

seen in the sample of parameter estimates in Dixon and Robinson (1998)) that there is a strong correlation between the attacking strength and the defensive strength of a team. Typically, the top teams have the best attacking and the best defensive strengths, and the bottom teams have the worst attacking and the worst defensive strengths. Thus, it seems sensible to instead consider a model which only uses a single parameter to represent the overall strength of a team.

3.2.2 The Bradley-Terry model

The Bradley-Terry model was first proposed by Bradley and Terry (1952) and has been used for forecasting association football results (see Cattelan et al. (2013); Knorr-Held (2000); Fahrmeir and Tutz (1994)). The Bradley-Terry model for a 3-outcome event (denoted $Y \in \{0, 1, 2\}$ where 0, 1, and 2 denote the events ‘away team win’, ‘draw’, and ‘home team win’ respectively) where team i plays at home to team j is specified by the following probability:

$$\mathbb{P}(Y \leq y) = \frac{\exp(\delta_y - (h + S_i - S_j))}{1 + \exp(\delta_y - (h + S_i - S_j))} \quad (3.5)$$

where $-\infty < \delta_0 < \delta_1 < \delta_2 = \infty$ are the threshold parameters and S_k represents the ability of team k . The constraint $\delta_0 = -\delta$ and $\delta_1 = \delta$ with $\delta \geq 0$ ensures that home and away teams have the same probability of winning if there is no home advantage ($h = 0$) and the team abilities are equal ($S_i = S_j$). Larger values of δ correspond to a larger probability of the draw outcome, $Y = 1$. The constraint $\sum_{k=1}^{20} S_k = 0$ is necessary for model identifiability. In contrast to the model of Dixon and Robinson (1998), the Bradley-Terry model represents only the probabilities of a home win, draw, or away win as opposed to modelling the goal times between two competing teams. Thus, it is a natural choice of model when considering the 1X2 betting market, as we do in this paper. However, in contrast with the model we propose, the richness of available data which includes comprehensive information on goal times, cannot be exploited directly by this class of model.

3.2.3 A new non-homogeneous Poisson process model

The model we propose is related to the approach taken by Dixon and Robinson (1998), in that it is also a non-homogeneous Poisson process model. However we replace the two parameters representing attacking and defensive strengths for each team with a single parameter, denoted R_k for team k , representing the total resource (i.e. capacity) of that team. We then define the function $\alpha_k(t)$ to be the proportion of resource team k puts into attacking at time t , leaving $1 - \alpha_k(t)$ as the proportion of

resource allocated to defence. Thus $\alpha_k(t)$ attempts to describe how teams typically behave, whether offensively or defensively, through time in a single match.

The intuitive notion behind this model is that teams divide a finite amount of resource between attacking and defending. Shifting the balance of resource towards the former tends to increase the rate of scoring but also conceding goals, while shifting towards the latter, tends to reduce both the chance of scoring and of conceding. The model is defined by the following instantaneous rates of scoring in match m where team i plays at home against team j :

$$\log(\lambda_m(t)) = h + \alpha_i(t)R_i - (1 - \alpha_j(t))R_j + \rho(t) \quad (3.6)$$

$$\log(\mu_m(t)) = a + \alpha_j(t)R_j - (1 - \alpha_i(t))R_i + \rho(t) \quad (3.7)$$

where $\lambda_m(t)$ and $\mu_m(t)$ are the instantaneous rates of scoring for the home and away teams respectively, h and a are parameters representing the baseline scoring rate for any home and away team respectively, and t is as described in Section 3.2.1. However we modify $\rho(t)$ slightly so that:

$$\rho(t) = \begin{cases} \rho_1 & \text{if } t \in (44/90, 45/90] \\ \rho_2 & \text{if } t \in (89/90, 90/90] \\ 0 & \text{otherwise.} \end{cases} \quad (3.8)$$

We also have the constraints $0 \leq \alpha_k(t) \leq 1$ for $t \in [0, 1]$ and $R_k \geq 0$ for all k . Any reasonable form for $\alpha_k(t)$, the proportion of resource team k puts into attacking at time t , may be proposed. Here we consider the formulation:

$$\alpha_k(t) = \begin{cases} c_1 + (d_1 - c_1)t & \text{if team } k \text{ is winning at time } t \\ c_0 + (d_0 - c_0)t & \text{if team } k \text{ is drawing at time } t \\ c_{-1} + (d_{-1} - c_{-1})t & \text{if team } k \text{ is losing at time } t \end{cases} \quad (3.9)$$

where $\alpha_k(t)$ is common to all teams k . This allows us to capture a linear change with time in a team's allocation of resource over three different states, which is a novel and interpretable approach for dealing with the change in a team's behaviour throughout a match. This model specification is also more parsimonious than that of Dixon and Robinson (1998) and more readily describes how teams adapt their behaviour in response to their current situation.

This specification also means that $\alpha_k(0) = c_i$ and $\alpha_k(1) = d_i$ where i is -1 , 0 , or 1 when the team is losing, drawing, or winning respectively. Thus, constraining c_i and d_i to be $\in [0, 1]$ will provide the constraint $0 \leq \alpha_k(t) \leq 1$ for $t \in [0, 1]$. Inferences on these parameters should offer insights into how teams typically react to losing, drawing, or winning throughout a match.

We also test the assumption of linearity in $\alpha_k(t)$ by considering two more general

functions. Firstly, a quadratic polynomial function which we denote $\alpha_k^{(1)}(t)$, specified in order to satisfy $\alpha_k^{(1)}(0) = c_i$ and $\alpha_k^{(1)}(1) = d_i$ (akin to $\alpha_k(t)$), but also $\alpha_k^{(1)}(0.5) = e_i$, giving three additional parameters e_{-1} , e_0 , and e_1 . The specification of $\alpha_k^{(1)}(t)$ is therefore:

$$\alpha_k^{(1)}(t) = \begin{cases} c_1 + (-3c_1 + 4e_1 - d_1)t + (2c_1 - 4e_1 + 2d_1)t^2 & \text{team } k \text{ winning} \\ c_0 + (-3c_0 + 4e_0 - d_0)t + (2c_0 - 4e_0 + 2d_0)t^2 & \text{team } k \text{ drawing} \\ c_{-1} + (-3c_{-1} + 4e_{-1} - d_{-1})t + (2c_{-1} - 4e_{-1} + 2d_{-1})t^2 & \text{team } k \text{ losing.} \end{cases} \quad (3.10)$$

We choose this form as opposed to the usual polynomial, $a + bt + ct^2$, so that $\alpha_k^{(1)}(t)$ reduces to $\alpha_k(t)$ when $e_i - 0.5(c_i + d_i) = 0$ for $i = -1, 0, 1$.

In order to satisfy the constraint, $0 \leq \alpha_k^{(1)}(t) \leq 1$ for $t \in [0, 1]$, we firstly constrain c_i , e_i , and d_i to be $\in [0, 1]$. We then consider the turning points t_i^* of $\alpha_k^{(1)}(t)$. If $t_i^* \notin [0, 1]$ then the constraint must be satisfied. Otherwise, when $t_i^* \in [0, 1]$, the constraint is only satisfied when $0 \leq \alpha_k^{(1)}(t_i^*) \leq 1$.

Secondly, we consider a piecewise linear function which we denote $\alpha_k^{(2)}(t)$. The function comprises 12 parameters, four for each of the three game states (losing, drawing, or winning) which define the value of the function at times 0/90, 30/90, 60/90, and 90/90. We denote the parameters f_i^j where i corresponds to the losing, drawing, or winning state, and j the times 0/90, 30/90, 60/90, and 90/90.

3.3 Bayesian inference for parameter estimation

We now discuss the necessary ingredients for inference in a Bayesian setting of the parameter vectors $\boldsymbol{\theta} = (h, a, c_{-1}, c_0, c_1, d_{-1}, d_0, d_1, \rho_1, \rho_2, R_1, \dots, R_{20})$, $\boldsymbol{\theta}^{(1)} = (\boldsymbol{\theta}, e_{-1}, e_0, e_1)$, and $\boldsymbol{\theta}^{(2)} = (h, a, f_{-1}^{0/90}, f_{-1}^{30/90}, \dots, f_1^{90/90}, \rho_1, \rho_2, R_1, \dots, R_{20})$. We use the 2010/2011 season data to aid with prior elicitation and then study posterior samples taken using data from the 2011/2012 season.

3.3.1 The log-likelihood

We consider data (H_1, \dots, H_{380}) and (A_1, \dots, A_{380}) where H_m is a set containing the times of the home team goals in match m and A_m is defined similarly for the away team. Then, conditioning on the parameter values and assuming that events in different matches are independent conditional on the parameters, the contribution

to the log-likelihood from match m where team i plays at home against team j is:

$$\begin{aligned}\log(L(\boldsymbol{\theta}; m)) = & - \int_0^1 \lambda_m(t) dt - \int_0^1 \mu_m(t) dt \\ & + \sum_{t \in H_m} \log(\lambda_m(t)) + \sum_{t \in A_m} \log(\mu_m(t)).\end{aligned}\quad (3.11)$$

The log-likelihood for an entire season of matches is then given by:

$$\log(L(\boldsymbol{\theta})) = \sum_{m=1}^{380} \log(L(\boldsymbol{\theta}; m)).\quad (3.12)$$

For a reference on the likelihood for such models, see Cox and Lewis (1966) and Dixon and Robinson (1998). The integrals in Equation (3.11) must be calculated piecewise between goal times which change the state of the match (home team winning, a draw, or the away team winning) or change-points in $\rho(t)$ (44/90, 45/90, 89/90, 90/90) so that the integrands are smooth, continuous functions. For start point t_1 and end point t_2 meeting these criteria, and considering the use of the linear function $\alpha_k(t)$, the integral $\int_{t_1}^{t_2} \lambda_m(t) dt$ can be calculated by first defining the following terms:

$$\lambda_m(t) = e^{g(t)}\quad (3.13)$$

$$\begin{aligned}g(t) &= h + \alpha_i(t)R_i - (1 - \alpha_j(t))R_j + \rho(t) \\ &= h + (c_i + (d_i - c_i)t)R_i - (1 - (c_j + (d_j - c_j)t))R_j + \rho(t)\end{aligned}\quad (3.14)$$

$$g'(t) = (d_i - c_i)R_i + (d_j - c_j)R_j.\quad (3.15)$$

We then have:

$$\int_{t_1}^{t_2} \lambda_m(t) dt = \int_{t_1}^{t_2} e^{g(t)} dt\quad (3.16)$$

$$= \frac{1}{g'(t)} [e^{g(t)}]_{t_1}^{t_2}\quad (3.17)$$

$$= \frac{1}{g'(t)} (e^{g(t_2)} - e^{g(t_1)})\quad (3.18)$$

$$= \frac{1}{g'(t)} (\lambda_m(t_2) - \lambda_m(t_1))\quad (3.19)$$

where c_i is c_{-1} , c_0 or c_1 when team i is losing, drawing or winning between the times t_1 and t_2 respectively, d_i follows similarly. The calculation of the integral $\int_{t_1}^{t_2} \mu_m(t) dt$ also follows in a similar fashion to that of $\int_{t_1}^{t_2} \lambda_m(t) dt$.

When considering the function $\alpha_k^{(1)}(t)$, the integral terms in Equation (3.11) must be calculated using numerical integration methods (for which we use Simpson's composite rule). For the piecewise linear function $\alpha_k^{(2)}(t)$, the integrals are calculated similarly to when using $\alpha_k(t)$, but with two additional points from which the integral

much be computed piecewise between, 30/90 and 60/90.

3.3.2 Simulation of goal times

We discuss here the methodology used for the simulation of goal times from the model, since the notation and ideas follow from Section 3.3.1.

The simulation of goal times can be performed by rescaling the time axis of a homogeneous process as is described in Lewis and Shedler (1979); Cinlar (2013). Thus, to simulate the time of a goal from given initial time t_1 , we solve the following equation for t_2 :

$$\int_{t_1}^{t_2} (\lambda_m(t) + \mu_m(t)) dt = \tau \quad (3.20)$$

where τ is a random variate drawn from an Exponential distribution with unit rate. We then have that a goal has occurred at time t_2 and the process is repeated from this point in time. If $t_2 > 1$ then no goal is scored before the game ends. The simulated goal is identified as being scored by the home team, i , with probability:

$$p_h = \frac{\lambda_m(t_2)}{\lambda_m(t_2) + \mu_m(t_2)} \quad (3.21)$$

and so the probability that the goal being scored by the away team, j , is $p_a = 1 - p_h$.

Again considering the use of the linear function $\alpha_k(t)$, we have already seen the form of the integral $\int_{t_1}^{t_2} \lambda_m(t) dt$ in Section 3.3.1, it follows similarly that:

$$\int_{t_1}^{t_2} (\lambda_m(t) + \mu_m(t)) dt = \frac{1}{g'(t)} ((\lambda_m(t_2) + \mu_m(t_2)) - (\lambda_m(t_1) + \mu_m(t_1))) \quad (3.22)$$

$$= \frac{\bar{\lambda}_m + \bar{\mu}_m}{g'(t)} (e^{g'(t)t_2} - e^{g'(t)t_1}) \quad (3.23)$$

where we consider only times t_1 to t_2 where there are no change-points in $\rho(t)$ and we define $\bar{\lambda}_m$ and $\bar{\mu}_m$ as:

$$\bar{\lambda}_m = e^{h+c_i R_i + (c_j-1) R_j + \rho} \quad (3.24)$$

$$\bar{\mu}_m = e^{a+c_j R_j + (c_i-1) R_i + \rho} \quad (3.25)$$

where ρ is the (constant) value of $\rho(t)$ between times t_1 and t_2 so:

$$\lambda_m(t) = \bar{\lambda}_m e^{g'(t)t} \quad (3.26)$$

$$\mu_m(t) = \bar{\mu}_m e^{g'(t)t}. \quad (3.27)$$

This yields the solution:

$$t_2 = \frac{1}{g'(t)} \log \left(\frac{\tau g'(t)}{\bar{\lambda}_m + \bar{\mu}_m} + e^{g'(t)t_1} \right). \quad (3.28)$$

Note the solution in Equation (3.28) is only valid for times t_1 and t_2 which do not cross change-points in $\rho(t)$. The solution for t_2 must be found by checking the integral in (3.20) from t_1 to future $\rho(t)$ change-points after t_1 and checking if the integral is bigger or smaller than τ . We can then determine between which change-points the goal time t_2 lies and use Equation (3.28) to find the exact time. Of course, if the goal time is beyond change-points, we must account for the value of the integral up to those change-points. For example, if (known) $t_1 < 44/90$ and the true (unknown) value of the simulated goal is $t_2 = 70/90$ we would have:

$$\int_{t_1}^{44/90} (\lambda_i(t) + \mu_j(t)) dt + \int_{44/90}^{45/90} (\lambda_i(t) + \mu_j(t)) dt + \int_{45/90}^{t_2} (\lambda_i(t) + \mu_j(t)) dt = \tau \quad (3.29)$$

then in a similar fashion to the derivation of Equation (3.28):

$$t_2 = \frac{1}{g'(t)} \log \left(\frac{\tau' g'(t)}{\bar{\lambda}_i + \bar{\mu}_j} + e^{g'(t)45/90} \right) \quad (3.30)$$

where:

$$\tau' = \tau - \left(\int_{t_1}^{44/90} (\lambda_i(t) + \mu_j(t)) dt + \int_{44/90}^{45/90} (\lambda_i(t) + \mu_j(t)) dt \right). \quad (3.31)$$

Since the model explicitly simulates goal times, it can be used in practice to estimate the probability of any event defined by the times of goals for the two competing teams in a match. Thus, assuming we can approximate the posterior distribution of the model parameters $\boldsymbol{\theta}$, we may use these simulation methods in order to approximate the posterior predictive distribution $\hat{p}(\Delta|\mathbf{D})$. We typically take Δ to be the random variable which denotes the outcome of a particular match (home team win, draw, or away team win), but also consider match scorelines in Section 3.3.6. Again, \mathbf{D} denotes the observed data.

In this chapter (and throughout the thesis) we use sampling methods to approximate the posterior distribution of the model parameters, for each sample we then simulate goal times for a single match and count the proportion of occurrences of our events of interest (the possible outcomes of Δ) in order to calculate $\hat{p}(\Delta|\mathbf{D})$.

When considering the function $\alpha_k^{(1)}(t)$ we must use numerical methods for the simulation of goal times, which follows from the integration in Section 3.3.1 being

analytically intractable. When considering the piecewise linear function $\alpha_k^{(2)}(t)$, the simulation of goal times is similar to $\alpha_k(t)$ when accounting for the two additional change-points in time, 30/90 and 60/90.

3.3.3 Prior choice

We use data from the 2010/2011 season to determine suitable prior distributions as follows:

$$\begin{aligned} h &\sim N(0.4, 0.5^2) & c_i &\sim B(1.5, 1.5) \text{ for } i \in \{-1, 0, 1\} \\ a &\sim N(0.08, 0.5^2) & e_i &\sim U(0, 1) \text{ for } i \in \{-1, 0, 1\} \\ \rho_1 &\sim N(1.098, 0.5^2) & f_i^j &\sim U(0, 1) \text{ for } i \in \{-1, 0, 1\}, \text{ for all } j \\ \rho_2 &\sim N(1.504, 0.5^2) & d_i &\sim B(3, 1) \text{ for } i \in \{-1, 0, 1\}. \end{aligned} \tag{3.32}$$

We note here that the prior parameters are chosen to reflect knowledge on past EPL data, under the Bayesian paradigm. The choice of specific prior parameter values is explained below.

We recall that parameters ρ_1 and ρ_2 account for higher counts of goals (in reality scored in injury time) recorded as occurring in the 45-th and 90-th minute. A reasonable estimate, in the absence of data (injury time minutes are not recorded in the data), for the amount of injury time typically played in the first and second half is 2 and 3.5 minutes respectively. This suggests that there is typically a 3 and 4.5 minute window in which goals can be recorded as 45 and 90 respectively (for example in the first half, the minute of 44 to 45 plus the 2 minutes extra injury time). Thus, we choose ρ_1 and ρ_2 to increase the rate of scoring by an average factor of 3 and 4.5, which is achieved by proposing prior means of $\log(3) = 1.098$ and $\log(4.5) = 1.504$ respectively.

The prior standard deviations for ρ_1 and ρ_2 were given the value 0.5 which results in fairly vague prior distributions, in that 95% prior credible intervals are (0.1186, 2.0786) and (0.5241, 2.4841), corresponding to an increase in the rates of scoring by a factor of (1.1260, 7.9932) and (1.6889, 11.9898) respectively.

We can then define a base scoring rate for the home and away teams defined here by h and a . In the absence of all other parameters, h and a would provide a constant scoring rate for home and away teams respectively, effectively reducing the model to a simple Poisson model. Maximum likelihood estimates of h and a are then 0.48470 (0.04026) and 0.16015 (0.04735) respectively (standard errors shown in brackets) using data from the 2010/2011 season. However, these point estimates do not account for parameters ρ_1 and ρ_2 which both serve to increase the rate of scoring for 1 minute each. We therefore round the prior mean for h down to 0.4 and

decrease the prior mean for a a similar amount to 0.08. We also inflate the standard error estimates by a factor of over 10 to 0.5 for use as the prior standard deviation. This is to account for additional uncertainty in how the prior means of h and a were estimated.

The prior distributions for c_i and d_i are chosen to be fairly non-informative. Other authors (Dixon and Robinson (1998)) and betting markets have suggested that there are more goals in the second half than the first, consistent with teams becoming more attack-minded as a match progresses ($c_i < d_i$). We use the model parameter estimates to determine whether this phenomenon is indicated by our data. Since the parameter e_i is used to test the assumption of linearity in $\alpha_k(t)$, we choose to give it a non-informative prior over the allowable parameter range. However, when using the function $\alpha_k^{(1)}(t)$ we must consider the joint prior $p(c_i, e_i, d_i)$ which is zero when the constraints in Section 3.2.3 are not satisfied and otherwise proportional to $p(c_i)p(e_i)p(d_i)$. Similarly, we place a non-informative uniform prior on the f_i^j parameters.

We now discuss the prior that we propose for the resource parameters $\mathbf{R} = (R_1, \dots, R_{20})$.

3.3.4 A prior for \mathbf{R} using ranking information

For the resource parameter vector \mathbf{R} , a prior which uses only the team's ranking from a previous season is adopted. This small amount of data provides enough information to create an informative prior, producing a simple method to incorporate previous team rankings into parameter inference. The joint prior distribution for the teams' resource parameters is given by:

$$p(\mathbf{R}) \propto f(R_1) \dots f(R_{20}) e^{-\gamma_1 D_1(\mathbf{R})} e^{-\gamma_2 D_2(\mathbf{R})} \quad (3.33)$$

where $D_1(\mathbf{R})$ and $D_2(\mathbf{R})$ are functions which measure the distance between the ordering of \mathbf{R} and a pre-specified prior order. In this example, our prior order is simply the ordering of the previous season's league table, where $D_2(\mathbf{R})$ relates to newly promoted teams, $D_1(\mathbf{R})$ relates to 'surviving' teams, and $f(R_k)$ is the gamma PDF at point R_k with hyperparameters α and β .

Here, $e^{-\gamma_1 D_1(\mathbf{R})}$ and $e^{-\gamma_2 D_2(\mathbf{R})}$ are terms (whose magnitude is determined by the value of γ_1 and γ_2) that penalise differences between the order induced by \mathbf{R} and the previous season's league positions. Conceivably, the performance of this model is improved by having a smaller penalty on the resource of newly promoted teams ($\gamma_1 > \gamma_2$) reflecting our uncertainty in how the newly promoted teams will perform in the EPL. A plausible option for the placing of teams promoted from the Championship in positions 1, 2, and 3 is in positions 18, 19, and 20 of the Premiership.

We choose the functions $D_1(\mathbf{R})$ and $D_2(\mathbf{R})$ to count the minimum number of adjacent swaps made to transform the ordering of \mathbf{R} into the same order as the previous season's league table. Swaps which involve one of the three newly promoted teams contribute to $D_2(\mathbf{R})$, otherwise swaps contribute to $D_1(\mathbf{R})$. For example, if the previous season ended with Manchester United and Chelsea in 1st and 2nd place respectively, and in \mathbf{R} the top two values were $R_{ManchesterUnited} = 1.2$ and $R_{Chelsea} = 1.5$ then there would be 1 swap (contributing to $D_1(\mathbf{R})$) since neither of these teams was newly promoted) to put both of these teams in the correct order.

Algorithm 3 implements the functions $D_1(R)$ and $D_2(R)$ in a similar fashion to a bubble swap algorithm.

We set the gamma density parameters in $f(R_k)$ to $\alpha = 2$ and $\beta = 4$. This choice of prior parameters ensures that the difference between the resource parameters is not too large, as one would expect in a highly competitive league such as the EPL. Furthermore, the gamma density also naturally provides the constraint $R_k \geq 0$ for all k .

To illustrate the properties of the joint prior on the team resources, we simulate from it using a random-walk MH algorithm. Figure 3.1 and Figure 3.2 portray the characteristics of the prior with $\gamma_1 = 1$ and $\gamma_2 = 0.5$ from 200,000 samples for randomly chosen teams. We also show that the marginal density of the prior for R_k is continuous. Let us consider $p(R_1)$, the marginal distribution of R_1 :

$$\begin{aligned} p(R_1) &\propto \int_0^\infty \dots \int_0^\infty f(R_1) \dots f(R_{20}) e^{-\gamma_1 D_1(\mathbf{R})} e^{-\gamma_2 D_2(\mathbf{R})} dR_2 \dots dR_{20} \\ &= f(R_1) \int_0^\infty \dots \int_0^\infty f(R_2) \dots f(R_{20}) e^{-\gamma_1 D_1(\mathbf{R})} e^{-\gamma_2 D_2(\mathbf{R})} dR_2 \dots dR_{20}. \end{aligned} \quad (3.34)$$

Now, this is the expectation of the penalty term $e^{-\gamma_1 D_1(\mathbf{R})} e^{-\gamma_2 D_2(\mathbf{R})}$ with respect to the PDF $f(R_2) \dots f(R_{20})$ and so by defining $\mathbf{R}' = (R'_2, \dots, R'_{20})$ with $R'_k \sim \Gamma(\alpha, \beta)$ for all k independently we have:

$$p(R_1) \propto f(R_1) \mathbb{E}_{\mathbf{R}'}(e^{-\gamma_1 D_1((R_1, \mathbf{R}'))} e^{-\gamma_2 D_2((R_1, \mathbf{R}'))}) \quad (3.35)$$

where we use the notation $\mathbb{E}_{\mathbf{R}'}$ to clarify that the expectation is with respect to \mathbf{R}' . The gamma density function and the expectation of the penalty term must be continuous in R_1 and thus $p(R_1)$ must be also. In contrast to the continuous marginals (samples from which are displayed in Figure 3.1), Figure 3.2 highlights the discontinuity of the prior density for \mathbf{R} in a bivariate setting, there is a clear discontinuity on the line $y = x$.

We treat γ_1 and γ_2 as tuning parameters which are determined using past data (here from the 2010/2011 season, as with the prior parameters discussed in Section 3.3.3)

Algorithm 3 Swap algorithm function

```

1: function SWAPS(D1, D2,  $\mathbf{R}$ )
2:   D1 = 0
3:   D2 = 0
4:
5:    $\triangleright$  index i of  $\mathbf{P}$  is previous league position of team i
6:    $\mathbf{P}$  = previousSeasonPositions()
7:
8:    $\triangleright$  index i of  $\mathbf{O}$  is ranking of team i in  $\mathbf{R}$ 
9:    $\mathbf{O}$  = positions( $\mathbf{R}$ )
10:
11:  madeSwaps = true
12:  while madeSwaps do
13:    madeSwaps = false
14:    for i = 1 to 19 do
15:      firstTeam = teamInPosition(i,  $\mathbf{O}$ )
16:      secondTeam = teamInPosition(i + 1,  $\mathbf{O}$ )
17:      if  $P_{\text{firstTeam}} > P_{\text{secondTeam}}$  then
18:        madeSwaps = true
19:         $O_{\text{firstTeam}} = O_{\text{firstTeam}} + 1$ 
20:         $O_{\text{secondTeam}} = O_{\text{secondTeam}} - 1$ 
21:        if  $P_{\text{firstTeam}} < 18 \&\& P_{\text{secondTeam}} < 18$  then
22:          D1 = D1 + 1;
23:        else
24:          D2 = D2 + 1;
25:        end if
26:      end if
27:    end for
28:  end while
29: end function

```

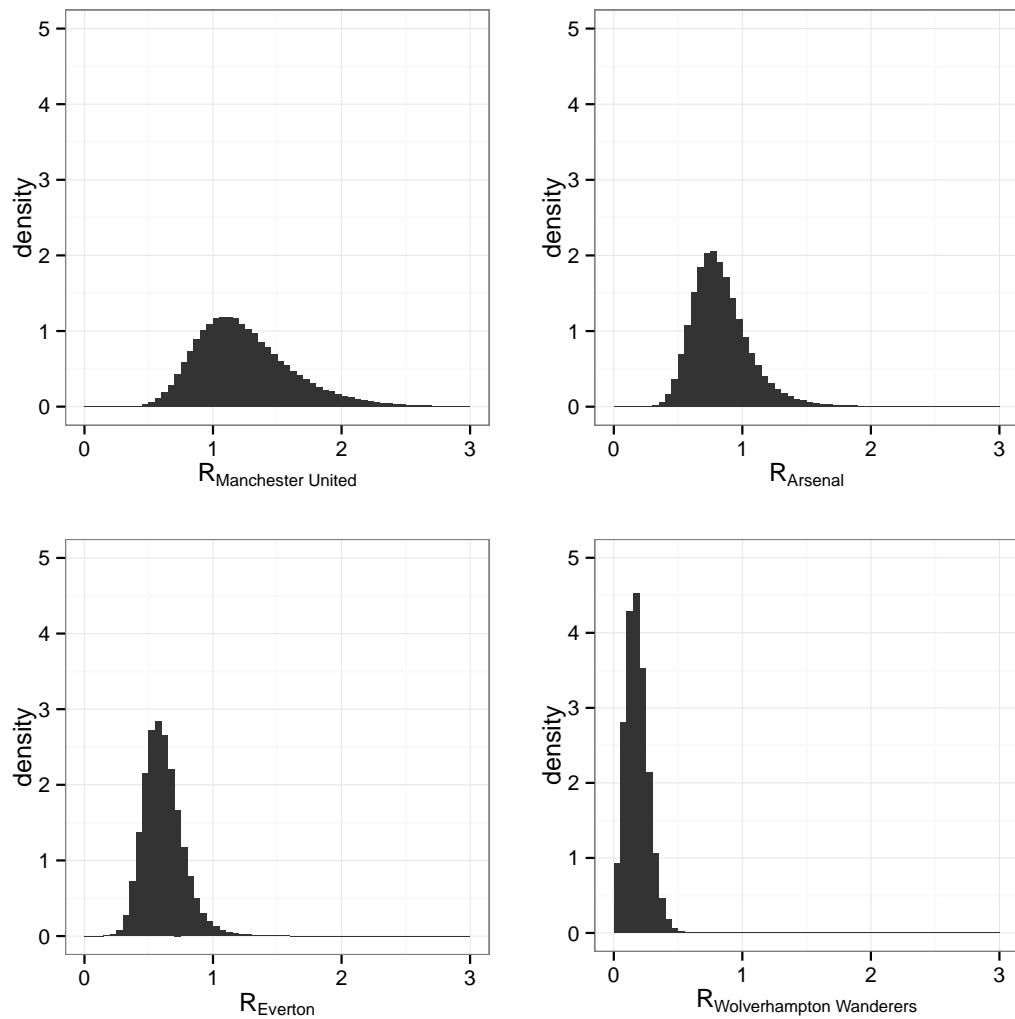


Figure 3.1: Marginal plots of samples from R_k for Manchester United, Arsenal, Everton, and Wolverhampton Wanderers

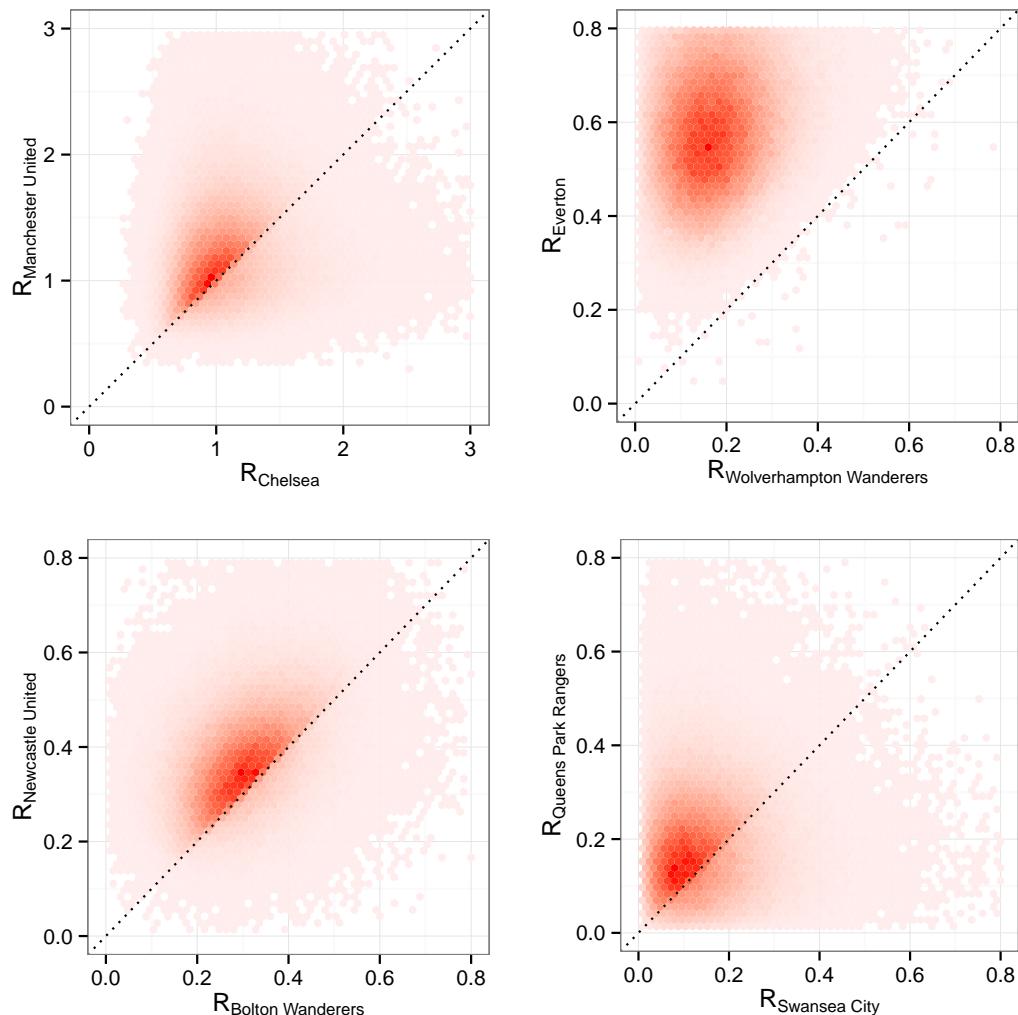


Figure 3.2: Bivariate plots of samples from \mathbf{R} plotted using hexagonal bins. The darker red hexagons represent a higher count of samples in that bin

in an analogous approach to the tuning parameter, λ , in ridge or lasso regression (see for example Tibshirani (1996) or Hastie et al. (2009) which suggest determining the optimal value of λ by cross-validation). In our context it is natural to choose a fixed value of γ_1 and γ_2 in order to maximise the model's one-week-ahead predictive ability. The methodology for doing so is as follows:

For each of the 38 match weeks in the 2010/2011 EPL season we sample from the joint posterior distribution of the parameter space $\boldsymbol{\theta}$ using a random-walk MH algorithm where the log-likelihood contains data up to but not including the match week w for which we wish to make predictions. We then estimate the posterior probability of the observed match results in week w (as per Section 3.3.2) and record the value of the logarithmic scoring rule (see Dawid et al. (2012); Gneiting and Raftery (2007)), LSR_w for each week w :

$$LSR_w = \sum_{m \in M_w} \log(\hat{\mathbb{P}}(O_m | \mathbf{D}_{w-1})) \quad (3.36)$$

where M_w is the set of 10 matches in week w , O_m is the observed outcome of match m (either a home team win, draw, or away team win), and \mathbf{D}_{w-1} is the observed data up to but not including week w . This process is repeated for differing combinations of γ_1 and γ_2 .

We estimated the quantity LSR_w for each week $w = 1, \dots, 38$ with 40,000 posterior samples (after a 5,000 sample burn) and thus estimated $\sum_{w=1}^{38} LSR_w$ which represents the model's forecasting power over the whole season. Optimal values were found empirically when $\gamma_1 = \log(2)$ and $\gamma_2 = \log(1.5)$. We also noted that once γ_1 and γ_2 become sufficiently large the difference in $\sum_{w=1}^{38} LSR_w$ becomes negligible as the probability of accepting any orderings of the teams that contradict the previous seasons order in the MCMC is effectively zero.

3.3.5 Inference results

Firstly, in Figure 3.3 we display the posterior distributions of $e_i - 0.5(c_i + d_i)$ for $i = -1, 0, 1$ using 100,000 posterior samples after a 5,000 sample burn. The function $\alpha_k^{(1)}(t)$ reduces to $\alpha_k(t)$ when $e_i - 0.5(c_i + d_i) = 0$ and the plots show no compelling evidence against this for any i . Secondly, we show plots of 1,500 posterior samples from the functions $\alpha_k(t)$, $\alpha_k^{(1)}(t)$, and $\alpha_k^{(2)}(t)$ in Figure 3.4. Only a small number of samples are shown to avoid plot rendering issues. The plots do not display any real evidence against the suitability of the linear function. Thus for the remainder of the thesis we solely consider the use of the linear function $\alpha_k(t)$.

We now explore the posterior distribution of the model parameters, again, using data from the 2011/2012 season. Visualisations of 100,000 posterior samples (again

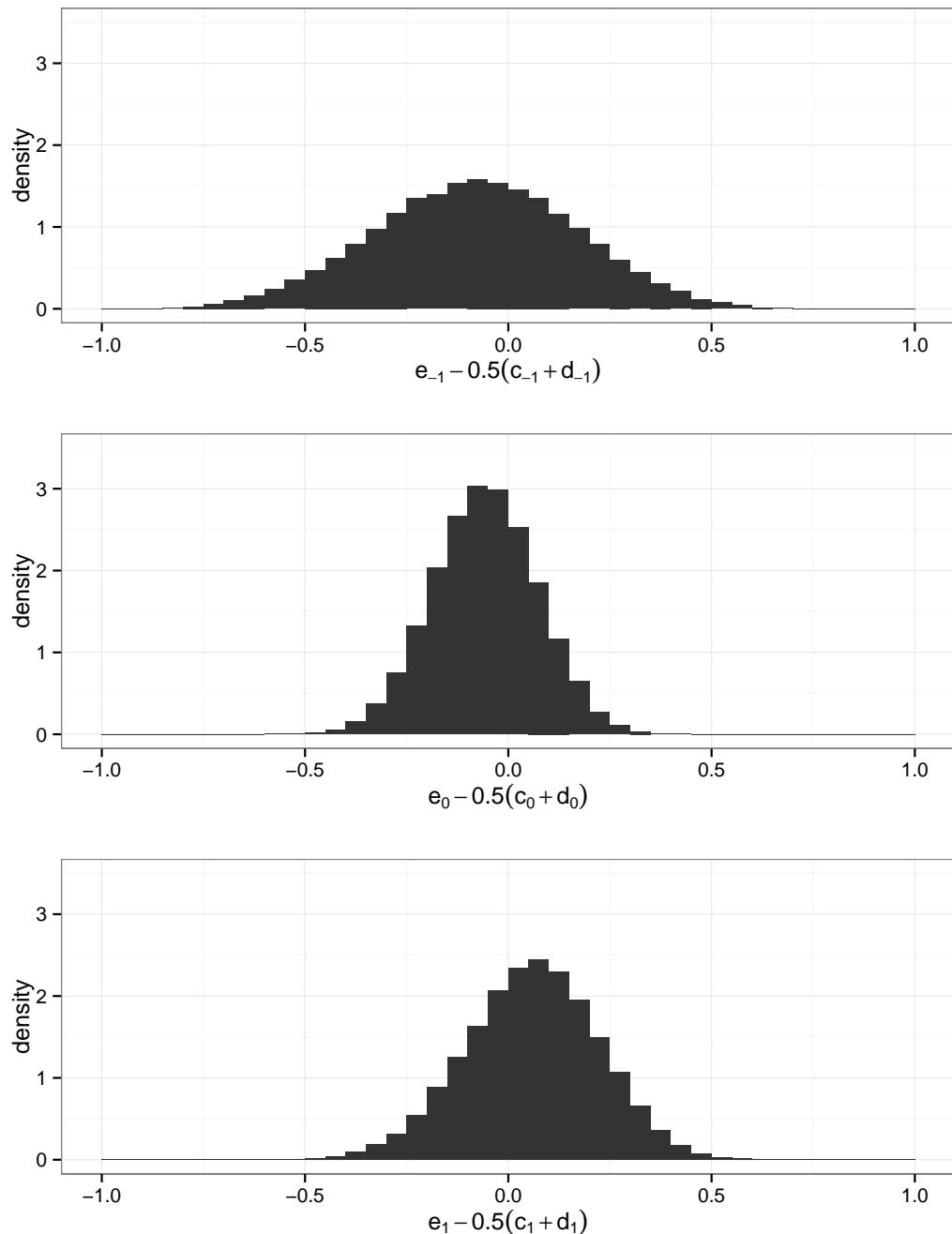


Figure 3.3: Density histograms of the posterior samples for $e_i - 0.5(c_i + d_i)$ for $i = -1$ (top), $i = 0$ (middle), and $i = 1$ (bottom)

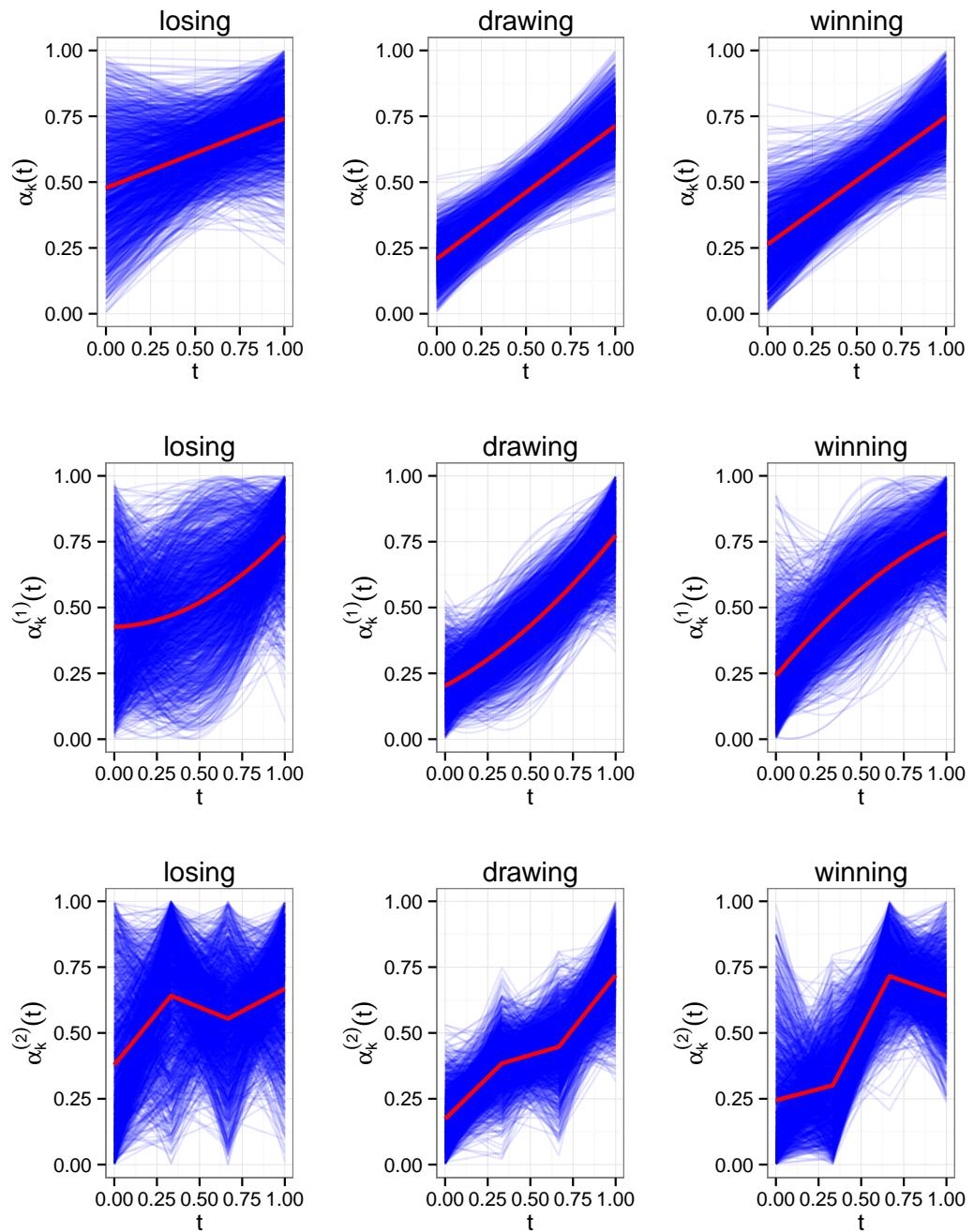


Figure 3.4: Transparent plots of 1,500 posterior samples from the functions $\alpha_k(t)$ (top), $\alpha_k^{(1)}(t)$ (middle), and $\alpha_k^{(2)}(t)$ (bottom) for the states losing, drawing, and winning. — the posterior mean

with 5,000 sample burn) of the team resource parameters are displayed in Figure 3.5 and Figure 3.6. Histogram estimates and trace plots are displayed in Figure 3.7 and Figure 3.8 for the 10 non-resource model parameters, with a corresponding summary shown in Table 3.1.

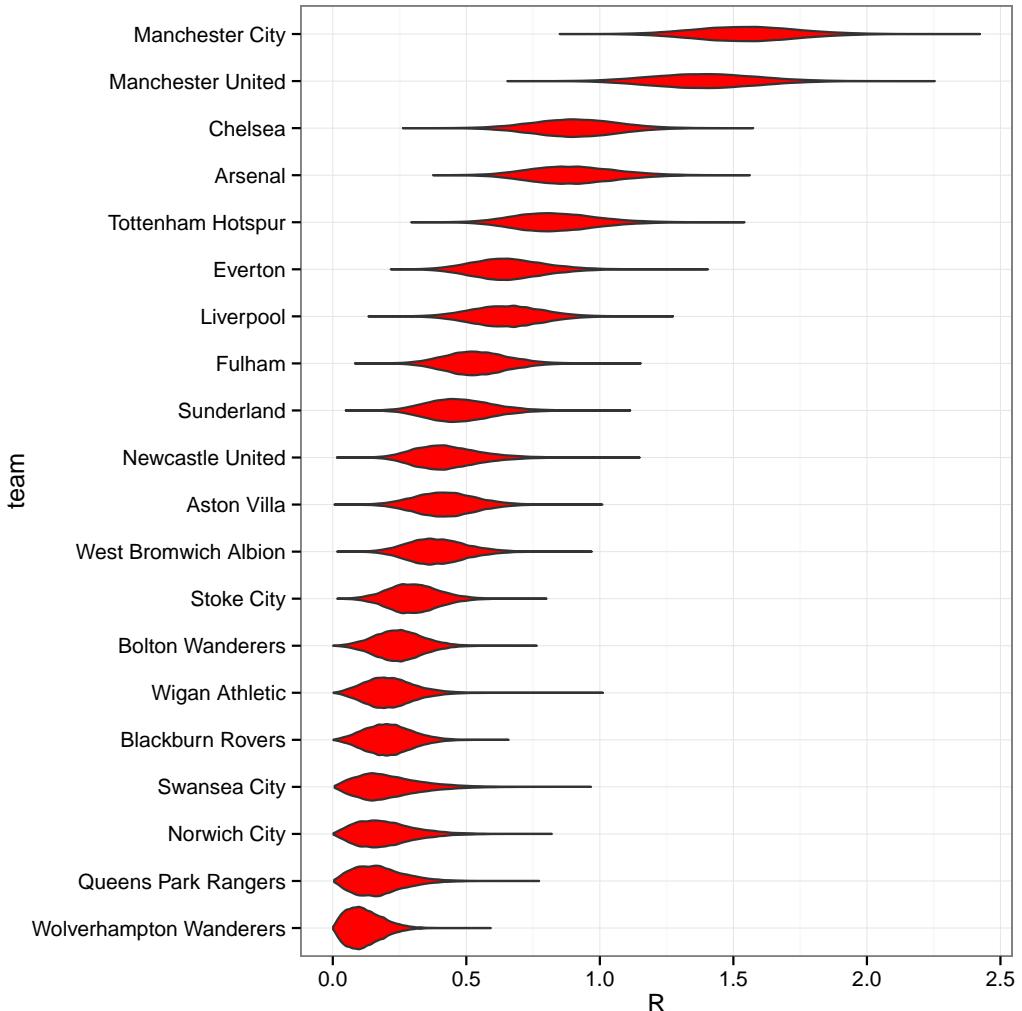


Figure 3.5: A violin plot of the marginal posterior distribution of the teams' resources (R_k for team k). The teams have been ordered by the posterior mean of their resource

The relative ability of each team, as suggested by the marginal posterior distribution of each team's resource, R_k , is displayed in Figure 3.5 through a violin plot. A violin plot is similar to a box-plot but includes a kernel density estimate reflected in the horizontal axis to provide a clearer comparison of multiple densities on a single plot.

Figure 3.6 displays the ranking of each team according to the posterior distribution of the parameter vector \mathbf{R} . That is, each posterior sample taken indicates a ranking of the teams from best to worst, and the posterior distribution of the ranking can be estimated from the entire MCMC output. For each team, this distribution is shown in the plot using shading across each row of the graph. It can be seen that each team's resource largely lies in a single ranking position, although the newly promoted

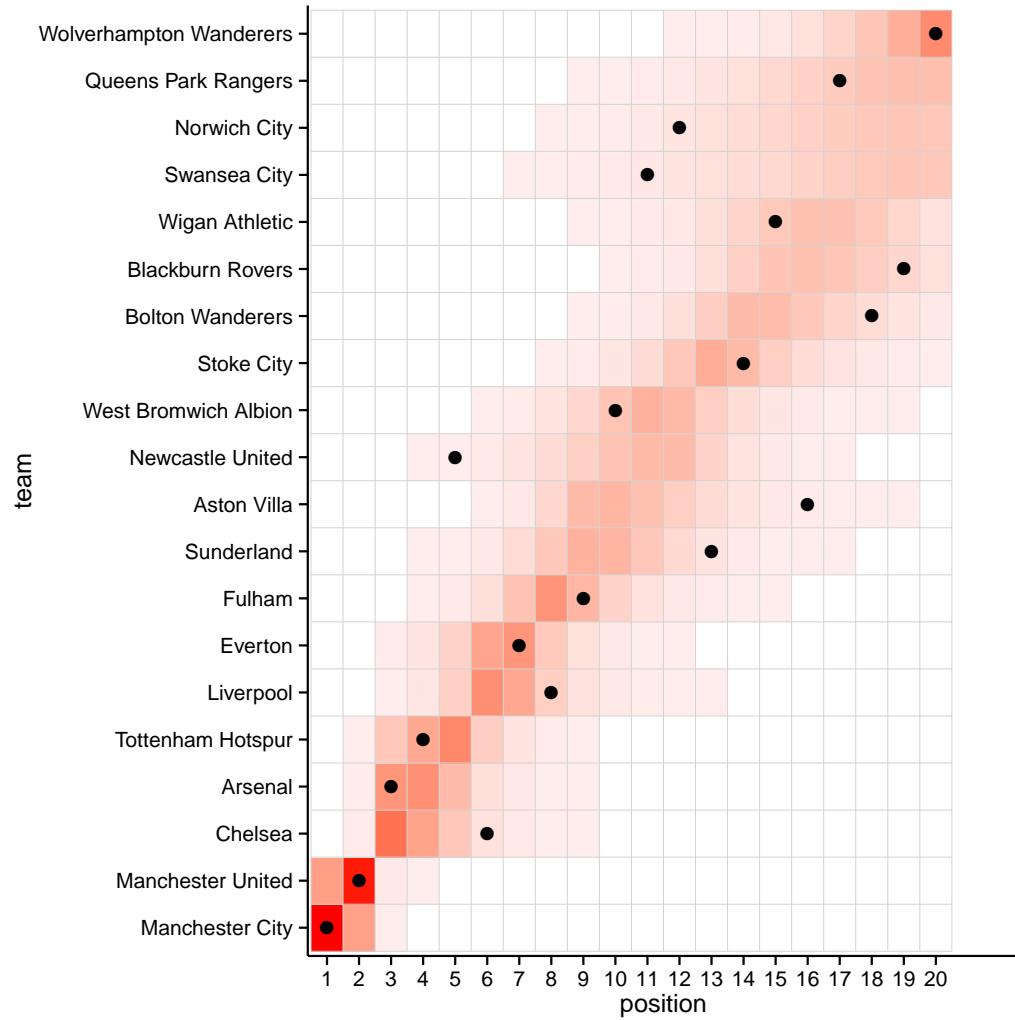


Figure 3.6: A plot displaying the posterior distribution of a ranking of the teams based on the parameter vector \mathbf{R} . Darker regions represent areas of higher posterior probability and again, the teams have been ordered by the posterior mean of their resource. • denote the final league position of the teams in the 2011/2012 season

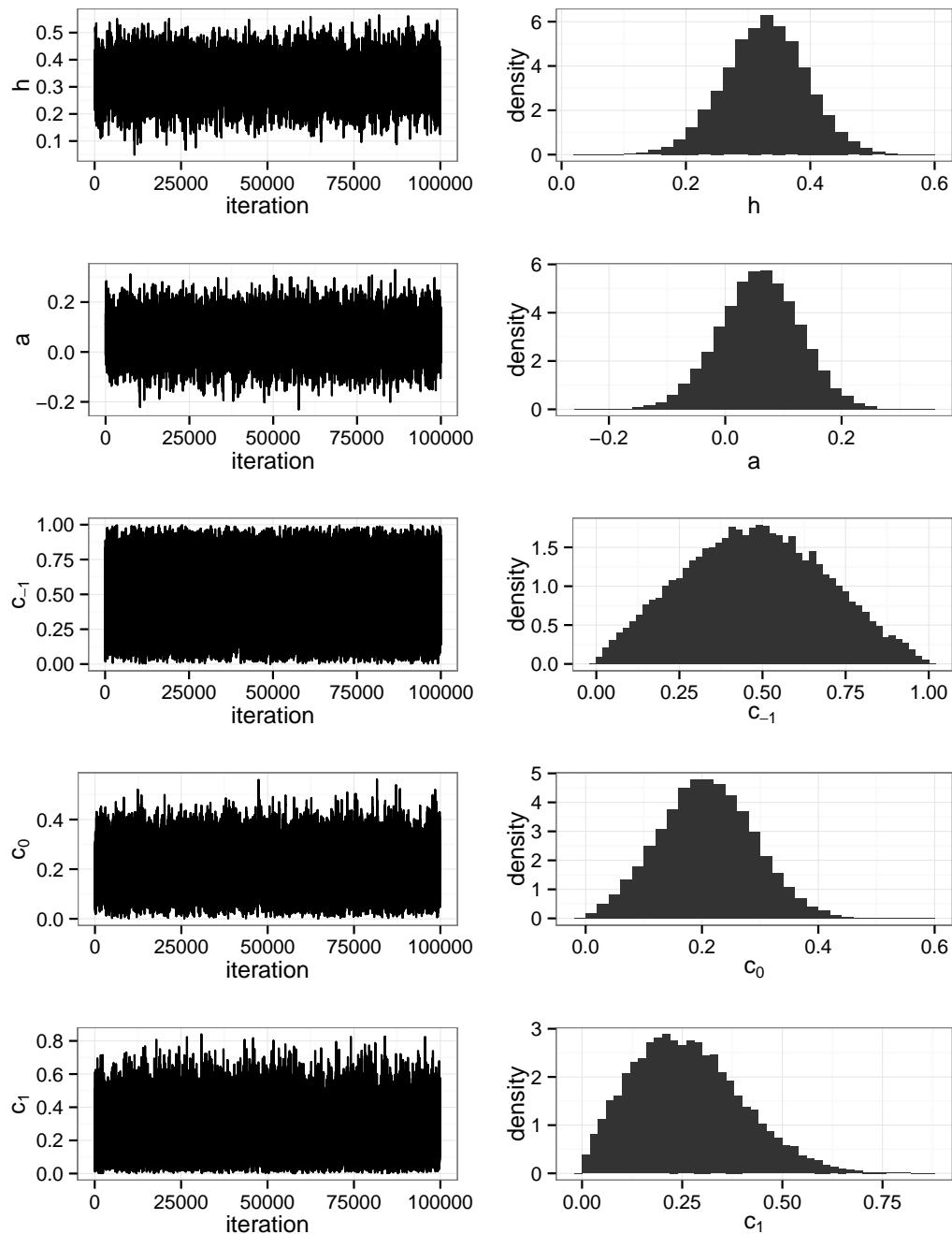


Figure 3.7: Trace plots and density histograms of the posterior samples for parameters h , a , c_{-1} , c_0 , and c_1

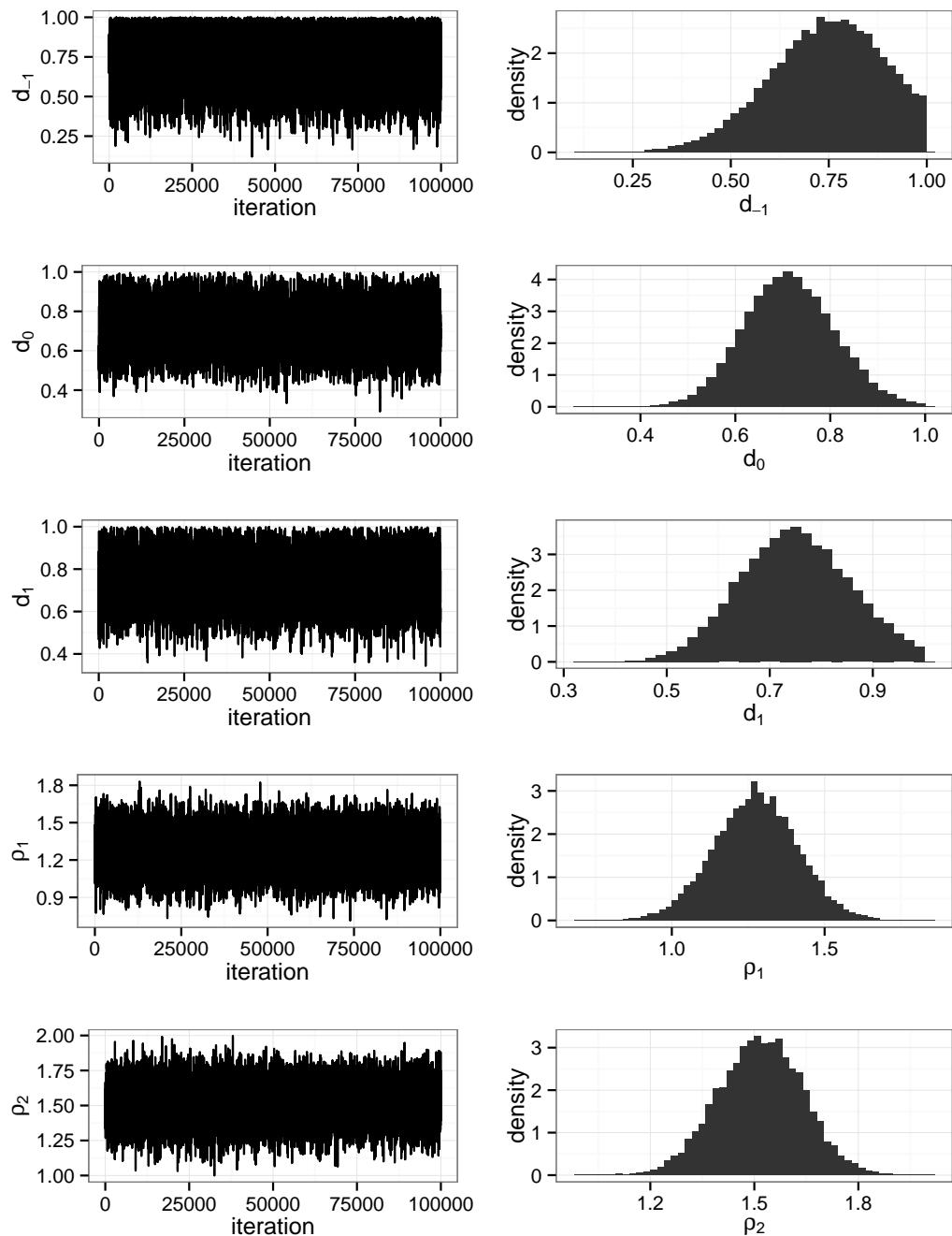


Figure 3.8: Trace plots and density histograms of the posterior samples for parameters d_{-1} , d_0 , d_1 , ρ_1 , and ρ_2

parameter	mean	median	95% BCI
h	0.3277	0.3281	(0.1965, 0.4542)
a	0.0591	0.0598	(−0.0783, 0.1923)
c_{-1}	0.4779	0.4766	(0.0906, 0.8836)
c_0	0.2083	0.2071	(0.0552, 0.3695)
c_1	0.2635	0.2526	(0.0416, 0.5569)
d_{-1}	0.7415	0.7501	(0.4400, 0.9787)
d_0	0.7137	0.7119	(0.5319, 0.9047)
d_1	0.7481	0.7474	(0.5435, 0.9525)
ρ_1	1.2740	1.2747	(1.0045, 1.5391)
ρ_2	1.5169	1.5188	(1.2766, 1.7480)

Table 3.1: A summary of posterior estimates for the 10 non-resource model parameters

teams (Norwich, Swansea, and Queens Park Rangers) are more free to explore other ranks. The plot suggests that Norwich and Swansea may be better teams than suggested by Figure 3.5, where their posterior mean resource was comparatively low.

To test whether or not there is a significant home advantage, that is, $h > a$, it is not sufficient to simply check the posterior summaries in Table 3.1. There may be correlation in the joint posterior distribution of $\boldsymbol{\theta}$ so we consider the posterior probability that $h > a$. This probability is found to be unity from which we conclude that a home advantage clearly exists. The same applies for comparison of the parameters c_i and d_i , for which we estimate posterior probabilities of 0.8210, 0.99996, and 0.9914 in $d_i > c_i$ for $i = -1, 0$, and 1 respectively. Thus, it is clear that teams generally allocate more resource to attack as a match progresses, showing that this model contains the observation that more goals are typically scored later in a match (as found by Dixon and Robinson (1998)). Finally, Table 3.1 displays a suggestion that teams play differently when losing near the start of a match. One might expect teams to play more offensively when they are in a losing state, and this is captured by the parameter c_{-1} which on averages is around twice that of its drawing and winning counterparts, c_0 and c_1 .

In Figure 3.9 we display the dynamics of the model by showing the instantaneous rates of scoring, $\lambda_m(t)$ and $\mu_m(t)$, for one particular match m between Manchester City (the home team) and Queens Park Rangers (the away team) in the last week of the 2011/2012 season. The posterior mean, $\bar{\boldsymbol{\theta}}$, was used to calculate the rates for simplicity of display. We admit that using such a point estimate is somewhat against Bayesian methodology, but it does allow us to convey the dynamics of the model clearly. The plot visually conveys the extent at which the rates of scoring change

when the winning, losing, or drawing state changes, along with a gradual increase in scoring rates throughout the match. Furthermore, this match featured two home team goals in second half injury time (recorded as having occurred at minute 90), which coincide with increased rates of scoring due to the parameter $\rho(t)$. We revisit

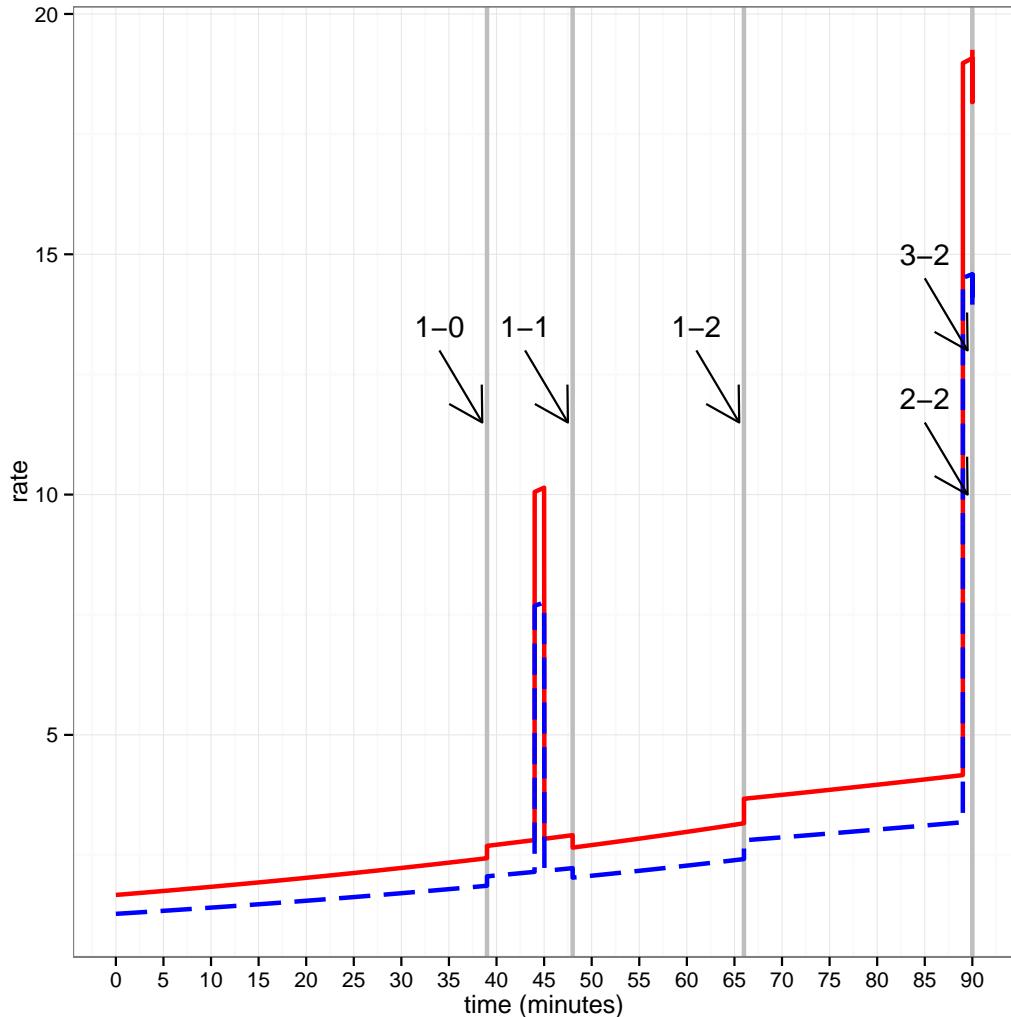


Figure 3.9: A plot of $\lambda_m(t)$ (—) and $\mu_m(t)$ (---) for match m where Manchester City (i) played Queens Park Rangers (j). — denotes goal times, the resulting score in the form $i - j$ is annotated

this particular match in more detail in Chapter 5 Section 5.7.

MCMC convergence was checked via Gelman-Rubin diagnostics (Gelman and Rubin (1992); Brooks and Gelman (1998)) and can be seen in Table 3.2. The diagnostics show no indication of the MCMC procedure not converging to the desired stationary distribution. The multivariate potential scale reduction factor was 1.0024 - sufficiently close to 1.

parameter	point estimate	upper CI
h	1.00012	1.00029
a	1.00042	1.00123
c_{-1}	1.00010	1.00036
c_0	1.00007	1.00025
c_1	1.00001	1.00005
d_{-1}	1.00013	1.00050
d_0	1.00032	1.00084
d_1	1.00036	1.00125
ρ_1	1.00006	1.00021
ρ_2	1.00024	1.00082
$R_{Arsenal}$	1.00021	1.00040
$R_{AstonVilla}$	1.00021	1.00072
$R_{BlackburnRovers}$	1.00033	1.00115
$R_{BoltonWanderers}$	1.00118	1.00412
$R_{Chelsea}$	1.00039	1.00133
$R_{Everton}$	1.00019	1.00053
R_{Fulham}	1.00064	1.00188
$R_{Liverpool}$	1.00053	1.00176
$R_{ManchesterCity}$	1.00013	1.00049
$R_{ManchesterUnited}$	1.00034	1.00115
$R_{NewcastleUnited}$	1.00110	1.00338
$R_{NorwichCity}$	1.00026	1.00087
$R_{QueensParkRangers}$	1.00028	1.00089
$R_{StokeCity}$	1.00092	1.00294
$R_{Sunderland}$	1.00068	1.00188
$R_{SwanseaCity}$	1.00094	1.00289
$R_{TottenhamHotspur}$	1.00029	1.00082
$R_{WestBromwichAlbion}$	1.00044	1.00155
$R_{WiganAthletic}$	1.00025	1.00086
$R_{WolverhamptonWanderers}$	1.00043	1.00159

Table 3.2: Potential scale reduction factors from Gelman-Rubin's convergence diagnostic obtained from three chains

3.3.6 Goodness of fit

Using the methods described in Section 3.3.2, we can use our posterior samples to simulate match scorelines for all 380 matches of the season and examine their distribution, effectively providing a posterior predictive check on the goodness of fit of the model. For each of 40,000 posterior samples we simulate the scores of a single season and thus estimate the posterior predictive expectation of the number of occurrences of each score in a season:

$$\mathbf{E} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 26.4008 & 28.52860 & 18.51570 & 8.86340 & 3.48027 \\ 35.3997 & 42.09020 & 24.84920 & 11.17730 & 4.08060 \\ 28.5913 & 30.80390 & 18.48740 & 7.53418 & 2.60033 \\ 16.8721 & 17.14380 & 9.36785 & 3.78510 & 1.17815 \\ 8.1698 & 7.77698 & 3.97843 & 1.45443 & 0.44690 \end{pmatrix} \end{matrix}. \quad (3.37)$$

The expected frequencies of scores \mathbf{E} can be compared to the observed frequency of score lines:

$$\mathbf{O} = \begin{matrix} & \begin{matrix} 0 & 1 & 2 & 3 & 4 \end{matrix} \\ \begin{matrix} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{matrix} & \begin{pmatrix} 27 & 20 & 21 & 7 & 3 \\ 33 & 45 & 34 & 8 & 2 \\ 30 & 30 & 14 & 11 & 1 \\ 21 & 23 & 7 & 5 & 0 \\ 9 & 4 & 2 & 1 & 2 \end{pmatrix} \end{matrix} \quad (3.38)$$

where entry i, j in the matrix refers to the score i to j , so the home team scores are across rows and the away team scores are across columns. Scores where a team has scored more than four goals have not been displayed to save space, but are used in all calculations.

Informally, apart from score lines of 3-0 and 3-1, the observed and expected scores appear to conform. A more formal test however, is obtained by calculating a χ^2 type statistic for each of our simulated seasons based on the difference between the simulated score lines in that season (denoted \mathbf{S}_i from simulation i) and the expected score lines in matrix \mathbf{E} . We define the following χ^2 type statistic between matrices \mathbf{X} and \mathbf{Y} representing the observed and expected scores respectively:

$$\chi^2(\mathbf{X}, \mathbf{Y}) = \sum_{i=0}^{20} \sum_{j=0}^{20} I(Y_{i,j} \geq 5) \frac{(X_{i,j} - Y_{i,j})^2}{Y_{i,j}} + \frac{(X' - Y')^2}{Y'} \quad (3.39)$$

where:

$$X' = \sum_{i=0}^{20} \sum_{j=0}^{20} I(Y_{i,j} < 5) X_{i,j} \quad (3.40)$$

$$Y' = \sum_{i=0}^{20} \sum_{j=0}^{20} I(Y_{i,j} < 5) Y_{i,j} \quad (3.41)$$

so that score lines with expectation less than five are grouped.

For a further 40,000 simulations we calculate $\chi_i^2 = \chi^2(\mathbf{S}_i, \mathbf{E})$. In essence, we are treating each \mathbf{S}_i as an observed set of score lines and examining the distribution of χ_i^2 for $i = 1, \dots, 40,000$. We can then see how the statistic $\chi^2 = \chi^2(\mathbf{O}, \mathbf{E})$ fits into this distribution, or rather if the observed score lines \mathbf{O} , fit in with a typical simulation from the model. Figure 3.10 shows no evidence against the model's

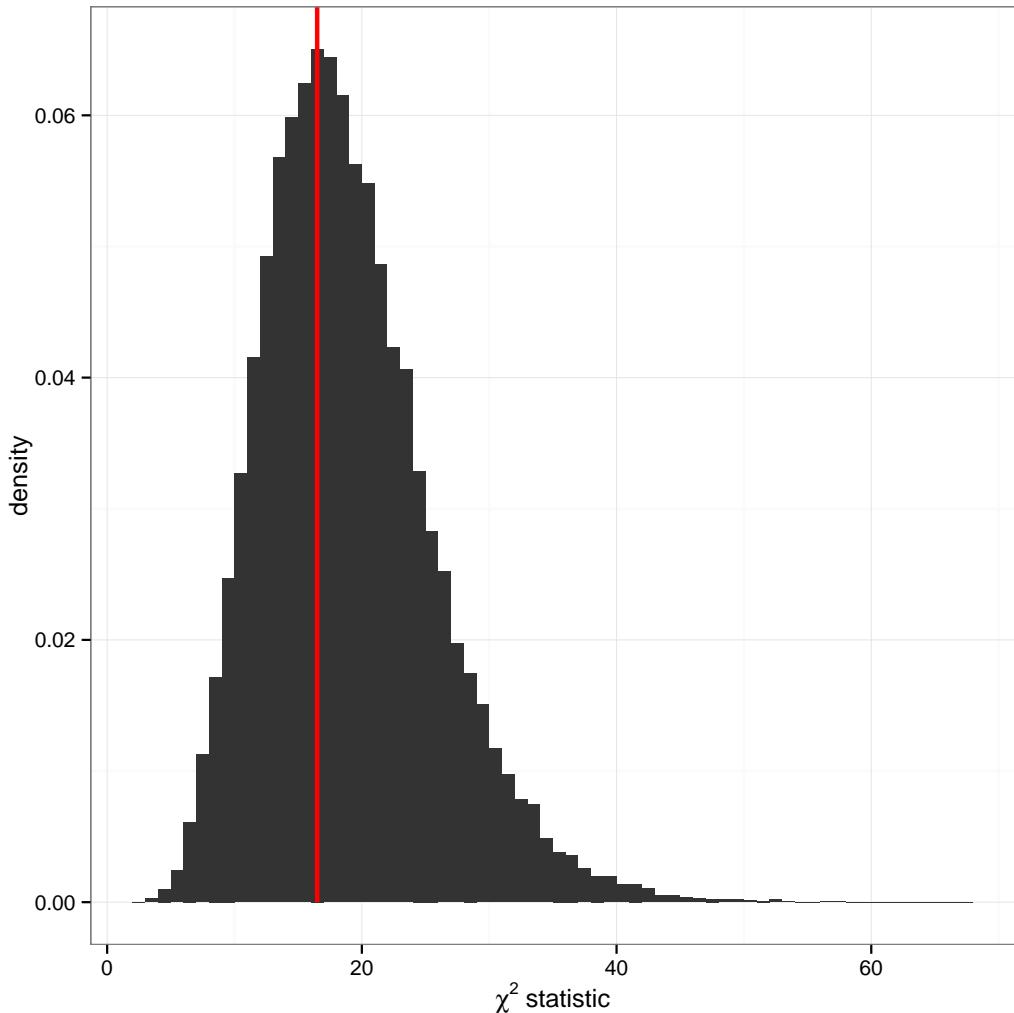


Figure 3.10: A density histogram of the simulated χ_i^2 statistics. — the overall χ^2 statistic

ability to capture the structure of the distribution of score lines, since the observed χ^2 statistic lies at the mode of the distribution of χ_i^2 .

3.4 Model comparisons

We now compare our proposed non-homogeneous Poisson process model (denoted M) to three other models on the basis of their ability to forecast match outcome results one-week-ahead in the EPL 2011/2012 season. The competing models are as follows: the model of Dixon and Robinson (1998) (described in Section 3.2.1) with parameter estimation in a classical framework as in their paper (denoted DR), the Bradley-Terry model (described in Section 3.2.2) with parameter estimation in a classical framework (denoted BTC), and the Bradley-Terry model with parameter estimation in a Bayesian framework using a similar ranking prior to that described in Section 3.3.4 (denoted BTB).

To reiterate, the DR model was chosen as a competing model as it is the closest relative of our proposed model, and the Bradley-Terry type models have been selected since we are assessing model performance based on ability to forecast the home win, draw, away win outcomes, which the Bradley-Terry models directly calculate.

To obtain maximum likelihood estimates of the parameters in models BTC and DR we use the C++ optimisation library ‘NLopt’ (Johnson (2010)). To obtain posterior samples of the parameters in the BTB model we use following prior distributions:

$$p(\mathbf{S}) \propto f(S_1) \dots f(S_{20}) \exp(-\gamma_1 D_1(\mathbf{S})) \exp(-\gamma_2 D_2(\mathbf{S})) \quad (3.42)$$

where $\mathbf{S} = (S_1, \dots, S_{20})$ is the team ability parameter and $f(S_k)$ is the normal density given by $f(S_k) \propto \exp(-\frac{(S_k - \mu)^2}{2\sigma^2})$. The hyperparameters μ and σ^2 are common for all teams with $\mu = 0$ and $\sigma^2 = 5^2$ being used. Prior distributions for parameters h and δ are given by $h \sim N(0.35, 1^2)$ and $\delta \sim \Gamma(3, 5)$. These are chosen to reflect the belief that a home advantage ($h > 0$) is likely and to give realistic probabilities of a draw. If $h = 0.35$ and the competing teams have equal ability, $S_i = S_j$, then a 95% prior credible interval for δ is $(0.1237, 1.4449)$ which corresponds to draw probabilities of 0.0599 and 0.6068, with a reasonable value being 0.2833 when δ is 0.6 (the prior mean). Optimal values for γ_1 and γ_2 for model BTB were found at $\gamma_1 = \log(3.5)$ and $\gamma_2 = \log(1.5)$ using the same methods as in Section 3.3.4.

We chose to begin the comparisons on week six, since after five weeks of matches, the graph, in which vertices are teams and edges link teams that have played each other, becomes connected - a necessary condition for identifiability of the team-ability model parameters. That is, pairs of teams who have yet to play each other are nevertheless comparable by virtue of their links to teams that they have both already played against. Thus, the comparisons use 330 matches of the EPL 2011/2012 season. The home team win, draw and away team win probabilities are estimated by simulation of match outcomes for model M and model DR, and calculated directly

model	$\sum_{w=6}^{38} LSR_w$	$GM_{6,38}$
M	-324.26	0.3743
DR	-337.24	0.3599
BTC	-357.04	0.3389
BTB	-334.56	0.3628

Table 3.3: A comparison of the four competing models in terms of the sum of the logarithmic scoring rule and the geometric mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed, for weeks 6 to 38

for the two Bradley-Terry models BTC and BTB.

3.4.1 Comparison using a scoring rule

We calculate the modelling approach performance metric LSR_w (Equation (3.36)) for weeks 6 to 38 in the season. The overall forecasting ability of each of the models for the whole season can be seen in Table 3.3. We also show the value of:

$$GM_{6,38} = \exp\left(\frac{1}{330} \sum_{w=6}^{38} LSR_w\right) \quad (3.43)$$

which is the geometric mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed - enabling perhaps a more intuitive notion of magnitude of the difference in forecasting ability between the models.

It is clear that the model we have proposed (M) exhibits the best performance as measured by this particular procedure. The results also highlight potential advantages from adopting a Bayesian approach when informative prior distributions can be used, which is often the case in modelling of sporting events, as evidenced by the comparative performance of models BTC and BTB.

We also suggest a test of the significance of the differences in $\sum_{w=6}^{38} LSR_w$ by considering the sampling distribution of LSR_w for each of the four models. This is non-trivial since the LSR_w values do not form an IID sample. That is, for weeks $w = 7, \dots, 38$ we cannot assume the sampling distribution of LSR_w is the same as LSR_{w-1} . The pairs of competing teams are different in each week w and furthermore, the model prediction for the observed outcome of match m in week w ($\hat{\mathbb{P}}(O_m | \mathbf{D}_{w-1})$) which is used to calculate LSR_w depends on the results in the previous weeks (\mathbf{D}_{w-1}). We thus consider H_0 : the distribution of LSR_w given \mathbf{D}_{w-1} is approximately equal for all four models each week w , against H_1 : the distribution of LSR_w is different for at least one of the models each week w . H_0 may only state that the distributions are approximately equal since each model predicts different

probabilities for the outcomes of each match, and so one could immediately prove that the distribution of $\sum_{w=6}^{38} LSR_w$ is slightly different under each model.

We sample from the distribution of $\sum_{w=6}^{38} LSR_w$ under H_0 by iterating through weeks $w = 6, \dots, 38$ and for each w we randomly choose LSR_w under one of the four models. This was repeated 100,000 times and the resulting histogram estimate is shown in Figure 3.11. There is clear evidence that $\sum_{w=6}^{38} LSR_w$ calculated using

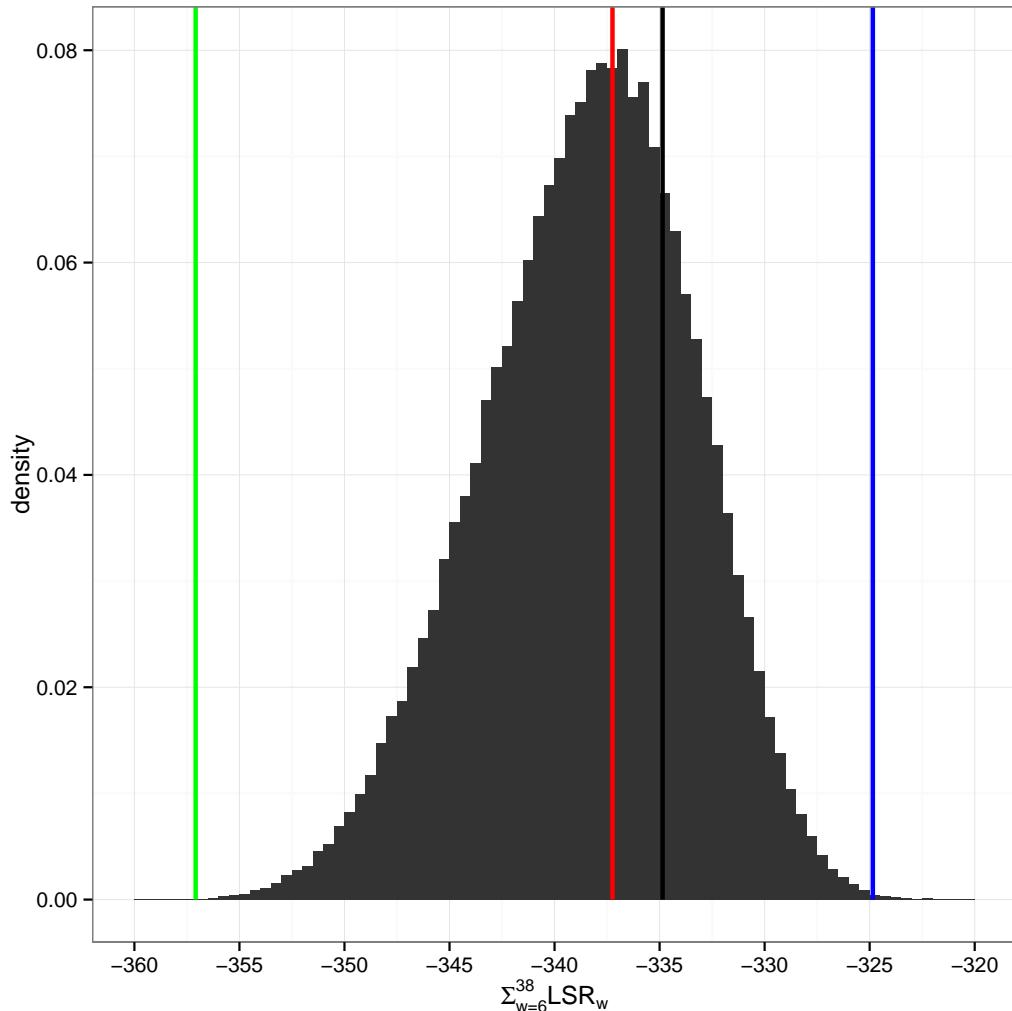


Figure 3.11: A density histogram of $\sum_{w=6}^{38} LSR_w$ under H_0 : the distribution of LSR_w given D_{w-1} is approximately equal for all four models each week w . Individual model estimates of $\sum_{w=6}^{38} LSR_w$ are denoted by — M, — DR, — BTC, — BTB

the estimated probabilities of model M is different (and preferable in terms of performance) from the other models, as the value -324.26 sits very much on the upper tail of the distribution of $\sum_{w=6}^{38} LSR_w$ under H_0 .

A natural question that one might ask after seeing Figure 3.11 is ‘what happens if you remove model BTC which is clearly performing worse than the others?’ We display the corresponding plot, for which model BTC was omitted from the sampling procedure, in Figure 3.12. Again, $\sum_{w=6}^{38} LSR_w$ calculated using the estimated

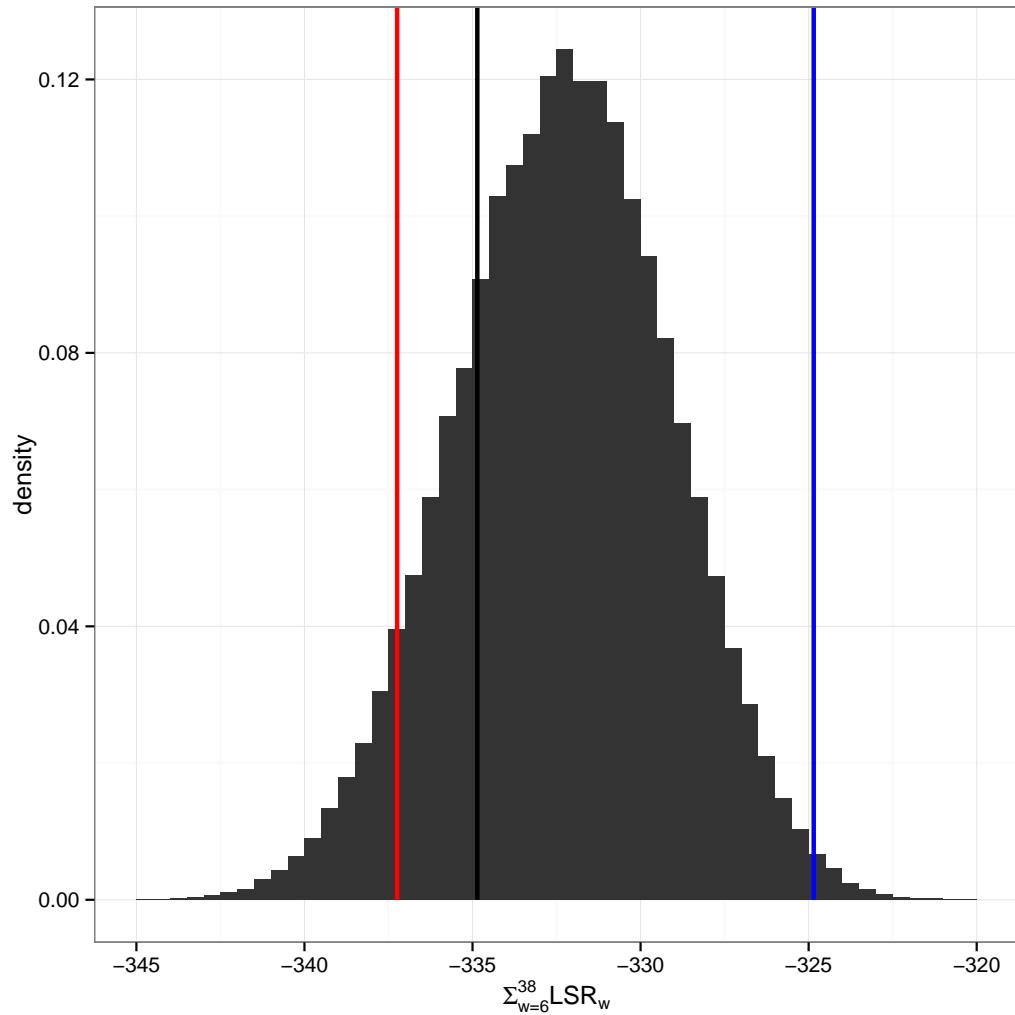


Figure 3.12: A density histogram of $\sum_{w=6}^{38} LSR_w$ under H_0 : the distribution of LSR_w given D_{w-1} is approximately equal for the three best performing models each week w . Individual model estimates of $\sum_{w=6}^{38} LSR_w$ are denoted by — M, — DR, — BTB

probabilities of model M sits on the upper tail of the distribution of $\sum_{w=6}^{38} LSR_w$ under H_0 . The corresponding p-value, $\mathbb{P}(\sum_{w=6}^{38} LSR_w > -324.26 | H_0)$, is 0.0072.

The results of this test also show that if one wished to maximise $\sum_{w=6}^{38} LSR_w$ by using some combination of the four models over the weeks w , then selecting to use model M each every week w provides an almost optimal solution.

3.4.2 A Hosmer-Lemeshow type test

A further test of models involves grouping based on the values of estimated probabilities with the aim of assessing the performance of the models over different ranges of predicted probability. The test is similar to multiple separate Hosmer-Lemeshow tests (Hosmer Jr et al. (2013)) in its grouping based on predicted probabilities and comparison of observed and expected occurrences in each group.

We define the following notation. Let $\mathbb{P}(E_m)$ denote a model's predicted probability of event E in match m , I denote a probability interval (of the form $(x, y]$), $\mathcal{M}_{I,E}$ denote the set of matches m such that $\mathbb{P}(E_m) \in I$ (note that the set $\mathcal{M}_{I,E}$ may contain different matches for the same I when a different model's predicted probabilities are used), and $X_{\mathcal{M}_{I,E}}$ denote a random variable which counts the frequency of event E in the set of matches $\mathcal{M}_{I,E}$, of which we observe the value $O_{\mathcal{M}_{I,E}}$. Using a particular model we have expectation and variance:

$$\mathbb{E}(X_{\mathcal{M}_{I,E}}) = \sum_{m \in \mathcal{M}_{I,E}} \mathbb{P}(E_m) \quad (3.44)$$

$$\mathbb{V}(X_{\mathcal{M}_{I,E}}) = \sum_{m \in \mathcal{M}_{I,E}} \mathbb{P}(E_m)(1 - \mathbb{P}(E_m)). \quad (3.45)$$

We use a normal approximation to obtain a 95% prediction interval around $\mathbb{E}(X_{\mathcal{M}_{I,E}})$ and compare this to $O_{\mathcal{M}_{I,E}}$. The end-points in each probability interval I are chosen to ensure that for each of the four models, $\mathcal{M}_{I,E}$ always contained at least 34 matches, in line with the Hosmer-Lemeshow test which determines the end-points based on probability deciles (33 observations in each of 10 probability intervals). Results of the test are summarised in Figure 3.13. The test does not show any evidence against the fit of model M since the observed frequencies are consistent with the prediction interval for all probability intervals I and types of event E . However it can be seen that the other models may be underperforming in certain ranges of predicted probability where the observed frequencies are outwith the 95% prediction intervals.

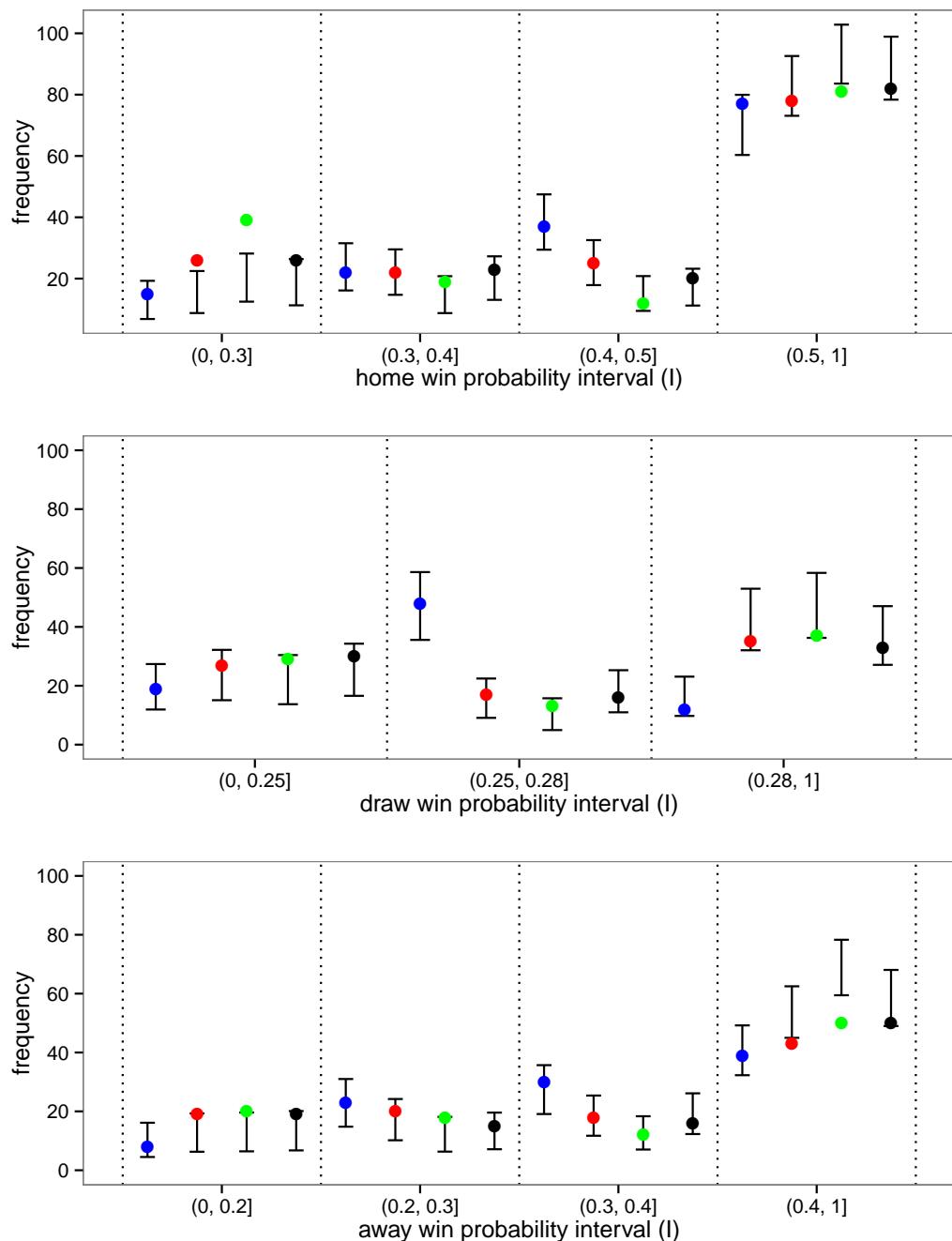


Figure 3.13: A plot of $O_{\mathcal{M}_{I,E}}$ and a 95% prediction interval around $\mathbb{E}(X_{\mathcal{M}_{I,E}})$ for the events H (top), D (middle), and A (bottom). In each probability interval I , $O_{\mathcal{M}_{I,E}}$ is denoted by (•), (●), (○), and (●) for models M, DR, BTC, and BTB respectively, the 95% prediction interval follows similarly

3.4.3 Use of models to inform betting strategies

As mentioned in Chapter 1 Section 1.3, bookmakers add over-round into their estimated probabilities on which their published odds are based. There may nevertheless be scope to make profit when betting against bookmakers when their estimated probability of an event is low.

We study the performance of the models when they are used to inform a naive strategy for selecting bets against UK bookmaker Bet365. We simply place a single unit bet on event E in match m if $\mathbb{P}_m(E) > 1/d_{m,E}$ where $\mathbb{P}_m(E)$ is a model's estimated probability and $1/d_{m,E}$ is the bookmaker's estimated probability plus over-round ($p_{m,E} + o_{m,E}$). We discuss the significance of any of the model's betting profits via a hypothesis test.

Consider H_0 : the bookmaker's estimated probabilities are correct, against H_1 : the use of model probabilities gives the bettor an advantage over the bookmaker. We sample from the distribution of betting profit for each of the model's portfolio of bets under H_0 . In order to do so we firstly uncover the bookmaker's underlying probabilities by assuming $o_{m,H} = o_{m,D} = o_{m,A} = K_m$ and thus:

$$K_m = \frac{1}{3} \left(\frac{1}{d_{m,H}} + \frac{1}{d_{m,D}} + \frac{1}{d_{m,A}} - 1 \right) \quad (3.46)$$

$$p_{m,E} = \frac{1}{d_{m,E}} - K_m. \quad (3.47)$$

The probabilities are then used to simulate match outcomes for the 330 matches under consideration. This is repeated 100,000 times, recording each time the profit from each model's portfolio of bets, providing samples from the distributions of profit using the respective models. We also calculate the profit of each model's portfolio of bets for the observed match outcomes, determining the quantile of this value with respect to each model's distribution of profit under H_0 . This yields a p-value, $\mathbb{P}(P > O_P)$, the probability that the profit under H_0 exceeds the observed profit, O_P .

Results are displayed in Table 3.4 which also shows the expected profit using each model under H_0 . Model M achieves the highest profit betting against the bookmaker and furthermore is the only model which shows significant evidence against H_0 at the 5% level. What is somewhat counter-intuitive, is that model BTB performs particularly poorly on this test. We previously noted in Section 3.4.1 that BTB performed better than BTC based on forecasting ability - and one might then expect model BTB to achieve greater betting profits than BTC. In this instance the opposite is true, but we note that betting profits are very sensitive to the placement of only a small number of bets. For example the difference in betting profits between models

model	O_P	$\mathbb{E}(P)$	$\mathbb{P}(P > O_P)$
M	60.33	-37.61	0.0045
DR	17.47	-30.50	0.0640
BTC	1.43	-25.07	0.1685
BTB	-6.09	-25.36	0.2310

Table 3.4: The results of the hypothesis test H_0 : the bookmaker's estimated probabilities are correct, against H_1 : the use of model probabilities gives the bettor an advantage over the bookmaker

BTC and BTB of 7.52 could be due to the differing placement of a single bet. It is for this reason we suggest a test of the statistical significance of any observed betting profits.

The naive betting strategy can be generalised by considering a difference in the product of the model estimated probability and the bookmaker's odds. We can impose a stricter betting strategy which only places bets when the product exceeds a certain value. That is we place a single unit bet on event E in match m if:

$$\mathbb{P}_m(E)d_{m,E} = \frac{\mathbb{P}_m(E)}{p_{m,E} + o_{m,E}} > r. \quad (3.48)$$

A similar betting strategy is discussed in Dixon and Coles (1997) where the model estimated probabilities are only profitable when $r > 1.1$. In Figure 3.14 we display the profit $O_{P,r}$ for differing values of r using each of the four model's estimated probabilities. As would be expected after the results of the simpler betting strategy, model M almost always provides the largest profit. The plot however shows that a greater profit is achievable for all models using a value of r in the region of 1.2, of course, one would need to determine the optimal value of r before the test to avoid using the data twice.

Figure 3.14 also displays the dynamics of the betting strategy as r varies, in that as $r \rightarrow 0$, the strategy selects all possible bets, and due to the over-round present in the odds, makes a certain loss. Also, as r becomes larger, the strategy selects fewer bets, to the point of selecting none, and the profit is 0.

We are also aware of more complex betting strategies, such as the Kelly betting criterion (Kelly Jr (1956)) which suggests a fraction of the betters bankroll to bet dependant on the estimated probability and the bookmaker's odds - with the aim of maximising the expected log-utility of the bettor's bankroll. More recently, work has appeared in the literature extending the Kelly betting criterion to the case of multiple simultaneous events (Whitrow (2007)). For a bettor wishing to place bets on matches in the EPL, this would prove to be most useful since, as mentioned in Chapter 1 Section 1.4, matches often take place concurrently.

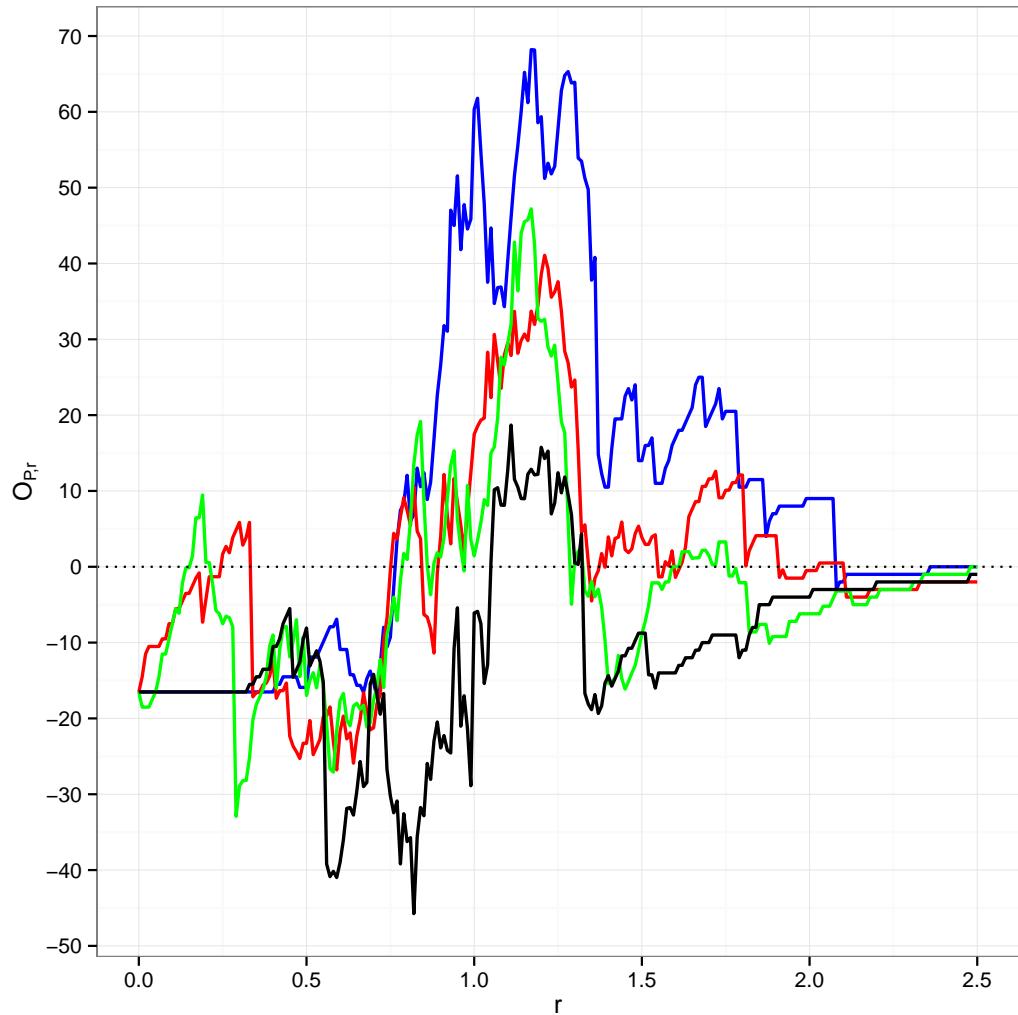


Figure 3.14: A plot of the observed profit $O_{P,r}$ for varying levels of r and when using estimated probabilities from each of the four models. — M, — DR, — BTC, — BTB

We choose however to not investigate the optimal betting strategies further, since the focus of this thesis is the statistical modelling which can be used to inform betting strategies, rather than the design of the strategies themselves.

3.5 Concluding remarks

In this chapter we have described a non-homogeneous Poisson process model for modelling the goal times of competing teams in the EPL. The model is able to capture a linear change with time in a team’s behaviour over the states losing, drawing, and winning. It is also more parsimonious than earlier models in the literature, thus more readily lending itself to novel inference methods.

We have shown that there may be several advantages using the ranking-based prior proposed in Section 3.3.4, most notably by the difference in performance between the Bradley-Terry models in the classical and Bayesian frameworks. One other advantage is that an appropriate team ranking could be easily given by combining expert opinion and/or past data and could therefore be chosen in order to account for team changes between seasons. Thus the ranking-based prior provides a simple method for an expert to specify prior distributions. The ranking-based prior can also easily be adapted to other sports, for example tennis where rankings could simply be determined from the tennis world rankings at the start of the analysis.

The model developed here also outperformed a related model described by Dixon and Robinson (1998) based on one-week-ahead predictive accuracy, and it was also shown that our model is capable of making a significant profit when used to bet against a large UK bookmaker.

We now move onto more practical considerations for the inference of our non-homogeneous Poisson process model, and also present an addition to the model which allows the team resource parameters to vary dynamically throughout a season, as was deemed beneficial by Owen (2011).

Chapter 4

Fast updating of dynamic and static parameters using particle filters

4.1 Introduction

In the domain of sports modelling, several authors have suggested benefits in allowing parameters related to team strength to follow a dynamic system whereby the parameters are assumed to evolve throughout time. Owen (2011) presented a dynamic generalised linear model which allowed parameters representing each team's attacking and defensive strengths to follow a random walk through time, and found that the dynamic model performed better than its non-dynamic counterpart when compared on forecasting ability. Koopman and Lit (2015) modelled the team related parameters as an auto-regressive process and again suggested increased forecasting power when comparing their model with the time invariant version. Several Bradley-Terry type models with an added dynamic component have also been presented in the literature. For example Fahrmeir and Tutz (1994) considered models which contained the team ability parameter following a first-order, second-order, and local linear trend, with parameter inference via empirical Bayes. Knorr-Held (2000) allowed the team ability parameter to follow a Gaussian first-order random walk, with parameter inference via the extended Kalman filter (see, for example, Einicke (2012)). Finally, Cattelan et al. (2013) proposed modelling the team abilities via an exponentially weighted moving average and performed inference via a two step maximisation of the likelihood.

Here we apply methods described in Liu and West (2001) which update parameter beliefs regarding dynamic and non-dynamic (static) parameters as new data arrive. We aim to show that particle filtering methods are computationally fast, accurate, are not limited by assumptions of normality or linearity, and can be straightforward

to implement. Moreover, we discuss whether particle filtering methods and the additional dynamic model component perform better than the static model with inference in a more traditional MCMC framework when comparisons are based on one-week-ahead predictive ability. In addition, we propose that the computational speed of the particle filtering methods may allow for in-play updating of parameter estimates, so one could for example use the most up-to-date posterior distributions to inform in-play betting strategies.

The chapter is organised as follows: Section 4.2 presents a number of particle filtering algorithms and discusses their practical implementation. Section 4.3 describes how the particle filtering algorithms can be modified in order to deal with the inference of dynamic and static model parameters. Section 4.4 presents a modification to our non-homogeneous Poisson process model presented in Chapter 3 Section 3.2.3 which allows the team resource parameters to follow a dynamic system throughout a season, and deals with the inference of all model parameters. Lastly, concluding remarks are presented in Section 4.5.

4.2 Particle filtering methods

We firstly reiterate some of the theory and notation that was introduced in Chapter 2 Section 2.5. The aim of particle filtering methods is to quickly update the posterior belief of a dynamic model parameter \mathbf{x}_t (which we assume to be Markovian) as data \mathbf{y}_t is observed sequentially at each time point t with $\mathbf{D}_t = \{\mathbf{D}_{t-1}, \mathbf{y}_t\}$ denoting the entire data observed up to time t . The methods assume a known model transition density $p(\mathbf{x}_{t+1}|\mathbf{x}_t)$ and likelihood function $p(\mathbf{y}_t|\mathbf{x}_t)$.

4.2.1 The bootstrap filter

A simple but effective first algorithm is the *bootstrap filter* proposed by Gordon et al. (1993), which implements the theory discussed in Chapter 2 Section 2.5. Suppose at time t we have a set of equally weighted samples $\{\mathbf{x}_t^{(1)}, \dots, \mathbf{x}_t^{(n)}\}$ which are approximately distributed as the posterior distribution $p(\mathbf{x}_t|\mathbf{D}_t)$, the bootstrap filter is an algorithm for propagating and updating these samples in order to obtain a new set of equally weighted samples which are approximately distributed as $p(\mathbf{x}_{t+1}|\mathbf{D}_{t+1})$. The algorithm consists of three steps:

1. *Prediction.* The samples are used to generate an approximation of the time $t + 1$ prior, $p(\mathbf{x}_{t+1}|\mathbf{D}_t)$. That is, a new set of samples $\{\mathbf{x}_{t+1}^{(1)*}, \dots, \mathbf{x}_{t+1}^{(n)*}\}$ are drawn from the model transition densities $p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(i)})$

2. *Update.* The data at time $t + 1$, \mathbf{y}_{t+1} , is used to give each prior sample i a normalised weight:

$$\omega^{(i)} = \frac{p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{(i),*})}{\sum_{j=1}^n p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{(j),*})} \quad (4.1)$$

3. *Resampling.* Sample with replacement n times from $\{\mathbf{x}_{t+1}^{(1),*}, \dots, \mathbf{x}_{t+1}^{(n),*}\}$ with probability $\omega^{(i)}$ of picking $\mathbf{x}_{t+1}^{(i),*}$. The new sample $\{\mathbf{x}_{t+1}^{(1)}, \dots, \mathbf{x}_{t+1}^{(n)}\}$ has approximate distribution $p(\mathbf{x}_{t+1} | \mathbf{D}_{t+1})$

The algorithm is initialised at time $t = 0$ by drawing samples from the known prior $p(\mathbf{x}_1 | \mathbf{D}_0) = p(\mathbf{x}_1)$. These samples feed directly into the *update* stage of the filter.

After the *resampling* stage of the filter, particles with larger relative weights will be replicated and particles with smaller weights will be discarded. It is the resampling step which distinguishes the bootstrap filter from a Sequential Importance Sampling (SIS) scheme as it implies repeated applications of the importance sampling and resampling steps. Particle filtering algorithms like the bootstrap filter are thus known as SIR schemes. The SIS scheme updates the weights $w_t^{(i)}$ at each time point t (with no resampling) and typically suffers from a problem known as *particle degeneracy* where only a small proportion of the particles contain nearly all of the weight, see Cappé et al. (2007) for an overview on particle degeneracy and the SIS scheme.

The resampling step described by Gordon et al. (1993) is a simple multinomial sampling procedure. There are however resampling strategies with smaller variance such as residual resampling, stratified resampling and systematic resampling (see Carpenter et al. (1999); Douc and Cappé (2005)). A discussion of residual resampling is in Section 4.2.5.

4.2.2 General particle filter

The more general particle filter can propagate particles with importance (or umbrella) function $q(\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{y}_{t+1})$ as opposed to using the model transition density $p(\mathbf{x}_{t+1} | \mathbf{x}_t)$. This may mean that particles can more adequately move around the space we wish to explore. There is also potential benefit from using an adapted particle filter where the importance density uses the newly arrived data \mathbf{y}_{t+1} (see the ‘likelihood filter’ in Sanjeev Arulampalam et al. (2002)). The weights given to

each particle i in the *update* stage of the bootstrap filter are modified to:

$$\bar{\omega}^{(i)} = \frac{p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1}^{(i)}) p(\mathbf{x}_{t+1}^{(i)} | \mathbf{x}_t^{(i)})}{q(\mathbf{x}_{t+1}^{(i)} | \mathbf{x}_t^{(i)}, \mathbf{y}_{t+1})} \quad (4.2)$$

$$\omega^{(i)} = \frac{\bar{\omega}^{(i)}}{\sum_{j=1}^n \bar{\omega}^{(j)}} \quad (4.3)$$

where $\bar{\omega}^{(i)}$ are the un-normalised weights, so that $\bar{\omega}^{(i)} \propto \omega^{(i)}$. We now consider the use of un-normalised weights, which can always be normalised by dividing each weight by the sum of the weights. If the importance function $q(\mathbf{x}_{t+1} | \mathbf{x}_t, \mathbf{y}_{t+1}) = p(\mathbf{x}_{t+1} | \mathbf{x}_t)$ then the general particle filter reduces to the bootstrap filter.

Pitt and Shephard (1999) describe two basic weaknesses of the particle filtering algorithm:

1. When \mathbf{y}_{t+1} is an outlier, the weights $\bar{\omega}^{(i)}$ will be very unevenly distributed and so the algorithm will require a very large number of particles or a more efficient sampling process. This is of particular concern when the likelihood is very peaked, that is, $p(\mathbf{y}_{t+1} | \mathbf{x}_{t+1})$ is very sensitive to \mathbf{x}_{t+1}
2. Due to the particle filter mixture approximation, the tails of $p(\mathbf{x}_{t+1} | \mathbf{D}_t)$ may be poorly approximated. This can lead to a poor approximation of $p(\mathbf{x}_{t+1} | \mathbf{D}_{t+1})$ when an outlier is observed

A solution to the first point is the *auxiliary particle filter* which has an additional resampling stage which favours particles that are more likely to be consistent with the next data points and is explained in the following section. We also note the problem of selecting an adequate number of particles, as discussed by Boers (1999). In order to alleviate these problems, we opt to run algorithms with a very high number of particles - 100,000.

4.2.3 Auxiliary particle filter

Proposed by Pitt and Shephard (1999), the idea of the auxiliary particle filter is to add an additional step to the general particle filtering method so that particle locations are typically more consistent with the likelihood of the new data \mathbf{y}_{t+1} . Before propagation of the particles (moving them under the importance function), there is an additional resampling step, we assign the first-stage weights for particle i as:

$$\bar{\omega}_1^{(i)} = p(\mathbf{y}_{t+1} | \boldsymbol{\mu}_{t+1}^{(i)}) \quad (4.4)$$

where $\boldsymbol{\mu}_{t+1}^{(i)}$ is some likely value that particle i will evolve to at time $t+1$ such as the mean or the mode of $p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(i)})$. We then sample (with replacement) ‘auxiliary’ indicators j with probabilities proportional to $\bar{\omega}_1^{(i)}$. The particles are then propagated via sampling from the model transition density $p(\mathbf{x}_{t+1}|\mathbf{x}_t^{(j)})$ (based on the auxiliary indicators j) and second-stage weights are:

$$\bar{\omega}_2^{(j)} = \frac{p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^{(j)})}{p(\mathbf{y}_{t+1}|\boldsymbol{\mu}_{t+1}^{(j)})}. \quad (4.5)$$

After a second case of resampling based on the weights $\bar{\omega}_2^{(j)}$, the particles form an equally weighted approximation to $p(\mathbf{x}_{t+1}|\mathbf{D}_{t+1})$.

4.2.4 Practical implementation

In implementation of particle filters, weights are naturally stored on a log scale since calculating the likelihood is often not computationally feasible (its value can be numerically 0 when stored on a computer using double precision). An example of representing weights on the log scale for the general particle filter is as follows:

$$\log(\bar{\omega}^{(j)}) = \log(p(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^{(j)})) + \log(p(\mathbf{x}_{t+1}^{(j)}|\mathbf{x}_t^{(j)})) - \log(q(\mathbf{x}_{t+1}^{(j)}|\mathbf{x}_t^{(j)})). \quad (4.6)$$

We then calculate $m = \max(\log(\bar{\omega}^{(j)}))$ and calculate new un-normalised weights as:

$$\hat{\omega}^{(j)} = e^{\log(\bar{\omega}^{(j)})-m} \quad (4.7)$$

resampling can then be based on the new weights $\hat{\omega}^{(i)}$. The largest weight now has a value of 1 and therefore it and other significant weights can be easily stored on a computer using double precision. Weights that are very small relative to the largest weight may be computationally 0 after we use the exponential function and will not be chosen in the resampling process.

The propagation and weighting steps of the particle filtering algorithm are also good candidates to make use of parallel processing. We use the parallel processing library **OpenMp** (see Dagum and Menon (1998)) which very easily allows a programmer to write for loops which are executed in parallel. Other parallel processing libraries are available, for example **MPI** (see Gropp et al. (1996)) which may result in quicker program run-times, but are notoriously harder to work with.

4.2.5 Resampling methods

Douc and Cappé (2005) describe four methods for particle filter resampling. The

most simple of these is the standard multinomial resampling method, where each particle is selected with probability equal to its weight (or proportional to its unnormalised weight). Stratified and systematic resampling methods are mentioned, but we choose to implement the residual resampling method, which has no mentioned drawbacks and is very intuitive.

Let $N^{(i)}$ count the number of times particle i is chosen in the resampling process. We can then split each $N^{(i)}$ into two parts, one which is deterministic (which lowers the variance of this resampling method compared to multinomial sampling) based on the expectation of $N^{(i)}$, and one which is random (the residual contribution). We define:

$$N^{(i)} = \lfloor n\omega^{(i)} \rfloor + N^{(i),*} \quad (4.8)$$

where $n\omega^{(i)}$ is the expectation of $N^{(i)}$ (since $\omega^{(i)}$ is the normalised weight) and $N^{(i),*}$ follows a multinomial distribution which draws $n - \sum_{i=1}^n \lfloor n\omega^{(i)} \rfloor$ (n is the total number of particles) counts so in total there are n counts. The weights of the multinomial resampling step are:

$$\omega^{(i),*} = n\omega^{(i)} - \lfloor n\omega^{(i)} \rfloor \quad (4.9)$$

so the weights take into account how many draws of that particle have already occurred in the deterministic stage of the resampling method, also note that these weights are not normalised and they sum to $n - \sum_{i=1}^n \lfloor n\omega^{(i)} \rfloor$. Figure 4.1 shows an example of how residual sampling can work when a sample of size 100,000 is taken from a $U(-5, 5)$ distribution and given weights according to a $N(0, 1)$ distribution.

4.3 A mixture of dynamic and static parameters

We have seen how particle filter methods can be used to update our belief regarding dynamic parameters, that is, parameters which are thought to vary throughout time, which we have denoted \mathbf{x}_t . We now consider the case where we have a mixture of dynamic (time varying) and static (non time varying) parameters. The filtering algorithms previously discussed do not work for static parameters as they have no model transition density and thus become stuck in position, not able to explore the posterior distribution. We represent our belief of the static parameter \mathbf{z} at time t via the posterior distribution $p(\mathbf{z}|\mathbf{D}_t)$, and the joint distribution $p(\boldsymbol{\theta}_t|\mathbf{D}_t)$ where $\boldsymbol{\theta}_t = \{\mathbf{z}, \mathbf{x}_t\}$. Note we may use the notation \mathbf{z}_t to denote the time t posterior of the parameter \mathbf{z} - this does not indicate time variation in the parameter \mathbf{z} .

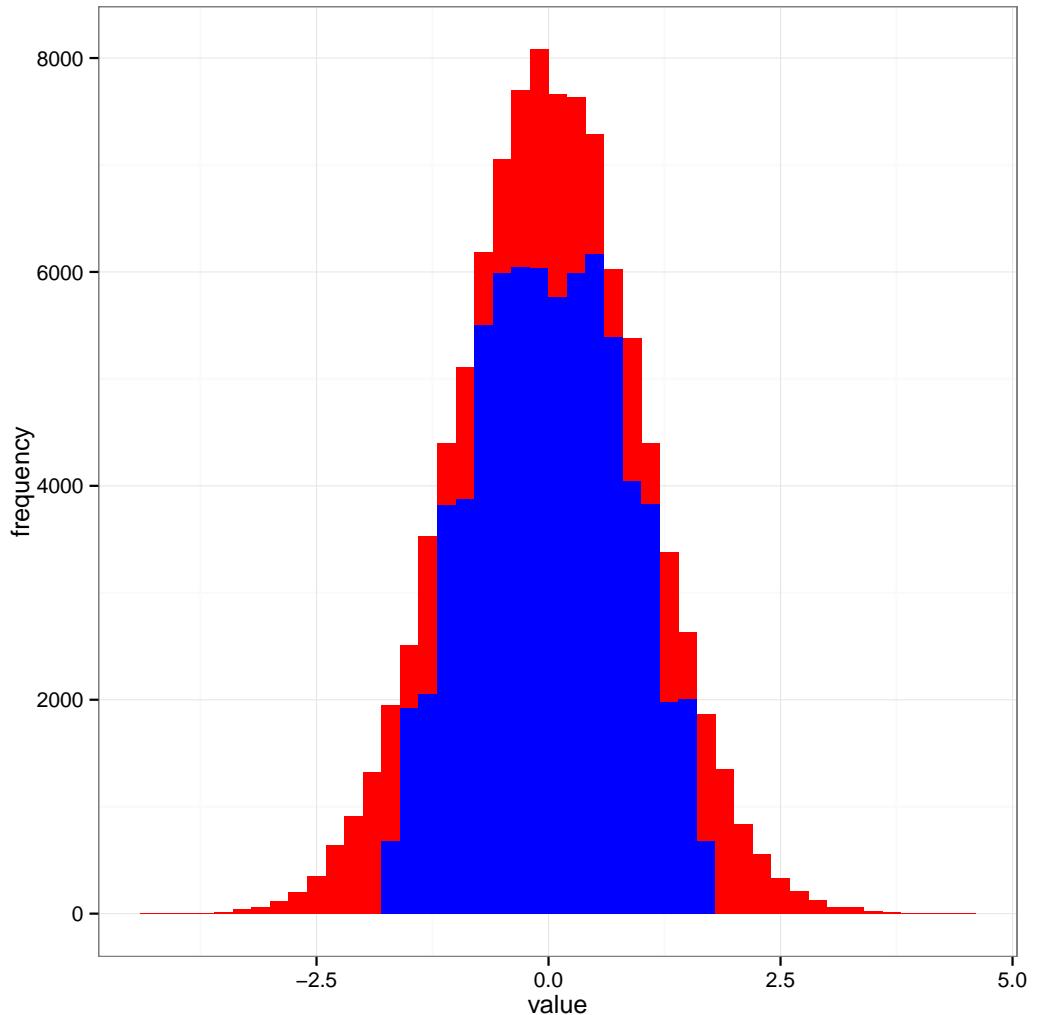


Figure 4.1: An example of residual resampling to gain a sample from a $N(0, 1)$ distribution. ■ particles which counts have been deterministic from the floor of the particle count expectation, ■ particles which have been included from the multinomial resampling step

4.3.1 Artificial evolution

One simple solution described in Liu and West (2001) based on ideas originally by Gordon et al. (1993) is to add an artificial evolution to the parameter \mathbf{z} so that \mathbf{z} is replaced by \mathbf{z}_t at time t . We than have the dynamic system:

$$\mathbf{z}_{t+1} = \mathbf{z}_t + \boldsymbol{\epsilon}_{\mathbf{z},t} \quad (4.10)$$

where $\boldsymbol{\epsilon}_{\mathbf{z},t} \sim N(0, \mathbf{W}_t)$ and \mathbf{W}_t is typically small providing a minimal perturbation to \mathbf{z}_{t+1} . This method allows the particles to move as the posterior distribution changes with each new observation of data. If the particles did not move we would have problems with particle degeneracy (Cappé et al. (2007)). The method however introduces the problem that the variance of the posterior $p(\mathbf{z}_t | \mathbf{D}_t)$ will be larger than the correct theoretical posterior $p(\mathbf{z} | \mathbf{D}_t)$ and will compound at each time step t . So there is a loss of information introducing the artificial evolution of the parameter \mathbf{z} .

In order to tackle the problem of the loss of information in artificial evolution, we first discuss the kernel smoothing methods of West (1993) in Section 4.3.2. We then discuss how Liu and West (2001) use these methods for the artificial evolution problem in Section 4.3.3.

4.3.2 Kernel smoothing

Consider a particle approximation of the posterior $p(\mathbf{z} | \mathbf{D})$, which is $\mathbf{z}^* = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(n)}\}$ with importance weights $\mathbf{w} = \{w^{(1)}, \dots, w^{(n)}\}$. A typical approximation of $p(\mathbf{z} | \mathbf{D})$ is then:

$$\hat{p}_1(\mathbf{z} | \mathbf{D}) = \sum_{i=1}^n \omega^{(i)} \mathbf{z}^{(i)}. \quad (4.11)$$

West (1993) however developed smooth kernel density approximations of the form:

$$\hat{p}_2(\mathbf{z} | \mathbf{D}) = \sum_{i=1}^n \omega^{(i)} g_i(\mathbf{z}) \quad (4.12)$$

where $g_i(\mathbf{z})$ denotes an elliptically symmetric PDF centered at $\mathbf{m}^{(i)}$ with variance $h^2 \mathbf{V}$, h is a smoothing parameter, $\mathbf{m}^{(i)}$ is the kernel location of density $g_i(\mathbf{z})$, and \mathbf{V} is an estimate of the variance of $p(\mathbf{z} | \mathbf{D})$. We solely consider the use of a Gaussian density for $g_i(\mathbf{z})$, which we will see relates to the Gaussian perturbation in Section 4.3.1, although similar results will hold for other kernels.

We follow West (1993) in using the Monte Carlo estimate of the variance:

$$\begin{aligned}\mathbf{V} &= \sum_{i=1}^n \omega^{(i)} (\mathbf{z}^{(i)} - \bar{\mathbf{z}})^2 \\ &= \sum_{i=1}^n \omega^{(i)} (\mathbf{z}^{(i)})^2 - \left(\sum_{i=1}^n \omega^{(i)} \mathbf{z}^{(i)} \right)^2\end{aligned}\quad (4.13)$$

where $\bar{\mathbf{z}} = \sum_{i=1}^n \omega^{(i)} \mathbf{z}^{(i)}$ is the Monte Carlo mean.

For $h > 0$, we show that the kernel density approximation may have an undesirable variance greater than the variance of the target posterior distribution. We first consider that the moments of the kernel density approximations in Equation (4.12) are given by:

$$\begin{aligned}\mathbb{E}_{\hat{p}_2}(\mathbf{z}^k | \mathbf{D}) &= \int_{-\infty}^{\infty} \mathbf{z}^k \sum_{i=1}^n \omega^{(i)} g_i(\mathbf{z}) d\mathbf{z} \\ &= \sum_{i=1}^n \omega^{(i)} \int_{-\infty}^{\infty} \mathbf{z}^k g_i(\mathbf{z}) d\mathbf{z} \\ &= \sum_{i=1}^n \omega^{(i)} \mathbb{E}_{g_i}(\mathbf{z}^k)\end{aligned}\quad (4.14)$$

where we use the notation $\mathbb{E}_{\hat{p}_2}(\mathbf{z}^k | \mathbf{D})$ to denote the expectation of $\mathbf{z}^k | \mathbf{D}$ with respect to the PDF \hat{p}_2 (similarly for the variance $\mathbb{V}_{\hat{p}_2}(\mathbf{z} | \mathbf{D})$). The variance of the kernel density approximation is thus:

$$\begin{aligned}\mathbb{V}_{\hat{p}_2}(\mathbf{z} | \mathbf{D}) &= \sum_{i=1}^n \omega^{(i)} \mathbb{E}_{g_i}(\mathbf{z}^2) - \left(\sum_{i=1}^n \omega^{(i)} \mathbb{E}_{g_i}(\mathbf{z}) \right)^2 \\ &= \sum_{i=1}^n \omega^{(i)} (h^2 \mathbf{V} + (\mathbf{m}^{(i)})^2) - \left(\sum_{i=1}^n \omega^{(i)} \mathbf{m}^{(i)} \right)^2 \\ &= h^2 \mathbf{V} + \sum_{i=1}^n \omega^{(i)} (\mathbf{m}^{(i)})^2 - \left(\sum_{i=1}^n \omega^{(i)} \mathbf{m}^{(i)} \right)^2.\end{aligned}\quad (4.15)$$

Using a natural first choice of kernel location $\mathbf{m}^{(i)} = \mathbf{z}^{(i)}$, we immediately see from Equations (4.15) and (4.13) that $\mathbb{V}_{\hat{p}_2}(\mathbf{z} | \mathbf{D}) = h^2 \mathbf{V} + \mathbf{V}$, larger than the target variance of \mathbf{V} . West (1993) describes one method of alleviating the problem of over estimating the target variance by specifying:

$$h = \frac{c}{n^{\frac{1}{1+4d}}}\quad (4.16)$$

where:

$$c = \left(\frac{4}{1 + 2d} \right)^{\frac{1}{1+4d}} \quad (4.17)$$

and d is the dimension of \mathbf{z} (the number of parameters). This enables $h \rightarrow 0$ as $n \rightarrow \infty$ and so $\hat{p}_2(\mathbf{z}|\mathbf{D})$ approaches $p(\mathbf{z}|\mathbf{D})$ (and $\hat{p}_1(\mathbf{z}|\mathbf{D})$) as n increases.

However for fixed n , $\hat{p}_2(\mathbf{z}|\mathbf{D})$ is always over-dispersed relative to $p(\mathbf{z}|\mathbf{D})$. To correct the over-dispersion, West (1993) suggested shrinking Gaussian kernel locations $\mathbf{m}^{(i)}$ from $\mathbf{z}^{(i)}$ closer to the weighted mean of \mathbf{z}^* , $\bar{\mathbf{z}}$. Using parameter $\alpha \in [0, 1]$ we define the kernel locations as:

$$\mathbf{m}^{(i)} = \alpha \mathbf{z}^{(i)} + (1 - \alpha) \bar{\mathbf{z}}. \quad (4.18)$$

Following from Equation (4.15) the variance of the kernel density approximation with the new kernel locations is then:

$$\mathbb{V}_{\hat{p}_2}(\mathbf{z}|\mathbf{D}) = h^2 \mathbf{V} + \alpha^2 \mathbf{V} \quad (4.19)$$

and so the kernel density approximation will have the desired variance \mathbf{V} when $\alpha = \sqrt{1 - h^2}$.

Figure 4.2 shows an example of the effect of the two different kernel locations (with shrinkage: $\mathbf{m}^{(i)} = \alpha \mathbf{z}^{(i)} + (1 - \alpha) \bar{\mathbf{z}}$, and without: $\mathbf{m}^{(i)} = \mathbf{z}^{(i)}$) when $\hat{p}_2(\mathbf{z}|\mathbf{D})$ is used to approximate a $N(0, 1)$ distribution via an equally weighted random sample of size 5,000. When h is low, there is little effect of adding shrinkage to the kernel locations, and the resulting kernel density estimate is very un-smooth. However, when h becomes larger, the kernel density estimate with shrunk kernel locations more accurately approximates the theoretical distribution. The density estimate in red, which simply takes the kernel locations as the particle locations, has clear over-dispersion.

4.3.3 Kernel smoothing methods for variance reduction in artificial evolution

The problem with a mixture of Gaussian distributions with kernel locations at each particle ($\mathbf{m}^{(i)} = \mathbf{z}^{(i)}$) is very similar in nature to each particle being given an artificial evolution via Gaussian noise. In the case of the mixture of Gaussian distributions the particle sample has Monte Carlo variance \mathbf{V} , but the smooth approximation has variance $\mathbf{V} + h^2 \mathbf{V}$. In the case of artificial evolution the particle sample has Monte Carlo variance \mathbf{V}_t at time t , but after artificial evolution the particles have variance $\mathbf{V}_t + \mathbf{W}_t$ (\mathbf{W}_t is the variance of the artificial perturbation $\boldsymbol{\epsilon}_{\mathbf{z},t}$). Both methods lead

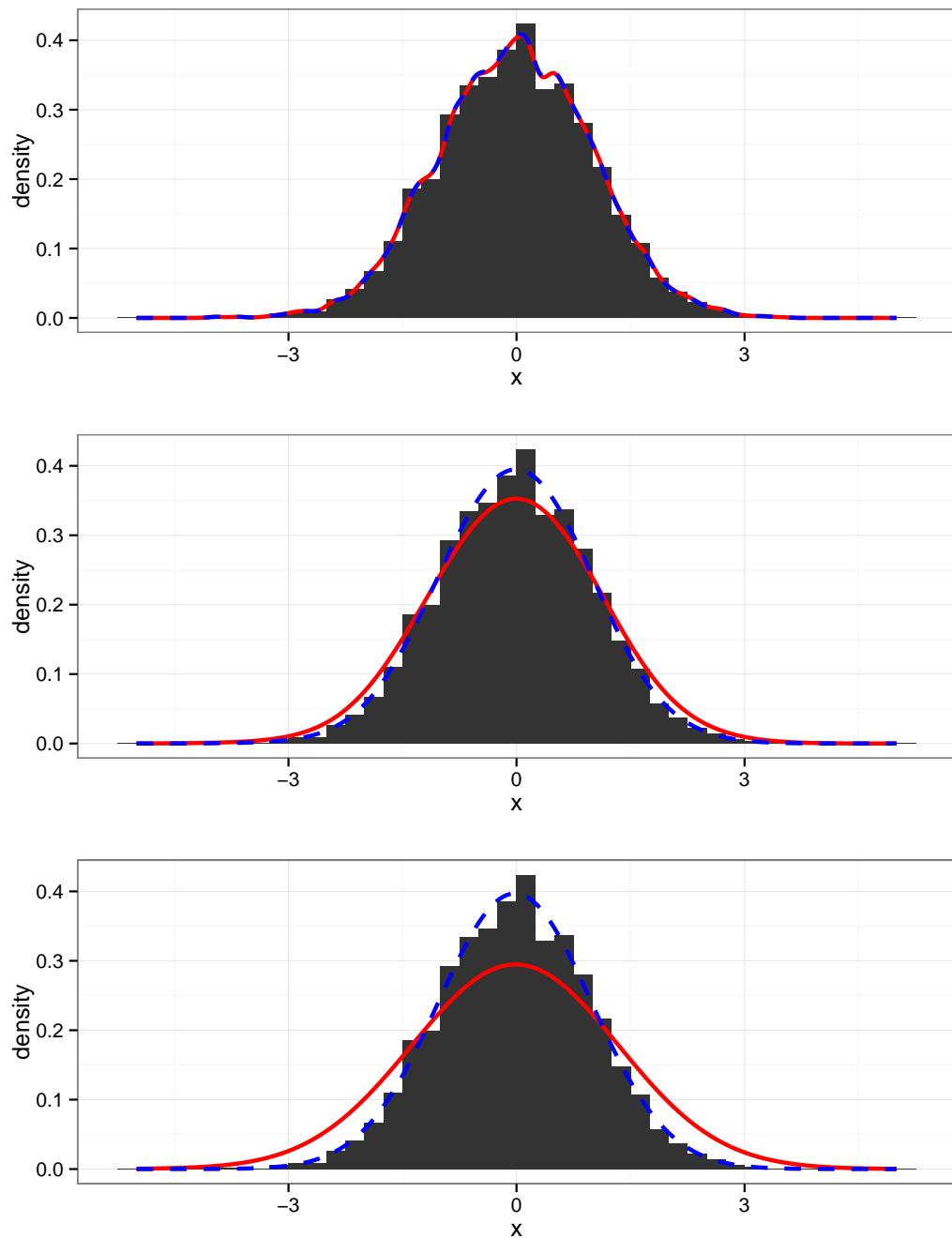


Figure 4.2: Smooth kernel density estimates when $h = 0.1$ (top), $h = 0.5$ (middle), and $h = 0.9$ (bottom). — the kernel density estimate with no shrinkage, - - - the kernel density estimate with shrinkage

to over-dispersion of the target distribution.

The solution to over-dispersion in the kernel smoothing case was to shrink the kernel locations towards the sample mean, and a very similar solution is possible in the case of artificial evolution. The artificial evolution in Equation (4.10) implies:

$$\begin{aligned}\mathbb{V}(\mathbf{z}_{t+1}|\mathbf{D}_t) &= \mathbb{V}(\mathbf{z}_t + \boldsymbol{\epsilon}_{\mathbf{z},t}|\mathbf{D}_t) \\ &= \mathbb{V}(\mathbf{z}_t|\mathbf{D}_t) + \mathbf{W}_t + \mathbb{C}(\mathbf{z}_t, \boldsymbol{\epsilon}_{\mathbf{z},t}|\mathbf{D}_t) + \mathbb{C}(\boldsymbol{\epsilon}_{\mathbf{z},t}, \mathbf{z}_t, |\mathbf{D}_t).\end{aligned}\quad (4.20)$$

Previously $\mathbb{C}(\mathbf{z}_t, \boldsymbol{\epsilon}_{\mathbf{z},t}|\mathbf{D}_t) = \mathbb{C}(\boldsymbol{\epsilon}_{\mathbf{z},t}, \mathbf{z}_t|\mathbf{D}_t) = \mathbf{0}$ since there was independence between the current state \mathbf{z}_t and the artificial perturbation $\boldsymbol{\epsilon}_{\mathbf{z},t}$. We allow these covariance terms to be non-zero to enable the artificial perturbation to not increase the variance of the particle approximation ($\mathbb{V}(\mathbf{z}_{t+1}|\mathbf{D}_t) = \mathbb{V}(\mathbf{z}_t|\mathbf{D}_t)$). We thus set:

$$\begin{aligned}\mathbb{C}(\mathbf{z}_t, \boldsymbol{\epsilon}_{\mathbf{z},t}|\mathbf{D}_t) &= \mathbb{C}(\boldsymbol{\epsilon}_{\mathbf{z},t}, \mathbf{z}_t|\mathbf{D}_t) \\ &= -\frac{1}{2}\mathbf{W}_t\end{aligned}\quad (4.21)$$

for symmetric variance matrix \mathbf{W}_t (for random vectors \mathbf{X} and \mathbf{Y} , $\mathbb{C}(\mathbf{X}, \mathbf{Y}) = \mathbb{C}(\mathbf{Y}, \mathbf{X})'$ so the covariance matrices may only be equal when they are symmetric). This implies:

$$\begin{aligned}\mathbb{C}(\mathbf{z}_{t+1}, \mathbf{z}_t|\mathbf{D}_t) &= \mathbb{E}(\mathbf{z}_{t+1}\mathbf{z}_t|\mathbf{D}_t) - \mathbb{E}(\mathbf{z}_{t+1}|\mathbf{D}_t)\mathbb{E}(\mathbf{z}_t|\mathbf{D}_t) \\ &= \mathbb{E}((\mathbf{z}_t + \boldsymbol{\epsilon}_{\mathbf{z},t})\mathbf{z}_t|\mathbf{D}_t) - \bar{\mathbf{z}}_t^2 \\ &= \mathbb{E}(\mathbf{z}_t^2 + \boldsymbol{\epsilon}_{\mathbf{z},t}\mathbf{z}_t|\mathbf{D}_t) - \bar{\mathbf{z}}_t^2 \\ &= \mathbb{E}(\mathbf{z}_t^2|\mathbf{D}_t) + \mathbb{E}(\boldsymbol{\epsilon}_{\mathbf{z},t}\mathbf{z}_t|\mathbf{D}_t) - \bar{\mathbf{z}}_t^2 \\ &= (\mathbf{V}_t + \bar{\mathbf{z}}_t^2) + (-\frac{1}{2}\mathbf{W}_t + 0) - \bar{\mathbf{z}}_t^2 \\ &= \mathbf{V}_t - \frac{1}{2}\mathbf{W}_t.\end{aligned}\quad (4.22)$$

It follows similarly that $\mathbb{C}(\mathbf{z}_t, \mathbf{z}_{t+1}|\mathbf{D}_t) = \mathbf{V}_t - \frac{1}{2}\mathbf{W}_t$ which is automatic only in the case of scalar \mathbf{z}_{t+1} and \mathbf{z}_t .

Under the assumption of approximate joint normality of $\boldsymbol{\theta}_t, \boldsymbol{\epsilon}_{\mathbf{z},t}|\mathbf{D}_t$, we may use standard results regarding the conditional distribution of a multivariate normal distribution. An outline of the results are as follows:

For a multivariate normal vector $\mathbf{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, consider partitioning $\boldsymbol{\mu}$ and \mathbf{Y} into:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix} \quad (4.23)$$

with a similar partition of Σ into:

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \quad (4.24)$$

The conditional distribution of the first partition given the second is then:

$$\mathbf{y}_1 | \mathbf{y}_2 \sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}). \quad (4.25)$$

In our problem:

$$\begin{aligned} \boldsymbol{\mu}_1 &= \boldsymbol{\mu}_2 = \bar{\mathbf{z}}_t \\ \Sigma_{11} &= \Sigma_{22} = \mathbf{V}_t \\ \Sigma_{12} &= \Sigma_{21} = \mathbf{V}_t - \frac{1}{2}\mathbf{W}_t \end{aligned}$$

and so the conditional normal evolution, as stated by Liu and West (2001), is:

$$\begin{aligned} \mathbf{z}_{t+1} | \mathbf{z}_t &\sim N(\bar{\mathbf{z}}_t + (\mathbf{V}_t - \frac{1}{2}\mathbf{W}_t)\mathbf{V}_t^{-1}(\mathbf{z}_t - \bar{\mathbf{z}}_t), \mathbf{V}_t - (\mathbf{V}_t - \frac{1}{2}\mathbf{W}_t)\mathbf{V}_t^{-1}(\mathbf{V}_t - \frac{1}{2}\mathbf{W}_t)) \\ &= N(\mathbf{A}_t \mathbf{z}_t + (\mathbf{I} - \mathbf{A}_t) \bar{\mathbf{z}}_t, (\mathbf{I} - \mathbf{A}_t^2) \mathbf{V}_t) \end{aligned} \quad (4.26)$$

where:

$$\mathbf{A}_t = \mathbf{I} - \frac{1}{2}\mathbf{W}_t\mathbf{V}_t^{-1}. \quad (4.27)$$

Liu and West (2001) restrict to the special case where the artificial evolution variance matrix is specified by a standard discount factor, and use the symmetric matrix:

$$\mathbf{W}_t = \mathbf{V}_t(\frac{1}{\delta} - 1) \quad (4.28)$$

the conditional distribution in Equation (4.26) then simplifies further to:

$$\mathbf{z}_{t+1} | \mathbf{z}_t \sim N(\alpha \mathbf{z}_t + (1 - \alpha) \bar{\mathbf{z}}, h^2 \mathbf{V}_t) \quad (4.29)$$

where $\delta \in [1/3, 1]$ and is typically 0.95 - 0.99 (Liu and West (2001) suggest around 0.99), $\alpha = \frac{3\delta-1}{2\delta}$, and $h = \sqrt{1 - \alpha^2}$. That is, a very similar form to as was seen to correct over-dispersion for the kernel smoothing in Section 4.3.2.

Using the transition density $p(\mathbf{z}_{t+1} | \mathbf{z}_t)$ implied by Equation (4.29) means that if $p(\mathbf{z}_t | \mathbf{D}_t)$ has finite mean $\bar{\mathbf{z}}_t$ and variance \mathbf{V}_t then $p(\mathbf{z}_{t+1} | \mathbf{D}_t)$ will also have finite mean $\bar{\mathbf{z}}_t$ and variance \mathbf{V}_t . We therefore have a solution to the problem of increasing variance and loss of information when using artificial evolution.

4.3.4 An example of variance reduction in artificial evolution

Here we present a simple example where we wish to perform inference on a single static parameter μ using the methods described in Section 4.3.3. We observe data points from a $N(\mu, 2^2)$ distribution, the true value of μ being 1. We start with the (poor choice of) prior $\mu \sim N(5, 1^2)$ and at each time point t , we observe the data vector of length two, $\mathbf{y}_t = (y_t^{(1)}, y_t^{(2)})$. The theoretical posterior distribution of $\mu | \mathbf{D}_t$ is available:

$$\mu | \mathbf{D}_t \sim N\left(\frac{20 + \sum_{i=1}^t y_i^{(1)} + y_i^{(2)}}{4 + 2t}, \left(1 + \frac{t}{2}\right)^{-1}\right). \quad (4.30)$$

We use the discount factor $\delta = 0.99$ as suggested by Liu and West (2001), and 50,000 particles updated via the Auxiliary Particle Filter. Despite the large sample, the program run times were very quick. Results can be seen in Figure 4.3, which conveys how accurately the sampled particles follow the theoretical posterior distribution.

4.4 An updated model

We present a modification to the non-homogeneous Poisson process model of Chapter 3 Section 3.2.3, which allows the log of a team's resource to follow a random walk. We therefore consider the model:

$$\log(\lambda_i(t, w)) = h + \alpha_i(t)e^{LR_{i,w}} - (1 - \alpha_j(t))e^{LR_{j,w}} + \rho(t) \quad (4.31)$$

$$\log(\mu_j(t, w)) = a + \alpha_j(t)e^{LR_{j,w}} - (1 - \alpha_i(t))e^{LR_{i,w}} + \rho(t) \quad (4.32)$$

where $LR_{k,w}$ represents the log of team k 's resource in week w of the season (the log-resource) and $\lambda_k(t, w)$ is team k 's instantaneous rate of scoring at time t of the match in week w , similarly for the away team regarding $\mu_k(t, w)$. The random walk process for team k is:

$$LR_{k,w+1} = LR_{k,w} + \epsilon_k \quad (4.33)$$

where $\epsilon_k \sim N(0, \sigma^2)$ for all $k = 1, \dots, 20$, so the log-resource for each team follows an independent random walk. Inference is now concerned with the parameter vector $\boldsymbol{\theta}_w = (\mathbf{z}, \mathbf{LR}_w)$ where $\mathbf{z} = (h, a, c_{-1}, c_0, c_1, d_{-1}, d_0, d_1, \rho_1, \rho_2)$ is the vector of 10 static parameters and $\mathbf{LR}_w = (LR_{1,w}, \dots, LR_{20,w})$ is the vector of 20 dynamic log-resource parameters.

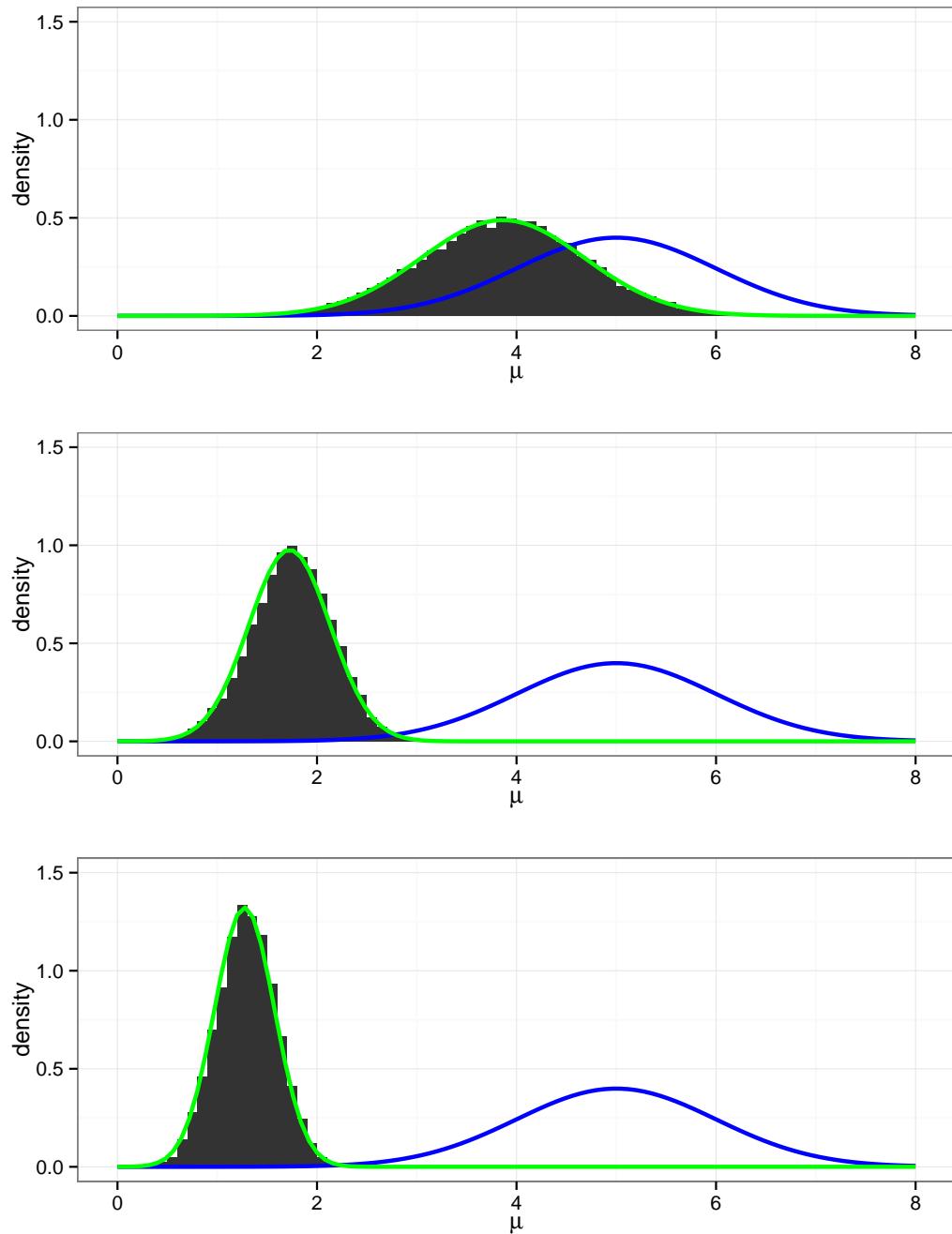


Figure 4.3: Example of the Auxiliary Particle Filter at time 1 (top), 10 (middle), 20 (bottom). The histogram represents the particle posterior approximation of the static parameter μ . — the prior density, — the theoretical posterior

4.4.1 Updated model inference

We now follow through a full example of how the particle filter updates the posterior belief of the parameter vector $\boldsymbol{\theta}_w$ for the season 2011/2012 in the EPL. We employ the method of the auxiliary particle filter as discussed in Section 4.2.3, and also show how an optimal value of σ^2 (the variance of the random walk) can be found. Note the subtle difference in notation used to denote the week w and the weight ω .

We start the inference process by obtaining a particle approximation of $p(\boldsymbol{\theta}_1|\mathbf{D}_0)$, that is, the prior distribution of the week 1 parameters before we have observed any data. We begin by using a random-walk MH algorithm, assuming static resource parameters (as was in Chapter 3 Section 3.3) using all data from the previous season, 2010/2011. We used the following prior distributions:

$$\begin{aligned} h &\sim N(0.4, 0.5^2) & c_i &\sim \beta(1.5, 1.5) \text{ for } i \in \{-1, 0, 1\} \\ a &\sim N(0.08, 0.5^2) & d_i &\sim \beta(3, 1) \text{ for } i \in \{-1, 0, 1\} \\ \rho_1 &\sim N(1.098, 0.5^2) & LR_k &\sim N(-0.7, 1^2) \text{ for } k \in \{1, \dots, 20\} \\ \rho_2 &\sim N(1.504, 0.5^2) \end{aligned} \quad (4.34)$$

using the same reasoning as in Chapter 3 Section 3.3. The prior on the log-resource parameters, LR_k , is quite vague, in that a 95% prior credible interval for a team's resource is (0.0700 3.5253).

The resulting posterior distribution was then used as our particle approximation of $p(\boldsymbol{\theta}_0|\mathbf{D}_0)$, and we obtain an approximation of $p(\boldsymbol{\theta}_1|\mathbf{D}_0)$ by passing the particles through the system transition model. This means that the prior for the log-resource parameters at week $w = 1$, $p(\mathbf{LR}_1|\mathbf{D}_0)$, will have higher variance than $p(\mathbf{LR}_0|\mathbf{D}_0)$ - in order to take account of uncertainty regarding how the team's abilities change between seasons. For the static parameters, the variance of $p(\mathbf{z}_1|\mathbf{D}_0)$ will be equal to that of $p(\mathbf{z}_0|\mathbf{D}_0)$ since the system model for these parameters is static through each week.

The three teams who were promoted into the 2011/2012 EPL, (Queens Park Rangers, Norwich City, and Swansea City) were given the particle representation for the log-resource from the three teams whom they replaced (Birmingham City, Blackpool, and West Ham United).

We observe data for week w sequentially for $w = 1, \dots, 38$. The posterior distribution of $\boldsymbol{\theta}_w$ after observing data from week w is $p(\boldsymbol{\theta}_w|\mathbf{D}_w)$ and is approximated by a set of n equally weighted particles $\{\boldsymbol{\theta}_w^{(1)}, \dots, \boldsymbol{\theta}_w^{(n)}\}$.

Upon the arrival of data from week $w+1$, \mathbf{y}_{w+1} , we firstly identify the current kernel

location vector of the 10 static parameters:

$$\mathbf{m}_w^{(i)} = \alpha \mathbf{z}_w^{(i)} + (1 - \alpha) \bar{\mathbf{z}}_w. \quad (4.35)$$

In order to perform the auxiliary resampling step, we calculate first-stage weights for each particle i as:

$$\bar{\omega}_1^{(i)} = p(\mathbf{y}_{w+1} | \mathbf{L}\mathbf{R}_w^{(i)}, \mathbf{m}_w^{(i)}) \quad (4.36)$$

since $\mathbf{L}\mathbf{R}_w^{(i)}$ is the mean of the random log-resource parameter vector when evolving to $\mathbf{L}\mathbf{R}_{w+1}^{(i)}$ for particle i . We then sample with replacement the auxiliary indicators j with probabilities proportional to $\bar{\omega}_1^{(i)}$.

Now the particles must be propagated based on the auxiliary indicators j , the static parameters are propagated via the method of artificial evolution corrected for information loss (as described in Section 4.3.3):

$$\mathbf{z}_{w+1}^{(j)} \sim N(\mathbf{m}_w^{(j)}, h^2 \mathbf{V}_w) \quad (4.37)$$

where \mathbf{V}_w is the covariance matrix of the 10 static parameters, estimated from the sample covariance matrix of the particle approximation of $p(\mathbf{z} | \mathbf{D}_w)$. We use the discount factor $\delta = 0.99$ as suggested by Liu and West (2001), which implies $h^2 = 0.010$. The dynamic parameters are more simply propagated as per the system transition model:

$$LR_{k,w+1}^{(j)} = LR_{k,w}^{(j)} + \epsilon_k \quad \text{for } k = 1, \dots, 20. \quad (4.38)$$

Finally, we calculate second-stage weights for particle j proportional to:

$$\bar{\omega}_2^{(j)} = \frac{p(\mathbf{y}_{w+1} | \mathbf{L}\mathbf{R}_{w+1}^{(j)}, \mathbf{z}_{w+1}^{(j)})}{p(\mathbf{y}_{w+1} | \mathbf{L}\mathbf{R}_w^{(j)}, \mathbf{m}_w^{(j)})}. \quad (4.39)$$

A resampling step of the particles j based on these second-stage weights provides a set of equally weighted particles $\{\boldsymbol{\theta}_{w+1}^{(1)}, \dots, \boldsymbol{\theta}_{w+1}^{(n)}\}$ which provide an approximation of the posterior $p(\boldsymbol{\theta}_{w+1} | \mathbf{D}_{w+1})$ where $\boldsymbol{\theta}_{w+1}^{(k)} = (\mathbf{z}_{w+1}^{(k)}, \mathbf{L}\mathbf{R}_{w+1}^{(k)})$.

4.4.2 Finding the optimal variance

We now have sufficient tools to give a particle approximation of $p(\boldsymbol{\theta}_w | \mathbf{D}_w)$. We use these particles to simulate match results, providing one week ahead forecasts. We can do this for different values of σ^2 , and in a similar fashion to determining the optimal value of the parameters γ_1 and γ_2 in Chapter 3 Section 3.3.4, record the value of a performance metric. We follow Owen (2011) who used the geometric

mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed:

$$GM_{1,38} = \exp\left(\frac{1}{380} \sum_{w=1}^{38} \sum_{m \in M_w} \log(\hat{\mathbb{P}}(O_m | \mathbf{D}_{w-1}))\right) \quad (4.40)$$

where again, M_w is the set of 10 matches in week w , O_m is the observed outcome of match m (either a home team win, draw, or away team win), and \mathbf{D}_{w-1} is the observed data up to but not including week w . The interpretation of the metric $GM_{1,38}$ is then perhaps simpler than the metric $\sum_{w=1}^{38} LSR_w$ previously considered in Chapter 3 Section 3.3.4, although the two methods will be totally in agreement with regards to determining the optimal σ^2 .

Using 100,000 particles, a plot of various different values for σ^2 and the corresponding $GM_{1,38}$ can be seen in Figure 4.4. Again, we follow Owen (2011) and show σ^2 in the range of 0 to 0.02, there is however no concrete evidence that there is an optimal $\sigma^2 > 0$, as highlighted by the smooth Local Regression (LOESS) estimate (see Cleveland (1979)) which is typically decreasing in σ^2 . An optimal value of $\sigma^2 > 0$ would have suggested that there is benefit in allowing the team log-resource parameters to follow a dynamic system, but somewhat unfortunately, we cannot deduce this from our results.

This is in contrast to the findings of Owen (2011), who found an optimal value of σ^2 near 0.004 based on the metric $GM_{1,38}$. This is likely due to one (or both) of the following reasons:

1. The different data used, in our case the 2011/2012 season from the EPL, and in the case of Owen (2011), seasons 2003/2004, 2004/2005, and 2005/2006 from the Scottish Premier League (SPL)
2. The different model specifications used, in our case a dynamic system for a single ‘resource’ parameter for each team, and in the case of Owen (2011), a dynamic system for an ‘attack’ and ‘defence’ parameter for each team

That is, it may be that the teams in the SPL vary in ability more throughout a season(s), or that attack and defense strengths may vary throughout a season, but overall team abilities/resources do not.

We continue analysis of the particle filtering methods and the dynamic team log-resource model using $\sigma^2 = 0.0001$, a very small value which should still allow adequate propagation of the particles, not unlike the artificial evolution methods described in section 4.3.1.

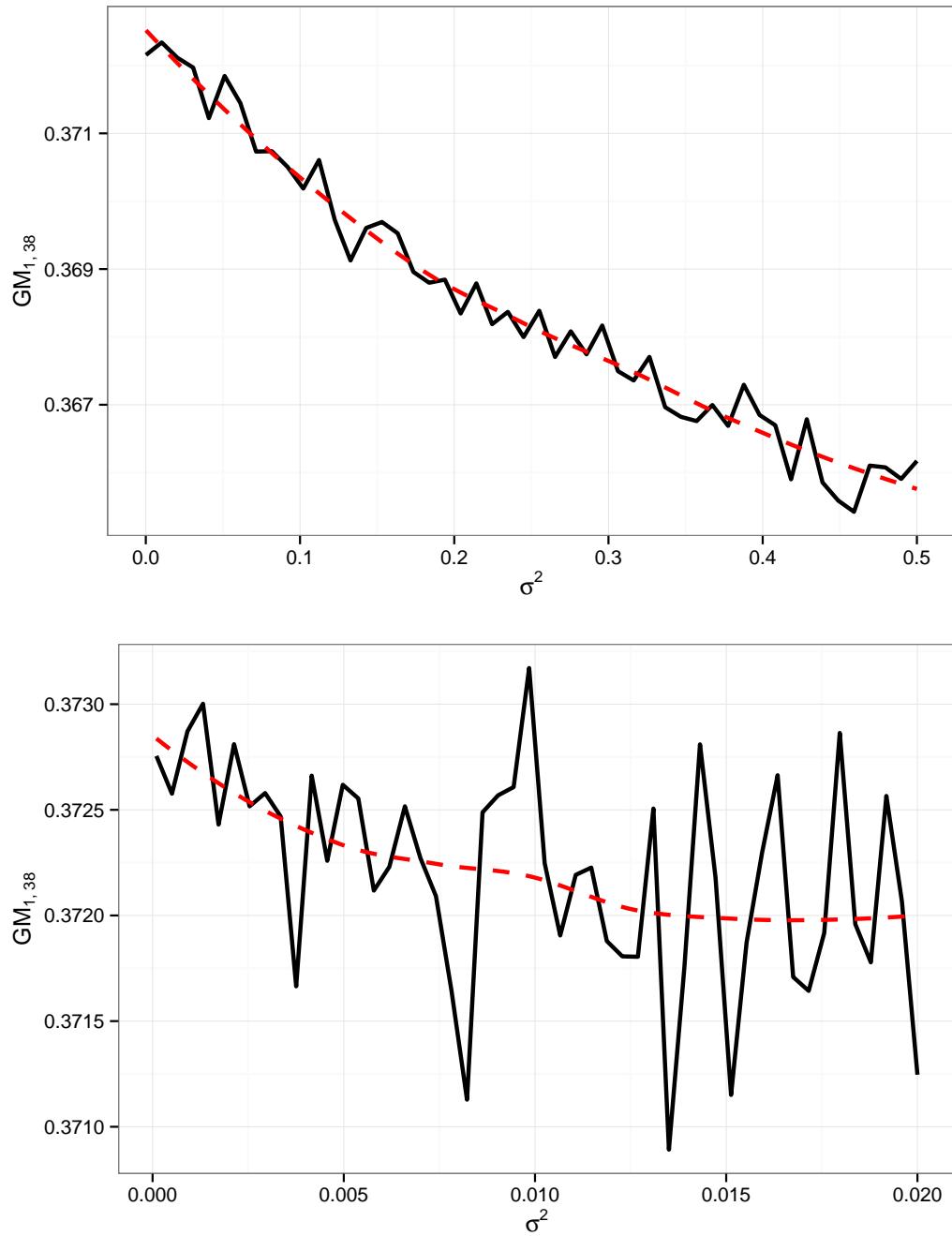


Figure 4.4: Values of the metric $GM_{1,38}$ for different values of σ^2 in the range 0 to 0.5 (top) and 0 to 0.02 (bottom). — the calculated values, - - - a smooth LOESS estimate

4.4.3 Inference results

Here we show time series plots for the 0.025 and 0.975 posterior quantiles, along with the posterior mean of the particle approximation of $p(\boldsymbol{\theta}_w | \mathcal{D}_w)$ using 100,000 particles, for each parameter in $\boldsymbol{\theta}_w$. The plots are shown in Figures 4.5 to 4.9.

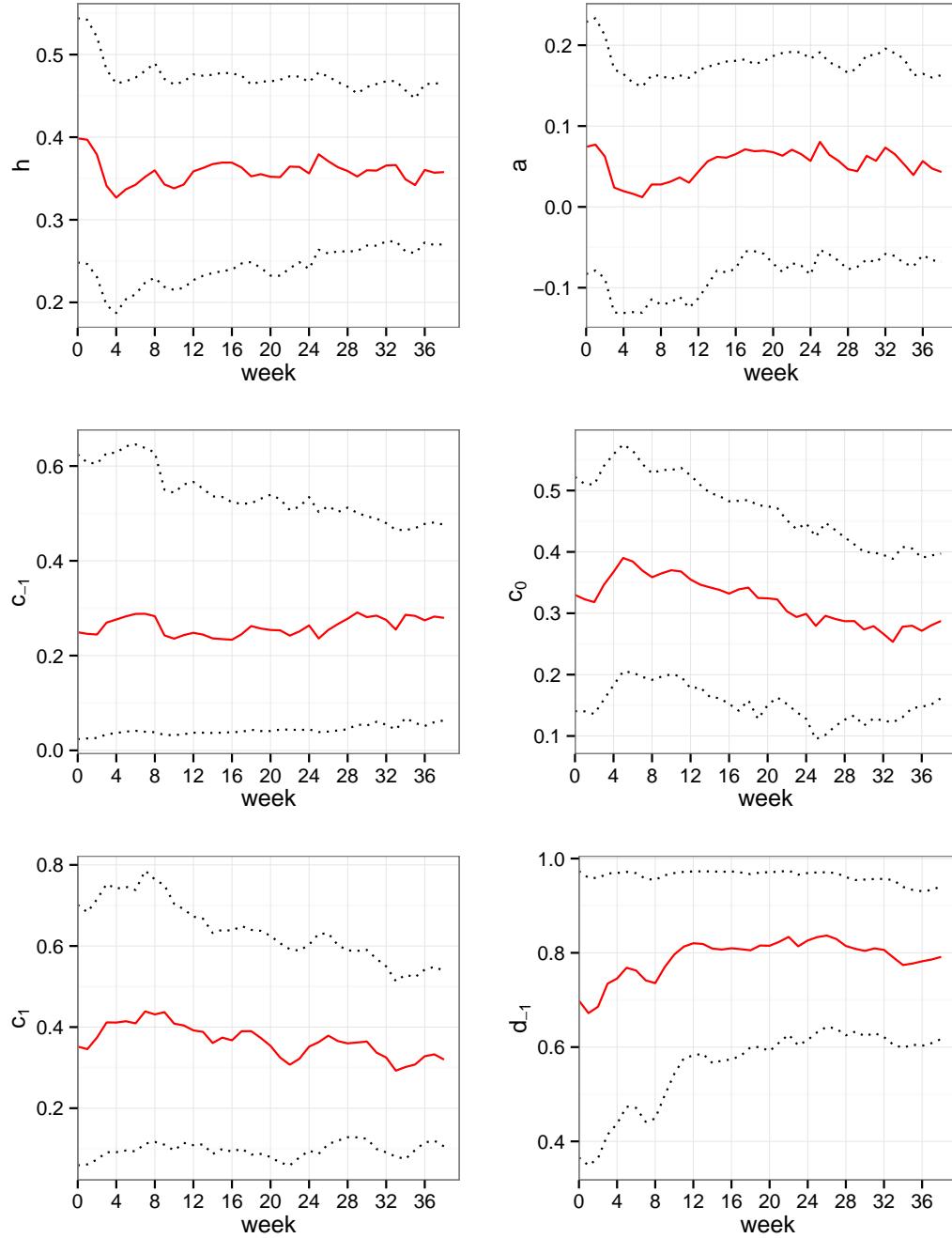


Figure 4.5: Time series plot of parameters h , a , c_{-1} , c_0 , c_1 , and d_{-1} . — posterior mean, - - - 95% BCI

There are some notable fluctuations in posterior probability of the team log-resource parameters which correspond to large match scores. For example week $w = 9$ featured a score of Manchester United 1 - 6 Manchester City and there is clear

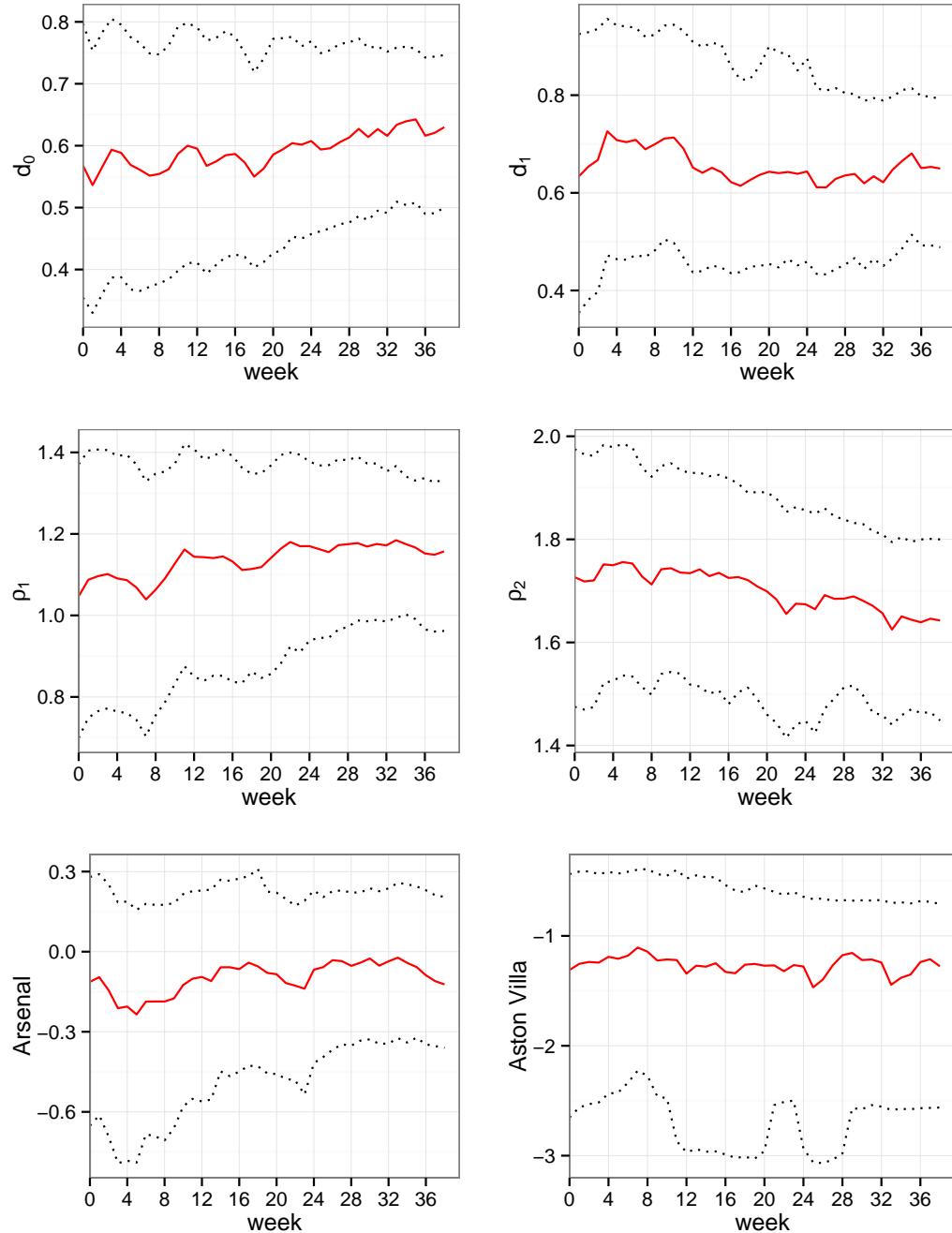


Figure 4.6: Time series plot of parameters d_1 , d_0 , ρ_1 , ρ_2 and the log-resource team parameters. — posterior mean, - - - 95% BCI

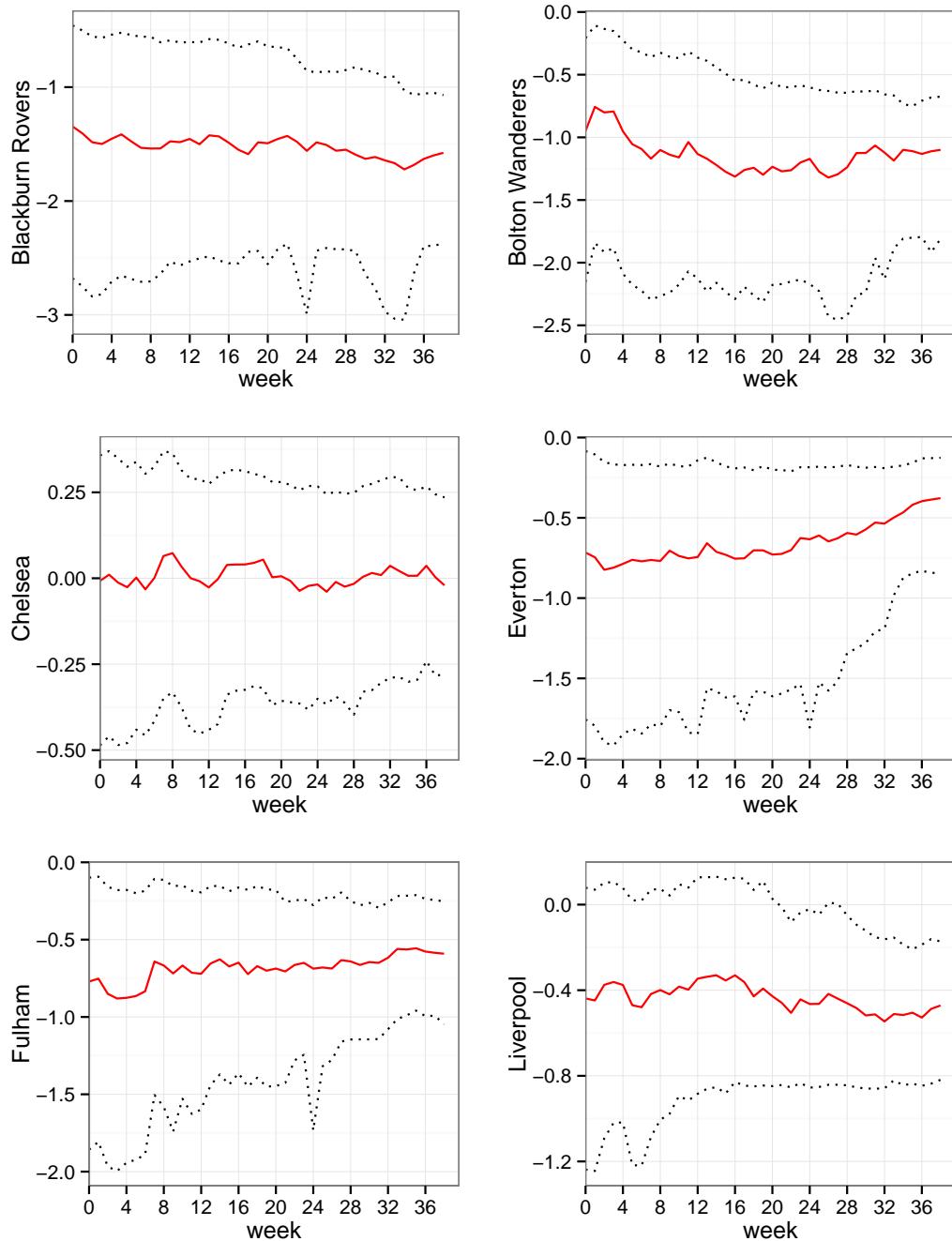


Figure 4.7: Time series plot of the team log-resource parameters. — posterior mean, - - - 95% BCI

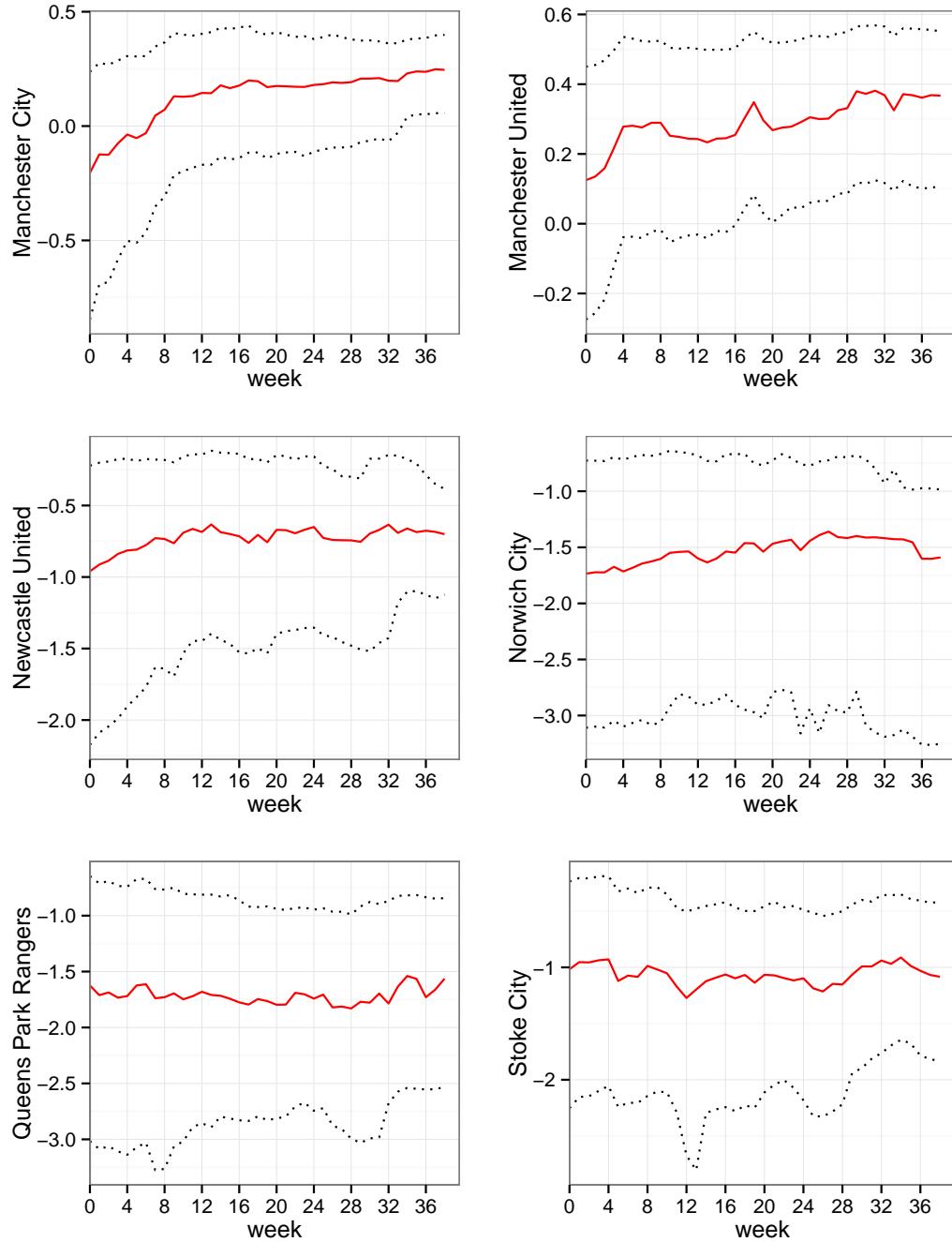


Figure 4.8: Time series plot of the team log-resource parameters. — posterior mean, - - - 95% BCI

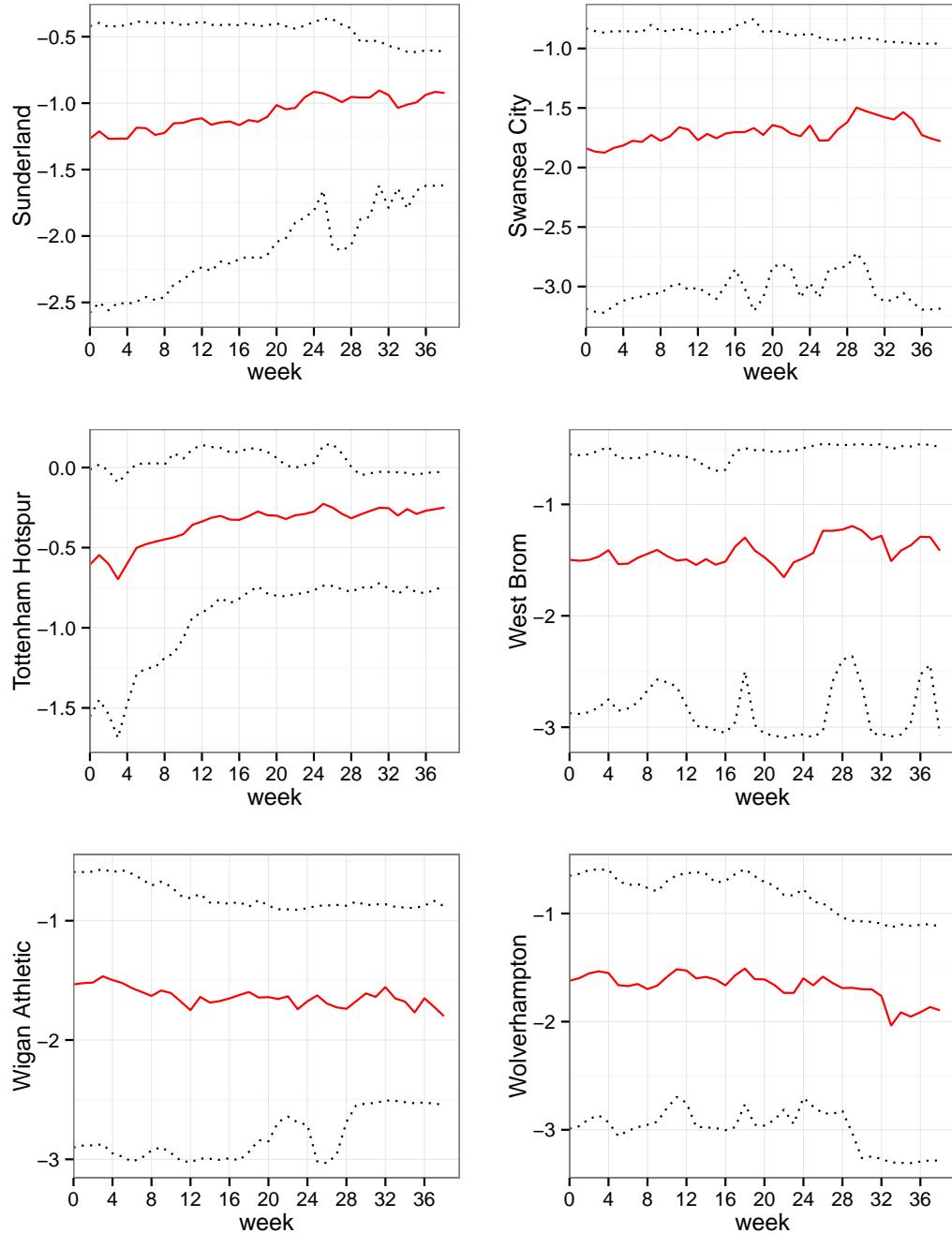


Figure 4.9: Time series plot of the team log-resource parameters. — posterior mean, - - - 95% BCI

model	$\sum_{w=6}^{38} LSR_w$	$GM_{6,38}$	O_P
M	-324.26	0.3743	60.33
DM	-324.23	0.3744	48.18

Table 4.1: A comparison of the sum of the logarithmic scoring rule, the geometric mean of the one-week ahead predicted probabilities for the match outcomes that were actually observed, and the betting profit, for models M and DM for weeks 6 to 38

spike in Manchester City’s log-resource at this time, conversely, the log-resource of Manchester City appears to be decreasing at this time. Other notable results include: Manchester United 5 - 0 Wigan Athletic in week 18, Newcastle United 3 - 0 Manchester United in week 20, and West Brom 1 - 0 Chelsea in week 27. We also comment on the parameter estimates with comparison to parameter estimates from simulated data in Section 4.4.6.

4.4.4 Model and inference performance

Here we make comparisons regarding the one-week-ahead predictive performance of the dynamic log-resource model (which we will denote DM), with our proposed model (denoted M) in Chapter 3 Section 3.2.3. In order for model DM to be comparable to the models shown in Chapter 3 Section 3.4, we again consider matches from weeks 6 to 38. These weeks were previously considered since after 5 weeks of matches, all the teams were comparable in that they had all played a common team. The results of the comparison can be seen in Table 4.1. The results are almost identical, as one would expect - the models and inference methods are in many ways the same. With regards to the models, for model DM we used a very small value of σ^2 which effectively reduces the dynamic model to the static model of M. With regards to the inference methods, we used posterior samples from the previous season’s data (the 2010/2011 season) to give the week $w = 1$ prior distribution approximation for the model DM parameters. Likewise, for the parameters in model M, we used the previous season data for prior elicitation, albeit in a much simpler way.

In terms of betting against the UK bookmaker Bet365 on matches in weeks 6 to 38, the models did not perform so similarly, which may be at first somewhat unintuitive given the previous comments. The betting profits are shown in Table 4.1, under the notation O_P . We thus choose to examine the differences in one-week-ahead predicted probabilities between the two models, a plot of which can be seen in Figure 4.10. The values in the plot do largely lie on the line $y = x$, but there are some subtle differences which lead to the disparity in the betting profits - which can be influenced by only a single bet being placed or not placed. The smooth LOESS estimate helps

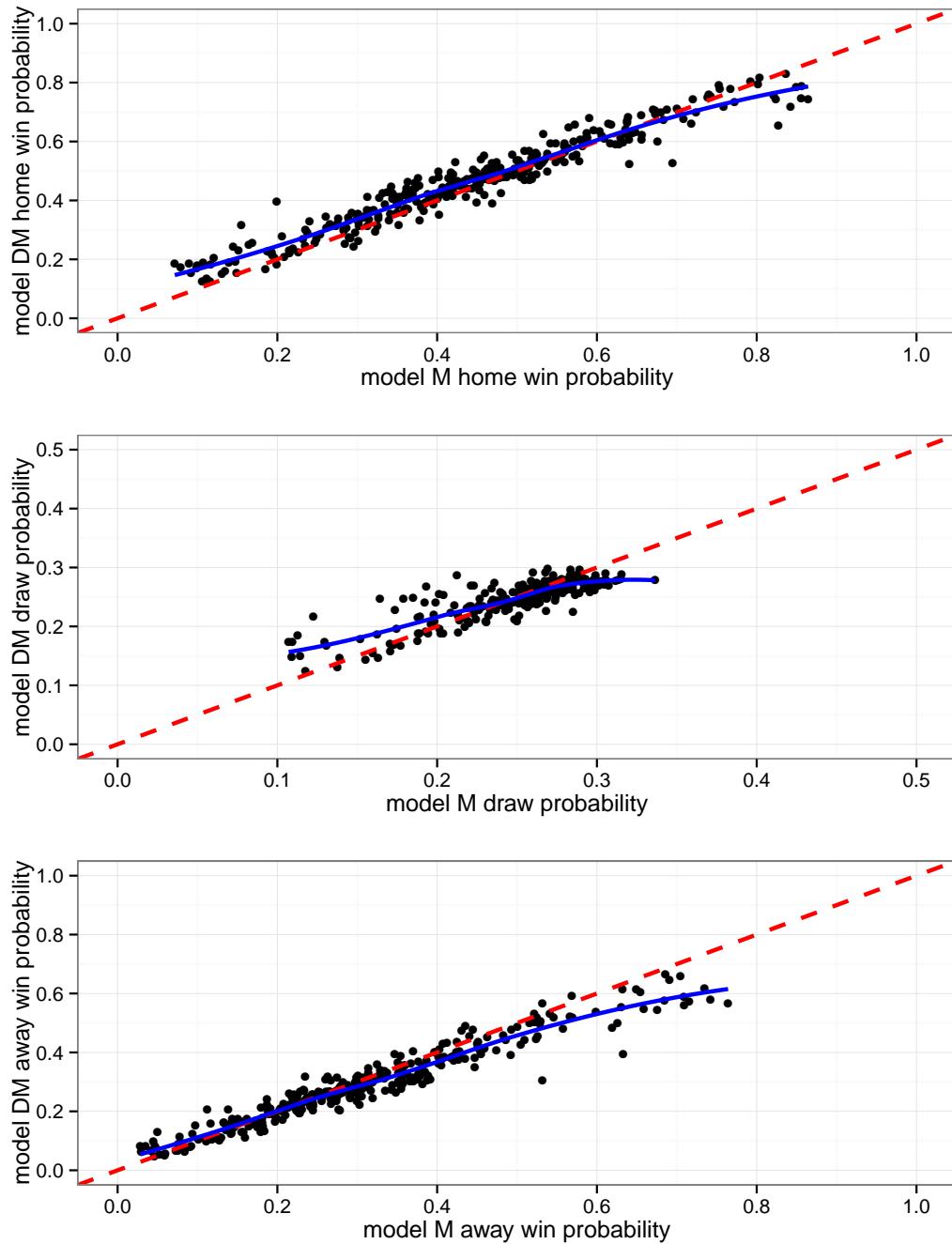


Figure 4.10: Scatter plot of the home win (top), draw (middle), and away win (bottom) probabilities predicted by the models M and DM for 330 matches from weeks 6 to 380. --- the line $y = x$, — a smooth LOESS estimate

to display that the models are in the most disagreement for events which are at the tail end of being likely or unlikely. This is most probably due to the choice of prior used for the inference of the model M parameters, the ranking prior may mean that model M places higher probability on stronger teams winning (and likewise, less on weak teams).

In this case the most notable difference is between the computational efficiency of the inference of model M and DM. In situations when data are observed sequentially, MCMC methods with MH updates are very inefficient. Given a posterior sample, MH does not immediately provide a method of updating the sample based on any new data. That is, we cannot use the sample as a prior which may be combined with the likelihood of new data in order to sample from a new posterior. One must simply discard the posterior sample and start again using all the available data. Particle filtering methods retain the posterior sample at each point and it is simply updated in accordance with the likelihood of the new data from each week as it arrives.

Furthermore, to take a single sample using MH we must calculate the log-likelihood up to 30 times (the full log-likelihood may not need to be taken if terms may be cancelled in the MH acceptance ratio) for each of the 30 model parameters. A single sample using particle filtering methods may only require the log-likelihood to be calculated once or twice (twice with algorithms which have two stages of weights for example the auxiliary particle filter).

Program run-times are displayed in Figure 4.11 which show the cumulative time taken for the inference process throughout the season. That is the time to generate 100,000 particles/samples from the posteriors $p(\boldsymbol{\theta}_M | \mathbf{D}_w)$ using random-walk MH, and $p(\boldsymbol{\theta}_{DM} | \mathbf{D}_w)$ using the auxiliary particle filter, for each week $w = 1, \dots, 38$. $\boldsymbol{\theta}_M$ denotes the parameter vector for model M, $\boldsymbol{\theta}_{DM}$ is the parameter vector for model DM, and \mathbf{D}_w is all the data up to and including week w . It is clear that the computational time for the particle filtering sampling method scales linearly as the amount of data increases each week. Conversely, due to the nature of repeating the inference process for the entirety of the data each week, the MH sampling method scales exponentially.

There is thus clear benefit in the particle filter method for time critical applications, in particular when the total amount of observed data is large. To perform inference for matches in-play, MH may be useful for the first few weeks, but will soon become too slow to be useful, as the match will have changed significantly (or even finished) by the time the inference process is complete. Particle filtering methods are however capable of updating a whole week's worth of data in less than 10 seconds - fast enough to update posterior distributions for in-play matches.

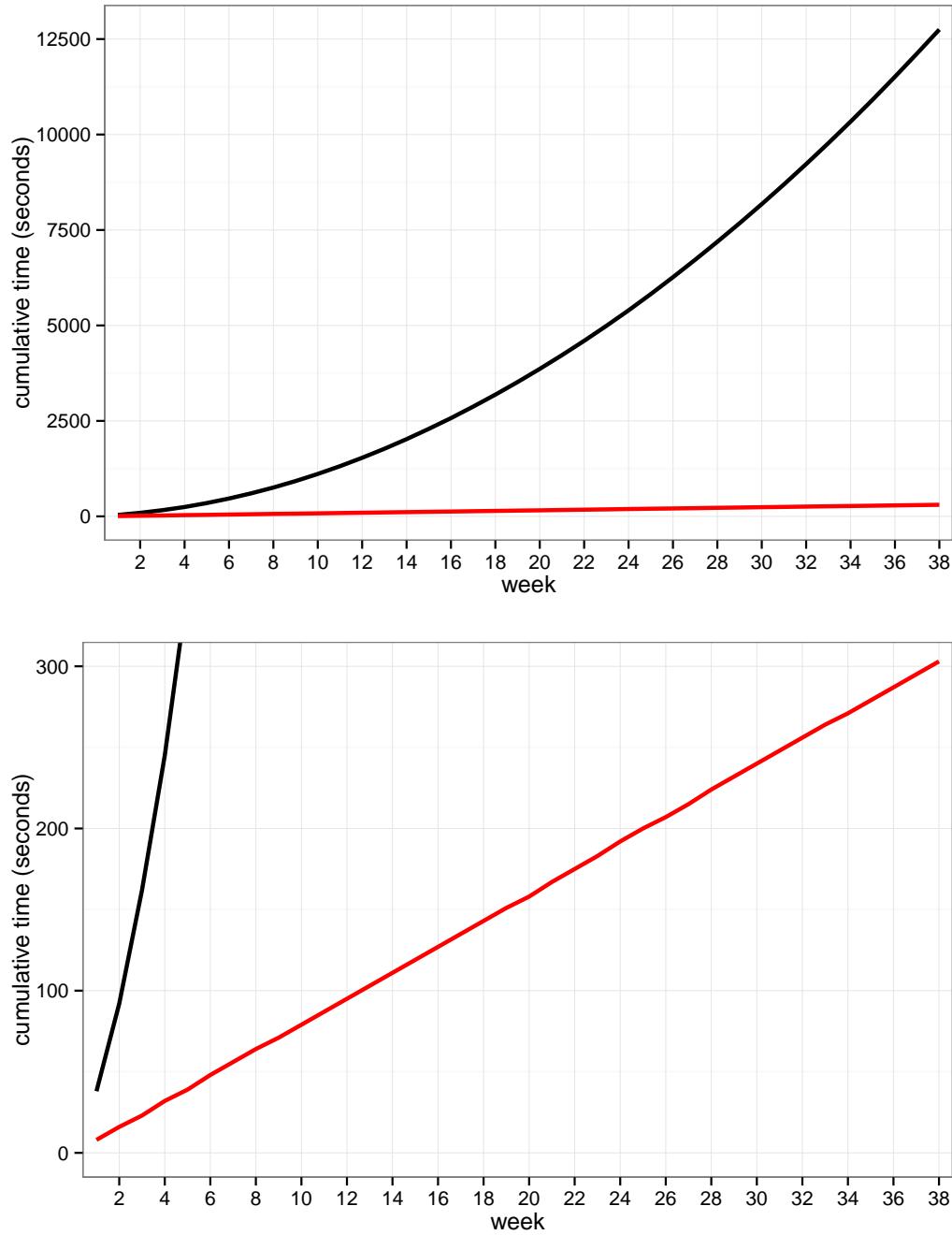


Figure 4.11: The cumulative time to take 100,000 samples from the posteriors $p(\boldsymbol{\theta}_M | \mathcal{D}_w)$ (—) and $p(\boldsymbol{\theta}_{DM} | \mathcal{D}_w)$ (—) for weeks $w = 1, \dots, 38$

4.4.5 Combining MH and particle filtering methods

As noted by Liu and West (2001), sequential filtering methods inherently induce approximation errors which can build up over time. While it does not appear that there are problems with model performance or particle degeneracy in our example, as noted by the performance indicators in Table 4.1 and the time series plots of the model parameters in Figures 4.5 to 4.9, it may be the case for longer time periods or different data/models. In this case, one may choose to ‘refresh’ the particle sample in an off-line setting when there is sufficient time using more standard MCMC methods.

We thus propose that in a real setting, the optimal strategy may be to perform an off-line analysis, using for example MH, in order to generate posterior samples when there is ample time after matches have ended, and then update the samples using on-line particle filtering methods while matches are in-play. This would allow the user to be informed by the most up-to-date posterior distribution during a match.

4.4.6 Inference using simulated data

We also study the performance of the particle filtering method when the true value of θ_w is known for all w . Match result data from a single season are simulated from these true parameter values, we then use the particle filtering algorithm described in Section 4.4.1 to recover the true parameter values. The true values for the dynamic log-resource parameters are as follows:

$$LR_{k,w} = \cos\left(U_k + \frac{w}{15}\right) + \epsilon_{k,w} \quad (4.41)$$

where U_k is a random variate drawn from a $U(0, 2\pi)$ distribution and $\epsilon_{k,w}$ is random noise added at each week w via a $N(0, 0.05^2)$ distribution. The true values for the static parameters are:

$$\begin{aligned} h &= 0.4 & c_0 &= 0.4 \\ a &= 0.1 & c_1 &= 0.6 \\ \rho_1 &= 1 & d_{-1} &= 0.9 \\ \rho_2 &= 1.5 & d_0 &= 0.5 \\ c_{-1} &= 0.7 & d_1 &= 0.3 \end{aligned} \quad (4.42)$$

These true parameter values were chosen so that match simulations result in score-lines which are in line with our expectations of matches in the EPL.

Determination of σ^2 is the same as in Section 4.4.2, but since the underlying team log-resources clearly vary throughout a season, the optimal value of σ^2 is much clearer, as is shown in Figure 4.12. Thus, for the simulated league data, we use $\sigma^2 = 0.06131$ which corresponded to $GM_{1,38} = 0.4487$. The inference process was started

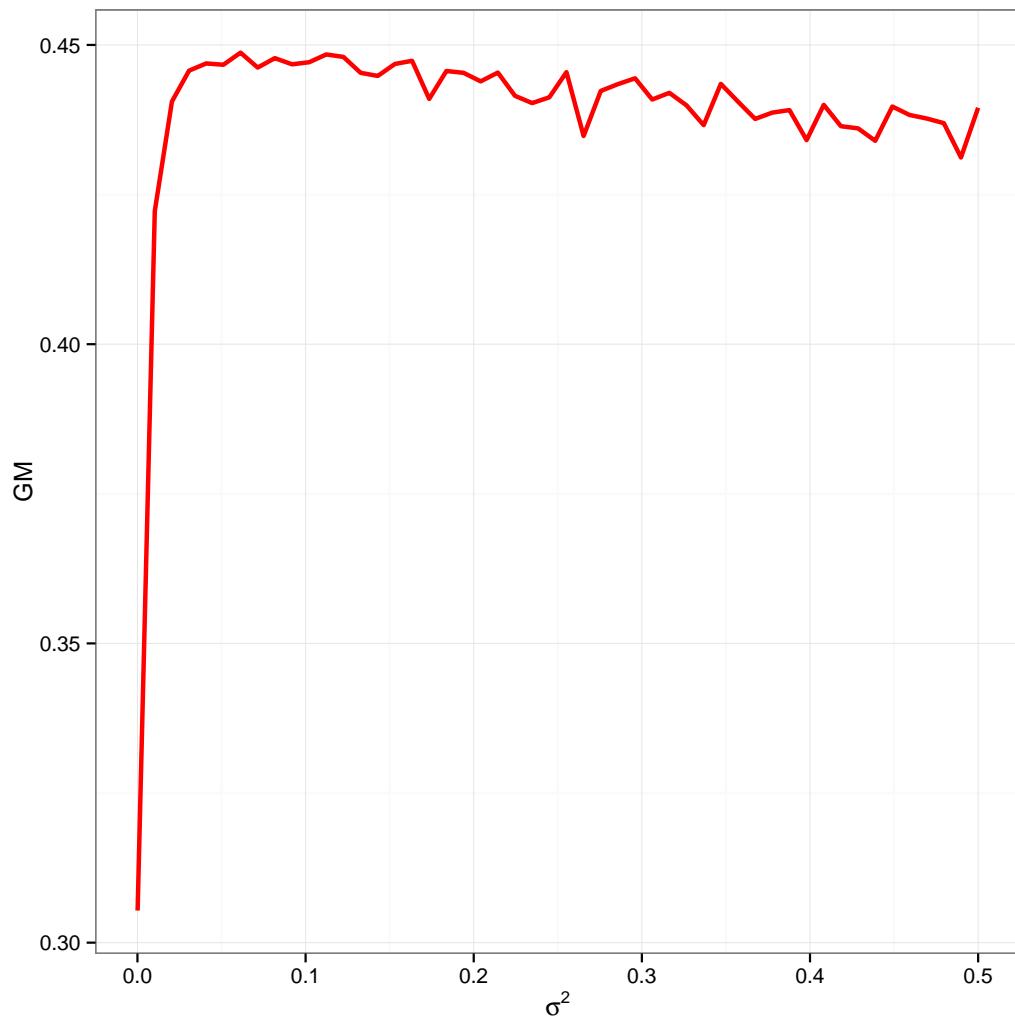


Figure 4.12: Values of $GM_{1,38}$ corresponding to different values of σ^2

by drawing a random sample with mean equal to the true parameter value and it was tested how the particles, approximating the posterior distribution $p(\boldsymbol{\theta}_w | \mathbf{D}_w)$, behaved throughout the weeks $w = 1, \dots, 38$. Again, we display time series plots for the 0.025 and 0.975 posterior quantiles, along with the posterior mean of the particle approximation of $p(\boldsymbol{\theta}_w | \mathbf{D}_w)$, but we can also now display the true underlying parameter values. The plots are shown in Figures 4.13 to 4.17.

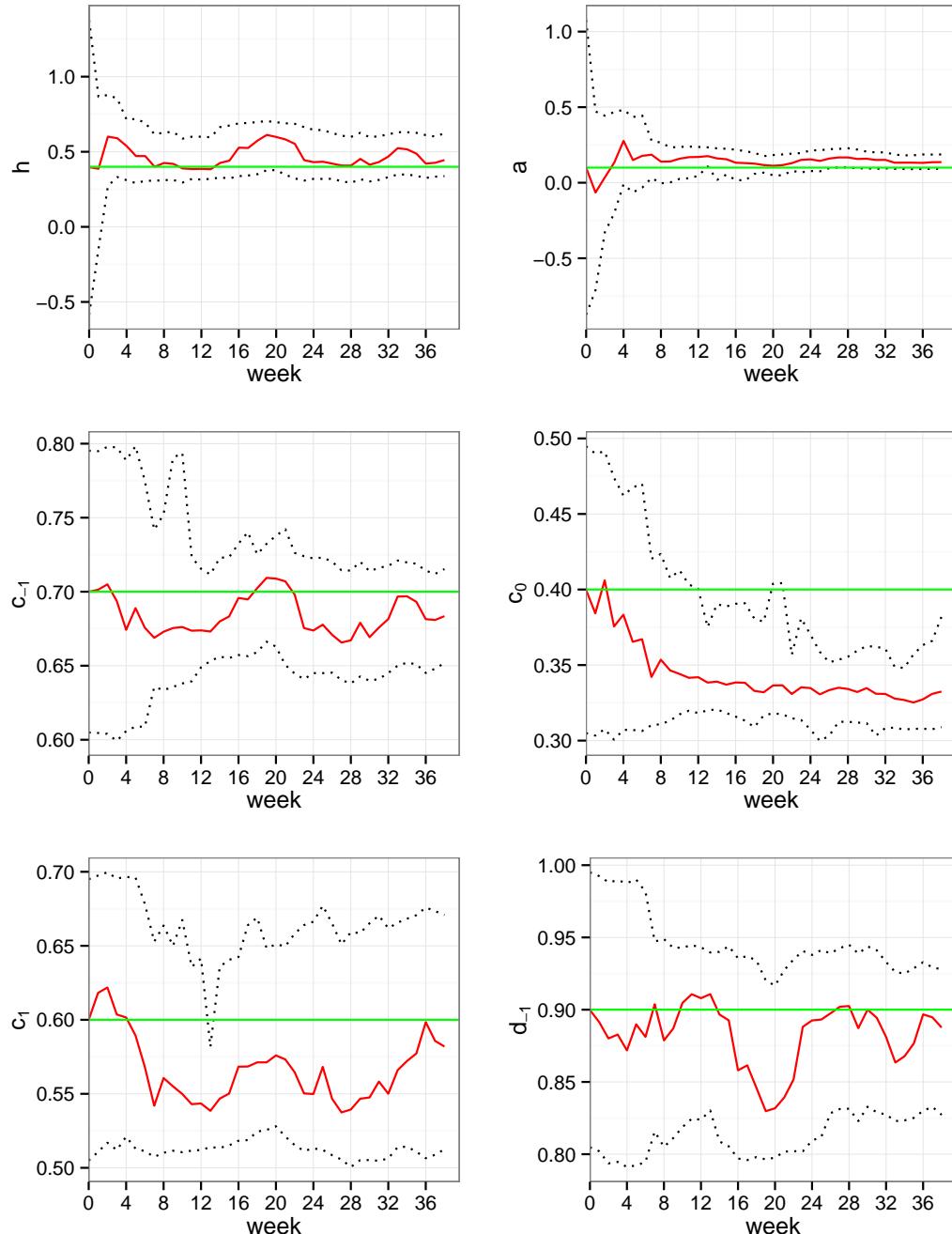


Figure 4.13: Time series plot of parameters h , a , c_- , c_0 , c_1 , and d_- . — posterior mean, --- 95% BCI, — green the true parameter value

Firstly, let us consider the tracking ability of the 10 static parameters. It is clear that the variance of the posterior approximation is deceasing throughout the season,

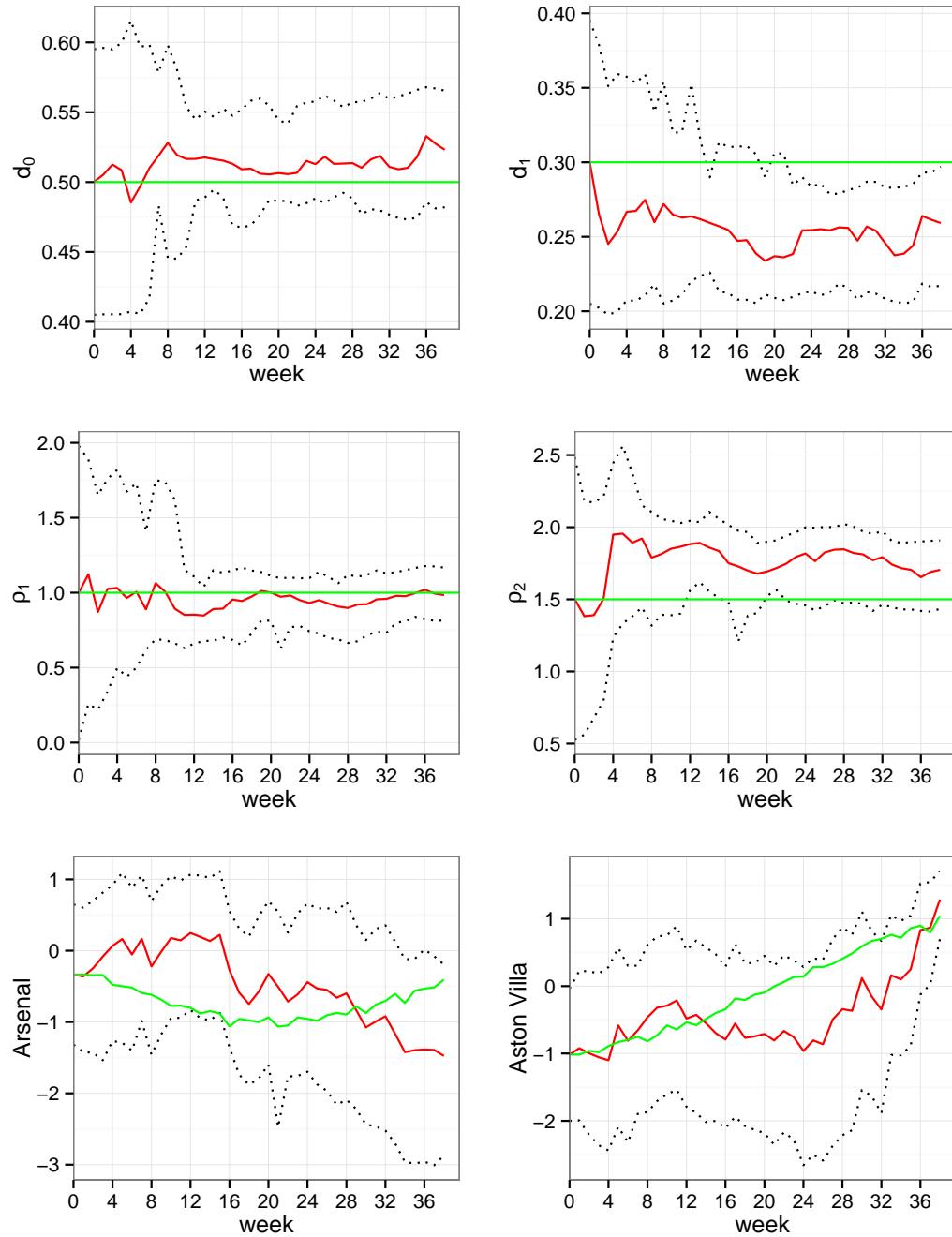


Figure 4.14: Time series plot of parameters d_0 , d_{-1} , ρ_1 , ρ_2 , and the team log-resource parameters. — posterior mean, - - - 95% BCI, — the true parameter value

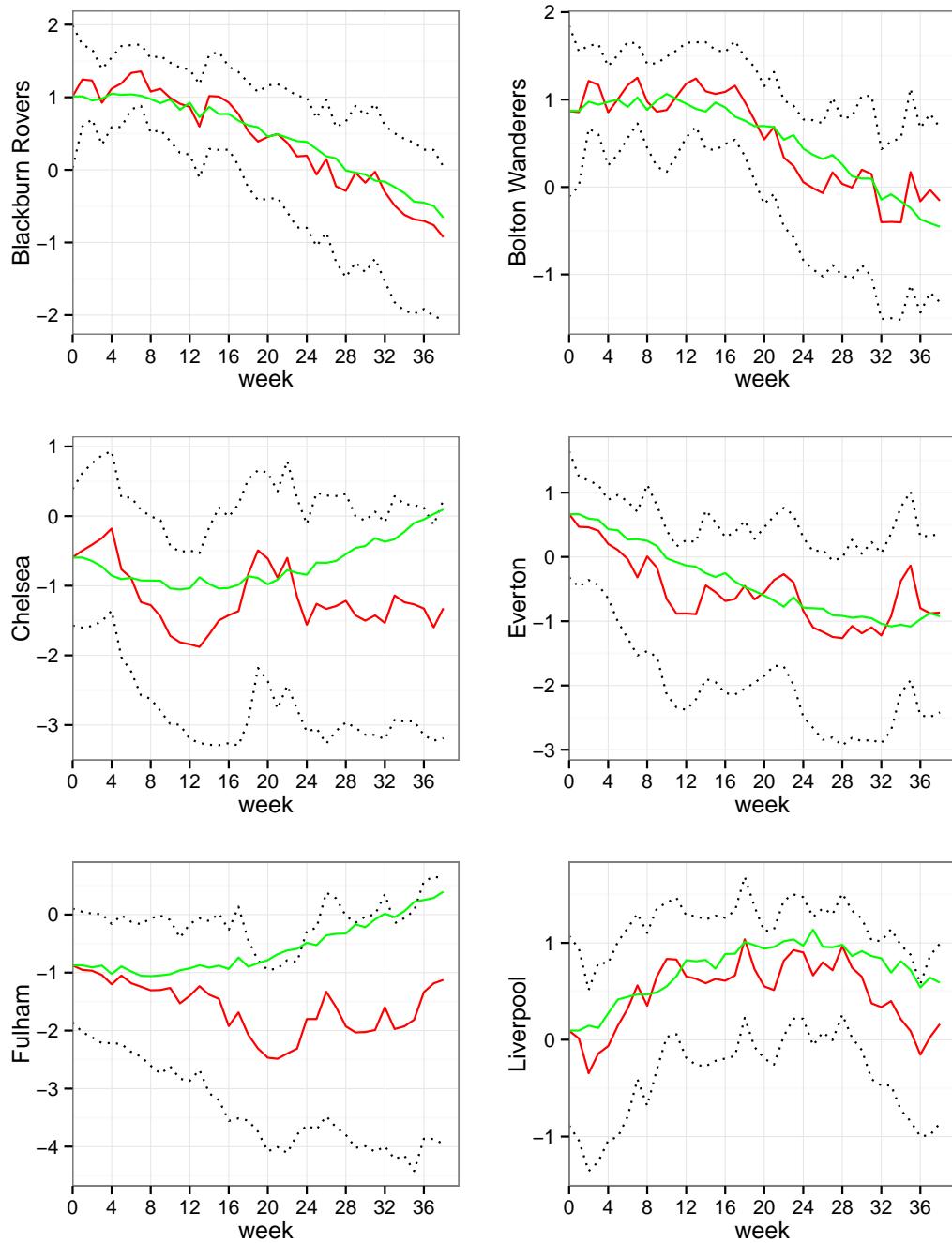


Figure 4.15: Time series plot of the team log-resource parameters. — posterior mean, - - - 95% BCI, — the true parameter value

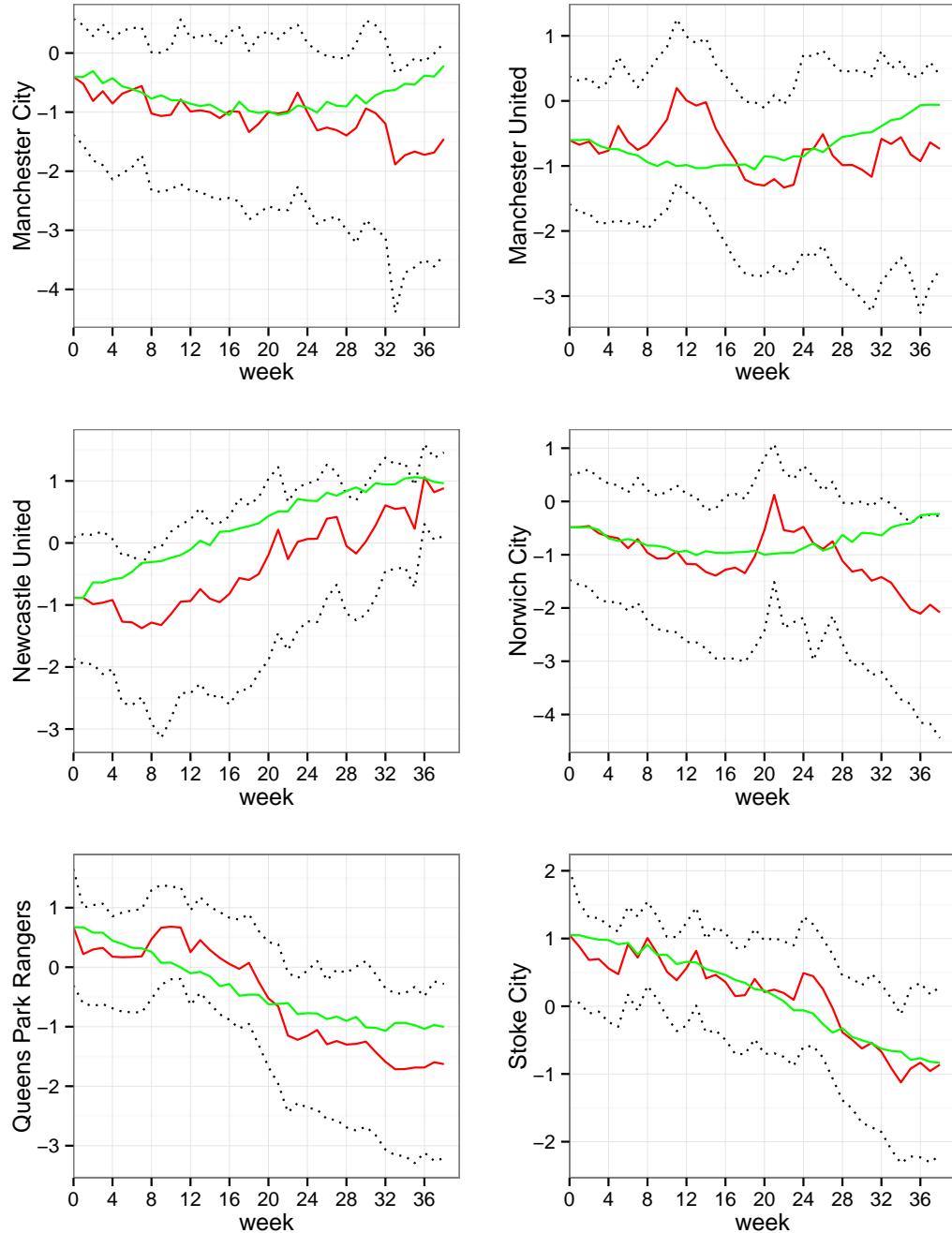


Figure 4.16: Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value

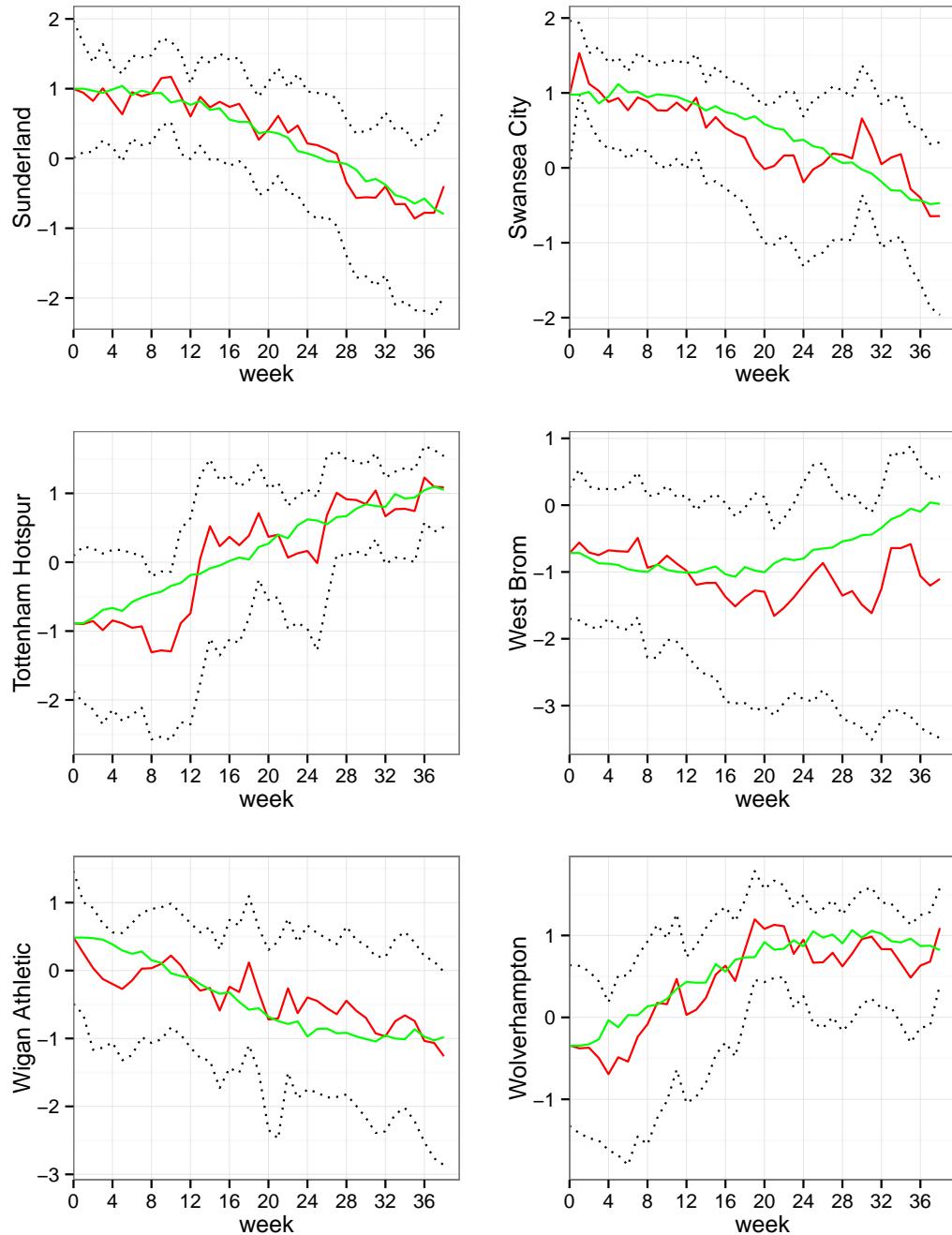


Figure 4.17: Time series plot of the team log-resource parameters. — posterior mean, -·- 95% BCI, — the true parameter value

most notably in the parameters h , a , ρ_1 and ρ_2 - likely due to the fact that these parameters are not constrained to be $\in [0, 1]$ like the c_i and d_i parameters. It is possible to ‘home in’ on the true static parameter value due to the methods of Liu and West (2001) for correcting loss of information in artificial evolution. A similar pattern is not observed in the time series plots for the static parameters using real data (Figure 4.5 and Figure 4.6 in Section 4.4.3) because there the particles at week $w = 1$ were obtained from data from the previous season, and were likely already a very good posterior approximation.

Secondly, the time series plots for the dynamic team log-resource parameters in Figures 4.6 to 4.9 show that the particle filtering methods do provide accurate tracking of the underlying true log-resource parameters. We must also bear in mind that the information of the underlying process is taken from match results, which is a rather noisy process. For example it is not unlikely that a team can be improving and still lose a match to a worse team, in addition sometimes there may be few, or even no goals so information can be scarce. It should also be noted that, in contrast to the 10 static parameters, the posterior variance of the log-resource parameters appears fairly constant throughout the season.

4.5 Concluding remarks

In this chapter we have presented an update to the model in Chapter 3 Section 3.2.3 by allowing the 20 team resource parameters to vary dynamically throughout a season. We then presented efficient particle filtering algorithms which could readily handle such a dynamic underlying system and showed how the methods could be used to also perform inference on other model parameters which were deemed static throughout the season.

We displayed the computational efficiency of the particle filtering methods when compared to more traditional MCMC methods, in the case of when data are observed sequentially. In particular, if one wished to perform in-play posterior updating for betting purposes during a match, the MCMC methods would not be able to provide a posterior sample approximation using all the available data within a reasonable time frame (or likely even before the match finished) whereas the particle filtering methods would simply/quickly update the pre-match posterior samples based on the small amount of data observed in the match. Thus, although the dynamic system model component which particle filtering methods readily handle being somewhat wasted on our tested data, there is still benefit in employing efficient particle filtering methods.

We thus note that one ‘best of both worlds’ solution would be to use traditional

off-line MCMC methods (for example MH) to approximate posterior distributions when there is ample time after matches have finished, and then use on-line particle filtering methods to update the posterior distribution for matches in-play when time is limited.

In Chapter 5 a further modification to our non-homogeneous Poisson process model is presented, with the aim of specifying the model so that model inference is able to capture how teams may modify their behaviour in light of their current league position.

Chapter 5

Incorporating league position into team behaviour

5.1 Introduction

So far in the literature, all statistical models concerned with predicting the outcome of association football matches have made the assumption that the outcomes of distinct concurrent matches are independent. Furthermore, while Poisson process models have included current state of the match as predictors of the scoring rate, none have taken account of league situation.

It is intuitive that football teams should be influenced by their league situation, most notably at the end of a season. Modern technology has also meant that teams are almost instantaneously aware of any changes to their league situation arising from events in other concurrent matches. Some notable examples of this include the occasion when Manchester City scored two goals in injury time to beat Queens Park Rangers 3-2 in the last game of the 2011/2012 season. When information that Manchester United were beating Sunderland in a concurrent game became available, Manchester City, who trailed 1-2, were currently 2nd in the league and needed to win their match to secure 1st place in the league. It appeared that while in this losing match state, Manchester City devoted a large proportion of their resource into attack in order to increase the chance they could win the match and thus the league.

We propose an extension of our non-homogeneous Poisson process model which aims to uncover the extent to which teams change their behaviour depending on their current league situation and analyse the implications of these behavioural changes with regard to model predictions. For example, we will consider questions such as ‘do teams devote all of their resource into attack if they need a goal to prevent league relegation?’ Or, ‘do teams devote all their resource into defense if they only need a draw in order to secure 1st place in the league?’ Furthermore, any evidence

suggesting teams change their behaviour based on their current league situation would also imply that matches are not necessarily independent, since goals in one match can affect the league situation of teams in different matches, and therefore, their tactics.

Evidence of dependence between separate matches in association football suggests a considerable overhaul needs to be made in how bookmakers offer odds. With one of the most popular bets being the ‘1X2 accumulator’ where bettors select one of the home win, draw, or away win events for multiple matches, it should be of utmost interest to bookmakers to ensure that the odds they offer for these bets are correct, that is, profitable. From a bettor’s point of view, if a certain bookmaker offers odds based on the assumption of independence between matches, then there may exist certain situations where a net betting profit is expected when the independence assumption is clearly invalid.

The chapter is organised as follows: Section 5.2 discusses two models (which are not Poisson-process models) from the literature which take account of the team’s league position. Section 5.3 then presents an extension of our non-homogeneous Poisson process model which under the assumption that teams are instantly aware of any league situation changes, allows the rates of scoring to depend on league considerations. Section 5.4 discusses our motivation in selecting the data used to infer model parameters. Section 5.5 considers four different models which we compare using DIC (Spiegelhalter et al. (2002)). Section 5.6 presents an alternative approach to hand-selecting models and comparing them with DIC, via the use of RJMCMC (Green (1995)) to allow for Bayesian model choice. Section 5.7 penultimately shows an example of how the rates of scoring implied by the non-homogeneous Poisson process model change in response to events in concurrent matches, and concluding remarks are presented in Section 5.8.

5.2 Association football models using league information

Scarf and Shi (2008) developed a quantitative measure of ‘match importance’ using Bradley-Terry type models. The idea is that with respect to outcome X , a team would deem a match important if there exists favourable and unfavourable results of the match, conditioned on which the difference in probability of achieving X is large. X can then be taken to denote favourable league outcomes, for example winning the league, qualifying for a European tournament, or not being relegated.

The importance of match m to team k at current week of the season w with respect

to outcome X is defined as:

$$S_k(X)_{w,m} = \mathbb{P}(X_k|F_{k,m}, \mathbf{D}_w) - \mathbb{P}(X_k|U_{k,m}, \mathbf{D}_w) \quad (5.1)$$

where X_k is the event that team k achieves X , $F_{k,m}$ and $U_{k,m}$ denote a favourable and unfavourable outcome of match m for team k respectively, and again, \mathbf{D}_w denotes the data observed up to week w . We note that team k is not necessarily one of the competing teams in match m , but typically team k will deem that matches in which it plays the most important, in which the favourable outcome will be team k winning match m , and the unfavourable outcome will be team k losing match m . For match m in which team k is not competing, the draw outcome can potentially be the favourable or unfavourable outcome, for example if team k is top of the league with 80 points, and teams i and j who both have 78 points compete in match m , the favourable outcome of match m for team k is a draw, for which they will retain the top league position (again, teams are awarded 3 points for a win, 1 for a draw, and 0 for a loss).

$S_k(X)_{w,m}$ is typically calculated by simulating the remainder of the season from current week w using a probabilistic model which does not depend on $S_k(X)_{t,m}$, and fixed results of $F_{k,m}$ and $U_{k,m}$. Hence this method is able to characterise how important specific matches are for broadcasting or tournament design purposes, but is unfortunately not directly useful for prediction. If teams change their behaviour based on the match importance, then the probabilistic model (which does not depend on match importance) used to calculate match importance is invalid - as noted by Scarf and Shi (2008) with regards to using match importance measures for optimising competitor effort.

Goddard and Asimakopoulos (2003) were able to include the concept of match importance into a probabilistic model. They proposed adding covariates denoted $SIGH_{i,j}$ and $SIGA_{i,j}$ (which indicate if the match is deemed significant for the home and away teams respectively) into a Bradley-Terry type model for a match in which team i plays at home to team j as follows:

$$SIGH_{i,j} = \begin{cases} 1 & \text{if match has championship, promotion, or relegation} \\ & \text{significance for team } i \text{ but not for team } j \\ 0 & \text{otherwise} \end{cases} \quad (5.2)$$

$$SIGA_{i,j} = \begin{cases} 1 & \text{if match has championship, promotion, or relegation} \\ & \text{significance for team } j \text{ but not for team } i \\ 0 & \text{otherwise.} \end{cases} \quad (5.3)$$

A match was deemed to be significant if it is still possible for the team in question

to win the championship (league), be promoted, or be relegated, assuming that all other teams currently in contention for the same outcome take one point on average from their remaining matches. It is hoped that this covariate can indicate a difference in incentive between the competing teams i and j , which will typically only become apparent later in the season. Goddard and Asimakopoulos (2003) described the coefficients of $SIGH_{i,j}$ and $SIGA_{i,j}$ as significant at the 1% and 10% levels respectively based on Wald tests (see for example Hosmer Jr et al. (2013)). Data from seasons 1986/1987 to 2000/2001 for four English leagues were used for the analysis.

5.3 A non-homogeneous Poisson process model using a concept of utility

We propose that a team's behaviour should be influenced by the respective values of league positions, which naturally suggests the use of a utility function. An example of how such a utility function might look is given in Figure 5.1. This particular utility function is defined by eight 'knots', between which the function is linear. The knots allow 'jumps' in utility between certain league positions. For example there is a jump in utility when moving between positions 1 and 2 in the league, and similarly between positions 17 and 18 (position 18 is in the relegation zone). Also of importance is the league positions which permit entry to play in further European leagues, the Champions League for positions 1-4, and the Europa League for position 5 (in normal circumstances and ignoring domestic cups which may also grant entry to the European leagues). The governing body, UEFA, released that there was a prize fund of €904.6m to be shared amongst the 32 teams participating in the 2012/2013 Champions League (UEFA (2013a)) and (a still substantial, but much less) €209m to be shared amongst the 56 teams participating in the 2012/2013 Europa League (UEFA (2013b)). Thus, there is a clear monetary incentive for teams to reach league positions which provide qualification for the European leagues, in particular the Champions League.

The first challenge we address is how to formulate a model specification which uses a utility function to determine how the teams behave, that is, how to sensibly incorporate utility values in a formula which defines the function $\alpha_k(t)$. We choose to add an extra term to our existing formulation of $\alpha_k(t)$ so the function is now a mixture of information from the current game situation (as with the previous formulation which considered winning, drawing, or losing in a match) and the current

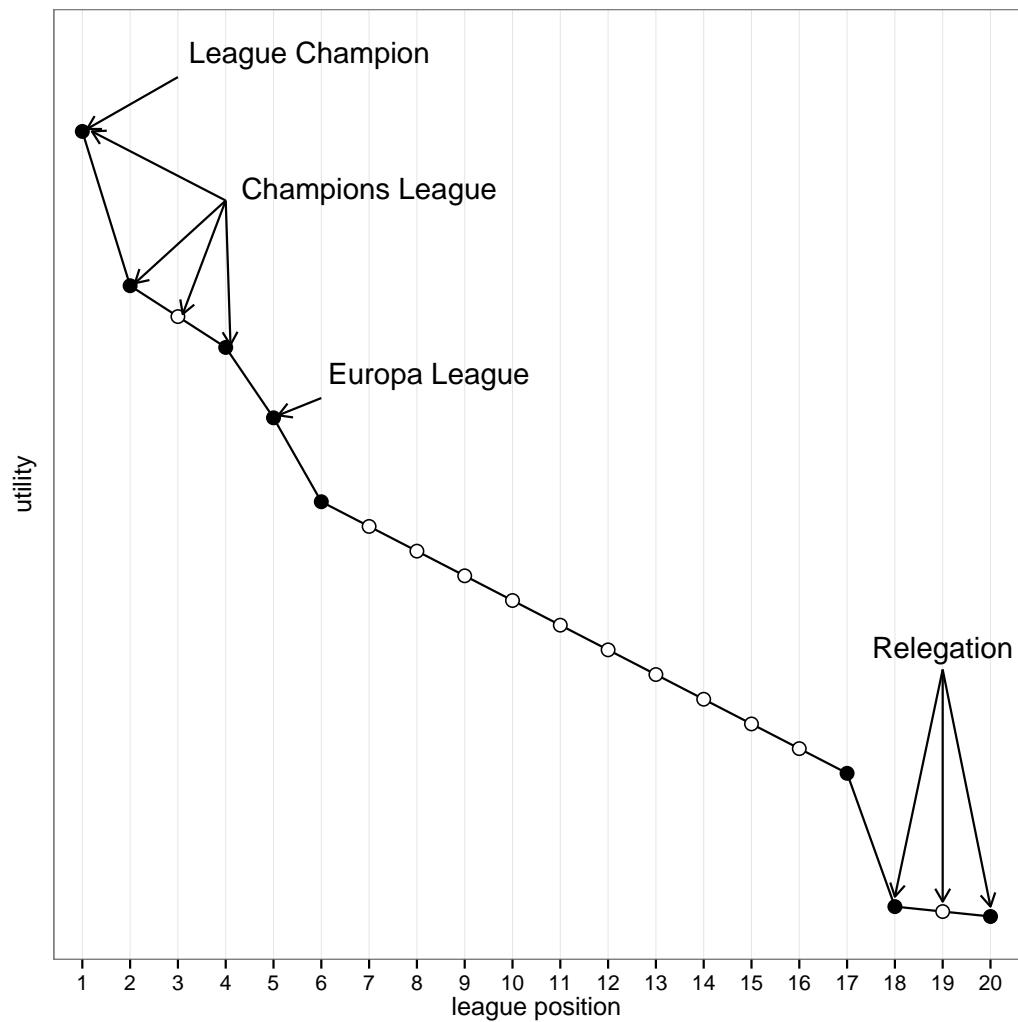


Figure 5.1: An illustration of how a utility function which values each league position might look. Knot locations are denoted by •

league situation:

$$\alpha_k(t, w) = \beta(w)\alpha_{l,k}(t) + (1 - \beta(w))\alpha_{m,k}(t) \quad (5.4)$$

where $0 \leq \beta(w) \leq 1$ is a mixing parameter between the allocation of resource based on the league (l) situation, $\alpha_{l,k}(t)$, and the resource allocation based on the match (m) situation, $\alpha_{m,k}(t)$ (previously $\alpha_k(t)$), in week w .

We expect that $\beta(w)$ will increase from week $w = 1$ to week $w = 38$ as league positions become more important in the later stages of the season, and we aim to capture the extent of the increase by formulating:

$$\beta(w) = g(A + Bw) \quad (5.5)$$

where g is the logistic function:

$$g(x) = \frac{1}{1 + e^{-x}}. \quad (5.6)$$

Note that if the parameter A becomes large and negative and the parameter B is near 0, $\beta(w)$ quickly approaches zero and the resulting model reduces to the previously proposed model in Chapter 3 Section 3.2.3. Inference on the values of $\beta(w)$ will thus demonstrate whether the data show any evidence of teams changing their behaviour towards the end of a season, a time in which Sir Alex Ferguson (a well known Manchester United manager) famously referred to as ‘squeaky-bum time’ (Ferguson (2003)).

Lastly, and somewhat most importantly, we specify:

$$\alpha_{l,k}(t) = g\left(\sum_{p \in P} (U(p) - U(c))\right) \quad (5.7)$$

where P is the set containing the unique league positions of team k if they were to concede 2, 1, or score 1 or 2 goals at time t (the reasoning for this choice is explained later in this section), c is their current position, $U(p)$ denotes the value of the utility function at position p , and again, $g(.)$ denotes the logistic function as in Equation (5.6). This formulation allows the model to capture any evidence of teams changing their behaviour based on the league positions which are realistically attainable (1 or 2 goals away). When none of the 4 scoring scenarios would result in a league position change, the value of $\alpha_{l,k}(t, w)$ will be 0.5. However, when the scoring scenarios result in position changes, $\alpha_{l,k}(t)$ can be greater than 0.5 (as the team plays offensively) or less than 0.5 (as the team plays defensively) depending on the values of utility $U(p)$ for the different obtainable positions p .

The intuition behind this formulation of $\alpha_{l,k}(t)$ is that if scoring a goal would result

in a move up to a new, more desirable position p^+ , then the team may choose to play offensively in order to have an increased chance of scoring. We observe this offensive behaviour via an increase of goals scored but also conceded, and use it to suggest that the utility of position p^+ , $U(p^+)$, is greater than the utility of their current position $U(c)$ and thus the team is playing in order to maximise their expected utility. That is, $U(p^+) - U(c) > 0$ which suggests (ignoring other scoring scenarios) $\alpha_{l,k}(t, w) > 0.5$, corresponding to offensive behaviour.

In a similar fashion, if conceding a goal would move the team down to position p^- , then the team may choose to play defensively in order to decrease the chance of conceding. We observe this defensive behaviour and use it to suggest that $U(p^-) - U(c) < 0$ which suggests $\alpha_{l,k}(t, w) < 0.5$, corresponding to defensive behaviour.

One assumption in this model is that teams play according to their league position if they were to concede 2, 1, or score 1 or 2 goals. We made this assumption with the thought that scoring 1 or 2 goals in a short time frame is quite reasonable in the EPL whereas 3 is quite rare. For example, suppose team k is losing 3-0 which results in 0 league points while a 3-3 draw would result in 1 point and would boost their league position by a single place. Should team k devote more of their resource to attack (play more offensively)? They are almost surely going to lose and we feel that league considerations will be quite irrelevant for the team at that point. Should team k however score, so the match score is 3-1, they may then start playing offensively as they have a chance to score 2 more goals and achieve a 3-3 draw which would result in them moving up a place in the league.

We follow the model specification presented in Chapter 4 Section 4.4 in that we consider the model specification:

$$\log(\lambda_i(t, w)) = h + \alpha_i(t, w)e^{LR_i} - (1 - \alpha_j(t, w))e^{LR_j} + \rho(t) \quad (5.8)$$

$$\log(\mu_j(t, w)) = a + \alpha_j(t, w)e^{LR_j} - (1 - \alpha_i(t, w))e^{LR_i} + \rho(t). \quad (5.9)$$

The model is however not dynamic in that the team resources are considered fixed throughout a season (it is just our estimate of the team resource parameters which changes). Nevertheless we still consider the log-resource which makes computation slightly easier (there is no restriction on the value of LR_k for all teams k).

5.4 Data

As indicated previously, we suspect that the effects of utility decisions will only become apparent at the end of a season, where teams will be fighting last minute to avoid relegation or for one of the top league positions. This not only suggests that the mixing parameter $\beta(w)$ be a function of the week of the season w , but

model	knot positions
m_0	(1, 17, 18, 20)
m_1	(1, 5, 17, 18, 20)
m_2	(1, 2, 5, 17, 18, 20)
m_3	(1, 2, 4, 5, 6, 17, 18, 20)

Table 5.1: A summary of the different knot positions in each of the four models

also that only a small portion of each season's data may display the effects we are looking for - making inference difficult. With this in mind we chose to use five seasons of data, from 2007/2008 to 2011/2012. The data contain the day of each match, but unfortunately not the time, so we make the assumption that matches are played at the same time every day since the model must be aware of the current league standings at all times. These five seasons of data include 29 teams (teams are relegated and promoted each season) and thus we now consider the team log-resource parameters $\mathbf{LR} = (LR_1, \dots, LR_{29})$ which are assumed to be constant from season to season.

5.5 Four competing models

We now consider the problems of model inference and also model choice as we attempt to determine the number and location of knots in the utility function. In this section we consider four competing models, each defined by a different choice of positions for the knots. The models are hand-selected based on our prior belief of what the league utility function might look like and have varying complexity. For example we believe there should be a jump in utility between positions 17 and 18 (again, position 18 is the highest relegation position) which implies we place knots at these positions. We then use DIC as a model selection criterion to choose between the four hand-selected models.

We begin by considering the model which places knots at the same positions as in Figure 5.1, that is, positions (1, 2, 4, 5, 6, 17, 18, 20). We consider this as our most complex model, m_3 , and in addition consider three less complex models which have knots in positions shown in Table 5.1. For all models, the utility function is linear between knot positions, as was in Figure 5.1. The knot positions for each model are fixed, but the corresponding utility value at each knot is a model parameter, which we denote \mathbf{U}_{m_i} for model m_i . However, for all models the value of utility at position 20 is constrained, $U(20) = 0$, to ensure model identifiability. For example, for model m_0 the parameter space is $(\boldsymbol{\theta}, \mathbf{U}_{m_0})$ where $\mathbf{U}_{m_0} = (U_1, U_{17}, U_{18})$ (the utility values at positions 1, 17, and 18 respectively) and $\boldsymbol{\theta} = (h, a, \dots)$ contains the remaining

p	1	2	4	5	6	17	18
U_p^0	10	8	7.5	7	6.5	4	1

Table 5.2: The means of the Γ prior density for the utility value at the knot positions p used within models m_0 , m_1 , m_2 , and m_3

model parameters, common to all models.

5.5.1 Prior choice

The parameters in θ are assigned the following prior distributions which follow from prior distributions used for model inference in Chapters 3 and 4 where possible:

$$\begin{aligned}
 h &\sim N(0.4, 0.5^2) & d_i &\sim B(3, 1) \text{ for } i \in \{-1, 0, 1\} \\
 a &\sim N(0.08, 0.5^2) & LR_k &\sim N(-0.7, 1^2) \text{ for } k \in \{1, \dots, 29\} \\
 \rho_1 &\sim N(1, 0.5^2) & A &\sim N(0, 2^2) \\
 \rho_2 &\sim N(1.5, 0.5^2) & B &\sim Exp(0.5). \\
 c_i &\sim B(1.5, 1.5) \text{ for } i \in \{-1, 0, 1\} &
 \end{aligned} \tag{5.10}$$

We have little information on plausible values of A , and thus use the relatively non-informative prior $N(0, 2^2)$ which should still allow for A to become large and negative, allowing the model to reduce effectively to that of Chapter 3 if the data suggest such a value. B however suggests a change throughout the season in how important utility decisions are, and so we constrain it to be positive (utility decisions become more important towards the end of the season).

For the utility value at position p , we choose the prior $U_p \sim \Gamma(2.5, 2.5/U_p^0)$ so the priors at each position share a common shape parameter and have differing means, U_p^0 . The mean at each position is shown in Table 5.2. We also display a density plot of the utility value prior distributions in Figure 5.2. Of course, the simpler models do not make use of the prior distributions at positions where they do not contain knots.

The variance of the utility value is notably larger for the higher positions when compared to position 18. The model is constrained so that the utility of position 20 is zero ($U(20) = 0$) and so we can be fairly sure that neighbouring positions (for example position 18 which is also a relegation position) have a utility fairly near 0. As we move further away from position 20 however, it is harder to reasonably estimate values of the utility function, and hence, we allow the variance of the utility values to increase with the mean.

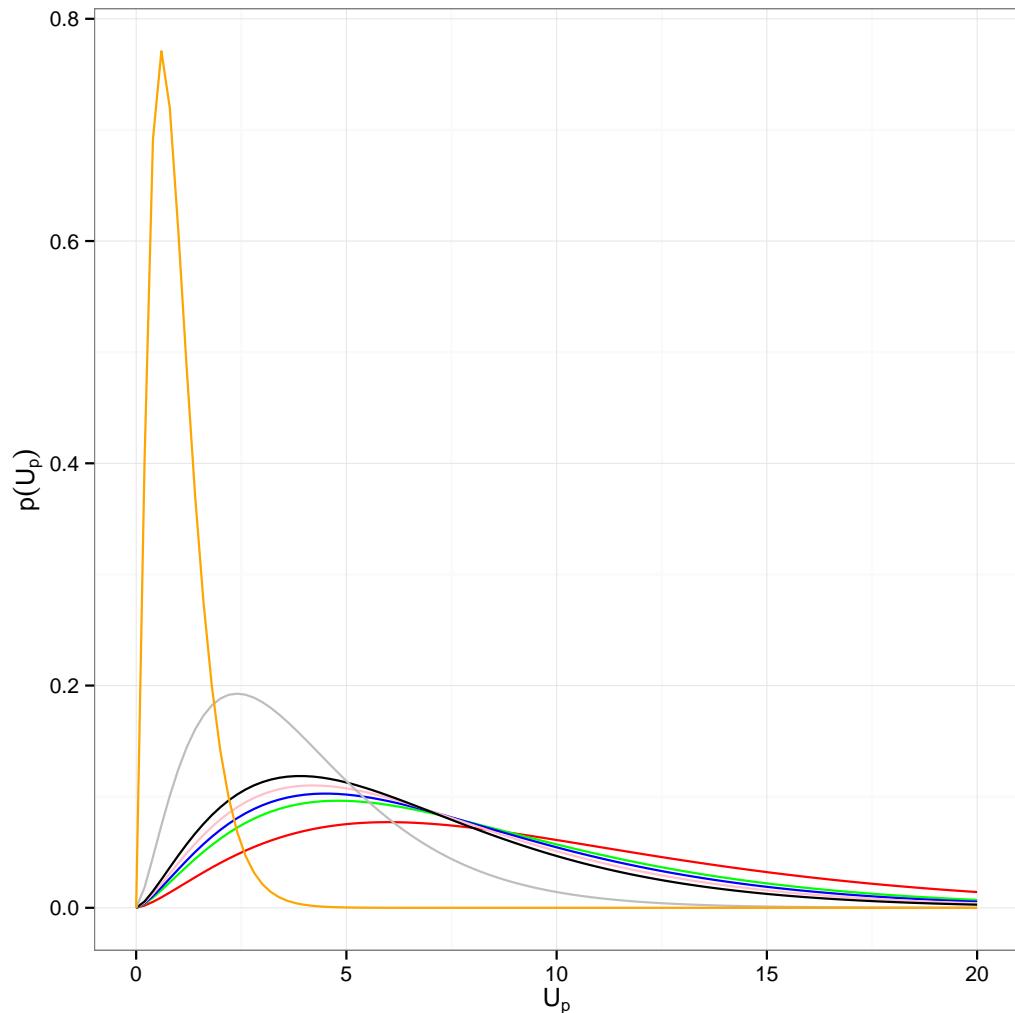


Figure 5.2: A plot of the prior density of the utility value for the possible knot locations. — position 1, — position 2, — position 4, — position 5, — position 6, — position 17, — position 18

5.5.2 Results of model fitting

For each model m_i we take 40,000 posterior samples from the joint posterior distribution of $(\boldsymbol{\theta}, \mathbf{U}_{m_i})$ using a random-walk MH algorithm. An optimisation routine for the log-likelihood was performed in order to find the MLE $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{U}}_{m_i})$ which was used as a starting value. For a similar model, we have already considered the posterior distribution of the parameters $\boldsymbol{\theta}$ in Chapter 3 Section 3.3. Here we only report the posterior distributions of the new parameters related to the utility function. We do however note that our posterior estimate of $\boldsymbol{\theta}$ will in general now be slightly different to that reported in Chapter 3 Section 3.3, due to the different model and data employed. Figures 5.3, 5.4, 5.5, and 5.6 display plots of posterior samples from the functions $U(p)$ and $\beta(w)$ for each of the four models. The plots display a random sample of size 1,500 from the 40,000 posterior samples taken to prevent plot rendering problems.

The plots showing posterior samples of the function $\beta(w)$ are all quite similar for each of the four models, and show that in the first match of the season the function is in the region of 0.05, increasing to values in the region of 0.4 in the last match of the season.

The plots showing posterior samples of the utility function $U(p)$ are largely as we would expect. There is a clear difference between the utility value of relegation positions (18, 19, and 20) and the first non-relegation position (17). The value of utility then in general increases upwards towards the league winner (position 1). However, there is one very noticeable, and unexpected result shown in the inference of model m_3 . The plot suggests that $U(6) > U(5)$, that is, the value of position six is greater than the value of position five, as shown by the behaviour of teams. In fact, the posterior probability $\mathbb{P}(U(6) > U(5)) = 0.9804$.

We mentioned in Section 5.3 that position 5 granted qualification to the Europa League, and thought this would provide incentive for teams to reach this position. We naturally then placed a prior distribution on the utility function values which suggested that $U(5) > U(6)$. The posterior distribution however contradicts our prior belief, and suggests that teams will play offensively in order to get out of position 5, and play defensively to defend position 6 instead of moving up to position 5. This contradiction between the prior and posterior distributions suggests that the data may contain strong evidence of an unexpected effect in team behaviour related to position 5, which we discuss in further detail.

In 2015 several on-line articles appeared suggesting that EPL teams may prefer not to play in the Europa League, for example the aptly named ‘The Race to Avoid Europa League Qualification Starts Now’ (Potton (2015)), ‘The battle to avoid the Europa League’ (Lea (2015)), and ‘Tottenham’s Mauricio Pochettino: avoiding Eu-

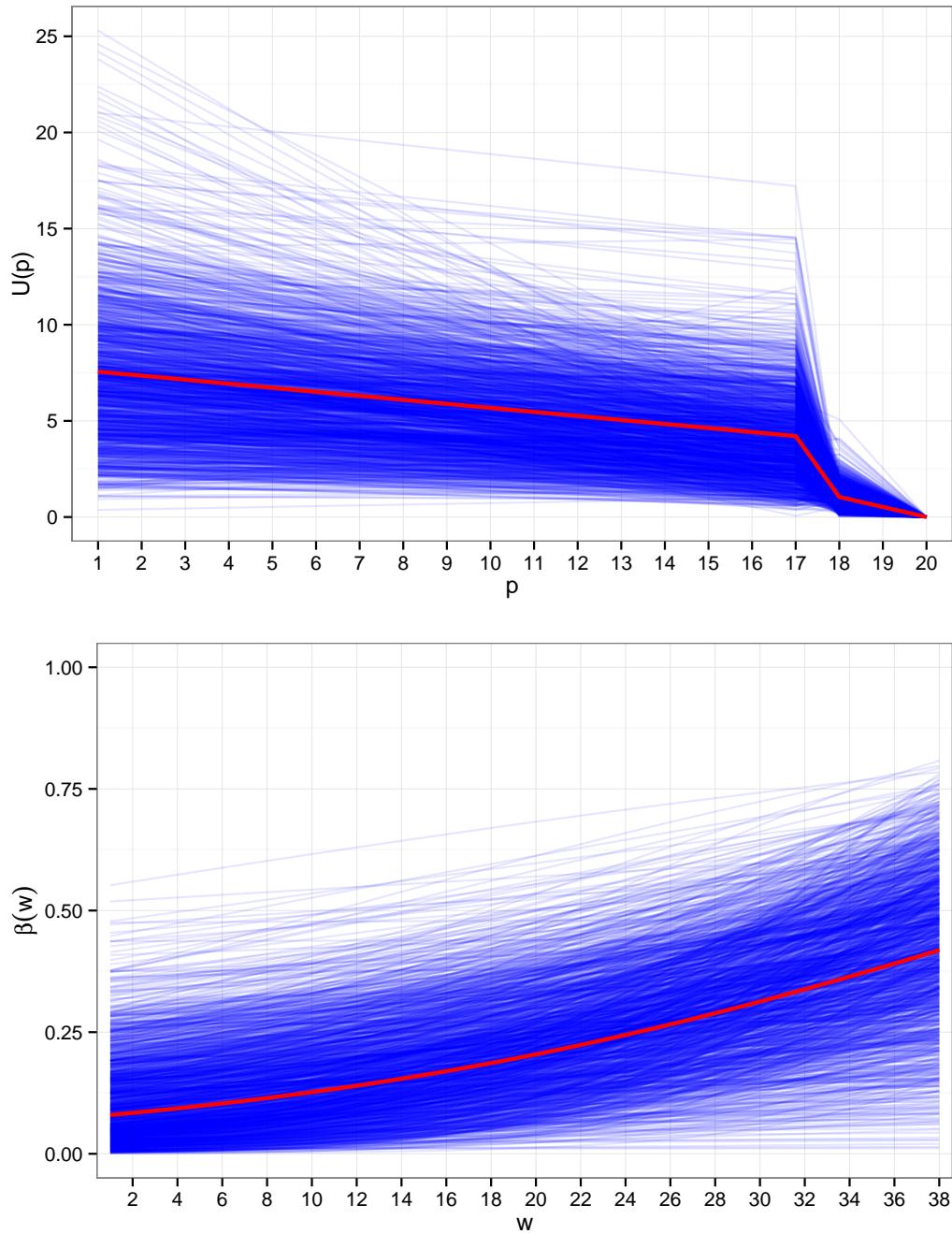


Figure 5.3: A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_0 . — the posterior mean

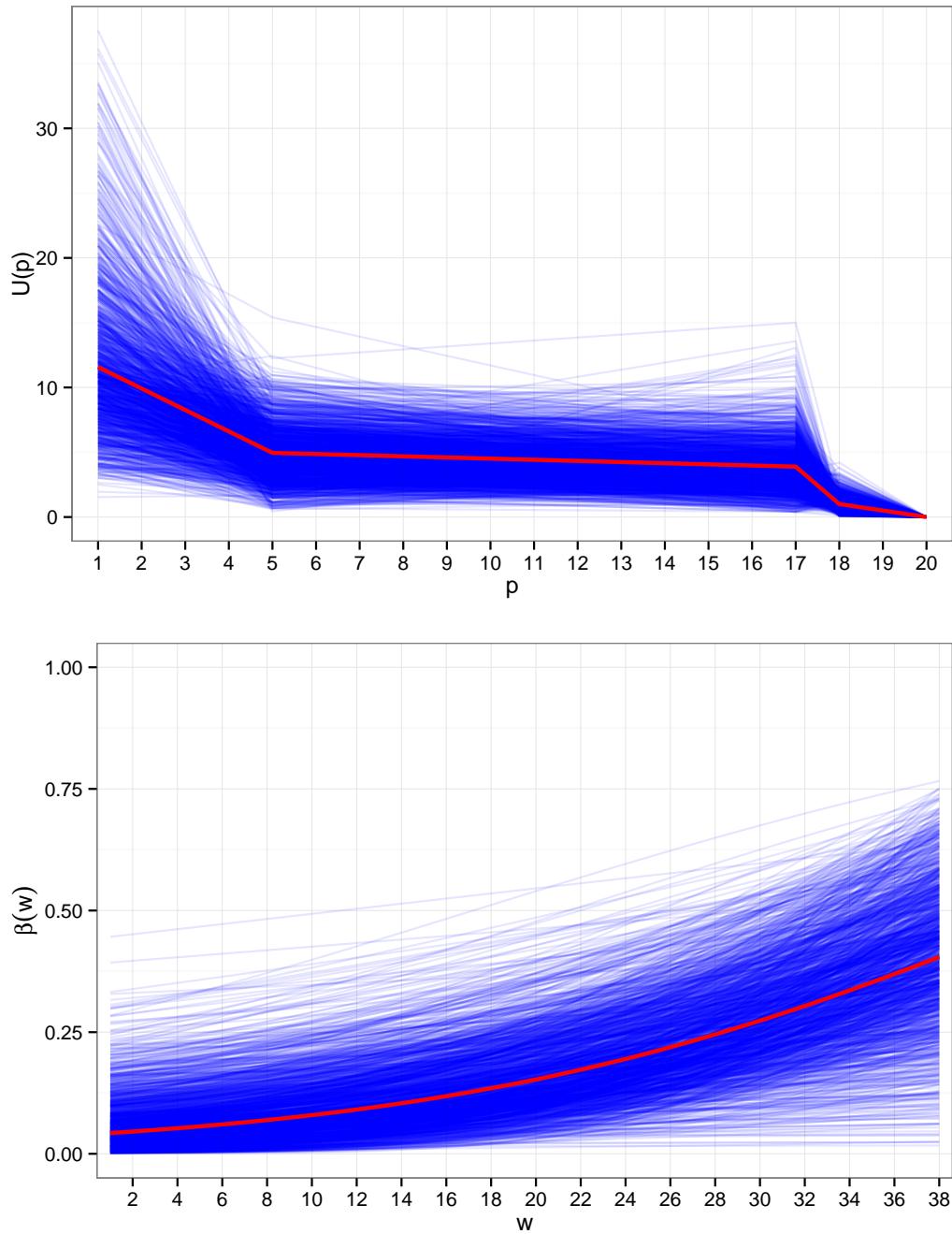


Figure 5.4: A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_1 . — the posterior mean

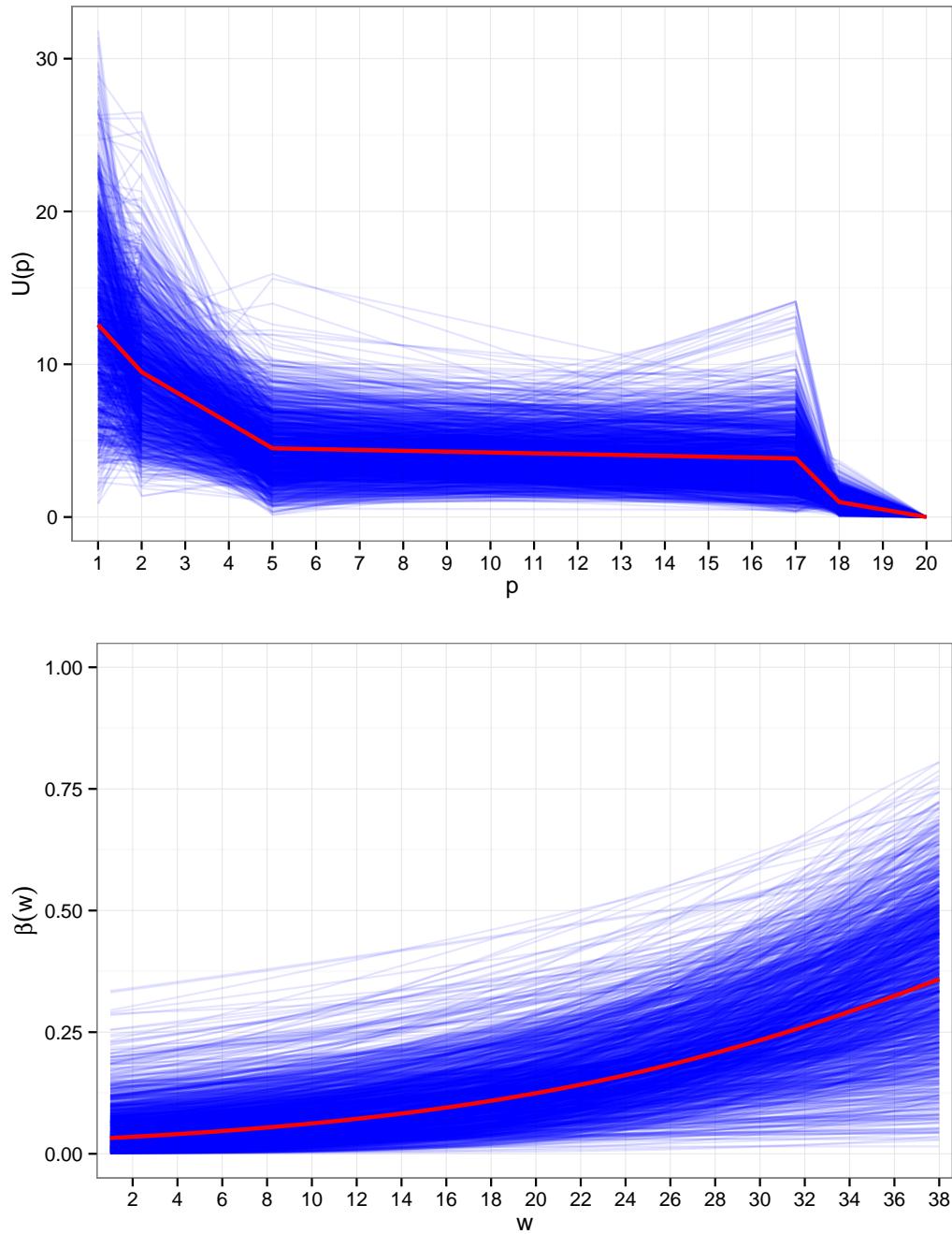


Figure 5.5: A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_2 . — the posterior mean

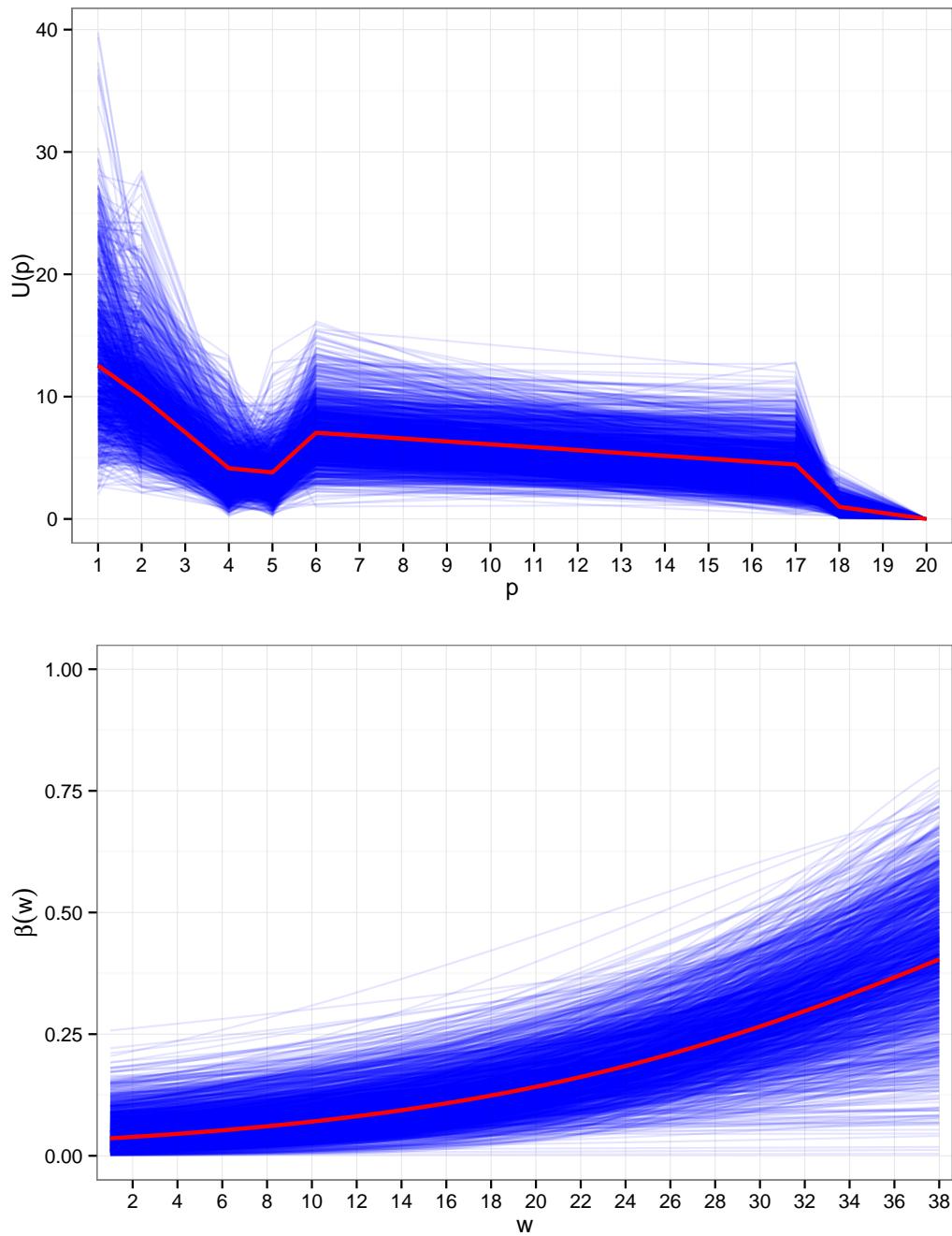


Figure 5.6: A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) for model m_3 . — the posterior mean

model	p_D	DIC
m_0	35.7211	5650.54
m_1	36.7981	5647.54
m_2	36.4681	5647.96
m_3	37.3923	5642.70

Table 5.3: p_D and DIC for the four competing models

ropa League could help us' (Hytner (2015)). The articles all imply that the extra travelling around Europe and matches (up to 23) is a potential burden on teams. An EPL team typically plays a Europa League match on a Thursday, and then a EPL match on a Sunday, and it is widely thought that players struggle to recover physically and/or mentally between the frequent matches. Thus playing in the Europa League may have a negative affect on a teams performance in the EPL and furthermore, as mentioned in Section 5.3, the monetary rewards of the Europa League are not huge (in European association football terms).

On reflection it does seem plausible that teams behave in order to avoid qualification for the Europa League - and our model inference is consistent with this thought. We do however realise that there could be other explanations for observing the locally low value of $U(5)$ in Figure 5.6. One thought is that a particular team may have been in position 5 for a large amount of time throughout the data, and it may be that this team had a particularly offensive style of play. Thus, the team may not have been devoting a larger proportion of resource into attack in order to move out of position 5 (to 4 or 6), they are behaving this way simply because it is their style of play. The same thoughts apply to a particularly defensive team which may have spent a large amount of time in position 6.

5.5.3 Model choice using DIC

DIC values calculated from 40,000 posterior samples are shown in Table 5.3. The differences in values of DIC are not dramatic between the models, but do suggest that model m_3 is preferred.

We also note the rather counter-intuitive finding of the effective number of parameters p_D being greater for model m_1 when compared to model m_2 . We might not expect this since $\dim(\mathbf{U}_{m_2}) > \dim(\mathbf{U}_{m_1})$, but as can be seen in Figure 5.4 and Figure 5.5, it appears that the extra parameter in the utility function for model m_2 actually serves to decrease the variance in the utility function (note the differing y-axis scales on each figure).

Taking the approach of comparing models on DIC, one might stop here and base

all analysis on the ‘best’ model, model m_3 , which as mentioned in Section 5.5.2, displayed unexpected evidence of teams playing in order to not be in position 5 (Europa League qualification position). We however present an alternative approach to choosing a single model and basing all analysis upon it in the following section using the method of Bayesian model averaging first presented in Chapter 2 Section 2.6.1.

5.6 Model inference using RJMCMC

In Section 5.5 we considered four models, each defined by a different set of knot positions. There are however a total of $2^{18} = 262,144$ possible models, which each either does or does not contain a knot at any of the 18 available positions, $2, 3, \dots, 19$. We design a sampling procedure using RJMCMC (reviewed in Chapter 2 Section 2.4) which explores the posterior distribution on a space of models and the corresponding model parameters at the same time. Thus using a data-driven approach to decide what knot positions should be considered.

This approach allows us to use Bayesian model averaging in order to infer the utility function $U(p)$ by averaging over the space of possible models (possible knot locations). We shall then see how close the utility function obtained from Bayesian model averaging is to the utility function from simply using model m_3 . For example, is it possible that we did not even consider a model defined by a choice of knot locations that would better fit our test data?

The parameter space of interest is $(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta})$ where $\boldsymbol{\kappa}$ is the league positions of the knots, \mathbf{U}_κ are their corresponding utility values, and $\boldsymbol{\theta} = (h, a, \dots)$ as previously contains the remaining model parameters. Furthermore, we denote $U_m(p)$ to be the value of utility under model m at position p , and denote the number of knots as $K = \dim(\boldsymbol{\kappa})$. We follow a similar method to that of Punska et al. (1999), in that for each iteration of the RJMCMC we propose either a birth of a knot, the death of a knot, or the movement of an existing knot, followed by random-walk MH updating of the parameters $\boldsymbol{\theta}$ which are not related to the utility function. As in Section 5.5, we fix the knot at position 20 with $U(20) = 0$ and constrain that there is always a knot at position 1 to ensure model identifiability. The allowable values of K are thus the integers from 2 to 20. An outline of a single iteration of the algorithm after initialisation of the parameters is as follows:

1. Sample $s \sim U(0, 1)$
2. If $s < p_{birth}(K)$ then propose the birth of a new knot
3. Else if $s < p_{birth}(K) + p_{death}(K)$ then propose the death of an existing knot

4. Else propose a movement to one of the existing knot utility values, \mathbf{U}_κ
5. Update $\boldsymbol{\theta}$ using random-walk MH

where:

$$p_{birth}(K) = \begin{cases} 1/2 & \text{if } K = 2 \\ 0 & \text{if } K = 20 \\ 1/3 & \text{otherwise} \end{cases} \quad (5.11)$$

$$p_{death}(K) = \begin{cases} 0 & \text{if } K = 2 \\ 1/2 & \text{if } K = 20 \\ 1/3 & \text{otherwise} \end{cases} \quad (5.12)$$

so the functions $p_{birth}(K)$ and $p_{death}(K)$ give the probability of a proposal of a birth or death respectively depending on the current number of knots, (minimum 2 and maximum 20). Parameters $\boldsymbol{\theta}$ are updated using random-walk MH methods, the birth and death proposals are however slightly more non-standard and are thus explained in Sections 5.6.1 and 5.6.2.

5.6.1 Knot birth

If the current state of the chain is $(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta})$ for which we denote corresponding model m , the methodology for the proposal for the birth of a new knot is as follows:

1. Propose model m' which places a knot in one of the $20 - K$ currently available positions with equal probability, denote this position p' and thus $\boldsymbol{\kappa}' = (p', \boldsymbol{\kappa})$
2. Sample $u' \sim N(U_m(p'), \sigma_\kappa^2)$, our proposal for the utility value at the new position p' , and set $\mathbf{U}'_\kappa = (u', \mathbf{U}_\kappa)$
3. Accept model m' with probability α_{birth}

where:

$$\alpha_{birth} = \min \left(1, \frac{p(\mathbf{D}|\boldsymbol{\kappa}', \mathbf{U}'_\kappa, \boldsymbol{\theta})}{p(\mathbf{D}|\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta})} \times \frac{p(\boldsymbol{\kappa}', \mathbf{U}'_\kappa)}{p(\boldsymbol{\kappa}, \mathbf{U}_\kappa)} \times \frac{p_{death}(K')}{p_{birth}(K)} \times \frac{\frac{1}{K'-2}}{\frac{1}{20-K}} \times \frac{1}{p(u')} \right). \quad (5.13)$$

In this case the proposal state is $(\boldsymbol{\kappa}', \mathbf{U}'_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu}')$ with $\dim(\boldsymbol{\mu}') = 0$. The additional parameters $\boldsymbol{\mu} = (p', u')$ (which are included in the vectors $\boldsymbol{\kappa}'$ and \mathbf{U}'_κ) are drawn directly from probability distributions and all remaining parameters are the same

from models m to m' . We may write the parameter transformation function as:

$$g_{m,m'}(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu}) = (\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, p', u', \boldsymbol{\mu}') \quad (5.14)$$

that is, it does not actually perform any transformations. It is fairly straightforward to see that:

$$\left| \frac{\partial g_{m,m'}(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu})}{\partial(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu})} \right| = \left| \frac{\partial(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, p', u', \boldsymbol{\mu}')}{\partial(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu})} \right| = 1. \quad (5.15)$$

As was noted in Punkska et al. (1999), if the proposal (in this example $\boldsymbol{\mu}$) is made directly in the new parameter space (as opposed to using dimension matching random variables) the Jacobian term is equal to 1.

Figure 5.7 shows a graphical illustration of the knot birth process. Firstly, a birth of a knot in position $p' = 10$ is chosen uniformly from the available positions $2, 3, \dots, 19$. The utility value at the new knot position is then sampled from the proposal density $N(U_m(p'), \sigma_\kappa^2)$. Model m' containing the new knot is then accepted with probability α_{birth} .

5.6.2 Knot death

We follow from the knot birth and explain how to reverse the birth move. At current state $(\boldsymbol{\kappa}', \mathbf{U}'_\kappa, \boldsymbol{\theta}')$ for which we denote corresponding model m' and using the notation \mathbf{A}_{-a} to denote the vector \mathbf{A} with value a removed, the methodology for the proposal for the death of an existing knot is as follows:

1. Propose model m which removes one of the $K' - 2$ current knot positions with equal probability, denote the removed position p' with corresponding utility value u' . Thus the proposal parameters are $\boldsymbol{\kappa} = \boldsymbol{\kappa}'_{-p'}$ and $\mathbf{U}_\kappa = \mathbf{U}'_{\kappa, -u'}$
2. Accept model m with probability α_{death}

where:

$$\alpha_{death} = \min \left(1, \frac{p(\mathbf{D}|\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta})}{p(\mathbf{D}|\boldsymbol{\kappa}', \mathbf{U}'_\kappa, \boldsymbol{\theta})} \times \frac{p(\boldsymbol{\kappa}, \mathbf{U}_\kappa)}{p(\boldsymbol{\kappa}', \mathbf{U}'_\kappa)} \times \frac{p_{birth}(K)}{p_{death}(K')} \times \frac{\frac{1}{20-K}}{\frac{1}{K'-2}} \times \frac{p(u')}{1} \right). \quad (5.16)$$

In this case the proposal parameters are $(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu})$ with $\boldsymbol{\mu} = (p', u')$ containing the removed parameters and $dim(\boldsymbol{\mu}') = 0$. The parameter transformation function is the inverse of that seen for the birth:

$$g_{m',m}(\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, p', u', \boldsymbol{\mu}') = (\boldsymbol{\kappa}, \mathbf{U}_\kappa, \boldsymbol{\theta}, \boldsymbol{\mu}) \quad (5.17)$$

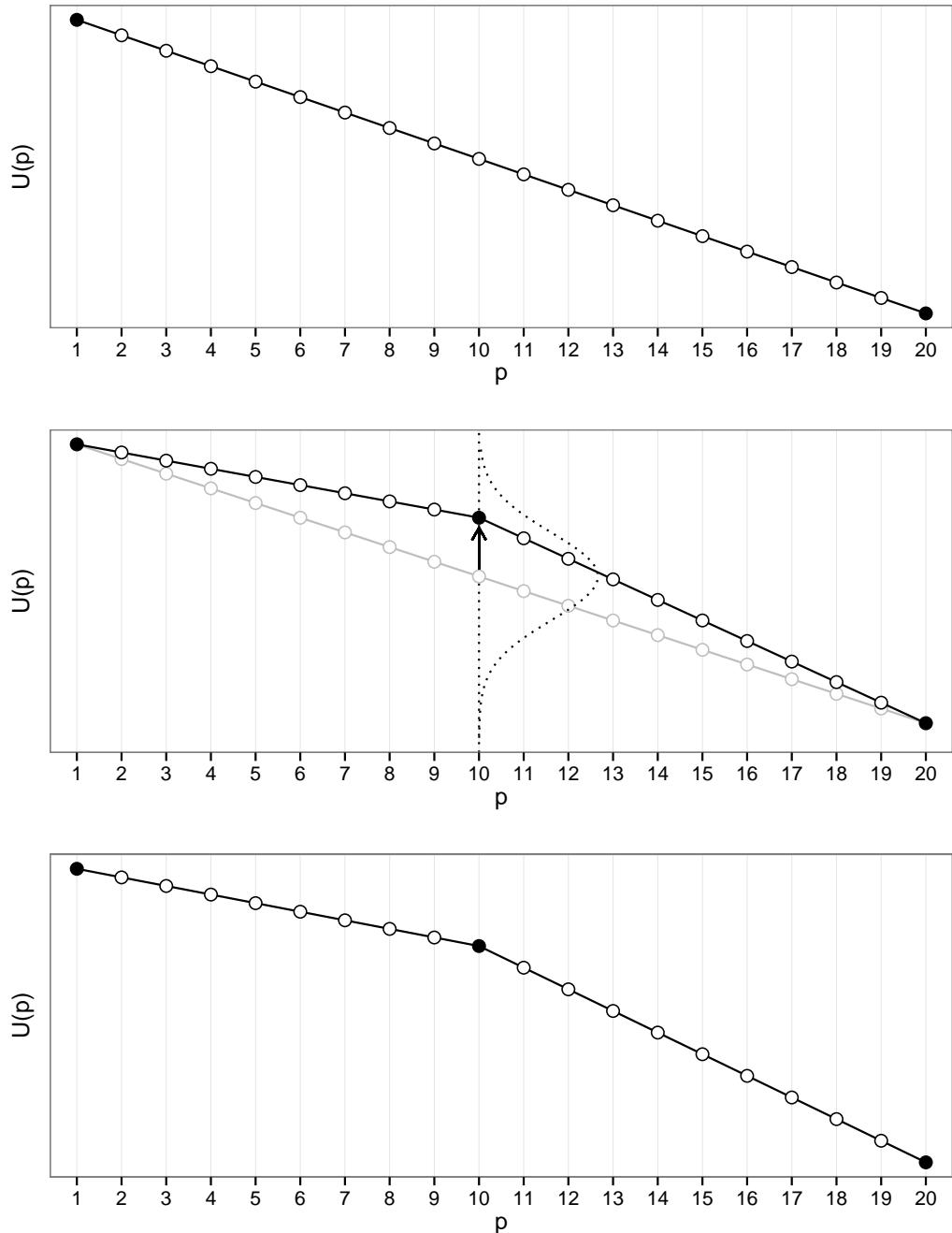


Figure 5.7: The utility function given by model m (top), the proposal of a new knot in position 10 (middle), the resulting utility function of model m' (bottom). • utility function knot locations, - - - the superimposed proposal density

p	1	2	3	4	5	6	7	8	9	10
U_p^0	10	8	7.75	7.5	7	6.5	6.27	6.05	5.82	5.59
p	11	12	13	14	15	16	17	18	19	
U_p^0	5.36	5.14	4.91	4.68	4.45	4.23	4	1	0.5	

Table 5.4: The means of the Γ prior density for the utility value at each possible knot position

and again, does not perform any transformations so the Jacobian term is equal to 1. Also it can be seen that α_{birth} and α_{death} are reciprocals, as is needed in order for the Markov Chain to satisfy detailed balance (Hastie and Green (2012)).

Figure 5.8 shows a graphical illustration of the knot death process. Firstly, the death of the knot in position $p' = 3$ is chosen uniformly from the available knot positions 3, 5, 10. Model m containing one fewer knot is then accepted with probability α_{death} .

5.6.3 Prior choice

The prior distributions used for $\boldsymbol{\theta}$ follows from Section 5.5.1, we now however place a prior on the number of knots K , the knot values \mathbf{U}_κ , and the knot positions κ .

The position of two knots in κ are fixed at 1 and 20. There is then a uniform prior on the position of the potential remaining knots in positions 2 to 19. We extend Table 5.2 with similar reasoning as previously discussed to show the Γ density prior means at all the possible knot positions, which can be seen in Table 5.4, again, $U_p \sim \Gamma(2.5, 2.5/U_p^0)$. Finally, we place a prior on the number of knots, K :

$$\mathbb{P}(K) = \begin{cases} \frac{f(K)}{1-F(1)} & \text{if } 2 \leq K \leq 20 \\ 0 & \text{otherwise} \end{cases} \quad (5.18)$$

where $f(K)$ is the discrete binomial probability mass function at point K with parameters n and α , and $F(1)$ is the corresponding cumulative distribution function at point 1. $\mathbb{P}(K)$ is then a conditional binomial distribution, the condition being $K > 1$. $n = 20$, and we choose $\alpha = 0.05$ which means the prior favours models with a smaller number of knots (this prior distribution can be seen in Figure 5.10). We thus argue that a high posterior probability of the presence of a knot is due to significant evidence in the data.

The prior (as seen in Equations 5.13 and 5.16) for the vector of knot positions and

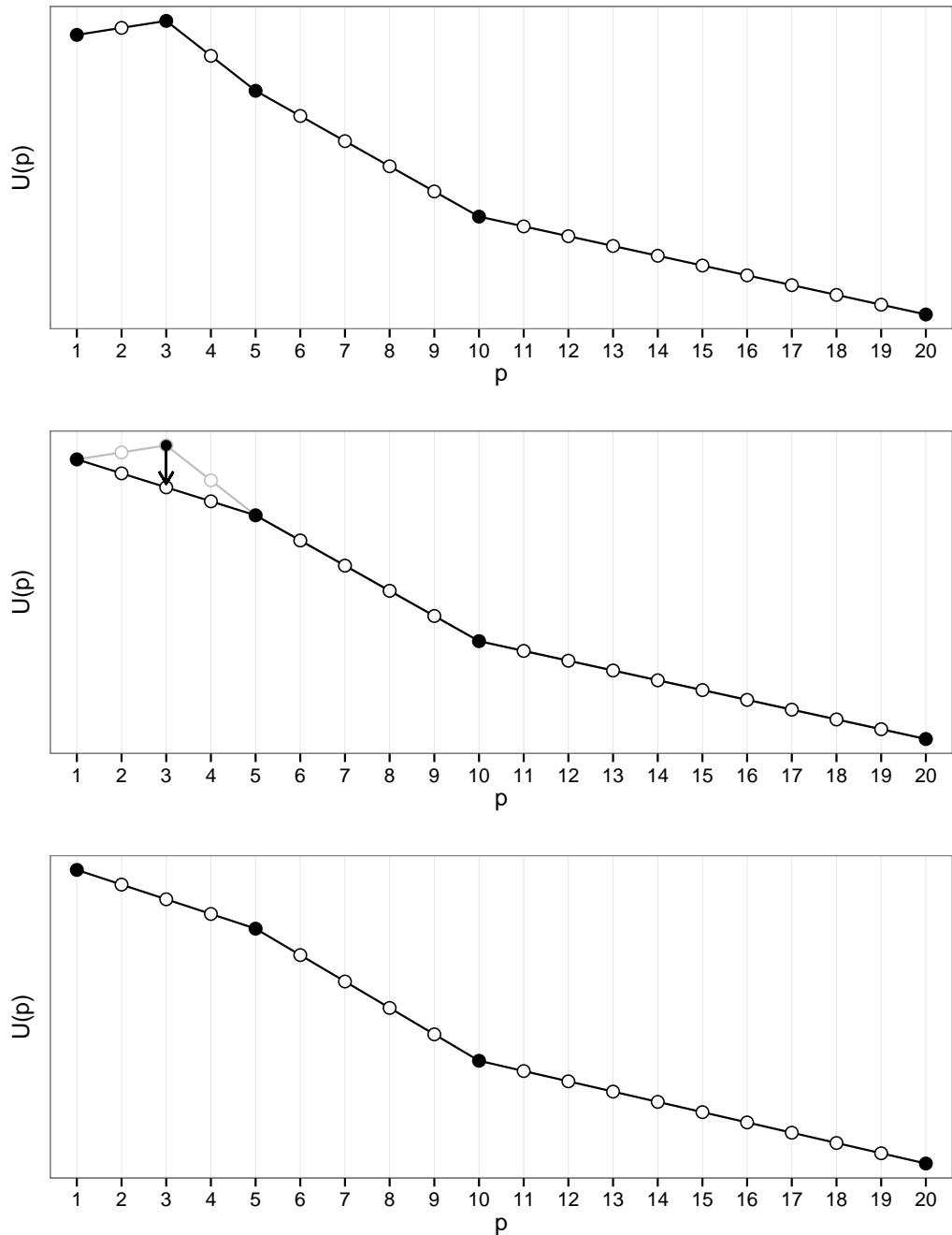


Figure 5.8: The utility function given by model m' (top), the proposal of the death of the knot in position 3 (middle), the resulting utility function of model m (bottom). • utility function knot locations

corresponding knot values is thus:

$$p(\boldsymbol{\kappa}, \mathbf{U}_\kappa) \propto \mathbb{P}(K) \prod_{i=1}^K p(U_{\kappa_i}) \quad (5.19)$$

where again, $K = \dim(\boldsymbol{\kappa})$ (the number of knots), and $p(U_{\kappa_i})$ is the value of the $\Gamma(2.5, 2.5/U_{\kappa_i}^0)$ density at point U_{κ_i} with $\mathbf{U}_\kappa = (U_{\kappa_1}, \dots, U_{\kappa_K})$.

5.6.4 Model inference results

We opt to take a large amount of samples from the RJMCMC process since the posterior space we wish to explore is very high in dimension. We initialise the chain with a sample from the last iteration of the MCMC procedure for model m_3 from Section 5.5 and take 100,000 posterior samples. Figure 5.9 displays plots of posterior samples from the functions $U(p)$ and $\beta(w)$ - which are Bayesian model averaging estimates. Again, the plots only display a random sample of size 1,500 of the 100,000 to prevent plot rendering problems. Figure 5.10 displays plots showing the prior and posterior distribution of the number of knots, K , and the posterior distribution of the knot positions, $\boldsymbol{\kappa}$.

The most prominent knot positions are 4 and 5, which as can be seen in Figure 5.10 have the highest posterior probabilities of the non-fixed knot positions. Also, the posterior probability $\mathbb{P}(U(6) > U(5)) = 0.81887$, again showing an unexpected decrease in utility going from position 6 to position 5, which we attribute to the same reasons as discussed in Section 5.5.2. The difference here is that we imposed no knowledge of a knot in any of the positions, the knot position was chosen purely by the data. Thus, in some ways this suggests even stronger evidence of teams adapting their behaviour based on their league situation, in order to avoid position 5 which grants entry to the Europa League.

Figure 5.9 also shows an, at first, counter-intuitive finding of a locally high utility value around the mid-table positions (position 10). We expect that this is due to the fact that when utility decisions become important (later in the season due to $\beta(w)$) it is highly likely that teams are only able to move a few positions at any one time. For example it would certainly be very unlikely that a team would be able to move from position 10 to position 1 by winning a single match. And so we must be careful how we interpret the utility function, as the utility values are only comparable locally. It does then seem that the mid-table position may be rather comfortable for some teams, which makes sense in that these teams are safe from the pitfall of relegation.

In comparison to the inferred utility function of model m_3 , the utility function here

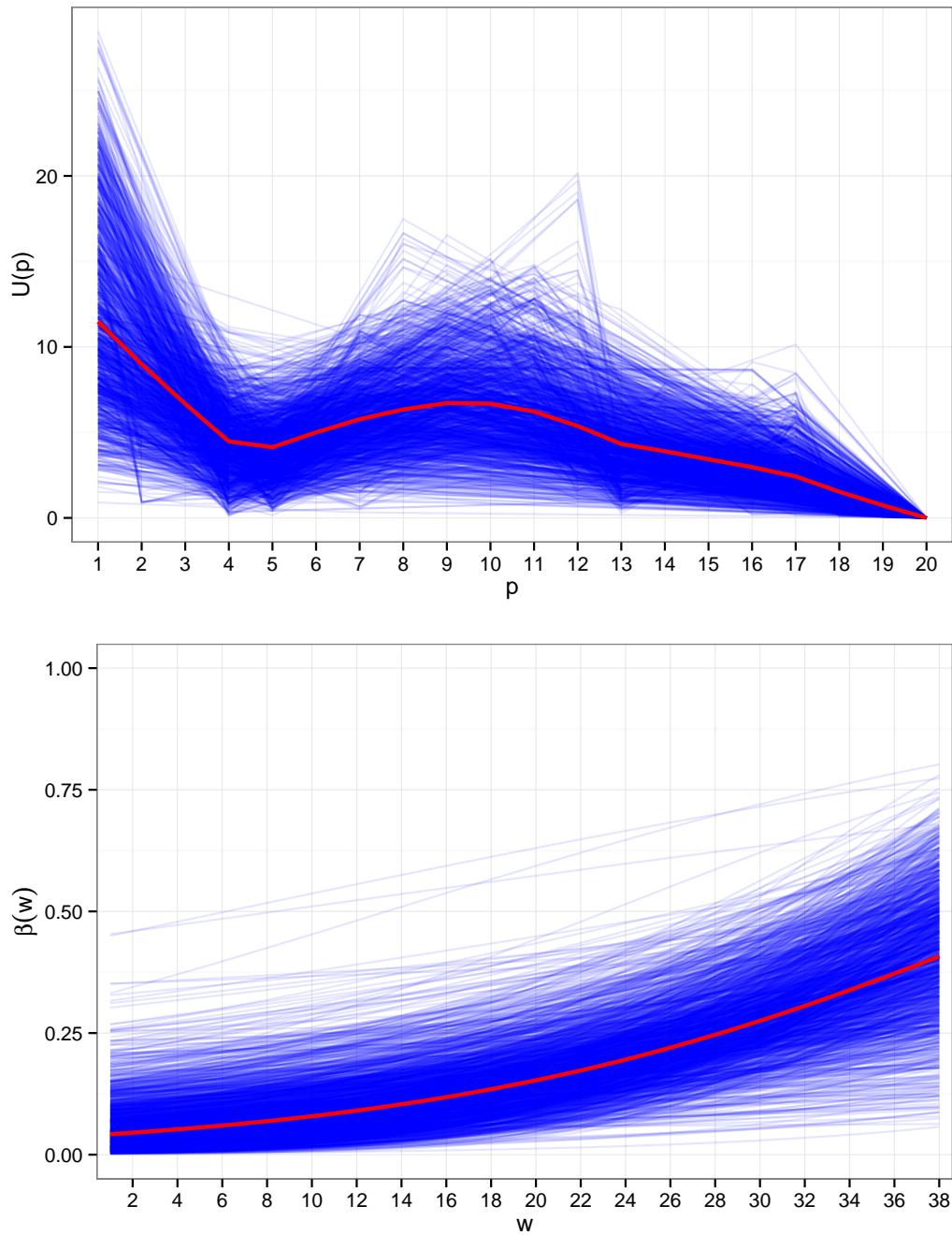


Figure 5.9: A plot of 1,500 samples from the posterior distribution of the utility function $U(p)$ (top) and the function $\beta(w)$ (bottom) using RJMCMC. — the posterior mean

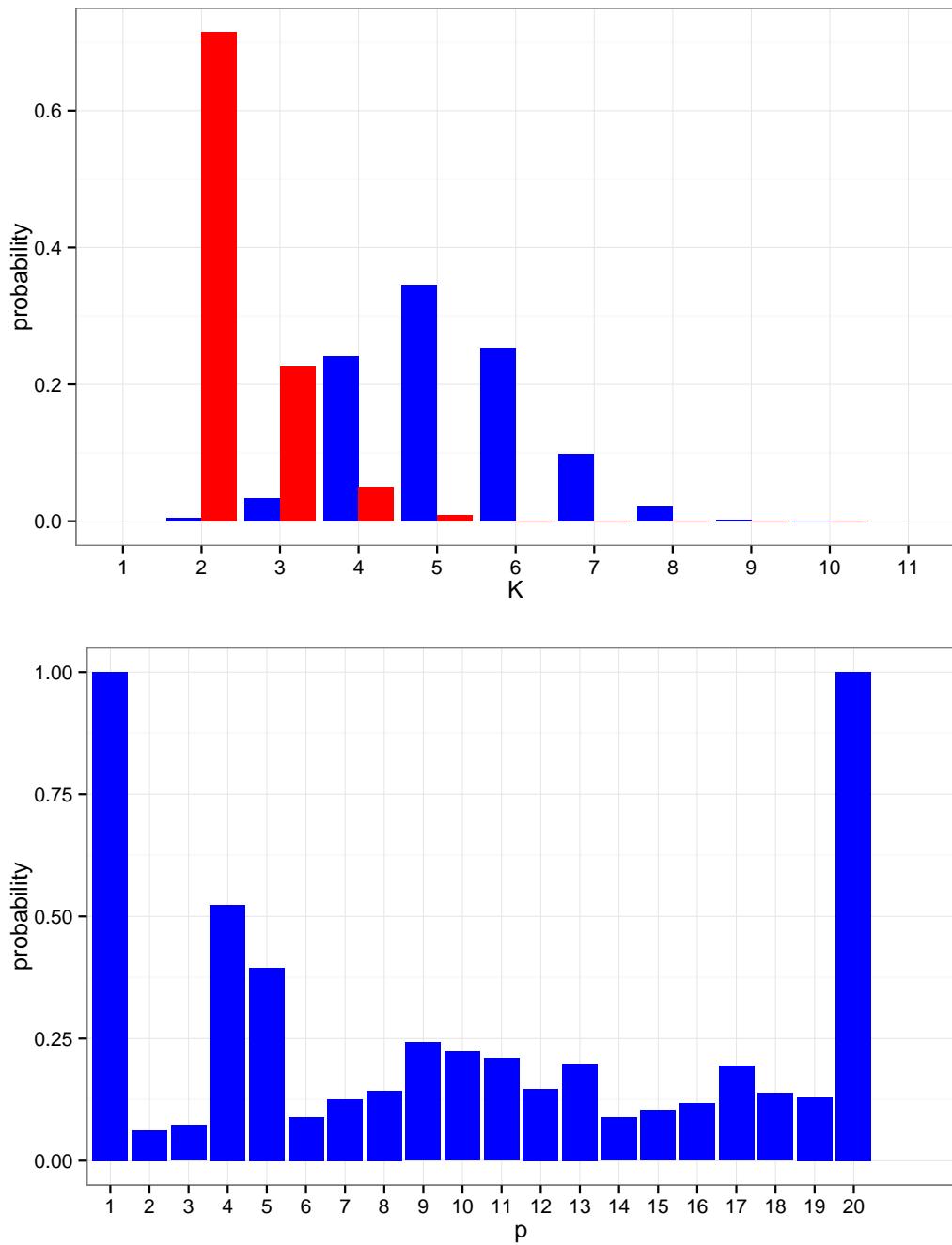


Figure 5.10: A bar plot of the prior and posterior probabilities for the total number of knots (top). A bar plot of the posterior probability of a knot in each position (bottom). ■ the prior probabilities, ■ the posterior probabilities

is very similar in most positions apart from the relegation positions. The RJMCMC procedure did not produce a high posterior probability of a knot at positions 17 and 18 which would be required to separate the utility values of the relegation positions to the non-relegation positions, and thus the utility function shown in Figure 5.9 is reasonably linear around the relegation positions. This does suggest that teams play offensively in order to escape relegation, but not as obviously or dramatically as the utility function of model m_3 (Figure 5.6) would suggest. One possible explanation for this is that the teams which are typically in the relegation zone at the end of the season are of lesser ability and thus low resource, effectively meaning that changing their behaviour has little effect on the rates of scoring - making the effects harder to uncover.

It should be noted however that posterior model probabilities (which we have implicitly used for the Bayesian model averaging estimate of the utility function $U(p)$) are susceptible to the Jeffreys-Lindley paradox as discussed in Chapter 2 Section 2.6.3, whereas methods like model choice based on DIC are not. Thus, we might expect the Bayesian model averaging estimate of the utility function to be less complex than one determined by comparison of models based on DIC. We do however restrict our susceptibility to the Jeffreys-Lindley paradox via the use of relatively informative prior distributions.

5.7 An exemplar match

In a similar fashion to as shown in Chapter 3 Section 3.3, we show the rates of scoring and the resource allocation for the competing teams throughout an EPL match in Figure 5.11. Again, the rates of scoring and the resource allocations are based on posterior mean estimates. The match saw Manchester City play at home to Queens Park Rangers in the last match of the season on 13th May 2012. What was particularly interesting about this match, and was not mentioned in any detail in Chapter 3 Section 3.3, was that this match saw Manchester City fluctuate between league positions 1 and 2, and at the other end of the league, Queens Park Rangers fluctuate between positions 16, 17, and 18 (a relegation position). Before the 10 final matches, the league table was as is shown in Table 5.5.

We list in detail the main events (marked with vertical grey dashed and solid lines in Figure 5.11) which affected the resource allocation (behaviour) of the teams and thus the rates of scoring:

1. Minute 9: Wolverhampton Wanderers score against Wigan Athletic, meaning if Queens Park Rangers are able to win their match, they could move up to position 15.

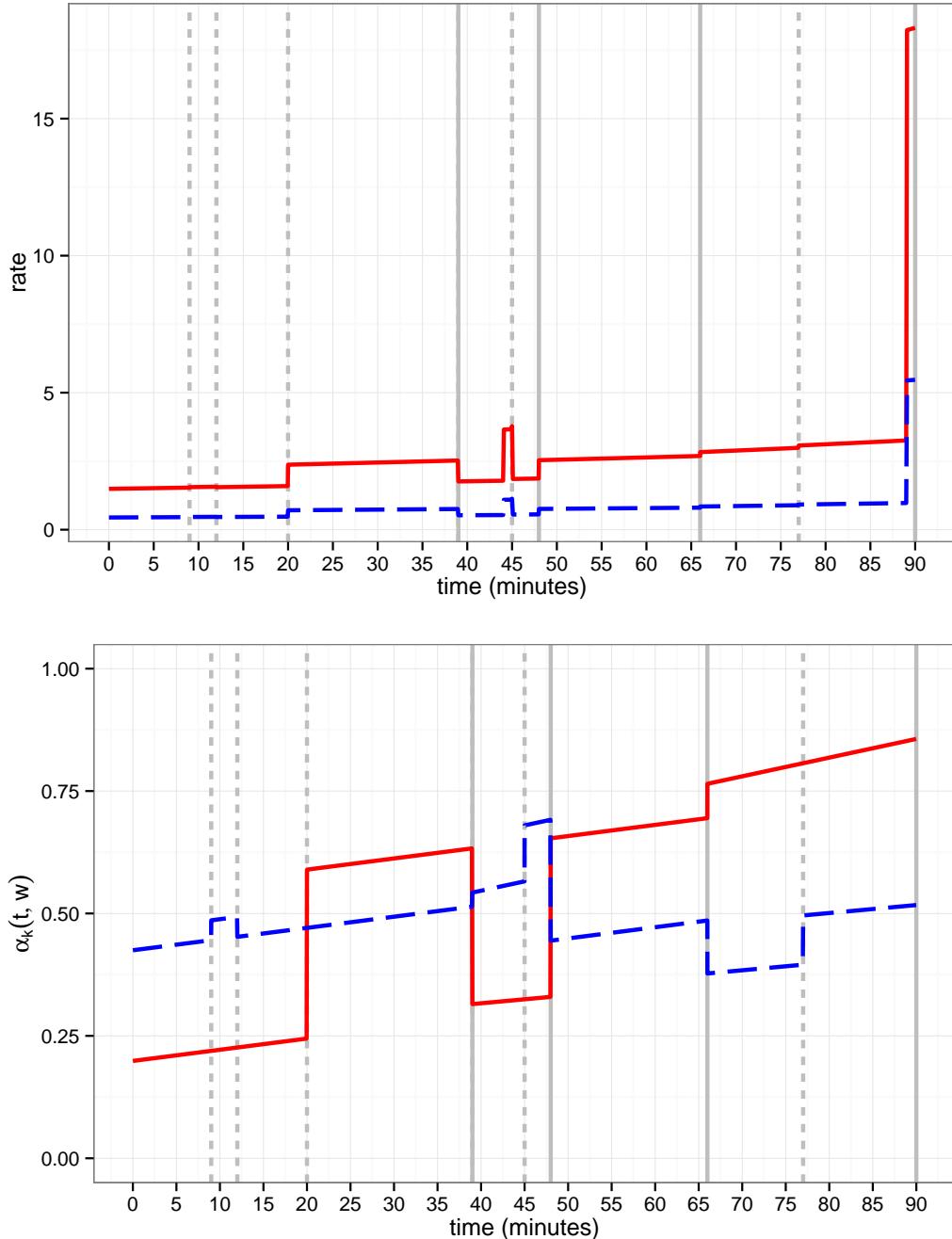


Figure 5.11: A plot of the rates of scoring ($\lambda_m(t)$ and $\mu_m(t)$) (top) and the resource allocation $\alpha_k(t, w)$ (bottom) for match m where Manchester City played at home to Queens Park Rangers. — denotes Manchester City, -·- denotes Queens Park Rangers, — goals scored in this match, --- goals scored in other concurrent matches

p	team	pld	w	d	l	gf	ga	gd	pts
1	Manchester City	37	27	5	5	90	27	+63	86
2	Manchester United	37	27	5	5	88	33	+55	86
3	Arsenal	37	20	7	10	71	47	+24	67
4	Tottenham Hotspur	37	19	9	9	64	41	+23	66
5	Newcastle United	37	19	8	10	55	48	+7	65
6	Chelsea	37	17	10	10	63	45	+18	61
7	Everton	37	14	11	12	47	39	+8	53
8	Liverpool	37	14	10	13	47	39	+8	52
9	Fulham	37	14	10	13	48	49	-1	52
10	West Bromwich Albion	37	13	8	16	43	49	-6	47
11	Sunderland	37	11	12	14	45	45	0	45
12	Swansea City	37	11	11	15	43	51	-8	44
13	Norwich City	37	11	11	15	50	66	-16	44
14	Stoke City	37	11	11	15	34	51	-17	44
15	Wigan Athletic	37	10	10	17	39	60	-21	40
16	Aston Villa	37	7	17	13	37	51	-14	38
17	Queens Park Rangers	37	10	7	20	41	63	-22	37
18	Bolton Wanderers	37	10	5	22	44	75	-31	35
19	Blackburn Rovers	37	8	7	22	47	76	-29	31
20	Wolverhampton Wanderers	37	5	10	22	38	79	-41	25

Table 5.5: The league table before the last match of the EPL 2011/2012 season. ‘p’ denotes the league position, ‘pld’ the number of matches played (note this is one less than the week of the season), ‘w’ the number of matches won, ‘d’ the number of matches drawn, ‘l’ the number of matches lost, ‘gf’ the number of goals for (scored), ‘ga’ the number of goals against (conceded), ‘gd’ the goal difference (‘gf’ - ‘ga’), and ‘pts’ the league points

2. Minute 12: Wigan Athletic score against Wolverhampton Wanderers, returning Queens Park Rangers into the same situation they were in at the beginning of the match.
3. Minute 20: Manchester United score against Sunderland, Manchester City thus move into position 2, behind Manchester United.
4. Minute 39: Manchester City score, and thus regain the top position.
5. Minute 45: Bolton score, moving Queens Park Rangers into the relegation zone (position 18).
6. Minute 48: Queens Park Rangers react with a shock goal, moving themselves into position 17, and also Manchester City back into position 2.
7. Minute 66: Queens Park Rangers score again, moving themselves into a rather safe position 16 while Manchester City remain in position 2. The league title appears to have slipped away.
8. Minute 90: In one of the most dramatic season ends, Manchester City score two goals in injury time to win 3-2 and claim league victory.

We note two main points here, firstly, the model suggests that the rates of scoring can be greatly altered by Manchester City's behavioural changes. This is since they are a strong team with a large amount of resource. Conversely, Queens Park Rangers have little resource, and so their behavioural changes have little effect on the rates of scoring of the match. Secondly, we can see how the teams alter their behaviour as goals in concurrent matches change their league situation. For example at minute 20 when Manchester United score against Sunderland, Manchester City move down to position 2 and immediately play more offensively in an effort to score and regain position 1. Similarly, the most offensive behaviour from Queens Park Rangers is seen between minutes 45 and 48, where they are in position 18 (a relegation position).

Furthermore, at minute 20 when Manchester United score against Sunderland, the combined instantaneous rate of scoring ($\lambda_m(t) + \mu_m(t)$) in the Manchester City - Queens Park Rangers match increases by almost 50% from 2.0662 to 3.0804, a clearly significant amount which has dramatic implications for in-play betting markets such as 'total number of goals in the match' or 'time of next goal'.

5.8 Concluding remarks

In this chapter we have presented a substantial modification to the model first presented in Chapter 3 Section 3.2.3 by allowing the resource allocation of teams to depend on their league situation. We were able to specify a model which could

capture how league considerations become more important later in a season, and how teams play offensively when scoring means moving to a ‘better’ league position, or play defensively when conceding means moving to a ‘worse’ league position. We were also thus able to place an effective value of ‘utility’ on each of the league positions.

Two different methods of model inference and model choice were proposed:

1. Comparison of four hand-selected models via DIC
2. Bayesian model choice using RJMCMC to consider the entire space of models

The RJMCMC method allowed posterior model probabilities over the whole possible model space, over which we could average in order to infer a Bayesian model averaging estimate of the utility function $U(p)$. The estimate of the utility function was in fact quite similar to our best (according to DIC) hand-selected (from 262,144 choices) model.

Posterior estimates of the utility function $U(p)$ suggested a locally minimal value at position 5, a position granting Europa League qualification which a priori, we believed would be valuable. This finding is actually consistent with many on-line articles (Potton (2015); Lea (2015); Hytner (2015)) regarding the Europa League and how teams may wish to avoid the extra matches and travelling it brings - in order to focus on the much more prestigious/profitable EPL.

Finally, we were able to display (via an example) the magnitude of the effect league considerations had on the actual model rates of scoring. Using our models, we estimated that at one point, a single goal in a match between Manchester United and Sunderland affected the combined scoring rate of a separate match between Manchester City and Queens Park Rangers by almost 50%. We thus conclude that these effects may have considerable implications for in-play betting markets for select matches.

Chapter 6

Conclusion

6.1 Results and discussion

At the beginning of this thesis, we described four main aims regarding the development of modelling approaches in association football. To reiterate in short, they are:

1. The creation of models which are widely applicable, have improved predictive power, and are more parsimonious than models currently in literature
2. Inference of models in a Bayesian framework so that prior knowledge can be exploited and parameter uncertainty accounted for
3. Methods of inference which are quick to compute
4. The creation of models which can capture behavioural aspects of teams

Throughout this thesis we have shown examples of numerous Bayesian methods of inference, for example MH, particle filtering methods, and RJMCMC. We have suggested benefit in these methods which allow the addition of extra information into the model inference process via prior distributions, perform parameter updating very quickly, and also naturally inform model choice/model averaging decisions. Not only this but the Bayesian framework easily accounts for uncertainty in parameter estimation through the posterior distribution. We strongly feel that the Bayesian framework should be strongly considered when performing inference of sports related models - and hope that this thesis has shown some of the many benefits.

In Chapter 3 of the thesis we presented a novel non-homogeneous Poisson process model which we then presented several modifications to in Chapters 4 and 5 (some more potential modifications are discussed below). The model is more parsimonious than similar models in literature by defining each team's ability by a single team-specific parameter (which we called the 'resource'), as opposed to a team-specific attack and defense parameter (Maher (1982); Dixon and Coles (1997); Dixon and

Robinson (1998)). This simpler model was shown to outperform other models in the literature based on a number of tests, largely concerning one-week-ahead forecasting ability. Furthermore, since the model can be used to simulate match goal times, it can be used to make predictions around any event which is defined by goal times - of which the vast majority of betting markets comprise.

We displayed how particle filtering methods could be employed to update posterior distributions extremely quickly and efficiently, and also how they could cope with a mixture of static and dynamic model parameters. Similarly to Owen (2011), we proposed a model whereby the team resource parameters were allowed to vary dynamically throughout a season, in contrast however, we found no compelling evidence that this addition to the model was beneficial for our (different to that of Owen (2011)) test data. The superior computational speed of particle filtering methods however allows users to practically update posterior distributions while matches are being played. Clearly useful if one wished to have the most up-to-date parameter estimates during matches for betting or other purposes.

Finally, in terms of the results of the thesis, we presented a final extension to the model which included modifying the team resource allocation parameter ($\alpha_k(t)$) in order to account for current league position and obtainable league positions. This is indeed a new and novel behavioural aspect which previous models in literature have not been able to fully capture. Model inference then suggested that teams may indeed play in order to avoid ending the season in position 5 (which grants qualification to the Europa League) by noting a local minimum in the estimated utility function at this position. That is, there is some evidence to suggest that controversially, teams in the EPL would rather finish the season in position 6 than 5. Furthermore, we showed (via an example of a match between Manchester City and Queens Park Rangers in the final week of the season) how the model captures the change in behaviour of teams as they react to news from other concurrent matches which affects their league situation. The change in behaviour was shown to have a large real effect on the rates of scoring, in one example increasing the scoring rate by almost 50%, and thus has clear implications for certain betting markets.

Statistical models all make assumptions, either explicit or implicit, and indeed ours is no exception to this. We note some of our modelling assumptions here:

1. Red cards have no effect
2. The perception of utility is common across all teams
3. With regards to the match situation (and not the league situation), teams only consider whether they are winning, drawing, or losing
4. With regards to the league, teams consider positions which can be reached following the scoring of at most two more goals for either team

The first point portrays the main limitation of the model, that it should not be used for prediction during in-play matches for which a player has been dismissed via a red card, and that the model likelihood does not explicitly account for matches which have contained red cards. Whilst relatively rare events, and typically occurring at later times in matches, red cards *surely* effect the scoring rate during matches. Somewhat surprisingly however, Volf (2009) did consider red cards as a covariate in a Poisson process model for goal times and found it to be non-significant (a confidence interval for the coefficient contained the value 0).

Having the scoring rates of the competing teams depend on red cards would add another dimension of complexity to the model. Simulation of matches would require not only the simulation of goal times, but also cards. With our aim to create parsimonious models, for which inference and prediction was computationally efficient, we chose to not add red card information - in line with many authors in the literature concerning the modelling of association football outcomes (for example Dixon and Coles (1997); Dixon and Robinson (1998); McHale and Scarf (2011)).

The remaining three points concern our formulation of the team resource allocation parameter ($\alpha_k(t)$) for which we discuss potential modifications for below. With one of the aims of this thesis to create parsimonious models which are still capable of capturing a large variety of effects, we always opted for the simplest sensible formulation possible. We thus assumed that all teams react the same way to league or match position, as opposed to specifying up to 20 team-specific behavioural parameters. At first we also only consider whether a team was losing, drawing, or winning during a match, which is intuitive and also largely agrees with the modelling of Dixon and Robinson (1998) who additionally considered different formulations for when the match score was 0-1 or 1-0. However with the addition of the league utility considerations into the model, the model considers all score lines attainable through the scoring of up to two more goals for either team, as well as the winning, drawing, or losing state.

6.2 Future work

Evolution of statistical models in sport often comes as experts in the sport suggest effects which statistical models should be able to capture. In this thesis, we, for example, hypothesised that teams should change their behaviour based on the league situation, and were able to specify a model formulation which could capture said effects. There are many other effects which an expert would suggest a statistical model for association football should be able to capture. This is both a blessing and a curse. On one hand there are always ideas for how to specify your next model, on the other, whatever model you can think of will have something that it doesn't

account for or is misspecified or misrepresented. This ties in quite nicely with a well-known quote from Box and Draper (1987), ‘all models are wrong, but some are useful’. This is why research on statistical models for sport will evolve as long as sport itself does.

To extend the utility based model presented in Chapter 5 one might look at ways of adding an idea of ‘match importance’ (Scarf and Shi (2008)) into the function $\beta(w)$ so that teams play in accordance to their league position when the match is considered important with regards to the league. Unfortunately, the raw ideas of Scarf and Shi (2008) would not be usable, some other form of ‘match importance’ would need to be formulated.

More generally, there are many factors which the function resource allocation function $\alpha_k(\cdot)$ could account for, for example any linear or non-linear interaction of in-play data such as number of corners kicks, number of throw-ins, number of shots, number of passes, or number of yellow/red cards. One might also expect that different teams react differently depending on the match state, and also whether or not they are the home or away team. For example we would expect a strong team at home to react differently to weak team away when they are in the losing state. Many sporting experts will have (likely differing) opinions on how these data imply changes to the rates of scoring/the team behaviour in a match - and as these data become more readily available it is likely that statistical models will appear in literature which account for them all.

More complex models may also require more complex MCMC algorithms in order to perform inference efficiently. Hamiltonian MCMC (see for example Neal (2011)) is one such algorithm, and it is becoming more popular due to recent developments in the program Stan (Stan Development Team (2015)) which implements Hamiltonian MCMC while maintaining a very similar user interface to WinBUGS. In addition, model comparison will always be a difficult topic, and information criteria other than DIC could also be considered, for example the Widely Applicable Information Criterion (WAIC) (see Vehtari et al. (2015)).

Somewhat unfortunately, historical in-play betting odds data are not currently easily accessible. This is where one might expect betting strategies informed by statistical models to be most profitable, due to the continuously changing odds as the match unfolds and particular events (namely goals) occur. In particular we would expect the model to be profitable if the bookmaker odds were manually driven. Our non-homogeneous Poisson process model would naturally be able to provide predicted probabilities for events of matches as they are in-play, although one would need to be careful in the cases where matches contain ‘red-card’ events, as mentioned above. The first research to appear in the literature showing the potential profits of in-play, statistical model informed, betting strategies could be truly ground breaking, and

could also be worrying news for bookmakers.

Finally, one might also consider the use of a similar non-homogeneous Poisson process model for the occurrences of points in other sports. For example Rugby Union where teams divide their resource between attempting to score a try or a penalty kick/drop goal. In a similar fashion, American football teams may divide their resource between attempting to score field goals or touchdowns, which suggests an potential extension to the point-process model of Baker and McHale (2013). Ice Hockey is another popular sport where such a model could be effective - in particular since Ice Hockey teams are known to behave particularly offensively when losing in the final minutes of a match, and often deploy their goalkeeper as an extra attacking player in an attempt to score (this tactic however most often leads to conceding a goal).

Bibliography

- Aitkin, M. (1991), ‘Posterior Bayes factors’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 111–142.
- Baker, R. D. and McHale, I. G. (2013), ‘Forecasting exact scores in national football league games’, *International Journal of Forecasting* **29**(1), 122–130.
- Barford, V. and Judah, S. (2013), ‘The street with 18 betting shops’. [Online; accessed 8-September-2015].
www.bbc.co.uk/news/magazine-22934305
- Bayes, M. and Price, M. (1763), ‘An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS’, *Philosophical Transactions (1683-1775)* pp. 370–418.
- Bedard, M. (2008), ‘Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234’, *Stochastic Processes and their Applications* **118**(12), 2198–2222.
- Berger, J. O. and Bernardo, J. M. (1992), ‘On the development of reference priors’, *Bayesian statistics* **4**(4), 35–60.
- Berger, J. O., Bernardo, J. M. and Sun, D. (2009), ‘The formal definition of reference priors’, *The Annals of Statistics* pp. 905–938.
- Bernardo, J. M. (1979), ‘Reference posterior distributions for Bayesian inference’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 113–147.
- Boers, Y. (1999), On the number of samples to be drawn in particle filtering, in ‘Target Tracking: Algorithms and Applications (Ref. No. 1999/090, 1999/215), IEE Colloquium on’, IET, pp. 5–1.
- Box, G. E. and Draper, N. R. (1987), *Empirical model-building and response surfaces*, Vol. 424, Wiley New York.
- Bradley, R. A. and Terry, M. E. (1952), ‘Rank analysis of incomplete block designs: I. the method of paired comparisons’, *Biometrika* **39**(3/4), 324–345.

BIBLIOGRAPHY

- Brooks, S. P. and Gelman, A. (1998), ‘General methods for monitoring convergence of iterative simulations’, *Journal of computational and graphical statistics* **7**(4), 434–455.
- Burnham, K. P. and Anderson, D. R. (2004), ‘Multimodel inference understanding AIC and BIC in model selection’, *Sociological methods & research* **33**(2), 261–304.
- Cappé, O., Godsill, S. J. and Moulines, E. (2007), ‘An overview of existing methods and recent advances in sequential Monte Carlo’, *Proceedings of the IEEE* **95**(5), 899–924.
- Carpenter, J., Clifford, P. and Fearnhead, P. (1999), ‘Improved particle filter for nonlinear problems’, *IEE Proceedings-Radar, Sonar and Navigation* **146**(1), 2–7.
- Casella, G. and Berger, R. L. (1987), ‘Reconciling Bayesian and Frequentist evidence in the one-sided testing problem’, *Journal of the American Statistical Association* **82**(397), 106–111.
- Cattelan, M., Varin, C. and Firth, D. (2013), ‘Dynamic Bradley-Terry modelling of sports tournaments’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **62**(1), 135–150.
- Chib, S. and Jeliazkov, I. (2001), ‘Marginal likelihood from the Metropolis-Hastings output’, *Journal of the American Statistical Association* **96**(453), 270–281.
- Chopin, N. (2002), ‘A sequential particle filter method for static models’, *Biometrika* **89**(3), 539–552.
- Cinlar, E. (2013), *Introduction to stochastic processes*, Courier Dover Publications.
- Cleveland, W. S. (1979), ‘Robust locally weighted regression and smoothing scatterplots’, *Journal of the American statistical association* **74**(368), 829–836.
- Colquhoun, D. (2014), ‘An investigation of the false discovery rate and the misinterpretation of p-values’, *Royal Society Open Science* **1**(3).
- Constantinou, A. C., Fenton, N. E. and Neil, M. (2012), ‘pi-football: A Bayesian network model for forecasting association football match outcomes’, *Knowledge-Based Systems* **36**, 322–339.
- Cox, D. R. and Lewis, P. A. (1966), ‘The statistical analysis of series of events’, *Monographs on Applied Probability and Statistics, London: Chapman and Hall* **1**.
- Dagum, L. and Menon, R. (1998), ‘OpenMP: An industry standard API for shared-memory programming’, *Computational Science & Engineering, IEEE* **5**(1), 46–55.
- Dawid, A. P., Lauritzen, S. and Parry, M. (2012), ‘Proper local scoring rules on discrete sample spaces’, *The Annals of Statistics* **40**(1), 593–608.

BIBLIOGRAPHY

- Del Moral, P., Doucet, A. and Jasra, A. (2006), ‘Sequential Monte Carlo samplers’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(3), 411–436.
- 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.
- Dixon, M. and Robinson, M. (1998), ‘A birth process model for association football matches’, *Journal of the Royal Statistical Society: Series D (The Statistician)* **47**(3), 523–538.
- Douc, R. and Cappé, O. (2005), Comparison of resampling schemes for particle filtering, in ‘Image and Signal Processing and Analysis, 2005. ISPA 2005. Proceedings of the 4th International Symposium on’, IEEE, pp. 64–69.
- Draper, D. (1995), ‘Assessment and propagation of model uncertainty’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 45–97.
- Einicke, G. A. (2012), ‘Smoothing, filtering and prediction: Estimating the past, present and future’, *New York: InTech*.
- Fahrmeir, L. and Tutz, G. (1994), ‘Dynamic stochastic models for time-dependent ordered paired comparison systems’, *Journal of the American Statistical Association* **89**(428), 1438–1449.
- Ferguson, A. (2003), ‘Sir Alex Ferguson’s best quotes’. [Online; accessed 26-August-2015].
www.theguardian.com/football/2013/may/08/sir-alex-ferguson-best-quotes
- Gelman, A. (2013), ‘Two simple examples for understanding posterior p-values whose distributions are far from uniform’, *Electronic Journal of Statistics* **7**, 2595–2602.
- Gelman, A., Roberts, G. and Gilks, W. (1996), ‘Efficient Metropolis jumping rules’, *Bayesian statistics* **5**(599-608), 42.
- Gelman, A. and Rubin, D. B. (1992), ‘Inference from iterative simulation using multiple sequences’, *Statistical science* pp. 457–472.
- Geman, S. and Geman, D. (1984), ‘Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images’, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **6**(6), 721–741.
- Geweke, J. et al. (1991), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Vol. 196, Federal Reserve Bank of Minneapolis, Research Department.

BIBLIOGRAPHY

- Geyer, C. (2011), ‘Introduction to Markov chain Monte Carlo’, *Handbook of Markov Chain Monte Carlo* pp. 3–48.
- Gibson, G. J. and Renshaw, E. (1998), ‘Estimating parameters in stochastic compartmental models using Markov chain methods’, *Mathematical Medicine and Biology* **15**(1), 19–40.
- Gneiting, T. and Raftery, A. E. (2007), ‘Strictly proper scoring rules, prediction, and estimation’, *Journal of the American Statistical Association* **102**(477), 359–378.
- Goddard, J. and Asimakopoulos, I. (2003), Modelling football match results and the efficiency of fixed-odds betting, Technical report, Working Paper, Department of Economics, Swansea University.
- Gordon, N. J., Salmond, D. J. and Smith, A. F. (1993), Novel approach to nonlinear/non-Gaussian Bayesian state estimation, in ‘IEE Proceedings F (Radar and Signal Processing)’, Vol. 140, IET, pp. 107–113.
- Green, P. J. (1995), ‘Reversible jump Markov chain Monte Carlo computation and Bayesian model determination’, *Biometrika* **82**(4), 711–732.
- Gropp, W., Lusk, E., Doss, N. and Skjellum, A. (1996), ‘A high-performance, portable implementation of the MPI message passing interface standard’, *Parallel computing* **22**(6), 789–828.
- Gustafsson, F., Gunnarsson, F., Bergman, N., Forssell, U., Jansson, J., Karlsson, R. and Nordlund, P.-J. (2002), ‘Particle filters for positioning, navigation, and tracking’, *Signal Processing, IEEE Transactions on* **50**(2), 425–437.
- Hastie, D. I. and Green, P. J. (2012), ‘Model choice using reversible jump Markov chain Monte Carlo’, *Statistica Neerlandica* **66**(3), 309–338.
- Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2009), *The elements of statistical learning*, second edn, Springer.
- Hastings, W. K. (1970), ‘Monte Carlo sampling methods using Markov chains and their applications’, *Biometrika* **57**(1), 97–109.
- Hitchcock, D. B. (2003), ‘A history of the Metropolis-Hastings algorithm’, *The American Statistician* **57**(4).
- Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999), ‘Bayesian model averaging: A tutorial’, *Statistical science* pp. 382–401.
- Hosmer Jr, D. W., Lemeshow, S. and Sturdivant, R. X. (2013), *Applied logistic regression*, Vol. 398, John Wiley & Sons.
- Hytner, D. (2015), ‘Tottenham’s Mauricio Pochettino: avoiding Europa League could help us’. [Online; accessed 28-August-2015].

BIBLIOGRAPHY

- www.theguardian.com/football/2015/apr/23/tottenham-mauricio-pochettino-europa-league
- Ioannidis, J. P. (2005), ‘Why most published research findings are false’, *Chance* **18**(4), 40–47.
- Jeffreys, H. (1939), *Theory of Probability*, Oxford University Press.
- Jeffreys, H. (1946), An invariant form for the prior probability in estimation problems, in ‘Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences’, Vol. 186, The Royal Society, pp. 453–461.
- Johnson, S. G. (2010), *The NLOpt nonlinear-optimization package*.
<http://ab-initio.mit.edu/nlopt>
- Kalman, R. E. (1960), ‘A new approach to linear filtering and prediction problems’, *Journal of Fluids Engineering* **82**(1), 35–45.
- Karlis, D. and 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.
- Kass, R. E., Carlin, B. P., Gelman, A. and Neal, R. M. (1998), ‘Markov chain Monte Carlo in practice: A roundtable discussion’, *The American Statistician* **52**(2), 93–100.
- Kass, R. E. and Raftery, A. E. (1995), ‘Bayes factors’, *Journal of the American Statistical Association* **90**(430), 773–795.
- Kelly Jr, J. L. (1956), ‘A new interpretation of information rate’, *Information Theory, IRE Transactions on* **2**(3), 185–189.
- Knorr-Held, L. (2000), ‘Dynamic rating of sports teams’, *Journal of the Royal Statistical Society: Series D (The Statistician)* **49**(2), 261–276.
- Koopman, S. J. and Lit, R. (2015), ‘A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league’, *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **178**(1), 167–186.
- Kullback, S. and Leibler, R. A. (1951), ‘On information and sufficiency’, *The annals of mathematical statistics* pp. 79–86.
- Lea, G. (2015), ‘The battle to avoid the europa league’. [Online; accessed 28-August-2015].
www.goal.com/en-gb/news/2915/europa-league/2015/04/24/11060642/the-battle-to-avoid-the-europa-league
- Lewis, M. (2004), *Moneyball: The art of winning an unfair game*, WW Norton & Company.

BIBLIOGRAPHY

- Lewis, P. A. and Shedler, G. S. (1979), ‘Simulation of nonhomogeneous Poisson processes by thinning’, *Naval Research Logistics Quarterly* **26**(3), 403–413.
- Lindley, D. V. (1957), ‘A statistical paradox’, *Biometrika* pp. 187–192.
- Link, W. A. and Eaton, M. J. (2012), ‘On thinning of chains in MCMC’, *Methods in Ecology and Evolution* **3**(1), 112–115.
- Liu, J. and West, M. (2001), Combined parameter and state estimation in simulation-based filtering, in ‘Sequential Monte Carlo methods in practice’, Springer, pp. 197–223.
- Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000), ‘WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility’, *Statistics and Computing* **10**(4), 325–337.
<http://dx.doi.org/10.1023/A:1008929526011>
- Madigan, D. and Raftery, A. E. (1994), ‘Model selection and accounting for model uncertainty in graphical models using Occam’s window’, *Journal of the American Statistical Association* **89**(428), 1535–1546.
- Maher, M. J. (1982), ‘Modelling association football scores’, *Statistica Neerlandica* **36**(3), 109–118.
- McHale, I. and Scarf, P. (2011), ‘Modelling the dependence of goals scored by opposing teams in international soccer matches’, *Statistical Modelling* **11**(3), 219–236.
- Meng, X. (1994), ‘Posterior predictive p-values’, *The Annals of Statistics* pp. 1142–1160.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), ‘Equation of state calculations by fast computing machines’, *The journal of chemical physics* **21**, 1087.
- Montemerlo, D., Thrun, S. and Whittaker, W. (2002), Conditional particle filters for simultaneous mobile robot localization and people-tracking, in ‘Robotics and Automation, 2002. Proceedings. ICRA’02. IEEE International Conference on’, Vol. 1, IEEE, pp. 695–701.
- Neal, R. (2008), ‘The harmonic mean of the likelihood: Worst Monte Carlo method ever’. [Online; accessed 2-April-2015].
www.radfordneal.wordpress.com/2008/08/17/the-harmonic-mean-of-the-likelihood-worst-monte-carlo-method-ever
- Neal, R. M. (1993), ‘Probabilistic inference using Markov chain Monte Carlo methods’.

BIBLIOGRAPHY

- Neal, R. M. (2001), ‘Annealed importance sampling’, *Statistics and Computing* **11**(2), 125–139.
- Neal, R. M. (2011), ‘MCMC using Hamiltonian dynamics’, *Handbook of Markov Chain Monte Carlo* **2**.
- Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.
- Nummiaro, K., Koller-Meier, E. and Van Gool, L. (2003), ‘An adaptive color-based particle filter’, *Image and vision computing* **21**(1), 99–110.
- Owen, A. (2011), ‘Dynamic Bayesian forecasting models of football match outcomes with estimation of the evolution variance parameter’, *IMA Journal of Management Mathematics* **22**(2), 99–113.
- Pitt, M. K. and Shephard, N. (1999), ‘Filtering via simulation: Auxiliary particle filters’, *Journal of the American statistical association* **94**(446), 590–599.
- Potton, L. (2015), ‘The race to avoid Europa League qualification starts now’. [Online; accessed 28-August-2015].
www.90min.com/posts/2080889-the-race-to-avoid-europa-league-qualification-starts-now
- Punská, O., Doucet, C. A. A., Fitzgerald, W., Andrieu, C. and Doucet, A. (1999), ‘Bayesian segmentation of piecewise constant autoregressive processes using MCMC methods’.
- R Core Team (2015), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
<http://www.R-project.org>
- Rekleitis, I. M. (2004), ‘A particle filter tutorial for mobile robot localization’, *Centre for Intelligent Machines, McGill University, Tech. Rep. TR-CIM-04-02* .
- Robert, C. P. (2014), ‘On the Jeffreys-Lindley paradox’, *Philosophy of Science* **81**(2), 216–232.
- Roberts, G. O. (1996), ‘Markov chain concepts related to sampling algorithms’, *Markov chain Monte Carlo in practice* **57**.
- Roberts, G. O., Gelman, A., Gilks, W. R. et al. (1997), ‘Weak convergence and optimal scaling of random walk Metropolis algorithms’, *The annals of applied probability* **7**(1), 110–120.
- Roberts, G. O. and Polson, N. G. (1994), ‘On the geometric convergence of the Gibbs sampler’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 377–384.

BIBLIOGRAPHY

- Sanjeev Arulampalam, M., Maskell, S., Gordon, N. and Clapp, T. (2002), ‘A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking’, *Signal Processing, IEEE Transactions on* **50**(2), 174–188.
- Scarf, P. A. and Shi, X. (2008), ‘The importance of a match in a tournament’, *Computers & Operations Research* **35**(7), 2406–2418.
- Senn, S. (2015), ‘The pathetic p-value’. [Online; accessed 10-July-2015].
www.errorstatistics.com/2015/03/16/stephen-senn-the-pathetic-p-value-guest-post
- Spanos, A. (2013), ‘Who should be afraid of the Jeffreys-Lindley paradox?’, *Philosophy of Science* **80**(1), 73–93.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and Van Der Linde, A. (2002), ‘Bayesian measures of model complexity and fit’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(4), 583–639.
- Stan Development Team (2015), ‘Stan: A C++ library for probability and sampling, version 2.7.0’.
<http://mc-stan.org/>
- Streftaris, G. and Gibson, G. J. (2004), ‘Bayesian inference for stochastic epidemics in closed populations’, *Statistical Modelling* **4**(1), 63–75.
- Tibshirani, R. (1996), ‘Regression shrinkage and selection via the lasso’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 267–288.
- Tierney, L. (1996), ‘Introduction to general state-space Markov chain theory’, *Markov chain Monte Carlo in practice* pp. 59–74.
- Tierney, L. (1998), ‘A note on Metropolis-Hastings kernels for general state spaces’, *Annals of Applied Probability* pp. 1–9.
- Trafimow, D. and Marks, M. (2015), ‘Editorial’, *Basic and Applied Social Psychology* **37**(1), 1–2.
<http://dx.doi.org/10.1080/01973533.2015.1012991>
- UEFA (2013a), ‘Payments to 2012/13 Champions League clubs’. [Online; accessed 26-August-2015].
www.uefa.com/MultimediaFiles/Download/uefaorg/Finance/01/97/52/97/1975297_DOWNLOAD.pdf
- UEFA (2013b), ‘Payments to 2012/13 Europa League clubs’. [Online; accessed 26-August-2015].
www.uefa.com/MultimediaFiles/Download/uefaorg/Finance/01/97/53/15/1975315_DOWNLOAD.pdf

BIBLIOGRAPHY

- Vehtari, A., Gelman, A. and Gabry, J. (2015), ‘Efficient implementation of leave-one-out cross-validation and WAIC for evaluating fitted Bayesian models’, *arXiv preprint arXiv:1507.04544* .
- Volf, P. (2009), ‘A random point process model for the score in sport matches’, *IMA Journal of Management Mathematics* **20**(2), 121–131.
- West, M. (1993), ‘Approximating posterior distributions by mixture’, *Journal of the Royal Statistical Society: Series B (Methodological)* pp. 409–422.
- Whitrow, C. (2007), ‘Algorithms for optimal allocation of bets on many simultaneous events’, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **56**(5), 607–623.
- Wickham, H. (2009), *ggplot2: elegant graphics for data analysis*, Springer New York.
<http://had.co.nz/ggplot2/book>
- Wikipedia (2014), ‘Lindley’s paradox - Wikipedia, the free encyclopedia’. [Online; accessed 11-December-2014].
http://en.wikipedia.org/wiki/Lindley%27s_paradox