Statistical Methods for Detecting Match-Fixing in Tennis

Oliver Hatfield, MMath (Hons.), MRes

Submitted for the degree of Doctor of Philosophy at Lancaster University.

December 2019

STOR-i
excellence with impact

Abstract

Match-fixing is a key problem facing many sports, undermining the integrity and sporting spectacle of events, ruining players' careers and enabling the criminals behind the fixes to funnel funds into other illicit activities. Although for a long time authorities were reticent to act, more and more sports bodies and betting companies are now taking steps to tackle the issue, though much remains to be done. Tennis in particular has faced past criticism for its approach to combatting match-fixing, culminating in widespread media coverage of a leak of match-fixing related documents in 2016, although the Tennis Integrity Unit has since intensified its efforts to deal with the problem.

In this thesis, we develop new statistical methods for identifying tennis matches in which suspicious betting activity occurs. We also make some advancements on existing sports models to enable us to better analyse tennis matches to detect this corrupt activity. Our work is among the first to use both pre-match and in-play odds data to investigate match-fixing, and to also integrate betting volumes. Our pre-match odds are sampled at several intervals during the pre-match market, allowing for more detailed analysis than other work. Our in-play odds data are recorded during every game break along with live scores so that we can explore how the odds vary as the score progresses. In particular, we look for divergences between market odds and predictions coming both from sports models and from direct predictions of odds based on in-play events. Our methods successfully identify past matches that other external sources have found to contain suspicious betting activity, and are able

to quantify how unusual this activity was in relation to typical betting behaviour. This suggests that our methods, coupled with other sources of evidence, can provide a valuable quantification of suspicious betting activity in future matches.

Acknowledgements

There are many people I would like to thank for their roles in helping me complete my PhD, without whom this thesis would not have been possible. I would first like to thank my supervisors, Professor Jonathan Tawn and Dr. Chris Kirkbride from Lancaster University, and Dr. Tim Paulden, Dr. David Irons and Grace Stirling from ATASS Sports. Their support and knowledge was vital to help me get through this, and I am very grateful for receiving their time. Special mention must go to Professor Tawn in particular for providing such thorough comments on a greatl deal of drafts as my deadlines approached.

I'd like to thank STOR-i Centre for Doctoral Training for providing me with the opportunity to study, and to EPSRC for funding STOR-i. The work environment at STOR-i was particularly important, with the football and quiz teams proving giving shape to the week on (mostly) Tuesdays throughout. I'd also like to thank ATASS Sports for their provision of the odds data required for this project and their funding, as well as for warmly hosting me on my various visits.

My family have my sincere gratitude for their unwavering support throughout. Thanks must also go to my good friend George for his strong listening skills. Finally, acknowledgements must go to Dr. Emma Simpson, who was by my side the whole time and always wanted the best for me.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Oliver Hatfield

Word Count: approximately 73,000 words.

Contents

List of Figures

Chapter 1

Introduction

Match-fixing was once described by the Council of Europe as "one of the most serious threats to contemporary sport, undermining the fundamental values of integrity, fair play and respect for others", (Olfers et al., 2014). In addition, match-fixing poses other societal problems due to its frequent perpetration by criminal gangs, who use it as a "vehicle for... a number of other financial crimes, including money-laundering and tax evasion.", (United Nations Office on Drugs and Crime, 2016). A report by Olfers et al. (2014) details the various efforts of EU bodies to develop a cohesive, co-operative strategy to tackle match-fixing. Several leading bookmakers also united in 2005 to create the European Sports Security Association (ESSA), a betting integrity unit that aims to quickly assimilate information on suspicious betting patterns in an attempt to identify match-fixing behaviour. Individual sports are also taking responsibility for detecting match-fixing in their own domain, and in 2008 the Tennis Integrity Unit was one of the first specialist anti-corruption taskforces for a major sport, according to their website.

Forensic sports analytics is the science of detecting corruption in sport using statistical analysis. Taking its name and ideas from forensic economics, which performs a similar role in financial and industrial settings, it attempts to find evidence of corrupt activity in data. Sporting corruption takes many forms, but throughout this project

we focus only on match-fixing. Some teams fix matches to win, bribing referees or opponents to further their own sporting career. Other teams fix matches to lose, usually to help a third party profit by gambling using this knowledge. This corrupt betting activity can make the betting markets in such fixed matches stand out from clean ones. However, most work in this area is done privately by betting companies or sports corruption watchdogs.

Criticisms have been made of these organisations' effectiveness, and the methods used remain largely secret. The aim of this project is therefore to develop new methods of analysing of betting data in order to see whether suspicious activity that possibly conforms to match-fixing can be identified.

1.1 Match-Fixing in Tennis

This project aims to target match-fixing in tennis particularly. The first reason is because it is a major sport with an ostensibly large match-fixing problem. During a period in 2015, 36 out of 47 alerts for suspicious betting activity by the ESSA came on tennis matches, as reported in ESSA (2015a) and ESSA (2015b). This may or may not be skewed by the precise nature of the ESSA's investigations: nonetheless, it indicates that a clear match-fixing problem has existed in tennis at the highest level. Tennis is also an attractive sport to analyse from a statistical point of view. In team sports, the strength of a team can be difficult to model precisely when one or more key players are absent. By comparison, individual sports include just one player, and hence it can be easier to model their strength.

Pace is gathering in the battle against match-fixing, both in the efforts by sporting bodies and in academic literature. A joint investigation by the BBC and BuzzFeed in January 2016, (Blake and Templon, 2016) was particularly enlightening, with a multitude of documents relating to previous investigations into match-fixing released, albeit heavily redacted. Since then, the Tennis Integrity Unit (TIU) has appeared

keener to publicise information about sanctioned players, with more media releases and a "Currently Suspended" section on their website. However, the fight is not over, and there remains much to be done.

Much of the investigations by bodies such as the TIU is conducted in secret. This is for good reason, as it helps both avoiding making pre-emptive accusations against players, and preventing match-fixers from using knowledge of the TIU's methods to avoid detection. Nonetheless, a reasonable amount of work in detecting match-fixing exists in forensic sports analytics literature, which we can build on to help develop new methods for highlighting suspicious matches.

Statistics alone cannot prove whether a match is fixed or not - more tangible evidence is required, such as money transfers or covert messages. Indeed, Nigel Willerton, head of the TIU, once said "Betting data alone is not sufficient to bring forward a prosecution". However, investigating matches takes time and money, so methods to identify the most suspicious matches can be of great help to investigators deciding how to allocate time and resources.

The aim of this project is therefore to develop new methods for identifying potentially fixed matches for investigators to examine more thoroughly. We mainly do this by following the lead of current academic literature and using sports models to attempt to predict odds. Matches which do not conform to predictions are considered suspicious, and worthy of further investigation. We will be using historical data to test methods, and as such the matches we identify may already have been investigated, and the players involved found guilty or innocent. This is but one reason we must therefore be careful to not to imply guilt at any stage. Nothing can be proven from statistical analysis, and the reputations of the players are at stake. For this reason, all data used in this project has had the names of player removed.

1.2 Chapter Summaries

In Chapter 2, we shall perform a review of current literature in the area. This falls into two parts. Firstly, we shall review a wide range of literature on detecting match-fixing in sport, and secondly we shall review relevant literature in predicting the results of tennis matches. Chapter 3 is an extended proof of the fact that under certain assumptions there is a one-to-one relationship between the probability of a player winning a match and the difference in quality of the two players. Chapter 4 explains in greater detail the Glicko ratings, one of the tennis models discussed in the literature review, relates it to state space models, extends the model to allow for 5-set matches and applies it to tennis data. Chapter 5 is a brief interlude describing the data used in this thesis. The two different sources of data are the exchange data from pre-match and in-play markets provided by ATASS Sports, and the tennis results data from `github.com/JeffSackmann`. Chapter 6 develops a model for pre-match odds under normal betting behaviour and uses it to identify matches that do not conform to this pattern. The model is more sophisticated than other models in the literature in that it does not simply look at the difference in the opening and closing odds of the pre-match market, but looks at various intervals in between and permits greater flexibility when betting volumes are low and the market is not yet formed. Chapter 7 describes a model for the in-play odds in which we estimate player strengths throughout each match to estimate in-play odds and look for anomalies with respect to this model. Chapter 8 builds on this model by instead describing a Gaussian process for the in-play player strengths that is more flexible than Bayesian updating, allowing for better modelling of in-play odds by extension. This thesis concludes with Chapter 9, which summarises the major contributions of the thesis and the main opportunities for further research in the area.

1.3 Main Contributions

The main goal of this thesis is to design methods for identifying suspicious odds movements in matches, such that the matches can be flagged for further investigation by appropriate authorities. There is very little discussion in academic literature of in-play analysis of betting odds for suspicious activity, and so our two methods in Chapters 7 and 8 represent a significant step in this fledgeling research area. The Gaussian process method in Chapter 8 appears particularly promising.

We also look at swings in pre-match markets in Chapter 6, but attempt to advance on current literature by incorporating volume data and looking at interemediate odds data, rather than simply the opening and closing prices.

In order to perform these analyses, we also generated new advances in modelling tennis matches. Chapter 4 describes a new way of accounting for 5-set matches in Glicko ratings. Chapter 3, meanwhile, proves the invertibility of a function for estimating match-win probabilities using the quality difference of two players. The numerical inverse of this function is already used in the literature, but the proof that this inverse exists reassures us of the safety of using the numerical inverse, and presents some interesting mathematical ideas in its own right.

Chapter 2

Literature Review

This literature review is split into two main parts. First, we consider the literature on the detection of match-fixing in different sports. During this literature review, it will become apparent that the ability to forecast tennis matches may be very useful in identifying match-fixing. The second part of the literature review therefore concerns a review of methods for predicting tennis matches. This second part of the literature review is itself further split into two main sections, the first concerning pre-match predictions and the second concerning in-play predictions.

2.1 Match-fixing Literature

The literature on the detection of match-fixing can be divided into two main categories. Some papers examine individual matches to identify if they are fixed or not, while others look at matches in aggregate to identify the prevalence of match-fixing, or else find evidence that the prevalence is non-zero. We focus on the identification of individual matches, but it is still worth considering methods that examine prevalence to see what can be learned.

We begin by considering papers that analyse betting markets to detect unusual activity that may be indicative of match-fixing. The fundamental idea behind the

analysis of betting odds to detect match-fixing is that economic theory suggests that betting odds can be viewed as probabilistic forecasts of an event, provided the betting market is efficient. A betting market is efficient if it assimilates all available information about an event, and hence it is impossible to "beat the odds" through superior forecasting Wolfers and Zitzewitz (2004). To understand why, we consider how and why the odds move over time.

In a traditional bookmaker's, odds will be offered on each competitor. If enough people's perceived probability of a competitor winning suggests that the odds represent a good bet, a substantial imbalance in the amount bet on each competitor may arise. The bookmaker now stands to make substantial losses should this competitor win, and so will shift their odds to encourage betting on the opponent, mitigating the impact of the first player winning and hedging their position. As a result, the odds shift over time to reflect the public information available on the probability of each player winning. The odds therefore represent probabilistic forecasts of the match in themselves.

On betting exchanges the mechanism is different, but the outcome is much the same. On betting exchanges, punters offer (or lay) odds to each other, rather than relying on a bookmaker. As such, there may be a queue of different gamblers offering successively better odds, waiting for another gambler to match their bet. The best odds available are then the best market predictions available. Should new information arise about the probability of the players winning the match such as in injury, these best offers will be matched, removing offered odds from the front of the queue until a new equilibrium is met at a set of odds reflecting the new information. Croxson and Reade (2011) shows that this happens faster than at traditional bookmakers.

When a match is fixed, the fixer knows the outcome of the match with certainty and can bet accordingly. If they only wager small amounts of money, the impact on the market will be negligible. However, the costs of fixing matches through bribery can be high, and large financial gains can be made through wagering large stakes.

If the fixer places bets of sufficiently high value, the odds may shift for the reasons described above.

The goal then is to detect such shifts by predicting what the odds should be in a clean match. Since the odds also represent probabilistic forecasts, other strong forecasts of the match-result should be similar to the odds. Research suggests that betting markets in tennis are quite efficient, apart from some small biases concerning favourites and longshots, discussed by Forrest and McHale (2007) and briefly later in this literature review. Kovalchik (2016) show that bookmaker odds provide better forecasts of tennis matches than any known method. This suggests that the strategy of analysing betting odds may prove fruitful in investigating match-fixing. We shall now discuss several papers that forecast match outcomes, compare these forecasts to betting odds and closely examine matches in which any differences occur.

The works of Forrest and McHale (2015) and Forrest and McHale (2019) provide an intriguing analysis of one of many systems already in use for detecting suspect betting activity in football and tennis. SportRadar is a corporation that provides many services to combat match-fixing and preserve the integrity of sport. This independent review of SportRadar's Fraud Detection System provides a non-technical analysis of the procedures used to automatically flag betting activity that does not conform to expectations, and many of the core ideas are applicable to any sport.

The report of Forrest and McHale (2015) is clear throughout that the purpose of betting analysis is to flag matches for further investigation - it is not an end in itself. Depending on the scale of anomaly, a flag of appropriate severity (red, orange or green, from most severe to least severe) is raised, and the match data are manually inspected to see if an innocent explanation can be found. This could be due to injury, announcement of starting line-up, or another factor that cannot be picked up automatically by a model, but provides a plausible explanation for the anomaly. This increases the rate of false positives, but this is considered acceptable to ensure fewer false negatives.

Pre-match markets can be flagged for any of three reasons. Firstly, SportRadar look for large movements in fractional odds, as these can indicate excessive, potentially corrupt betting on one outcome, and secondly, they examine betting volumes on Betfair to see if more is gambled than expected for a similar match. Finally, the closing pre-match odds are compared to estimated match-win probabilities for the match, derived from an Elo model (Elo (1978), Section 2.2.3). Large discrepancies between the two are flagged as suspicious, as the market should be approximately efficient. Similarly, matches are flagged in-play if odds at any time differ significantly from their model's predictions. This model uses past matches to estimate how factors such as goals, red cards and time remaining shift the opening odds for the pre-match market (which should be roughly the same as the closing odds for the pre-match market.)

The models used by SportRadar to estimate odds and identify matches are proprietory, and are not discussed in either Forrest and McHale (2015) or Forrest and McHale (2019). Nonetheless, the articles provide useful information about the ways in which one fraud detection system considers the most important anomalies to search for in identifying suspicious matches, as well as reminding us that looking for anomalies in betting data is but one early step in a multi-stage process for identifying fixed matches.

Several other works take a similar approach to identifying fixed matches, but only look at pre-match markets. As part of a major news report about match-fixing in tennis Blake and Templon (2016), BuzzFeed also performed some statistical analysis looking at changes in pre-match odds, identifying matches with odds swings with at least one bookmaker of at least ten percentage points as potentially suspicious. This occurred in about 11% of all matches they considered. Since this could happen for innocent reasons, BuzzFeed focusses on players that have lost at least ten matches in which this happens. This includes matches which the players lose despite the odds swings suggesting the player will win - other sources, such as Rodenberg and Feustel

(2014), and Blake and Templon (2016) suggest that moves toward the eventual winner are far more suspicious.

Using the odds as estimated win probabilities, they simulate the outcomes of the matches with large swings for each player and find the probability they lose at least as many matches as they did. There are fifteen players for whom the probability is less than 5%. However, testing 39 players means some false positives would be expected as well, so a correction is applied to take this into account, after which four players remain. Since it is unlikely for these players to have lost so many matches with suspicious odds swings, these players are labelled as suspicious, and worthy of further scrutiny - though the authors still note that this is not enough to indicate guilt, and further investigation is required.

DW on Sport (2016) follow up on Blake and Templon (2016)'s research by de-anonymising BuzzFeed's data and examining some results to see if innocent explanations can be found for the suspicious matches. One of the players is highlighted in particular - DW on Sport (2016) randomly select eight of his fifteen matches and examine the odds and context in each case. In each case, another explanation is plausible, for reasons such as the player returning from a long injury (making the selection of opening odds very difficult), or one outlying bookmaker correcting their odds to be more similar to their competitors. One match even saw the pre-match market re-opened during an overnight rain delay, with the player's opponent one set ahead. This created the illusion of a large pre-match swing. This further emphasises the need to examine each match on a case-by-case basis, lest innocent players find themselves wrongly accused due to artefacts of the betting markets.

The work of Rodenberg and Feustel (2014) on identifying fixed matches focusses on changes in pre-match odds in tennis matches. Similar to the football model SportRadar use in their fraud detection system, as Forrest and McHale (2015) describes, Rodenberg and Feustel (2014) also use an Elo model to predict matches, and defines prediction error of the odds as the difference between the pre-match odds and

the Elo model prediction. Note that it is unusual to assume the predictive model gives the correct probabilities and look at the error of the odds with respect to that model, as many sources, such as Kovalchik (2016) and Feustel and Rodenberg (2015), find the odds tend to be better predictors than predictive models.

Nonetheless, this approach is taken so that the prediction error of the odds (the difference between the odds and model predictions) can be considered at the opening and closing of the pre-match market. By looking at changes in the error, rather than the odds themselves, the authors can differentiate between cases where odds move towards and away from the predictive model, since a move towards the predictive model is more likely to indicate mis-specified opening odds, whereas a move away is more likely to indicate a fix or the dissemination of new information.

Rodenberg and Feustel (2014) argues that first-round matches are most likely to be fixed, since a corrupt player knowing they will exit the tournament will conserve energy by fixing as soon as possible. When considering first-round matches with large increases in error, the odds were generally found to swing in favour of the eventual winner more frequently than in matches in later rounds. Since first-round swings are more informative of the eventual match-winner, this is presented as evidence that such swings are more likely to indicate fixes than swings in later rounds. It is not clear, however, the extent to which this behaviour might also relate to the phenomenon of unfit players retiring during first-round matches rather than beforehand in order to avoid forfeiting prize money, Agence France-Presse (2017). Should a player's lack of fitness become widespread knowledge before a match, such an odds swing might also be expected.

Feustel and Rodenberg (2015) instead focusses on football, and analyses the differences between closing pre-match odds and model predictions based on a bivariate Poisson model for goals, as in Dixon and Coles (1997). The goal-scoring rates of both teams were derived from offensive and defensive ratings assigned to each team by some unspecified method, using data in the relevant season only. Matches with the

biggest differences between odds and model probabilities were labelled as potentially suspicious.

Four seasons of data were separately analysed for each of four professional football leagues. France's Ligue 2 and Italy's Serie B were chosen due to known match-fixing scandals, whereas the Premier League and Major League Soccer were expected to have fewer fixed matches. The eight matches in each league with the biggest discrepancy between model predictions and odds were analysed, and the results were roughly as a match-fixing explanation might suggest: in Ligue 2 and Serie B, the winning team's odds were higher than expected rather than lower (suggesting gamblers potentially had more information than the model), the absolute differences between odds and predictions were bigger, and the matches occurred closer to the end of the season, where impact on promotion or relegation could be bigger, giving more incentive to fix.

Ötting et al. (2018) analyse both pre-match both odds and betting volumes in betting markets for football matches. Matches are flagged as suspicious if either of these observed quantities are significantly different to model predictions, based on seven years of data on Serie B matches, a number of which were known to be fixed. The model achieved a true positive rate of 79.2%, and a true negative rate of 64.5%

To model normal behaviour for both odds and volumes, generalised additive models for location, scale and shape, or GAMLSSs, are used. These extend generalized linear models in two ways. Firstly, they permit the response variable to depend on non-parametric and parametric functions of the predictor variables, and secondly, more parameters of the error distribution can be fitted than simply the mean. For example, if errors are normally distributed, the variance of the error terms could also depend on non-parametric functions of the predictor variables.

Betting volumes are modelled using a log-normal distribution for responses, since volumes are always positive. The mean and variance of volume distribution in each match are then fit using estimates for the quality of the two teams, the day of the

week and the stage of the season.

In order to model betting odds, it is assumed that these should be similar to modelling the probabilities of each team winning, as with previously discussed methods. These probabilities are modelled using the model of Karlis and Ntzoufras (2003), in which goals scored by each team scores are correlated Poisson random variables, with rates fitted using the GAMLSS method. The predictor variables chosen attempt to capture home advantage, the overall strength of each team and recent form.

Given these two models, the authors then attempt to classify matches depending on whether the difference (normalised by fitted conditional variance) between predicted values in each model and the observed values exceed some threshold. Thresholds are optimised to maximise true positive rate and true negative rates - the labels of the matches are known due to widely-publicised previous investigations into Italian match-fixing. They achieve best results by combining information from the model for volumes and odds, rather than considering each separately.

Two further papers that examine match-fixing in football Reade and Akie (2013) and Reade (2014). Both compare model predictions for pre-match odds with observed pre-match odds. The focus is on the probabilities of draws for a few reasons. Firstly, the authors claim (without proof) that the probability of draws is not as heavily affected as the probaility of a win for either team by information the model cannot capture, such as team or injury news. Similarly, the probabilities of a draw from bookmakers in both papers data sets appears less variable than the probabilities of wins, rarely rising much above one third making outliers significantly above this easy to identify. It is also claimed that since a draw is typically less interesting an event than a victory, and so fixing to draw may also attract less attention than fixing to win, making it easier for the fix to go unnoticed.

To model win and draw probabilities, an ordered probit regression model is used. This is an extension of probit regression, Section 2.2.1, which permits more than two ordered outcomes, which in this case allows for the inclusion of draws. The work dif-

fers from others discussed so far in that it is acknowledged that modelling odds is not necessarily the same as modelling match-win probabilities. Once model probabilities have been established, linear regression is used to establish a relationship between odds and model probabilities. This is to account for effects such as favourite-longshot bias, an inefficiency in some betting markets whereby strong favourites win more often than odds suggest, due to gamblers' preference for gambling on underdogs.

Williams (1999) and Forrest and McHale (2007) summarise much of the current understanding of favourite-longshot bias. It occurs markets where many gamblers have a preference for high-skewness bets, i.e. long-odds bets with a high potential return for little outlay. Competing explanations exist for the precise reasons behind the bets causing these biases, but most broadly centre around the fact that many people gamble recreationally, rather than as a profit-making exercise. Because of this, there may be greater value associated with the thrill of a long bet coming in rather than gambling on short odds for small but positive rewards, or simple overconfidence in the probability of unlikely events.

Because of these preferences for long-odds bets, the betting odds are then shifted to encourage betting on short-odds bets, and taking such bets can lead to greater long-term rewards than long-odds bets. The scale of the bias can depend heavily on the prevalence of strong favourites and underdogs in the sport. Forrest and McHale (2007) finds evidence of this bias in tennis, and also reviews research into the bias in other sports, discussing the types of horse race it is present or absent in, notes its absence in English Premier League football and a negative bias in some American sports.

Reade and Akie (2013) and Reade (2014) use their method to examine international football (including youth and women's competitions) and Serie B matches respectively. They look for matches with large difference between observed and predicted odds, and investigate the characteristics of such matches, such as stage of the season or whether the match is friendly or competitive.

A number of other papers choose to focus not on identifying suspicious matches through betting data, but on identifying other sources of potential evidence for match-fixing. Examples of this include seeing if certain variables that might indicate match-fixing help predict the outcome of sports matches, or seeing if matches hypothesised to be more susceptible to fixing differ from those less at risk in ways that might be consistent with fixing.

A key example in tennis is Rodenberg and Feustel (2014). As well as the aforementioned discussion of odds, they investigate match-fixing in two other ways. Crucial to their investigations is their hypothesis that first-round matches are more likely to be fixed than matches in later rounds. Their main argument is that prize money in tennis is heavily weighted to the latter rounds of a tournaments. This means that a player may have little financial incentive to progress through tournaments, given the risk of injury, burnout and lost training time, if they believe themselves to be unlikely to progress beyond the early rounds. The expected financial reward of attempting to win the match is low, and so players will be more tempted to fix first round matches than they would matches later in the tournament, when large financial awards await the winner. The authors use this hypothesis to estimate the prevalence of fixed matches by comparing certain features in first-round matches and later rounds. If these matches should behave the same under the assumption that no matches are fixed, but in fact behave differently, this may be evidence that some of the matches are fixed. The hypothesis appears reasonable, but it would be helpful if it were backed by a study of known fixed matches.

The first comparison rests on the claim that players fixing matches will exert less effort in order to avoid the strain and risky of injury from competitive matches. This theory is unverified. The authors state that if this claim is untrue then their methods underestimate the extent of match-fixing, but this is debatable. Were the claim untrue, and there were another explanation for the differences, then their methods would teach nothing about match-fixing. It could be hypothesised that the smaller

proportion of three-set matches in these early rounds of tournaments is due to the fact that the weakest players in the tournament are still present, meaning a greater average skill difference than in later rounds. It is claimed that this effect is controlled for by looking at "imputed win probabilities" without further explanation.

Based on this claim, the authors compare proxies for player effort in first-round matches and later rounds. Results show that proportion of three-set matches, tie breaks and breaks of serve by the loser is around around 1 percentage point lower in first-round matches, suggesting that the losers exert less effort in these matches. They conclude that around 1% of first-round matches involve "tanking", or playing to lose. This conclusion appears to be in error. Let p denote the proportion of fixed matches, and let x_f and x_c and x_1 probabilities of certain events (for example, the match going to three sets) in fixed matches, clean matches and first-round matches respectively. Then the equation $(1-p)x_c + p(x_f) = x_1$ is obtained. It is assumed that $x_f = 0$, and so fixed matches never go to three sets, x_1 is estimated from first-round data and x_c from matches after the first round, since all such matches are assumed clean. Solving for p then yields $p = (x_c - x_1)/x_c$. However, the authors simply estimate p using $p = x_c - x_1$, underestimating the rate of match-fixing by a factor of $1/x_c$. The authors' later examination of betting odds correctly uses this equation, so the reasons for its omission here are unclear.

The next set of tests attempts to use betting odds to similarly infer the degree of match-fixing. Since betting markets are expected to be approximately efficient, betting returns using sports models should be approximately equal in first-round matches and later matches. However, if first-round matches are being fixed then these matches become harder to predict, and so the betting returns using the model should be lower than in later rounds. The authors use two sports models and a simple betting strategy, and find both perform worse in first-round matches than later matches. They claim that their results suggest that between 1.58% and 2.71% of first-round matches are fixed. It is unclear the extent to which this effect could also be explained by players

carrying injuries into tournaments. If knowledge of such an injury was widespread, the model predictions would have less information than gamblers. The gamblers' knowledge would be reflected in the betting odds, resulting in lower betting returns for the sports model.

The work of Duggan and Levitt (2000) examines the sport of sumo wrestling. In a sumo wrestling tournament, wrestlers take part in fifteen matches. Winning more than half of these bouts will increase a wrestler's rank, while winning less than half will cause a wrestler's rank to fall. Hence, if a wrestler going into his final match having won seven matches of fourteen, there is a big motivation, both financially and in terms of ranking, to win the final match. (A wrestler going into his final match having both won and lost seven is said to be "on the bubble".) The authors show that wrestlers are much more likely to win such a match than lose it. To explore this effect Duggan and Levitt (2000) uses a linear regression of the form

$$\text{Win}_{ijtd} = \beta\ \text{Bubble}_{ijtd} + \gamma\ (R_i - R_j) + \lambda_{ij} + \delta_{it} + \epsilon_{ijtd}, \tag{2.1.1}$$

where Win_{ijtd} is the win probability of wrestler i against wrestler j in tournament t on day d. The value of Bubble_{ijtd} is 1 if wrestler i is on the bubble, -1 if wrestler j is, or 0 if both or neither are, and parameters β and γ are to be fitted. The rank of wrestler i is R_i, the residuals are ϵ_{ijtd}, and λ_{ij} and δ_{it} are optional wrestler-wrestler and wrestler-tournament terms respectively. Note it is more usual in the sports modelling literature to use a logistic regression than linear regression to model win probabilities to ensure results remain in the interval [0,1] - see Section 2.2.1.

This effect could, of course, be simply attributed to wrestlers on the bubble trying harder than their opponents with little to gain or lose. However, Duggan and Levitt (2000) also investigates a number of other potential factors that might indicate fixing being more likely than increased effort. These include a smaller bubble effect during periods of high media interest in corruption, and some statistical evidence of reciprocal arrangements between wrestlers in the same "heya", or stable. (Similar to horse-racing, wrestlers in the same stable are not considered team-mates, but the

success of all wrestlers in the stable benefits the stable, leading to the possibility of fixing being co-ordinated within a stable.) Additionally, whistle-blowers in the sport identified a number of wrestlers as corrupt, as well as declaring others innocent. Duggan and Levitt (2000) found the bubble effect to be especially prominent amongst those identified as corrupt, and negligible between wrestlers identified as clean.

This thorough investigation points to a number of factors that might indicate match-fixing - it is the agreement of multiple pieces of evidence that proves useful, indicating a strong potential for match-fixing in the sport.

The work of Deutscher et al. (2017) instead focusses on refereeing in football. Specifically, it aims to investigate whether betting markets might contain evidence that certain Bundesliga referees may be influencing football matches for financial benefit. They choose the "over/under 2.5 goals" market on the basis that it may be easier to influence this than the match-win outcome undetected.

The authors employ linear regression to model betting volumes in each match to see whether more than expected is being bet when certain referees are officiating matches. In order to do this, a range of other factors must be included in the linear regression, such as the identities of the teams, the year and week of the match and other match-specific variables. The only match-specific variables described observed referee performance, as measured by objective statistics, such as penalties and cards awarded to each team, and subjective ratings given to the referee for their performance by *Kicker* magazine. The idea is to identify referees on whose matches more than expected is being bet for no observable reason other than the identity of the referee, as this is a possible sign that fixers are bribing these referees to subtly influence whether or not at least 2.5 goals are scored in the match.

They find two or three referees, depending on combinations of variables used, on whose matches more is bet than expected. The authors do however concede that other explanations for this may be possible. One limitation they cite is that assignment of referees to matches is not wholly random, as more experienced referees are assigned

to more high-profile matches.

Other potential shortcomings include use of linear regression for positive, right-skewed betting volume data - log regression may have been more appropriate. Variable selection is also key to such analyses.

The scale of match-fixing in basketball is a much-debated topic in academic literature, and presents an interesting case study on the difficulties involved in using indirect methods to analyse match-fixing. It can be easy to identify strange behaviour which could be a result of match-fixing, but harder to rule out other explanations, as the wealth of disagreements in the literature on basketball match-fixing, or "point-shaving" shows.

The principal concern of the work of Wolfers (2006), Gibbs (2007), Bernhardt and Heston (2010) and Diemer and Leeds (2013) is the popular "point-spread" market. In basketball, it is common for matches to have heavy favourites, making betting on the winner uninteresting. Instead, many bookmakers offer an approximately even bet on whether or not a team will win by a certain number of points. This margin of victory is called the points spread. In this way, matches between even teams will have a spread of 0, matches with slight favourites may have a spread of around five or lower, whereas matches with a heavy favourite may have a spread of twelve points or even higher. In cases where bookmakers would normally change their odds to avoid losses, they instead change the spread. The fact that it remains an approximately even bet means it remains interesting, even in matches with big favourites.

However, players are still mainly concerned about the fact of winning rather than the margin of victory. This presents an opportunity for corruption, since teams in the lead can aim to win by less than the spread, achieving their sporting aims while still earning money by fixing the point-spread market. This is known as point-shaving, and Bernhardt and Heston (2010) highlight several high-profile examples of this happening.

In order to investigate whether this happens regularly, Wolfers (2006) choose to

view the spread as a "prediction-market-generated median forecast", Wolfers and Zitzewitz (2004) - in other words, the spread is a forecast of the score, and errors in this forecast (the difference between the spread and observed margin of victory) should be normally distributed. However, analysis of college basketball data reveals that while this held for matches without a strong favourite (where the spread, S, was less than 12), the forecast error was in fact asymmetric for matches with matches where the S was at least 12. In these cases, the favourites won by (strictly) between 0 and S points 46.2% of the time, but by between S and $2S$ only 40.7% of the time. After some exploration and dismissal of alternative hypotheses, such as market inefficiency or strong teams exerting less effort after establishing a lead, it is concluded that this is probably evidence of point-shaving. Gibbs (2007) conduct a similar analysis on NBA (National Basketball Association) data with similar findings.

However, Bernhardt and Heston (2010) explores these alternative hypotheses differently and finds a different conclusion. They build a model for spreads based on the teams' form in attempt to estimate spreads for matches without a betting market. They find the same asymmetries in differences between estimated spread and observed margin of victory in matches with strong favourites in matches both with and without betting markets. This, they claim, suggests that point-shaving is not the best example for this phenomenon, since there is no incentive to shave in matches without betting markets, and suggest decreased effort for winning teams as an explanation. On the other hand, Borghesi (2008) instead uncovers interviews suggesting that bookmakers raise spreads in matches with strong favourites to take advantage of basketball gamblers' predilection for betting on favourites, so that the bookmakers profit in the likely scenario that the favourites win but fail to cover the spread. This bias among gamblers suggests that the point spread market is not totally efficient. Diemer and Leeds (2013) disagrees with both, arguing that if there were an innocent explanation for this phenomenon, the distribution of forecast errors would still be symmetric, but not around 0 - this is not the case, however, with too many large wins

being observed. It seems conceivable, however, that the inflated probability of large wins could result from strong favourites sometimes conserving effort, but sometimes playing at full capacity, creating a bimodal distribution that leaves open the possibility for big wins.

Whatever the truth, the array of competing hypotheses show the importance of investigating alternative theories fully, but also the difficulty in explaining anomalies that do not conform to one's model of "usual behaviour". This is a fact that we must be acutely aware of throughout our analysis - caution must be applied to all conclusions, as there may be an innocent explanation behind any anomaly. SportRadar's system of only using betting data to flag matches for further investigation therefore seems very sensible, and is one we will aim to follow throughout.

2.2 Pre-Match Tennis Modelling

2.2.1 Regression and Machine Learning

One of the simplest approaches one can take to modelling the probability that player i wins a match against player j is using regression-based methods. This involves attempting to use information about the two players, as denoted by a covariate matrix X_{ij}, and a vector of fitted parameters β, to make inference about the probability that player i wins the match, π_{ij}. This is modelled via a link function, typically either an inverse logistic function, $\mathcal{L}^{-1}(\pi) = \log(\frac{\pi}{1-\pi})$, or a probit link function, $\Phi^{-1}(\pi)$, where $\Phi(z) = P(Z < z)$ for $Z \sim N(0, 1)$. Depending on which is chosen, this then involves fitting the model

$$\mathcal{L}^{-1}(\pi_{ij}) = \boldsymbol{\beta} X_{ij}, \text{ or}$$
$$\Phi^{-1}(\pi_{ij}) = \boldsymbol{\beta} X_{ij}.$$

In practice both functions are very similar, so the choice makes little difference. Boulier and Stekler (1999) use the difference between tournament seedings, as the

only covariate, while Clarke and Dyte (2000) use the difference between the logs of the player's ATP ranking, $\log(r_i)$. Klaassen and Magnus (2003) use the difference between transformed ranks, $R_i = 8 - \log_2(r_i)$, which should give equivalent predictions. Irons et al. (2014) suggest adding another covariate $r_i - r_j$ to Clarke and Dyte (2000)'s model to give more freedom, so that the difference between the 5th and 10th best players need not be the same as that between the 500th and 1000th. They also allow covariates to vary depending on whether a three or five-set match is being played. Meanwhile, McHale and Morton (2011) suggest using ranking points, instead of raw rankings, as covariates in Clarke and Dyte (2000)'s model.

Many papers have performed logistic regression with far more covariates, such as Gilsdorf and Sukhatme (2008), Del Corral and Prieto-Rodríguez (2010), Sipko and Knottenbelt (2015) and Hostačnỳ (2018). Each takes its own approach to covariate selection, often depending on the emphasis of the paper. Additionally, Hostačnỳ (2018) examines 342 covariates taken from across a broad range of past work of this type, and use penalised likelihood to select the most impactful.

The extra covariates generally fall into three broad categories, as defined by Del Corral and Prieto-Rodríguez (2010): past performance, player characteristics and match characteristics.

Past performance covariates give different information about players' previous results to official rankings and ranking points. Gilsdorf and Sukhatme (2008) look at head-to-head records and total career wins on the relevant surface, while Sipko and Knottenbelt (2015) combine points won on serve, return and aces in various ways, averaging with more weight on recent performance, as well as head-to-head records. Del Corral and Prieto-Rodríguez (2010) introduce a dummy variable to represent whether a player has previously been a top-10 ranked player, as this could be a sign that the current rankings underrate a top player returning from injury, for example. Lisi and Zanella (2017) uses ranking points, and also puts ranks into intervals instead of looking at raw ranks, since these ranking intervals are less highly correlated with

ranking points than the raw ranks. The problem of using highly correlated predictors is known as collinearity. Using correlated predictors can make the regression coefficients for these predictors hard to estimate, which can in induce unnecessary variance into out-of-sample predictions, though this is mostly an issue in cases where there predictors are not correlated as highly in the new sample. This is unlikely to be the case with rankings and ranking points, but this step to decorrelate the predictors could still be a sensible precaution nonetheless.

Physical characteristics covariates describe other player attributes. Using a linear and quadratic term for age is common to capture the rise and fall of a player over their career, while Del Corral and Prieto-Rodríguez (2010) also consider height and preferred hands of both players. Sipko and Knottenbelt (2015) also considers the potential fatigue of players, as measured by the number of games played in the past three days, as well as an indicator variable denoting whether the player's last match ended in a retirement.

Finally, match characteristics focus on information about the match that is not specific to the players. Surface is a common choice, while Del Corral and Prieto-Rodríguez (2010) also consider the tournament level and round. Gilsdorf and Sukhatme (2008) consider all of these to help investigate whether the difference in potential prize money for a player winning the tournament compared to losing in the current round affects probabilities, and conclude that a larger difference favours the stronger player.

Most machine learning methods used in the literature to model tennis matches work on a broadly similar basis, but using a much wider class of functions than the linear predictor and the logit and probit link functions. Somboonphokkaphan et al. (2009) and Sipko and Knottenbelt (2015) both use an artificial neural networks with surface and historical proportions of points won or lost on serve, while Sipko and Knottenbelt (2015) also extends to support vector machines. Hostačnỳ (2018) includes a much richer set of 342 covariates, and also considers Random Forests and other tree-based methods. Cornman et al. (2017) test all three methods and logistic

regression against each other and find limited differences in predictive performance for their implementation. However, this could be significantly affected by features selected and choice of hyperparameters.

2.2.2 Bradley-Terry Models

An alternative model for pairwise competitions is the Bradley-Terry model. Suggested by Bradley and Terry (1952), but also studied by Zermelo (1929), this model is popular for its simplicity and effectiveness.

Under the Bradley-Terry model, each competitor i is modelled as having a strength parameter $\alpha_i > 0$. If competitors i and j are compared (which in a sporting contest would be a match between the two), then the probability assigned to the event of i being found superior to j is

$$P(i \text{ beats } j) = \frac{\alpha_i}{\alpha_i + \alpha_j}.$$

Under the assumption that all matches are independent, the joint likelihood to be maximised is then

$$L(\boldsymbol{\alpha}|\boldsymbol{w}) = \prod_{(i,j)\in\Omega} \prod_{k=1}^{n_{ij}} \frac{\alpha_i^{w_{ijk}} \alpha_j^{1-w_{ijk}}}{\alpha_i + \alpha_j}, \tag{2.2.1}$$

where Ω is the set of all pairs (i, j) such that i plays a match against j, n_{ij} is the number of matches between i and j in Ω, and $w_{ijk} = 1$ if player i wins their k-th match against player j, or else equals 0. The vector $\boldsymbol{\alpha}$ denotes $(\alpha_1, \ldots, \alpha_n)$ if there are n players, and $\boldsymbol{w}$ describes all w_{ijk}.

It can be easily seen that this likelihood does not have a unique maximum. For any $\boldsymbol{\alpha}$ that maximises the likelihood and any constant C, then, $C\boldsymbol{\alpha}$ is also a maximum likelihood estimator. Hence a constraint must be placed on $\boldsymbol{\alpha}$. Bradley and Terry (1952) choose to specify that $\sum_{i=1}^{n} \alpha_i = 1$, but McHale and Morton (2011) instead specify $\alpha_1 = 1$. Bradley and Terry (1952) then go on to describe their algorithm for maximising this likelihood.

McHale and Morton (2011) extends the basic model provided in equation (2.2.1) to a tennis context. Of course, not all matches are equally relevant to a player's current strength. Matches played a long time ago may have almost no relevance to current beliefs about player strengths, so perhaps not as much importance should be placed on these earlier matches. The likelihood of these past matches is therefore downweighted by McHale and Morton (2011), following on from ideas suggested by Dixon and Coles (1997). If the current time is t, let the likelihood of the k-th match between players i and j, which occurs in the past at time t_{ijk}, be downweighted by factor $\exp\big(\epsilon(t - t_{ijk})\big)$ for some parameter $\epsilon > 0$.

Surface can also play a large part in modelling player strengths, and matches on a different surface to the current match may also not be as relevant. Hence, if the current match is taking place on surface S, McHale and Morton (2011) instead define player i's current strength on that surface as α_{itS}. If the k-th match between i and j takes place on surface S_{ijk}, its likelihood is downweighted by a factor $\Gamma_{S,S_{ijk}}$ which takes value 1 if $S_{ijk} = S$, or else some value between 0 and 1. Let Γ be a 3×3 matrix containing all possible values of $\Gamma_{S,S_{ijk}}$, as there are three main surfaces: hard, grass and clay).

McHale and Morton (2011) further improve the model by noting that simply looking at whether player i wins or loses the match throws away readily available, and potentially very useful, data about tennis matches. For example, a player that has lost 6-0, 6-7, 6-7 has clearly played better than one that has lost 0-6, 0-6, and it is possible to incorporate this into the updates of the players' strengths. The approach taken by McHale and Morton (2011) to account for this is to essentially view a tennis match as a series of contests, where each contest is a game, instead of one match contest. If g_i and g_j are the numbers of games won by each of players i and j, the likelihood of observing a scoreline can then be calculated accordingly as

$$L(\alpha_{itS},\ \alpha_{jtS}|g_i, g_j) \propto \frac{\alpha_{itS}^{g_i}\ \alpha_{jtS}^{g_j}}{(\alpha_{itS} + \alpha_{jtS})^{g_i+g_j}}.$$

Although this does not account for which player is serving, since both players serve a roughly equal number of games, it is hoped that this effect will not be substantial. Let A_t denote the set of matches that happens before time t, and let n_{ijt} denote the number of matches players i and j play before time t. For some ϵ and $\Gamma_{S,S_{ijk}}$, the likelihood to be maximised over $\boldsymbol{\alpha}_{tS} := (\alpha_{1tS}, \ldots, \alpha_{ntS})$ is then expressed as

$$L(\boldsymbol{\alpha}_{tS}|A_t) = \prod_{(i,j)\in A_t} \prod_{k=1}^{n_{ijt}} \Big(\frac{\alpha_{itS}^{g_{ijk}} \; \alpha_{jtS}^{g_{jik}}}{(\alpha_{itS} + \alpha_{jtS})^{g_{ijk}+g_{jik}}} . \Big)^{e^{\epsilon(t-t_{ijk})}\Gamma_{S,S_{ijk}}}. \tag{2.2.2}$$

Note that in this expression, the current strength of player i, α_{itS}, is applied to all matches played by that player - even those in the past on a different surface. However, by downweighting these past matches where players have the "wrong" strength α_{itS}, these matches' contribution to the likelihood of the current strength is minimised, meaning the more recent matches on the correct surface bear the most relevance, while older matches gradually become less and less important. If a new set of matches is observed, the maximisation of this likelihood must be performed from scratch, as the weights of all matches will change.

McHale and Morton (2011) briefly discuss an extension that incorporates actual time dynamics in the same way as Dixon and Coles (1997) proceeds to use, but believe that such an extension only brings minor benefits to modelling power for a significantly higher computational cost. Selecting values for ϵ and $\Gamma_{S,S_{ijk}}$ is not as straightforward as simply allowing them to be selected by maximum likelihood in equation 2.2.2. McHale and Morton (2011) note that simply taking $\epsilon = 0$ and $\Gamma_{S,S'} = 1$ for all surfaces $S \neq S'$ would increase the likelihood, without necessarily being more useful for predicting current tennis matches. This is further discussed by Dixon and Coles (1997). Instead, the approach employed is to choose these values maximise the predictive accuracy of the model.

Let $\hat{\boldsymbol{\alpha}}_{tS}(\epsilon, \Gamma)$ denote the vector $\boldsymbol{\alpha}_{tS}$ that maximises equation 2.2.2 given ϵ and Γ, and let $\boldsymbol{W}_{t+1,S}$ be a random variable denoting the results of all matches at time $t+1$ and on surface S, with $\boldsymbol{w}_{t+1,S}$ a realisation of that random variable. Then

$P\big(\boldsymbol{W}_{t+1,S} = \boldsymbol{w}_{t+1,S}|\epsilon, \Gamma, \hat{\boldsymbol{\alpha}}_{t,S}(\epsilon, \Gamma)\big)$ is the predicted probability of observed results $\boldsymbol{w}_{t+1,S}$ occurring given ϵ, Γ and $\hat{\boldsymbol{\alpha}}_{tS}(\epsilon, \Gamma)$. The goal is to select ϵ and Γ that maximise

$$\prod_{S} \prod_{t=t^*}^{t^{max}-1} P\Big(\boldsymbol{W}_{t+1,S} = \boldsymbol{w}_{t+1,S}|\epsilon, \Gamma, \hat{\boldsymbol{\alpha}}_{tS}(\epsilon, \Gamma)\Big),$$

which is the predictive probability assigned to all matches after some time t^* using all data up to one time step before each match. The value t^* is selected to be high enough that the model can assign good predictions to all matches considered (McHale and Morton (2011) suggest one year into the data), and t^{max} is the last time point in the data.

The higher this product of probabilities is, the better the model has forecast future tennis matches, and hence it is better to try to maximise this than the likelihood. A grid search over possible values of ϵ and Γ is employed to achieve this.

The work of McHale and Morton (2011) is a key starting point for the work of Irons et al. (2014). The primary focus is not forecasting accuracy, but to develop an alternative to the current ATP/WTA ranking system that better reflects players' strength, but also has several features desirable of official rankings. One example is that it must incentivise attendance at major tournaments (chiefly Grand Slams). Also, a player whose strength remains constant over time but varies on different surfaces should ideally keep a constant rank through the year, and should not be rewarded simply for having recently played on the player's favourite surface.

To investigate these affects, Irons et al. (2014) explore different downweighting functions for time and surface, as well as different forms of likelihood, and compare forecasting accuracy to McHale and Morton (2011)'s model and the official rankings, as well as exploring seasonality, surface bias and other effects in all three. They find a trade-off exists between forecasting accuracy and several desirable properties of ranking systems, but using the framework of McHale and Morton (2011) still manage to devise a ranking system that forecasts better than the official rankings while having several key properties.

2.2.3 Dynamic Pairwise Comparisons

An alternative set of models are dynamic pairwise comparison models. In these models, each player has an unknown strength, and inference about that strength is made purely based on observed results from matches between two players, or pairwise comparisons.

One of the first and most famous of these models was the Elo model, Elo (1978), developed to rank chess players. It was adopted by the USCF in 1960, and FIDE in 1970. The Elo model is often presented purely as an update algorithm for ratings, but it is interesting to note its statistical origins, and important to do so to give context to other methods that have come since. The version featuring only wins and losses is discussed here, as it is most relevant for our tennis context, but some subtle additions were made in the original to account for draws in chess.

Elo noted that chess players do not play at a constant level, but there is in fact some variation in how well players play. Thus Elo decided to assume that chess player i's performance level in a given match was a Gaussian random variable θ_i, and that an appropriate marker of their ability would be their "average" level of performance, r_i. Under the assumption that each player had the same variance in performance levels σ^2, the distribution of their performance in a given match is then

$$\theta_i \sim N(r_i, \sigma^2).$$

The parameters are of course never observed, and hence inference about them can only be made through the results of matches between players. If one player beats another once, it suggests the winning player was the stronger of the two. Repeated wins for one player strengthens this belief. In order to form a statistical basis to make this inference, an assumption was required about the probability player i beat player j. Let the random variable S_{ij} take the value 1 if player i beats j, or 0 otherwise. Given some arbitrary scaling constant q, this probability was then expressed as

$$E_{ij} := P(S_{ij} = 1) = \frac{1}{1 + e^{-q(\theta_i - \theta_j)}}.$$

Given this probability of observing results and observation of a match result, inference about the players can be made using Bayes Theorem. The posterior distribution of θ_i and θ_j becomes

$$\pi(\theta_i, \theta_j | S_{ij} = s_{ij}) \propto \pi(\theta_i)\pi(\theta_j)P(S_{ij} = s_{ij}|\theta_i, \theta_j)$$

Elo made a number of simplifying assumptions about this posterior in order to approximate it by another Gaussian random variable. This allowed him to arrive at a very simple and well-known update formula for the posterior mean, r_i', given by

$$r_i' = r_i + K(S_{ij} - E_{ij}), \tag{2.2.3}$$

where the parameter $K > 0$ is chosen to alter the speed at which ratings update, and should also depend on q and σ. A large value of K means a large adjustment is made for the match that has just occurred, whereas a small K means smaller updates, making older results more important.

Two other factors should be taken into account. Firstly, in this Bayesian parameter update, the posterior variance should also be updated to some value σ'^2. Secondly, players do not have constant ability, but instead improve and get worse over time, going through various peaks and troughs. Both of these should be taken into account, but Elo decides to treat both of these together.

Let r_{it} denote player i's rating at time t. If it were assumed that ratings evolved as a Gaussian random walk, so that $r_{i,t+1} \sim N(r_{it}, \gamma^2)$, then it would follow that each player's variance after playing a match and evolving would be defined as $\sigma^{*2} := \sigma'^2 + \gamma^2$. This would have to be updated after each match, it might be different for different players, and it would also affect K.

Instead of following this approach, Elo decided a similar effect could be achieved by choosing $\sigma^{*2} = \sigma^2$ instead, restoring the posterior variance to its previous value. This meant that it was never necessary to store or update σ, making storage and computation easier, as well as removing any potential dependence of K on specific σ values. Instead, updating the mean, r_i, is all that is required. This gave a powerful

statistically based model for chess ratings that was very easy to update and understand.

Although the ease of usage and understanding for chess players is as relevant today as it was in 1960 when the Elo method was first developed, the advantage of computational ease is not, with the more advanced computational power that now exists. Hence, many have come up with more advanced adaptations of Elo's work that try to better model players' abilities.

Several sources find Elo updates new players too slowly, and have made attempts to rectify this by tweaking the K-factor. FIDE themselves implement a K-factor that halves after a player's first 30 games, with a few exceptions, FIDE (2017). Similarly, Glickman and Doan (2017) describe how in most cases USCF update ratings after a tournament using $K = \frac{800}{N'+m}$, where m is the number of games played in the tournament, and N' is a player's "effective number of games played". We do not discuss the definition of N' here, but note that it is capped at 50. Meanwhile, the analysis performed by Morris and Bialik (2015) for fivethirtyeight.com on tennis players uses a system that gives each player i a different K_i value that depends on the player's number games played, N_i. For fitted parameters K_0, ξ and η, this is equal to

$$K_i = \frac{K_0}{(N_i + \xi)^\eta}.$$

In all of these examples, this reduction in K over time is analagous to newer players having a higher value of σ, representing the large uncertainty in their ability. Information from more recent matches for these newer players therefore carry comparatively more weight, and the posterior distribution of players' ratings is less heavily weighted to the prior.

Gorgi et al. (2018) derive an alternative dynamic paired comparison model for updating player strengths of tennis players using a generalised autoregressive score (GAS) model, Creal et al. (2013). Under close inspection, it transpires that the GAS model is also equivalent to the basic Elo model as given by equation 2.2.3. However, it does provide an alternative framework for including extra information such as surface,

number of sets and Grand Slam effects.

The work of Glickman (1999) aims to stick closer to the original idea of Bayesian updating by removing some relevant assumptions. It is designed to be a general method, but the original paper discusses examples with chess and tennis. The authors no longer assumed that each player has constant variance σ^2. Instead, each player i has their own uncertainty parameter σ_i which is updated after each match along with r_i. This means that new players or those that have taken long breaks can update their ratings at different speeds to those with more stable ratings. Additionally, fewer assumptions and simplifications were made in approximating the posterior distribution by a new Gaussian. However, no discussion is given as to whether this provided an improvement in predictive performance over standard Elo ratings.

Subsequently, the work of Glickman (1999) has been further adapted by adding extra parameters, such as in the Glicko-2 ratings, Glickman (2012), and Stephenson ratings, Stephenson and Sonas. (2016), and alternative chess ratings system.

2.3 In-Play Tennis Modelling

Most of the methods previously described here work on the basis of directly modelling the probability of each player winning. However, another approach is possible. Tennis matches, like for most other racquet sports, have a hierarchical structure. Matches are played as a series of points - a player must win a certain number of points to win a game, a certain number of games to win a set, and a certain number of sets to take the match. Because of this structure, one could instead choose to model the probabilities of players winning points at different times of the match. The idea is to model this micro-scale behaviour first, and see how this affects the "emergent behaviour", the probabilities of the players winning matches.

This gives the ability to estimate the probabilities of players winning from any scoreline in the match, and allows for the analysis of in-play match-fixing. This is

extremely useful in that it unlocks a huge additional market to scour for corrupt activity, and distinguishes our work from most other efforts in the literature, which focus on pre-match markets. Additionally, it is also possible to update the estimates of the relative strengths of two players mid-way through a match based on how they have performed, allowing for more accurate predictions. It even permits for easy analysis of far more markets than simply which player wins or loses, such as how many sets each player wins, or "spread betting" on the difference of the number of games each player wins. However, this is not an area touched on further in this thesis.

2.3.1 Markov Chains for Tennis Matches

In order to model tennis matches in-play, two simplifying assumptions are generally made. First of all, it is assumed that the outcomes of different service points in tennis by the same player are independent, so that the outcome of one point does not influence the next. Secondly, points are identically distributed. In a match between player i and player j, player i will win each service point with probability p_{ij} independent of all other points. Similarly, player j wins points on serve with probability p_{ji}.

Making these assumptions allows a tennis match to be formulated as a Markov chain, in which states are scores, and transition probabilities are governed by p_{ij} and p_{ji}. In this setting, it is relatively straightforward to calculate the probabilities of each player winning the match, given p_{ij} and p_{ji}. Formulae for this are derived by O'Malley (2008) using combinatorial arguments, while Barnett et al. (2002) iteratively calculates probabilities with recurrence relations. However, the basic properties of absorbing Markov chains, as described for example in Grinstead and Snell (2012), are also sufficient to give probabilities of either player winning.

Recall that a state in a Markov chain is transient if there is a non-zero probability that the Markov chain never returns to that state, or else it is recurrent. A state is absorbing if the Markov chain can never leave that state. If every state can reach an absorbing state, then the Markov chain is called an absorbing Markov chain. In the

example of a game in tennis, as shown in Figure 2.3.1, the states in which each player has won the game are absorbing, and all other states are transient.

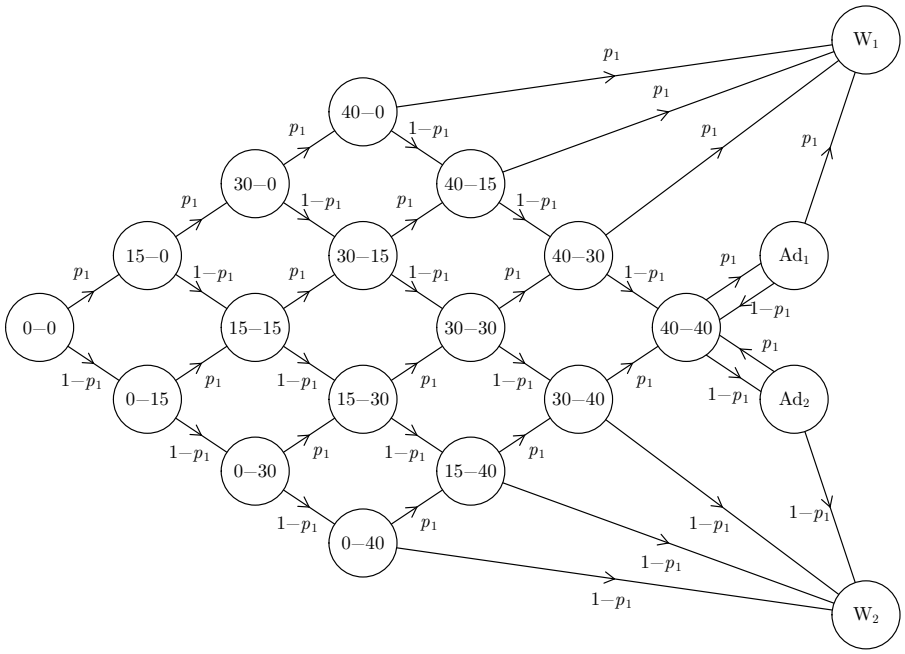

Figure 2.3.1: A Markov chain representing a game of tennis with player 1 serving. State W_i represents player i having won the game, Ad_i denotes advantage to player i for $i = 1, 2$, and D denotes deuce. The probability of player 1 winning a point on serve is p_1.

Suppose the matrix has n_t transient states and n_a absorbing states. Then the transition matrix $\mathbf{P}$ takes the form

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_{n_a} \end{pmatrix}.$$

In this transition matrix, $\mathbf{Q}$ is of size $n_t \times n_t$ and $\mathbf{I}_{n_a}$ is an identity matrix of size $n_a \times n_a$. The matrix $\mathbf{R}$ is non-zero, and $\mathbf{0}$ is a zero matrix of size $n_a \times n_t$.

It can then be proven (for example by Grinstead and Snell (2012)) that to find the matrix $\mathbf{B}$ of probabilities to absorption from each state to each absorbing state, we take

$$\mathbf{B} = (\mathbf{I}_{n_t} - \mathbf{Q})^{-1}\mathbf{R}.$$

This can easily be applied in the tennis example to provide probabilities of absorption, or each player winning the game.

While calculating match-win probabilities in practice, one could create a huge Markov chain featuring all possible scores in the match, though in general this proves inconvenient as the resulting Markov chain is so large. It is instead generally considered more convenient to create smaller Markov chains relating to games, sets and matches, and use conditional probability to combine them appropriately. In order to win a match, players must win a certain amount of sets, and a state space model featuring the number of sets won as states is also a Markov chain. Transition probabilities can be found by making a Markov chain for a set, in which games won by each player are states, and so on. Using this, the probability of a player winning a match can be found by using conditional probabilities of winning games and sets from these smaller Markov chains with the equations

$$\begin{aligned} P(i \text{ wins match}) = \quad & (2.3.1) \\ & P(i \text{ wins game})P(i \text{ wins set}|i \text{ wins game})P(i \text{ wins match}|i \text{ wins set}) \\ +& P(i \text{ wins game})P(i \text{ loses set}|i \text{ wins game})P(i \text{ wins match}|i \text{ loses set}) \\ +& P(i \text{ loses game})P(i \text{ wins set}|i \text{ loses game})P(i \text{ wins match}|i \text{ wins set}) \\ +& P(i \text{ loses game})P(i \text{ loses set}|i \text{ loses game})P(i \text{ wins match}|i \text{ loses set}). \end{aligned}$$

Exactly the same logic can be applied should the match be in a tie-break instead of a game.

Plots of these smaller Markov chains are shown in Figures 2.3.2, 2.3.3, 2.3.4 and 2.3.5, including a different version of the Markov chain for a game to that shown in

Figure 2.3.1. When observing the Markov chain for a game, it can be noticed that the state "40-30" has exactly the same transition probabilities as state Ad_i, which denotes advantage to player i. Should player i win the point, they also win the game, or else the deuce state is reached. These states can therefore be merged to reduce the size of the Markov chain without losing any information. Similarly, state "30-40" can be merged with state Ad_j, which leaves state "30-30" with identical transition probabilities to the deuce state, and thus these can be merged too. Similar arguments can be made to reduce the size of the set Markov chain too.

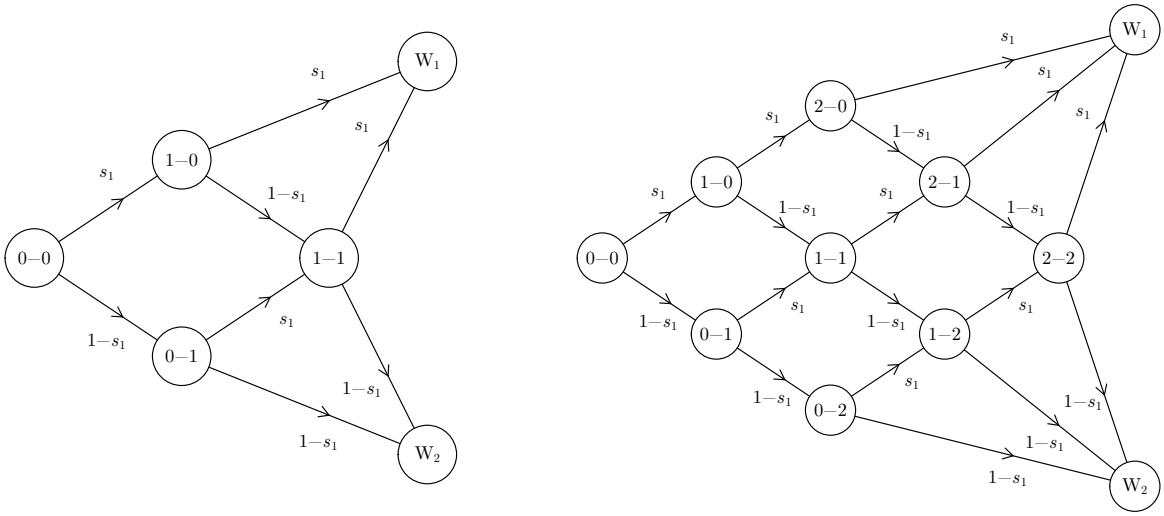

Figure 2.3.2: Markov chains representing tennis matches of three sets (left) or five sets (right). State W_i represents player i having won the match, and the probability of player i winning a set is s_i for $i = 1, 2$.

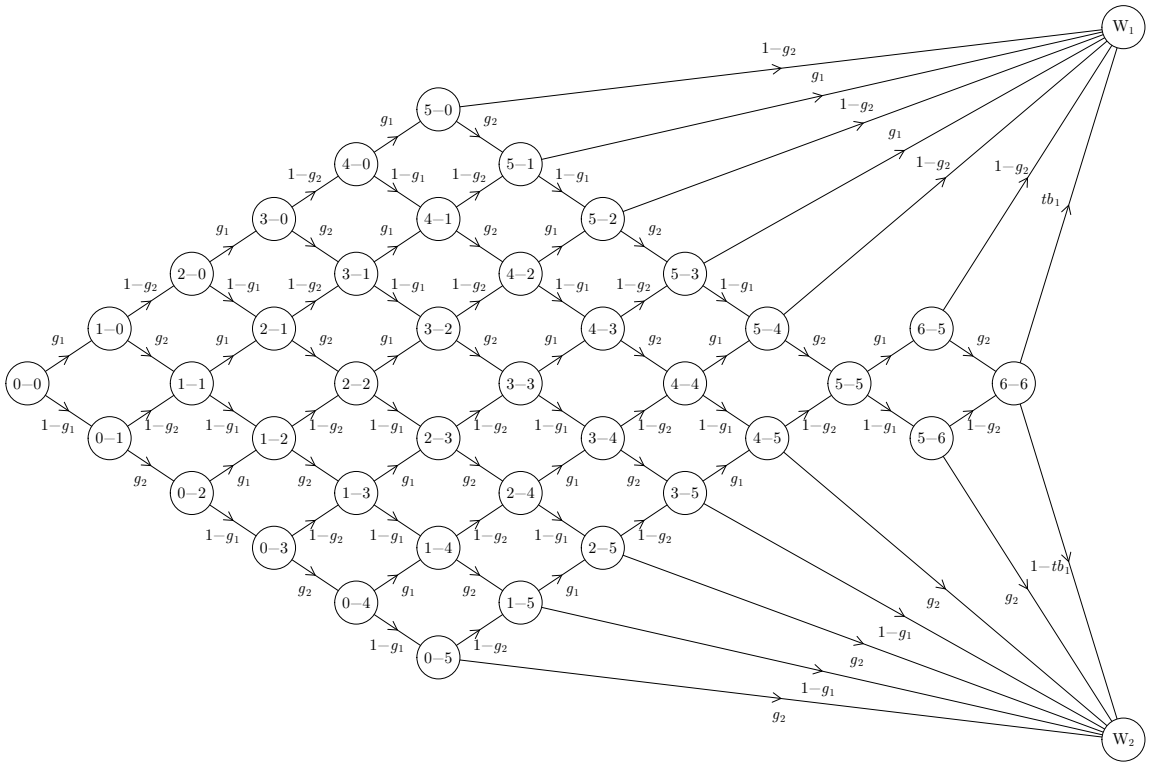

Figure 2.3.3: A Markov chain representing a normal set in a tennis match. State W_i represents player i having won the set, and the probability of player i winning a game on serve is g_i for $i = 1, 2$.

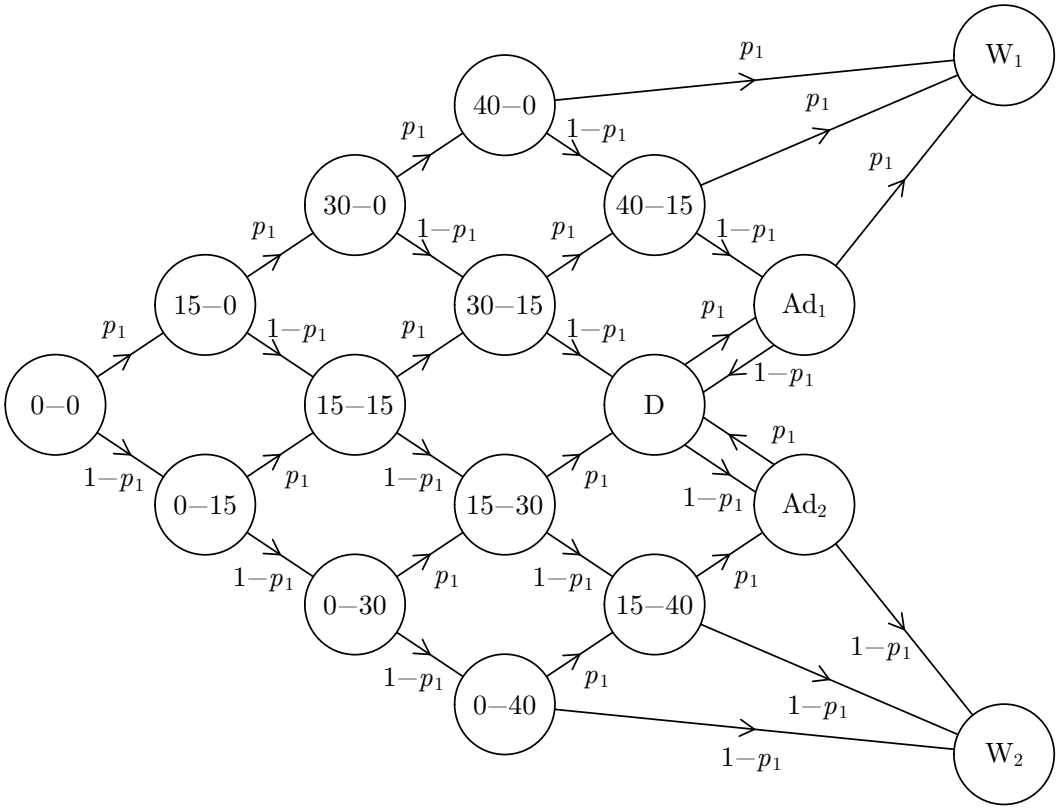

Figure 2.3.4: An alternative but equivalent Markov chain to 2.3.1 representing a game of tennis in which player 1 is serving. States W_i represents player i having won the game, Ad_i represents advantage to player i and D represents deuce for $i = 1, 2$. The probability of player 1 winning a point on serve is p_1.

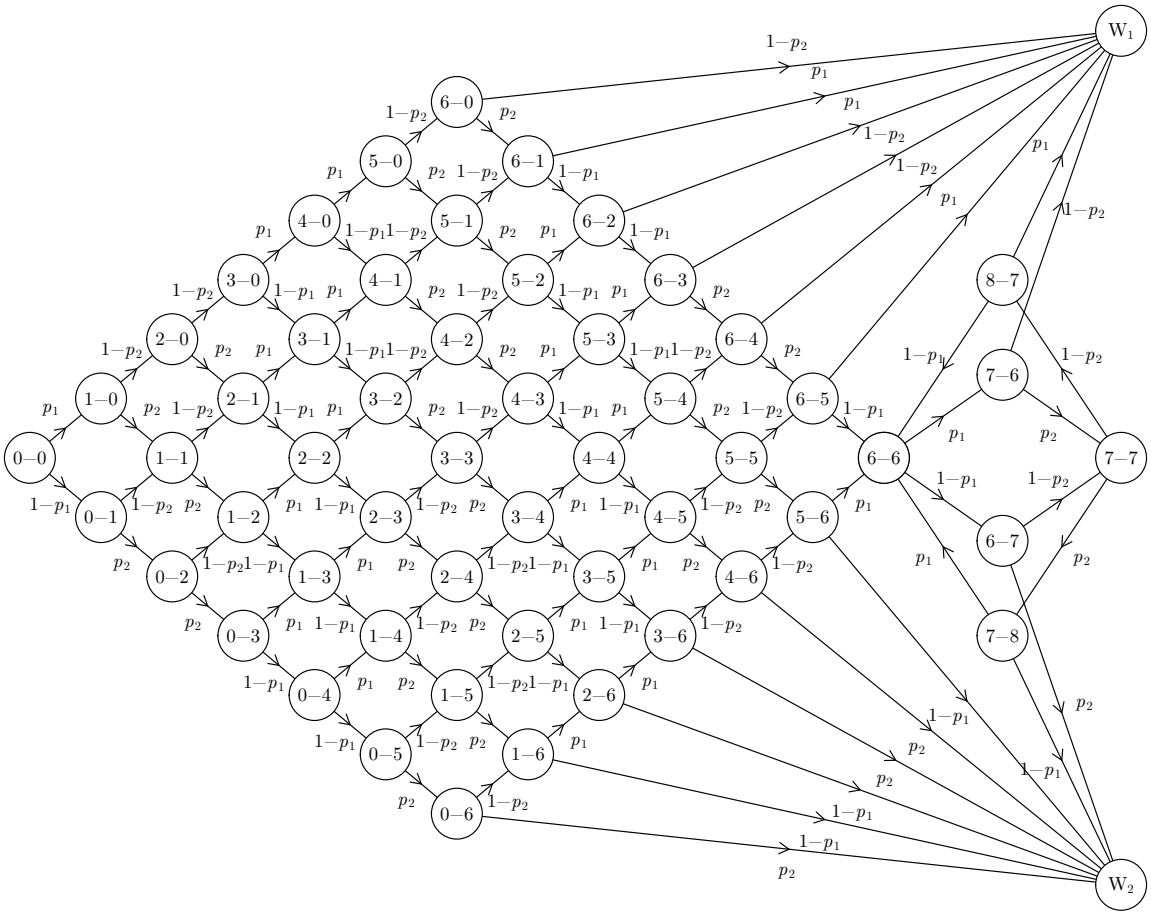

Figure 2.3.5: A Markov chain representing a tie-break in a tennis match. State W_i represents player i having won the tie-break, and the probability of player i winning a point on serve is p_i for $i = 1, 2$.

The assumptions of independence and identical distribution are important ones, and greatly ease the calculation of in-play probabilities. However, there are legitimate causes for questioning both. For example, independence may be violated if players let losing a point negatively (or indeed, positively) impact the next point. Similarly, it has been hypothesised that players react differently to the most important points in the match - some may thrive, while others crumble under the pressure, leading to different probabilities of winning depending on the current score. A study of these assumptions was conducted by Klaassen and Magnus (2001) using a binary panel data

method.

The study examined independence by considering whether the winner of one point affected the probability that each player won the next point (after controlling for player quality). Identical distribution was tested for by considering the importance of the point, defined as how much a player's probability of the match is affected by the next point, assuming the iid model is correct. It has been suggested that players may perform worse under at important points due to the pressure caused by the high stakes involved. The importance of points will be discussed in greater detail in Section 2.3.3.

The study rejects both the hypotheses of independence and identical distribution, suggesting dependence between successive points and a reduction in server quality at important points. However, they also suggest that, based on other results in the paper, that making these assumptions in forecasting tennis matches is "relatively harmless", even in-play, given that the divergence from the iid assumptions is only small. Given that these assumptions allow us to build a Markov chain framework, rendering the calculating of in-play probabilities much easier, the trade-off seems reasonable, as long as it is acknowledged that this may lead to some minor discrepancies.

There are some ways in which the study could be updated and improved upon. The study was conducted using four years' of Wimbledon data from 1992-1995. The authors admit that it is unclear whether the results are generalisable to other surfaces, and the divergences from the iid assumption may also have changed in the 25 years since. One difficulty the authors note is the availability of point-by-point data. In the years that have passed since, this may have become less of an issue. For example, the data available at `github.com/JeffSackmann` appears to be a promising source of point-by-point data for Grand Slams from 2011-2018.

Additionally, there may be scope to measure player quality in other ways. The definition used was $R_i = 8 - \log_2(r_i)$, as in Klaassen and Magnus (2003), where r_i denots a player's world ranking. Our literature has discussed many different ways of

measuring a player's quality to ranking based methods, and an alternative measure may lead to different results. Finally, the only form of independence considered is the dependence of one point on the next. While this effect may be small, the effect of longer term dependence is not studied, such as whether the winner of a point affects the winner of the second or third next point to be played, or whether losing multiple points can dent a player's confidence further down the line. It is possible that while the effect from one point to the next is small, dependence may build up over time in a manner not captured by considering adjacent points.

The work of Klaassen and Magnus (2001) therefore represents an important study of the assumptions that points are independent and identically distributed, and their findings give us some confidence that these assumptions are reasonable in the context of making in-play predictions for the results of tennis matches. However, with appropriate point-by-point data there is scope to further test some of the effects on more modern data, which may shed additional light on the validity of the iid assumptions.

2.3.2 Estimating Point-Win Probabilities

This hierachical Markov chain model gives a framework for estimating probabilities of players winning a match given the players' probabilities of winning points. However, in order to apply this in practice, one obviously needs good estimates for these probabilities of winning points.

One simplification that can be made is to assume a relationship between p_{ij} and p_{ji}. It has been noticed by several authors, for example Klaassen and Magnus (2003), that the average probability of the two players winning a point on serve, $\mu_{ij} := \frac{1}{2}(p_{ij} + p_{ji})$, is not very informative about the pre-match probability of players winning points. Instead, it is mainly the difference between these probabilities, $\lambda_{ij} := \frac{1}{2}\big(p_{ij} - p_{ji}\big)$, that is informative about match-win probabilities. This can be seen in Figure 2.3.6. Using the same four years of Wimbledon data, Klaassen and Magnus (2003) and Magnus and Klaassen (1999) both suggested the average point-win probability was around

64.5% for men and around 56% for women. These therefore provide sensible values to use for μ_{ij}, particularly on grass courts, the surface used at Wimbledon.

In Chapter 3, we shall prove that for fixed μ_{ij}, the pre-match win probability

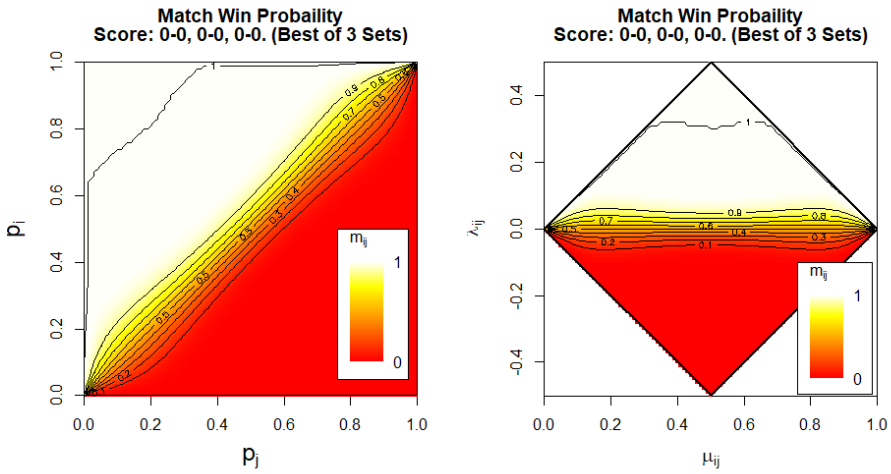

Figure 2.3.6: The probability m_{ij} of player i winning a 3-set tennis match compared with the parameters p_i and p_j (left) and μ_{ij} and λ_{ij} (right) at the start of a tennis match. (Each player has 0 sets, 0 games and 0 points).

function $m(\lambda_{ij}|\mu_{ij}, \boldsymbol{s}, b) := P(i \text{ wins}|\mu_{ij}, \lambda_{ij}, \boldsymbol{s}, b)$ is invertible, for all scores $\boldsymbol{s}$ in a best-of-b-sets match. This means that if the pre-match probability of i winning the match is known, this specifies unique λ_{ij} that can be used to estimate the probability i wins from all other scorelines.

We must be careful about using this method in-play though, when μ_{ij} can affect match-win probabilities, especially if one player is winning near the end of a match. This effect can be seen in Figure 2.3.7. If μ_{ij} is high, and hence p_{ij} and p_{ji} are also high, then player who is losing is unlikely to obtain the break of serve required to level the scores. On the other hand, if μ_{ij} is low, there is a much greater chance of serve being broken, giving the losing player a better chance of catching up.

Barnett and Clarke (2005) uses data from the ATP website to find a players' career averages of points won and lost on serve. (At the time, return averages had to

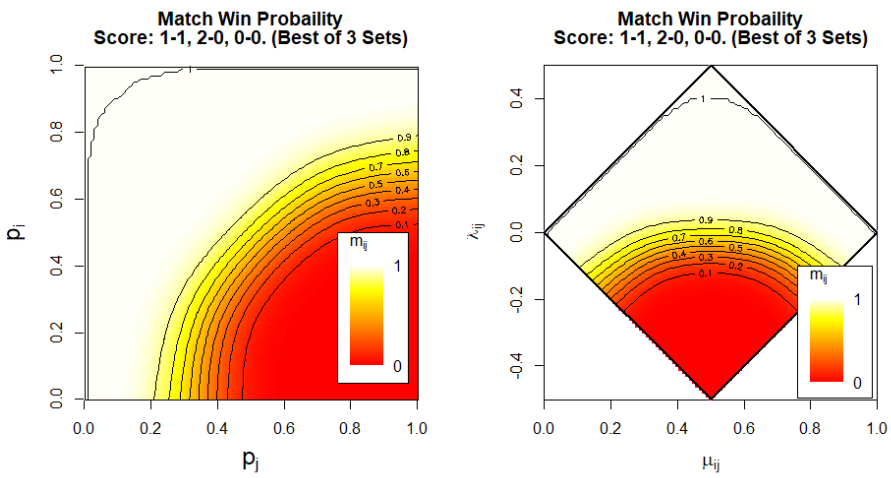

Figure 2.3.7: The probability m_{ij} of player i winning a 3-set tennis match compared with the parameters p_i and p_j (left) and μ_{ij} and λ_{ij} (right) when each player has 1 set, player i is winning the final set by 2 games to 0, and the points score is 0-0.

be estimated from a combination of relevant data and some minor assumptions , but extra data now available on the website means players' career average for points won on return is now also readily available.)

Each player i has parameters $\bar{p}_i$ and $\bar{q}_i$, which respectively denote proportion of points won on serve and return across all matches against all opponents their whole career. To instead estimate p_{ij} - the probability player i wins a point on serve against player j - the quality of player i's serve and player j's return must be taken into account. Even a weak server may win many points on serve against a player with an even worse return. Surface may also play an important factor in the amount of points players win on serve.

In order to account for the above when modelling a match between two players, Barnett and Clarke (2005) begin by taking μ_{ij} as the average points won on serve and return for all players at the same tournament the previous year, $\bar{p}_t$ and $\bar{q}_t = 1 - \bar{p}_t$. This provides a baseline mode which should account for differences in surface, as well as any differences that may caused by the level of the tournament (for example,

whether it's a Grand Slam or a lower level event). For each of player i and j, the values $\bar{p}_t$ and $\bar{q}_t$ are modified by how much better players i and j are than the average player, where average return and serve percentages for all players are denoted by p_{av} and q_{av}. The formulae for this are

$$p_{ij} = \bar{p}_t + (\bar{p}_i - p_{av}) - (\bar{q}_j - q_{av}),$$
$$q_{ij} = \bar{q}_t + (\bar{q}_i - q_{av}) - (\bar{p}_j - p_{av}).$$

Note that since $\bar{p}_t + \bar{q}_t = 1$ by definition, the essential property that $p_{ij} + q_{ji} = 1$ (and vice versa) is also achieved. On a technical level, it may be worth noting that there is no mechanism to ensure that all of the estimated probabilities lie in [0,1], though the statistics available for each player are typically similar enough to each other that this is not an issue.

One drawback is in the implicit assumption that players' average points won and lost on serve are directly comparable with other players'. Weak players may win many matches at low-level tournaments, but consistently get knocked out by the first opponent they face in large tournaments. Stronger players, on the other hand, may play in fewer small tournaments and advance further in larger tournaments, only to lose to opposition of a far higher calibre than the weaker player ever encountered. The fact that strong players play against stronger opponents than weaker players mean comparing points won and lost on serve between the two may not always be valid. A player's career average of points won and lost on serve is also not ideal, as it gives equal weight to older data and to newer, more relevant data.

A slightly different approach is taken by Knottenbelt et al. (2012). To counteract the fact that the average opponent for a given pair of players may be quite different (since stronger players will play more matches against strong players), Knottenbelt et al. (2012) choose to compare pairs of players by only considering other players that both have played recently. These are the players' "common opponents". To avoid the problem of mixing older, irrelevant data with newer data, only the two players' last 50 matches on the appropriate surface are used.

For a common opponent k, the proportion of points each of player i and j won or lost against k is found. Let $\hat{p}_{ik}$ be the observed proportion of points won on serve by player i against player k, and so on for other probabilities. Let $\Delta_{ij|k}$ be the estimate for the superiority of player i over player j given how well they both did against player j. This is defined as

$$\Delta_{ij|k} = (\hat{p}_{ik} - \hat{p}_{jk}) - (\hat{q}_{jk} - \hat{q}_{ik}).$$

Whereas Barnett and Clarke (2005) decided to use the previous year's tournament's average, p_t, as a baseline value, Knottenbelt et al. (2012) use 0.6, but apply it in a different manner. To obtain point-win probabilities for player i, Knottenbelt et al. (2012) considers that $\Delta_{ij|k}$ could be added either to the baseline serve probability, 0.6, or the baseline return probability, 0.4. The approach taken is to average the match-win probabilities obtained by doing both, so that

$$\begin{aligned} p^{(1)}_{ij|k} &:= 0.6 + \Delta_{ij|k}, & q^{(1)} &:= 0.4, \\ p^{(2)}_{ij|k} &:= 0.6, & q^{(2)} &:= 0.4 + \Delta_{ij|k}, \\ m_{ij|k} &:= \frac{1}{2}\big(m(p^{(1)}_{ij|k}, 1 - q^{(1)}, b) + m(p^{(2)}_{ij|k}, 1 - q^{(2)}, b)\big). \end{aligned}$$

To link this back to our earlier notation of μ_{ij} and λ_{ij}, we would let $\lambda^{(n)}_{ij|k} = \frac{1}{2}(p^{(n)}_{ij|k} - p^{(n)}_{ji|k})$ and $\mu^{(n)}_{ij|k} = \frac{1}{2}(p^{(n)}_{ij|k} + p^{(n)}_{ji|k})$, for $n = 1, 2$. This would give $\lambda^{(n)}_{ij|k} = \frac{1}{2}\Delta_{ij|k}$, for $n = 1$ and $n = 2$, while $\mu^{(1)}_{ij|k} = 0.6 + \frac{1}{2}\Delta_{ij|k}$ and $\mu^{(2)}_{ij|k} = 0.6 - \frac{1}{2}\Delta_{ij|k}$.

If players i and j have n_{ij} common opponents, this process is repeated for all common opponents k to get n_{ij} predictions for the match. The model estimate for the probability player i beats player j, which we call $\bar{m}_{ij}$, is the average of all of these predictions, so that

$$\bar{m}_{ij} = \sum_{k=1}^{n_{ij}} m_{ij|k}.$$

The main example described by Knottenbelt et al. (2012) is a match between Vania King and Greta Arn. Bookmakers gave King an implied probability of 48%

of winning the match, suggesting the two players were well-matched. However, for the players' ten common opponents, the values of $m_{ij|k}$ ranged between 0.62% and 99.99%. Of the predictions, only 6 were between 1% and 99%, with none between 30% and 70%. The average was 59%.

No comment is offered on why the probabilities are so extreme. However, in the example provided of King and Arn's common opponent, an examination of the serve and return proportions against there common opponents provided suggests that extreme probabilities can easily arise when one player performs well against a common opponent and the other performs poorly.

With such a disparate range of probabilities provided by these $m_{ij|k}$, it is surprising that they can be relied upon to consistently provide reasonable average predictions for the matches. The authors note the extreme probabilities, and as a result suggest that predictions made with a small number of common opponents should be treated with caution, though they believe this is typically not an issue in matches between active professional players. However, they were only able to predict 1228 of the matches (56.5%) they considered if the common opponents needed the same surface, or 1873 (86.2%) if not, due to a lack of common opponents. This suggests a lack of common opponents is frequently an issue. This may be acceptable if one wishes simply to make a positive return on betting, but it presents a serious obstacle if one wishes to make predictions for all matches, as we would need in attempting to detect match-fixing.

This model was tested with a simple betting strategy, and obtained returns of 3.8% over 2173 matches in one year if keeping the same surface, or 3.41% using any surface. At different men's and women's Grand Slam tournaments in 2011, betting returns ranged from -23.94% to 32.50%, suggesting returns are very variable over modest numbers of matches. Each Grand Slam tournament has 127 matches. McHale and Morton (2011) sound a note of caution in using betting returns as a measure of model performance, observing that betting strategy and shopping around for favourable odds can be as important for obtaining positive returns as the qual-

ity of model predictions. Nevertheless, the positive returns obtained suggests that the model has promise, though its inability to consistently provide predictions on all matches limits its usefulness in detecting match-fixing.

2.3.3 Variations to the IID Markov Chain Model

A few authors have suggested alterations to this basic Markov chain model. However, none that we have found do it with the specific goal of improving in-play forecasting. Some focus on pre-match forecasting, while others explore the assumptions of independence and identical distribution in general rather than the specifics of applying it to forecasting individual matches.

Some papers aim to approximate the idea of momentum in a match - the idea that a player that performs better than expected earlier on will continue to do so later in the match. Barnett et al. (2006) suggest that a player that has won more sets than their opponent increases their chance of winning future sets by a small amount $\alpha > 0$ common to all matches. This does not affect pre-match probabilities, and appears to improve the fit for the lengths of sets and matches, though the effect on in-play probabilities is not examined.

Meanwhile, Madurska (2012) attempts to achieve a similar effect by looking at how individual player's scores in a first set affected their second-set scores, and so on. For example, if a player wins their first set 7-6, do they tend to win the second set more convincingly, or crumble under the pressure? What if they won 6-0 instead? To fit this for a given player, Madurska (2012) looks at a player's past matches to find the average maximum likelihood estimator for λ in set $n+1$, conditional on each possible scoreline in set n. In a match between two players, the authors then lay out a procedure for how to combine the conditional maximum likelihood estimates for these players in a way that takes into account both of their past behaviours, in the hope that how the players reacted to winning or losing sets in past matches can be used to make predictions about how those players might react in future matches.

To ensure the matches considered are relevant to the current time, only a player's last 50 matches are used to fit the model. However, since there are 14 unique possible scorelines in a set, each is observed infrequently in those 50 matches, making estimation of the sizes of these effects difficult. In addition, when looking at past data, no alteration is made based on the strength of the opponent. This means it is not clear how much of the dependence between scores in successive sets is simply down to the quality of the opponent, rather than a player-specific reaction to adversity or triumph. The authors do, however, note some improvements in predictive performance when using this model compared with using iid Markov chains and the common-opponent model of Knottenbelt et al. (2012).

Other papers attempt to include the idea of the importance of certain points in the match. This is the notion that at key points which can shift the balance of the whole match, probabilities might be different to typical points. This is potentially due to players trying harder on these points, as they realise the huge benefit that can be gained at these key moments. It seems sensible that a player who is serving at 30-40 to prevent a break of service in a tight match will see more benefit in winning that point than a player returning while losing 40-0 in a match they are already losing heavily. However, while the importance of points has been studied, very little has been done to fit a predictive model that incorporates this idea.

Morris (1977) was the first to introduce a mathematical definition of the importance of a point. Let $G_i(p_s, s|a, a_i)$ denote the probability player i wins a game given the server is player $s \in \{i, j\}$, and has point-win probability p_s, (so that the returning player has point-win probability $1-p_s$), and players i and j have won a_i and a_j points in this game respectively. The importance of a given point to a typical game to player i is then defined as

$$I_i^P(a_i, a_j|p_s, s) = G_i(p_s, s|a_i + 1, a_j) - G_i(p_s, s|a_i, a_j + 1).$$

This is difference in the probability of winning the game depending on whether player i wins or loses the current point. If the probability of winning the game

after winning the point is much greater than if the point is lost, then the point is important due to its huge value, whereas if the probability of winning the game is largely unaffected by who wins the next point, then it is unimportant.

The importance of a game to a typical set and the importance of the set to a match are defined similarly, given the current score of b_i and b_j games apiece in the current set, and c_i and c_j sets each. To find the importance of a point to the match as a whole, the relevant point, game and set probabilities are multiplied, yielding the definition

$$\mathcal{I}_i(a_i, a_j, b_i, b_j, c_i, c_j | p_i, p_j, s) = I_i^P(a_i, a_j | p_s, s) I_i^G(b_i, b_j | p_i, p_j)) I_i^S(c_i, c_j | p_i, p_j)$$

Morris (1977) explores a number of properties of this definition. Importance is always positive, and each point is equally important to each player. Morris (1977) also finds that the most and least important point in each game are 30-40 and 40-0 respectively, for typical point-win probability p_i=0.64. This leads to the interesting result that if a player manages to increase their probability of winning the most important point in each game by some small $\epsilon > 0$, and decrease their probability of winning the least important point by the same ϵ, then their probability of winning the match increases. Of course if the most important point generally occurred more frequently in the match than the least important, then it would be possible that this result was simply due to the average probability of player i winning a point becoming higher than p_i (since the probability player i won a point was more likely to be $p_i + \epsilon$ than $p_i - \epsilon$). Morris (1977) account for this by multiplying the increases and decreases in point-win probability by the expected frequency of the most and least important points in a game respectively, so that the average point-win probability remains at p_i, and find that the results still hold.

Newton and Aslam (2006) and Viney (2015) both do more work on this idea, exploring how much the value of ϵ affects match-win probabilities though they do not account for how often the most and least important points occur.

Though these results are interesting from a strategy perspective, they do not seem

as applicable to fitting a model. Should both players attempt to raise their point-win probability in this manner, what would happen? A model would be required to establish which players are better able to raise their probabilities of winning points.

Klaassen and Magnus (2001)'s findings suggest that at important points, it is generally the returning player that manages to increase their point-win probability, with the server's probability decreasing. More importantly, they also find that this effect is weaker among stronger players: that the very best players can avoid ceding too much advantage on the most important points. However, their focus is on investigating the iid hypothesis rather than predictive modelling, and thus there is no further discussion on how to fit such a predictive model. To do so, one would probably either require player-specific parameters that describe reaction to important points, or at the very least a model common to all players that describes the relationship between players' strength and their reaction to important points.

One paper that attempts to provide a predictive model that touches on the idea of important points is that of Carrari et al. (2017). They choose to relax the assumption that each player i has point-winning probability p_{ij} while serving throughout the match against player j. Instead, the player normally wins a point while serving with probability p_{ij}, or with an adjusted probability $\tilde{p}_{ij}$ if the score has reached 30-30 or later. A Markov chain representing this system is shown in Figure 2.3.8. This change represents how players might behave differently under pressure at the end of a close game, and bears many similarities with the idea of importance as defined by Morris (1977). Recall that 30-30 can be considered to be the same state as a deuce, as explained in Section 2.3. This model shall henceforth be referred to as the "deuce model".

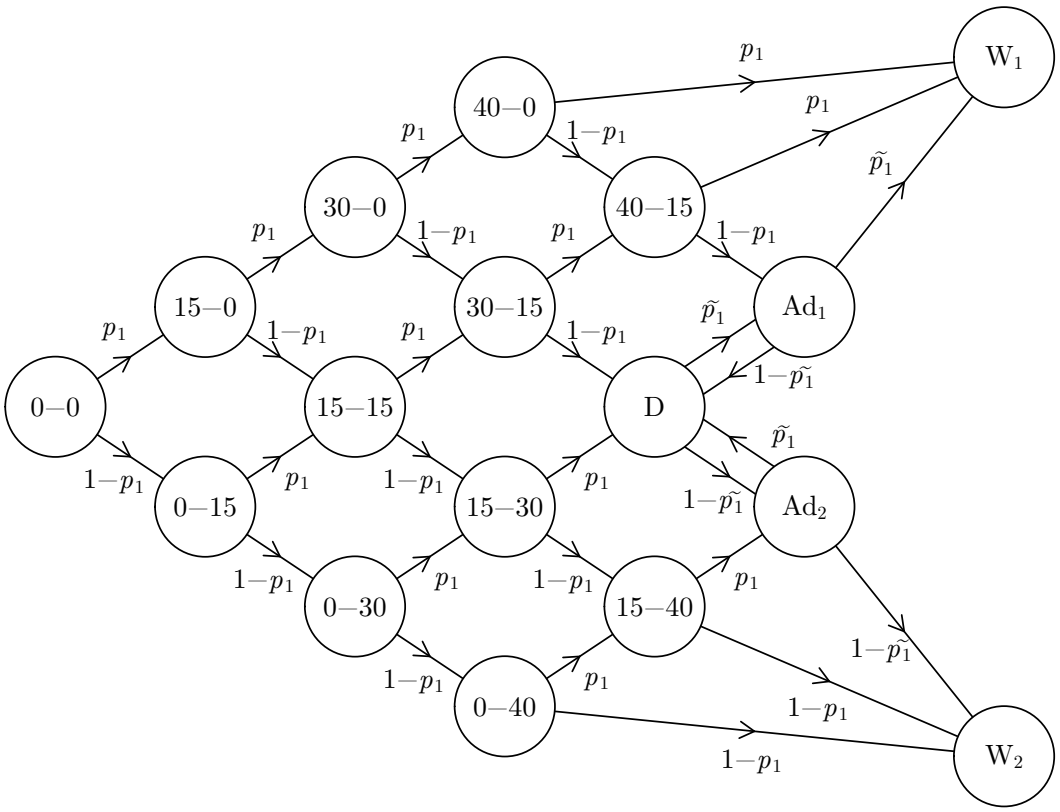

Figure 2.3.8: A Markov chain representing a game of tennis. States W_i represents player i having won the game, and Ad_i denotes advantage to player i. The probability of player 1 winning a point on serve is p_1 before both players have reached 30 points, and $\tilde{p}_1$ after.

To fit this model for a match between a given pair of players, only matches between those two players in the past are counted. The probabilities p_{ij} and $\tilde{p}_{ij}$ are estimated using the historical proportion between the players with data available from tennisearth.com, which provides point-by-point updates of ATP and WTA matches up to 2014.

Carrari et al. (2017) fit and test this model using matches between just three tennis players - Djokovoic, Federer and Nadal - as these three have played a large number of matches between each other, and estimates using a small number of matches are unreliable. These three have played a total of 57 matches in the dataset. It would

also appear to be helpful that these three players have remained broadly consistent in quality for around ten years. If one were to take a pair of players whose rankings have fluctuated more wildly over the years, then averaging across their whole career might not be as sensible.

On this small sample, the results section notes that the deuce model produces game-win probabilities closer to the observed game-win proportions for 5 out of the 6 possible combinations of each of the three players serving to another other. They also compare average game lengths with expected game lengths. Expected game length is different depending on whether the server or returner wins the game. With 3 players, there are therefore 12 possible combinations of server, returner and game winner, and the deuce model is found to produce closer expected game lengths to the observed average in 9 of these cases. These results are promising, but it seems as if a more formal likelihood-based approach would provide significantly more evidence for these claims.

The model has yet to be extended to predicting outcomes between all possible pairs of players. One of the previously discussed strategies for estimating point-win probabilities could probably be adapted to this setting using data from `tennisearth.com`, splitting points into those before and after 30-30. However, many of the problems discussed with previous methods (such as ensuring old data does not skew averages, while gathering enough data to get meaningful results) would persist. Moreover, having to estimate two probabilities per player, p_{ij} and $\tilde{p}_{ij}$, would only leave around half as much data to estimate each. It is therefore not yet known whether these results hold over a wider range of players.

While the idea of the deuce model remains an intriguing and promising idea, the current lack of a large-scale set of results and the challenges involved in fitting the model mean it is not an idea explored further in this thesis. Our own results, in general, are of sufficient quality that the need to investigate this idea further is not pressing.

2.4 A Comparative Study - Kovalchik (2016)

In order to compare some of the different predictive models available, Kovalchik (2016) took several different models and applied them to ATP data from the 2014 season. The author considered several of the regression-based models considered in Section 2.2.1, the paired comparison methods of McHale and Morton (2011) and Morris and Bialik (2015) discussed in Sections 2.2.2 and 2.2.3, and a few different Markov chain models described in Section 2.3. These models were also compared with a model that took the average of several bookmaker's odds in order to compare these model's performance to market predictions.

The models were all trained using data from the 2013 season with the aim of providing a fair comparison between models. For Markov chain based models, in which player strengths are essentially estimated using an average of recent form, there is a challenge involved in choosing enough data that a representative picture of a player's abilities can be created without including data from so long ago that it no longer reflects current form. For regression based methods that use player's ranks and the like as covariates, information is pooled from across all different matches in the training data, irrespective of the identities of the individual players. As such, even though the use of more data may be relevant and useful, it may also be true that one year's data is sufficient to reliably estimate patterns across all tennis players.

However, the paired comparison methods of McHale and Morton (2011) and Morris and Bialik (2015) are designed to include more data from player's entire careers, and so using a single year's data does not represent best practice. They are designed to focus on recent matches but also use information from older matches. Because of this, the method of McHale and Morton (2011) was also tested using two years of data, and the method of Morris and Bialik (2015) was tested using players' entire career histories. Both models were also tested using just a single year's training data to provide a fair comparison with the other methods that used one year's data.

The most accurate of all models was the bookmaker consensus model, indicating

that the odds remain one of the most reliable predictors of tennis matches. One implication of this for our work is that if we wish to identify suspicious matches by comparing odds to match-win probability estimates, there will be matches when differences occur simply because the odds provide a better reflection of player strength than our models. This is typically because they can utilise information that is difficult to incorporate into models, such as injury news. As such, not all differences between odds and predictions are suspicious.

Of the tennis models used, the paired comparison method of Morris and Bialik (2015) with entire career histories was the best by every performance criteria where objective ranking was possible. The precise ranking of the next few models depended on the performance criteria used, but some of the stronger models that came the closest were the Markov chain-based method of Barnett and Clarke (2005), several of the regression-based methods, and Morris and Bialik (2015)'s model using just one year of data.

Kovalchik (2016) draws several conclusions from their results. First and foremost, they suggest that Elo-based can be very strong in modelling tennis matches. The ability to incorporate career histories can give them an edge over other models, but are not necessary to provide comparable performance. The best Markov chain-based method, used by Barnett and Clarke (2005), adjusted the point-win probability on serve based on the quality of their opponent, which appeared to give a significant advantage, as one might expect. Among regression-based methods, using player rankings as a predictor provided the best performance, but the use of further predictors provided no tangible benefit. Interestingly, all of the methods provided better predictions for high-ranked players than low-ranked players. The author does not hypothesise why, but possible explanations may be a greater availability of data, and greater consistency in performance, for high-ranked players.

2.5 Summary

This literature review has focused on two main areas, match-fixing and tennis modelling. Our match-fixing review began with a broad discussion of the tools available to detect match-fixing. The main idea underpinning all statistical investigations was to construct models for the usual behaviour of data in clean matches, analyse the data to see if it conforms to expectations, before carefully examining the possible causes of any differences.

Odds data are one promising source of information about potential match-fixing. When matches are clean, the odds for a given player or team in a betting market should reflect the perceived probability that that player wins. As such, the probabilities inferred from the odds should be close to good predictions of the probability that that player or team wins. In a fixed match, the money wagered by the fixed match may swing the market, causing observable differences that we seek to identify.

Several papers analysed betting markets to find evidence of suspicious odds movements, but almost all focussed on the pre-match markets. Some works focussed on the differences between the closing pre-match odds and model predictions in their respective sports, such as the work of Reade and Akie (2013) Reade (2014) and Ötting et al. (2018) in football, as these should be similar in efficient markets. Ötting et al. (2018) also analyses betting volumes in football, as fixed matches may see significantly higher betting volumes. A similar approach could work in tennis, and could represent an interesting alternative avenue of work to the research into odds we present, but it we chose not to pursue it any further.

Other works focus instead on swings in pre-match markets, which suggest a sustained pattern of gambling contrary to bookmakers' original expectations - a possible sign of match fixing. Rodenberg and Feustel (2014) and Blake and Templon (2016) analyse swings in pre-match markets in tennis matches, while Feustel and Rodenberg (2015) considered the same issue football matches.

Focussing only on the swing in pre-match markets neglects the fact that odds may

naturally move very rapidly when betting volume is low on betting exchanges, as the first tentative gambles are made before the market settles. As such, we hope to use time-stamped pre-match market data with volume information to ignore these early fluctuations, and instead analyse large swings once the market has settled.

Another form of pre-match market analysis is the study of point-spread markets in basketball, with a comparative multitude of papers discussing this issue. In contrast to the other analysis of betting markets we have considered, these attempted to estimate the rate of match-fixing by analysing the shape of the distribution of the differences between observed margins of victories and points spreads. The challenge lies in describing the behaviour of this distribution in both the presence and absence of match-fixing, with different authors hotly disputing the evidence available.

However, apart from the work of Forrest and McHale (2019), there appears to be no significant discussion of the detection of in-play match-fixing in any sport. This appears to be a significant oversight, given how large the in-play markets in sport have grown. Forrest and McHale (2019) focusses on a general discussion of the issue with some examples, but is unable to disclose the proprietary algorithms used to analyse market anomalies. The analysis of in-play markets in tennis therefore represents a significant potential research area, as so little analysis in the area has been conducted, and the potential for finding corrupt activity is strong.

We also considered a few works which analysed factors other than betting markets. Rodenberg and Feustel (2014) suggested low player effort in early rounds of tennis tournaments could be a sign of match-fixing. However, the focus is on the rate of match-fixing rather than identifying individual matches. We therefore chose not to pursue this avenue further, as we felt that the analysis of betting markets could more easily provide evidence of potential match-fixing by individuals, so as to give the best chance of removing fixers from the sport. Deutscher et al. (2017) analysed whether the choice of referee impacted the average number of goals in German football matches, suggesting that the market for the number of goals may be fixed, while

Duggan and Levitt (2000) examined whether sumo wrestlers may fix their final bouts in tournaments when only one player had any meaningful incentive to win. These last two papers do not easily translate to a tennis context, and so need no further direct consideration in this thesis. However, all three of these works are illustrative of the wider spectrum of methods available to match-fixing investigators, who need not necessarily rely on betting markets alone to identify corruption.

The second literature review concerned tennis modelling. We considered a range of models, with some focussing on pre-match predictions while others some focussed on predictions in-play.

Kovalchik (2016) performed the most comprehensive study of tennis models to data, considering a range of regression based models, Markov chain models and paired comparison models. Morris and Bialik (2015)'s method provided the strongest predictive accuracy, partly due to its ability to incorporate more data from players' career histories, though it still performed well using just a single year of data. This suggests strong potential for other Elo-based methods that can also use players' career histories. The authors note the existence of other Elo-based methods such as the Glicko ratings, and suggested further work may include adapting such methods to a tennis context.

In a match-fixing context, Glicko ratings have the intriguing property of being able to assign different variances to the ratings of different players depending on the availability recent information. This in turn can be used to assign uncertainty to the estimated probability of each player winning. In a match-fixing context, this gives a statistical basis for assessing how much variance in the win probabilities implied by the odds is usual, based on the probabilistic distributions of win probability provided by the Glicko ratings. Additionally, if a player has returned from a long injury, it may be difficult to estimate their strength, and thus a correspondingly large amount of uncertainty may be assigned to the estimate of their strength. DW on Sport (2016) note that some large odds swings can be explained by players returning from a long

injury. Glicko ratings therefore seem a strong candidate for use in our work, providing they can be adapted appropriately to a tennis context.

In-play modelling in tennis is centred around the assumption that points are independent and identically distributed. Under the assumption of independence, it is easy to construct a Markov chain representing a tennis match, in which each state represents a score. Klaassen and Magnus (2001) performed a key study on the validity of these assumptions and found evidence for dependence between successive points and different behaviour at important points in the match. However, they note that the impact of assuming that points are iid is small, and makes little practical difference for in-play forecasts of tennis matches.

Other studies have considered whether predictive performance can be included by relaxing these assumptions. Morris (1977), Newton and Aslam (2006) and Viney (2015) studied the importance of points without attempting to use it in predictive modelling. Madurska (2012) attempted to model the dependence of successive set scores, and Carrari et al. (2017) considered modelling points after a deuce differently to points before. While both present intriguing ideas, we felt that both of these methods required more robust testing before being put to use, and hence we felt it best to use the assumptions of independence and identical distribution in all of our work to follow. There is, however, substantial scope to further research the iid assumptions and ways to improve in-play tennis modelling, especially given that point-by-point data are more readily available than when Klaassen and Magnus (2001) performed their original study.

Chapter 3

Proofs of Results About Tennis Match Markov Chains

In Section 2.3 we discussed Markov chains representing tennis matches. By assuming that the outcomes of all points in a match are independent and identically distributed, and representing each possible score as a state in a Markov chain, it is possible to estimate the probability of each player winning the match given the probabilities, p_1 and p_2, that each of the two players wins a point on serve. Section 2.3.2 discussed the common simplification of reparameterising the model so that $p_1 = \mu+\lambda$ and $p_2 = \mu-\lambda$, since for fixed λ, the value μ has little impact on the pre-match probability of either player's victory.

Later work in this thesis requires that the function $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible in λ on the interval (0,1) for all μ, $\boldsymbol{s}$ and b, where $m(\lambda|\mu, \boldsymbol{s}, b)$ is the probability that player 1 wins a tennis match given μ, λ and the current score is $\boldsymbol{s}$ in a best of b-sets match. A sufficient condition for a function to be invertible is that it is continuous and increasing in λ. Some papers in the literature perform numerical inversion of $m(\lambda|\mu, \boldsymbol{s}, b)$ without proving the existence of an inverse - for example, see Klaassen and Magnus (2003). A proof that this is possible does not appear to be available. Numerical inversion of this function has led to no practical problems, mainly due to

the fact that the function is indeed invertible. However, we prove this result and present it here for a few reasons. Firstly, the existence of a proof is reassuring to those uncomfortable inverting a function without proving the existence of an inverse. Additionally, the proof is interesting in itself, and it adds to the current literature on modelling matches in tennis and other sports.

This chapter therefore details a proof of this for a general class of Markov chains using inductive arguments, and we then use these results to show that the proof also holds for games, tie-breaks, sets and matches, thus enabling us to prove the general result that $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible for use elsewhere in this thesis.

Proving that $m(\lambda|\mu, \boldsymbol{s}, b)$ is increasing in λ is essentially proving that the higher a player's probability of winning a point on serve is than their opponent's, the more likely they are to win the match. In order to prove this, we shall also prove that under the Markov chain model, winning a point always improves a player's chance of winning the match. While both of these facts appear obvious in a tennis context, proving their truth mathematically under the Markov chain model is not trivial, and poses some interesting challenges.

3.1 Continuity of $m(\lambda|\mu, \boldsymbol{s}, b)$

To prove that $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible, we must first prove that it is continuous. In order to do this, we recall that

$$m(\lambda|\mu, \boldsymbol{s}, b) = m(p_1, p_2|\boldsymbol{s}, b) := P(1 \text{ wins match } |p_1 = \mu + \lambda,\ p_2 = \mu - \lambda,\ \boldsymbol{s}, b).$$

We must therefore prove that $m(p_1, p_2|\boldsymbol{s}, b)$ is continuous in μ and λ.

Recall from the earlier equation (2.3.1) that $m(p_1, p_2|\boldsymbol{s}, b)$ can be broken down if

players 1 and 2 are currently playing a game during a match as

$$\begin{aligned} P(1 \text{ wins match}|\boldsymbol{s}, b) = & \qquad (3.1.1) \\ & P(1 \text{ wins game})P(1 \text{ wins set}|1 \text{ wins game})P(1 \text{ wins match}|1 \text{ wins set}) \\ & +P(1 \text{ wins game})P(1 \text{ loses set}|1 \text{ wins game})P(1 \text{ wins match}|1 \text{ loses set}) \\ & +P(1 \text{ loses game})P(1 \text{ wins set}|1 \text{ loses game})P(1 \text{ wins match}|1 \text{ wins set}) \\ & +P(1 \text{ loses game})P(1 \text{ loses set}|1 \text{ loses game})P(1 \text{ wins match}|1 \text{ loses set}), \end{aligned}$$

or similarly if the players are playing a tie-break. Each of these sub-probabilities can be calculated by using the appropriate Markov chain from Section 2.3 for games, sets, tie-breaks or matches. Section 2.3 described how if the appropriate Markov chain has transition matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I} \end{pmatrix},$$

then a matrix $\mathbf{B}$ giving absorption probabilities from each state to the states in which each player won the contest is given by

$$\mathbf{B} = (\mathbf{I} - \mathbf{Q})^{-1}\mathbf{R}. \qquad (3.1.2)$$

Since all probabilities in $\mathbf{P}$ are continuous functions of p_1 and p_2, the absorption probabilities in $\mathbf{B}$ are also continuous functions of p_1 and p_2, due to being compositions of continuous functions. Inputting these into equation (3.1.1), we can therefore see that $m(p_1, p_2|\boldsymbol{s}, b)$ is continuous in both p_1 and p_2, and hence $m(\lambda|\mu, \boldsymbol{s}, b)$ is continuous in λ for all μ, $\boldsymbol{s}$ and b.

3.2 Monotonicity of $m(\lambda|\mu, \boldsymbol{s}, b)$

In order to prove that $m(\lambda|\mu, \boldsymbol{s}, b)$ is increasing in λ, we first note by the chain rule that

$$\begin{aligned}\frac{dm(\lambda|\mu, \boldsymbol{s}, b)}{d\lambda} &= \frac{dm(p_1, p_2|\boldsymbol{s}, b)}{dp_1}\frac{dp_1}{d\lambda} + \frac{dm(p_1, p_2|\boldsymbol{s}, b)}{dp_2}\frac{dp_2}{d\lambda} \\ &= \frac{dm(p_1, p_2|\boldsymbol{s}, b)}{dp_1} - \frac{dm(p_1, p_2|\boldsymbol{s}, b)}{dp_2}.\end{aligned}$$

To prove that $\frac{dm(\lambda|\mu,\boldsymbol{s},b)}{d\lambda} > 0$, it is sufficient to prove that $\frac{dm(p_1,p_2|\boldsymbol{s},b)}{dp_1} > 0$ and $\frac{dm(p_1,p_2|\boldsymbol{s},b)}{dp_2} \leq 0$. We shall only prove that $\frac{dm(p_1,p_2|\boldsymbol{s},b)}{dp_1} > 0$ - the proof that $\frac{dm(p_1,p_2|\boldsymbol{s},b)}{dp_2} < 0$ follows similarly.

In order to prove this, we first formalise the concepts in equation (3.1.1). Let $\boldsymbol{s} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$, be the current score, where $\boldsymbol{x}_1$ is a vector of the number of sets each player is on, $\boldsymbol{x}_2$ the number of games and $\boldsymbol{x}_3$ the number of points. We then let $m_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ be the probability player 1 wins the match from score $\boldsymbol{s} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$, dropping the dependence on p_1 and p_2 for brevity. Then equation (3.1.1) becomes

$$\begin{aligned}m_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3) = {} & g_1(\boldsymbol{x}_3)\; s_1(\boldsymbol{x}_2 + (1,0), \mathbf{0})\; m_1(\boldsymbol{x}_1 + (1,0), \mathbf{0}, \mathbf{0}) \\ & + g_1(\boldsymbol{x}_3)\big(1 - s_1(\boldsymbol{x}_2 + (1,0), \mathbf{0})\big)\; m_1(\boldsymbol{x}_1 + (0,1), \mathbf{0}, \mathbf{0}) \\ & + \big(1 - g_1(\boldsymbol{x}_3)\big) s_1(\boldsymbol{x}_2 + (0,1), \mathbf{0})\; m_1(\boldsymbol{x}_1 + (1,0), \mathbf{0}, \mathbf{0}) \\ & + \big(1 - g_1(\boldsymbol{x}_3)\big)\big(1 - s_1(\boldsymbol{x}_2 + (0,1), \mathbf{0})\big)\; m_1(\boldsymbol{x}_1 + (0,1), \mathbf{0}, \mathbf{0}),\end{aligned}$$

where $\mathbf{0} = (0, 0)$. Note also that each of these probabilities also depends on the player that is currently serving, but without loss of generality we can assume that player 1 is serving, and so we also drop this notation for brevity.

We wish to prove that

$$\frac{dm_1(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)}{dp_1} > 0 \text{ for all } \boldsymbol{s} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3).$$

Taking this derivative and grouping terms, we see that

$$\begin{aligned}\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} = {} & \frac{dg_1(\boldsymbol{x}_3)}{dp_1}\Big(s_1(\boldsymbol{x}_2+(1,0),\mathbf{0}) - s_1(\boldsymbol{x}_2+(0,1),\mathbf{0})\Big)\times \\ & \qquad \Big(m_1(\boldsymbol{x}_1+(1,0),\mathbf{0},\mathbf{0}) - m_1(\boldsymbol{x}_1+(0,1),\mathbf{0},\mathbf{0})\Big) \\ & + \frac{ds_1(\boldsymbol{x}_2+(1,0),\mathbf{0})}{dp_1} g(\boldsymbol{x}_3)\Big(m_1(\boldsymbol{x}_1+(1,0),\mathbf{0},\mathbf{0}) - m_1(\boldsymbol{x}_1+(0,1),\mathbf{0},\mathbf{0})\Big) \\ & + \frac{ds_1(\boldsymbol{x}_2+(0,1),\mathbf{0})}{dp_1}(1-g(\boldsymbol{x}_3))\Big(m_1(\boldsymbol{x}_1+(1,0),\mathbf{0},\mathbf{0}) - m_1(\boldsymbol{x}_1+(0,1),\mathbf{0},\mathbf{0})\Big) \\ & + \frac{dm_1(\boldsymbol{x}_1+(1,0),\mathbf{0},\mathbf{0})}{dp_1}\Big(g_1(\boldsymbol{x}_3)\ s_1(\boldsymbol{x}_2+(1,0),\mathbf{0}) + \\ & \qquad \big(1-g_1(\boldsymbol{x}_3)\big)s_1(\boldsymbol{x}_2+(0,1),\mathbf{0})\Big) \\ & + \frac{dm_1(\boldsymbol{x}_1+(0,1),\mathbf{0},\mathbf{0})}{dp_1}\Big(g_1(\boldsymbol{x}_3)\ \big(1-s_1(\boldsymbol{x}_2+(1,0),\mathbf{0})\big) + \\ & \qquad \big(1-g_1(\boldsymbol{x}_3)\big)\big(1-s_1(\boldsymbol{x}_2+(0,1),\mathbf{0})\big)\Big).\end{aligned}$$

Many of these terms are probabilities, and so are positive. There are a few exceptions, and so to prove that $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$, we also need to prove that the following are all positive:

$$\frac{dg_1(\boldsymbol{x}_3)}{dp_1} > 0 \text{ for all } \boldsymbol{x}_3. \tag{3.2.1}$$

$$\frac{ds_1(\boldsymbol{x}_2,\mathbf{0})}{dp_1} > 0 \text{ for all } \boldsymbol{x}_2. \tag{3.2.2}$$

$$\frac{dm_1(\boldsymbol{x}_1,\mathbf{0},\mathbf{0})}{dp_1} > 0 \text{ for all } \boldsymbol{x}_1. \tag{3.2.3}$$

$$s_1(\boldsymbol{x}_2+(1,0),\mathbf{0}) - s_1(\boldsymbol{x}_2+(0,1),\mathbf{0}) > 0 \text{ for all } \boldsymbol{x}_2. \tag{3.2.4}$$

$$m_1(\boldsymbol{x}_1+(1,0),\mathbf{0},\mathbf{0}) - m_1(\boldsymbol{x}_1+(0,1),\mathbf{0},\mathbf{0}) > 0 \text{ for all } \boldsymbol{x}_1. \tag{3.2.5}$$

If the match is in a tie-break instead of a game, it is easy to show that this simply requires the extra property

$$\frac{dt_1(\boldsymbol{x}_3)}{dp_1} > 0 \text{ for all } \boldsymbol{x}_3. \tag{3.2.6}$$

In order to prove these, we will show that the Markov chains for games, sets, matches (and tie-breaks) fall into a general category of Markov chains and prove general results

about all Markov chains in this class. This will help us obtain the desired results, proving that $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$ for all $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$.

3.3 First to $(M+1, N+1)$ Markov Chain

Suppose that there is a discrete-time Markov chain in which $M \times N$ of the states are co-ordinates on a two-dimensional finite discrete grid, (m, n). These co-ordinates can be thought to represent the number of points that two players have scored respectively, and M and N are the largest number of points each player can win without winning the contest. If a player 1 reaches $M+1$ points while player 2 has less than N points, they win the contest, represented by state W_1. Similarly, if player 2 reaches $N+1$ points while player 2 has less than M points, they win the contest, represented by state W_2. Other states may exist, but they can only be reached from the state (M, N). From state (M, N), we assume that another point is played, and so the Markov chain reaches either state $(M+1, N)$ or state $(M, N+1)$. These states may simply be W_1 and W_2 respectively, as in the left-hand plot in Figure 3.3.1, or they may be distinct states, from which yet other states may be reached, as in the right-hand plot. This permits different behaviour in the Markov chain if after $M+N$ points no player has won.

At each time τ, the probability each player wins a point is a function of τ, as well as two parameters representing the strength of each player, α_1 and α_2. If the contest starts at time τ_0, then time τ is always equal to $\tau_0 + m + n$. Without loss of generality, we will use $\tau_0 = 0$. Importantly, the probabilities of winning points do not depend on m or n. We define the transition probabilities more formally in Section 3.3.1, state two theorems about the probability of player 1 winning the contest in Section 3.3.2 and use them to prove $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$ in Section 3.4 before proving the theorems in Section 3.6.

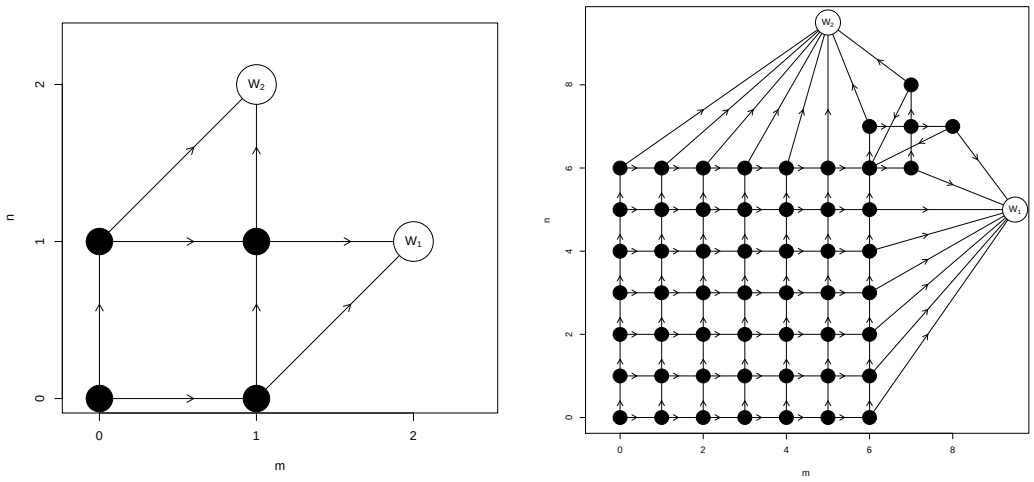

Figure 3.3.1: Two examples of "First to $(M+1, N+1)$" Markov chains.

3.3.1 Transition Probabilities

Let $P(x, y)$ denote the transition probability of the Markov chain from state x to state y. If $m < M$ and $n < N$, the transition probabilities are given by

$$\begin{aligned} P\Big((m,n),(m+1,n)\Big) &= q(\tau = m+n, \alpha_1, \alpha_2), \\ P\Big((m,n),(m,n+1)\Big) &= 1 - q(\tau = m+n, \alpha_1, \alpha_2), \\ P\big((m,n),x\big) &= 0 \text{ for all other states } x. \end{aligned}$$

If $m = M$ and $n < N$, then

$$\begin{aligned} P\Big((M,n),W_1\Big) &= q(\tau = M+n, \alpha_1, \alpha_2) \\ P\Big((M,n),(M,n+1)\Big) &= 1 - q(\tau = M+n, \alpha_1, \alpha_2) \\ P\big((M,n),x\big) &= 0 \text{ for all other states } x. \end{aligned}$$

If $m < M$ and $n = N$, then

$$\begin{aligned} P\Big((m,N),(m+1,N)\Big) &= q(\tau = m+N, \alpha_1, \alpha_2), \\ P\Big((m,N),W_2\Big) &= 1 - q(\tau = m+N, \alpha_1, \alpha_2), \\ P\big((m,N),x\big) &= 0 \text{ for all other states } x. \end{aligned}$$

States W_1 and W_2 are absorbing states, that is

$$\begin{aligned} P(W_1, W_1) &= 1, \\ P(W_1, x) &= 0 \text{ for all other states } x \\ \text{and} \quad P(W_2, W_2) &= 1, \\ P(W_2, x) &= 0 \text{ for all other states } x. \end{aligned}$$

From state (M, N), we assumed that the Markov chain can only reach two states $(M+1, N)$ and $(M, N+1)$. In some cases, these states may be simply be W_1 and W_2 respectively, but in other cases they may be distinct new states. As with other states of type (m, n), we assume that the transition probabilities from (M, N) are

$$\begin{aligned} P\Big((M,N),(M+1,N)\Big) &= q(\tau = M+N, \alpha_1, \alpha_2), \\ P\Big((M,N),(M,N+1)\Big) &= 1 - q(\tau = M+N, \alpha_1, \alpha_2), \\ P\big((M,N),x\big) &= 0 \text{ for all other states } x. \end{aligned}$$

We also make the following assumptions about $q(\tau, \alpha_1, \alpha_2)$:

$$\begin{aligned} 0 < q(\tau, \alpha_1, \alpha_2) &< 1, \\ \frac{dq(\tau, \alpha_1, \alpha_2)}{d\alpha_1} &\geq 0, \\ \frac{dq(\tau, \alpha_1, \alpha_2)}{d\alpha_2} &\leq 0. \end{aligned} \tag{3.3.1}$$

The first condition is partly in place to ensure that $q(\tau, \alpha_1, \alpha_2)$ is a probability, but note the inclusion of strict inequalities rather than permitting equality. The proofs that follow have not been completed in cases where $q(\tau, \alpha_1, \alpha_2) = 0$ or 1 for some values of τ. The second and third conditions ensure that player i's probability of winning a point is increasing in α_i. For convenience, we tend to drop the dependence on α_1 and α_2 and simply write $q(\tau, \alpha_1, \alpha_2) = q_\tau$.

In order to prove the desired results, we will need to consider the absorption probabilities of the Markov chain into state W_1, that is the probability that player

1 wins the contest. We let $Q(m,n)$ denote the probability that player 1 wins the contest from state (m,n), with the additional property that $Q(M+1,n) = 1$ for $n < N$ and $Q(m,N+1) = 0$ for $m < M$. These additional properties are required since technically the state $(M+1,n)$ and $(m,N+1)$ do not exist. In this chapter we will use the important property that this probability can be found by conditioning on the winner of the next point, giving

$$Q(m,n) = q_{m+n}Q(m+1,n) + (1-q_{m+n})Q(m,n+1).$$

3.3.2 Statement of Two Theorems on First to $(M+1, N+1)$ Markov chains

We define two assertions about the properties of $Q(m,n)$, which we call $A(m,n)$ and $B(m,n)$. These are defined as

$$A(m,n) : Q(m+1,n) > Q(m,n+1), \tag{3.3.2}$$

$$B(m,n) : \frac{dQ(m,n)}{d\alpha_1} > 0. \tag{3.3.3}$$

$A(m,n)$ means that from score (m,n), a player 1's chance of winning the contest will be better if they win the next point than if they lose. $B(m,n)$ means that at score (m,n), a player 1's chance of winning the contest gets higher as their chance of winning the point increases (since q_{m+n} is increasing in α_1). These are very intuitive ideas in a tennis context, but require some effort to prove in "First to $(M+1,N+1)$" Markov chains.

We define two theorems and prove that they are true for all "First to $(M+1,N+1)$" Markov chains. We will then discuss how this relates to tennis matches.

Theorem 3.3.1. *If $A(M,N)$ is true, then $A(m,n)$ is true for all $m \leq M$ and all $n \leq N$.*

Theorem 3.3.2. *If $B(M,N)$ is true, then $B(m,n)$ is true for all $m \leq M$ and all $n \leq N$.*

These are proven at the end of this chapter in Section 3.6. In the meantime, we use these theorems to prove equations (3.2.1) to (3.2.6) and hence that $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$.

3.4 Applying Theorems 3.3.1 and 3.3.2 to Tennis

Having proven these results, we wish to make use of them to prove that each of equations (3.2.1) to (3.2.6) are true, hence proving the desired result, $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$ for all $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$.

3.4.1 Sets and Matches

A Markov chain for a match in which the states are the number of sets each player has won is a very simple example of a "First to $(M+1, N+1)$" Markov chain. In a best-of-b sets match that includes a final set tie-break, a player can win $(b-1)/2$ sets without winning, and player 1 wins sets with probability s_1, where $s_2 = 1 - s_1$. Figure 3.4.1 shows plots of the relevant Markov chains. In summary,

- $M = N = (b-1)/2$.
- $\alpha_1 = p_1$, $\alpha_2 = p_2$.
- $q(\tau, \alpha_1, \alpha_2) = s(p_1, p_2)$ for all τ.

This satisfies all of the conditions for the Markov chain, provided the conditions on q_τ in equations (3.3.1) are satisfied, namely $0 < s(p_1, p_2) < 1$, (which simply requires $0 < p_1 < 1$ and $0 < p_2 < 1$) and $\frac{ds(p_1,p_2)}{dp_1} > 0$. This is equivalent to proving equation (3.2.2).

Under the assumption that equation (3.2.2) holds, then in order to prove that $A(m,n)$ and $B(m,n)$ are true everywhere in this Markov chain, we must simply prove that they are true at (M,N). In any tennis match, the winner of the final set wins the match. To prove $A(M,N)$, we note that $Q(M+1,N) = 1$ and $Q(M,N+1) = 0$,

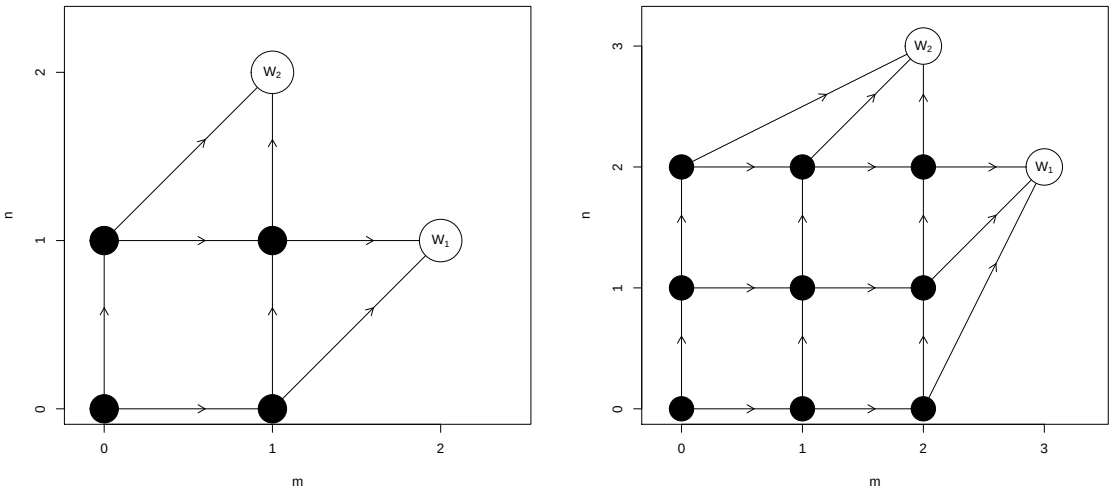

Figure 3.4.1: Markov chains representing 3-set and 5-set tennis matches.

and since $1 > 0$, $A(M, N)$ is true. Hence $A(m, n)$ would be true for all m and n, proving equation (3.2.5). To prove $B(M, N)$, we note that $Q(M, N) = s(p_1, p_2)$, and then so that $\frac{ds(p_1,p_2)}{dp_1} > 0$, which again requires equation (3.2.2) to be true, and which shall be proven in Section 3.4.2. (Proving that this also holds in matches without a final set tie-break is trivial.) This would prove that $B(m, n)$ is true for all m and all n, which is the property required in equation (3.2.3).

In summary, if equation (3.2.2) is true, then equations (3.2.3) and (3.2.5) are also true.

3.4.2 Games and Sets

A set of tennis can be shown to fall into the class of "First to $(M+1, N+1)$" Markov chains by observing Figure 3.4.2 and defining the parameters of the Markov chain as follows:

- $M = N = 5$.
- $\alpha_1 = p_1, \alpha_2 = p_2$.

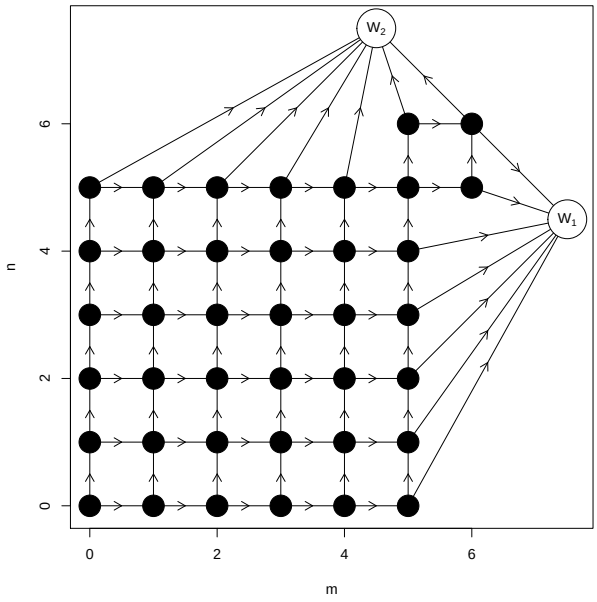

Figure 3.4.2: A Markov chain representing a set of a tennis match.

- $q(\tau, \alpha_1, \alpha_2) = g(p_1) := g_1$ if τ is even and less than 12.
- $q(\tau, \alpha_1, \alpha_2) = 1 - g(p_2) := 1 - g_2$ if τ is odd and less than 12.
- $q(\tau, \alpha_1, \alpha_2) = t(p_1, p_2)$ if $\tau = 12$.

From state (5,5), an additional two games are played to see if either player can win by two clear games - if not, a tie-break is played.

Observe that $q(\tau, \alpha_1, \alpha_2)$ satisfies the conditions in equations (3.3.1) for all τ provided that $\frac{dg(p_1)}{dp_1} > 0$ and $\frac{dt(p_1,p_2)}{dp_1} > 0$, which are equations (3.2.1) and (3.2.6) respectively, and shall be proven in Sections 3.4.3 and 3.4.4 respectively.

To prove $A(m, n)$ and $B(m, n)$ are true for states before (5,5), we must only prove that they are true at (5,5). However, the Markov chain from (5,5) is also a "First to $(M + 1, N + 1)$" Markov chain. Hence proving that $A(6, 6)$ and $B(6, 6)$ are true is sufficient to prove $A(m, n)$ and $B(m, n)$ are true for $5 \leq m \leq 6$ and $5 \leq n \leq 6$. Since this would mean that $A(5, 5)$ and $B(5, 5)$ are true, this would mean that $A(m, n)$ and

$B(m, n)$ are also true for all other states in the Markov chain.

It is easy to prove $A(6, 6)$ is true, since $Q(7, 6) = 1$ and $Q(6, 7) = 0$. Since $Q(6, 6) = t(p_1, p_2)$, then $B(6, 6)$ is true if we can prove that $\frac{dt(p_1,p_2)}{dp_1} > 0$. This is simply equation (3.2.6), and shall be proven in Section 3.4.3. Assuming this is true, however, $B(m, n)$ is true for all m and n in the Markov chain, which is equivalent to proving equation (3.2.2).

Hence provided equations (3.2.1) and (3.2.6) are true, then equations (3.2.2) and (3.2.4) are true, and hence so are equations (3.2.3) and (3.2.5).

3.4.3 Points and Tie-Breaks

Similarly, a tennis tie-break also falls into this class of "First to $(M + 1, N + 1)$" Markov chains. We use the information from Figure 3.4.3 and say that

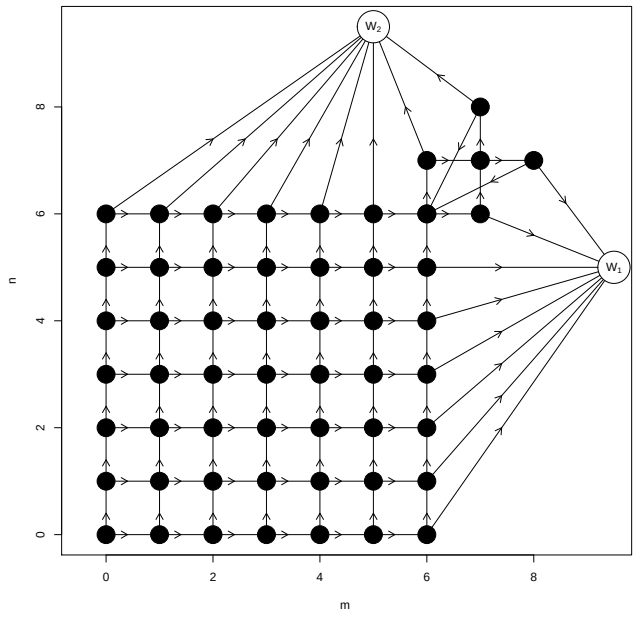

Figure 3.4.3: A Markov chain representing a tie-break in a tennis match.

- $M = N = 6$.

- $\alpha_1 = p_1$, $\alpha_2 = p_2$.
- $q(\tau, p_1, p_2) = p_1$ if $\tau \bmod 4 = 0$ or 1.
- $q(\tau, p_1, p_2) = 1 - p_2$ if $\tau \bmod 4 = 2$ or 3.

The transition probabilities $q(\tau, p_1, p_2)$ follow this pattern because the order in which players 1 and 2 serve in a tie-break is (1,2,2,1), which is repeated throughout the tie-break. From (6,6) onwards, players must be two points ahead of their opponents to win the tie-break. Therefore if at (8,7) player 1 loses the point, or if player 1 wins the point at (7,8), the Markov chain returns to state (6,6) as the serving cycle starts again. Note that state (7,7) is different to state (6,6) as player 2 serves at the former, and player 1 serves at the latter.

To prove that $A(m,n)$ and $B(m,n)$ hold for every m and every n, we must first prove it individually for every state with $m \geq 6$ and $n \geq 6$ - proving the statements for $m = n = 6$ then implies that the statements are true for all $m \leq 6$ and $n \leq 6$.

To find $Q(m,n)$ for each $m \geq 6$ and $n \geq 6$ we look at the Markov chain from state (6,6) onwards. To find $Q(m,n)$ for the states in this Markov chain, we solve the recurrence relation

$$
\begin{aligned}
Q(6,6) &= p_1 Q(7,6) + (1-p_1) Q(6,7) \\
Q(6,7) &= (1-p_2) Q(7,7) \\
Q(7,6) &= (1-p_2) \cdot 1 + p_2 Q(7,7) \\
Q(7,7) &= (1-p_2) Q(8,7) + (p_2) Q(7,8) \\
Q(7,8) &= p_1 Q(7,7) \\
Q(8,7) &= p_1 \cdot 1 + (1-p_2) Q(6,6)
\end{aligned}
$$

This is easy enough to do by hand, giving

$$
\begin{aligned}
Q(6,6) = Q(7,7) &= \frac{p_1(1-p_2)}{p_1(1-p_2)+p_2(1-p_1)} \\
Q(6,7) &= (1-p_2)Q(6,6) \\
Q(7,6) &= 1-p_2+p_2Q(6,6) \\
Q(7,8) &= p_1Q(6,6) \\
Q(8,7) &= p_1+(1-p_1)Q(6,6).
\end{aligned}
$$

The properties $A(m,n)$ and $B(m,n)$ can then be proven easily for each individual state with $m \geq 6$ and $n \geq 6$, thus proving that they hold everywhere in the Markov chain so that $t(p_1, p_2, \boldsymbol{x}_3)$ is increasing in p_1 and decreasing in p_2 for all scores $\boldsymbol{x}_3$. This proves equation (3.2.6).

3.4.4 Points and Games

Under the assumption that player 1 is serving, a Markov chain representing a game of tennis, as shown in Figure 3.4.4 is given by

- $M = N = 3$.
- $\alpha_1 = p_1$, $\alpha_2 = 1 - p_1$.
- $q(\tau, p_1, p_2) = p_1$.

From state (3,3), (which is more properly called 40-40 or deuce in tennis), players must win two consecutive points to win the game. The Markov chain therefore moves to state (4,3) (or advantage for player 1) with probability p_1, and thence on to W_1 with probability p_1 or back to deuce with probability $1-p_1$. Similarly, player 2 must win a point with probability $1-p_1$ to reach (3,4) and then win again to win the game, or else return to (3,3).

To prove $A(m,n)$ and $B(m,n)$ are true for $m \geq 3$ and $n \geq 3$), we find $Q(m,n)$ for

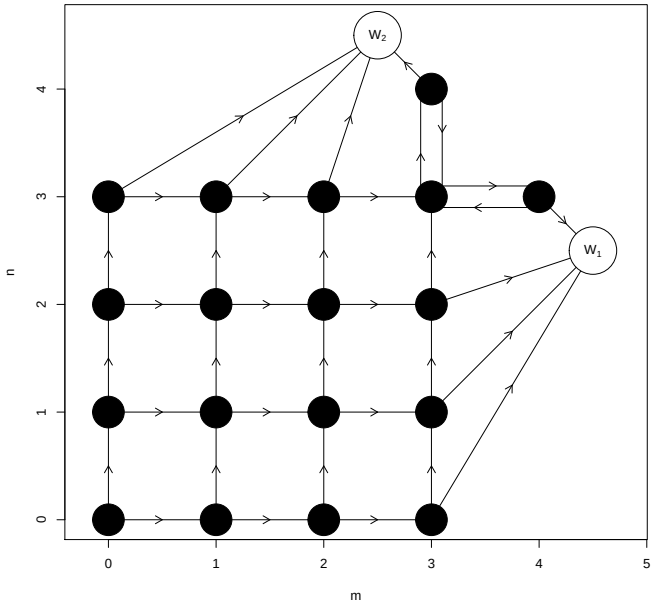

Figure 3.4.4: A Markov chain representing a game of a tennis match.

each such state. Doing so proves that

$$Q(3,3) = \frac{p_1^2}{p^2 + (1-p_1)^2}$$
$$Q(4,3) = p_1 + (1-p_1)Q(3,3),$$
$$Q(3,4) = p_1 Q(3,3).$$

The properties $A(m,n)$ and $B(m,n)$ can then be proven for each of these states individually, and thus since they hold at (M,N), they hold for all $m \leq M$ and $n \leq N$ too. Hence $g(\boldsymbol{x}_3)$ is strictly increasing in p_1 for all scores $\boldsymbol{x}_3$. This proves equation (3.2.1).

3.4.5 Summarising Remarks

To prove that Since equations (3.2.1) and (3.2.6) are true, then all of equations (3.2.2) to (3.2.5) must also be true. This is exactly what was required to prove that $\frac{dm_1(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3)}{dp_1} > 0$ for all $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, which in turn was required to prove

that $m(\lambda|\mu, \boldsymbol{s}, b)$ is increasing in λ for all μ, $\boldsymbol{s}$ and b, thus proving that the function $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible. We can therefore reasonably use this inverse in the rest of this thesis, having proven its existence. We shall invert this function numerically, due to the complexity of this inverse function.

3.5 Further Work

While we have proven everything we wish to in Markov chains related to tennis, we suspect that the results described also hold for more general topologies on a grid than simple $M \times N$ rectangles, should one wish to explore other Markov chains. This could help write a more general proof that our results hold for games, sets, matches and tie-breaks without having to deal with the behaviour after state (M, N) individually for each contest, but it could also be useful to extend the proofs Markov chains for applications outside of tennis.

Suppose instead that for each n, the largest value that m can take without player 1 winning the match a function of n, which we call $M(n)$. Similarly, the largest value n can take without winning the match $N(m)$, a function of m. This would mean that the state (m, n) is only part of the Markov chain if $m \leq M(n)$ and $n \leq M(n)$. An example of such a grid is shown in Figure 3.5.1. A set in a tennis match would be another such example, with $N(m) = 5$ if $n \leq 4$, or else $N(m) = 6$, and $M(n) = 5$ if $m \leq 4$, or else $M(n) = 6$ - see Figure 3.4.2. The conditions on q would be as before, and the new transition probabilities would be

$$
\begin{aligned}
P\Big((m,n),(m+1,n)\Big) &= q(m+n,p_1,p_2), \quad \text{for } m < M(n),\\
P\Big((m,n),(m,n+1)\Big) &= 1-q(m+n,p_1,p_2), \quad \text{for } n < N(m),\\
P\Big((M(n),n),W_1\Big) &= q(m+n,p_1,p_2),\\
P\Big((m,N(m)),W_2\Big) &= 1-q(m+n,p_1,p_2),\\
P\Big((m,n),(x,y)\Big) &= 0.
\end{aligned}
$$

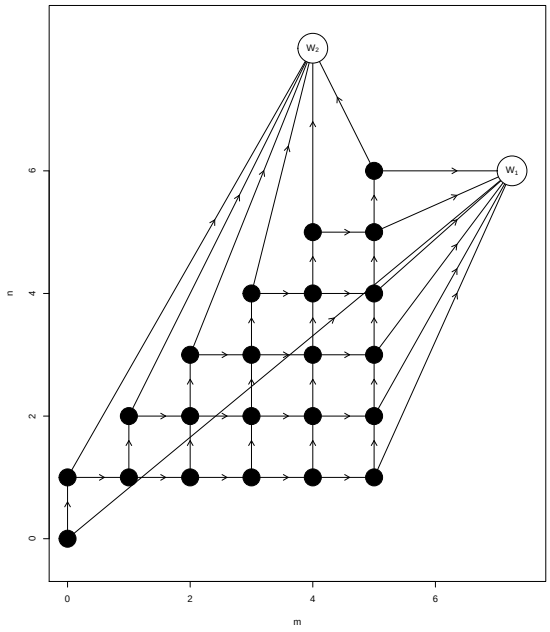

Figure 3.5.1: A Markov chain in which $N(m) = m + 1$ for $m \leq 5$ and $M(n) = 0$ if $n = 0$, or else $M(n) = 5$ for $1 \leq n \leq 6$.

For each location (m, n), proofs of the statements $A(m, n)$ and $B(m, n)$ relies only on the proofs for the statements at $(m, n + 1)$ and $(m + 1, n)$, as well as the upper and right-hand edges of the rectangle. It stands to reason that for Markov chains on square grids of other shapes that, provided the statements hold everywhere on the Pareto boundary of the grid - that is, for the locations $(m, N(m))$ for all m and $(M(n), n)$ for all n - that the double backwards induction would still be possible.

However, note that the proofs that $A(m + 1, n)$ and $A(m, n + 1)$ imply $A(m, n)$ rely on $Q(m+1, n+1)$ being well defined. For this reason, we would require $N(m)$ to be non-decreasing in m, and $M(n)$ to be non-decreasing in n. Consider for example the state (2,1) in the Markov chain in Figure 3.5.2, in which $q_\tau = p_1$ for all τ. Even though $A(3, 1)$ and $A(2, 2)$ may be true, we see that $Q(3, 1) = p_1^4$ and $Q(2, 2) = p_1$, which is higher, hence $A(2, 1)$ is untrue. This is because $Q(3, 2)$ not well defined, due to state (3,2) not existing. It is assumed in the proof that the probability of

winning the contest after winning one point and losing one point is the same, yet this probability is 0 if one wins a point and then loses a point, or 1 if a point is lost and then won. It would therefore be advantageous to lose the point at (2,1). It would be interesting to explore further what shape Markov chains are permitted to ensure that properties $A(m, n)$ and $B(m, n)$ hold everywhere in the Markov chain.

Another interesting point of exploration would be relaxing the fact that $P((m, n), (m+$

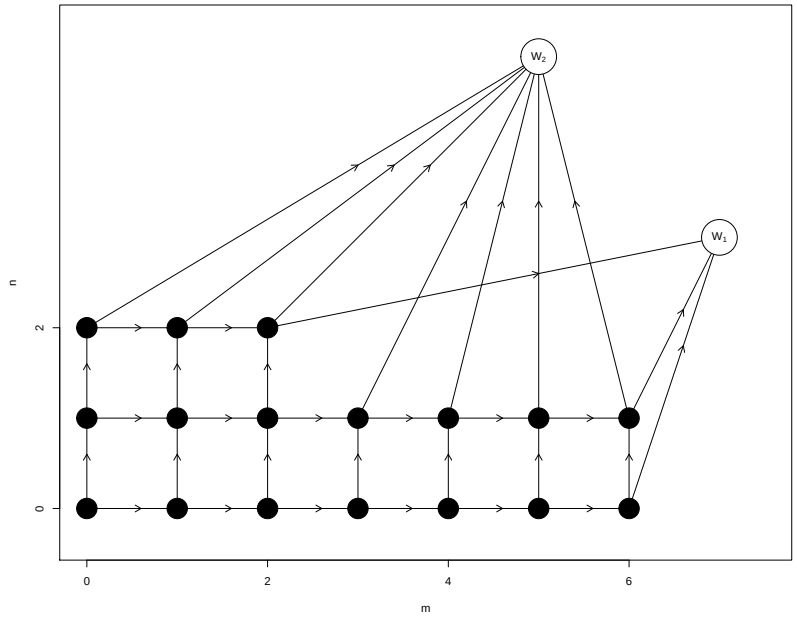

Figure 3.5.2: A Markov chain in which $A(2, 1)$ would not hold.

$1, n))$ must only a function of time, $\tau = m + n$. This was to ensure that if a player starts winning points, their probabilities of winning points do not decrease. Were this not true, it could be possible that losing a few points could be beneficial in order to gain large increases in the probability of winning future points. However, we speculate that as long $P((m, n), (m + 1, n)$ is non-decreasing in m, the proofs should still hold. This could be helpful if we wished to implement a model in which a tennis player gained momentum in a game if they took an early lead and their chances of winning points improved.

3.6 Proof of Theorems 3.3.1 and 3.3.1

3.6.1 Proof of Theorem 3.3.1

We will prove that if $A(M, N)$ is true then $A(m, n)$ is true for all $m \leq M$ and $n \leq N$ by double backwards induction. In order to prove a statement by double induction, we must prove each of the following:

$$A(M, n) \text{ is true for all } n \leq N. \tag{3.6.1}$$

$$A(m, N) \text{ is true for all } m \leq M. \tag{3.6.2}$$

$$\text{If } A(m, n+1) \text{ and } A(m+1, n) \text{ are true, then so is } A(m, n). \tag{3.6.3}$$

We will prove each of these in turn.

Proof that $A(M, n)$ is true for all $n < N$

We have assumed as one of the conditions of Theorem 3.3.1 that $A(M, N)$ is true. Therefore, if $A(M, n+1)$ implies that $A(M, n)$ is true for some general n, then by induction $A(M, n)$ is also true for all $n \leq N$. From point (M, n), we see that

$$Q(M+1, n) = 1.$$

We then examine $Q(M, n+1)$ by conditioning on the winner of the point at time $\tau + 1 = M + n + 1$, obtaining

$$\begin{aligned} Q(M, n+1) &= q_{\tau+1} Q(M+1, n+1) + (1 - q_{\tau+1}) Q(M, n+2) \\ &= q_{\tau+1} \cdot 1 + (1 - q_{\tau+1}) Q(M, n+2) \end{aligned}$$

If $Q(M, n+2) < Q(M+1, n+1) = 1$, which is simply the condition $A(M, n+1)$, and $0 < q_{\tau+1} < 1$, which we have assumed in equation (3.3.1), then

$$Q(M, n+1) < q_{\tau+1} + (1 - q_{\tau+1}) = 1 = Q(M+1, n),$$

and so $A(M, n+1) \Rightarrow A(M, n)$. Since $A(M, N)$ is also true, this implies $A(M, n)$ is true for all $n \leq N$.

Proof that $A(m, N)$ is true for all $m < M$

We know that $A(M, N)$ is true. Therefore, if $A(m+1, N) \Rightarrow A(m, N)$ for some general m, then $A(m, N)$ is also true for all $m < M$. From point (m, N), we see that

$$\begin{aligned} Q(m, N+1) &= 0 \\ Q(m+1, N) &= q_{\tau+1} Q(m+2, N) + (1 - q_{\tau+1}) Q(m+1, N+1) \\ &= q_{\tau+1} Q(m+2, N) + (1 - q_{\tau+1}) \cdot 0. \end{aligned}$$

The statement $A(m+1, N)$ gives us that $Q(m+2, N) > Q(m+1, N+1) = 0$, and we have assumed that $0 < q_{\tau+1} < 1$, and so

$$Q(m+1, N) > 0 = Q(M+1, n),$$

and so $A(m+1, N) \Rightarrow A(m, N)$. Since $A(M, N)$ is true by the statement of the theorem, this implies that $A(m, N)$ is true for all $m \leq M$.

Proof that if $A(m, n+1)$ and $A(m+1, n)$ are true, then so is $A(m, n)$

We begin again by looking at the values of $Q(m+1, n)$ and $Q(m, n+1)$ conditioning on the winner of the next point. Doing so gives

$$\begin{aligned} Q(m+1, n) &= q_\tau Q(m+2, n) + (1 - q_\tau) Q(m+1, n+1), \\ Q(m, n+1) &= q_\tau Q(m+1, n+1) + (1 - q_\tau) Q(m, n+2). \end{aligned}$$

Taking the difference of these two probabilities, we see that

$$\begin{aligned} Q(m+1, n) - Q(m, n+1) = q_\tau \Big(Q(m+2, n) - Q(m+1, n+1) \Big) \\ + (1 - q_\tau) \Big(Q(m+1, n+1) - Q(m, n+2) \Big). \end{aligned}$$

In assuming that $A(m, n+1)$ and $A(m+1, n)$ are true, we have assumed that $Q(m+2, n) > Q(m+1, n+1)$ and $Q(m+1, n+1) > Q(m, n+2)$. Therefore, assuming that $0 < q_\tau < 1$, then

$$Q(m+1, n) - Q(m, n+1) > 0.$$

and hence $Q(m+1,n) > Q(m,n+1)$. This proves that if $A(m,n+1)$ and $A(m+1,n)$ are true, then so is $A(m,n)$.

Having proven that each of equations (3.6.1), (3.6.2) and (3.6.3) are true, we have thus proven that $A(m,n)$ is true for all m and n, and that winning points always increases one's chances of winning the contest.

3.6.2 Proof of Theorem 3.3.2

We will again prove that $B(m,n)$ is true for all $m \leq M$ and $n \leq N$ by double backwards induction. As before, we must prove each of the following:

$$B(M,n) \text{ is true for all } n \leq N, \tag{3.6.4}$$

$$B(m,N) \text{ is true for all } m \leq M, \tag{3.6.5}$$

$$\text{If } B(m,n+1) \text{ and } B(m+1,n) \text{ are true, then so is } B(m,n). \tag{3.6.6}$$

We will prove each of these in turn.

Proof that $B(M,n)$ is true for all $n \leq N$

We have assumed as one of the conditions of Theorem 3.3.2 that $B(M,N)$ is true. Therefore, if $B(M,n+1)$ implies $B(M,n)$ for some n, then by single backwards induction, $B(M,n)$ is also true for all $n < N$. From point (M,n), we condition on the winner of the next point and see that

$$\begin{aligned} Q(M,n) &= q_\tau Q(M+1,n) + (1-q_\tau)Q(M,n+1) \\ &= q_\tau \cdot 1 + (1-q_\tau)Q(M,n+1). \end{aligned}$$

Differentiating with respect to α_1, we obtain

$$\begin{aligned} \frac{dQ(M,n)}{d\alpha_1} &= \frac{dq_\tau}{d\alpha_1} - \frac{dq_\tau}{d\alpha_1}Q(M,n+1) + (1-q_\tau)\frac{dQ(M,n+1)}{d\alpha_1} \\ &= \frac{dq_\tau}{d\alpha_1}\Big(1 - Q(M,n+1)\Big) + (1-q_\tau)\frac{dQ(M,n+1)}{d\alpha_1}. \end{aligned} \tag{3.6.7}$$

By assumption, we have that $\frac{dq_\tau}{d\alpha_1} \geq 0$, and it is clearly also true that $(1 - Q(M, n + 1)) \geq 0$, and hence the first half of equation (3.6.7) is greater than or equal to 0. We also assumed that $0 < q_\tau < 1$. If the assumption $B(M, n + 1)$ is true, then $\frac{dQ(M,n+1)}{d\alpha_1} > 0$, and the second half of equation (3.6.7) is strictly greater than 0, and hence $\frac{dQ(M,n)}{d\alpha_1} > 0$. Therefore, if $B(M, n + 1)$ is true then $B(M, n)$ is true. Since we know $B(M, N)$ to be true, then $B(M, n)$ is true for all $n \leq N$.

Proof that $B(m, N)$ is true for all $m \leq M$

We have assumed that $B(M, N)$ is true. Therefore, if $B(m + 1, N)$ implies $B(m, N)$ for some general m, then by single backwards induction, $B(m, N)$ is also true for all $m \leq M$. From point (m, N), we see that

$$\begin{aligned} Q(m, N) &= q_\tau Q(m + 1, N) + (1 - q_\tau)Q(m, N + 1) \\ &= q_\tau Q(m + 1, N) + (1 - q_\tau)0. \\ &= q_\tau Q(m + 1, N). \end{aligned}$$

Differentiating with respect to α_1, we obtain

$$\frac{dQ(m, N)}{d\alpha_1} = \frac{dq_\tau}{d\alpha_1} Q(m + 1, N) + q_\tau \frac{dQ(m + 1, N)}{d\alpha_1}. \tag{3.6.8}$$

By assumption, we have that $\frac{dq_\tau}{d\alpha_1} \geq 0.$, and it is clearly also true that $Q(m+1, N) \geq 0$, and hence the first half of equation (3.6.8) is greater than or equal to 0. We also assumed that $0 < q_\tau < 1$. If the assumption $B(m + 1, N)$ is true, then $\frac{dQ(m+1,N)}{d\alpha_1} > 0$, and the second half of equation (3.6.8) is strictly greater than 0, and hence $\frac{dQ(m,N)}{d\alpha_1} > 0$. Therefore, if $B(m + 1, N)$ is true then $B(m, N)$ is true. Since we know $B(M, N)$ to be true, then $B(m, N)$ is also true for all $m \leq M$.

Proof that if $B(m, n+1)$ and $B(m+1, n)$ are true, then so is $B(m, n)$

We begin by looking at the value of $Q(m, n)$ conditioning on the winner of the next point. Doing so gives

$$Q(m,n) = q_\tau Q(m+1,n) + (1-q_\tau)Q(m,n+1).$$

We differentiate with respect to α_1 and see that

$$\begin{aligned}\frac{dQ(m,n)}{d\alpha_1} &= q_\tau \frac{dQ(m+1,n)}{d\alpha_1} + \frac{dq_\tau}{d\alpha_1}Q(m+1,n) + (1-q_\tau)\frac{dQ(m,n+1)}{d\alpha_1} - \frac{dq_\tau}{d\alpha_1}Q(m,n+1) \\ &= q_\tau \frac{dQ(m+1,n)}{d\alpha_1} + (1-q_\tau)\frac{dQ(m,n+1)}{d\alpha_1} + \frac{dq_\tau}{d\alpha_1}\Big(Q(m+1,n) - Q(m,n+1)\Big)\end{aligned} \tag{3.6.9}$$

We have assumed that $0 < q_\tau < 1$. The assumptions $B(m+1, n)$ and $B(m, n+1)$ tell us that the two derivatives, $\frac{dQ(m+1,n)}{d\alpha_1}$ and $\frac{dQ(m,n+1)}{d\alpha_1}$, are strictly positive. Hence the sum of the first two terms in equation (3.6.9) is also strictly positive. We also assumed that $\frac{dq_\tau}{d\alpha_1} \geq 0$. In order to prove that $\big(Q(m+1,n) - Q(m,n+1)\big) > 0$, we use statement $A(m, n)$ from equation (3.3.2), which we have proven to be true. Hence $\frac{dQ(m,n)}{d\alpha_1}$ can be written as a sum of a mixture of positive and strictly positive elements, and so is also strictly positive, proving the statement $B(m, n)$ is true, provided $B(m+1, n)$ and $B(m, n+1)$ are also true.

Having now proven that each of equations (3.6.4), (3.6.5) and (3.6.6) is true, we have thus proven that the statement $B(m, n)$ is true for all $m \leq M$ and $n \leq N$, and that improving one's chances of winning points always increases one's chances of winning the contest.

Chapter 4

Glicko Ratings with an Application to Tennis

Later in this thesis we shall be attempting to identify suspicious betting activity in tennis by comparing the odds in tennis matches with predictions of the outcomes of matches. Under normal circumstances, we would expect strong agreement between the odds and model predictions. However, in matches with suspicious betting activity, differences can arise between the odds and our predictions. To look for these differences, we need to generate predictions of our own.

This chapter uses the Glicko ratings, Glickman (1999), as seen in Section 2.2.3, to generate probabilistic predictions of the outcomes of tennis matches. Among the advantages of Glicko ratings is that they are powerful, quick to implement due to the approximations involved in updating rankings. Kovalchik (2016) found the Elo-based model of Morris and Bialik (2015) to be the best among the predictive models of tennis they considered. With Glicko ratings also being essentially Elo-based, there is hope therefore that with some adaptations to suit a tennis context they may also provide a strong predictive model.

However, crucially for our purposes, Glicko ratings also allow for different players to have different uncertainties attached to their ratings. The work of DW on Sport

(2016) discussed how some of the large odds swings observed by Blake and Templon (2016) could be explained by a player returning after a long injury. We hope that using Glicko ratings would alleviate this problem, attaching less certainty to the ratings of such players and making our algorithms less likely to flag the match as suspicious.

In this chapter, we discuss the Glicko ratings described in Glickman (1999) and Section 2.2.3 of this thesis. Section 4.1 describes the basic set-up of the Glicko ratings, giving an in-depth explanation of the update steps involved in calculating the ratings. Section 4.2 then demonstrates the link between Glicko ratings and Gaussian state space models, linking the Glicko ratings to a much wider body of literature.

In the second half of this chapter, we them look to apply the Glicko ratings to our tennis data in order to be able to use them to estimate player strengths, which will help us look for anomalies in betting odds in later chapters. In Section 4.3, we extend Glicko ratings to be able to treat five-set matches differently to three-set matches, since five-set matches tend to favour the stronger player, and hence the results of five-set matches provides different information to the results of three-set matches about the quality of the players involved. We are then ready to apply the Glicko ratings to our tennis data. In Section 5.2.1, we discuss the tennis data we apply the ratings to and decide whether or not it is best to use all of our available data, before in Section 4.4.2 we look at the results of using Glicko ratings to model the strengths of players in this data and examine a few interesting features.

4.1 Glicko Model for Player Strengths

4.1.1 Glicko ratings system: basic setup

Suppose that there are N tennis players to model, and T time periods of matches are observed. The length of these time periods should be chosen appropriately depending on context - this could be in the order of days, weeks or even months. A simplifying assumption of the model is to assume all matches in a time period are assumed to occur

simultaneously. Choosing the length of the time period involves a trade-off between conflicting factors. Choosing time periods that are too long means the short-term change in players' strength is lost. On the other hand, Glickman suggests that some of the approximations in this model are more dependable for longer time periods, since more data are collected in each time period. Additionally, choosing a short time period requires more updates, and hence it can take slightly longer to optimise parameters.

Given this setup, the core assumption of the Glicko model is that each player i has strength θ_{it} during time period t. We then seek to model the probability that player i beats player j during time period t using a logistic model based on the difference in the two players' strengths. Let W_{ijt} be a random variable that takes value 1 if player i beats player j at time t, and takes value 0 otherwise. Similarly, w_{ijt} is an observation of that random variable. Under the Glicko model, the formula that describes the probability distribution of W_{ijt} is then

$$\begin{gathered} P(W_{ijt} = w_{ijt}|\theta_{it}, \theta_{jt}) = \frac{\left(e^{q(\theta_{it}-\theta_{jt})}\right)^{w_{ijt}}}{1 + e^{q(\theta_{it}-\theta_{jt})}}, \\ q = \frac{\log(10)}{400}, \\ w_{ijt} \in \{0, 1\},\ i = 1, \ldots, N,\ j = 1, \ldots, N,\ i \neq j,\quad t = 1, \ldots, T. \end{gathered} \tag{4.1.1}$$

Note that the overall analysis is unaffected by the choice of q - it merely scales the ratings. The value of q in equation (4.1.1) was originally chosen so that the ratings are roughly on the same scale as chess' Elo ratings. Additionally, the log(10) term is included by Glickman so that the analysis can easily be conducted using exponentials of base 10 instead of base e, if one chooses. This may be to make it easier to interpret for chess players who are not specialist mathematicians.

Note also that this model is also of the form of the widely-used Bradley-Terry model, Bradley and Terry (1952), as seen in Section 2.2.2. These give players strength α_{it} and α_{jt}, with $P(W_{ijt} = 1) = \frac{\alpha_{it}}{\alpha_{it}+\alpha_{jt}}$. The connection can be seen by simply using the transformation $\alpha_{it} = \exp(q\theta_{it})$.

An important factor to consider in modelling tennis matches is the fact that players' strengths alter over time. One way this can be modelled is to assume each θ_{it} moves according to a symmetric random walk each time period, independently of all other θ_{jt}, with variance γ^2 at all times and for all players. Given a rating at time $t-1$, each player's rating at time t therefore has distribution

$$\theta_{it}|\theta_{i,t-1},\gamma \sim N(\theta_{i,t-1},\gamma^2), \quad i=1,\ldots,N,\ t=1,\ldots,T. \tag{4.1.2}$$

Reasonable questions can be asked about whether this model describes player movements sufficiently comprehensively. While local movements may exhibit random structure, we would also expect a player to improve during the early part of their career before declining as they age. However, this model may in reality be sufficiently flexible to naturally account for these career trajectories.

In order to make inference about $\boldsymbol{\theta}_t$, the vector of all θ_{it} for $i=1,\ldots,n$, the Glicko ratings system uses a state-space model. The players' ratings can never be known precisely, we can only make inference about them with uncertainty. In each time period t, our beliefs about $\boldsymbol{\theta}_t$ are represented by a distribution, and Bayesian updates of our beliefs about $\boldsymbol{\theta}_t$ are performed after observing $\boldsymbol{w}_t$, the vector of all match results during time period t.

Ideally in Bayesian analysis, a prior would be chosen that is conjugate to the likelihood. However for the likelihood function in equation (4.1.1), there are no easy examples. Other reasonable likelihood functions that could be chosen do not exhibit nice conjugacy properties either. The options are therefore to model the distribution of each $\boldsymbol{\theta}_t$ using non-conjugate priors exactly, which is very difficult; use computational methods to get non-parametric approximations of the distribution of $\boldsymbol{\theta}_t$; or to repeatedly perform approximations of the distribution of $\boldsymbol{\theta}_t$, so that updates can be performed easily, albeit with a possible cost to the accuracy of the model. Glicko opts for the latter option, devising a series of closed-form approximations for the posterior distribution of $\boldsymbol{\theta}_t$ given the results of the latest set of matches, $\boldsymbol{w}_t$, and a prior distribution for $\boldsymbol{\theta}_t$. In particular, the Glicko model attempts to describe the beliefs about

each player using independent Gaussian distributions in each time period. We use $\boldsymbol{\nu}_t$ and $\boldsymbol{\sigma}_t$ to denote the vectors of prior parameters for our beliefs about $\boldsymbol{\theta}_t$, but use $\boldsymbol{\nu}_t^*$ and $\boldsymbol{\sigma}_t^*$ for our posterior parameters after observing scores, $\boldsymbol{w}_t$. We use $\boldsymbol{w}_{it}$ to denote the vector of all of player i's scores during time period t.

The above can be summarised as

$$\theta_{it}|\nu_{it},\sigma_{it} \sim N(\nu_{it},\sigma_{it}^2), \qquad i=1,\ldots,N,\ t=0,\ldots,T, \tag{4.1.3}$$

$$\theta_{it}|\boldsymbol{w}_{it},\nu_{it},\sigma_{it} \sim N(\nu_{it}^*,\sigma_{it}^{*2}), \qquad i=1,\ldots,N,\ t=0,\ldots,T. \tag{4.1.4}$$

It is typical to pick common initial prior parameters ν_0 and σ_0 for all players, so that $\nu_{i0}=\nu_0$ and $\sigma_{i0}=\sigma_0$ for all i. If more information is known about the players at time 0, then the players could be given different prior parameters. For example, players' world rankings could be used to generate more informative priors.

Using common prior parameters generally means player strengths in early time periods are poorly modelled. The ratings will be heavily influenced by the prior parameters until enough matches have been played to make them irrelevant. Glickman describes a method to make better inference about player strengths in these early time periods by using information from future matches. However, this is unnecessary in our case since the matches we want to investigate occur between 2013 and 2016, which is significantly after the first match results we use in 1991. We therefore decided that further investigation of this was not required, but it appears that it would be worth considering in any application in which it is important to model early player strengths well.

Glickman's paper summarises a set of update steps they use to provide a closed form approximation for the distributions of each $\boldsymbol{\theta}_t$. We will now describe them in slightly more detail.

The distribution in equation (4.1.2) can be used to find the distribution of $\boldsymbol{\theta}_t$ given the posterior parameters of $\boldsymbol{\theta}_{t-1}$, those being $\boldsymbol{\nu}_{t-1}^*$ and $\boldsymbol{\sigma}_{t-1}^*$. This is done by

conditioning on the unknown $\boldsymbol{\theta}_{t-1}$, giving

$$\begin{aligned}\pi(\boldsymbol{\theta}_t|\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1}) &= \int_\Omega \pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1},\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1})\pi(\boldsymbol{\theta}_{t-1}|\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1})d\boldsymbol{\theta}_{t-1} \\ &= \int_\Omega \pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})\pi(\boldsymbol{\theta}_{t-1}|\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1})d\boldsymbol{\theta}_{t-1} \qquad (4.1.5)\end{aligned}$$

The support of $\boldsymbol{\theta}_{t-1}$ is $(-\infty,\infty)^N$, which is denoted by Ω. If $\boldsymbol{\theta}_{t-1}$ is normally distributed, then using properties of conditional normals also gives a normal prior distribution for $\boldsymbol{\theta}_t$. Specifically,

$$\boldsymbol{\theta}_t|\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1},\gamma \sim MVN(\boldsymbol{\nu}^*_{t-1},\boldsymbol{\sigma}^*_{t-1}+\gamma^2). \qquad (4.1.6)$$

Combining this with equation (4.1.3) shows how the posterior parameters at time $t-1$ link to the prior parameters at time t, namely

$$\begin{aligned}\boldsymbol{\theta}_t|\boldsymbol{\nu}_t,\boldsymbol{\sigma}_t &\sim MVN(\boldsymbol{\nu}_t,\boldsymbol{\sigma}_t), \\ \boldsymbol{\nu}_t &= \boldsymbol{\nu}^*_{t-1}, \\ \boldsymbol{\sigma}_t &= \boldsymbol{\sigma}^*_{t-1}+\gamma^2.\end{aligned}$$

It is common to cap each σ_{it} at σ_{max}, (generally taken to be σ_0), so that we are never less certain about a player's rating than when they are new to the system. In practice, however, this rarely proves an issue after a player has played their first match.

4.1.2 Glicko ratings system: using match outcomes to make inference

The next step to discuss how to make inference about $\boldsymbol{\theta}_t$ given $\boldsymbol{w}_t$. If $\boldsymbol{\theta}_{t-1}$ were also known, Bayes Theorem, along with the fact that the distribution of $\boldsymbol{W}_t$ depends only on $\boldsymbol{\theta}_t$, can be used to find the posterior distribution for $\boldsymbol{\theta}_t$. The updating formula is then

$$\begin{aligned}\pi(\boldsymbol{\theta}_t|\boldsymbol{w}_t,\boldsymbol{\theta}_{t-1}) &\propto P(\boldsymbol{W}_t=\boldsymbol{w}_t|\boldsymbol{\theta}_t,\boldsymbol{\theta}_{t-1})\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}), \\ &\propto P(\boldsymbol{W}_t=\boldsymbol{w}_t|\boldsymbol{\theta}_t)\pi(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}).\end{aligned}$$

Note that if $\boldsymbol{w}_t = \emptyset$, no update is required.

In reality, of course, $\boldsymbol{\theta}_{t-1}$ is not known, but belief about it is expressed through $\boldsymbol{\nu}_{t-1}$ and $\boldsymbol{\sigma}_{t-1}$, as discussed. We therefore express belief about $\boldsymbol{\theta}_t$ through $\boldsymbol{\nu}_t$ and $\boldsymbol{\sigma}_t$, and get the posterior distribution of $\boldsymbol{\theta}_t$ by simply applying Bayes theorem, saying

$$\begin{aligned}\pi(\boldsymbol{\theta}_t|\boldsymbol{w}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) &\propto P(\boldsymbol{W}_t = \boldsymbol{w}_t|\boldsymbol{\theta}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)\pi(\boldsymbol{\theta}_t|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t),\\ &\propto P(\boldsymbol{W}_t = \boldsymbol{w}_t|\boldsymbol{\theta}_t)\pi(\boldsymbol{\theta}_t|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t),\end{aligned}$$

which can be found easily using (4.1.1) and (4.1.6).

However, this posterior distribution is not very convenient to work with. To demonstrate why, we shall begin by finding the marginal posterior distribution for each player i. The marginal posterior distribution will also be needed if we are to provide simple update rules for the parameters of each θ_{it} after observing $\boldsymbol{w}_{it}$.

In order to do this, we begin by defining some more notation. Let E_t be the set of pairs of players (j, k) who play against each other at time t. From a graph theory perspective, this is the set of edges at time t in a graph featuring all players as nodes, with an edge at time t if player j plays player k. Let $opp_t(j)$ be the set of opponents of player j in time period t.

It can be noted that $\pi(\theta_{it}|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) = \pi(\theta_{it}|\nu_{it}, \sigma_{it})$, due to conditional independence. This helps factorise the likelihood, $L(\boldsymbol{\theta}_t|\boldsymbol{w}_t) := P(\boldsymbol{W}_t = \boldsymbol{w}_t|\boldsymbol{\theta}_t)$, which yields

$$\begin{aligned}L(\boldsymbol{\theta}_t|\boldsymbol{w}_t) = P(\boldsymbol{W}_t = \boldsymbol{w}_t|\boldsymbol{\theta}_t) &= \prod_{(i,j)\in E_t} P(W_{ijt} = w_{ijt}|\boldsymbol{\theta}_t)\\ &= \prod_{(i,j)\in E_t} P(W_{ijt} = w_{ijt}|\theta_{it}, \theta_{jt}) := \prod_{(i,j)\in E_t} L(\theta_{it}, \theta_{jt}|w_{ijt}).\end{aligned}$$

This in turn helps rearrange the full posterior into factors that depend on θ_{it} and factors that do not,

$$\begin{aligned}\pi(\boldsymbol{\theta}_t|\boldsymbol{w}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) \propto \pi(\theta_{it}|\nu_{it}, \sigma_{it}) &\prod_{j\in opp_t(i)} \pi(\theta_{jt}|\nu_{jt}, \sigma_{jt}) \prod_{k\notin\{i, opp_t(i)\}} \pi(\theta_{kt}|\nu_{kt}, \sigma_{kt})\\ &\times \prod_{j\in opp_t(i)} L(\theta_{it}, \theta_{jt}|w_{ijt}) \prod_{(k,l)\in E_t\,:\,k,l\neq i} L(\theta_{kl}, \theta_{lt}|w_{klt}).\end{aligned}$$

To find the marginal distribution of θ_{it}, we note that the final term on each line is not dependent on θ_{it} and can be treated as constant. Finally, we integrate over uncertainty in all other θ_{jt} where $j \neq i$. This leaves us with

$$\pi(\theta_{it}|\boldsymbol{w}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) \propto \pi(\theta_{it}|\nu_{it}, \sigma_{it}) \prod_{j \in opp_t(i)} \int_{-\infty}^{\infty} L(\theta_{it}, \theta_{jt}|w_{ijt}) \pi(\theta_{jt}|\nu_{jt}, \sigma_{jt}) \; d\theta_{jt}.$$

We introduce the natural definitions

$$\begin{aligned} \int_{-\infty}^{\infty} L(\theta_{it}, \theta_{jt}|w_{ijt}) \; \pi(\theta_{jt}|\nu_{jt}, \sigma_{jt}) \; d\theta_{jt} &:= L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) \\ &:= P(W_{ijt} = w_{ijt}|\theta_{it}, \nu_{jt}, \sigma_{jt}). \end{aligned} \tag{4.1.7}$$

and

$$\prod_{j \in opp_t(i)} L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) := L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}).$$

for reference later, so that

$$\begin{aligned} \pi(\theta_{it}|\boldsymbol{w}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) &\propto \pi(\theta_{it}|\nu_{it}, \sigma_{it}) \prod_{j \in opp_t(i)} L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) \\ &\propto \pi(\theta_{it}|\nu_{it}, \sigma_{it}) \; L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}). \end{aligned} \tag{4.1.8}$$

The integral required to calculate $L(\theta_{it}, \nu_{it}, \sigma_{it}|w_{ijt})$ in (4.1.7) cannot be solved analytically, and hence the posterior distribution in (4.1.8) is difficult to work with. In the short term, properties such as the means and variance of the posterior distribution of θ_{it} can be calculated computationally. However, successive posterior distributions become more and more difficult as time advances, as the Gaussian prior at time $t = 0$ becomes less and less informative and the amount of matches that players have played increases. There are computational methods for working with these models - Glickman (1999) cite papers using empirical Bayes methods and Markov chain Monte Carlo simulation as examples - but these can be very computationally expensive for large amounts of players and time periods.

Glickman instead proposes a method for approximating the posterior distribution

$\pi(\theta_{it}|\boldsymbol{w}_t, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)$ with a Gaussian distribution with parameters expressed in terms of the prior parameters $\boldsymbol{\nu}_t$ and $\boldsymbol{\sigma}_t$. This means that $\theta_{i,t+1}$ also has a Gaussian prior, and exactly the same algorithm can be applied to approximate $\pi(\theta_{i,t+1}|\boldsymbol{w}_{t+1}, \boldsymbol{\nu}_{t+1}, \boldsymbol{\sigma}_{t+1})$ with a Gaussian distribution, and so on through the whole of time. This approximation makes updating posterior distributions very quick and allows for effective analysis of players strengths at each time period. Section 4.1.3 describes these approximation steps in some detail and discusses their validity.

4.1.3 Glicko ratings system: approximating the posterior distribution of θ_{it}

The approximation Glickman uses for the posterior distribution in (4.1.8) involves multiple steps, and we will describe each here. The key idea is to use the fact that the conjugate prior for a Gaussian density is also a Gaussian density. Looking at (4.1.8), we see that θ_{it}'s prior distribution, $\pi(\theta_{it}|\nu_{it}, \sigma_{it})$ is a Gaussian density. Hence if θ_{it}'s likelihood, $L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})$, were also Gaussian, then θ_{it} would also have a Gaussian posterior density. The goal of these approximation steps is therefore to approximate the likelihood of θ_{it} by a Gaussian density, so that the posterior is also Gaussian. This will mean the prior distribution of θ_{it+1} is also Gaussian, and so on through all time periods, making updating and using the successive distributions of θ_{it} much easier.

The first step is to approximate the logistic functions representing the likelihood of each match result, $L(\theta_{it}, \theta_{jt}|w_{ijt})$ by Gaussian CDFs. This allows the joint likelihood of all matches for each player, $L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})$, to be represented by a product of single Gaussian CDFs. These Gaussian CDFs will then be approximated by logistic functions again, so that a link can be drawn between these logistic functions and the original ones in equation (4.1.1). The product of these logistic functions will then be approximated by a single Gaussian density, so that the product of this Gaussian likelihood with the Gaussian prior will yield a Gaussian posterior. This will provide a simple rule for updating the parameters of the distribution of θ_{it} to reflect the match

results $\boldsymbol{w}_{it}$, and provide a Gaussian prior for $\theta_{i,t+1}$.

Step 1. Approximating $L(\theta_{it}, \theta_{jt}|w_{ijt})$, the likelihood for each match, by a Gaussian CDF

Inputting the expressions for a match likelihood in equation (4.1.1) and the prior density of θ_{jt} in equation (4.1.2), into the likelihood for player i alone in equation (4.1.7) gives

$$\begin{aligned} L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) &= \int_{-\infty}^{\infty} L(\theta_{it}, \theta_{jt}|w_{ijt})\pi(\theta_{jt}|\nu_{jt}, \sigma_{jt})\ d\theta_{jt}. \\ &\propto \int_{-\infty}^{\infty} \frac{\left(e^{q(\theta_{it}-\theta_{jt})}\right)^{w_{ijt}}}{1+e^{q(\theta_{it}-\theta_{jt})}} \exp\left(-\frac{1}{2\sigma_{jt}^2}(\theta_{jt}-\nu_{jt})^2\right) d\theta_{jt} \\ &\propto \int_{-\infty}^{\infty} \frac{\left(e^{q(\theta_{jt}-\theta_{it})}\right)^{1-w_{ijt}}}{1+e^{q(\theta_{jt}-\theta_{it})}} \exp\left(-\frac{1}{2\sigma_{jt}^2}(\theta_{jt}-\nu_{jt})^2\right) d\theta_{jt}. \qquad (4.1.9) \end{aligned}$$

The last step is found by multiplying both the numerator and denominator by $\exp(q(\theta_{jt}-\theta_{it}))$ and rearranging.

In order to approximate the expression obtained for the integrand in equation (4.1.9), we begin by noting the form for the CDF of a logistic distribution. Suppose X has a logistic distribution with mean m and variance v. If we let $\delta^2 = 3v/\pi^2$, then

$$P(X < x) = \frac{e^{\frac{(x-m)}{\delta}}}{1+e^{\frac{x-m}{\delta}}} \text{ for } -\infty < x < \infty.$$

It can be noted that if $w_{ijt} = 0$, the first term in equation (4.1.9) takes this form, where $m = \theta_{it}$ and $\delta = 1/q$. Glicko approximates this logistic CDF by the CDF of a normally distributed random variable, Y, with the same mean, θ_{it}, as suggested by Cox (1987) and Aitchison and Begg (1976), for example. These references then suggested either matching a quantile or matching the variance - Glickman favours matching the variance. The variance of a logistic distributed random variable is $v = \delta^2/3\pi^2$. This means that if we are to consider the first term in equation (4.1.9) as the CDF of a logistic random variable in θ_{it}, then $\delta = 1/q$ and the variance is equal to $(1/q)^2\pi^2/3 = \pi^2/(3q^2)$.

In summary, to approximate the CDF of a logistic random variable X with mean m and scale parameter δ, we say

$$P(X < x) \approx P(Y < x) = \Phi\left(\frac{x - m}{\frac{\delta\pi}{\sqrt{3}}}\right).$$

A comparison between a logistic CDF and a Gaussian approximation is shown in Figure 4.1.1. It can be seen that the two are very similar. Using equal variances balances the needs for the functions to be close near the centre of the distribution and in the tails - a different variance can be used in the Gaussian random variable if a function is desired that is more similar to the logistic CDF near the centre of the distribution or at a particular quantile, as in Cox (1987) and Aitchison and Begg (1976).

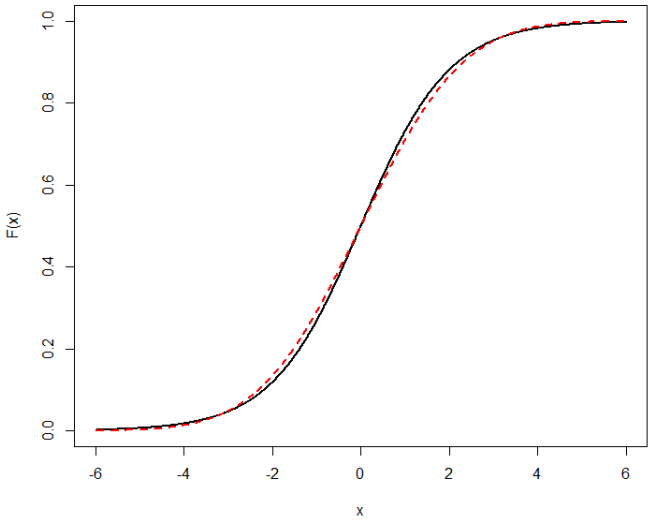

Figure 4.1.1: The CDFs of a logistic random variable (solid black line) and a Gaussian variable (dotted red line), both with mean 0 and variance 1.

If $\Phi(z)$ is used to denote the CDF of a standard normal random variable, the

above information can be summarised by saying

$$\begin{aligned}\frac{\left(e^{q(\theta_{jt}-\theta_{it})}\right)}{1+e^{q(\theta_{jt}-\theta_{it})}} &= P(X<\theta_{jt})\\ &\approx P(Y<\theta_{jt})\\ &= \Phi\left(\frac{(\theta_{jt}-\theta_{it})}{\sqrt{(\pi^2/(3q^2)}}\right)\\ &= \Phi\left(\frac{q\sqrt{3}}{\pi}(\theta_{jt}-\theta_{it})\right),\end{aligned}$$

using the properties of normal distributions.

Similarly, if $w_{ijt}=1$, this term is instead the survival function of the logistic distribution, since $\frac{1}{1+e^x}=1-\frac{e^x}{1+e^x}$. Noting that $1-\Phi(x)=\Phi(-x)$, the same approximation can be performed with the survival functions,

$$\begin{aligned}\frac{1}{1+e^{q(\theta_{jt}-\theta_{it})}} &= 1-P(X<\theta_{jt})\\ &\approx 1-P(Y<\theta_{jt})\\ &= 1-\Phi\left(\frac{q\sqrt{3}}{\pi}(\theta_{jt}-\theta_{it})\right)\\ &= \Phi\left(-\frac{q\sqrt{3}}{\pi}(\theta_{jt}-\theta_{it})\right).\end{aligned}$$

Combining this information gives

$$\frac{\left(e^{q(\theta_{jt}-\theta_{it})}\right)^{1-w_{ijt}}}{1+e^{q(\theta_{jt}-\theta_{it})}} \approx \Phi\left((-1)^{w_{ijt}}\frac{q\sqrt{3}}{\pi}(\theta_{jt}-\theta_{it})\right).$$

This completes the approximation of $L(\theta_{it},\nu_{jt},\sigma_{jt}|w_{ijt})$ with a Gaussian CDF. Inputting this approximation into equation (4.1.9) yields

$$L(\theta_{it},\nu_{jt},\sigma_{jt}|w_{ijt}) \approx \int_{-\infty}^{\infty}\Phi\left((-1)^{w_{ijt}}\frac{q\sqrt{3}}{\pi}(\theta_{jt}-\theta_{it})\right)\exp\left(-\frac{1}{2\sigma_{jt}^2}(\theta_{jt}-\nu_{jt})^2\right)d\theta_{jt}. \tag{4.1.10}$$

Step 2: Proving that this approximation to $L(\theta_{it},\nu_{jt},\sigma_{jt}|w_{ijt})$, the likelihood for each match, is equal a single Gaussian CDF

Attention now turns to the integral in equation (4.1.10) - the next section will prove it is equal to a Gaussian CDF. To do this, the substitution $z=(\theta_{jt}-\nu_{jt})/\sigma_{jt}$ is

required, as well as $b_{ijt} = \frac{\theta_{it}-\nu_{jt}}{\sigma_{jt}}$, and $a_{jt} = \frac{\pi}{\sigma_{jt}q\sqrt{3}}$ and a dummy variable y, recalling that $\Phi(x) = \int_{-\infty}^{x} e^{-\frac{y^2}{2}}\, dy$. This leads to

$$\begin{aligned}
L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) &\approx \int_{-\infty}^{\infty} \Phi\bigg((-1)^{w_{ijt}} \frac{q\sqrt{3}}{\pi}(\theta_{jt} - \theta_{it})\bigg) \exp\bigg(-\frac{1}{2\sigma_{jt}^2}(\theta_{jt} - \nu_{jt})^2\bigg) d\theta_{jt}. \\
&\approx \int_{-\infty}^{\infty} \Phi\bigg((-1)^{w_{ijt}} \frac{q\sqrt{3}}{\pi}(\sigma_{jt}z + \nu_{jt} - \theta_{it})\bigg) \exp\bigg(-\frac{z^2}{2}\bigg) dz \\
&\approx \int_{-\infty}^{\infty} \Phi\bigg((-1)^{w_{ijt}} \frac{(z - \frac{\theta_{it}-\nu_{jt}}{\sigma_{jt}})}{\frac{\pi}{q\sigma_{jt}\sqrt{3}}}\bigg) \exp\bigg(-\frac{z^2}{2}\bigg) dz \\
&\approx \int_{-\infty}^{\infty} \int_{-\infty}^{(-1)^{w_{ijt}} \frac{z-b_{ijt}}{a_{jt}}} \exp\bigg(-\frac{y^2}{2}\bigg) \exp\bigg(-\frac{z^2}{2}\bigg) dy dz. \qquad (4.1.11)
\end{aligned}$$

We then note that this expression takes the same form as $P(Y < (-1)^{w_{ijt}}(Z - b_{ijt})/a_{jt})$, where Y and Z are independent standard normal random variables. This is because we integrate the products of their densities over the appropriate region.

If $w_{ijt} = 0$ then this probability is equal to $P(Z - a_{jt}Y > b_{ijt})$. Since Y and Z are standard Gaussian random variables, we note that

$$Z - a_{jt}Y \sim N(0, 1 + a_{jt}^2) \Rightarrow P(Z - a_{jt}Y > b_{ijt}) = \Phi\bigg(-\frac{b_{ijt}}{\sqrt{1 + a_{jt}^2}}\bigg).$$

On the other hand, if $w_{ijt} = 1$, we instead get $P(Z + a_{jt}Y < b_{ijt})$. Similarly,

$$Z + a_{jt}Y \sim N(0, 1 + a_{jt}^2) \Rightarrow P(Z + a_{jt}Y < b_{ijt}) = \Phi\bigg(\frac{b_{ijt}}{\sqrt{1 + a_{jt}^2}}\bigg).$$

These two results can be summarised jointly as

$$P\bigg(Y < (-1)^{w_{ijt}} \frac{Z - b_{ijt}}{a_{jt}}\bigg) = \Phi\bigg((-1)^{1-w_{ijt}} \frac{b_{ijt}}{\sqrt{1 + a_{jt}^2}}\bigg).$$

This can then be substituted back into the approximation in equation (4.1.11), yielding

$$\begin{aligned}
L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) &\approx \int_{-\infty}^{\infty} \int_{-\infty}^{(-1)^{w_{ijt}} \frac{z-b_{ijt}}{a_{jt}}} \exp\bigg(-\frac{y^2}{2}\bigg) \exp\bigg(-\frac{z^2}{2}\bigg) dy dz \\
&\approx \Phi\bigg((-1)^{1-w_{ijt}} \frac{b_{ijt}}{\sqrt{1 + a_{jt}^2}}\bigg), \quad \text{for } b_{ijt} = \frac{\theta_{it} - \nu_{jt}}{\sigma_{jt}} \text{ and } a_{jt} = \frac{\pi}{\sigma_{jt}q\sqrt{3}}.
\end{aligned} \qquad (4.1.12)$$

Step 3: Approximating the Gaussian approximation to $L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt})$ with a logistic CDF

Having approximated the likelihood of each match for each player in equation (4.1.12), Glickman then seeks to further approximate the likelihood for each match with a logistic CDF. The reasoning behind this is not explained, but appears to stem from a desire with logit link functions rather than Gaussian - see also the likelihood initially chosen in equation (4.1.1), where use of a Gaussian function would have been equally as valid. This may be simply be because the audience of chess players that the Glicko ratings system is designed for would find the simple analytic formula of a logistic function easier to understand than a Gaussian CDF, which is slightly more opaque in meaning to a non-mathematical audience. Either way, there appears to be no strict need for this step, as the steps that follow would work fine were this step omitted.

In order to approximate the likelihood in (4.1.12) with a logistic CDF, recall that a logistic CDF is of the form $P(X < x) = \frac{e^{(x-m)/\delta}}{1+e^{(x-m)/\delta}}$, where $\text{Var}(X) = \delta^2\pi^2/3$, and $\text{E}(X) = m$. Since we are approximating the CDFs of standard Gaussian random variables, with mean 0 and variance 1, we require $m = 0$ and $\delta^2 = 3/\pi^2$. Performing this approximation, cancelling terms and using the notation

$$g(\sigma_j) = \left(1 + \frac{3q^2\sigma_j^2}{\pi^2}\right)^{-1/2}$$

leads to a final approximation for the likelihood of a single match result

$$L(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) \approx \frac{\left(e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}\right)^{w_{ijt}}}{1 + e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}. \tag{4.1.13}$$

This approximation will be referred to with the notation

$$\begin{aligned}\tilde{L}(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt}) &:= \frac{\left(e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}\right)^{w_{ijt}}}{1 + e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}} \\ \tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}) &:= \prod_{j \in opp_t(i)} \frac{\left(e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}\right)^{w_{ijt}}}{1 + e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}.\end{aligned} \tag{4.1.14}$$

Step 4: Approximating $\tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})$ with a Gaussian density

We begin by inputting the approximation for the likelihood of a single match in equation (4.1.14) into the marginal posterior distribution for θ_{it} in equation (4.1.8) to approximate the marginal posterior distribution of θ_{it}. Here $\propto\!\!\!\!\!\sim$ is again used to denote approximate proportionality, then

$$\begin{aligned}\pi(\theta_{it}|\boldsymbol{w}_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) &\propto \pi(\theta_{it}|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)\ L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}) \\ &\propto\!\!\!\!\!\sim \pi(\theta_{it}|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) \prod_{j \in opp_t(i)} \frac{\left(e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}\right)^{w_{ijt}}}{1 + e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}. \qquad (4.1.15)\end{aligned}$$

Since the prior distribution is Gaussian, then were it possible to approximate the joint likelihood by a Gaussian, this would also give an approximately Gaussian posterior distribution. This would be very desirable, as then the posterior distribution could be updated at each time step by simply using the parameters of other players at that time, and the results of the matches between players.

Glickman does not justify this approximation, but a little experimentation shows that as long as a player wins and loses a match in this time period, the product of these CDFs is reasonably well approximated by a Gaussian density, as shown in Figure 4.1.2. The left of these two plots shows the likelihood of player i's results against four opponents of random strengths, with player i winning three and losing one. The right hand graph demonstrates that the product of these likelihoods is well approximated by a Gaussian distribution. However, if a player wins or loses all of their matches in a time period, then clearly a Gaussian density is a poor approximation, as shown in Figure 4.1.3. In the left of these plots, the likelihood of player i winning four matches against the same opponents is plotted, and the right-hand graph demonstrates that the products of these likelihoods would be very poorly approximated by a Gaussian density.

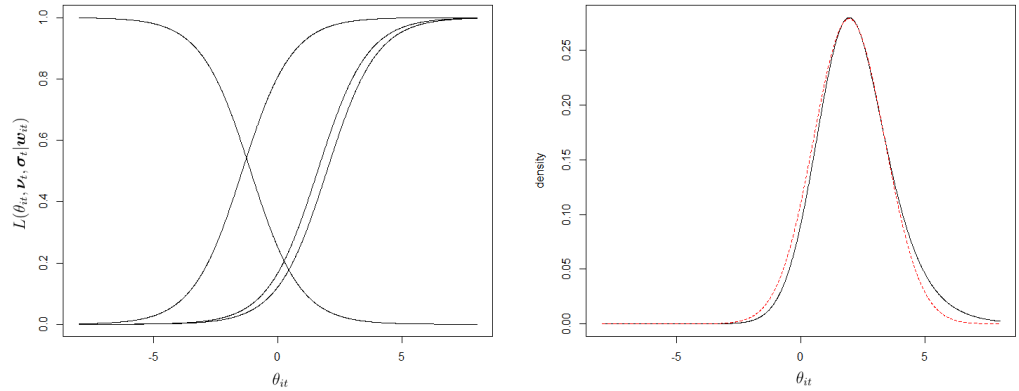

Figure 4.1.2: Likelihoods of player i's results against four opponents of random strengths, with three wins and one defeat. The normalised product of these likelihoods (solid black) can be closely approximated by a Gaussian density (dotted red).

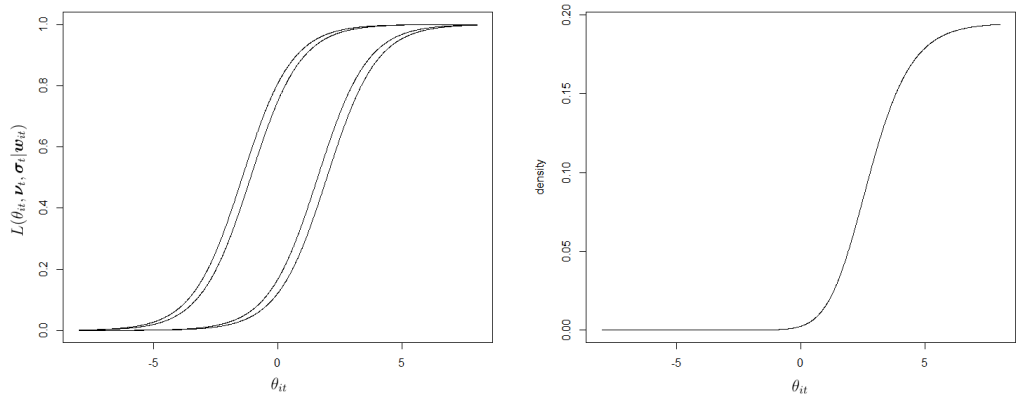

Figure 4.1.3: Likelihoods of player i's results against the same four opponents as in Figure 4.1.2, with four wins. The normalised product of these likelihoods (solid black) can no longer be well approximated by a Gaussian density.

In cases where all w_{ijt} are equal for some i and t, there are therefore legitimate concerns about approximating the likelihood by a Gaussian density. However, further experimentation shows that the Glicko model's updating steps can be derived in a different way that avoids the problem of approximating a Gaussian CDF by a Gaus-

sian density, should a player win (or lose) all of these matches in a time period. We will discuss this in Section 4.2 after describing Glickman's own steps.

In order to approximate the approximate joint likelihood of all of player i's matches, $\tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})$, as seen in equation (4.1.14), with a Gaussian density, Glickman suggests using a Gaussian distribution with mean $\hat{\theta}_{it}$, the maximum likelihood estimator of θ_{it}, and variance $-1/I(\hat{\theta}_{it})$, where $I(\theta_{it})$ is the Fisher information of θ_{it} using this joint (approximate) likelihood. Note, however, that if $w_{ijt} = 1$ for all a player's opponents j, then $\hat{\theta}_{it} = \infty$. Similarly, if $w_{ijt} = 0$ for all opponents j, then $\hat{\theta}_{it} = -\infty$. When w_{ijt} are not all equal, however, the steps can reasonably be performed.

Step 4a: Finding $\hat{\theta}_{it}$, which will be the mean of the Gaussian density

We will need to find $\hat{\theta}_{it}$ by taking the derivative of the approximate joint log likelihood in equation (4.1.14). The approximate log-likelihood is denoted by

$$\begin{aligned}\tilde{l}(\theta_{it}|\boldsymbol{w}_{it}) :&= \log\big(\tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})\big).\\ &= \sum_{j\in opp_t(i)} \Big(qg(\sigma_{jt})(\theta_{it}-\nu_{it})w_{ijt} - \log\big(1+e^{qg(\sigma_{jt})(\theta_{it}-\nu_{jt})}\big)\Big).\end{aligned}$$

Note that for simplicity, the dependence on $\boldsymbol{\nu}_t$ and $\boldsymbol{\sigma}_t$ is dropped from the notation $\tilde{l}(\theta_{it}|\boldsymbol{w}_{it})$, since context makes it obvious.

Upon differentiating $\tilde{l}(\theta_{it}|\boldsymbol{w}_{it})$, it can then be seen that

$$\begin{aligned}\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it}) := \frac{\partial\tilde{l}(\theta_{it}|\boldsymbol{w}_{it})}{\partial\theta_{it}} &= q\sum_{j\in opp_t(i)} g(\sigma_{jt})\Big(w_{ijt} - \frac{e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}{1+e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}\Big)\\ &= q\sum_{j\in opp_t(i)} g(\sigma_{jt})\Big(w_{ijt} - \frac{1}{1+e^{-(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}}\Big). \qquad (4.1.16)\end{aligned}$$

The maximum likelihood estimator, $\hat{\theta}_{it}$, is the value of θ_{it} at which this derivative equals 0. While this value could be found numerically, this would be time-consuming and would not provide easily calculable update rules. Glickman therefore approximates $\hat{\theta}_{it}$ instead.

In order to do this the definition

$$h(\theta_{it}) = \sum_{j \in opp_t(i)} \frac{g(\sigma_{jt})}{1 + e^{-(\theta_{it} - \nu_{jt})qg(\sigma_{jt})}}. \tag{4.1.17}$$

is introduced. In order to approximate $\hat{\theta}_{it}$, Glickman takes a Taylor expansion of $h(\theta_{it})$ around ν_{it} and evaluates it at $\theta_{it} = \hat{\theta}_{it}$. It should be expected that $\hat{\theta}_{it}$ will be close to ν_{it}, unless there are huges amounts of information about $\hat{\theta}_{it}$ contained in $\boldsymbol{w}_{it}$.

By setting the derivative of the approximate log likelihood in equation (4.1.16) to 0, we obtain an equality involving $h(\hat{\theta}_{it})$,

$$h(\hat{\theta}_{it}) = \sum_{j \in opp_t(i)} g(\sigma_{jt}) w_{ijt}, \tag{4.1.18}$$

so that

$$\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it}) = q\big(h(\hat{\theta}_{it}) - h(\theta_{it})\big). \tag{4.1.19}$$

If a Taylor expansion is then taken around ν_{it} and rearranged, this yields

$$\begin{aligned} h(\hat{\theta}_{it}) &\approx h(\nu_{it}) + (\hat{\theta}_{it} - \nu_{it})h'(\nu_{it}) \\ \Rightarrow \hat{\theta}_{it} &\approx \nu_{it} + \frac{h(\hat{\theta}_{it}) - h(\nu_{it})}{h'(\nu_{it})}. \end{aligned}$$

We already know from equation (4.1.19) that

$$\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\Big|_{\theta_{it}=\nu_{it}} = q\big(h(\hat{\theta}_{it}) - h(\nu_{it})\big),$$

and this means

$$h'(\nu_{it}) = 0 - \frac{1}{q}\, \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\Big|_{\theta_{it}=\nu_{it}},$$

where

$$\begin{aligned} \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it}) &= -q^2 \sum_{j \in opp_t(i)} \big(g(\sigma_{jt})\big)^2 \frac{e^{qg(\sigma_{jt})(\theta_{it} - \nu_{jt})}}{(1 + e^{(\theta_{it} - \nu_{jt})qg(\sigma_{jt})})^2} \\ &= -q^2 \sum_{j \in opp_t(i)} \big(g(\sigma_{jt})^2\big)\ \tilde{L}(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt} = 1)\ \tilde{L}(\theta_{it}, \nu_{jt}, \sigma_{jt}|w_{ijt} = 0). \end{aligned}$$

This all combines together gives a new equation in $\hat{\theta}_{it}$,

$$\hat{\theta}_{it} \approx \nu_{it} - \left.\frac{\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})}{\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})}\right|_{\theta_{it}=\nu_{it}}.$$

This approximation of $\hat{\theta}_{it}$ will be labelled $\tilde{\theta}_{it}$, defined as

$$\tilde{\theta}_{it} := \nu_{it} - \left.\frac{\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})}{\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})}\right|_{\theta_{it}=\nu_{it}}.$$

Note that this step can be stated more simply by instead taking a first order Taylor expansion of $\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})$ and rearranging for $\hat{\theta}_{it}$. Doing this is equivalent to performing one method of Newton-Raphson's method for finding roots of functions.

To prove this, consider a function $f(x)$ we wish to find the root of. Taking a first-order Taylor expansion of this function around some $x^{(0)}$ yields

$$f(x) \approx f(x^{(0)}) + (x - x^{(0)})f'(x^{(0)}).$$

If this is evaluated at some $\hat{x}$ such that $f(\hat{x}) = 0$, then this can be rearranged to solve for $\hat{x}$, which gives

$$\begin{aligned} f(\hat{x}) &\approx f(x^{(0)}) + (\hat{x} - x^{(0)})f'(x_0) \\ \Rightarrow \hat{x} &\approx x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}. \end{aligned}$$

This approximation can be labelled $x^{(1)} = x^{(0)} - f(x^{(0)})/f'(x^{(0)})$, and is the general iterative formula for Newton-Raphson's method of finding roots of functions.

Rearranging the Taylor expansion of $\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})$ in this way gives

$$\begin{aligned} \tilde{l}'(\theta_{it}|\boldsymbol{w}_{it}) &\approx \left.\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}} + (\theta_{it} - \nu_{it})\left.\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}} \\ \Rightarrow \theta_{it} &\approx \nu_{it} + \frac{\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it}) - \left.\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}}}{\left.\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}}}. \end{aligned}$$

If this is evaluated at $\theta_{it} = \hat{\theta}_{it}$, then since $\left.\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\hat{\theta}_{it}} = 0$, this implies

$$\hat{\theta}_{it} \approx \nu_{it} - \frac{\left.\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}}}{\left.\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\right|_{\theta_{it}=\nu_{it}}},$$

as before.

When finding the roots of $f(x)$ in general, after finding $x^{(1)}$ one can also Taylor expand $f(x)$ around $x^{(1)}$ to give an even closer approximation of $\hat{x}$. Repeatedly doing so with newer approximations $x^{(2)}$, $x^{(3)}$, ... will give closer and closer approximations to $\hat{x}$ (under certain technical conditions).

It is therefore reasonable to ask whether taking only one iteration of Newton-Raphson's method gives a sufficiently useful approximation to $\hat{\theta}_{it}$ in the Glicko ratings system. To answer this, note first that if $f(x) = ax + b$ is linear, it is easy to prove that one step is sufficient to find the only root, $x = -2a/b$, no matter what $x^{(0)}$ is chosen. Note also that if one is maximising the likelihood of a density $w(x)$, then is equivalent to finding the roots of the derivative of the log likelihood, $f(x) = \frac{d\log(w(x))}{dx}$. For a Gaussian density $w(x)$, it can be seen that

$$\begin{aligned} w(x) &= \frac{1}{\sigma\sqrt{\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu^2)} \\ \Rightarrow f(x) &= \frac{d\log(w(x))}{dx} \\ &= -\frac{1}{\sigma^2}(x-\mu). \end{aligned}$$

Since $f(x)$ is a linear function, this means that using Newton-Raphson's method to find the roots of $f(x)$, and hence maximise the log-likelihood of the Gaussian density $w(x)$ will provide the correct maximum likelihood estimator after one iteration. Hence, since the Glicko ratings system assumes that $\tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t | \boldsymbol{w}_{it})$ is approximately equal to a Gaussian density, it is also reasonable to assume that one iteration of Newton-Raphson's method gives a good approximation of $\hat{\theta}_{it}$. While this assumption does not hold when all w_{ijt} are equal for some i and t, we shall see later why this may be unimportant.

Step 4b: Finding $1/I(\hat{\theta}_{it})$, which will be the variance of the Gaussian density

The next step is to find the variance associated with this likelihood. This is approximated by $1/I(\hat{\theta}_{it}) = -1/\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\big|_{\theta_{it}=\hat{\theta}_{it}}$. In this case, the maximum likelihood

estimator, $\hat{\theta}_{it}$, is approximated by ν_{it}, the prior mean of θ_{it}, instead of $\tilde{\theta}_{it}$. It should not make much difference which is used, since Glickman's approximations assumes $\tilde{l}(\theta_{it}|\boldsymbol{w}_{it})$ is well approximated by the log of a Gaussian likelihood. For of a Gaussian likelihood $\frac{1}{\sigma\sqrt{\pi}}e^{-\frac{(x-\mu)^2}{\sigma^2}}$, the second derivative of the log-likelihood has constant second derivative $-1/\sigma^2$ for all x. Hence $\tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})$ is also approximately constant, and so $\hat{\theta}_{it}$ and ν_{it} should be close enough to $\hat{\theta}_{it}$ that it is fair to say $1/I(\hat{\theta}_{it}) \approx 1/I(\tilde{\theta}_{it}) \approx 1/I(\nu_{it})$.

After steps 4a and 4b, we have approximated the joint likelihood of all of player i's matches by a Gaussian density. If the notation $\phi(x|m, v)$ is used to denote the density of a Gaussian random variable X with mean m and variance v, this density is then

$$L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}) \approx \prod_{j \in opp_t(i)} \frac{\left(e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}\right)^{w_{ijt}}}{1 + e^{(\theta_{it}-\nu_{jt})qg(\sigma_{jt})}} \tag{4.1.20}$$

$$\propto\!\!\!\propto \phi\big(\theta_{it}|\tilde{\theta}_{it}, 1/I(\nu_{it})\big). \tag{4.1.21}$$

Step 5: The prior and likelihood of θ_{it} are Gaussian, and hence so is the posterior

Now that the joint likelihood of all of player i's matches has been approximated by a Gaussian density, the posterior distribution of θ_{it} can be seen to be the product of two Gaussian densities. If $\propto\!\!\!\propto$ is again used to denote approximate proportionality, then by using equations (4.1.15), (4.1.3) and (4.1.21), it can be seen that

$$\begin{aligned}\pi(\theta_{it}|\boldsymbol{w}_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) &\propto \pi(\theta_{it}|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t) L(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it}) \\ &\propto\!\!\!\propto \phi\big(\theta_{it}|\nu_{it}, \sigma_{it}^2\big)\ \phi\big(\theta_{it}|\tilde{\theta}_{it}, 1/I(\nu_{it})\big),\end{aligned}$$

Step 6: Posterior parameters in terms of prior parameters: simple update rules.

It is straightforward to use the properties of Gaussian variables to find the mean and variance parameters in this Gaussian approximation of the posterior. Doing so yields

$$\begin{aligned}\theta_{it}|\boldsymbol{w}_{it},\boldsymbol{\nu}_t,\boldsymbol{\sigma}_t \sim\ & N(\nu^*_{it},\sigma^{*2}_{it}) \\ \sigma^{*2}_{it} &= \left(\sigma_{it}^{-2}+I(\nu_{it})\right)^{-1} \qquad (4.1.22)\\ \nu^*_{it} &= \sigma^{*2}_{it}\left(\frac{\nu_{it}}{\sigma^2_{it}}+\frac{\tilde{\theta}_{it}}{1/I(\nu_{it})}\right) \\ &= \nu_{it}+\sigma^{*2}_{it}\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})\Big|_{\theta_{it}=\nu_{it}}. \qquad (4.1.23)\end{aligned}$$

This completes the steps for the Glicko model's method of updating beliefs about $\boldsymbol{\theta}_t$ from prior parameters $\boldsymbol{\nu}_t$ and $\boldsymbol{\sigma}_t$ to posterior parameters $\boldsymbol{\nu}^*_t$ and $\boldsymbol{\sigma}^*_t$ given the scores of all matches in a time period. This process can be repeated every time new data comes in to reflect the new information. By applying these simple update steps, posterior distributions can be approximated very quickly and easily.

4.2 Links between the Glicko ratings system and Gaussian state space models

Performing the last few steps of the Glicko ratings slightly differently leads to two main conclusions. Firstly, the concerns about approximating the likelihood with a Gaussian PDF in cases where all w_{ijt} are equal for some i and t are not as serious as they might first appear, and secondly that there are very strong similarities between the Glicko ratings system and a Gaussian state space model.

Instead of approximating the likelihood with a Gaussian PDF, we will instead approximate the posterior distribution. This will lead to exactly the same update steps, but is more justifiable than approximating the likelihood in cases where all w_{ijt} are equal for some i and t. This is because a Gaussian PDF multiplied by a Gaussian

CDF still appears approximately Gaussian provided the variance of the CDF is not too small by comparison, and so the approximation by a Gaussian PDF in this case is justifiable, as shown in Figure 4.2.1.

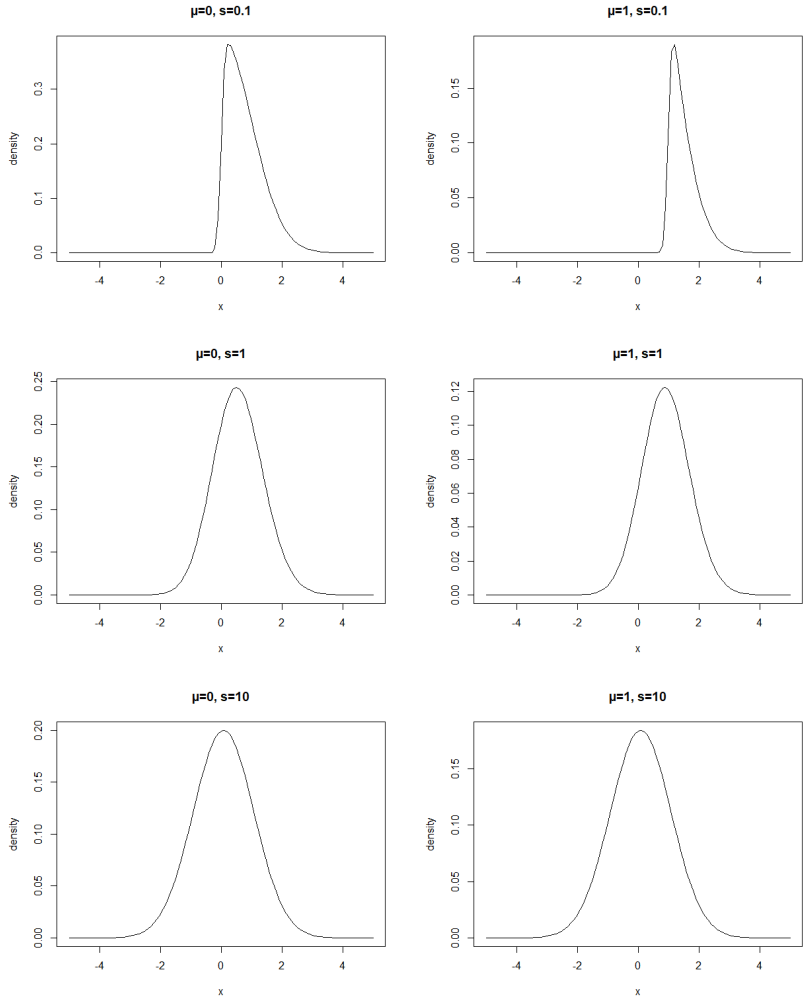

Figure 4.2.1: A standard normal density multiplied by a normal CDF with mean μ and standard deviation s. Approximating these by a normal density would provide the poorest approximation when s is small, but still not wholly unreasonable.

Consider first the mean of the Laplace approximation to the posterior distribution. Recall first that the Laplace approximation of a function $w(x)$ is a normal density with mean $\hat{x}$ that maximises $w(x)$, and variance $-1/\frac{d^2 \log(w(x))}{dx^2}$. If $w(x)$ is a likelihood

function, then this mean is equivalent to the maximum likelihood estimator and the variance is equal to the Fisher information.

Recall that to find the root of a function $f(x)$ given some initial estimate $x^{(0)}$, we can use one step of Newton-Raphson's method to find a better estimate $x^{(1)}$ given by

$$x^{(1)} = x^{(0)} - \frac{f(x^{(0)})}{f'(x^{(0)})}.$$

In our case, we wish to find the maximum of the log density, so let

$$\begin{aligned} f(\theta_{it}) :&= \frac{\partial}{\partial \theta_{it}} \log\left(\pi(\theta_{it}|\boldsymbol{w}_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)\right) \\ &= \frac{\partial}{\partial \theta_{it}} \left(\log\left(\pi(\theta_{it}|\boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)\right) + \log\left(\tilde{L}(\theta_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t|\boldsymbol{w}_{it})\right) \right) \\ &= -\frac{(\theta_{it} - \nu_{it})}{\sigma_{it}^2} + \tilde{l}'(\theta_{it}|\boldsymbol{w}_{it}) \\ f'(\theta_{it}) &= -\frac{1}{\sigma_{it}^2} + \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it}). \end{aligned}$$

Hence using one step of Newton-Raphson's method to approximate the root of $f(\theta_{it})$ (which is equivalent to approximating the maximum of the posterior density) gives

$$\begin{aligned} \theta_{it}^{(1)} &= \theta_{it}^{(0)} - \frac{f(\theta_{it}^{(0)})}{f'(\theta_{it}^{(0)})} \\ &= \nu_{it} - \frac{\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})}{-1/\sigma_{it}^2 + \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})}\Big|_{\theta_{it}=\nu_{it}} \\ &= \nu_{it} + \frac{\tilde{l}'(\theta_{it}|\boldsymbol{w}_{it})}{1/\sigma_{it}^2 - \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})}\Big|_{\theta_{it}=\nu_{it}}. \end{aligned}$$

This gives exactly the same mean as the update step in the original Glicko formulation in equation (4.1.23), noting the form of σ_{it}^* given in equation (4.1.22).

Similarly, the variance of the Gaussian density found by using Laplace's approximation is $-1/\frac{d^2 \log(w(\theta_{it}))}{d\theta_{it}^2}$ evaluated at $\hat{\theta}_{it}$, where $w(\theta_{it}) = \pi(\theta_{it}|\boldsymbol{w}_{it}, \boldsymbol{\nu}_t, \boldsymbol{\sigma}_t)$. Comparison with equation (4.1.22) shows that this is also equivalent to Glickman's updating step if $\theta_{it} = \theta_{it}^{(0)} = \nu_{it}$, since

$$\begin{aligned} \frac{d^2 \log(w(\theta_{it}))}{d\theta_{it}^2} &= -1/\sigma_{it}^2 + \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it}) \\ -\left(\frac{d^2 \log(w(\theta_{it}))}{d\theta_{it}^2}\right)^{-1}\Big|_{\theta_{it}=\theta_{it}^{(0)}} &= \left(1/\sigma_{it}^2 - \tilde{l}''(\theta_{it}|\boldsymbol{w}_{it})\Big|_{\theta_{it}=\nu_{it}}\right)^{-1}. \end{aligned}$$

This demonstrates how the final step in the approximations for the Glicko ratings system is equivalent to using Laplace's approximation, while using one iteration of Newton's method to approximate the posterior mean. Using Laplace's method in this way alleviates concerns about the approximations Glickman uses when all w_{ijt} are equal for some i and t.

In summary, the Glicko ratings system is, at its heart, a Gaussian state space model for player ratings in which the posterior density is approximated using a Laplace approximation. The posterior mean of each player's marginal posterior density is found using one step of Newton-Raphson's method. The choice of a logit link function for the likelihoods of players beating each other, $L(\theta_{it}, \theta_{jt}|w_{ijt})$ means Glicko must approximate this likelihood by a Gaussian to obtain easy update steps. Were the probability of players beating each other modelled with a Gaussian CDF instead, this would be unnecessary, and there would essentially be no difference between the Glicko ratings system and a Gaussian state space model.

This links the Glicko ratings system into the existing literature on Gaussian state space models with Laplace approximations, and is useful to know for any who wish to further study the properties of the Glicko ratings system.

One issue that this link does raise is common one in state space modelling, and that is identifiability of parameters. In observing the results of matches, we only observe information about the difference between two player's ratings $\theta_{it} - \theta_{jt}$, but never make direct observations about the ratings θ_{it} themselves. As such, there is potential for ratings inflation or deflation to occur, in which all players' ratings slowly rise or fall over time, and so the relationship between new and current players changes over time. We hope that there are sufficient new players entering our data to "anchor" player ratings near new players' prior distributions, but nonetheless we must be careful to observe whether any inflation or deflation occurs when we apply Glicko ratings to our tennis data in order to see whether identifiability poses an issue for our modelling accuracy.

4.3 Extension to Five Sets

In a tennis context, we must consider how match-win probabilities are different in 5-set matches and 3-set matches. Anecdotally, longer matches favour the better player, as isolated incidents of bad luck will have a smaller effect. Similarly, in the standard iid points model, as discussed in Section 2.3, if a player's probability of winning a set is greater than 0.5, then his probability of winning a match is greater in a 5-set match than a 3-set match, as shown in Figure 4.3.1.

This section will describe our new method for accounting for this difference between 3 and 5 set matches in the Glicko ratings system by giving greater weight to 5-set matches. We will first explore this issue under the assumption that each set is independent, but will then observe that the transformation this gives can be changed by altering a parameter in the model. This parameter can be chosen to maximise predictive accuracy, essentially meaning we can tweak the model based on the data. This is useful if the data in fact suggest that the assumption that sets are independent gives too much or too little weight to 5-set matches.

An alternative option to tackle account for 5-set matches would be to consider the number of sets won by each player, instead of simply noting the match winner. This would give greater weight to 5-set matches, as more the fact that more sets are played means that more is learned about the players. It would also allow ratings to shift more when the margin of victory is large, but shift less in close matches when both players have won a different number of sets, which could help better reflect the larger skill differences implied by heavy wins than narrow. This was not something we considered in this thesis, and would represent a valid avenue for further work.

Let $m^{(b)}(s_{ij})$ be a function denoting the probability player i beats player j in a best-of-b-sets match given that player i wins a set against player j with probability s_{ij}. Additionally, let $m^{(b)\leftarrow}(\cdot)$ be the inverse function of $m^{(b)}(\cdot)$. We use this notation rather than the more conventional $m^{(b)-1}(\cdot)$ to avoid creating the impression that $(b)-1$ is a numerical equation to be evaluated and to make it clear that $m^{(b)\leftarrow}(\cdot)$

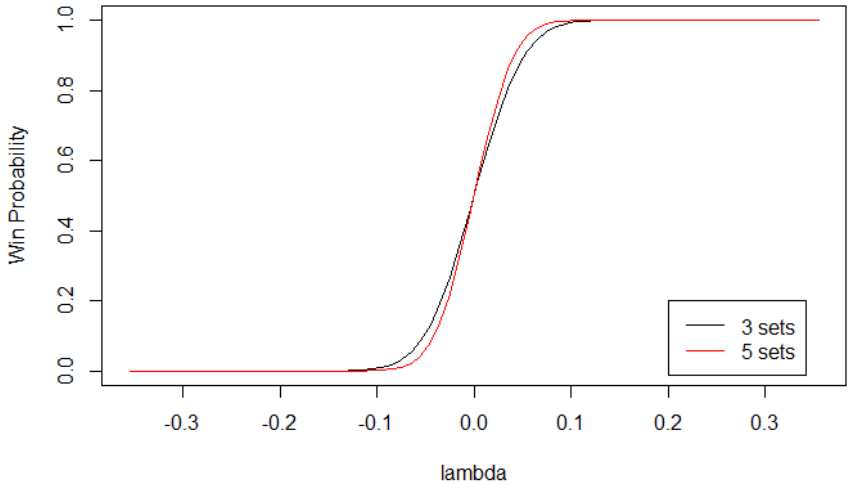

Figure 4.3.1: A comparison of match-win probabilities given dominance parameter $\lambda = \frac{1}{2}(p_{ij} - p_{ji})$ for 3 and 5 set matches.

refers to an inverse function, not a reciprocal.

Given these, recall from Section 2.3 that the formulae describing how the probability of one player winning a set against another player relates to the probability of the player winning the match, assuming sets are independent, are

$$m^{(3)}(s_{ij}) = 3s_{ij}^2 - 2s_{ij}^3$$
$$m^{(5)}(s_{ij}) = 10s_{ij}^3 - 15s_{ij}^4 + 6s_{ij}^5.$$

In order to relate the probability of winning a 5-set match to the probability of winning a 3-set match, we will invert the formula for three set matches to obtain

$$m_{ij}^{(5)} = m^{(5)}\left(m^{(3)\leftarrow}\left(m_{ij}^{(3)}\right)\right) \tag{4.3.1}$$

Although $m^{(3)\leftarrow}(s_{ij})$, the inverse of $m^{(3)}(s_{ij})$, is not easily derived analytically, it is easily implemented in many software programmes. This is because $m^{(3)}(s_{ij})$ is in fact the cumulative distribution function of a Beta(2,2) distribution. This can be found by differentiating it with respect to s_{ij}, and comparing it to the probability

density function of a Beta random variable. This means that $m^{(3)\leftarrow}(s_{ij})$ is just the quantile function of a Beta (2,2) random variable, which is easily implemented in many software packages. Similarly, $m^{(5)}(s_{ij})$ is the CDF of a Beta(3,3) distribution.

In order to incorporate the number of sets, b, into our previous notation, we naturally write

$$L(\theta_{it}, \theta_{jt}, b|w_{ijt}) := P(W_{ijt} = w_{ijt}|\theta_{it}, \theta_{jt}, b), \tag{4.3.2}$$
$$L(\theta_{it}, \nu_{jt}, \sigma_{jt}, b|w_{ijt}) := P(W_{ijt} = w_{ijt}|\theta_{it}, \nu_{jt}, \sigma_{jt}, b),$$

and so on.

This means that simply modelling the probabilities of players beating each other in five sets is straightforward. However, we also want to be able to use the results of five-set matches to update Glicko ratings. Looking at the form of the update equations in equations (4.1.22) and (4.1.22) shows that to do this, derivatives of the log likelihood are required. This becomes complicated if this transformation to five sets is applied. Incorporating equations (4.3.2) into (4.3.1) naturally yields

$$L(\theta_{it}, \nu_{jt}, \sigma_{jt}, b = 5|w_{ijt}) := m^{(5)}\Big(m^{(3)\leftarrow}\big(L(\theta_{it}, \nu_{jt}, \sigma_{jt}, b = 3|w_{ijt})\big)\Big). \tag{4.3.3}$$

The Glicko update step for the mean requires the first derivative of the log of this, which gives

$$\begin{aligned} l'(\theta_{it}, b = 5|w_{ijt}) &= \frac{\partial}{\partial \theta_{it}} l(\theta_{it}, b = 5|w_{ijt}) \\ &:= \frac{\partial}{\partial \theta_{it}} \log(L(\theta_{it}, \nu_{jt}, \sigma_{jt}, b = 5|w_{ijt})) \qquad (4.3.4) \\ &:= \frac{\partial}{\partial \theta_{it}} \log\Big(m^{(5)}\big(m^{(3)\leftarrow}\big(L(\theta_{it}, \nu_{jt}, \sigma_{jt}, b = 3|w_{ijt})\big)\big)\Big). \qquad (4.3.5) \end{aligned}$$

While this can be calculated with repeated application of the chain rule, doing so adds a complexity to the update steps that detracts from one of the main attractions of the Glicko ratings: the simplicity of the updates. We therefore propose a very close approximation that preserves the simplicity of the equations.

Recallling that $m^{(3)}(\cdot)$ and $m^{(5)}(\cdot)$ are CDFs, we can attempt to approximate both

of these functions with a different CDF, such as the logistic CDF, for which we write $F^{(b)}(\cdot)$ if approximating $m^{(b)}(\cdot)$. Doing so will allow $m^{(5)}\big(m^{(3)\leftarrow}(\cdot)\big)$ to be represented approximately a simple form that will be easy to implement into the Glicko parameter update formulae.

One potential problem with choosing a logistic CDF is that its input range is unbounded, whereas a Beta CDF is only non-trivial on [0,1]. However, recall that we are in fact attempting approximate $m^{(5)}\big(m^{(3)\leftarrow}(\cdot)\big)$, which accepts inputs on [0,1] and outputs on [0,1] - just as $F^{(5)}(F^{(3)\leftarrow}(x))$ would. It is the quality of this overall approximation that matters more than whether the individual CDFs are well approximated.

For both cases, $b = 3$ or 5, we want to match the means and variances of the random variables that have $F^{(b)}(x)$ and $m^{(b)}(x)$ as their CDFs. The Beta(2,2) and Beta(3,3) distributions both have mean 0.5, with variances 1/20 and 1/28 respectively. The variance of a logistic random variable X is $\delta^2\pi^2/3$ for scale parameter δ. Therefore, we let $X^{(b)}$ be a logistically distributed random variable with CDF $F^{(b)}(x)$, and with scale parameter $\delta_n^2 = 3\mathrm{Var}(X^{(b)})/\pi^2$. As such, $\delta_3 = \sqrt{3/20\pi^2}$ and $\delta_5 = \sqrt{3/28\pi^2}$. To find $F^{(5)}(F^{(3)\leftarrow}(x))$, we then note

$$\begin{aligned}
F^{(b)}(x) &= \frac{1}{1+\exp\big(-\frac{x-\frac{1}{2}}{\delta_n}\big)},\\
F^{(b)\leftarrow}(x) &= \frac{1}{2} + \delta_n \log\left(\frac{x}{1-x}\right),\\
\text{and therefore } F^{(5)}(F^{(3)\leftarrow}(x)) &\approx \frac{1}{1+\exp\left(-\frac{F^{(3)\leftarrow}(x)-\frac{1}{2}}{\delta_5}\right)}\\
&\approx \frac{1}{1+\exp\Big(-\frac{\delta_3}{\delta_5}\log\big(\frac{x}{1-x}\big)\Big)}\\
&= \frac{1}{1+\big(\frac{x}{1-x}\big)^{-\frac{\delta_3}{\delta_5}}}. \qquad (4.3.6)
\end{aligned}$$

If we compare this function with the truth, $m^{(5)}\big(m^{(3)\leftarrow}(x)\big)$, as in Figure 4.3.2, we observe that $F^{(5)}(F^{(3)\leftarrow}(x))$ appears to approximate $m^{(5)}\big(m^{(3)\leftarrow}(x)\big)\big)$ well.

Equation (4.3.6) provides a simple approximate formula relating the probability of winning three and five-set matches. Of course, however, given that it is indeed

only an approximation based on the assumption that each set is independent, it is possible that better approximations exist. However, equation (4.3.6) gives an easy framework from which to experiment with other approximations by using a different power instead of $\delta_3/\delta_5 = \sqrt{1.4}$ - any constant K could be input in its place, and this parameter could be optimised to maximise predictive performance in a given set of matches in order to better model the probability of winning five-set matches.

We can apply the equation (4.3.6) to our likelihood to find the probability of

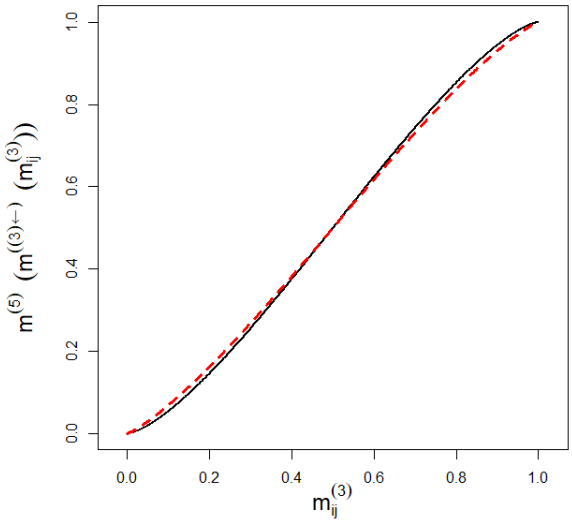

Figure 4.3.2: The exact expression $m^{(5)}\left(m^{(3)\leftarrow}(x)\right)\big)$ (black line) compared with the approximation $F^{(5)}(F^{(3)\leftarrow}(x))$ (red dashed line) using $K = \delta_3/\delta_5$.

winning a five-set match. Doing so gives

$$\begin{aligned}
L(\theta_{it},\theta_{jt},b=5|r_{ijt}) &= m_{ij}^{(5)}\big(m_{ij}^{(3)\leftarrow}(L(\theta_{it},\theta_{jt},b=3|r_{ijt}))\big)\\
&\approx \frac{1}{1+\left(\frac{L(\theta_{it},\theta_{jt},3|r_{ijt})}{1-L(\theta_{it},\theta_{jt},3|r_{ijt})}\right)^{-K}}\\
&\approx \frac{1}{1+\left(\frac{L(\theta_{it},\theta_{jt},3|r_{ijt})}{L(\theta_{it},\theta_{jt},3|1-r_{ijt})}\right)^{-K}}\\
&\approx \frac{1}{1+\left(\frac{e^{q(\theta_{it}-\theta_{jt})w_{ijt}}}{1+e^{q(\theta_{it}-\theta_{jt})}}\frac{1+e^{q(\theta_{it}-\theta_{jt})}}{e^{q(\theta_{it}-\theta_{jt})(1-w_{ijt})}}\right)^{-K}}\\
&\approx \frac{1}{1+\frac{e^{-Kq(\theta_{it}-\theta_{jt})w_{ijt}}}{e^{-Kq(\theta_{it}-\theta_{jt})(1-w_{ijt})}}} \qquad (4.3.7)\\
&\approx \frac{e^{Kq(\theta_{it}-\theta_{jt})w_{ijt}}}{1+e^{Kq(\theta_{it}-\theta_{jt})}}, \quad \text{when } w_{ijt}=0 \text{ or } 1, \qquad (4.3.8)\\
&\approx L(K\theta_{it},K\theta_{jt},b=3\ |w_{ijt}) \quad \text{when } w_{ijt}=0 \text{ or } 1.
\end{aligned}$$

We therefore see that this approximation of the likelihood takes exactly the same form as the original likelihood in equation (4.1.1), but with the distance between θ_{it} and θ_{jt} multiplied by a constant K, so that the difference between players becomes more significant and the stronger player is favoured, if $K > 1$.

Note that equations (4.3.7) and (4.3.8) are only equal when $w_{ijt} = 0$ or 1. For other values, such as $w_{ijt} = \frac{1}{2}$, equality does not hold - however, such values of w_{ijt} are impossible in this tennis setting.

Careful application of the approximation steps in Section 4.1.3 reveals that

$$\begin{aligned}
\tilde{L}(\theta_{it},\nu_{jt},\sigma_{jt},b=5\ |w_{ijt}) &= \frac{e^{Kqg(K\sigma_{jt})(\theta_{it}-\nu_{jt})w_{ijt}}}{1+e^{Kqg(K\sigma_{jt})(\theta_{it}-\nu_{jt})}} \qquad (4.3.9)\\
&= \tilde{L}(K\theta_{it},K\nu_{jt},K\sigma_{jt},b=3\ |w_{ijt}),
\end{aligned}$$

which in some ways is intuitive due to the fact that $K\theta_{jt} \sim N(K\nu_{jt}, K^2\sigma_{jt}^2)$.

The new approximate likelihood for five-set matches in equation (4.3.9) can then be applied easily in the Glicko parameters update equations (4.1.22) and (4.1.23), allowing for five-set matches to be easily incorporated into the existing Glicko framework.

4.4 Application of Glicko Ratings to Tennis Data

In this section, we discuss the implementation of Glicko ratings to tennis data for matches from 1991 to 2016. The data are described in more detail in Chapter 5. In order to implement Glicko ratings, we must pick appropriate parameters. The main parameters to set are the lengths of each time period and the increase in ratings variance per time period, γ^2, as well as the parameter K that governs the relationship between three and five-set matches. We have decided to select these parameters to optimise predictive performance, as given by the log-likelihood. This gave $\gamma^2 = 115.394$ and K=1.186. It is interesting that K is almost exactly equal to $\sqrt{1.4} = \delta_3/\delta_5$, as used in equation (4.3.6). The implication is that the amount of extra information learned from 5-set matches compared with 3-set matches is similar to that implied by an independent sets assumption. (This is not evidence in itself that sets are indeed independent).

To select the period length l, we performed a grid search over the values l=1,2,4 and 8 weeks. We chose not to consider time lengths less than a week as the tennis data only records the first date of the tournament. We felt if we used a period length of one day, it might give a false impression that results were updated at the end of each day, whereas tournaments typically last around a week or two weeks, and so each "day" typically contains at least a week's worth of information. In any case, using a period length of one day gives very similar predictive performance to using one week.

The decision to only consider whole numbers of weeks was made due to the weekly cycle of tournaments, as it seemed sensible to ensure the periods always began on the same day of the week. Of the options l=1,2,4 or 8 weeks, a period length of one week gave the best predictive performance.

In setting the period length to one week, it was important to consider what day of the week to begin a week on. Jeff Sackman's data does not give the date on which matches occur, only the date on which the tournament starts. Figure 4.4.1 shows how many tournaments start on each day of the week. Monday is by far the

most popular choice, with Friday and Saturday as the next two most popular. Most tournaments last around one week. Therefore, there would be significant overlap between a tournament starting on a Saturday and one on a Monday, and as such it would be sensible for such tournaments to be included in the same week for the purpose of Glicko ratings. As such, our implementation of Glicko ratings considers Friday to be the first day of the week, such that tournaments starting on Friday and the following Monday are considered simultaneous. Selecting Sunday or Monday as the first day of the week, as would be more traditional, would not achieve this desirable property.

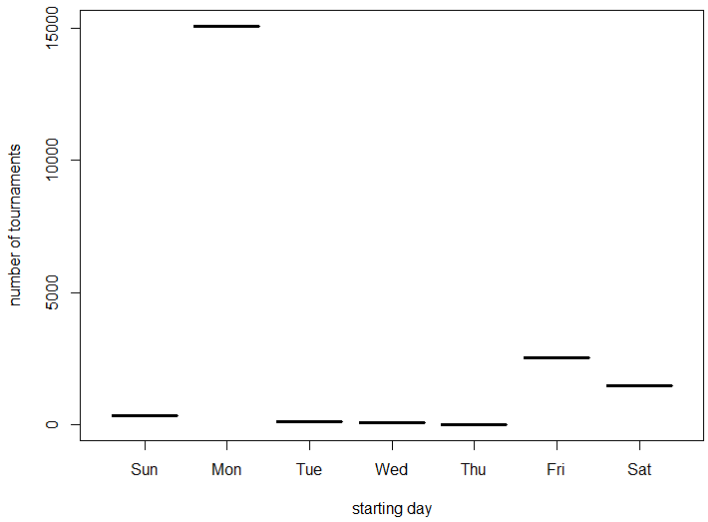

Figure 4.4.1: The number of tournaments in Jeff Sackman's data starting on each day of the week since 1991.

4.4.1 Model Calibration

With any predictive model, it is important to assess whether or not the predictions are well-calibrated, as discussed for example by Hosmer et al. (2013). This means that whenever, for example, our model predicts a player to win with probability

60%, then we would expect the favoured player to win in 60% of such matches. It is entirely possible for a model to have high predictive accuracy, in that it assigned high probability to the events that occurred, while still being badly calibrated. It is important for the model to be well calibrated in order for it to be useful in predicting the odds for individual matches.

Since the predicted probabilities of each player winning can fall anywhere in [0,1], we grouped the predicted win probabilities into bins of length 0.1 and compared the average predicted win probability in each bin with the observed proportion of players who won their match. The results are shown separately for three and five-set matches in Figure 4.4.2. This shows the results fall almost exactly on a straight line, and so we believe that calibration poses no problems for our modelling.

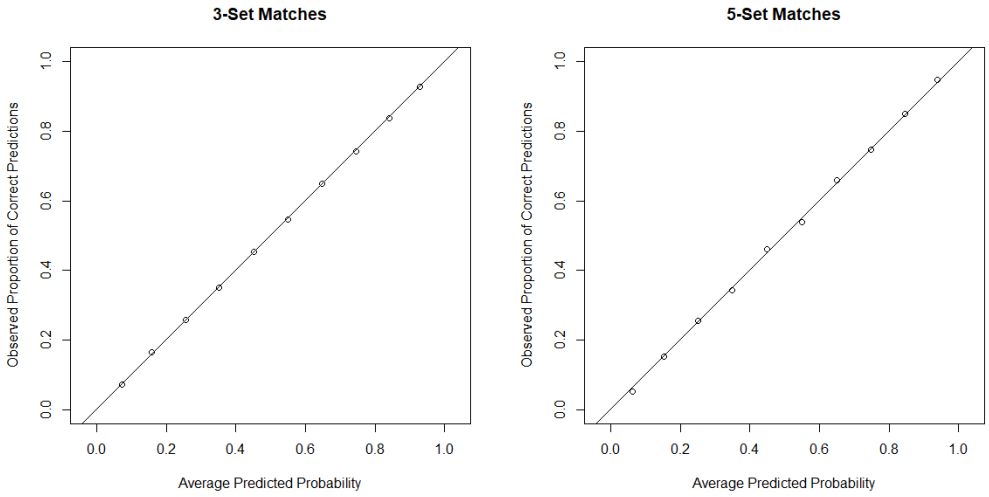

Figure 4.4.2: Average predicted win probabilities in bins of length 0.1 compared with observed proportion of wins for players with predicted win probabilities in those bins.

4.4.2 Examples of Player Ratings

Let's look now at some individual players and how their ratings evolve over time. Figure 4.4.3 shows Andy Murray's Glicko ratings over time. Before he makes his

debut, there is obviously substantial uncertainty about his strength. As he plays, this uncertainty quickly shrinks, and Murray's rise can be steadily tracked over his career, with his highest rating coming at the end of 2016, coinciding with him becoming world number 1 for the first time in November 2016.

Djokovic, Federer and Nadal's ratings are shown in Figure 4.4.4. The mean of the

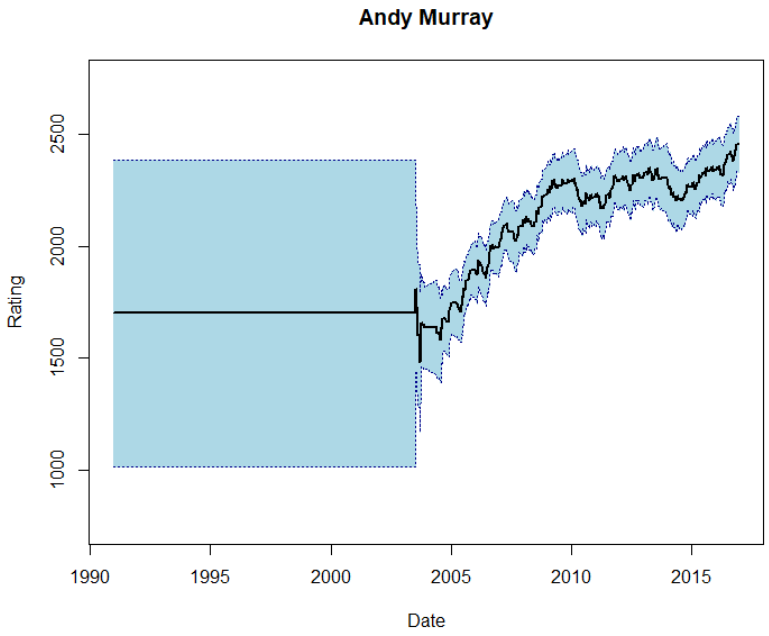

Figure 4.4.3: Andy Murray's Glicko ratings mean over time (black), with a 95% confidence interval (light blue) based on his ratings standard deviation.

estimate of Djokovic's ratings mean at the end of January 2016 is the highest of any player at any time, coinciding with Djokovic setting the all-time record ATP ranking points total at the same time. His rating decreases in the second half of 2016 due to a dip in performance after winning the French Open to hold all four Grand Slams simultaneously. Federer, meanwhile, hits his ranking peak in 2007 before taking a dip in 2008 after suffering from illness. Federer missed the second half of 2016 through injury, and this can be seen on the plot - his ratings mean estimate remains flat, but his ratings variance can be seen to grow. This reflects the uncertainty about the level

Roger would return at - perhaps he would return as strong as ever, but perhaps his performance would dip after his return from injury, or even improve after an opportunity for rest. His increased ratings variance will allow the ratings to quickly capture any changes after his return.

Figure 4.4.5 shows the ratings of two players who did not scale the heights of

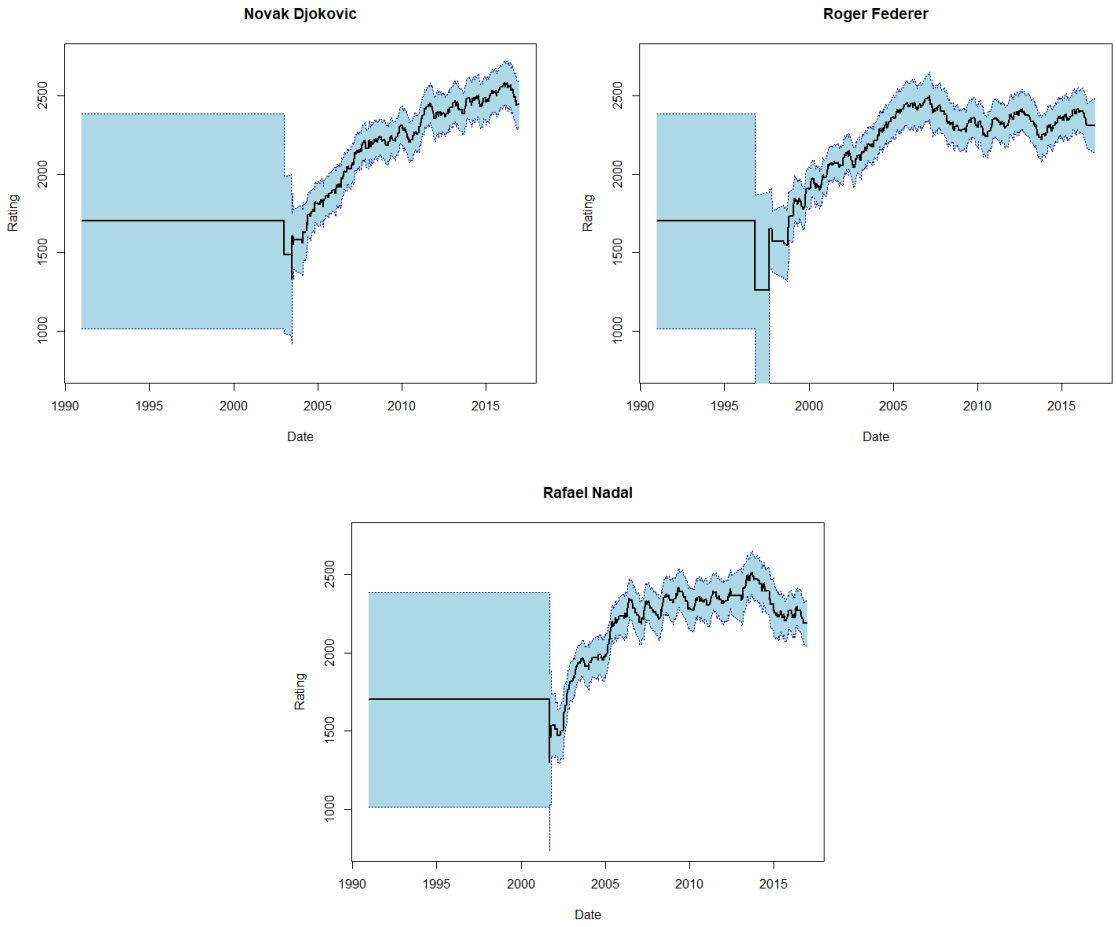

Figure 4.4.4: Novak Djokovic, Roger Federer and Rafael Nadal's Glicko ratings means over time (black), with 95% confidence intervals (light blue) based on their ratings standard deviations.

the world number 1 ranking. Greg Rusedski's highest ever world ranking was number 4, and his highest ratings mean is 2139. For comparison, Murray peaks at 2455 and Federer at 2499. Scott Willinsky reached his highest world ranking of 980 in 2002, and played the majority of his matches in Futures tournaments. His relatively low

win-rate means his ratings mean remains low throughout, but it is also worth noting that by qualifying for fewer tournaments, and progressing less far in them, he has a higher ratings variance from playing fewer matches.

It is also worth briefly looking at the ratings of Thomas Muster in Figure 4.4.6,

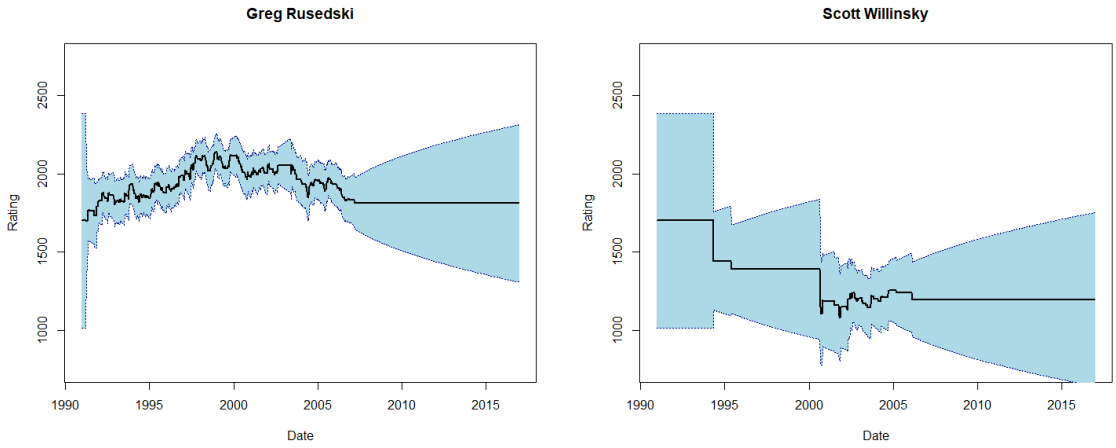

Figure 4.4.5: Greg Rusedski and Scott Willinsky's Glicko ratings means over time (black), with 95% confidence intervals (light blue) based on their ratings standard deviations.

as there are a couple of interesting features in his Glicko ratings to discuss. The first key feature is his long career break from 1999 to 2010. Most players only take relatively short breaks from tennis, and so the increase in their ratings variation are relatively subtle. Even Roger Federer's aforementioned absence in the second half of 2016 only sees his ratings standard deviation increase by about 24. Muster took a career break of 11 years before returning to professional tennis at the age of 42, and it is worth looking at how the Glicko ratings deal with this to highlight how the ratings cope with breaks in general. Figure 4.4.6 shows that while Muster's ratings mean remains constant throughout his career break, in the absence of any matches, his ratings variance grows significantly. This reflects the large amount of uncertainty about how strong he would be on his return. This large variance allowed the Glicko ratings to quickly capture that he was not the same level as he was after his first retirement, and his ratings mean quickly decays before his second retirement in 2011.

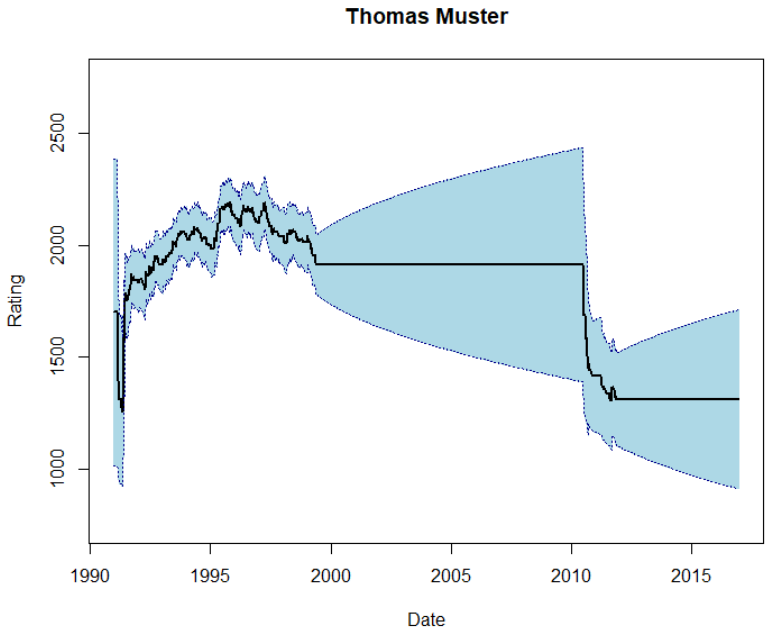

Figure 4.4.6: Thomas Muster's Glicko ratings mean over time (black), with a 95% confidence interval (light blue) based on his ratings standard deviation.

The second key feature of Muster's ratings is the appearance of some minor seasonality in his ratings, however this is probably due to his preferred choice of surface. Muster was a clay court specialist, who first became world number one in February 1996 before losing it for the final time in April of the same year. His ranking was not without controversy, as Andre Agassi and Pete Sampras, among others, criticised his over-reliance on winning ranking points from clay tournaments without distinguishing himself on other surfaces. In Muster's defence, it should be noticed all of Agassi's tournament wins in 1995 also came on only one surface - the hard court - and Muster beat Sampras on carpet at the 1995 German Masters' tournament.

However, our model appears to agree with criticism of Muster's ranking in part, with Muster ranked 5th at the time. A major factor in this is the fact that Glicko ratings naturally weigh recent matches more heavily, whereas ATP ratings weight the last 12 months' tournaments equally. As a result, while Muster's 1995 French

open win contributed heavily to his official ATP number 1 ranking, the Glicko ratings "punish" him for his more recent modest success in the hard court and carpet season, and Muster reaches his 1996 Glicko ratings peak in 1996 in July, further into his favoured clay court season. As such, the Glicko ratings effectively underrate Muster on clay courts at the start of the clay season, when his recent mixed record on other surfaces pull his ratings down despite previous years' form on clay indicating a strong season to come. Similarly, it overrates him on other surfaces at the end of the clay court season, when he will struggle to sustain the level of his success on clay on other surfaces.

4.4.3 Further Work: Glicko Ratings and Surface Information

The issue of incorporating surfaces into Glicko ratings is currently an open research topic. It is widely known that different players prefer different surfaces, and the issue is considered in the models of McHale and Morton (2011) and Irons et al. (2014), for example. Ideally a ratings method should be able to give different ratings to players on different surfaces to account for this issue, thus correctly placing Muster as the strongest clay player at the time of his rankings peak while acknowledging his lesser ability on other surfaces. The difficulty lies in learning players' surface preferences whilst also figuring out how much inference about a player's strength on one surface can be made from matches on another. Should Muster have had a bad season on hard court, how might this affect next season's clay court performance on average?

A simplistic approach to incorporate surface information into Glicko ratings could be to give each player a different rating for each surface, with the ratings being completely independent of each other. For example, we would only update a player's clay rating when they played on clay. If a player played on grass instead, nothing would be learned of the player's ability on clay, and no update of the player's clay rating would be performed. In some ways, it would be as if each player were replaced by three new players, one for each surface, each of which were entirely separate and independent

of each other. All of the theory from standard Glicko ratings would therefore still be applicable, making this model very simple to implement.

However, clearly this approach is slightly unrealistic, in that if a player shows an upturn in form during the clay court season, it might be expected that this form could also continue into the grass court season to follow. A player might make improvements in their game that would be applicable to all surfaces, and we would wish to account for this.

A possible improvement could be to model each player's three ratings as correlated. This would mean that a player winning a clay court match would see an increase in their clay court rating as before, but would also see a smaller increase in their grass and hard court ratings. This would mean that if players performed differently on different surfaces over a long period of time, the ratings on different surfaces would diverge, but it would allow us to use improved performance on one surface to predict potential improvement on other surfaces too. It should be possible to accomplish easily enough within the existing Glicko ratings framework, but we have yet to explore this fully to see whether this is true.

Issues with this approach could include identifying the correlation structure of players' ratings on different surfaces, and the fact that this approach may lead to there being too many parameters to effectively estimate. Maximum likelihood estimation could be used to estimate the correlation matrix, but a good starting point for estimating correlation structure would be based on Irons et al. (2014), which suggest the correlation matrix

$$M = \begin{pmatrix} 1 & 0.25 & 0.5 \\ 0.25 & 1 & 0.01 \\ 0.5 & 0.01 & 1 \end{pmatrix}$$

for their model, with surface $\in$ {hard, clay, grass}. Their model is closer to the Bradley-Terry type model of McHale and Morton (2011) than a Glicko model, and so the different models might need slightly different correlation matrices, but nevertheless

it would provide a good basis for further investigation.

4.4.4 Ratings Inflation

In order to determine whether the ratings this implementation of the Glicko ratings system are useful, it is important to consider with Glicko ratings is whether ratings have inflated or deflated over time. The ratings of players can of course never be observed directly, but only information about the difference between two players' ratings through the outcome of a tennis match. As such, if each players' rating were increased by 100, the model would function exactly as before, as the parameters are non-identifiable. As such, it is possible for the ratings of some or all players to steadily climb or fall over time, when no real change in the actual strength of the players has occurred, while still modelling the outcomes of these players' matches well. However, it is usual to give all new players the same rating whenever they play their first match. If the ratings of one group of players has become inflated, then their matches against new players will therefore not be modelled well. We investigate briefly whether ratings inflation of deflation is likely to be causing any problems in our implementation of Glicko ratings on tennis data.

One way of assessing whether ratings deflation or inflation has occurred is to look at the average ratings of players. However, while a player has played only very few matches, their rating is not particularly informative of their quality (as is represented by the high ratings standard deviation). We therefore only consider the average ratings of players who have played at least a certain number of matches - 50 seems a reasonable amount for a young player to establish themselves. The results of this are shown in Figure 4.4.7, and we can see an overall downward trend over time. Note that the average rating is initially much higher than 1700, the rating of new players, since the first players to play 50 matches are those who win tournaments, and hence gain many ranking points from doing so. This effect soon disappears, as more players play their 50th match.

However, while the average rating for such players is decreasing, the composition

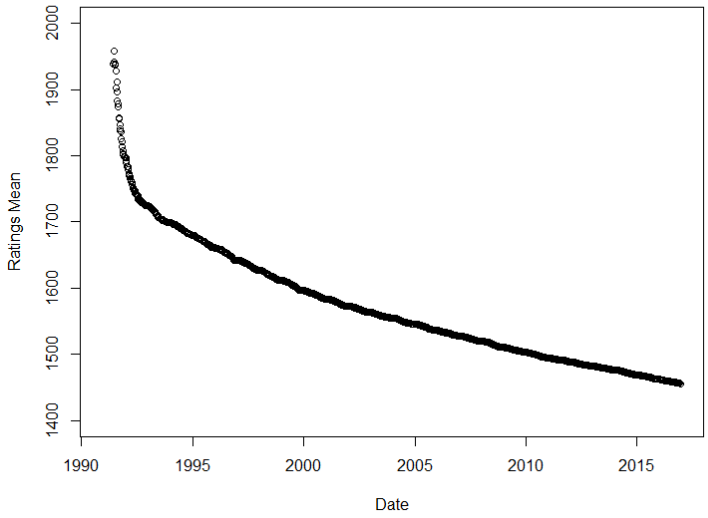

Figure 4.4.7: The number of matches of each type in each year in our data. (Data on Davis cup and ATP tour finals omitted from this plot, since the number of matches is small and varies very little.)

of the data is also changing. Figure 5.2.1 showed how the types of matches occurring change from year to year in our data.[1] While the number of ATP tour-level matches is roughly constant throughout, the number of Challenger matches climbs throughout, while the number of Futures matches increases hugely between the start and the end of the data. Additionally, ATP qualifying data are only available after 2007, and almost all of our Challenger qualifying data comes from 2016. Additionally, omitted from Figure 5.2.1 are the number of ATP tour finals matches and Davis Cup matches, since these numbers are relatively small (about 15 and 300 respectively) and are almost constant throughout the data, and would thus clutter the plot without adding useful information.

Players who participate in Futures are typically young and inexperienced, and similarly top players tend not to plays as many Challenger or qualifying matches.

[1]Some matches in the 1991 season actually occurred in late 1990, hence their inclusion.

This means that there are far more matches between low - rated players as time goes on, while the number of matches between highly rated players remains steady. Having data on a much greater range of low-level matches means we see far more low-level players, and as such it is unsurprising that the average rating decreases over time when the notion of the "average player" in our data is not constant.

This means that looking merely at the average rating of players over time cannot indicate whether players' ratings have inflated or deflated over time. One way to get around this is to instead look at stronger players. We decided to focus on a core of professional players. Various studies from different sources across the years have suggested different minimum ranking to make as much prize money as is spent in expenses while competing. For example, Bialik (2014) suggests that in 2014 a ranking of 336 was required for men, while Russell and USTA (2010) suggests that in 2010 the ranking was 164. Figure 4.4.8 shows how the average ratings mean of the top 100, 200 and 300 has changed over time. For each of these rankings, the average ratings mean appears to have remained relatively steady compared with how much the average ratings mean of all players that have played 50 matches has changed. Indeed, the average rating has risen by only about 25 rating points between the start and end of the data. This small change could even be explained by slight improvements in professional standard in the last 25 years, particularly during the era of dominance by Djokovic, Federer and Nadal, widely considered to be three of the strongest players ever.

The purpose of looking for ratings inflation or deflation is to consider whether new players are incorrectly rated compared to older ones. If all players (even new ones) received a boost of 100 points to their ratings mean, it would not affect our ratings. However, we give all new players a ratings mean of 1700. If all players except new ones received a boost to their points, then the new players would be underrated compare to older ones.

We look at how the probability of a new player beating an average top 200 player

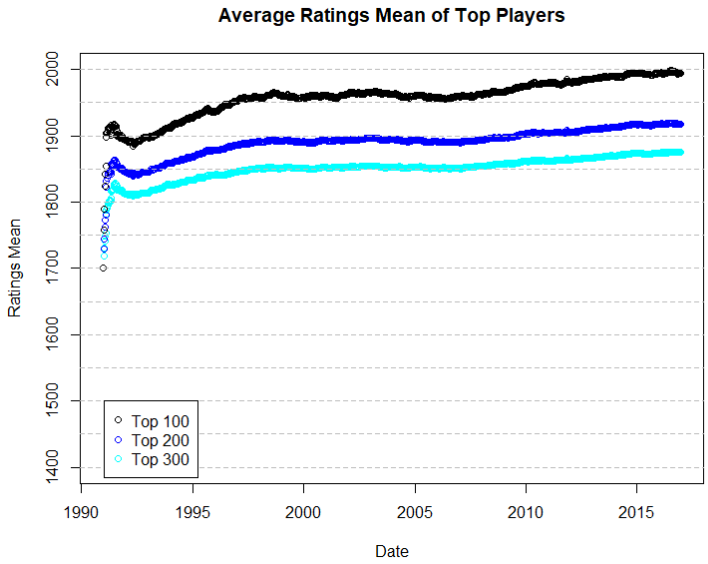

Figure 4.4.8: The average ratings mean of players up to a certain rank over time.

over time. Examination of the data suggests 615 to be a reasonable ratings standard deviation to take for a typical top 200 player. According to our Glicko model, increasing the rating of the top 200 player from 1900 to 1925 reduces the chance of the new players winning a three-set from 0.318 to 0.298 - a reduction of 0.02. This is not a large enough difference to cause concern, especially given the fact that a plausible explanation is the improvements of high-ranked players. We therefore conclude that no further action is required to combat drift in player ratings over time.

Chapter 5

Data: Odds and Results

This chapter contains a brief discussion of the two main sources of data used in the chapters that follow. The odds data provided by ATASS Sports describes the odds on a single betting exchange for 274 matches. We only used 274 for matches for this initial investigation, as the data required a significant amount of processing. We will analyse this odds data to see whether it behaves in a manner consistent with match-fixing or not. In order to help assess this, we use the results of past tennis matches to estimate the strengths of the players involved in the matches in the odds data. These results data came from a public GitHub repository, `https://github.com/JeffSackmann/tennis_atp`, which records a wide variety of data from tennis matches on the ATP tour, as well as Challenger and Futures series. These data will help model the strengths of players, which in turn help model the expected odds. When the odds differ from our expectations, it may be a sign that match-fixing is afoot.

5.1 Odds Data

The main method by which this project aims to highlight suspicious matches is by investigating whether betting odds behave as expected. It is widely known that

anomalies in odds data can occur in fixed matches due to fixers betting large amounts of money to profit on their additional information. Odds data are therefore required to investigate this. The odds data for this project are collected from a single betting exchange by ATASS Sports, who then also pre-processed the data.

In practice, one would not just monitor a single exchange for anomalous activity, but a wide range of bookmakers and exchanges to ensure that the suspicious betting activity is not missed. In mitigation, Reade and Akie (2013) note that traditional bookmakers adjust their odds based in part on the prices on betting exchanges, while Croxson and Reade (2011) consider one major betting exchange and two major bookmakers and claim that information, in the form of odds movements, appears on the exchange first. This suggests that odds in different bookmakers and exchanges may be highly correlated.

Nonetheless, it would be prudent to monitor as many sources as possible to avoid missing evidence. However, the focus in this thesis is on developing algorithms and tools for analysing odds in general. As such, we thought it sufficient to use odds from a single exchange to demonstrate the capabilities of our methods. In order to be of use in detecting suspicious betting activity, we would use our tools to monitor many betting markets and report suspicious behaviour on any, taking particular note when multiple markets are flagged as unusual.

ATASS Sports' odds data include both pre-match and in-play data for 274 matches from 2013 to 2016. They include matches from a range of different rounds and tournaments, from Challenger matches to Grand Slam level, with a mixture of hard and clay court matches. However, all player names and match details are omitted from this final thesis to avoid any accusations of besmirching individual players' reputations.

Table 5.1.1 shows how many matches of each surface and level were included in the data. We chose to focus only on two surfaces, clay and hard, in case we wished to do a comparative study of the differences between these surfaces. For such a study, we felt it would be better to have more matches on two surfaces than fewer matches across

	Hard	Clay	Total
Grand Slam	23	13	36
Masters 1000	38	26	64
Other ATP Tour	96	41	137
Challenger	15	22	37
Total	172	102	274

Table 5.1.1: A breakdown of the 274 matches for which we have odds data by surface and tournament level.

three surfaces by also including grass court matches. On the other hand, we decided to take a spread of matches from Grand Slam level all the way down to Challenger level, in case we wanted to investigate any features that gradually changed as the importance of the tournament increased or decreased. In this case, we felt it would be important to cover every different level in order to best highlight any gradual changes, even if that reduced the number of matches in each category. However, we eventually decided not to prioritise the analysis of surface or tournament level in order to focus instead more on odds data due to time constraints.

The in-play data are processed to provide a summary of the betting odds and volumes in each match during the breaks between games on a single betting exchange. This will allow us to explore how the odds vary throughout the match as the score changes from break to break.

To obtain the summary of the odds in each game break, the data used for this project underwent some pre-processing before use. In ATASS' original raw dataset, the best available price for each player is recorded at regular intervals on a single betting exchange. These prices will change gradually as the match develops, but also fluctuate randomly, even when the score is not changing. Because of this variation, it is necessary to summarise the odds recorded at different times in a game break to

get a sense of the state of the market during that break in play. Using a summary statistic for this helps cut out extraneous information.

Crucially, while the starts and end times of games are time-stamped to the nearest second in the data, the reaction of the exchange to the results of new points is swift, but not immediate. This can be due to the different times at which people observe the outcome of a point because of delays in internet or television signal. We do not have access to the original data to demonstrate the extent of this effect, but Croxson and Reade (2011) study the speed at which information is reflected in one exchange compared with bookmakers and show that while the exchange reacts much quicker, even there the change is not instantaneous.

As such, it is common for the odds data at the starts of the intervals between games, and to a lesser extent the ends, to fail to accurately reflect the current score. At the ends of important games in the match, such as breaks of serve and then ends of sets points, the differences in predicted win-probability after the game can be very large, and as such we want to omit all odds that reflect incorrect score information.

We therefore chose to calculate the median of the odds recorded in each game interval, as this would remove any drift at the starts and ends of the intervals, and outliers have less influence over the median than the mean. However, not all of the odds recorded during a game break will be equally informative, and so we wish to perform a further processing step. In order to explain why, it is necessary to introduce the concept of overround.

5.1.1 Overrounds in Betting Exchanges

The overround of the odds for all mutually exclusive outcomes at a given time is defined as the difference between 1 and sum of the reciprocals of the decimal odds. In a fair market, the overround would be 0, as the reciprocals of the odds would be probabilities. However, it is almost always greater than 0 in practice to provide a profit-making opportunity for bookmakers or odds layers on betting exchanges.

The over-round is a concept very closely linked to the bid-ask spread. The term "bid-ask spread" arises from financial markets, in which it is the difference between the lowest price currently available any seller is offering to sell the asset at (the ask price) and the highest price any buyer is offering for the asset (the bid price). While a spread exists, a stand-off of sorts ensues, as no transaction occurs until a buyer raises their bid to meet the ask price or a seller stoops to the bid price.

In betting exchanges with two participants, backing one player to win is the equivalent of laying the other, and the prices should reflect this. As such, it can be shown that the overround is a form of bid-ask spread, as discussed by Brown (2012) and Marginson (2013). The assets are bets on each player, and the odds form ask and bid prices.

There are many factors that can affect a bid-ask spread, as per Marginson (2013). One issue is the transaction cost, which in betting exchanges often takes the form of commission on profits. Buyers and sellers will choose prices to offset any such costs. A second issue is differences in opinion on the value of an asset. If those with an asset believe it to be more valuable than those wanting to buy, a bid-ask spread will emerge. Crucially, however, a bid-ask spread can also be closely linked to the liquidity in the market.

Liquidity is the ease with which transactions can be made in a market. If there is a large bid-ask spread, it is difficult for buyers and sellers to agree on prices, whereas it is much easier for to agreements on price to be made when the bid-ask spread is small. The markets for currencies are typically very liquid, as they have very low spreads, whereas the markets for rare paintings are much less liquid, as the value assigned to them varies much more among individuals, Kirkpatrick and Dahlquist (2010).

These examples also illustrate the interplay of liquidity and bid-ask spreads with supply and demand. Bid prices are linked to demand, in that if demand is high, bidders will have to raise their prices to beat their competition. If demand is low, they can afford to bid low prices and wait patiently for a seller to meet their demands.

Similarly, if supply is high a seller may need to lower the ask price, whereas low supply means prices can remain high.

As such, in betting exchanges the bid-ask spread can reflect the amount of competition in the market. If betting activity is generally low, prices (odds) may remain far apart, whereas if more gamblers are interested in an event, gamblers will compete to offer the best prices, driving the bid-ask spread (and the overround) down.

One consequence of low liquidity can be volatility, in that prices can move much more rapidly. Mike and Farmer (2008) cite low liquidity as one of the main causes of short-term market volatility in financial markets. This can be seen easily in betting exchanges. If the bid-ask spread is high, and there is a large difference between 1 and the sum of the odds, it is easy to undercut the best odds on offer. Once this bet has been matched it will no longer be available, and the odds may swing back to the previous best odds, or to a second new offer between the two. By contrast, if the bid-ask spread is low, much more sustained betting activity is required to shift the odds, and as such the odds remain more stable. Sarkissian (2016) and Roll (1984) discuss the links between bid-ask spread and market volatility.

When volatility is high and the odds are prone to sudden movements, it can be very hard to identify any signal (the implied probability of the event) from amongst the noise. Our goal is to establish what values of odds best represent the state of the market in each game break. How are we to do this when the odds are so variable? When the odds are more stable, it is much easier to identify clear signals. Kirkpatrick and Dahlquist (2010) discuss the difficulty in performing technical analysis of financial markets when liquidity is low.

As an example, suppose the best prices available for two players, A and B, are 1.25 and 2 respectively. The fair probabilities for these odds are hence $1/1.25 = 0.8$ and $1/2 = 0.5$, giving an overround of 0.3. Suppose a gambler believed player B had a 0.4 probability of winning. Offering any odds up to 1/0.4=2.5 for player B to win would give positive expected profit. As such, they can comfortably offer odds of 2.3,

for example, while expecting to profit, causing a large shift in the odds available. Another gambler with similar beliefs could then better this by offering odds of 2.4. Some new gamblers may then favour player A, snapping up these prices as quickly as they appeared. What then should we infer to be representative odds for this match? It is difficult to infer an "average" market opinion. For these reasons, when the overround is high the odds may provide unreliable information, and we seek to filter out these reported odds.

To summarise, high overrounds can be a sign of low competition in the betting market, which can in turn result in high volatility in the odds. When the odds are volatile, they provide less information about the state of the market due to the fact that they move so quickly from one moment to the next. As such, it can be very difficult to infer representative odds for the event. When the odds have low overround, they are typically therefore more informative.

One more cause of high bid-ask spreads to be aware of is the presence of informed traders in the market, Brown (2012), Marginson (2013). Should a traditional bookmaker believe a substantial number of informed traders to be operating in a market, they will tend to use higher overround to protect themselves against potential losses. This effect increases in markets with more uncertain outcomes, and the fear of missing information that informed traders have increases. This is one reason why horse races with more runners tend to have higher overrounds. Marginson (2013) also argue that this effect may also take place on betting exchanges, to the effect that odds layers may protect themselves against informed traders by offering odds more cautiously when the potential for information asymmetry is high. A corollary of this is that overrounds could be higher in fixed matches, when substantial information asymmetry exists, but we did not examine this further. Brown (2012) investigates overrounds at Wimbledon and concludes that high overrounds may be caused by traders with inside information in the form of some viewers' ability to observe the results of points before others, possibly due to a delay in television signal.

5.1.2 Pre-processing Odds Data

In order to summarise the state of the betting market in each game break, we therefore want to filter out sets of odds with a high overround before taking the median of odds recorded in each game break. However, there is a balance to be struck in deciding how to filter out sets of odds with high overround. If too many sets of odds with high overround are used, they may pollute the information provided by sets of odds with low overround. However, removing too many sets of odds with high overround will leave too few data points to give a representative summary of the market.

In order to strike this balance, we first split the sets of odds into bins according to their overround. The bins were [0, 0.02), [0.02, 0.05), [0.05, 0.1), [0.1, 0.2) and [0.2, ∞). We found that the best compromise between including unhelpful data and using too little data was found by using the ten sets of odds with the lowest overround, but only if they are in the lowest two bins. If there are less than ten odds in these bins, but more than zero, then all of the odds in those bins will be used, but no others. If there are no odds with overrounds in those bins, the next bin will be used in the same way, and so on.

The bottom two bins are therefore effectively one large bin for the purposes of this filtering. The processing step looks for the lowest bin (or merged bin) with any overrounds in, and uses all odds in that bin, up to a maximum of ten. No further bins are used. For example, if there were seven overrounds in the first bin and three in the fourth bin, only the first bin would be used. The unhelpful points in the fourth bin would be ignored. However, if there were only points in the fourth bin, these would be used. This means that sets of odds with low overround were targeted, and those with high overround were only used where necessary.

To give an idea of the reliability of the odds in each game break, the number of overrounds in each bin was recorded in the data. The sets of odds selected in each game break were then summarised by using the median of the selected odds for each player.

An alternative strategy to filtering for obtaining information from sets of odds with high and low overrounds would be to take a weighted average of the different odds recorded, weighting the different odds for each player according to the overround. We did not consider this for this thesis, however.

The pre-match data are collected and processed in exactly the same way, except the odds are organised according to different time windows before the start of the match, as obviously there are no game breaks pre-match. These windows are selected so that they are short just before the start of the match, but get longer the further away the match is. This is because odds move very little when the match is far away, but can move more as the match approaches and more and more gamblers become interested and start betting. Adding extra time windows far in advance of the match would hence add little value, but require extra storage space and processing time. The thresholds of the windows were (in minutes before match start) 0, 5, 20, 60, 120, 240, 480, 720 and 10,000.

5.2 Tennis Results Data

The second set of data used is a set of results of tennis matches. In order to assess whether betting odds are suspicious, it is common in the match-fixing literature to attempt to predict the odds by estimating the strength of both players (or teams in other sports). Betting odds and sports models can both predict sports matches well, so it should also be possible to make predictions about betting odds using sports models. Matches where odds do not conform to predictions can then be regarded as suspicious. In order to estimate the quality of tennis players the results of their past matches are required, as these are the only (publicly available) demonstrations of the players' quality. Our match data comes from `https://github.com/JeffSackmann/tennis_atp`, a repository which regularly updates with all of the latest tennis results. It is licensed for use under a Creative Commons Attribution-NonCommercial-ShareAlike

4.0 International License.

The data contain all results of ATP tour-level matches since 1968 (the start of the Open Era) as well as qualifying, challenger and futures matches since 1991. For each match, data include the identities of the players, the venue, the tournament, the round, and the final score in each set. In matches where they are available, there are also match statistics such as the number of points won and lost on serve by each player. This information is available for matches after 1991 for tour-level matches, but only more recently for the other matches. Some data are not recorded as the matches do not pass some sanity checks. The two examples cited are when the loser wins 60% of points or when the match time is under 20 minutes.

We also made some minor edits to the original data. We corrected the identities of players in four matches using data from tennislive.net when the original data recorded the same name as the winner and loser, and we also deleted one match in which the name of one player was not recorded.

The repository also contains results from WTA matches, but these are not used in this thesis. It is most convenient to focus on approximately homogeneous data, but men on average win more points on serve than women, as observed for example by Klaassen and Magnus (2001). It is therefore best to focus only on one gender. Since more data are available for men's tennis, this is what was used.

5.2.1 Data Selection

An important question to address is how much of the data to include. While it is tempting to use all available data, there are a few conflicting factors that must be balanced. First and foremost, it is important to include sufficient data such that players are ranked as accurately as possible at the times of the matches whose odds we wish to investigate for suspect activity. These all occur in from 2013 to 2016. We must therefore include enough historical matches to ensure players' ratings are modelled as well as possible by this time.

On the other hand, using additional matches introduces a computational cost for calculating Glicko ratings and optimising parameters to maximise predictive performance. In the most extreme case, were data available from 100 years ago or more it is highly doubtful it would improve our ability to model modern players. The earliest data available from Jeff Sackman's data are from 1968, so we must balance the data's ability to improve the quality of predictions against the computational burden of using them.

The other main factor to consider is that the composition of the data we have available changes over time. Figure 5.2.1 shows how the types of matches varies by number over time. Critically, the year 1991 sees the introduction of a large amount of Futures matches for the first time. The inclusion of these matches is important to be able to model the ratings of low-ranked players, but must be considered with caution. The addition of these matches may suddenly introduce a flood of players of lower quality than those playing in ATP matches before 1991, and the quality of the "average" player will take a sharp decrease. As such, players who play many matches before 1991 will be initially underrated compared to those that join after 1991, and it may be some time before this issue is resolved

While the composition of the data continues to alter over time, with more and more Futures, Challenger and qualifying matches over time, the biggest step chance is in 1991, and so we decided to use all data from 1991 onwards. We could possibly have used less data and still obtained very similar ratings for players in the matches from 2013-2016 for which we have odds data. Indeed, of all the players who played at least 10 matches in 2016, only 19.5% of players made their debuts before 2007, and Radek Stepanek was the first to make his debut, which happened in 1995. However, since the computational speed of our methods were not slowed significantly by using all data from 1991, we saw no reason to remove the early data.

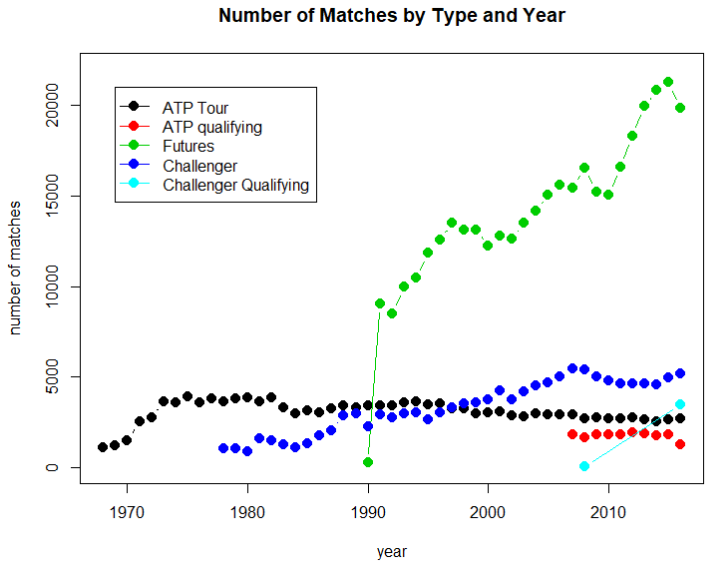

Figure 5.2.1: The number of matches of each type in each year in Jeff Sackman's data. (Data on Davis Cup and ATP tour finals are omitted from this plot, since the number of matches is small and varies very little by year.)

Chapter 6

Pre-Match Odds Modelling

In this chapter our objective is to build a model that describes the odds in the pre-match market of a tennis match as the start of the match approaches. Previous work in the literature focusses on the difference between closing and opening odds, but we aim to build a more sophisticated model that avoids some of the biases this can cause. It is possible for the opening odds to be poorly set, and the market quickly converges to a new position that gamblers believe is more representative of the two players. This is not suspicious behaviour, and is not particularly important to flag, as the odds can generally be quite variable at low volumes. However, further into the pre-match market once the odds are better established, it is more interesting when swings occur. This may be due to innocent reasons, such as injury news, or it may be evidence of a fix. We want to develop a model that flags such matches as anomalous so that they could potentially be investigated further for suspicious activity.

Our method will advance on current literature by investigating the use of betting volumes to identify low-volume swings that are of less interest, as well as looking at swings at various intervals throughout the pre-match market to help identify later, potentially more informative swings.

In our model, for each match separately the pre-match odds are randomly distributed around some constant mean, designed to represent "fair odds". However,

as the market becomes more liquid as the match approaches as more people gamble, we expect the odds to become more informative, and hence the variance of the odds will decrease as the match approaches. We shall investigate the use of both time and betting volumes to model this decreasing variance, and pool information across matches to estimate the correlation structure and usual amounts of variability in the odds.

We begin with a brief description of our odds data before defining Gaussian processes, which will be key to our model, and outlining a specific Gaussian process to represent the pre-match betting market, with the goal of flagging matches that do not fit this model well as being potentially worthy of further investigation.

6.1 Comparing Odds Data with Probabilities

In order to model pre-match odds data, we use odds data for 274 matches from ATASS Sports, described in Section 5.1. In this chapter and throughout this thesis we wish to use decimal odds to estimate probabilities of events occurring, or to compare odds to probabilistic predictions from those events. In a fair bet the probability corresponding to the odds would be the reciprocal of the decimal odds, but the bookmakers quote odds in their favour to make profit, and so the sum of these reciprocals is generally greater than 1. The difference between 1 and this sum is the overround, discussed in more detail in Section 5.1.1. Since these reciprocals sum to greater than 1, they will be consistenly higher than corresponding probability estimates. Attempting to estimate these odds by first estimating match-win probabilities will therefore lead to estimates that are typically too low. We therefore wish to account for this overround by transforming these reciprocals into probabilities that best represent the odds. For our tennis matches where the only events possible are either of the two players winning, we call these probabilities the implied win probabilities. There is no single way of obtaining these implied win probabilities, but two common methods exist to estimate

them.

Suppose there are n mutually exclusive outcomes where event i has decimal odds o_i, and we wish to find a probability p_i to represent the odds o_i for each i. The simplest method to turn odds into probabilities would be to find some constant k such that $\sum_{i=1}^{n} p_i = 1$, where

$$p_i = k o_i^{-1}, \quad i = 1, \ldots, n.$$

This would yield $k = \sum_{i=1}^{n} o_i^{-1}$.

We instead use Khutsishvili's formula, Vovk and Zhdanov (2009), which dictates that

$$p_i = o_i^{-\zeta}, \quad i = 1, \ldots, n$$

for some ζ such that $\sum_{i=1}^{n} p_i = 1$. Vovk and Zhdanov (2009) find that using Khutsishvili's formula provides better predictive accuracy than multiplicative normalisation. The other advantage of using Khutsishvili's formula concerns bets on multiple independent events, though this is not of particular use in this thesis. When bookmakers offer odds on two independent events, $i = 1$ and $i = 2$, with odds o_1 and o_2, both occurring, the odds quoted will typically be $o_1 o_2$, and the probability of both events occurring is $p_1 p_2$. Under the slightly simplified assumption that a bookmaker generates odds using some function $o_i = f(p_i)$ for all events i, then this function should therefore have the property that $f(p_1 p_2) = f(p_1) f(p_2)$. Such a function can always be found by taking $f(p_i) = p_i^{\zeta}$ for some ζ. To instead convert from odds to implied win probabilities, this function is inverted, giving Khutsishvili's formula.

Recall that the overround is the difference between 1 and $\sum_{i=1}^{n} o_i^{-1}$, the unnormalised reciprocals of the odds. If the overround is very small, and the reciprocals already sum to a value close to 1, the method use to obtain probabilities makes little difference. However, if the overround is higher, the different methods may yield very different results. The probabilities obtained when the overround is high may therefore be less informative than when the overround is low, in addition to the reasons

described in 5.1.1

6.2 Gaussian Processes

In order to model pre-match odds, we shall use Gaussian processes, Diggle and Ribeiro (2007). A Gaussian process in one dimension is a stochastic process $\{Y(x); x \in \mathbb{R}\}$ such that at every finite set of points $(x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}$ for $i = 1, \ldots, n$, the vector $\big(Y(x_1), \ldots, Y(x_n)\big)$ has a multivariate normal distribution for all n. For some mean vector $\boldsymbol{\mu}$ and variance matrix Σ, this means

$$\big(Y(x_1), \ldots, Y(x_n)\big) \sim MVN(\boldsymbol{\mu}, \Sigma).$$

A higher-dimensional spatial Gaussian process is one in which on any finite set of locations $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with each $\boldsymbol{x}_i \in \mathbb{R}^m$ for some $m \in \mathbb{Z}^+$, the joint distribution of the random variables $Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_n)$ is multivariate normal. Any such process is defined by its mean function, $\mu(\boldsymbol{x})$, and covariance function $\zeta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \text{Cov}(Y(\boldsymbol{x}_i), Y(\boldsymbol{x}_j))$ for all i and j.

A Gaussian process is stationary if $\mu(\boldsymbol{x}) = \mu$ for all $\boldsymbol{x}$, and $\zeta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \zeta(\boldsymbol{d}_{ij})$, where $\boldsymbol{d}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$, and so the covariance only depends on the difference between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The Gaussian process is isotropic if $\zeta(\boldsymbol{d}_{ij}) = \zeta(||\boldsymbol{d}_{ij}||)$, where $||\cdot||$ denotes Euclidean distance.

There are many different possible options for functions ζ. In this project, we have chosen to focus on the exponential correlation function,

$$\zeta(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho^{||\boldsymbol{x}_i - \boldsymbol{x}_j||},$$

since we believed this would be simple enough, and sufficient to model our data, given that we do not sample Y often at locations that are sufficiently close to reliably identify a more complex structure. However, investigating the use of other correlation functions could be a sensible extension to our work.

6.3 A Gaussian Process for Pre-Match Odds

We wish to model the pre-match odds with a constant mean and decreasing variance, and so fit a Gaussian process in each match k. We record odds at different times τ, with the rate of sampling getting higher as the match approaches, and also record betting volumes at each such time. (A fuller description of the data is provided in Section 5.1). These betting volumes increase with time, as they correspond to the total gambled up until that time. We expect odds recorded at times that are close together to be similar, but also expect that if the betting volumes have not changed much, the odds will not have changed much either. Hence, either time or volume could be useful to model the correlation of odds recorded at different times. As such, we will investigate using either of these or both as a "space" on which the odds are recorded. We also believe the overround might affect variances, but not correlation, because of our discussion in 6.1 about how high overrounds can lead to less reliable implied win-probabilities, which might result in a higher variance in our model

We therefore let $\boldsymbol{x}_{k,i}$ be a vector including the time, volume and overround of the i-th recorded data point in match k, (even though the overround does not have a spatial interpretation) and let $\boldsymbol{x}_k$ be a matrix containing all vectors $\boldsymbol{x}_{k,i}$ for match k.

We let $Y_k(\boldsymbol{x}_{k,i})$ be a random variable for the logit of the implied-win probabilities at location $\boldsymbol{x}_{k,i}$, and let $y_k(\boldsymbol{x}_{k,i})$ be an observation of this random variable. We take the logit of the implied win-probabilities to ensure that $Y_k(\boldsymbol{x}_{k,i})$ is unbounded, as the implied win probabilities are in [0,1], and having unbounded variables enables Gaussian process models to give a better approximation to the data.

We hence model the logit implied win probabilities such that at each $\boldsymbol{x}_{k,i}$, the logit implied win-probability has constant mean ω_k and marginal variance $\delta^2 \exp(\boldsymbol{x}_{k,i}\boldsymbol{\beta})$, so that

$$Y_k(\boldsymbol{x}_{k,i}) \sim N\big(\omega_k, \delta^2 \exp(\boldsymbol{x}_{k,i}\boldsymbol{\beta})\big) \text{ for all } k \text{ and } i. \tag{6.3.1}$$

The parameter δ^2 is a multiplier to the variance common to all matches, and the vector of parameters $\boldsymbol{\beta}$ controls how $\boldsymbol{x}_{k,i}$ affects the variability of the odds. Having these parameters common to all matches allows us to pool information across the different matches about how much variation in the pre-match odds is common, and at what level it becomes worthy of further investigation.

As an alternative parameterisation, we could let $\boldsymbol{x}^+_{k,i} = (1, \boldsymbol{x}_{k,i})$ and $\boldsymbol{\beta}^+ = (\log(\delta^2), \boldsymbol{\beta})$. This parameterisation leads to

$$Y_k(\boldsymbol{x}_{k,i}) \sim N\big(\omega_k, \exp(\boldsymbol{x}^+_{k,i}\boldsymbol{\beta}^+)\big) \text{ for all } k \text{ and } i.$$

This looks more like a traditional model using linear regression to fit the variance. However, we prefer the format in equation (6.3.1) due to the fact that later we can find the maximum likelihood estimator for δ^2 conditional on other parameters.

In order to model correlation between implied win-probabilities at successive times, we chose to use an exponential correlation function, since we believed this would be simple and powerful enough for our relatively simple Gaussian process. We wanted to explore whether modelling correlation over time or betting volumes provided better results, so considered both separately. Therefore to measure correlation in the logit of the implied win-probabilities using the data from $\boldsymbol{x}_k$, we let

$$\mathrm{Cor}\big(Y_k(\boldsymbol{x}_{k,i}), Y_k(\boldsymbol{x}_{k,j})\big) = \zeta(\boldsymbol{x}_{k,i}, \boldsymbol{x}_{k,j}) = \rho^{||\boldsymbol{x}_{k,i} - \boldsymbol{x}_{k,j}||}, \quad i, j = 1, \ldots, n_k.$$

As well as only considering correlation over time and volume separately, we could also have considered modelling correlation jointly over the two-dimensional space of time and volume. However, since time and volume are on very different scales, the correlation function would need adjusting to account for this. Doing so would be interesting to investigate further, but we decided not to do so for this project.

Combining this correlation function with the marginal distributions of each $Y_k(\boldsymbol{x}_{k,i})$, we let $\boldsymbol{Y}_k(\boldsymbol{x}_k)$ be a vector of all $Y_k(\boldsymbol{x}_{k,i})$ for $i = 1, \ldots, n_k$. Since $Y(\boldsymbol{x}_{k,i})$ has mean ω_k, note that $\boldsymbol{Y}_k(\boldsymbol{x}_k)$ has mean $\omega_k \mathbf{1}_{n_k}$, where $\mathbf{1}_n$ is a vector of length n in which every element has value 1, and n_k is the number of data points in match k. In our data, n_k

generally equals 8, but is sometimes smaller.

This leads to the model

$$\begin{aligned} \boldsymbol{Y}_k(\boldsymbol{x}_k) &\sim MVN\Big(\omega_k \mathbf{1}_{n_k},\ \ \delta^2 R_k\Big), \\ R_k &= D_k C_k D_k, \end{aligned} \tag{6.3.2}$$

where

$$D_k = \operatorname{diag}\Big(\exp\Big(\frac{\boldsymbol{x}_k\boldsymbol{\beta}}{2}\Big)\Big) = \begin{pmatrix} \exp\left(\frac{\boldsymbol{x}_{k,1}\boldsymbol{\beta}}{2}\right) & 0 & \dots & 0 \\ 0 & \exp\left(\frac{\boldsymbol{x}_{k,2}\boldsymbol{\beta}}{2}\right) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \exp\left(\frac{\boldsymbol{x}_{k,n_k}\boldsymbol{\beta}}{2}\right) \end{pmatrix}, \tag{6.3.3}$$

and C_k has (i,j)-th element

$$C_{k,ij} = \zeta(\boldsymbol{x}_{k,i}, \boldsymbol{x}_{k,j}) = \operatorname{Cor}\big(Y_k(\boldsymbol{x}_{k,i}), Y_k(\boldsymbol{x}_{k,j})\big) \text{ for all } k,\ i \text{ and } j. \tag{6.3.4}$$

Using $R_k = D_k C_k D_k$ in this way means that

$$\begin{aligned} \operatorname{Cov}(Y_k(\boldsymbol{x}_{k,i}), Y_k(\boldsymbol{x}_{k,j})) &= \operatorname{Cor}\big(Y_k(\boldsymbol{x}_{k,i}), Y_k(\boldsymbol{x}_{k,j})\big)\sqrt{\operatorname{Var}(Y_k(\boldsymbol{x}_{k,i}))\operatorname{Var}(Y_k(\boldsymbol{x}_{k,j}))} \\ &= \delta^2 \exp\Big(\frac{1}{2}(\boldsymbol{x}_{k,i} + \boldsymbol{x}_{k,j})\boldsymbol{\beta}\Big)\rho^{||\boldsymbol{x}_{k,i} - \boldsymbol{x}_{k,j}||} \text{ for all } k,\ i \text{ and } j. \end{aligned}$$

For match k this Gaussian process has constant mean, but not constant variance due to the matrix D_k. Note, however, that the Gaussian process

$$D_k^{-\frac{1}{2}}(\boldsymbol{Y}_k(\boldsymbol{x}_k) - \omega_k \mathbf{1}_{n_k}) D_k^{-\frac{1}{2}}$$

is stationary, isotropic and identically distributed for all k.

Additionally, we can add a so-called "nugget effect". This term permits some uncorrelated error between nearby points in the Gaussian process. This is often used to represent measurement error, and permits nearby points, or even points occurring at the same time, to exhibit differences, even though nearby points are also correlated. This was necessary when using volume to measure correlation, as some matches had

measurements at different times with the same volume but slightly different odds. Such realisations would be impossible for a Gaussian process without a nugget effect, as then two sites with distance 0 between them would have perfectly correlated odds. Although the odds mostly move when bets are matched and so the volume increases, it appears that the odds can also move when gamblers offer better odds than before without matching previous bets. We therefore require a nugget effect to allow for occasions when we have slightly different odds at sites at different times but with the same volume.

For an ordinary Gaussian process with constant marginal variance σ^2 and correlation matrix C, adding a nugget effect s^2 corresponds to replacing the variance matrix $\Sigma = \sigma^2 C$ with $\tilde{\Sigma} = \sigma^2 C + s^2 I_n$. This means that $\mathrm{Var}(Y(x_i)) = \sigma^2 + s^2$, and $\mathrm{Cor}(Y(x_i), Y(x_j)) = \sigma^2 C_{ij}/(\sigma^2 + s^2)$ for all i and j, where C_{ij} is the (i,j)-th element of C. It can be useful to also consider the relative nugget, namely $\eta^2 = s^2/\sigma^2$, so that $\tilde{\Sigma} = \sigma^2(C + \eta^2 I_n)$, and $\mathrm{Cor}(Y(x_i), Y(x_j)) = C_{ij}/(1 + \eta^2)$, as this helps us fit the nugget effect later.

For our model in equation (6.3.2), with $\Sigma_k = \delta^2 D_k C_k D_k$, there are two options for how to apply the nugget. These are

$$\tilde{\Sigma}_k = \delta^2 D_k (C_k + \eta^2 I_{n_k}) D_k, \tag{6.3.5}$$

$$\text{or} \quad \tilde{\Sigma}_k = \delta^2 (D_k C_k D_k + \eta^2 I_{n_k}). \tag{6.3.6}$$

The former option means that the effect of the nugget will be proportional to the marginal variance as specified by D_k, while the latter means the nugget will have constant effect everywhere in match k. We shall explore which fits best. Note that we use a common relative nugget parameter η across all matches, so that information about this parameter can be pooled from across different matches to help us identify which matches behave unusually. We shall let

$$\tilde{R}_k = \delta^{-2} \tilde{\Sigma}_k$$

denote either of the two covariance matrices in equations (6.3.5) and (6.3.6) divided by the common variance term δ^2, so that

$$\begin{aligned} \tilde{R}_k &= D_k(C_k + \eta^2 I_{n_k})D_k, \\ \text{or} \quad \tilde{R}_k &= (D_k C_k D_k + \eta^2 I_{n_k}). \end{aligned} \tag{6.3.7}$$

6.4 Fitting the Gaussian Process - Maximum Likelihood

We wish to optimise over the parameters $\boldsymbol{\beta}$, ρ, η, δ^2 and ω_k for $k = 1 \dots 274$ in order to maximise the likelihood

$$\begin{aligned} L(\boldsymbol{\omega}, \delta^2, \boldsymbol{\beta}, \rho, \eta | \boldsymbol{y}, \boldsymbol{x}) &= \prod_{k=1}^{274} L(\omega_k, \delta^2, \boldsymbol{\beta}, \rho, \eta | \boldsymbol{y}_k, \boldsymbol{x}_k) \\ &\propto \prod_{k=1}^{274} \det\left(\delta^2 \tilde{R}_k\right)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2\delta^2}(\boldsymbol{y}_k - \omega_k \mathbf{1}_{n_k})^\top \tilde{R}_k^{-1}(\boldsymbol{y}_k - \omega_k \mathbf{1}_{n_k})\Big), \end{aligned} \tag{6.4.1}$$

where $\boldsymbol{\omega} = (\omega_1, \dots, \omega_{274})$, $\boldsymbol{y}$ contains all $\boldsymbol{y}_k$, $\boldsymbol{x}$ contains all $\boldsymbol{x}_k$, and $\tilde{R}_k$ is defined by equations (6.3.3), (6.3.4) and (6.3.7).

It is not possible to use analytic methods to obtain maximum likelihood estimates for $\boldsymbol{\beta}$, ρ and η. However, given these parameters we can obtain profile maximum likelihood estimators for the parameters ω_k and δ, which shall be called $\hat{\omega}_k(\boldsymbol{\beta}, \rho, \eta)$ and $\hat{\delta}(\boldsymbol{\beta}, \rho, \eta)$, by maximising the log likelihood with respect to each ω_k and δ, while fixing $\boldsymbol{\beta}$, ρ and η, yielding

$$\begin{aligned} \hat{\omega}_k(\boldsymbol{\beta}, \rho, \eta) &= \frac{\mathbf{1}_{n_k}^\top \tilde{R}_k^{-1} \boldsymbol{y}_k}{\mathbf{1}_{n_k}^\top \tilde{R}_k^{-1} \mathbf{1}_{n_k}}, \\ \hat{\delta}^2(\boldsymbol{\beta}, \rho, \eta) &= \frac{\sum_{k=1}^{274} (\boldsymbol{y}_k - \hat{\omega}_k(\boldsymbol{\beta}, \rho, \eta)\mathbf{1}_{n_k})^\top \tilde{R}_k^{-1} (\boldsymbol{y}_k - \hat{\omega}_k(\boldsymbol{\beta}, \rho, \eta)\mathbf{1}_{n_k})}{\sum_{k=1}^{274} n_k}. \end{aligned}$$

In order to obtain maximum likelihood estimators for $\boldsymbol{\beta}$, ρ and η, we therefore use numerical methods to find values $\hat{\boldsymbol{\beta}}$, $\hat{\rho}$ and $\hat{\eta}$ that maximise the log likelihood

in equation (6.4.1) where ω_k is replaced by $\hat{\omega}_k(\boldsymbol{\beta}, \rho, \eta)$ for each $k = 1 \ldots 274$, and δ^2 by $\hat{\delta}^2(\boldsymbol{\beta}, \rho, \eta)$. It is here that using constant η across all matches proves helpful - maximising this likelihood over this smaller number of parameters is relatively straightforward, but would be much more challenging were separate parameters required for each of our 274 different matches. Extending this to other matches would be even more troublesome, given the thousands of matches that occur each year.

6.4.1 Results

Earlier we discussed some of the different features thought to affect the accuracy of the odds in the pre-match markets: the overround, betting volume, and time until match start. We shall explore how these three factors affect the model fit and choose a model that appropriately balances predictive performance with model complexity.

We shall also investigate whether transforming the overrounds and volumes by taking logarithms would provide better fit. However, there are some observations of 0 in both volume and overround, which makes taking logs problematic. Instead of taking overrounds, we shall therefore consider the sum of both player's unnormalised implied win probabilities, given by the overround + 1. We will also look at $\log(£1 + \text{volume})$ instead of volume. The volume is generally much bigger than £1 when it is non-zero, so the arbitrary shift of £1 should not differ much from a log transform for most volumes.

In order to select the most appropriate model, we evaluated AIC and BIC for each possible model. The options to consider were as follows:

- Which variables to include to model correlation.
- Which variables to include to model the changing variance of implied win-probabilities in the pre-match market for a match.
- How to include a nugget effect, or whether or not to include one.

Across our 274 matches, there were four occasions when odds data were available but volume data were not. So that models were evaluated over the same data we ensured that these were excluded from all our analyses, even in models where volume was not used.

After investigating all of the options, we obtained a few key findings. Firstly, overround had no real impact in our model, and so is henceforth excluded. Secondly, while including a mixture of time and volume (or a log(1+volume)) to model changing variance appeared to give high log likelihoods, this came at the expense of the model's conformation to expected behaviour. Instead of seeing variance decrease as the match approached, we instead saw an increase. The reasons for this are unclear.

We believe one key factor in this was the potential collinearity of the two variables. Time and volume both increased together, so including both made fitting the model difficult. The maximum volume varied greatly by match, but if volumes were standardised by dividing the volumes in each match by the final volume to look at how similarly time and volume increased across all matches, we found that time and standardised volumes had a correlation of 0.74, which may high enough to cause potential issues. However, the issues may also be due to confusing the two different sorts of correlation occurring.

We expect the odds to be similar at similar times, but also expect that the odds can only move significantly if gambling occurs, and will remain steady when little gambling activity occurs. In matches where the odds initially remain stable with little gambling activity, the model appears to consider these early odds to provide more of a pattern than the later, rapidly-shifting odds, and therefore weights the early odds more strongly, putting ω_k near these early odds. To compensate, the variance then increases as the match approaches, as shown in Figure 6.4.1. This plot is a post-hoc display of how the pre-match odds compare to the maximum likelihood parameters obtained after observing the entire pre-match market. It is striking how much weight is placed on the early odds observations, and how little weight is placed on the odds

near the end.

Unfortunately, we failed to conclusively determine the causes of these issues. It was unclear what inference to draw from the model in this scenario, and whether the problems were caused by mixing time and volume as covariates. As such, we found it more productive to avoid mixing time and volumes, instead pressing on with research that used one covariate or the other. It would be prudent, however, for future researchers to identify the cause of this strange behaviour and identify solutions.

The models that worked best were therefore those that used a single variable

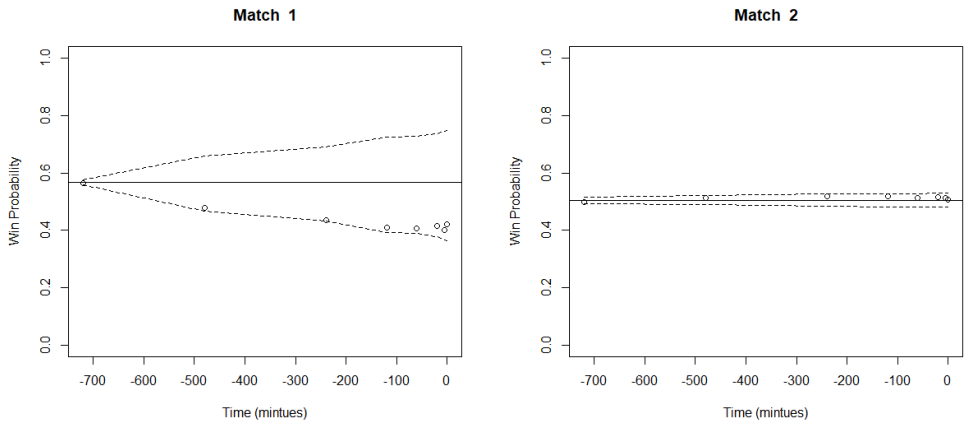

Figure 6.4.1: Maximum likelihood fits for two example matches using time to model correlation and both time and log(1+volume) to model changing variance. The solid line represents $\hat{\omega}_k$, the dashed lines represent 95% prediction intervals, and the dots represent $y_k(\tau)$.

to model changing variance and correlation. Time, volume and log(1+volume) all provided reasonable fits. However, log(1+volume) gave the best fit. We also found it best to include the nugget effect as in equation (6.3.5) rather than as in (6.3.6). A few examples of model fit are shown in Figure 6.4.2.

A word of warning must be applied, however. The parameter ω_k in each match is estimated using only at most eight data points. This provides substantial room for uncertainty in these estimated values. We therefore propose using Bayesian statistics to instead provide posterior distributions for the parameters in each match. This is

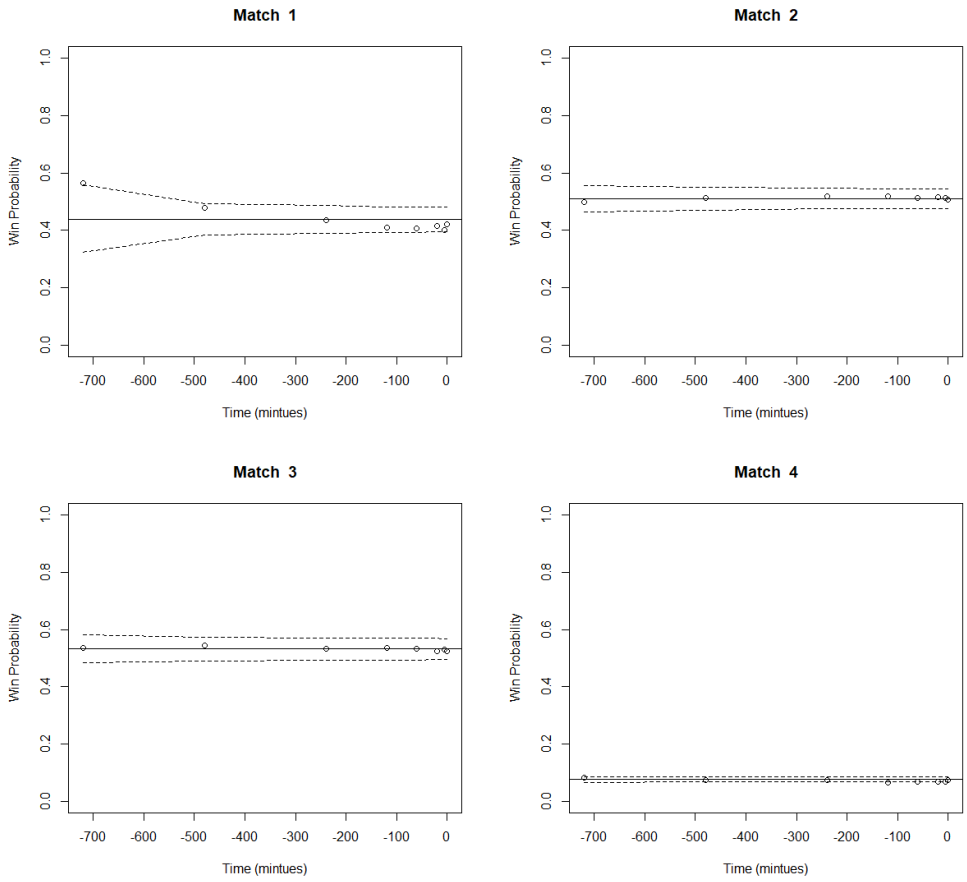

Figure 6.4.2: Maximum likelihood fits for four example matches using log(1+volume) to model correlation and changing variance. The solid line represents $\hat{\omega}_k$, the dashed lines represent 95% prediction intervals, and the dots represent $y_k(\tau)$.

helped by the fact that we have another source of information that should help us develop informative prior distributions for these parameters, namely the Glicko ratings. We shall use our implementation of Glicko ratings described in Chapter 4 to generate prior distributions for each ω_k, before updating these using the pre-match data, $\boldsymbol{y}_k$ to obtain posterior distributions over the parameters to account for the uncertainty involved.

6.5 Fitting the Gaussian Process - Bayesian Method

Using the likelihood for each match as in equation (6.3.2), we can see that a conjugate prior distribution for each parameter ω_k is

$$\omega_k \sim N(u_k, \gamma_k^2),$$

given hyperparameters u_k and γ_k^2.

Given these prior distributions, careful calculation reveals that the posterior distributions take the form

$$\begin{aligned} \omega_k|\boldsymbol{y}_k &\sim N(u_k^*, \gamma_k^{*2}), \\ \text{where} \quad u_k^* &= \gamma_k^{*2}\Big(\frac{\mathbf{1}_{n_k}^\top \tilde{R}_k^{-1}\boldsymbol{y}_k}{\delta^2} + \frac{u_k}{\gamma_k^2}\Big), \\ \text{and} \quad \gamma_k^{*2} &= \Big(\frac{\mathbf{1}_{n_k}^\top \tilde{R}_k^{-1}\mathbf{1}_{n_k}}{\delta^2} + \gamma_k^{-2}\Big)^{-1}. \end{aligned}$$

In order to set up a prior distribution for each ω_k, we can use the output of our Glicko ratings to provide information. We will do this by looking at the distributions of the fitted values $\hat{\omega}_k$.

Recalling notation from Section 4.1.1, the Glicko ratings θ_i for player i are distributed according to

$$\theta_i \sim N(\nu_i, \sigma_i^2).$$

(Note that we drop the dependence of these parameters on the time t that the match occurs (measured in weeks) for notational simplicity, to ensure it is not confused with time τ, measured in minutes to the start of the match).

Suppose match k is played between players i and j. In Figure 6.5.1 we plot ω_k against $\nu_i - \nu_j$ and find a strong relationship between the two. We hence let $u_k = \nu_i - \nu_j$, so that the prior for ω_k is

$$\omega_k \sim N(\nu_i - \nu_j, \gamma_k^2).$$

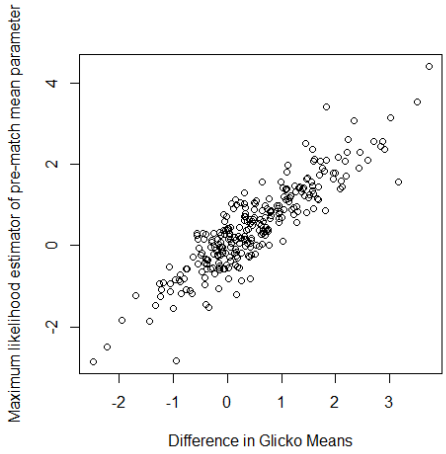

Figure 6.5.1: For each of our 274 matches with pre-match odds, we plot $\nu_i - \nu_j$ against $\hat{\omega}_k$.

We also want to see whether $(\sigma_i^2 + \sigma_j^2)$ helps us predict γ_k. We find that a slightly better log-likelihood is obtained by setting $\gamma_k^2 = \gamma^2(\sigma_i^2 + \sigma_j^2)$ for some estimated global constant γ^2. This means that in matches where the Glicko ratings are particularly uncertain, the variance in the pre-match prior mean parameter ω_k is also higher. Quantile-quantile plots for ω_k are shown in Figure 6.5.2, but the differences between the two are relatively minor.

Our prior distribution for ω_k is therefore

$$\omega_k \sim N(\nu_i - \nu_j, \gamma^2(\sigma_i^2 + \sigma_j^2)).$$

It is possible that if our Glicko ratings incorporated surface information, as in Section 4.4.3, and hence were able to better predict match outcomes, that the relationship between $\nu_i - \nu_j$ and ω_k would be even stronger, leading to a smaller estimate of γ^2, leading to more informative prior distributions.

6.6 Bayesian Model - Example Fits

We look at the fit of the implied win-probabilities for a few matches in Figure 6.6.1. These plots show 95% posterior confidence intervals for ω_k in each match k, as well

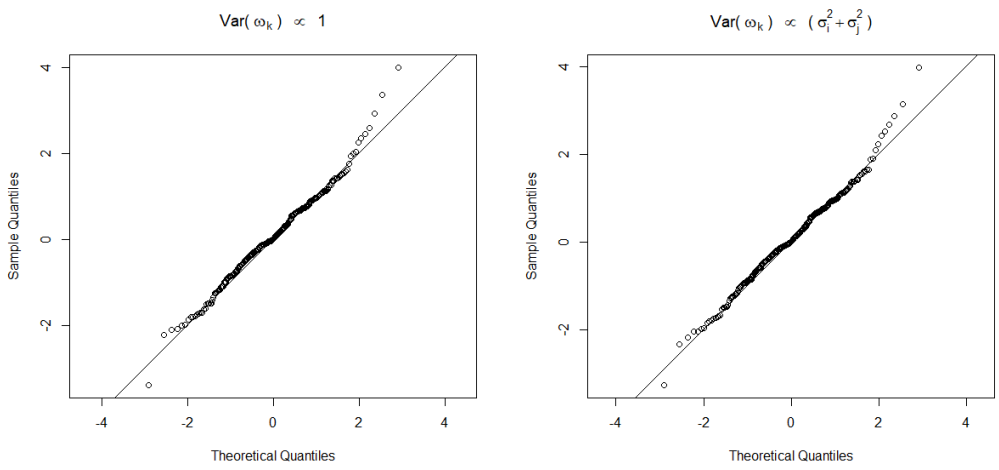

Figure 6.5.2: Quantile-quantile plots for $(\omega_k - (\nu_i - \nu_j))/\gamma_k$, where the variance γ_k^2 is proportional to different variables.

as 95% posterior predictive confidence intervals for $y_k(\boldsymbol{x}_k)$. These posterior predictive confidence intervals are obtained by considering

$$\int_{-\infty}^{\infty} f_Y(y_k(\boldsymbol{x}_{k,i+1})|\omega_k) f\big(\omega_k|y_k(\boldsymbol{x}_{k,1}),\ldots,y_k(\boldsymbol{x}_{k,i})\big)d\omega_k, \quad i=0,\ldots,n_k-1. \tag{6.6.1}$$

Note the important distinction between these and the more usual posterior predictive distributions for $y_k(\boldsymbol{x}_k)$. These would formally be given by looking at the distribution of $y_k(\boldsymbol{x}_{k,i+1})$ given $\{y_k(\boldsymbol{x}_{k,1}),\ldots,y_k(\boldsymbol{x}_{k,i})\}$, given by

$$f_Y\big(y_k(\boldsymbol{x}_{k,i+1})|y_k(\boldsymbol{x}_{k,1}),\ldots,y_k(\boldsymbol{x}_{k,i})\big) = \int_{-\infty}^{\infty} f_Y(y_k\big(\boldsymbol{x}_{k,i+1})|y_k(\boldsymbol{x}_{k,1}),\ldots,y_k(\boldsymbol{x}_{k,i}),\omega_k\big) f\big(\omega_k,|y_k(\boldsymbol{x}_{k,1}),\ldots,y_k(\boldsymbol{x}_{k,i})\big)d\omega_k.$$

and crucially takes into account the correlation between $y_k(\boldsymbol{x}_{k,i+1})$ and the previous observations. By contrast, the distribution in equation (6.6.1) ignores this correlation and looks only at the marginal distribution of $y_k(\boldsymbol{x}_{k,i+1})$ given the posterior parameters.

In each of these matches, the uncertainty in the win-probabilities and ω_k is large to begin with before decreasing as the pre-match market continues. In matches 1 and 2, the model is able to track the slow changes in pre-match odds, while in matches 3

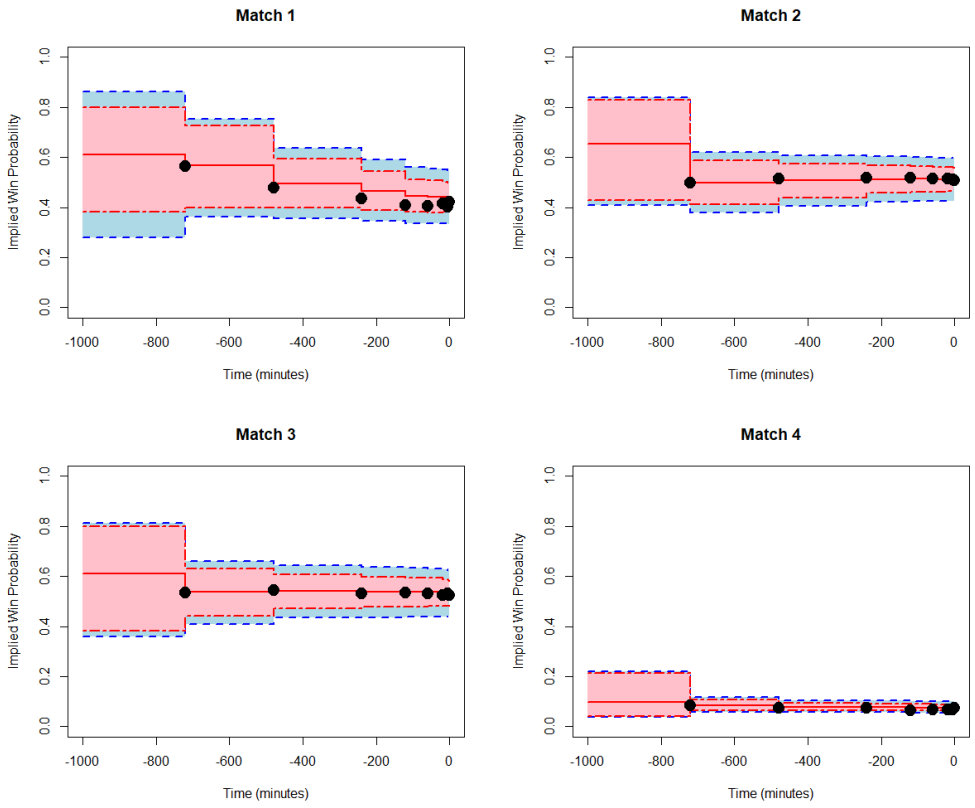

Figure 6.6.1: Successive 95% posterior confidence intervals for ω_k (in pink, surrounded by dots and dashes) and 95% unconditional posterior predictive confidence intervals *after* observing $y_k(\boldsymbol{x}_k)$ (in blue, surrounded by dashes). The solid red line is the posterior mean of ω_k, and black dots represent the values of $y_k(\boldsymbol{x}_k)$. The distribution at -1000 minutes represent prior distributions.

and 4 the odds move much less, and the uncertainty slowly decreases towards these values. In match 4, the uncertainty is very low due to the large amount gambled on that match.

It is interesting how much of the uncertainty in $y_k(\boldsymbol{x}_{k,i})$ is seemingly caused by the uncertainty in ω_k, given the relative proportions of their variances. From Figure 6.4.2, we see that if ω_k is well known then the fitted variance is quite small, particularly in matches with high betting volumes such as match 4. By comparison, the prior uncertainty in ω_k is quite large, but this can decrease quickly during the pre-match market.

In order to obtain an idea of how well the data in each match fit our model, we can look at posterior predictive p-values, that is, the p-values of the data $y_k(\boldsymbol{x}_{k,i+1})$ with respect to the unconditional posterior predictive confidence interval after the previous observation $y_k(\boldsymbol{x}_{k,i})$. These give an indication of how surprising each observation is with respect to the previously observed data, and hence mainly detects large shifts away from the mean at that single time instance. Looking at conditional predictive p-values would instead put the focus on large changes between successive values, focussing even more on short-term changes than the unconditional predictive p-values.

For each match, we look at the largest unconditional posterior predictive p-value to find the stage of the pre-match market that differs most from the mean, and we also look at the average p-value to get a broader picture of how the pre-match market differs from expected. We must be aware that the unconditional p-values may be highly correlated if there is sufficient swing away from the mean.

Ordinarily when assessing the fit of a Gaussian process we might expect to compare the log-likelihoods for different matches to properly account for the correlation structure of successive observations. However, this is complicated by the fact that we have used Bayesian statistics to put prior distributions on the parameters, and hence we only consider the p-values of individual points. It would be helpful if we could investigate more holistic summaries of how far the observed pre-match odds deviated from expected behaviour.

We plot the average and lowest p-values for each match in Figure 6.6.2. Figure 6.6.3 highlights the four matches with the lowest minimum p-value. Matches 209 and 73 are also the two matches with the lowest average p-values, so Figure 6.6.4 shows the two matches with the third and fourth lowest average p-values.

In match 73, we see a very large swing in the pre-match market that our model is unable to keep up with. This is precisely the sort of behaviour we wanted to highlight, and suggest that match 73 is worthy of further investigation. Of course, the swing could simply be due to injury news or other innocent factors, but the size of this swing

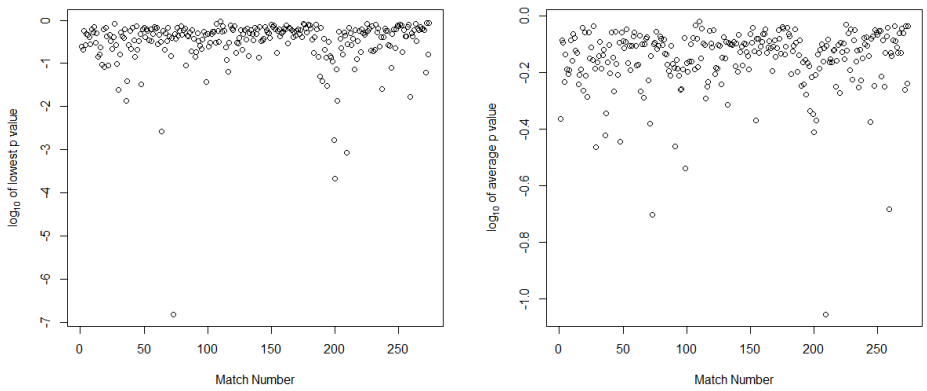

Figure 6.6.2: The lowest and average unconditional posterior predictive p-values for each of our 274 matches.

is sufficiently large and fast compared to the other matches in our data set to mark it out as anomalous.

On the other hand, matches 200 and 199 both have their lowest p-value on the very first observation. In these cases, the prior distributions given by the Glicko ratings disagree with the initial bookmaker assessments of the matches, but the pre-match markets otherwise behave very normally. If we have complete confidence in the Glicko rating's ability to predict pre-match odds in normal matches, we would be very suspicious of these differences and would wish to investigate further. However, with a sample size of 274 matches, we expect a couple of low p-values to occur randomly - there is some room for error in our Glicko ratings, so while it is still worth flagging these matches, the fact that the only issue is disagreement with the Glicko ratings means these matches are less of a cause for concern.

Match 209 presents a very interesting case. The opening pre-match odds disagree greatly with the pre-match priors, but the pre-match odds then move back almost exactly to what the Glicko priors suggested. The first p-value is therefore very small, but the speed of the market's return to the level expected is also much faster than expected. It is possible that the opening odds were very different to public opinion

about the match, so gamblers who agreed more with our model seized upon this potentially profitable opportunity and bet on the match much quicker than expected. On the other hand, the initial p-value could be due to poor modelling by Glicko, as we suggested as an explanation for matches 200 and 199, and the large swing after would then be much more worthy of investigation - though injury news could again supply an innocent explanation.

We next come to the matches with low average p-values in Figure 6.6.4. These

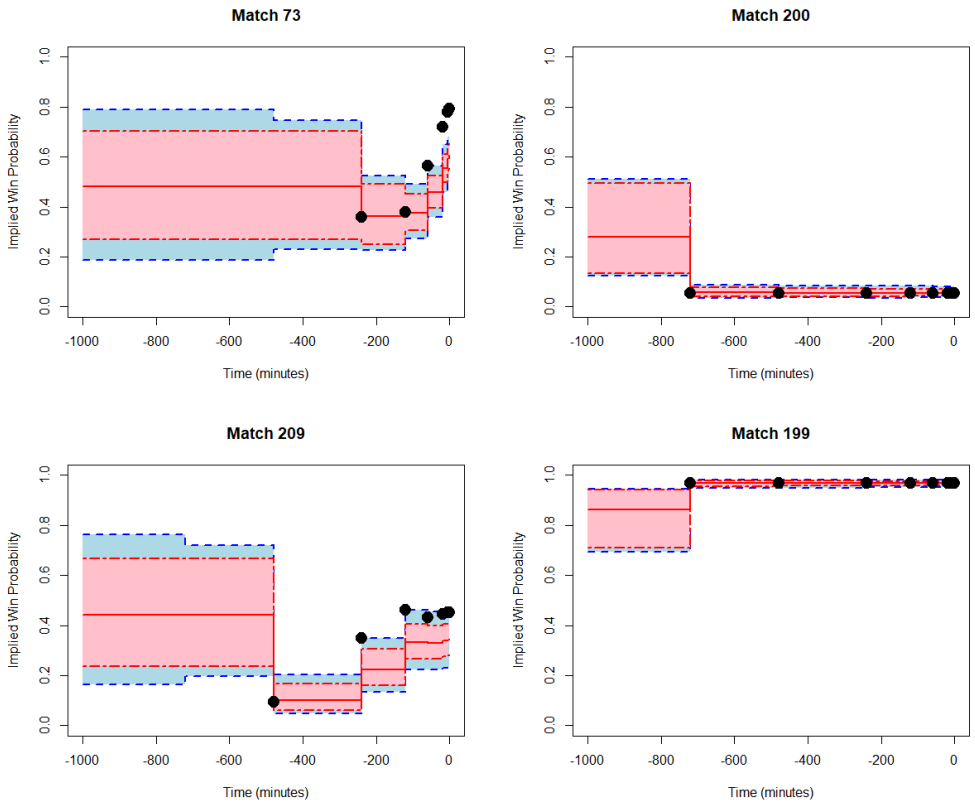

Figure 6.6.3: Four matches with the lowest unconditional posterior predictive p-values at any time. Successive 95% posterior confidence intervals for ω_k (in pink, surrounded by dashes) and 95% unconditional posterior predictive confidence intervals *after* observing $y_k(\boldsymbol{x}_k)$ (in blue, surrounded by dots and dashes). The solid red line is the posterior mean of ω_k, and black dots represent the values of $y_k(\boldsymbol{x}_k)$. The distribution at -1000 minutes represent prior distributions.

are the matches with the third and fourth-lowest average p-values, beaten by matches

73 and 209 (Figure 6.6.4), both of which had huge pre-match swings. The swing in match 259 is more modest. In 259, there's a swing of at least 15 percentage points, and while the model just about tracks this shift, the fact that every p-value is so low is enough to give this match the third-lowest average p-value.

The swings in these matches, while subtle, may be worthy of further investigation, but also serve to highlight another important issue to consider, which is the direction of the swing. If we believe a swing in pre-match odds is caused by corrupt betting activity, then we would expect far more than normal to gambled on the eventual winner, seeing an odds swing in their direction. Although it has not been important so far, our data are set up so that the probabilities reported are always for the eventual winner. Hence, if we are performing post-match analysis of the market, it may be that we are only interested in large pre-match increases in probability for the eventual winner, but not large decreases. A large pre-match decrease in implied win probability would suggest a surge of gambling on the eventual loser to win, suggesting that many people would lose their money. The reasons for such a swing may therefore be more complex. This could simply be due to inaccurate opening odds, or perhaps news of an injury caused a pre-match swing, but the injury was not as problematic as had been feared. On the other hand, it could also be caused by a failed attempt to fix a match, or false rumours of a fix. We have considered two-sided p-values so far. While it may be that considering only one-sided p-values in the direction of the eventual winner may be a stronger indication that suspicious betting activity has occurred, it is also not clear that one can completely disregard swings in the opposite direction, depending on their scale.

All of this only applies, of course, if the eventual winner is known. Sometimes it might be desirable to know pre-match or in-play whether a large shift in odds is occurring. Betting could then be suspended to prevent further opportunities for fixers to profit from the match. In this case, the direction of the swing is much less important. The eventual winner is not known, and so the direction of the swing is

meaningless.

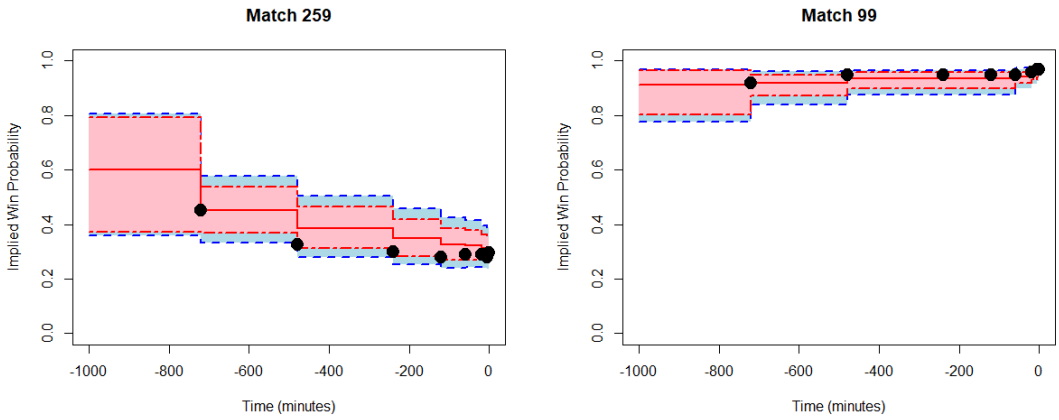

Figure 6.6.4: Two matches with among the lowest average unconditional posterior predictive p-values. Successive 95% posterior confidence intervals for ω_k (in pink, surrounded by dashes) and 95% unconditional posterior predictive confidence intervals *after* observing $y_k(\boldsymbol{x}_k)$ (in blue, surrounded by dots and dashes). The solid red line is the posterior mean of ω_k, and black dots represent the values of $y_k(\boldsymbol{x}_k)$. The distribution at -1000 minutes represent prior distributions.

6.7 Conclusion

In this chapter, we developed a Gaussian process model for pre-match implied win probabilities which fitted a constant mean to each match and a variance that decreased as the match start approached according to the increase in betting volumes. Matches with little pre-match odds movement fitted this model well, while matches with large pre-match swings fitted the model poorly. The goal was to identify these matches with the large pre-match swings. We pooled information from across all of the different matches in the data to establish how much variability could be expected in normal circumstances, and identified matches that were anomalous.

By using betting volumes to measure the decrease in variance, we were more sympathetic to odds movements in matches with little betting activity, provided they occurred over a long enough stretch of time. On the other hand, using volumes to

measure correlation of odds meant that huge swings in odds with very little money gambled were also identified as strange. Using volumes in this way represents an advancement on existing literature.

The rest of the literature also focusses only on the difference between the opening and closing odds. We were able to obtain extra information by looking at p-values at various intervals during the pre-match market. Focussing only on the difference between opening and closing odds also risks the opening odds being very poorly formed, leading to an unrealistic picture of the size of swing. Our use of volumes helped to mitigate this effect.

However, the behaviour of the model when combining time and volume as covariates remains perplexing. Further research should investigate the causes of this issue, and how it may affect the rest of our analyses.

We looked at the lowest p-values in each match to look for large short-term swings, and average p-values to look for matches which saw small but consistent swings in each time interval. There may, however, be better summaries of our p-values that highlight different behaviour that could also identify anomalous matches.

Chapter 7

A Bayesian Model for In-Play Odds

In this chapter we will describe a new Bayesian model for in-play implied win-probabilities with the goal of identifying matches which contain suspicious betting activity. Almost all current literature focusses on the pre-match market, and so developing models for in-play odds is a significant extension of the existing literature.

Having proven in Chapter 3 that the function $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible on λ, we will show how this allows us to model in-play odds by instead modelling λ, which is significantly easier to model. We will generate Bayesian prior distributions for λ using the Glicko ratings from 4, and update the posterior distributions of λ during matches based on the games won or lost in-play, so that our beliefs about the strengths of players are updated throughout matches, which allows for better fit than relying on our pre-match estimates. We shall show that the matches for which this model provides the worst fit are matches with large in-play swings, which could potentially be a sign of corrupt betting activity.

7.1 Introduction

Section 2.1 discussed the existing methods for detecting match-fixing in pre-match markets. We wish to extend this by considering match-fixing in-play, where it is believed a significant amount of match-fixing also occurs, as discussed for example by Blake and Templon (2016). This has yet to be considered in great detail in existing literature for any sport, except in general terms by Forrest and McHale (2015) and Forrest and McHale (2019), who discuss SportRadar's proprietary algorithms for detecting match-fixing in sport for both football and tennis.

This chapter will discuss a method to detect unusual odds behaviour by attempting to model the match-win probability throughout the match and comparing it with the odds. Other works such as Reade and Akie (2013) and Rodenberg and Feustel (2014) look for discrepancies between pre-match odds and match-win predictions to identify fixed matches, but we will extend this to an in-play setting.

In order to do this, we will need to extend the existing literature on in-play tennis modelling to suit our purposes. Current literature on in-play tennis modelling uses the Markov chain model of Section 2.3. This involves generating pre-match point estimates of p_1 and p_2, the probability that each of player 1 and player 2 win a point on serve, and using them in the Markov chain framework to generate point estimates of the match-win probability at various stages in the match. We find this unsatisfactory for investigating match-fixing for two reasons.

Firstly, using pre-match estimates throughout the match neglects the information available during the match. Despite our best pre-match estimates, it may become apparent during the match that one player is performing much better or worse than anticipated. This could be due to injury for example, or the fact that one player has nullified the tactics of the other player in a way that our pre-match models did not predict. Bettors will adapt their predictions in-play according to the flow of the match - it is important that our in-play predictions can do so too.

Note, however, that our ultimate goal is not to accurately predict odds, but to

flag matches with suspicious betting activity. As such, we must consider whether updating our estimates of player strengths in-play helps accomplish this.

If our predictions were updated to reflect the fact that a player is performing poorly, and yet the odds swing in their favour, this would magnify the difference between the odds and our predictions. This would make our model more sensitive to swings in the odds that disagree with player performance, helping to identify suspicious betting activity that follows this pattern.

On the other hand, if there is a swing in the odds in the same direction as is implied by the players' performance, the difference between the odds and predictions is less than if we did not update our predictions, making the method less likely to flag the match as anomalous. This will help avoid false alerts in some cases where some minor swing is to be expected based on player performance, but could arguably dampen evidence of a genuine fix in other cases. If the match is fixed, however, we would expect the swing caused by the fix to be in addition to any swing caused by an improvement in player performance. The extra odds movement should therefore still appear anomalous, and it should still be possible to detect it, even after accounting for player performance.

Balancing the two different cases, we believe updating our estimate of player strengths in play to be sensible, as it should help us identify minor swings in the direction contrary to expectations based on player performance, while we believe we should still be able to detect whether the odds are swinging more than expected in the same direction as player performance would suggest. With enough data we would be able to perform a study to examine whether this was true, but it may be challenging if the amount of suspicious matches in our data is low.

The second issue with generating point-estimates of the match-win probability is that it makes it difficult to objectively compare predictions with observed odds. How much discrepancy is acceptable? Does the level of acceptable discrepancy change from match to match? Some sources, such as DW on Sport (2016), discuss how large pre-

match swings may be more acceptable in cases where substantial uncertainty exists about the strength of particular players. The example DW on Sport (2016) provides is of a certain player whose matches featured large pre-match market swings more frequently than most, but has had many long injury breaks. DW on Sport (2016) claims that large swings after such injury breaks may be acceptable due to the difficulty involved for bookmakers, gamblers and statistical models alike in predicting how likely he was to win matches on his return, and so some extra leeway should be granted in the size of acceptable swings. We would like to formalise this idea using statistical uncertainty.

This chapter therefore develops a statistical model for predicting match-win probabilities in play that updates predictions based on how well the players are playing, and also includes statistical uncertainty around its predictions. Though we intend to use it do detect match-fixing, it should be equally capable of being used for other purposes, such as prediction or gambling.

7.2 A Bayesian Model for In-Play Odds

Modelling the temporal dynamics of match-win probabilities directly is hard. In a match between player 1 and player 2, let $m_1(\tau)$ denote a model prediction of the match-win probability for player 1 at time τ, and let $y(\tau)$ be the match-win probabilities implied by the observed odds at time τ. Some plots of $y(\tau)$ for different matches are shown in Figure 7.2.1. Clearly the length of a match is unknown beforehand, in terms of both time and the number of games, and although $y(\tau)$ must approach 1 as the match ends (or 0 if player 1 loses), it is hard to predict how quickly. Some matches are very close near the end, such as Match 6 in Figure 7.2.1, which ended in a tie-break, and so the probability just before the end is close to 1/2. Other matches are less close, and $y(\tau)$ quickly approaches 1. In Match 5 in Figure 7.2.1, the favourite established an early lead and held it, and so $y(\tau)$ quickly rose to near 1. The quality

of the players is also important - $y(\tau)$ will be much closer to 1 if one player dominates the other than if the match is close, even before any games have been played.

Instead of modelling $y(\tau)$ directly, we aim to model the quality of the two play-

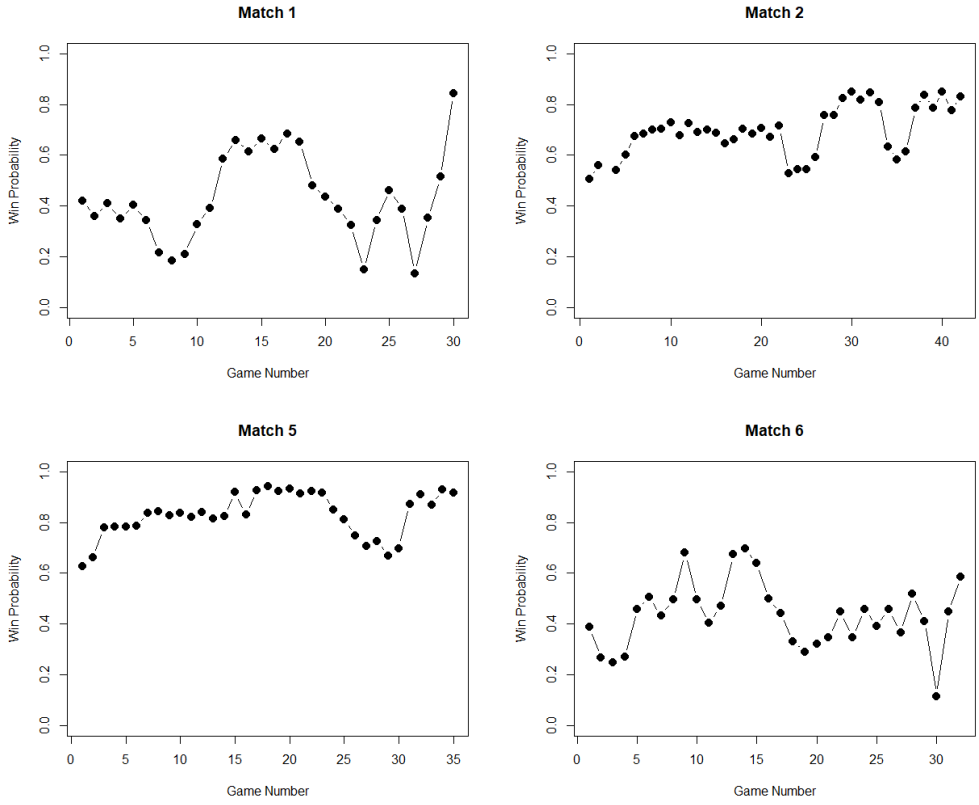

Figure 7.2.1: Plots of match-win probabilities implied by odds in four sample matches from our data.

ers and look at how this affects the match-win probability given the different scores throughout the match. According to the Markov chain model described in Section 2.3, the probability of a player winning a match at any given time in-play is controlled by just two factors - the probabilities of both players winning points while serving, and the current score. Section 2.3 also discussed and showed in Figures 2.3.6 and 2.3.7 how a further simplification can be made by reparameterising so that $p_1 = \mu + \lambda$ and $p_2 = \mu - \lambda$, and assuming that the average of the two players' point-win probabilities, μ, is some fixed value, meaning we only have to estimate λ. This means that,

given this Markov chain model, the problem of estimating the match-win probability is equivalent to modelling λ. This should be much easier to model since we will not also need to model the score or the length of the match.

As discussed and proven in Chapter 3, for fixed μ and current score $\boldsymbol{s}$ in a match played to the best of b sets, the function $m_1 = m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible since it is continuous and increasing. We shall call the inverse $\lambda = m^{\leftarrow}(m_1|\mu, \boldsymbol{s}, b)$, which we have proven the existence of in Chapter 3. As such, if we have $f_\lambda(\lambda|\mu)$, a distribution over λ, then the distribution of M_1, a random variable for the predicted match-win probability, at score $\boldsymbol{s}$, is given by a simple change of variables, yielding

$$f_{M_1}(m_1|\mu, \boldsymbol{s}, b) = f_\lambda\big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}, b)\big)\frac{dm^{\leftarrow}(m_1|\mu, \boldsymbol{s}, b)}{dm_1}. \tag{7.2.1}$$

Our goal therefore is to estimate the match-win probability at each stage in the match by estimating λ, which will give a probability distribution for m_1 according to expression (7.2.1).

7.2.1 Modelling in-play odds using a Bayesian model for λ

We will now introduce a new method for estimating in-play match-win probabilities by estimating λ. We do this by assuming that λ is constant throughout the match, but that λ can be learned about as the match progresses by observing the number of service games and tiebreaks won or lost by each player by using Bayesian modelling.

This works in exactly the same way as we would make inference about the mean parameter of a normal distribution, for example. Suppose we record successive observations from a normal distribution with unknown mean. With each new observation, we update our posterior beliefs about the mean, reflecting the new information, even though we do not believe that the underlying parameter is changing. In the same we, we describe our pre-match beliefs about λ, the strength difference of the two players, and update our beliefs as new information appears in-play, even though we do not believe that the underlying strength difference is changing - merely that we are learn-

ing more about it.

In order to perform Bayesian inference in this context, we require a few key features, each of which will be considered in more detail in the sections to follow.

- A prior distribution on λ.
- The likelihood of λ given games and tiebreaks won and lost during the match. (Recall that we do not have in-play data for points won and lost on serve, only games and tie-breaks).
- The resulting posterior distribution of λ.
- The posterior cumulative distribution function of λ given games won during the match. In particular, how this relates to quantiles and p-values.

7.2.2 Prior on λ

First of all, we shall describe different ways to specify a prior distribution on λ, which shall be denoted by $f_\lambda(\lambda|\mu)$. Due to the fact that $p_1 = \mu + \lambda$ and $p_2 = \mu - \lambda$, the domain of λ is $[-\min(\mu, 1-\mu), \min(\mu, 1-\mu)]$ to ensure that p_1 and p_2 lie in the interval [0,1]. Popular distributions with finite domains are the logistic normal and Beta distributions. Given μ, either of these could be transformed to have an appropriate domain instead of their usual domains of [0,1].

However, in general we may not have much information to make a pre-match model for λ directly, but will instead try to model the pre-match match-win probability m_1. If μ is fixed, then m_1, the pre-match win-probability for player 1, is an increasing function in λ, given by $m_1 = m(\lambda|\mu, \mathbf{0}, b)$. Hence $m(\cdot|\mu, \mathbf{0}, b)$ is invertible and a prior $f_{M_1}(m_1|b)$ on M_1 provides a prior on λ,

$$f_\lambda(\lambda|\mu) = f_{M_1}\big(m(\lambda|\mu, \mathbf{0}, b)|b\big)\frac{dm(\lambda|\mu, \mathbf{0}, b)}{d\lambda}. \tag{7.2.2}$$

7.2.3 Likelihood

The next task is to find the likelihood at time τ of λ, the dominance parameter, given μ, the average of the two-player's point-win probability, and the results of all games and tiebreaks up until time τ. To calculate the likelihood of λ in this setting, recall from Section 2.3 that under the standard assumption that points are independent and that a player always wins points with probability p, then the probability that player wins a game is

$$g(p) = p^4\Big(15 - 4p - \frac{10p^2}{p^2 + (1-p)^2}\Big).$$

At time τ, the number of games player i wins on serve is then a binomially distributed random variable $K_i^{(g)}(\tau)$, player i has played $n_i(\tau)$ service games, and has won $k_i^{(g)}(\tau)$ of them with probability $g_i = g(p_i)$ for each game. Then $K_i^{(g)}(\tau)$ has probability mass function $f_{K_i^{(g)}(\tau)}(k)$. In summary,

$$\begin{aligned} K_i^{(g)}(\tau) &\sim \text{Bin}\big(n_i(\tau), g_i\big), \\ f_{K_i^{(g)}(\tau)}(k) &= \binom{n_i(\tau)}{k} g_i^k (1-g_i)^{n_i(\tau)-k}, \qquad k = 0, \ldots, n_i(\tau). \end{aligned} \tag{7.2.3}$$

Similarly for tie-breaks, let $t_1 = t(p_1, p_2)$ be the probability player 1 wins a tiebreak given serve parameters p_1, p_2. Let $K_1^{(t)}(\tau)$ denote the number of tie-breaks won by player 1 by time τ and let $n_t(\tau)$ be the number of tie-breaks that have been played. Note that no subscript denoting player is required for $n_t(\tau)$, since both players play an equal number of tiebreaks. Similarly, $K_2^{(t)}(\tau) = n_t(\tau) - K_1^{(t)}(\tau)$, so we need only consider $K_1^{(t)}(\tau)$, which has probability mass function $f_{K_1^{(t)}(\tau)}(k)$. This leads to

$$\begin{aligned} K_1^{(t)}(\tau) &\sim \text{Bin}\big(n_t(\tau), t_1\big), \\ f_{K_1^{(t)}(\tau)}(k) &= \binom{n_t(\tau)}{k} t_1^k (1-t_1)^{n_t(\tau)-k}, \qquad k = 0, \ldots, n_i(\tau). \end{aligned} \tag{7.2.4}$$

The likelihoods (7.2.3) and (7.2.4) can then be combined to give one unified likelihood for p_1 and p_2. In order to condense notation, we let $\boldsymbol{n}(\tau) = \big(n_1(\tau), n_2(\tau), n_t(\tau)\big)$,

let $\boldsymbol{K}(\tau) = \big(K_1^{(g)}(\tau), K_2^{(g)}(\tau), K_1^{(t)}(\tau)\big)$ and let $\boldsymbol{k}(\tau) = \big(k_1^{(g)}(\tau), k_2^{(g)}(\tau), k_1^{(t)}(\tau)\big)$. We can then define

$$\begin{aligned} f_{\boldsymbol{K}(\tau)}\big((k_1,k_2,k_t) \mid \boldsymbol{n}(\tau), p_1, p_2\big) =& f_{K_1^{(g)}(\tau)}\big(k_1|p_1, n_1(\tau)\big)\ f_{K_2^{(g)}(\tau)}\big(k_2|p_2, n_2(\tau)\big) \\ & \times f_{K_1^{(t)}(\tau)}\big(k_t|p_1, p_2, n_t(\tau)\big) \\ \propto & \Big(g(p_1)\Big)^{k_1}\Big(1-g(p_1)\Big)^{n_1(\tau)-k_1}\Big(g(p_2)\Big)^{k_2}\Big(1-g(p_2)\Big)^{n_2(\tau)-k_2} \\ & \times \Big(t(p_1,\ p_2)\Big)^{k_t}\Big(1-t(p_1,\ p_2)\Big)^{n_t(\tau)-k_t}. \end{aligned}$$

By substituting in $p_1 = \mu + \lambda$ and $p_2 = \mu - \lambda$, we get the likelihood of μ and λ,

$$\begin{aligned} f_{\boldsymbol{K}(\tau)}((k_1,k_2,k_t) \mid \mu, \lambda, \boldsymbol{n}(\tau)) \ \propto\ & \Big(g(\mu+\lambda)\Big)^{k_1}\Big(1-g(\mu+\lambda)\Big)^{n_1(\tau)-k_1}\Big(g(\mu-\lambda)\Big)^{k_2} \\ & \times \Big(1-g(\mu-\lambda)\Big)^{n_2(\tau)-k_2}\Big(t(\mu+\lambda,\ \ \mu-\lambda)\Big)^{k_t} \\ & \times \Big(1-t(\mu+\lambda,\ \ \mu-\lambda)\Big)^{n_t(\tau)-k_t}. \end{aligned} \tag{7.2.5}$$

This gives the required likelihood to update beliefs about λ given the games and tie-breaks won and lost during a match.

From this, is easy to find the posterior distribution of λ in a given match at time τ by simply multiplying the prior and the likelihood, yielding

$$f_\lambda(\lambda|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)) \propto f_\lambda(\lambda|\mu) f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu, \lambda, \boldsymbol{n}(\tau)\Big). \tag{7.2.6}$$

7.2.4 The Posterior Density of M_1

The final step for analysing in-play match-win probabilities is converting the posterior distribution for λ into a posterior distribution for M_1. This involves a change of variables using the function $m_1 = m(\lambda|\mu, \boldsymbol{s}(\tau), b)$. In order to condense notation, let $\boldsymbol{\Omega}(\tau) = (\boldsymbol{n}(\tau), \boldsymbol{k}(\tau), \boldsymbol{s}(\tau))$. Applying the change of variables from λ to m_1 then yields

$$f_{M_1}(m_1|\mu, b, \boldsymbol{\Omega}(\tau)) = f_\lambda\big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\big)\frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)\Big).$$

Rather than differentiating $m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)$ with respect to m_1 directly, we use the property of the derivatives of inverses of functions,

$$\frac{df^{-1}(y)}{dy} = \frac{1}{\frac{df(x)}{dx}}\bigg|_{x=f^{-1}(y)}.$$

to obtain

$$\frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)\Big) = \frac{1}{\frac{d}{d\lambda}\Big(m(\lambda|\mu, \boldsymbol{s}(\tau), b)\Big)}\bigg|_{\lambda=m^{\leftarrow}(m_1|\mu,\boldsymbol{s}(\tau),b)}.$$

This can be put into the posterior distribution, with $\lambda = m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)$ everywhere possible to make the notation clearer. Doing so gives

$$f_{M_1}(m_1|\mu, b, \boldsymbol{\Omega}(\tau)) = \left(f_\lambda\big(\lambda|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\big)\frac{1}{\frac{d}{d\lambda}\Big(m(\lambda|\mu, \boldsymbol{s}(\tau), b)\Big)}\right)\bigg|_{\lambda=m^{\leftarrow}(m_1|\mu,\boldsymbol{s}(\tau),b)}.$$

If we want to compare the posterior distribution of M_1 with its prior distribution, we can use equations (7.2.2) and (7.2.6) to re-express $f_\lambda\big(\lambda|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\big)$, the posterior distribution for λ, and obtain

$$f_{M_1}(m_1|\mu, b, \boldsymbol{\Omega}(\tau)) \propto$$
$$\left(f_{M_1}\big(m(\lambda|\mu, \mathbf{0}, b)|b\big)f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu, \lambda, \boldsymbol{n}(\tau)\Big)\frac{\frac{d}{d\lambda}\Big(m(\lambda|\mu, \mathbf{0}, b)\Big)}{\frac{d}{d\lambda}\Big(m(\lambda|\mu, \boldsymbol{s}(\tau), b)\Big)}\right)\bigg|_{\lambda=m^{\leftarrow}(m_1|\mu,\boldsymbol{s}(\tau),b)}.$$

This is the product of the prior distribution of m_1, the likelihood of λ, and the quotient of the derivatives of the two equations to change variable from m_1 to λ at scores $\mathbf{0}$ and $\boldsymbol{s}(\tau)$.

7.2.5 The Distribution Functions of λ and m_1.

In order to consider whether the odds are behaving anomalously, we will also need to look at how extreme given odds and probabilities are with respect to $f_{M_1}(m_1|\mu, \boldsymbol{s}, b)$ at different scores $\boldsymbol{s}$ that occurs in the match. If the score at time τ is $\boldsymbol{s}(\tau)$, then we begin by considering the CDF of λ is given by

$$F_\lambda(a|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)) = \int_{\lambda_{min}}^{a} f_\lambda(\lambda|\mu)f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu, \lambda, \boldsymbol{n}(\tau)\Big)d\lambda,$$

where $\lambda_{min} = -\min(\mu, 1-\mu)$. Given the form of the likelihood of λ in equation (7.2.5), this integration cannot be performed analytically, therefore numerical methods must instead be used.

We then look at the CDF of M_1, given by

$$\begin{aligned} P(M_1 < z|\mu, b, \boldsymbol{\Omega}(\tau)) &= F_{M_1}(z|\mu, b, \boldsymbol{\Omega}(\tau)) \\ &= \int_0^z f_{M_1}(m_1|\mu, b, \boldsymbol{\Omega}(\tau))dm_1 \\ &= \int_0^z f_\lambda\big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\big)\frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)\Big)dm_1. \end{aligned}$$

The last step involves a change the variable of integration from m_1 to λ to help calculate this integral. Note that $m^{\leftarrow}(0|\mu, \boldsymbol{s}(\tau), b) = \lambda_{min}$. This is because $\lambda_{min} = -\min(\mu, 1-\mu)$, and so, either $p_1 = \mu + \lambda = 0$, or $p_2 = \mu - \lambda = 1$. Either option leaves player 1 unable to win the match, so $m_1 = 0$ too. Many terms cancel when using this change of variables, leaving

$$\begin{aligned} P(M_1 < z) &= \int_{\lambda_{min}}^{m^{\leftarrow}(z|\mu, \boldsymbol{s}(\tau), b)} f_\lambda(\lambda|\mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau))d\lambda \\ &= F_\lambda(m^{\leftarrow}(z|\mu, \boldsymbol{s}(\tau), b) \mid \mu, \boldsymbol{n}(\tau), \boldsymbol{k}(\tau)), \end{aligned}$$

and so, if Λ is a random variable of which λ is an observation, we obtain the result

$$P(M_1 < z) = P(\Lambda < m^{\leftarrow}(z|\mu, \boldsymbol{s}(\tau), b)).$$

In some ways, of course, this result is trivial given that $m(\lambda|\mu, \boldsymbol{s}(\tau), b)$ is invertible. However, it is helpful to note that this means that if we wish to get p-values for some observed values of m_1 with respect to predicted match-win probabilities, we can instead look at p-values for the posterior distribution of $\lambda = m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)$. This will be useful when we come to analysing whether the odds of matches conform to these distributions. Looking at m_1 provides information about both the strength of the two players and the current score. Looking at λ strips back information about the current score, allowing us to focus solely on the strength of the two players.

7.2.6 Priors on μ

So far, we have only dealt with the use of a fixed μ to model M_1 in-play. Now we will instead examine what would happen were we to have a prior distribution on μ instead. This would more realistically represent the fact that the average point-win probabilities of different players may be quite different, potentially giving more accurate predictions.

Suppose that there is a prior distribution over μ and m_1, given by $f_{\mu M_1}(\mu, m_1|b)$. We would instead like to transform this into a prior over λ and μ, $f_{\mu\lambda}(\mu, \lambda)$, in order to get posterior distributions for M_1 and λ at each time τ. In the same way as before, this will involve simple changes of variables, but over two variables instead of one. Using this, we will prove that

$$f_{\mu\lambda}\Big(\mu, \lambda|\boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\Big) \propto f_{\mu M_1}\Big(\mu, m(\lambda|\mu, \mathbf{0}, b)\Big|b\Big) f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu, \lambda, \boldsymbol{n}(\tau)\Big) \frac{dm(\lambda|\mu, \mathbf{0}, b)}{d\lambda} \tag{7.2.7}$$

and that

$$f_{M_1}(m_1|b, \boldsymbol{\Omega}(\tau)) \propto \int_0^1 f_{\mu\lambda}\Big(\mu, \lambda|\boldsymbol{n}(\tau), \boldsymbol{k}(\tau)\Big) \frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu, \boldsymbol{s}(\tau), b)\Big) d\mu. \tag{7.2.8}$$

In order to prove this, let the functions to transform from (μ, m_1) to (μ, λ) at score $\boldsymbol{s}$ and back be

$$H_{\boldsymbol{s}}(\mu, m_1) = \Big(\mu,\ m^{\leftarrow}(m_1|\mu, \boldsymbol{s}, b)\Big)$$
$$H_{\boldsymbol{s}}^{-1}(\mu, \lambda) = \Big(\mu,\ m(\lambda|\mu, \boldsymbol{s}, b)\Big).$$

Let $H_{\boldsymbol{s}}(\cdot, \cdot) = \big(H_{\boldsymbol{s},1}(\cdot), H_{\boldsymbol{s},2}(\cdot)\big)$, and similarly let $H_{\boldsymbol{s}}^{-1}(\cdot, \cdot) = \big(H_{\boldsymbol{s},1}^{-1}(\cdot), H_{\boldsymbol{s},2}^{-1}(\cdot)\big)$. Standard probability theory then gives a pre-match prior (i.e., when $\boldsymbol{s} = \mathbf{0}$) for μ and

λ,

$$\begin{aligned} f_{\mu\lambda}(\mu,\lambda) &= f_{\mu M_1}(\mu, m_1|b)\ \det\begin{pmatrix} \frac{dH^{-1}_{\mathbf{0},1}}{d\mu} & \frac{dH^{-1}_{\mathbf{0},1}}{d\lambda} \\ \frac{dH^{-1}_{\mathbf{0},2}}{d\mu} & \frac{dH^{-1}_{\mathbf{0},2}}{d\lambda} \end{pmatrix} \\ &= f_{\mu M_1}(\mu, m_1|b)\ \det\begin{pmatrix} 1 & 0 \\ \frac{dm(\lambda|\mu,\mathbf{0},b)}{d\mu} & \frac{dm(\lambda|\mu,\mathbf{0},b)}{d\lambda} \end{pmatrix} \\ &= f_{\mu M_1}(\mu, m_1|b)\frac{dm(\lambda|\mu,\mathbf{0},b)}{d\lambda}. \end{aligned}$$

A posterior distribution for μ and λ given data can then be obtained using this prior in the usual way,

$$\begin{aligned} f_{\mu\lambda}(\mu,\lambda|\boldsymbol{n}(\tau),\boldsymbol{k}(\tau)) &\propto f_{\mu\lambda}(\mu,\lambda) f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu,\lambda,\boldsymbol{n}(\tau)\Big) \\ &\propto f_{\mu M_1}(\mu, m_1|b) f_{\boldsymbol{K}(\tau)}\Big(\boldsymbol{k}(\tau)|\mu,\lambda,\boldsymbol{n}(\tau)\Big)\frac{dm(\lambda|\mu,\mathbf{0},b)}{d\lambda}, \end{aligned}$$

as claimed in equation (7.2.7).

In order to obtain a probability distribution over M_1, we first need to change variables back to μ and M_1, which is given by

$$\begin{aligned} f_{\mu M_1}(\mu, m_1|b,\boldsymbol{\Omega}(\tau)) &= f_{\mu\lambda}(\mu,\lambda|\boldsymbol{n}(\tau),\boldsymbol{k}(\tau))\ \det\begin{pmatrix} \frac{dH_{\boldsymbol{s}(\tau),1}}{d\mu} & \frac{dH_{\boldsymbol{s}(\tau),1}}{dm_1} \\ \frac{dH_{\boldsymbol{s}(\tau),2}}{d\mu} & \frac{dH_{\boldsymbol{s}(\tau),2}}{dm_1} \end{pmatrix} \\ &= f_{\mu\lambda}(\mu,\lambda|\boldsymbol{n}(\tau),\boldsymbol{k}(\tau))\frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu,\boldsymbol{s}(\tau),b)\Big). \end{aligned}$$

We then need the marginal distribution of M_1 with respect to this distribution, given by

$$\begin{aligned} f_{M_1}(m_1|b,\boldsymbol{\Omega}(\tau)) &= \int_0^1 f_{\mu M_1}(\mu, m_1|b,\boldsymbol{\Omega}(\tau))d\mu \\ &= \int_0^1 f_{\mu\lambda}(\mu,\lambda|\boldsymbol{n}(\tau),\boldsymbol{k}(\tau))\frac{d}{dm_1}\Big(m^{\leftarrow}(m_1|\mu,\boldsymbol{s}(\tau),b)\Big)d\mu. \end{aligned}$$

This corresponds with equation (7.2.8). The integration must be solved using numerical methods.

7.3 Results

7.3.1 Some plots of in-play predictions of m_1 and λ using Glicko priors

In order to illustrate the methods in Section 7.2, we shall now demonstrate its use on a couple of example matches and look at how it identifies suspicious-looking matches.

As well as looking at estimates of match-win probabilities compared to odds, we will also look at how estimated λ compares with $m^{\leftarrow}(y_\tau|\mu, \boldsymbol{s}(\tau), b)$. Looking at the odds on this alternative scale helps focus only on how the strength of the two players is modelled, while stripping out information about the current score that is included in the odds y_τ. Looking at both is key to examining our analyses and how they may be improved upon.

Figure 7.3.1 shows the values of λ implied by the odds for the matches in Figure 7.2.1. These vary much less than the odds, and looking at these plots will show how the behaviour of this implied λ is captured by our models for λ.

7.3.2 Example Match Fits

To begin with we will look at some results of using this Bayesian method with prior distributions generated from the Glicko ratings implementation as described in Chapter 4.

The easiest prior distributions to obtain for M_1 are obtained by simply using the formulae for match-win probability based on Glicko ratings as described in equation (4.1.1), along with parameters ν_i and σ_i obtained for each player i from implementing

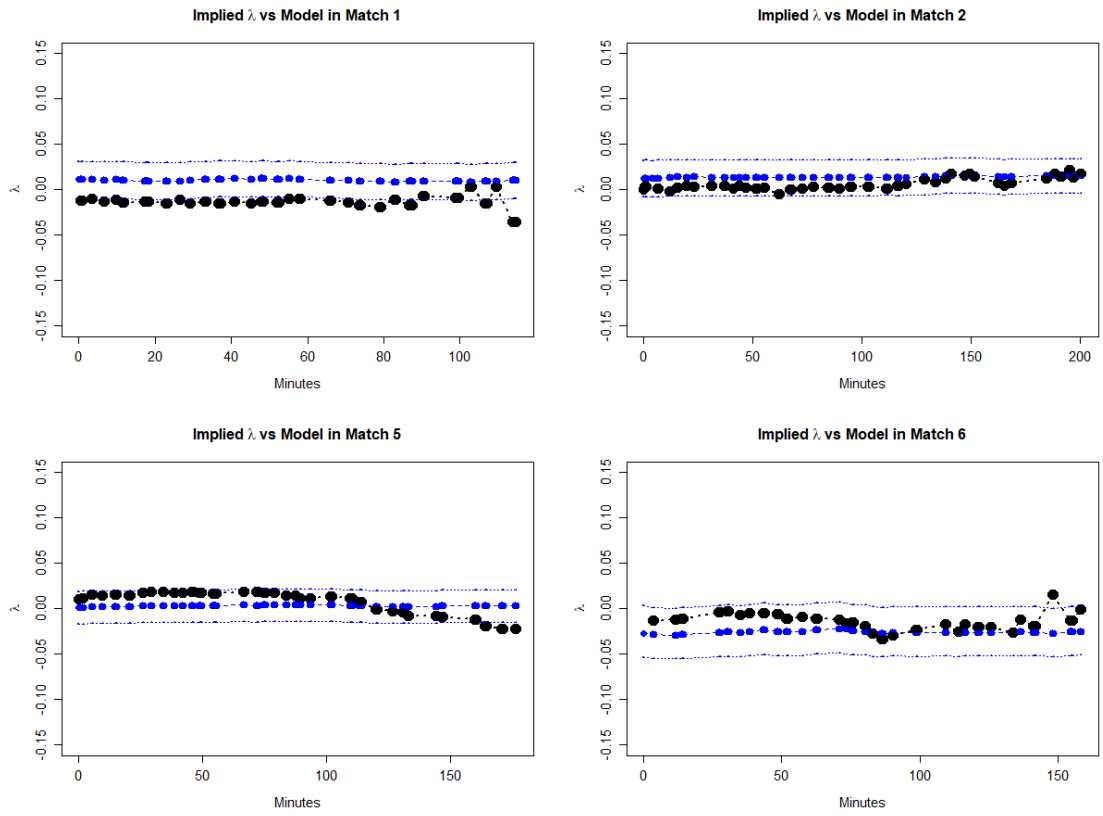

Figure 7.3.1: Plots of values of λ implied by odds in four sample matches from our data. Black dots represent our transformed odds data, while blue dots are the posterior median and successive 95% predictive intervals for λ. Time is measured in minutes from the start of the match.

Glicko ratings. The relevant formulae are

$$\begin{aligned} \theta_i &\sim N(\nu_i, \sigma_i^2) \\ M_1 &= \frac{e^{q(\theta_i - \theta_j)}}{1 + e^{q(\theta_i - \theta_j)}} \\ M_1 &\sim LN\Big(q(\nu_i - \nu_j), q^2(\sigma_i^2 + \sigma_j^2)\Big), \end{aligned}$$

where $M_1 \sim LN(\mu, \sigma^2)$ denotes that M_1 has logistic normal distribution with parameters μ and σ^2. This prior distribution for M_1 can then be used to provide a prior distribution for λ in the manner described in equation (7.2.2).

However, it is important to bear in mind that our real goal is to estimate the implied win probabilities, which we call $y(\tau)$, rather than simply the actual win prob-

abilities, $m_1(\tau)$. Although we expect $m_1(\tau) \approx y(\tau)$, the aim is to estimate $y(\tau)$ so that we can identify matches where the implied win probabilities do not behave as expected. In their work on pre-match odds, and for model prediction $\hat{m}_1$, Reade (2014) fit a linear regression model for the expected in-play opening odds, $E(Y(0)) = \alpha + \beta\hat{m}_1$ to account for any systematic biases in the odds compared with their predictive model. We took a similar approach, but want to ensure that $E(Y(0)) \in [0, 1]$. We therefore decided to examine whether $\alpha + (\beta + 1)q(\nu_i - \nu_j)$ is a better predictor for $Y(0)$ than $q(\nu_i - \nu_j)$. We also want to know if multiplying the variance, $q^2(\sigma_i^2 + \sigma_j^2)$, by a constant, c^2, better described the variability in $Y(0)$. We therefore wish to compare the two models,

$$Y(0) \sim LN(\alpha + (\beta + 1)q(\nu_i - \nu_j),\ c^2q^2(\sigma_i^2 + \sigma_j^2)\Big),$$
$$Y(0) \sim LN(q(\nu_i - \nu_j),\ q^2(\sigma_i^2 + \sigma_j^2)\Big),$$

looking at the significance of each of the parameters α, β and c.

A plot of $q(\nu_i - \nu_j)$ against $Y(0)$ is shown in Figure 7.3.2. We fitted the model without two matches known to exhibit peculiar odds behaviour and found no evidence to suggest that α or β made significant contributions to the model fit, but found that c was significantly different to 1 at the 5% level, with c being an estimated 0.914. We therefore included c in our model for $Y(0)$ to provide a better fit, meaning the model for $Y(0)$ used was

$$Y(0) \sim LN\Big(q(\nu_i - \nu_j),\ c^2q^2(\sigma_i^2 + \sigma_j^2)\Big). \tag{7.3.1}$$

This model was used to generate prior distributions for λ in each of the 274 matches for which we have in-play odds data, before we found posterior distributions in each game break using the methods described in Section 7.2.4. The next section will look in a little detail at some example matches, before overall behaviour from all of the matches is summarised.

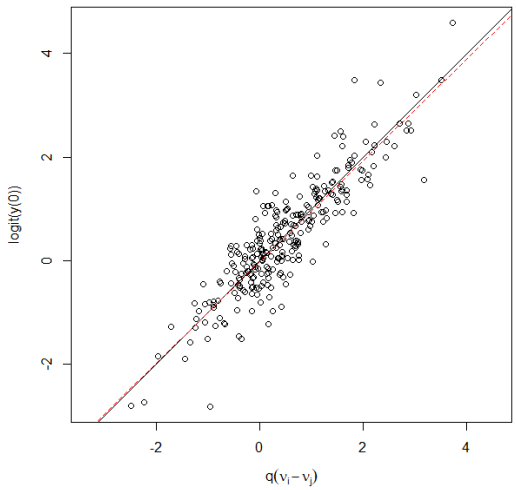

Figure 7.3.2: A plot showing the difference in means of Glicko ratings against logit opening in-play odds for each match. The fitted regression line (dashed red) is very similar to the line y=x (black).

7.4 Example Results

First, we consider a few matches where the odds have been very well predicted. Figure 7.4.1 shows two matches for which the odds have been very well estimated. The initial estimate for λ is strong, and the posterior estimate of λ changes little throughout the match. However, the implied λ also varies little, and so both λ and the match-win probabilities are well estimated throughout the match. However, it is interesting how wide the confidence intervals are, despite the fact that $c < 1$.

Figure 7.4.2 shows a few matches which have not been predicted as well. In both cases, the initial estimate of λ is not particularly accurate. In match 2, the implied λ drifts closer to the median estimate of λ, whereas in match 15 the implied λ drifts further away. It's unfortunate that the initial poor estimates of λ affect the estimates for the rest of the match.

There are several reasons why the initial estimates of λ may be poor. The first reason is the simple fact that our prediction methods are imperfect. The odds are known to be excellent predictors of In specific matches, one reason why the odds may

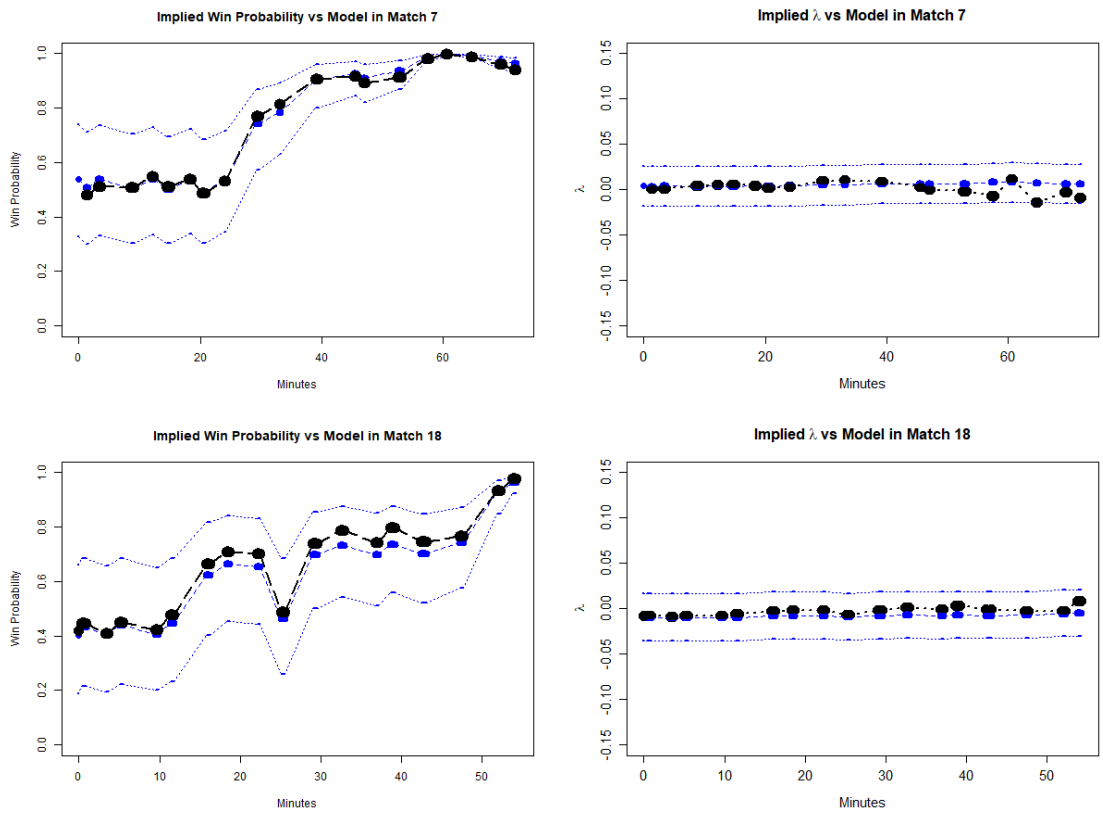

Figure 7.4.1: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in matches 7 and 18. Time is measured from the start of the match.

differ so much from our predictions is news of an injury to a player. This will shift market perceptions of the players' win probabilities, whereas our model's predictions will remain the same. Another issue may be the surfaces the matches are played on. It is well known that different players have different preferred surfaces, and if the markets account for this but our predictions do not, this will induce differences between the market and our predictions. Section 4.4.3 discussed possible remedies for this by incorporating surface information into Glicko ratings, but this remains an open area of research.

In examining why the posterior distributions of λ update so slowly, we look at the data for these two matches. In match 2, the final score is 6-4, 6-7, 7-6, with the winner

broke the loser's serve four times and was broken once in return. The evidence from the scores therefore suggest that this was a close game, with little information to be gained about the relative quality of the two players simply by looking at the number of games won or lost on serve. As such, the likelihood does not contain enough information to make the posterior distributions significantly different from the prior distributions. In match 15, the final score is 2-6, 7-5, 7-6. The winner therefore won fewer games than their opponent, breaking their opponent's serve 4 times and being broken 5 times in return. This suggests another close match, in which the loser has arguably played better than their opponent over the whole match. However, the rules of tennis are such that the loser's failure to win the last two close sets has cost them victory. The likelihood function again contains insufficient information to shift the posterior distributions. It is possible that the use of point-by-point data would help λ update faster, but equally even this data may not be enough to move estimates of λ closer to the odds.

One interesting example of a match that the model predicts poorly is shown in Figure 7.4.3. World rankings and betting odds both suggested that one player was a clear favourite. While the implied win probabilities appear to have been predicted fairly well staying near 1 throughout the match, it appears at first that our model has failed to capture a large decrease in implied λ. This decrease is surprising, given that by the end of the match the underdog had won just 3 out of his 8 service games, whereas the favourite won 7 out of 9.

Closer examination of the match data paints an interesting story. The favourite quickly established a commanding lead that he never relinquished. Looking at the odds themselves, rather than implied win probabilities, as shown in Figure 7.4.4 shows that the favourite's odds are so low that they change very little, even as his lead becomes more and more commanding. Meanwhile, although the underdog's odds lengthen, they do not change the implied win probability significantly either. This is what causes the apparent decrease in implied λ - if the win probability remains the

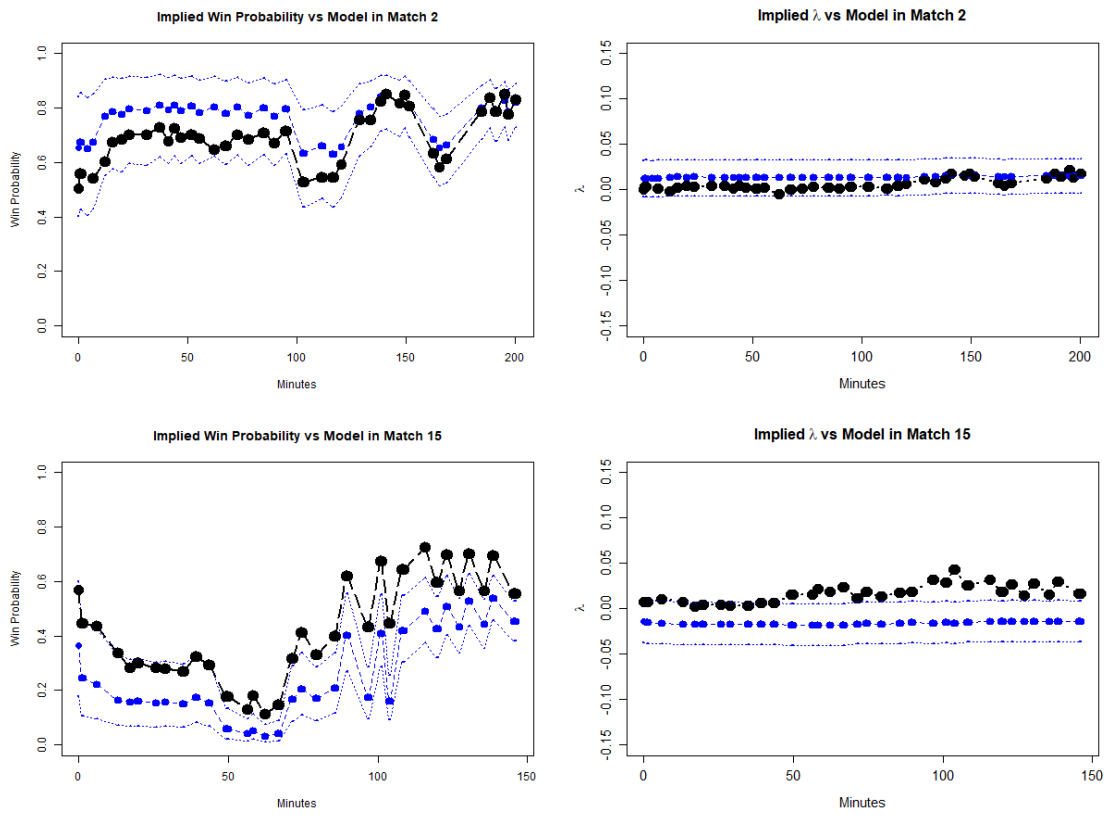

Figure 7.4.2: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in matches 2 and 15. Time is measured from the start of the match.

same while the score becomes more favourable, then λ must decrease. However, this is likely due to an issue with the liquidity of the market - once the result of the match essentially becomes a foregone conclusion, gamblers will risk very little further money on the outcome, and the odds will therefore not move. This is backed up by data about the volumes of money gambled on the match. By the end of the 12^{th} game interval 50 minutes into the match, £1,803,807 had been gambled on the match. By the end of the match 20 minutes later, this had increased by just £3,362. The odds are essentially censored such that if the market probability of an event dips below a certain value, the odds may not change as gamblers are unlikely to place a stake on such an unlikely event.

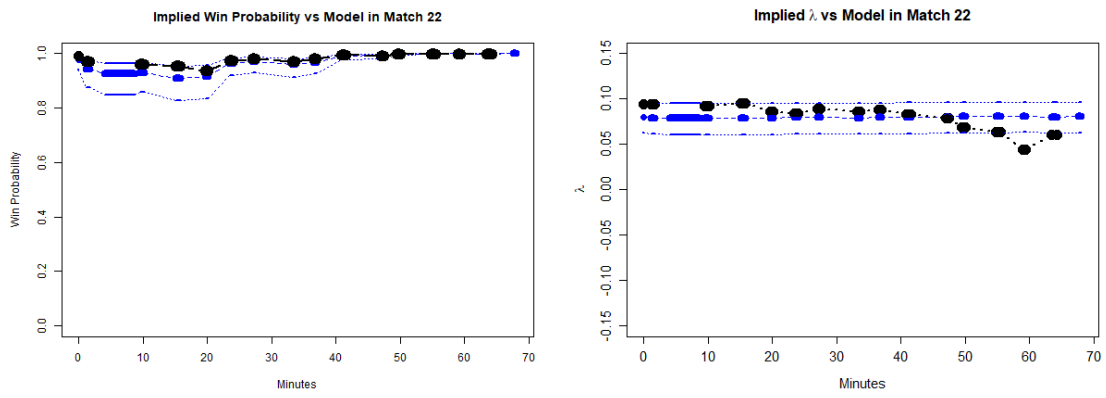

Figure 7.4.3: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in match 22. Time is measured from the start of the match.

This highlights the need to carefully examine matches which have been flagged

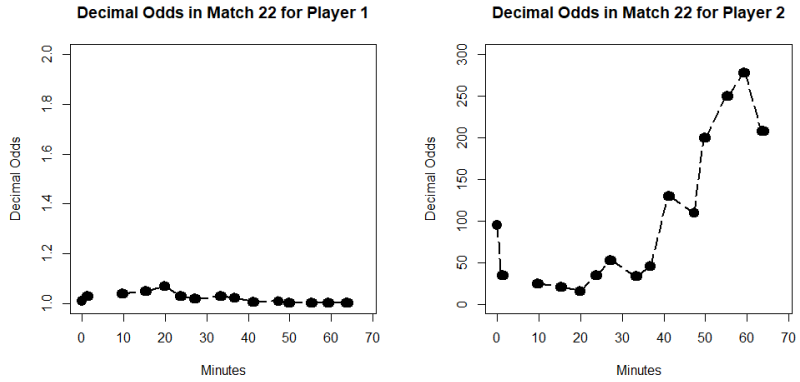

Figure 7.4.4: Plots of values of decimal odds for both players in match 22. (Note the difference in scales on the y-axes.) Time is measured from the start of the match.

up by our model to look for alternative explanations of the data. It is possible that better modelling would prevent matches such as this being flagged - however, it is sufficient for now to examine the match data in greater detail and see that an alternative explanation to match-fixing is more plausible, and thus disregard it from further investigation.

Next we come to two matches that have been cited in other sources due to the odds

being highly unusual. The plots for these matches are displayed in Figure 7.4.5. In match 136, the initial estimate for λ is good, whereas the initial estimate in match 73 is poor. However, in both cases the implied λ increases greatly throughout the match at a rate that is not justifiable by what is occurring in-play. Indeed, in match 136, the implied win-probability is increasing throughout the match, even though player 1 loses the first set. This win-probability behaviour is completely at odds with what the score would imply. It is precisely these sorts of matches that we wish to flag up as anomalous, and our method has successfully identified both matches.

Looking generally at our model estimates, two main features stand out. Firstly,

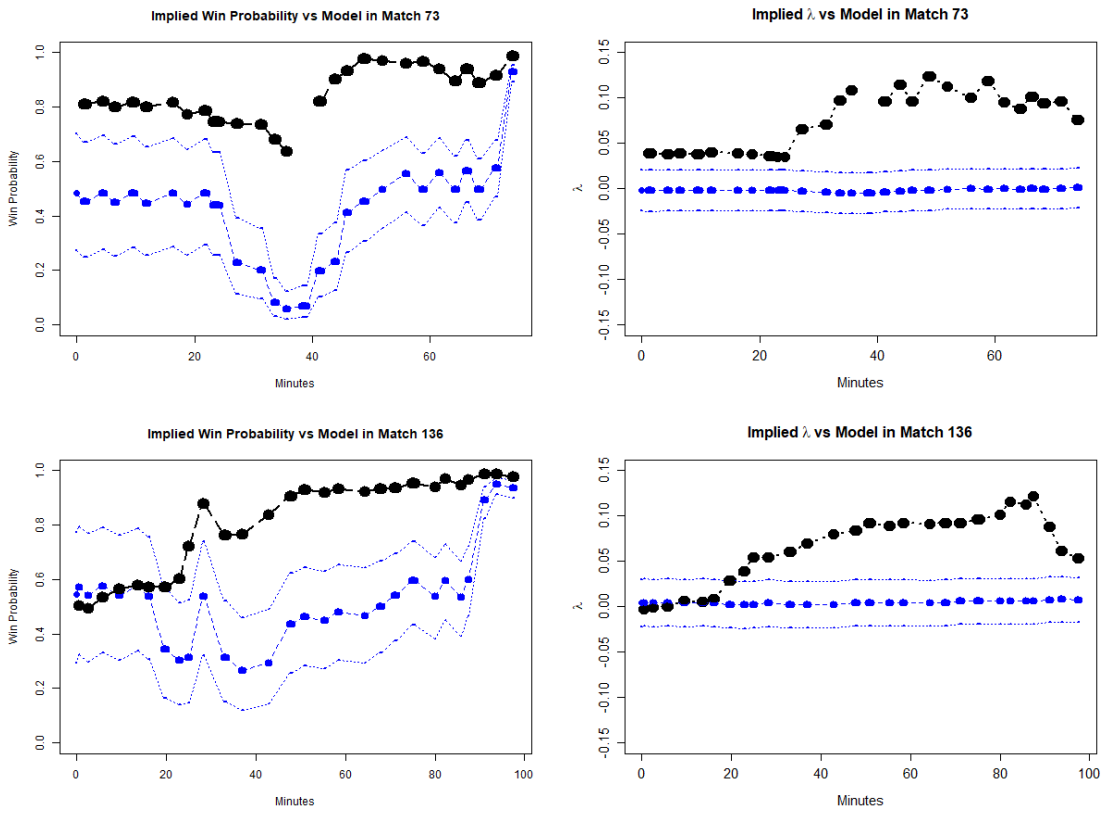

Figure 7.4.5: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in matches 73 and 136.

the predictive intervals are very wide, and secondly, the posterior distributions for λ change very little throughout the match. The first feature is due to in part due

to the large parameter uncertainty in the Glicko ratings, but also due to the quality of the predictions provided by the Glicko ratings. Figure 7.3.2 shows that although the Glicko ratings predict the opening implied win probabilities quite well, there is still a substantial amount of variation in the opening implied win probabilities that is unexplained. More accurate predictions would the term make c^2 in the model in equation 7.3.1 much smaller and provide much tighter predictive intervals.

The lack of movement in the posterior distributions of λ is due to the significant

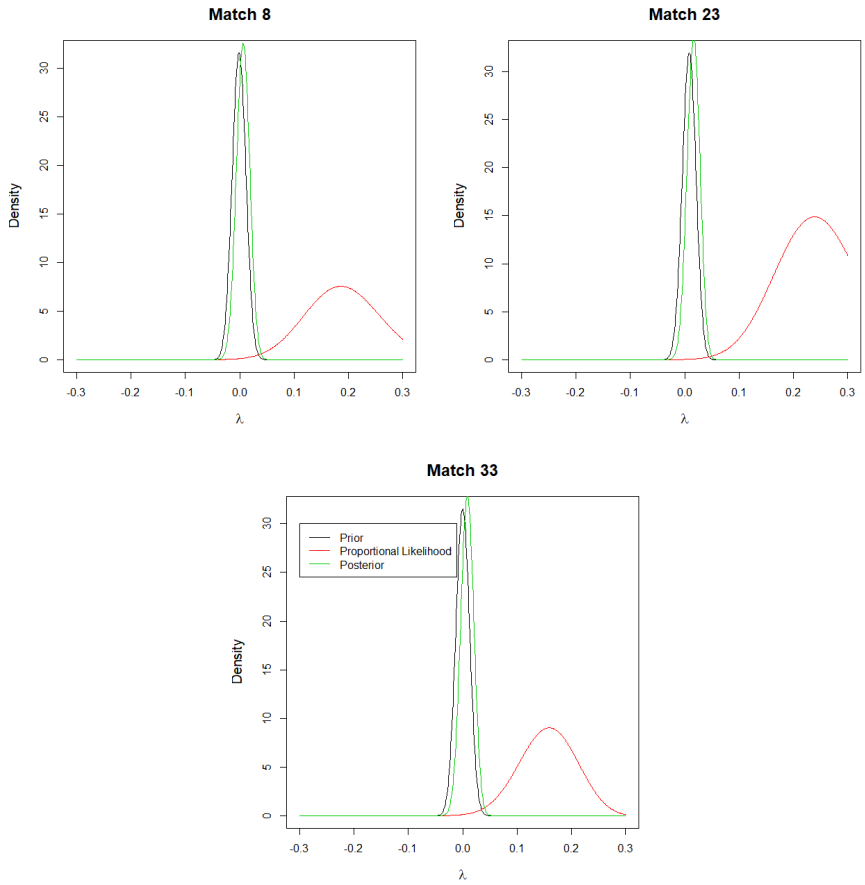

Figure 7.4.6: Prior distributions and end-of-match posterior distributions of λ for three matches, and a curve proportional to the likelihood. (Proportionality is used as the true likelihood is several orders of magnitude smaller than the prior and posterior densities of λ.)

uncertainty in the likelihood. Figure 7.4.6 shows the prior and end-of-match posterior distributions and the likelihood of λ for the three matches with the biggest change in

median prediction of λ between the first and last games. In these matches, it can be seen that although the maximum likelihood estimator of λ is very high, the fact that the uncertainty in the maximum likelihood estimator is so large limits how much the posterior distribution shifts, and hence how much can be learned about λ throughout the match. In these cases where one player dominates so heavily, subtle alterations are made to the posterior distributions, but in other matches very little can be learned about λ.

It would be interesting to explore whether the use of point-by-point data would help provide more useful likelihood information than the game-by-game data we currently use. Breaks of serve are rare in tennis, even when one player is playing much better than the other. Point-by-point data might help highlight other patterns, such as when one player is holding serve easily while the other is struggling to hold serve but still succeeding. The use of game-by-game data would mark both players as equal in this case. However, point-by-point data would highlight one player's superiority over the other, even if this superiority has yet to be converted into breaks of service.

In the absence of point-by-point data, it may be possible to use the time durations of the games as a proxy for the number of points each player wins. Holds of serve to love will be played over just four points, whereas a game with a deuce must feature at least eight points. It would be natural to expect that the games with deuces should take noticably more time than easy holds of serve. Naturally other factors will effect the time lengths of games, such as injuries and fatigue, so the relationship between the game duration and the score will be imperfect, but there may be a sufficiently strong relationship to get an idea of whether the closeness of games provides useful information.

It is of course possible that even point-by-point data would be insufficient to markedly alter the posterior distributions of λ. In many ways, it would be sensible if a player's career history proved much more informative about their strength than their performance at the start of the match - one set match should perhaps not greatly shift

our expectations of an excellent player. However, recall once more that our true goal is to predict the implied win-probabilities, rather than make predictions about the match results. Our previous plots show that implied λ can vary significantly through the match far more than our model currently allows. Our work in Chapter 8 will examine whether this variation in implied λ can be predicted from events occurring in-play, or whether it is simply due to randomness.

7.5 Analysis of in-play p values

Having considered a few matches individually, we must now consider how to identify the most suspicious matches automatically. Thousands of matches happen every year, and so it is not practical for each match to be examined individually. Instead, we seek criteria with which to flag matches as suspicious, so that they can be subjected to further investigation.

For each match, our model gives a p-value that explains how unusual the implied win probabilities, or equivalently, implied λ, are with respect to their respective posterior probability distributions. These must be summarised for each match in a way that flags the most suspicious matches while ignoring matches that behave as expected. We begin with two very simple summaries of all of the p-values in each match - the smallest p-value, and the average. The first will tell us the biggest discrepancy between our model and the implied win-probabilities, while the second will indicate a more consistent pattern of error.

Figure 7.5.1 shows the smallest p-value for implied λ in each match, as a measure of the biggest discrepancies between the odds and our model. The p-values were obtained using R's built-in numerical integration. The numerical accuracy of this integration is a specifiable parameter, and was chosen so that the error in each p-value was at most 0.01 or the size of the p-value, whichever is smaller. This ensures that the very smallest p-values are certainly not less than 0, and at most double their reported

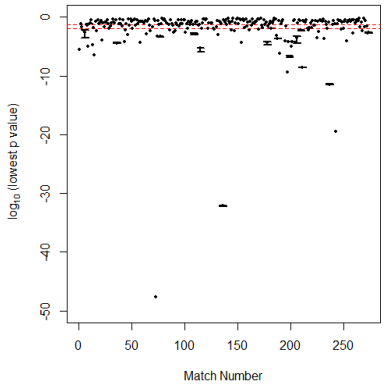

Figure 7.5.1: For each match, we take $\log_{10}$ of the lowest p-value of implied λ with respect to the posterior distribution of λ. Errors bars on the p-values for the accuracy of numerical integration are shown where appropriate. Red horizontal dashed lines represent $\log_{10}(0.05)$ and $\log_{10}(0.01)$.

value, although in most cases the error is much smaller.

The two matches with the most extreme p-values are matches 73 and 136, which have already been examined in Figure 7.4.5. The next three are matches 197, 237 and 243, which are shown in Figure 7.5.2. Matches 197 and 237 both follow a similar pattern, in that the opening estimate of λ is poor, and slowly worsens as the match progresses. The behaviour in Match 243 is rather more peculiar however, in that a huge swing in implied λ occurs in the first half of the match before staying relatively stable until some volatility creeps in in the last few games of the match.

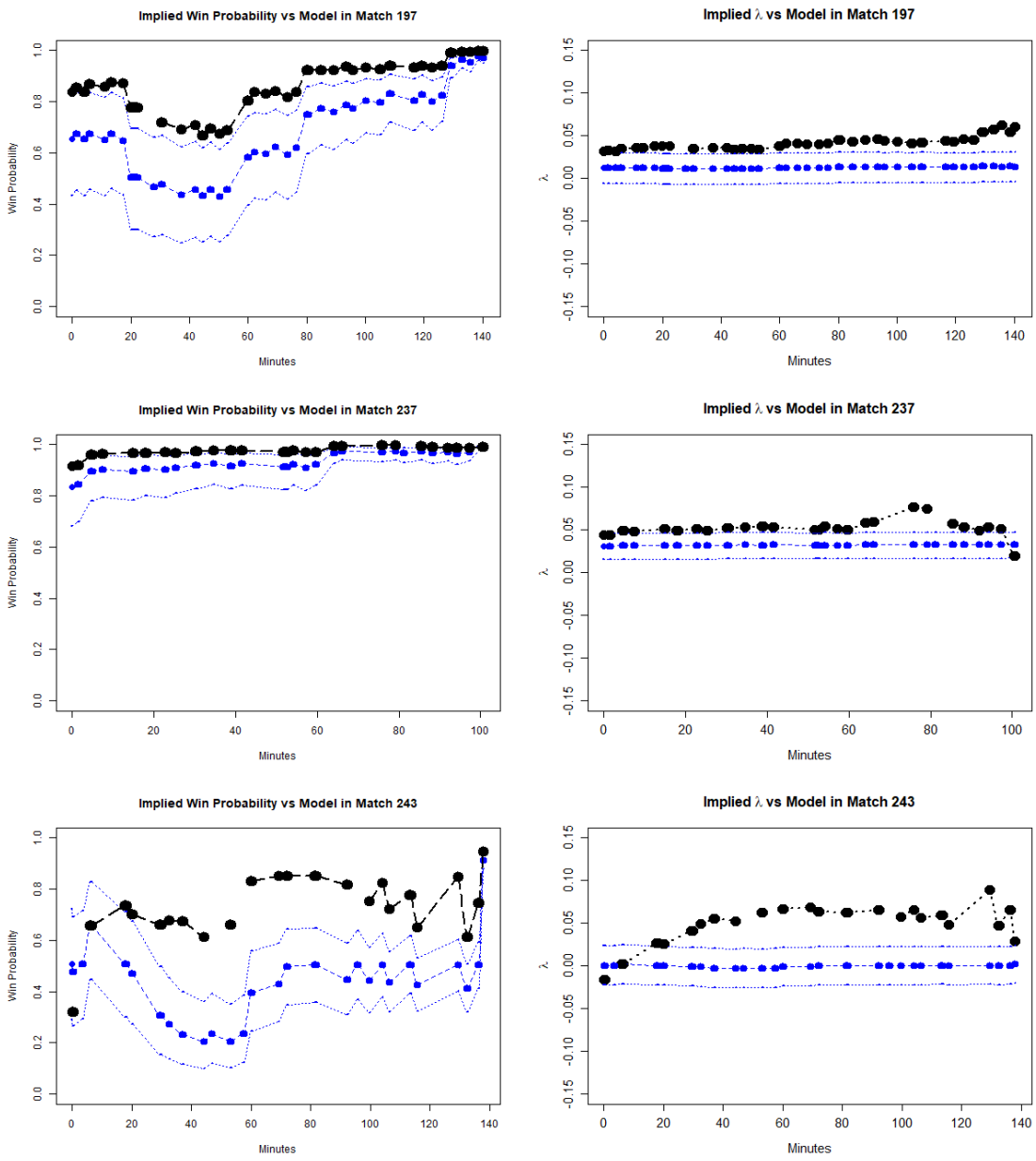

Figure 7.5.2: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in matches 197, 237 and 243.

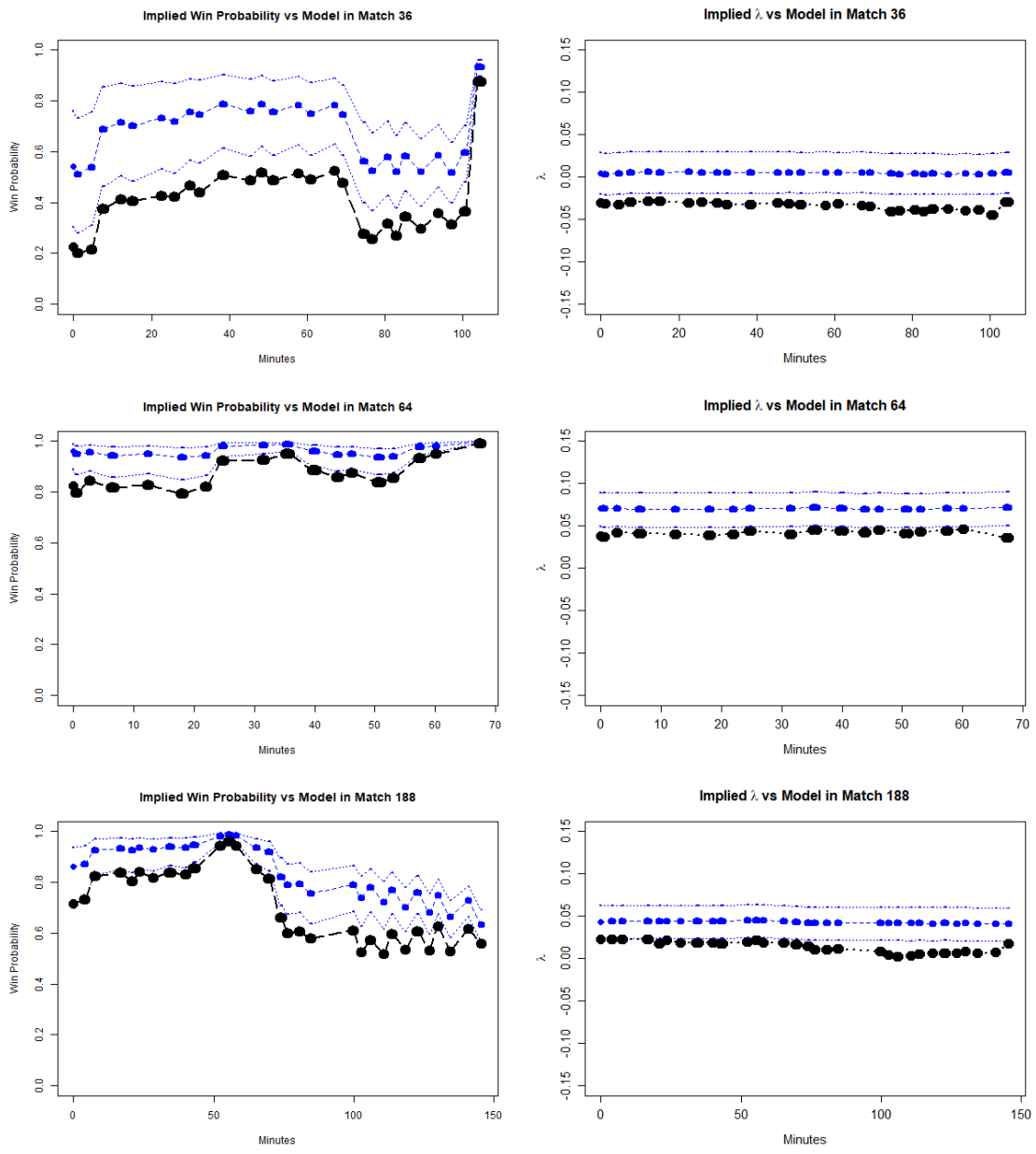

Figure 7.5.3: Plots of values of implied win probabilities and λ (black) vs the median of model predictions and a 95% predictive interval (blue) in matches 36, 64 and 188.

Next, we consider the average of all p-values in each match, the results of which are shown in Figure 7.5.4. Match 73 is still the most extreme, but interestingly, Match 136 has only the 30^{th} lowest average p-value. The next four lowest averages are from Matches 36, 197, 64 and 188. Match 197 has already been considered in Figure 7.5.2,

so we examine the other three in Figure 7.5.3.

In each of these matches, there is a consistent pattern that the initial estimate for λ is poor and remains poor throughout the match, as neither the model estimate of λ, nor implied λ itself, vary much during the match. Were the initial estimate of λ closer to that implied by the odds, these matches would probably not have been flagged. This could be a sign that suspicious activity has occurred in the pre-match market. However, if it is simply due to our model not being as accurate as hoped, it is more of a concern.

We now examine the extent to which the initial estimate of λ affects the average of

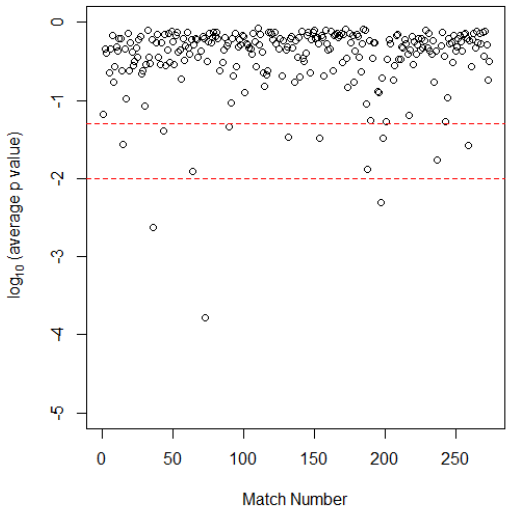

Figure 7.5.4: For each match, this plot displays $\log_{10}$ of the average of all p-values in each match. Red dashed lines represent $\log_{10}(0.05)$ and $\log_{10}(0.01)$.

the other p-values, shown in Figure 7.5.5. The two are highly correlated, particularly when both are near 0. Indeed, in the 13 matches with an average p-value below 0.05, all had an opening p-value below 0.09. These matches are flagged because the opening estimate is poor, but precious little can be said about the behaviour about the in-play odds, which is not entirely ideal. Looking for errors in the initial estimate is important, but errors may occur due to factors beyond our control - there could be

a minor injury in the lead-up to the game our model is unaware of, the surface may favour one player, or it may simply be a consequence of the fact that our prediction methods are imperfect. Therefore, while knowing about the error in the opening estimate is important, it would be helpful if we could also examine the in-play market in matches where there is an error in the opening estimate - something our model appears to be currently unable to do.

In other words, comparing the implied win-probabilities with model predictions

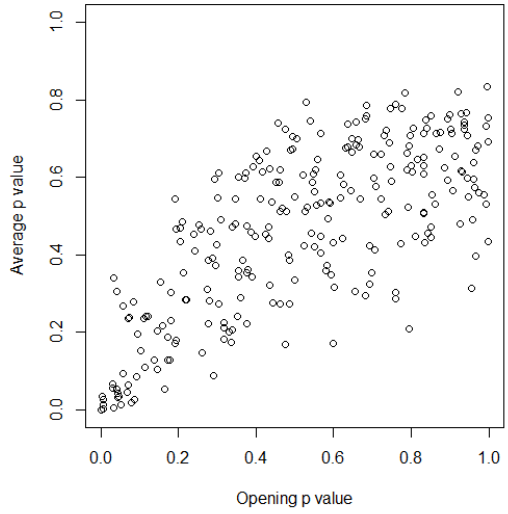

Figure 7.5.5: The first available p-value for implied λ with respect to model predictions compared with the average of all p-values in the match.

based on the Glicko ratings tells us whether or not the odds are as expected at each time in the match, conditioning only on the results of the players' previous matches and games won or lost in-play. This seems sensible, but it is important to note that this combines two distinct pieces of information - whether or not the pre-match market closes at a price we'd expect, and whether or not the odds develop as expected once the match starts. While there is no explicit problem with combining the two pieces of information in this way, it may be more helpful to separate the pieces of information in order to get a clearer picture of what is actually happening in the match. This is

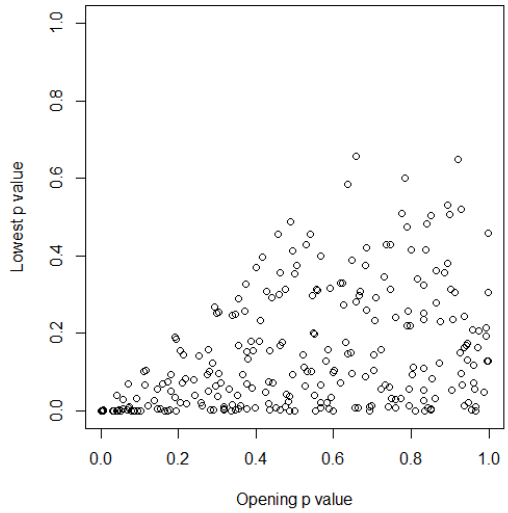

Figure 7.5.6: The first available p-value for implied λ with respect to model predictions compared with the lowest of all p-values in the match.

particularly true if the pre-match odds are not as expected. There may be a valid reason why there may be a discrepancy between the odds and our model, such as a recent injury or a public belief that one player's style will dominate the other in a way not capture by our model. If our pre-match model fails to capture this, the match will be flagged up, but we will be essentially unable to assess whether anything suspicious has also occurred in the in-play market. The match may then be dismissed due to injury being the probable explanation, and any genuine corrupt activity in-play may be missed. We therefore wish to separate the pre-match and in-play markets, assessing each separately for corrupt activity.

7.5.1 Using Pre-Match Odds to Generate Prior Distributions

Given that problems estimating the first λ may be hampering our analyses, some individuals have suggested to us that this problem may be resolved by simply shifting the mean of our prior distribution on the odds closer to the opening odds. This would

mean replacing the prior distribution in 7.3.1 with a prior distribution for $M_1(0)$, the probability player 1 wins at time 0, given by

$$M_1 \sim N(y(0), \gamma_k^2). \tag{7.5.1}$$

Some thought would have to given on how to estimate γ_k^2. We could then instead use this prior distribution to create a prior distribution for λ, update posterior distributions for $\lambda(\tau)$ at later times τ to reflect the incoming data, $\mathbf{\Omega}(\tau)$, and compare observed implied λ with these posterior distributions.

The previous work explored posterior distributions of λ conditonal on the players' career histories and the in-play data on player performance, $\mathbf{\Omega}(\tau)$. Using the prior distribution in equation 7.5.1 would instead be exploring the posterior distributions of λ conditonal on the pre-match market and the in-play data on player performance, ignoring career histories, and $\mathbf{\Omega}(\tau)$, which may also prove sensible.

We attempted this using $\gamma_k^2 = c^2(\sigma_i^2 + \sigma_j)^2$ to explore the outcomes. In doing so, we found that the resulting model was very poorly calibrated indeed, as highlighted in Figure 7.5.7. This figure shows boxplots of the CDFs, $P(M_1 < y(\tau)|\mu, b, \mathbf{\Omega}(\tau))$, for the observed data $y(\tau)$ from all of our matches with respect to our Bayesian posterior distributions, with the boxplots organised by the game index τ. (The first boxplot therefore corresponds to a boxplot of the values of $P(M_1 < y(1)|\mu, b, \mathbf{\Omega}(1))$ from all of our matches, and so on).

From Figure 7.5.7, we see that odds moved much less than expected in early matches than the odds-based prior predicted, resulting in a very narrow distribution of CDFs in early games, while the CDFs were over-dispersed in later games. The CDFs should be uniformly distributed. By contrast, the corresponding boxplots for the history-based priors in Figure 7.5.7 are distributed much more uniformly, although some problems persist in very late games, though the fact that not many matches have 30 games, resulting in small sample sizes in these cases, is a mitigating factor.

The implication of this is that the odds do not behave as predicted using the odds-based priors. The results are more consistent with a random walk effect of sorts,

in which small movements in the implied λ build up over time, yielding successively wider distributions. This means that from a fraud detection perspective, we would struggle to find large CDFs early in the match using odds-based priors, and hence may miss suspicious betting activity. Although a similar pattern is slightly visible in the lower plot of Figure 7.5.7 which uses career-based histories, the extent of the effect is much lesser, suggesting that in some way the data fits the model better in this case due to the more uniform distribution of CDFs (even though we expect that using odds-based priors would lead to more accurate point-estimate predictions, as odds are more informative than the Glicko ratings). This meaning that when using history-based priors, we are also able to observe large p-values early in the matches and thus are much more able to detect suspicious gambling.

It could be worth exploring further whether the issues in using odds-based priors could be circumvented. However, a wider view suggests that even though this Bayesian model saw some success in identifying matches with large in-play odds swings, a new model that more realistically captures the dynamics of the in-play odds may see even better success. Chapter 8 will introduce precisely such a model.

7.6 Conclusion

In this chapter, we described one of the first models to attempt to model in-play odds. We showed that we could model λ instead of modelling the implied win-probabilities directly. We then developed a prior distribution for λ based on Glicko ratings and demonstrated how to update the posterior distribution of λ during the match as games were won or lost, and used these posterior distributions to find posterior distributions for match-win probability. We compared implied match-win probabilities for observed odds data to these posterior distributions and analysed the $p-$values. We successfully identified two matches that other sources have also suggested have very suspicious betting activity. This model therefore represents an advancement on the methods in

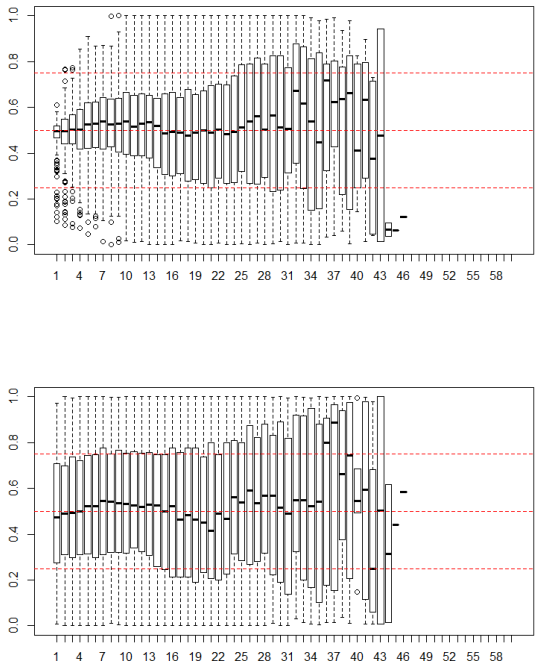

Figure 7.5.7: Boxplots of the CDFs of the odds with respect to our Bayesian posterior distributions according to each game. In the left plot the prior mean is shifted to match the data, while the right plot uses prior means based on Glicko ratings. Red dashed lines are at 0.25, 0.5 and 0.75

current literature.

However, some questions remain over the how well the model predicts λ in some matches. The Glicko ratings do not always predict the opening odds well, and we struggle to make inference about the suspiciousness of the betting activity when this occurs. We also found that our Bayesian updating did not track the movements in the observed implied λ process very closely, which seemed to be affected by games won and lost much more than our posterior distributions suggested. It is possible that point-by-point data would help provide better in-play updates, but by no means certain. For this Bayesian model to be successful, we would need to ensure our Glicko ratings provided much better predictions for the odds. Incorporating surface information would be a good start to this end.

It may be, however, that a better approach would be to try an alternative model. In the next chapter, we will attempt to model the observed implied λ process directly, so that we can base our predictions of how much the implied λ is affected by games won or lost based on the data, and so that we can circumvent the problems that arise from our implementation of Glicko ratings.

Chapter 8

A Gaussian Processes Model for In-Play Odds

In this chapter we develop a Gaussian process model for in-play odds based on modelling the values of implied λ. In Chapter 7 we built a Bayesian model for implied λ that built a prior distribution based on Glicko ratings and updated the posterior distribution based on the results of games won and lost. We found that the observed implied λ varied much more with games won and lost than this Bayesian model suggested, and also found that poor estimates of the opening implied λ hampered our analyses. In this chapter, we therefore develop a model in which the increase in implied λ based on observed games won and lost is estimated directly from the data. We estimate the size of this effect by pooling data from across our different matches, and we also pool information from across the different matches to estimate the variability and correlation structure of this process.

Note that modelling the in-play implied λ, and by extension the odds, in this way differs from how we have modelled the odds so far, and indeed all sports models, in an important way. Most sports models attempt to model match-win probabilities, and we have been using these to compare to the odds to look for anomalies. Now, however, we are constructing a model explicitly based on the odds data. We trans-

form the odds to implied win probabilities which in turn are transformed to implied λ, which we model directly. In doing so, it is possible that our model may provide less accurate predictions of the match-result, but this is a trade-off we are prepared to make. We do not necessarily believe that altering λ_k (the implied λ for match k) provides better predictions for the match-win probabilities in match k, but if it describes the market better then it is important to follow the market. Of course, if the market is perfectly efficient, then there is no philosophical difference between the two approaches. However, if the market is inefficient, (in ways that do not relate to match-fixing), we wish to follow the market.

Consider for example the so-called "favourite-longshot bias" observed in some sports, discussed in a football context in Reade (2014), in which gamblers prefer to bet on underdogs to favourites, causing some predictable inefficiencies. In this case, we would ideally wish to model this market ineffiecency rather than correcting for this bias to make better predictions. Forrest and McHale (2007) note the existence of a favourite-longshot bias in tennis in general, but we did not attempt to use their findings to predict the scale of the bias in individual matches in this thesis. This presents a potential area for further work.

8.1 The Gaussian Process In-Play Model

The model that we will propose has a major advancement on those previously discussed. We want to be able to track the odds and implied λ more closely in matches where no match-fixing occurs. We found that the Bayesian method reacted very little to players winning or losing games compared when compared to how the implied win probabilities reacted, so would like to rectify this. Our model will do this by exploiting knowledge over many different matches to look for anomalies, rather than simply looking at data from a single match.

To define this model, we first recall the definition of Gaussian processes from Sec-

tion 6.3. We want to build a Gaussian process for the implied λ_k in match k for all k, in which the mean of the implied λ_k during the match is some parameter α_k which is independent of the current match score, and the mean of implied λ_k then also rises and falls as games and tie-breaks are won or lost, with the sizes of these rises and falls consistent across all matches. We also believe that values of implied λ are correlated in some space. We chose not to consider correlation linked to the associated betting volumes, as we did in Section 6.3, since the principal reason for considering betting volumes in that section was to allow very variable odds at low volumes while the market was poorly formed, which we believed to be less of an issue in the in-play market.

Our data include the time between each game break, which would have been a sensible space to use to measure the correlation between implied λ in successive game breaks. However, for simplicity's sake we chose in the first instance not to use these and instead consider all game breaks to be an equal time apart from each other. Incorporating the actual times between games into this model would be a priority for further work in this area.

In order to model the movements of implied λ, let $\Lambda_k(\tau)$ be a random variable denoting the value of the implied λ at time τ in match k, and let $\mathbf{\Lambda}_k$ be a vector containing all $\Lambda_k(\tau)$ for match k. Let $\lambda_k(\tau)$ and $\boldsymbol{\lambda}_k$ be observations of the random variables $\Lambda_k(\tau)$ and $\mathbf{\Lambda}_k$ respectively. In match k there are n_k observations, and we let $\mathbf{1}_n$ be a vector with a 1 in each of its n entries.

The first model we considered for $\mathbf{\Lambda}_k$ was

$$\mathbf{\Lambda}_k \sim MVN(\alpha_k \mathbf{1}_{n_k} + \boldsymbol{x}_k \boldsymbol{\beta}, \delta^2(C_k + \eta^2 I_{n_k})), \tag{8.1.1}$$

given parameters α_k for $k = 1, \ldots, 274$, $\boldsymbol{\beta}$, δ^2, and a scaled nugget parameter η^2. The correlation matrix for each match k is C_k, and I_n is an identity matrix of size $n \times n$. We chose to use an exponential correlation function, where if $C_{k,ij}$ is the (i,j)th

element of C_k, then

$$C_{k,ij} = \rho^{||\tau_i - \tau_j||} \quad i, j = 1, \ldots, n_k \tag{8.1.2}$$

for some correlation parameter ρ.

The design matrix $\boldsymbol{x}_k$ includes three columns of data related to the differences in the number games and tie-breaks that each player has won or lost on serve. We use game data principally because we do not have points data, but it is also debatable whether points information would add accuracy were it available. It could be that the various tactical decisions tennis players make in deciding which points to strive for make the total number of points each player has won a potentially unreliable marker of the difference in quality of the players.

Recalling notation from Section 7.2.3 and supposing match k is played between players i and j, the three columns in $\boldsymbol{x}_k$ are

$$\begin{aligned}
\boldsymbol{x}_{k1}(\tau) &= k_i^{(g)}(\tau) - k_j^{(g)}(\tau), \\
\boldsymbol{x}_{k2}(\tau) &= \big(n_i^{(g)}(\tau) - k_i^{(g)}(\tau)\big) - \big(n_j^{(g)}(\tau) - k_j^{(g)}(\tau)\big), \\
\boldsymbol{x}_{k3}(\tau) &= k_i^{(t)}(\tau) - k_j^{(t)}(\tau).
\end{aligned} \tag{8.1.3}$$

The covariate $\boldsymbol{x}_{k1}$ is the difference in the number of games each player has won while serving, while $\boldsymbol{x}_{k2}$ is the difference in the number of games each player has lost while serving. Observe therefore that $\boldsymbol{x}_{k1} + \boldsymbol{x}_{k2} = n_i^{(g)}(\tau) - n_j^{(g)}(\tau)$, which should be -1, 0 or 1 at all times due to the fact that players alternate serve. The covariate $\boldsymbol{x}_{k3}$ is the difference in the number of tie-breaks each player has won. We expect these three variables representing the differences in the games and tie-breaks that the players have won to be the primary predictors of $\lambda_k = (p_i - p_j)/2$.

In the design matrix $\boldsymbol{x}_k$, we use differences in games won or lost to ensure that no unwanted effects are included in the model due to the ordering of the two players. For example, our model uses the eventual winner as the first player in all cases, but we do not necessarily want to boost implied λ for the eventual winner during play

before the winner is decided - the market does not know the eventual winner during play, and so the λ implied by the market should not use this information.

One consequence of this symmetry is that if both players have won and lost the same number of service games and tie-breaks after time τ, we see $E(\Lambda_k(\tau)) = E(\Lambda_k(0)) = \alpha_k$, and so we expect no movement in implied λ. However, if one player is believed to be much stronger than the other, perhaps this is undesirable. If we expect one player to dominate but is instead being matched by the underdog, it is possible that we expect $E(\Lambda_k(\tau))$ to drift toward 0 - though careful thought needs to be applied to what we mean by when one player is "expected" to dominate. To formalise this notion, we also try another model in which the shift in $E(\Lambda_k(\tau))$ is based on how the player performs compared to expectations, which is

$$\boldsymbol{\Lambda}_k \sim MVN(\alpha_k \mathbf{1}_{n_k} + (\boldsymbol{x}_k - E(\boldsymbol{X}_k))\boldsymbol{\beta}, \delta^2(C_k + \eta^2 I_{n_k})). \tag{8.1.4}$$

where $\boldsymbol{X}_k$ is a matrix of random variables $\boldsymbol{X}_k(\tau)$ of which $\boldsymbol{x}_k(\tau)$ are observations. We must decide what the expected value $E(\boldsymbol{X}_k(\tau))$ is conditional on. For example, consider the equation for $\boldsymbol{x}_{k1}(\tau)$ in equation (8.1.3) and consider the corresponding random variable

$$\boldsymbol{X}_{k1}(\tau) = K_i^{(g)}(\tau) - K_j^{(g)}(\tau),$$

where we recall from the earlier equation (7.2.3) that

$$K_i^{(g)}(\tau) \sim \text{Bin}(n_i^{(g)}(\tau),\ g(\mu_k + \lambda_k)),$$
$$K_j^{(g)}(\tau) \sim \text{Bin}(n_j^{(g)}(\tau),\ g(\mu_k - \lambda_k)),$$

for some values μ_k and λ_k in match k, so that

$$E(\boldsymbol{X}_{k1}(\tau)) = n_i^{(g)}(\tau)\ g(\mu_k + \lambda_k) - n_j^{(g)}(\tau)\ g(\mu_k - \lambda_k).$$

The expectation therefore depends on the information used to calculate μ_k and λ_k for match k. For μ, we take the global average suggested by Klaassen and Magnus (2003) of 0.645. For λ_k, we chose to condition on the last available information from

pre-match odds and take $\lambda = m^{\leftarrow}(y_k(0)|\mu_k, \mathbf{0}, b)$ for pre match implied win probability $y_k(0)$.

Whichever μ_k and λ_k are used, a quick examination of our variables reveals that

$$\boldsymbol{x}_{k1}(\tau) - E(\boldsymbol{x}_{k1}(\tau)) = \Big(k_i^{(g)}(\tau) - k_j^{(g)}(\tau)\Big) - \Big(n_i^{(g)}(\tau)g(\mu_k + \lambda_k) - n_j^{(g)}(\tau)g(\mu_k - \lambda_k)\Big),$$

and

$$\begin{aligned}\boldsymbol{x}_{k2}(\tau) - E(\boldsymbol{X}_{k2}(\tau)) &= \Big(\big(n_i^{(g)} - k_i^{(g)}(\tau)\big) - \big(n_i^{(g)} - k_j^{(g)}(\tau)\big)\Big) - \\ &\qquad \Big(\big(n_i^{(g)}(\tau) - n_i^{(g)}(\tau)g(\mu_k + \lambda_k)\big) - \big(n_j^{(g)}(\tau) - n_j^{(g)}(\tau)g(\mu_k - \lambda_k)\big)\Big) \\ &= -\Big(k_i^{(g)}(\tau) - k_j^{(g)}(\tau)\Big) - \Big(- n_i^{(g)}(\tau)g(\mu_k + \lambda_k) + n_j^{(g)}(\tau)g(\mu_k - \lambda_k)\Big) \\ &= -\Big(\boldsymbol{x}_{k1}(\tau) - E(\boldsymbol{X}_{k1}(\tau))\Big)\end{aligned}$$

Hence we no longer need to consider $\boldsymbol{x}_{k1}(\tau)$ and $\boldsymbol{x}_{k2}(\tau)$ separately, and we must instead remove one of the variables from the model. The variable $\boldsymbol{x}_{k3}(\tau)$ remains in the model as before.

If this model in equation (8.1.4) is used, should both players win the same number of games at time τ out of $n(\tau)$ service games each, then

$$E(\Lambda_k(\tau)) = \alpha_k - n(\tau)\big(g(\mu_k + \lambda_k) - g(\mu_k - \lambda_k)\big).$$

If λ_k is greater than 0, then this expected value will be less than α_k, and so $E(\Lambda_k(\tau)) < E(\Lambda_k(0)) = \alpha_k$.

We shall explore how the models in both equation (8.1.1) and equation (8.1.4) behave, but will first discuss fitting the parameters. This process will be described in terms of the model in equation (8.1.1), but the process is the same for the model in equation (8.1.4).

8.2 Parameter Estimation

In order to fit the model in (8.1.1), we want to find maximum likelihood estimators of the each of the parameters. The parameter α_k depends only on each match k,

but we want the parameters $\boldsymbol{\beta}$, δ^2, ρ and η^2 to be common to all matches. In order to estimate these parameters, we therefore need these parameters to maximise the product of the likelihood of all matches. In order to do this, we found it easiest to set up a new Gaussian process containing the data from each of our K matches, given by

$$\boldsymbol{\Lambda} \sim MVN\Big((Z, \boldsymbol{x})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}, \delta^2\tilde{C}\Big), \tag{8.2.1}$$

where

$$\boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_1\\ \vdots\\ \boldsymbol{\Lambda}_K\end{pmatrix}, \qquad \boldsymbol{x} = \begin{pmatrix}\boldsymbol{x}_1\\ \vdots\\ \boldsymbol{x}_K\end{pmatrix}, \qquad \boldsymbol{\alpha} = \begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_K\end{pmatrix},$$

$$Z = \begin{pmatrix} 1 & 0 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 1 & 0 & \dots & 0\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 1 & \dots & 0\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & 1\\ \vdots & \vdots & & \vdots\\ 0 & 0 & \dots & 1 \end{pmatrix}, \qquad \tilde{C} = \begin{pmatrix} \tilde{C}_1 & 0 & \dots & 0\\ 0 & \tilde{C}_2 & \dots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \dots & \tilde{C}_K \end{pmatrix}, \quad \text{and} \quad \tilde{C}_k = C_k + \eta^2 I_{n_k}$$

The idea behind this Gaussian process is to simultaneously model all $\boldsymbol{\Lambda}_k$ in one vector, $\boldsymbol{\Lambda}^\top = (\boldsymbol{\Lambda}_1^\top, \dots, \boldsymbol{\Lambda}_K^\top)$. This vector is of length $N = \sum_{k=1}^{K} n_k$, since each vector $\boldsymbol{\Lambda}_k$ is of length n_k. We also put each parameter α_k into a vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)^\top$, and stack the design matrices $\boldsymbol{x}_k$, each of which is of size $n_k \times 3$, to get an $N \times 3$ matrix which we call $\boldsymbol{x}$.

The matrix Z is an $N \times K$ matrix with values indicating which match each data point belongs to. In the k-th column of Z there are n_k values equal to 1 in the same locations as the locations of the values of $\boldsymbol{\Lambda}_k$ in $\boldsymbol{\Lambda}$. All other entries in column k are

equal to 0. Hence, each row i has exactly one entry equal to 1, indicating whether the i-th element of $\mathbf{\Lambda}$ corresponds to match k, and all other entries are 0.

The matrix product $Z\boldsymbol{\alpha}$ is then a vector of length N. For each k, there are n_k elements equal to α_k in the same locations as the entries corresponding to $\mathbf{\Lambda}_k$ in $\mathbf{\Lambda}$, and all other entries are 0. If the ith entry of $\mathbf{\Lambda}$ corresponds to time τ_j in match k, and $(Z, \boldsymbol{x})_i$ is the i-th row of $(Z, \boldsymbol{x})$, then the mean of $\Lambda_k(\tau_j)$ is therefore

$$E(\Lambda_k(\tau_j)) = (Z, \boldsymbol{x})_i \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} = \alpha_k + \boldsymbol{x}_{k,j}\boldsymbol{\beta},$$

as required.

To model the correlation between different data points, we want the correlation between data points in different matches to be unchanged from in, but the correlation between data points in different matches to be 0. To do this, the variance matrix needs to be a block-diagonal matrix with the matrices C_k on the diagonal for $k = 1, \ldots, K$.

When the model is formulated as in (8.2.1), it is then easy to find maximum likelihood estimators for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and δ^2 conditional on $\tilde{C}$, which in turn relies on the parameters ρ and η^2, recalling that ρ is the correlation parameter use to generate each matrix C_k in equation (8.1.2). Given the parameters ρ and η^2 we can find profile maximum likelihood estimators $\hat{\boldsymbol{\alpha}}(\rho, \eta^2)$, $\hat{\boldsymbol{\beta}}(\rho, \eta^2)$ and $\hat{\delta}^2(\rho, \eta^2)$ using standard regression formulae for normal distributions, namely

$$\begin{pmatrix} \hat{\boldsymbol{\alpha}}(\rho, \eta^2) \\ \hat{\boldsymbol{\beta}}(\rho, \eta^2) \end{pmatrix} = \Big((Z, \boldsymbol{x})^\top \tilde{C}^{-1} (Z, \boldsymbol{x}) \Big)^{-1} (Z, \boldsymbol{x})^\top \tilde{C}^{-1} \mathbf{\Lambda},$$

$$\hat{\delta}^2(\rho, \eta^2) = \frac{1}{N} \left(\mathbf{\Lambda} - (Z, \boldsymbol{x}) \begin{pmatrix} \hat{\boldsymbol{\alpha}}(\rho, \eta^2) \\ \hat{\boldsymbol{\beta}}(\rho, \eta^2) \end{pmatrix} \right)^\top \tilde{C}^{-1} \left(\mathbf{\Lambda} - (Z, \boldsymbol{x}) \begin{pmatrix} \hat{\boldsymbol{\alpha}}(\rho, \eta^2) \\ \hat{\boldsymbol{\beta}}(\rho, \eta^2) \end{pmatrix} \right)$$

where of course the easiest way to compute the large matrix inverse $\tilde{C}^{-1}$ is

$$\tilde{C}^{-1} = \begin{pmatrix} \tilde{C}_1^{-1} & 0 & \ldots & 0 \\ 0 & \tilde{C}_2^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \tilde{C}_K^{-1} \end{pmatrix}.$$

To find the maximum likelihood estimators for ρ and η, the parameters used to calculate $\tilde{C}$, we do a numerical search over these parameters, calculating the likelihood based on the profile maximum likelihood estimators $\hat{\boldsymbol{\alpha}}(\rho, \eta^2)$, $\hat{\boldsymbol{\beta}}(\rho, \eta^2)$ and $\hat{\delta}^2(\rho, \eta^2)$ and select the values that maximises this likelihood.

8.3 Results

We fit both the models in equations (8.1.1) and (8.1.4), and found that the model in equation (8.1.1) was better in terms of both AIC and BIC, and so we focus on this model from hereon in. (Excluding the nugget effect was also found to be inferior). It seems that while it may have been helpful to allow for the pre-match differences in perceptions of the player strengths to affect model fit in the model that incorporated $E(\boldsymbol{X}_k)$, the flexibility lost by considering holds of serves and breaks separately in (8.1.1) outweighed this gain. It is possible that updating the λ used to calculate $E(\boldsymbol{X}_k)$ may yield sufficiently improved model fit for the model to be worth using, but we have yet to explore this possibility further.

Our analyses therefore all focus on the model in equation (8.1.1). We begin by looking at a few example fits in Figure 8.3.1. These have been transformed onto the match-win probability space in Figure 8.3.2 using the method discussed in Section 6.1.

In matches 1 and 2 in particular it is notable how well the model tracks the data for the majority of the match. Matches 5 and 6 also start well before differences begin in the second half. Contrast these plots with those in Figure 7.3.1, in which the posterior distributions for λ barely move by comparison. In matches such as match 5 in which there is an in-play swing in $\boldsymbol{\lambda}_k$, the Gaussian process attempts to straddle the middle of the data to find the best possible fit.

There may be several reasons why $\boldsymbol{\lambda}_k$ appears to be more variable towards the ends of matches. One reason may be due to our choice of μ in each match. For this

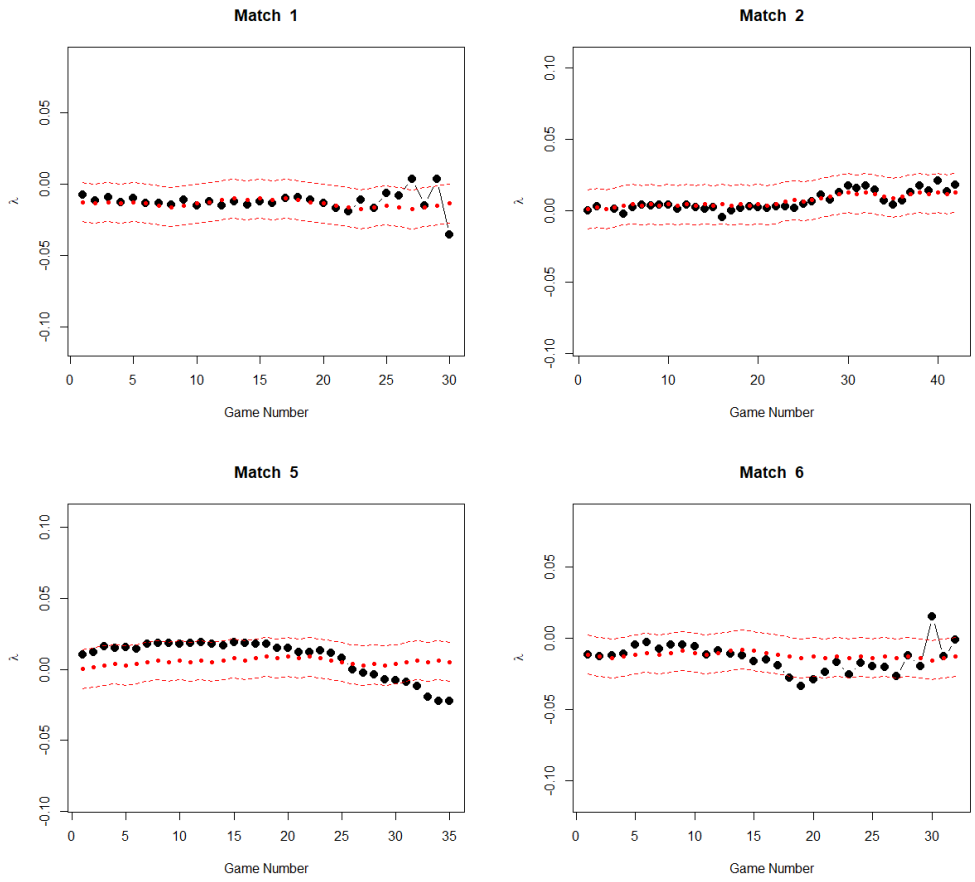

Figure 8.3.1: Gaussian process fits for four matches, based on the model in equation (8.1.1). Black dots represent $\boldsymbol{\lambda}_k$, and red dots represent the mean and a 95% predictive interval for $\boldsymbol{\lambda}_k$.

model we used the global average, $\mu = 0.645$. In Section 2.3, we discussed and saw in Figures 2.3.6 and 2.3.7 how the choice of μ made little difference early in matches, but could have a much more significant impact toward the end of matches. It is therefore possible that poor choice of μ has meant that the process of obtaining implied λ by taking $\lambda_k(\tau) = m^{\leftarrow}(y_k(\tau)|\mu, \boldsymbol{s}(\tau), b)$ has introduced irregularities.

However, there may also be other factors at play. When odds are very short, and the probability of a win is very close to 1, some artefacts may also creep into the data of the process of obtaining implied win-probabilities, similarly to what was discussed about match 22 and shown in Figure 7.4.3. Finally, the market may genuinely be overreacting to each individual game result if the final set is very close.

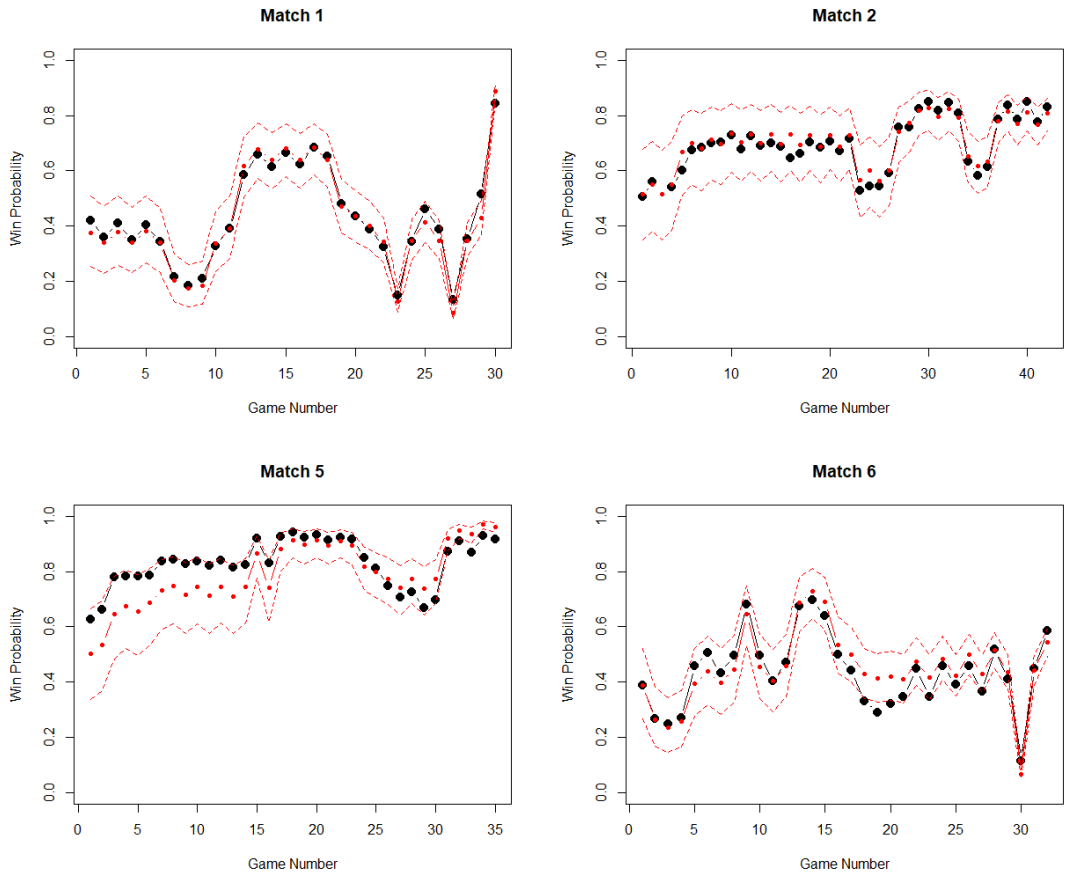

Figure 8.3.2: Gaussian process fits for four matches, based on the model in equation (8.1.1), transformed onto the match-win probability space. Black dots represent $\boldsymbol{y}_k$, and red dots represent the mean and a 95% predictive interval for match-win probability.

8.3.1 Mahalanobis Distance

In order to assess model fit we shall use Mahalanobis distance, Mardia et al. (1979). If an observation of a random variable $\boldsymbol{v}$ of length n is drawn from any distribution with mean $\boldsymbol{u}$ and variance Σ, then the Mahalanobis distance between $\boldsymbol{v}$ and $\boldsymbol{\mu}$ is

$$D^2(\boldsymbol{v}) = (\boldsymbol{v} - \boldsymbol{u})^\top \Sigma^{-1} (\boldsymbol{v} - \boldsymbol{u}). \tag{8.3.1}$$

Note therefore that if $\boldsymbol{v}$ is normally distributed, the likelihood $L(\boldsymbol{u}, \Sigma|\boldsymbol{v})$ is linked to the Mahalanobis distance by the equation

$$L(\boldsymbol{u}, \Sigma|\boldsymbol{v}) = 2\pi^{-\frac{n}{2}}\det(\Sigma)^{-\frac{1}{2}}e^{-\frac{1}{2}D^2(\boldsymbol{v})}.$$

The reason we use Mahalanobis distances is to help us to compare matches of different lengths. The log likelihoods and Mahalanobis distances are both affected by the length of the random vector, making comparison across different matches potentially difficult. By contrast, if the random vector $\boldsymbol{v}$ is normally distributed and of length n, then the Mahalanobis distance has the property that

$$D^2(\boldsymbol{v}) \sim \chi^2_n. \tag{8.3.2}$$

Under the assumption that our Gaussian process fit the data well, if we look at the p-values of the observed Mahalanobis distance for each match compared to the appropriate χ^2 distribution then we should be able to compare the fit in different matches without being concerned about the lengths of matches. Looking at the distribution of these p-values should also help us to assess model fit, since these p-values should be uniformly distributed across all matches. We begin by looking at the matches with the greatest Mahalanobis distances and then compare the results with the matches whose Mahalanobis distances have the smallest p-values with respect to their appropriate Mahalanobis distances.

Figure 8.3.3 shows the Mahalanobis distances for each match in our set of 274. We see that these values are generally fairly consistent, with a few significant outliers. We consider plot the data for these outliers in Figure 8.3.4. Each of these matches features a large swing in in-play implied λ. The Gaussian process fails to capture this swing, and the mean function sits in the centre of the data. It is also notable that all four swings are in the direction of the eventual winner, which as we have previously discussed can be an additional indicator that the swing could have come about due to gamblers having knowledge of the final result. In match 211, the implied λ swings back toward the mean at the end of the match, but when we look at the match-win

probabilities in Figure 8.3.5, we see this only happens once the win probabilities are very close to 1, and hence the fit may not be quite as reliable.

Matches 73 and 136 are two matches have been cited in other sources as containing very suspicious betting activity. The fact that our model can identify these matches and quantify how unusual the odds are relation to typical betting behaviour shows that our model has the potential to be an important tool in identifying and flagging matches with suspicious betting activity for further investigation.

The right-hand plot in Figure 8.3.3 shows $\log_{10}$ of the p values, $P(D(\mathbf{\Lambda}_k) >$

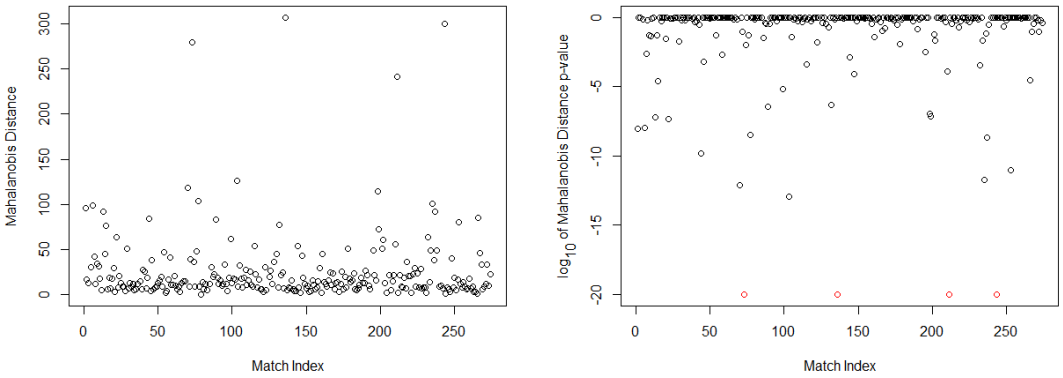

Figure 8.3.3: The left-hand plot shows Mahalanobis distances for each match based on the model in equation (8.1.1). The right-hand plot shows $\log_{10}$ of the p-values for the Mahalanobis distances in each match k with respect to a $\chi^2_{n_k}$ distribution. The matches in red have p-values computationally indistinguishable from 0, but are plotted on this figure to show their match index.

$D(\boldsymbol{\lambda}_k))$ with respect to a $\chi^2_{n_k}$ distribution given that each match is of length n_k games. In the four matches discussed in Figures 8.3.4 and 8.3.5 have got Mahalanobis distance p-values far smaller than the p-values in any other matches. (In fact, there p-values are so small that they are computationally indistinguishable from 0, though they cannot actually be 0 since the χ^2 distribution has no upper endpoint. Since $\log_{10}(0) = -\infty$, which we cannot plot, these matches are instead plotted at -20 in Figure 8.3.3 and marked in red.) These results based on p-values therefore agree with

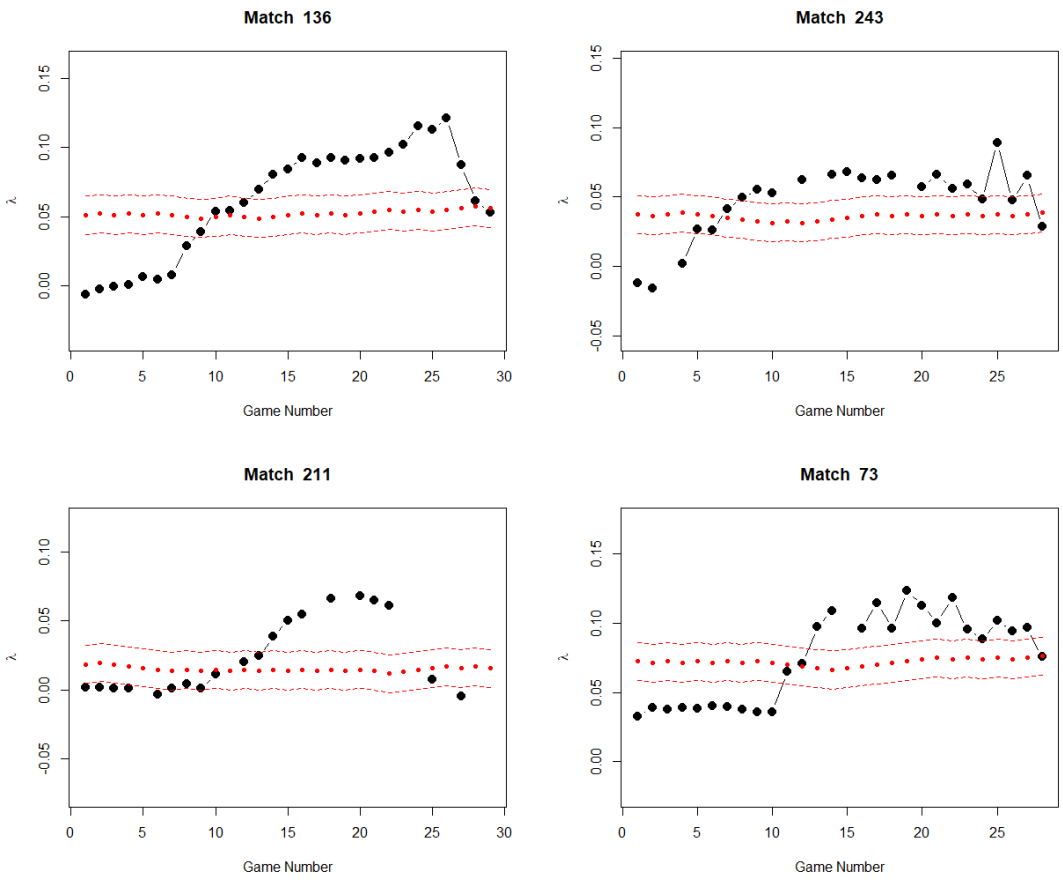

Figure 8.3.4: Gaussian process fits for the four matches with the largest Mahalanobis distances, based on the model in equation (8.1.1). Black dots represent $\boldsymbol{\lambda}_k$, and red dots represent the mean and a 95% predictive interval for match-win probability.

our results purely based on Mahalanobis distances.

A word of caution must be applied, however. In Figure 8.3.6 we show a histogram of the p-values of all of the Mahalanobis distances. We see that these p-values are certainly not uniformly distributed (i.e. as we would expect if our model were correct), with large peaks in frequency around 1 and 0. Were the implied λ perfectly normally distributed in each match, we would expect these p-values to be uniformly distributed, suggesting that there may be some further work to improve model fit, even though we have already successfully identified the worst offenders. This would help more accurately flag matches that are not as obviously suspect as our four worst

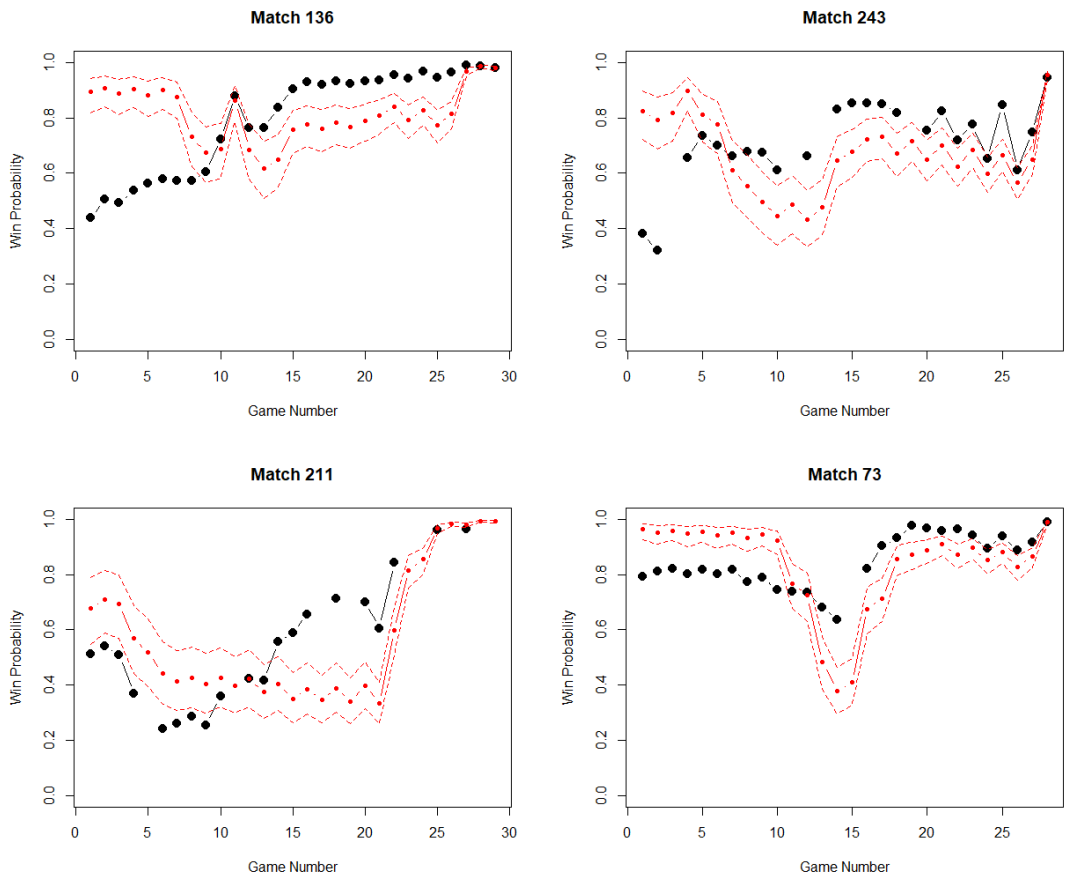

Figure 8.3.5: Gaussian process fits for the four matches with the largest Mahalanobis distances, based on the model in equation (8.1.1), transformed onto the match-win probability space. Black dots represent $\boldsymbol{y}_k$, and red dots represent the mean and a 95% predictive interval for match-win probability.

offenders so far.

If the values of implied λ were perfectly normally distributed, we would expect around 10% of our p-values to sit in each bin, which would correspond to about 27 matches. However, we see that around 150 matches for which the Mahalanobis distance is very small, and so the data are very close to the mean function, and around 50 matches in which the Mahalanobis distance is large and the data are far from the mean function.

Ensuring a common variance parameter δ^2 across all matches will undoubtedly

have had an impact on this. In some matches, such as match 2 in Figure 8.3.1, the data vary very little around the mean, and a variance parameter δ_k^2 fitted solely to that match would clearly be much smaller than the global δ^2, whereas a variance parameter δ_k^2 fitted to match 136 would clearly need to be much larger to account for this swing. Using a global variance parameter δ^2 is one of the features that allows us to identify matches with large swings, and so is a necessary sacrifice to accomplish our goals at the expense of modelling our data as accurately as possible.

The variability in $\boldsymbol{\lambda}_k$ toward the ends of matches is probably a larger concern. The global variance parameter δ^2 will undoubtedly be inflated by the frequent outliers in $\boldsymbol{\lambda}_k$ that occur toward the ends of otherwise well-behaved matches, such as matches 1 and 6 in Figure 8.3.1. We have performed some preliminary research into using a different parameter μ_k to represent the average of both players' serving ability that looked to make a promising start on reducing some of this variability, but this work requires further refinement. We also briefly explored using a student-t process rather than a Gaussian process to allow for longer tails in the distributions, as well as using some covariates to attempt to directly model this extra variability, but none of these appeared to resolve the issue entirely.

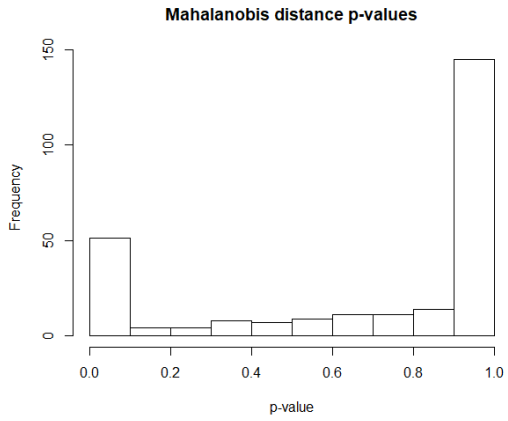

Figure 8.3.6: A histogram of the p-values for the Mahalanobis distances in each match with respect to a $\chi^2_{n_k}$ distribution.

8.4 Conclusion and Further Work

In this chapter, we have described a Gaussian process model for in-play implied λ in our matches in order to look for swings in in-play odds. Rather than simply using in-play data on games won and lost to update our estimates of λ, we modelled the values of λ implied by the data directly, using data from across our 274 matches to model how this implied λ process rises and falls as games are won and lost during play. This allowed us to model these movements more closely, making a more accurate model as a result. This differs significantly from other models in that it attempts to model odds directly rather than simply comparing odds with match predictions.

We compared a few different variations of our model, and found the one with the best balance between simplicity and modelling accuracy was the more complicated of our models. We looked at fit for a few different matches, and found that we successfully identified poor fit in the matches that have been cited elsewhere as suspicious. This suggest that our model could prove a helpful tool in flagging matches to help investigate match-fixing. Apart from the proprietary models described in Forrest and McHale (2019), we know of no other models that attempt to detect in-play match-fixing.

Some issues do remain with our model. Looking at the p-values of Mahalanobis distances suggests that our Gaussian model fit is not quite ideal. Some of this lack of fit is caused by the deliberate choice of using a common variance parameter δ^2 in all matches. Another issue is the extra variability that we see in implied λ toward the end of matches. Though we have begun to investigate possible ways to model this extra variability, we have yet to satisfactorily resolve this issue. Further work in this area should focus heavily on this issue.

Chapter 9

Conclusions

In this thesis, we gave ourselves the task of developing new methods for identifying potentially fixed matches in tennis. In doing this, we have extended the literature on predicting the outcomes of tennis matches to suit our needs before proceeding to research how best to separate the clean matches from the fixed. In this chapter we summarise the main contributions of our work and the key areas for further research.

9.1 Tennis Modelling

Chapters 3 and 4 were principally concerned with tennis modelling. The work in Chapter 3 proved that the probability of a player winning a match from any scoreline, given by the function $m(\lambda|\mu, \boldsymbol{s}, b)$ is invertible in the dominance parameter λ, for all scores $\boldsymbol{s}$ in a best-of-b sets match given that the players' average point-win probability is μ. Hence, the inverse of this function can be used in other sections of this thesis. This function is numerically inverted elsewhere in the literature without proof, and so it is reassuring for it to be definitively proven that this is possible. The proofs also helped formulate and prove intuitive ideas about Markov chains representing contests between two players in the class of Markov chains we defined as "first to

$(M+1, N+1)$" Markov chains. In order to prove these results, we proved that the absorption probability in this class of Markov chains was continuous and increasing in some parameter α. This proof could then be used to prove the same results for the Markov chains relating to games, sets, tie-breaks and matches, although these Markov chains did not fall directly into the relevant class. Further work in the area could involve extending the proofs to Markov chains of different shapes and with different transition probabilities, but this is not necessary for further analysis into match-fixing in tennis.

Chapter 4 expanded on the explanations of Glicko ratings in the literature and described our implementation of Glicko ratings, as well as a new way of incorporating 5-set matches into the Glicko ratings framework. There are a few main avenues for further work into Glicko ratings. Some concern the use of Glicko ratings in rating tennis players specifically, while there is also potential for a wider investigation into its effectiveness as a ratings system compared to other player rating systems.

We decided it was important to develop a method to lend extra weight to the outcomes of 5-set matches to reflect that fact that the increased length of such matches mean that luck plays less of a part, and that stronger players are more likely to triumph. Our method of weighting 5-set matches differently to 3-set matches gave an analytically-driven method to impart extra weight to matches played over 5 sets. This method includes a parameter that could be altered depending on the data, but our implementation suggested that the value given by the assumption of independent sets fits well.

An alternative model to this would be to model the number of sets, games or points won by each player instead of the match winner. This would naturally lend more weight to 5-set matches, as they feature more sets than 3-set matches, and would also allow the margin of victory to affect the shift in player ratings. Researching this further could potentially help the Glicko ratings give better predictive accuracy using data from fewer matches.

Another important area for further study would be accounting for the fact that players perform differently on different surfaces. One idea that may solve this issue would be to use correlated rankings for different surfaces, but significant further work is required to see whether this would prove successful.

There is also a wider issue that Glicko ratings appear not to have been widely compared to other ratings systems to see how well it performs. Glickman (1999) does not compare Glicko ratings to the Elo ratings that Glicko ratings are ostensibly built on, and it is not considered in the comparative study of Kovalchik (2016). It would be interesting to build on the work of Kovalchik (2016) to consider Glicko ratings and other methods to build a fuller picture of which ratings systems are best for modelling tennis players.

9.2 Match-Fixing

The remainder of the thesis concerns ways of attempting to identify potentially fixed matches. These expand on the existing literature in several ways. Our pre-match analysis uses betting volumes and odds recorded at more time periods than other works, whereas our in-play analyses features the first algorithms to be presented in academic literature to look for in-play match-fixing in tennis.

The work in Chapter 6 concerned our method for investigating the abnormality of pre-match odds in tennis matches. Most work in the literature looks simply at the difference between the opening and closing pre-match odds. We consider this unsatisfactory, as odds can be very volatile near the opening of the market, and it ignores the development of the pre-match odds between their opening and closing. We therefore grouped pre-match odds into several intervals to build a picture of the movement of odds as the market progressed. We also wanted to investigate whether betting volumes could be incorporated to provide additional information.

We built a model under the assumption that there was some "fair" value of odds,

which the actual odds converged to as the match approached. We did this by using a Gaussian process with constant mean and decreasing variance. On investigating whether to model this decrease using time or betting volumes, we found that betting volumes provided better fit. Mixing time and betting volumes provided some surprising results, the causes of which we failed to identify. Further research may shed light on this.

Fitting the model using maximum likelihood seemed unwise, as we had at most 8 data points per match. We therefore used a Bayesian approach, with the Glicko ratings used to generate a prior for the pre-match implied win-probability mean parameter. This allowed us to assess whether the opening pre-match odds were surprising with respect to the Glicko ratings, and use successive posterior ratings for the mean to examine the development of the pre-match odds.

We found that the method successfully flagged matches with large pre-match swings, achieving the desired result. It is hard to assess the improvement on existing literature as there is so little, and we cannot use common data. However, we believe that our results show strong potential that they may be useful for highlighting matches with suspicious odds patterns, taking a more nuanced approach than the existing comparisons of opening and closing odds.

The method also flagged several matches where the first recorded odds disagreed with our Glicko predictions. This was also interesting, but is much less likely to be caused by match-fixing, since very little gambling has happened by this stage. Stronger tennis modelling, especially incorporating surfaces, may account for some of these flagged matches. In one match in particular, the opening odds were far out from our Glicko predictions, but swung back towards our prediction, possibly indicating that the opening odds were merely mis-specified. Careful further thought must be provided to what sorts of matches we wish to flag or not, and how to alter the models to achieve this result.

In Chapter 7, we developed an in-play model for odds using Bayesian statistics. We

used Glicko ratings to generate prior distributions for player strengths, and updated these strengths throughout the match to provide posterior estimates for the match-win probabilities with predictive intervals in each interval between games. This was a new method of modelling in-play match-win probabilities. Using this model, we were able to find p-values for the implied win probabilities with respect to the posterior distributions of match-win probability in each game break. We used these p-values to identify the most suspicious matches, and found that while some matches that looked very peculiar were flagged, there were more other matches that were also flagged than hoped.

In Chapter 7, as with Chapter 6, we attempted to summarise the different p-values at each time-point into a unified statistic by looking at both the average of the p-values at each time point and the minimum. (The method in Chapter 8 lent itself naturally to the use of Mahalanobis distances to summarise the difference between the observed and expected implied win probabilities.) While these methods helped paint a picture of whether the betting activity was suspicious or not, further investigation is required to determine how best to summarise the information each method presents.

One issue was that the match-win probability estimates provided by our implementation of Glicko ratings sometimes disagreed strongly with the opening in-play odds. While this potentially provides useful information that there may be an issue with the pre-match market, we found it difficult to make inference about the in-play market in this scenario. One possible cause of the odds differing from the Glicko estimate could be the effect of this surface. Without having investigated this phenomenon further, it is not clear how much this could potentially affect our ratings. It is also possible that other models may provide stronger predictive performance than Glicko ratings, and alleviate this problem. The reason we chose Glicko ratings was so that we could have larger prior variance if a player was returning from a long absence. There may be another way of accomplishing this with a different model, though we have found no others that directly consider this issue.

Another issue with this model was that in-play posterior estimates of the players' relative strengths did not update as much as hoped in line with the movements in implied win probability. This was likely due to the fact that there is only so much information that can be learned from the number of games won or lost. It is possible that ratings would update quicker with point-by-point data.

These issues led us to develop the model in Chapter 8. The work in Chapter 3 allowed us to take implied win-probabilities and invert the function for turning the dominance parameter λ into win probabilities, $m(\lambda|\mu, \boldsymbol{s}, b)$, to directly model implied λ. Our Gaussian process-based model pooled data from across the different matches in our odds data to estimate how much the implied λ typically rises and falls with games won and lost. This model tracked the odds much more closely, providing better in-play estimates for reasonable in-play win probabilities. In doing so, we circumvented the issue of estimating player strengths (and by extension surface effects) by instead seeking to answer the question of whether the development of the odds was consistent with other matches. This meant we would only flag up matches where the odds swung in a manner not implied by the pattern of games won and lost, leaving the question of whether the pre-match odds were consistent with player strengths as a question purely for the pre-match methods.

While this model appeared to have had some success in modelling in-play odds, there are further improvements to be made. The distribution of the distance of the observed implied λ from our model fit, defined by Mahalanobis distances, was heavier in both tails than expected, suggesting that the marginal distribution of odds at some times was far more variable than our model implied at some times, and far less at other times. The causes of this are currently unclear. We briefly investigated using different average point-win probabilities μ in each match, but this needs significant further research. Other possible solutions could involve using a more heavy-tailed distribution than Gaussian for marginal distributions, such as a student-t distribution, or finding extra parameters to explain odds movement, particularly in the final sets

of matches.

Nonetheless, this model seemed to have greater success in filtering out matches that did and did not have large in-play swings than our Bayesian method, suggesting it has potential to form the basis of a valuable tool for flagging matches with suspicious in-play betting activity.

As well as these improvements to the match-fixing methods themselves, there remains a larger question of how to tie them together to give a unified strategy for flagging matches as potentially suspicious. The pre-match methods provide one indicator of suspicious activity, while we currently favour the Gaussian process method for identifying suspicious in-play betting patterns, as it broadly appears to be better at identifying matches with suspicious odds movements. However, in order to confirm whether this is true, and to provide more rigorous tests of the methods generally, it could be helpful to use data on more matches, and to find a way to more formally assess how well the methods identify fixed matches.

We should also consider how, and indeed whether, to combine the different information provided by the pre-match and in-play markets. One option would be to flag a match if its "suspiciousness" (however that is measured) rises above a certain threshold in the pre-match market, and raise a separate flag if the "suspiciousness" is sufficiently high in the in-play market. A decision would have to be made on how to select appropriate thresholds - there would be a trade-off here between filtering out clean matches and failing to identify fixed matches. It is possible, however, that the information from the pre-match and in-play could be combined in a cleverer way. If the pre-match and in-play markets are both almost suspicious enough raise an alert but not quite, would it be correct to discount the match? SportRadar's Fraud Detection System (FDS), as described by Forrest and McHale (2015), uses green, yellow and red alerts to label increasingly severe cases of abnormal betting activity. This seems like a sensible way of dealing with different levels of suspicious betting behaviour, and we could utilise something similar with our methods, with different thresholds

for different levels. If the abnormality of one market's activity reaches a certain high threshold, the match could automatically be flagged, but both the in-play and pre-match reach a lower threshold, this could also trigger an alert.

In order to more rigorously assess how well our models identify fixed matches, ideally it would be possible to set this up as a straightforward classification problem, in which we want to correctly identify as many known fixed matches as possible. There would be a trade-off between missing fixed matches and incorrectly flagging up clean matches. Failing to identify fixed matches would probably be more problematic than falsely flagging a few extra matches for investigation, but we would have to investigate this trade-off further.

One of the challenges we face in doing this is the fact that it is extremely difficult to find reliable records of which matches are fixed and which are not. Tennis authorities are less reticent to name players banned for fixing matches as they used to be, but they still do not specify which matches have been fixed when issuing punishments. If we regularly monitored odds, when a player is banned it might be possible to look through a player's match history and identify the matches with the most suspicious activity. This would be challenging, as we would not know whether the player had fixed one match or many, or even if the fixers had gambled enough to alter the betting market. Conversely, there is a danger of assuming matches to be clean when they are not. The tennis authorities may miss some matches, or fail to gather enough evidence for a prosecution.

Without a way of clearly identifying matches that are fixed, we rely on rumours of matches with suspicious betting activity, without confirmation that those matches are fixed. If the players involved are never banned, is that because they are innocent, or is it due to a lack of more concrete evidence? As such, it is difficult to formulate this problem as a simple classification task. Nonetheless, it would be worth investigating how best to use the information that we do have available in order best identify matches that appear suspicious.

9.3 Concluding Remarks

The goal of this thesis has been to investigate new ways of identifying tennis matches with suspicious betting activity. To accomplish this, we have had to develop new tools for modelling the strengths of players. Principally, however, we have described a method of detecting suspicious pre-match betting activity more sophisticated than any other in the literature, incorporating betting volumes and data at various time points right up until the match start, and have developed the first two methods in academic literature for detecting suspicious in-play betting activity in tennis. While there is further work to be done to refine these methods, and challenges to overcome in determining which matches are best to flag, we strongly believe that these methods could be extremely useful in flagging matches that could potentially be fixed. With extra development, these could provide another tool in the arsenal of investigators seeking to eradicate these crimes from the sport. We sincerely hope that with further research, the relevant authorities can stamp out this affliction, so that professional tennis matches can be played and watched free of the influences of match-fixers, so that the sport can be abused no more by those who would seek to corrupt players and the game for profit.

Bibliography

Agence France-Presse (2017). New AustralianOpen rules designed to prevent first round retirements and withdrawals. `https://www.thenational.ae/sport/tennis/new-australian-open-rules-designed-to-prevent-first-round-retirements-and-withdrawals-1.694505`. Retrieved December 15$^{\text{th}}$, 2019.

Aitchison, J. and Begg, C. B. (1976). Statistical diagnosis when basic cases are not classified with certainty. *Biometrika*, 63(1):1–12.

Barnett, T. and Clarke, S. R. (2005). Combining player statistics to predict outcomes of tennis matches. *IMA Journal of Management Mathematics*, 16(2):113–120.

Barnett, T. J., Clarke, S. R., et al. (2002). Using Microsoft Excel to model a tennis match. In *6th Conference on Mathematics and Computers in Sport*, pages 63–68.

Barnett, T. J. et al. (2006). *Mathematical modelling in hierarchical games with specific reference to tennis*. PhD thesis, Swinburne University of Technology.

Bernhardt, D. and Heston, S. (2010). Point shaving in college basketball: a cautionary tale for forensic economics. *Economic Inquiry*, 48(1):14–25.

Bialik, C. (2014). Tennis has an income inequality problem. `https://fivethirtyeight.com/features/tennis-has-an-income-inequality-problem/`. Retrieved December 15$^{\text{th}}$, 2019.

Blake, H. and Templon, J. (2016). The tennis racket. `https://www.buzzfeednews.com/article/heidiblake/the-tennis-racket`. Retrieved April 2$^{\text{nd}}$, 2019.

Borghesi, R. (2008). Widespread corruption in sports gambling: fact or fiction? *Southern Economic Journal*, pages 1063–1069.

Boulier, B. L. and Stekler, H. O. (1999). Are sports seedings good predictors?: an evaluation. *International Journal of Forecasting*, 15(1):83–91.

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.

Brown, A. (2012). Evidence of in-play insider trading on a uk betting exchange. *Applied Economics*, 44(9):1169–1175.

Carrari, A., Ferrante, M., and Fonseca, G. (2017). A new Markovian model for tennis matches. *Electronic Journal of Applied Statistical Analysis*, 10(3).

Clarke, S. R. and Dyte, D. (2000). Using official ratings to simulate major tennis tournaments. *International transactions in operational research*, 7(6):585–594.

Cornman, A., Spellman, G., and Wright, D. (2017). Machine learning for professional tennis match prediction and betting. Master's thesis, Stanford University.

Cox, D. (1987). *The Analysis of Binary Data*. Monographs on applied probability and statistics. Chapman and Hall.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Croxson, K. and Reade, J. J. (2011). Exchange vs Dealers: A High-Frequency Analysis of In-Play Betting Prices. Discussion Papers 11-19, Department of Economics, University of Birmingham.

Del Corral, J. and Prieto-Rodríguez, J. (2010). Are differences in ranks good predictors for grand slam tennis matches? *International Journal of Forecasting*, 26(3):551–563.

Deutscher, C., Dimant, E., Humphreys, B., et al. (2017). Match fixing and sports betting in football. empirical evidence from the german bundesliga. Technical report, Philosophy, Politics and Economics, University of Pennsylvania.

Diemer, G. and Leeds, M. A. (2013). Failing to cover: point shaving or statistical abnormality? *International Journal of Sport Finance*, 8(3):175–192.

Diggle, P. and Ribeiro, P. J. (2007). *Model-based Geostatistics*. Springer.

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.

Duggan, M. and Levitt, S. D. (2000). Winning isn't everything: corruption in sumo wrestling. Technical report, National Bureau of Economic Research.

DW on Sport (2016). No evidence of Lleyton Hewitt fixing matches. `http://www.sportdw.com/2016/01/tennis-fixing-buzzfeed-hewitt-innocent.html`. Retrieved April 2nd, 2019.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

ESSA (2015a). ESSA Q1 integrity report. `http://www.eu-ssa.org/wp-content/uploads/ESSA-AR2015-DEF.pdf`. Retrieved May 2nd, 2019.

ESSA (2015b). ESSA Q2 integrity report. `http://www.eu-ssa.org/wp-content/uploads/ESSA_2015-II-Q-integrity-report.pdf`. Retrieved May 2nd, 2019.

Feustel, E. D. and Rodenberg, R. M. (2015). Sports (betting) integrity: detecting match-fixing in soccer. *Gaming Law Review and Economics*, 19(10):689–694.

FIDE (2017). Fide handbook. `http://www.fide.com/component/handbook/?id=197&view=article`. Retrieved January 28th, 2019.

Forrest, D. and McHale, I. (2007). Anyone for tennis (betting)? *The European Journal of Finance*, 13(8):751–768.

Forrest, D. and McHale, I. (2015). An evaluation of SportRadar's fraud detection system. `https://integrity.sportradar.com/wp-content/uploads/sites/22/2018/11/Sportradar-Integrity-Services_University-of-Liverpool_An-Evaluation-of-the-FDS.pdf`. Retrieved April 2nd, 2019.

Forrest, D. and McHale, I. G. (2019). Using statistics to detect match fixing in sport. *IMA Journal of Management Mathematics*.

Gibbs, J. (2007). Point shaving in the NBA: An economic analysis of the national basketball association's point spread betting market.

Gilsdorf, K. F. and Sukhatme, V. A. (2008). Testing Rosen's sequential elimination tournament model: Incentives and player performance in professional tennis. *Journal of Sports Economics*, 9(3):287–303.

Glickman, M. E. (1999). Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 48(3):377–394.

Glickman, M. E. (2012). Example of the glicko-2 system. `http://www.glicko.net/glicko/glicko2.pdf`. Retrieved January 28th, 2019.

Glickman, M. E. and Doan, T. (2017). The US chess rating system. `http://www.glicko.net/ratings/rating.system.pdf`. Retrieved January 28th, 2019.

Gorgi, P., Koopman, S. J., and Lit, R. (2018). The analysis and forecasting of ATP tennis matches using a high-dimensional dynamic model.

Grinstead, C. M. and Snell, J. L. (2012). *Introduction to probability*. American Mathematical Soc.

Hosmer, D., Lemeshow, S., and Sturdivant, R. (2013). *Applied Logistic Regression*. Wiley Series in Probability and Statistics. Wiley.

Hostačnỳ, J. (2018). *Non-Linear Classification as a Tool for Predicting Tennis Matches*. PhD thesis, Univerzita Karlova, Fakulta sociálních věd.

Irons, D. J., Buckley, S., and Paulden, T. (2014). Developing an improved tennis ranking system. *Journal of Quantitative Analysis in Sports*, 10(2):109–118.

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.

Kirkpatrick, C. and Dahlquist, J. (2010). *Technical Analysis: The Complete Resource for Financial Market Technicians*. Pearson Education.

Klaassen, F. J. and Magnus, J. R. (2001). Are points in tennis independent and identically distributed? evidence from a dynamic binary panel data model. *Journal of the American Statistical Association*, 96(454):500–509.

Klaassen, F. J. and Magnus, J. R. (2003). Forecasting the winner of a tennis match. *European Journal of Operational Research*, 148(2):257–267.

Knottenbelt, W. J., Spanias, D., and Madurska, A. M. (2012). A common-opponent stochastic model for predicting the outcome of professional tennis matches. *Computers & Mathematics with Applications*, 64(12):3820 – 3827. Theory and Practice of Stochastic Modeling.

Kovalchik, S. A. (2016). Searching for the GOAT of tennis win prediction. *Journal of Quantitative Analysis in Sports*, 12(3):127–138.

Lisi, F. and Zanella, G. (2017). Tennis betting: can statistics beat bookmakers? *Electronic Journal of Applied Statistical Analysis*, 10(3).

Madurska, A. M. (2012). *A set-by-set analysis method for predicting the outcome of professional singles tennis matches*. PhD thesis, Imperial College London.

Magnus, J. R. and Klaassen, F. J. (1999). The effect of new balls in tennis: four years at Wimbledon. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 48(2):239–246.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press.

Marginson, D. (2013). What explains the existence of an exchange overround? In Vaughan-Williams, L. and Siegel, D. S., editors, *The Oxford Handbook of the Economics of Gambling*, chapter 16. Oxford University Press, Oxford.

McHale, I. and Morton, A. (2011). A Bradley-Terry type model for forecasting tennis match results. *International Journal of Forecasting*, 27(2):619 – 630.

Mike, S. and Farmer, J. D. (2008). An empirical behavioral model of liquidity and volatility. *Journal of Economic Dynamics and Control*, 32(1):200–234.

Morris, B. and Bialik, C. (2015). Serena Williams and the difference between all-time great and greatest of all time. `https://fivethirtyeight.com/features/serena-williams-and-the-difference-between-all-time-great-and-greatest-of-all-time/#fn-2`. Retrieved January 28$^{\text{th}}$, 2019.

Morris, C. (1977). The most important points in tennis. *Optimal strategies in sport*, pages 131–140.

Newton, P. K. and Aslam, K. (2006). Monte Carlo tennis. *SIAM review*, 48(4):722–742.

Olfers, M., Spapens, T., and Lodder, A. (2014). Study on the sharing of information and reporting of suspicious sports betting activity in the eu 28. `http://ec.europa.eu/sport/news/2014/docs/study_oxford_en.pdf`. Retrieved May 2$^{\text{nd}}$, 2019.

O'Malley, A. J. (2008). Probability formulas and statistical analysis in tennis. *Journal of Quantitative Analysis in Sports*, 4(2).

Ötting, M., Langrock, R., and Deutscher, C. (2018). Integrating multiple data sources in match-fixing warning systems. *Statistical Modelling*, 18(5-6):483–504.

Reade, J. (2014). Detecting corruption in football. In *Handbook on the Economics of Professional Football*, Chapters, chapter 25, pages 419–446. Edward Elgar Publishing.

Reade, J. and Akie, S. (2013). Using forecasting to detect corruption in international football. Technical report, The George Washington University, Department of Economics, Research Program ....

Rodenberg, R. and Feustel, E. D. (2014). Forensic sports analytics: Detecting and predicting match-fixing in tennis. *Journal of Prediction Markets*, 8(1):77–95.

Roll, R. (1984). A simple implicit measure of the effective bid-ask spread in an efficient market. *The Journal of Finance*, 39(4):1127–1139.

Russell, T. and USTA (2010). Going to college or turning pro? Making an informed decision! `http://assets.usta.com/assets/1/15/USTA%20College%20Varsity%20Analysis%20of%20College%20vs%20Pro%20FAQ.pdf`. Retrieved December 15$^{\text{th}}$, 2019.

Sarkissian, J. (2016). Spread, volatility, and volume relationship in financial markets and market maker's profit optimization. *Available at SSRN 2799798*.

Sipko, M. and Knottenbelt, W. (2015). Machine learning for the prediction of professional tennis matches. *MEng computing-final year project, Imperial College London*.

Somboonphokkaphan, A., Phimoltares, S., and Lursinsap, C. (2009). Tennis winner prediction based on time-series history with neural modeling. In *Proceedings of the*

International MultiConference of Engineers and Computer Scientists, volume 1, pages 18–20. Citeseer.

Stephenson, A. and Sonas., J. (2016). *Comparing Predictive Performance Of Chess Ratings With The PlayerRatings Package*. From R package "PlayerRatings: Dynamic Updating Methods for Player Ratings Estimation", version 1.0-1.

United Nations Office on Drugs and Crime (2016). Resource guide on good practices in the investigation of match-fixing. `https://www.unodc.org/documents/corruption/Publications/2016/V1602591-RESOURCE_GUIDE_ON_GOOD_PRACTICES_IN_THE_INVESTIGATION_OF_MATCH-FIXING.pdf`. Retrieved May 2nd, 2019.

Viney, M. (2015). *Prediction of in-play tennis*. PhD thesis, RMIT University.

Vovk, V. and Zhdanov, F. (2009). Prediction with expert advice for the Brier game. *Journal of Machine Learning Research*, 10(Nov):2445–2471.

Williams, L. V. (1999). Information efficiency in betting markets: A survey. *Bulletin of Economic Research*, 51(1):1–39.

Wolfers, J. (2006). Point shaving: Corruption in NCAA basketball. *The American economic review*, pages 279–283.

Wolfers, J. and Zitzewitz, E. (2004). Prediction markets. Technical report, National Bureau of Economic Research.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.