Sports statistics in the data age: betting fraud detection and performance evaluation

Marius Ötting

Dissertation

presented for the degree *Doctor rerum politicarum* (Dr. rer. pol.)
at the Faculty of Economics and Business Administration,
Bielefeld University

Bielefeld, April 2020

First Examiner: Prof. Dr. Roland Langrock
Second Examiner: Prof. Dr. Christian Deutscher
Third Examiner: Prof. Dr. Christiane Fuchs
Thesis defense: 14.08.2020

Printed on acid-free and non-aging paper (ISO 9706)

Für meine Familie

Contents

Acknowledgements **vii**

Statement of contribution **ix**

1 Introduction **1**

1.1 Betting fraud detection 4

1.2 Evaluation of in-game performance 5

2 The demand for English Premier League football betting **9**

2.1 Introduction 9

2.2 Literature review 11

2.3 Empirical analysis 12

2.3.1 Methodology 19

2.3.2 Linear model 19

2.3.3 Flexible approach 21

2.4 Final remarks 25

3 Betting market inefficiencies are short-lived in German professional football **27**

3.1 Introduction 27

3.2 Biases in betting odds 28

3.3 Evidence from the German Bundesliga 30

3.4 Concluding remarks 35

4 Integrating multiple data sources in match-fixing warning systems **37**

4.1 Introduction 37

4.2 Building models for betting volumes and odds 38

4.2.1 Modelling betting volumes 39

4.2.2 Modelling betting odds . . . 45
4.3 Detection of match fixing . . . 50
4.3.1 Classification results based on betting volumes . . . 50
4.3.2 Classification results based on betting odds . . . 51
4.3.3 Combining the classification based on volumes and odds . . . 52
4.3.4 Discussion of the results . . . 53
4.4 Conclusions . . . 54

5 The hot hand in professional darts 57
5.1 Introduction . . . 57
5.2 Data . . . 60
5.3 Modelling the hot hand in darts . . . 62
5.3.1 State-space model of the hot hand . . . 62
5.3.2 Model specifications . . . 64
5.3.3 Maximum likelihood estimation . . . 66
5.4 Results . . . 67
5.5 Discussion . . . 73

6 A regularised hidden Markov model for analysing the 'hot shoe' in football 77
6.1 Introduction . . . 77
6.2 Data . . . 78
6.3 Methods . . . 80
6.3.1 Hidden Markov models . . . 80
6.3.2 Variable selection by the LASSO . . . 82
6.4 A short simulation study . . . 83
6.5 Results . . . 88
6.6 Discussion . . . 90

7 A copula-based multivariate hidden Markov model for modelling momentum in football 93
7.1 Introduction . . . 93
7.2 Data . . . 94
7.3 Modelling momentum . . . 95
7.3.1 A baseline model . . . 96

7.3.2 Modelling within-state dependence using copulas 99
7.3.3 A model including covariates 101
7.4 Results 101
7.5 Discussion 107

8 Performance under pressure in skill tasks: An analysis of professional darts 111
8.1 Introduction 111
8.2 Performance under pressure 113
8.2.1 Terminology 113
8.2.2 Potential effects of pressure 113
8.2.3 Empirical findings for performance under pressure in skill tasks . 115
8.2.4 Task features of the darts setting 119
8.3 Pressure situations in darts 120
8.4 Empirical analysis 123
8.4.1 Descriptive statistics 124
8.4.2 Modelling checkout performance 126
8.5 Discussion 129

9 Predicting play calls in the National Football League using hidden Markov models 133
9.1 Introduction 133
9.2 Data 134
9.3 Modelling and forecasting play calls 137
9.4 Results 139
9.5 Discussion 140

10 Summary and outlook 143
10.1 Betting fraud detection 143
10.2 Evaluation of in-game performance 144

Appendices 146
A Further betting-related information 147
A.1 Details on Betfair 147
A.2 Betting volumes per team (Chapter 2) 148

A.3 List of fixed matches (Chapter 4) . . . 148
B Additional details on Chapter 4 . . . 148
B.1 Gradient boosting GAMLSS . . . 148
B.2 Classification results for cut-off values chosen via the PPV . . . 151
C Additional results for Chapter 7 . . . 153
C.1 Coefficients in the model for Borussia Dortmund . . . 153
C.2 Additional analysis of Hannover 96 data . . . 153
D Additional results for Chapter 8 . . . 154

Bibliography **159**

Acknowledgements

I wish to express my deepest gratitude to my supervisors, Prof. Roland Langrock and Prof. Christian Deutscher, for their guidance through each stage of my dissertation. My thanks to them have to include even the time *before* I started my dissertation, as especially Prof. Roland Langrock encouraged me a lot to do a PhD. I am deeply grateful for him convincing me that this is the right decision. With the start of my dissertation, I have encountered a new discipline, and I am indebted to Prof. Christian Deutscher for enabling this opportunity and for his trust in me. I was very lucky in having a team of two supervisors, who both supported me in every phase. As both come from different disciplines, I extremely benefited from their diverse inputs. I always had the feeling that I could pick the best of "both worlds".

I would further like to show my gratitude to all my inspiring co-authors. I am more than thankful to Prof. Bernd Frick, Dr. Sebastian Gehrmann, Prof. Andreas Groll, Dr. Vianey Leos-Barajas, Prof. Antonello Maruotti, Dr. Sandra Schneemann, and Hendrik Scholten for their helpful and critical comments through each stage of our respective joint projects. In addition, I would like to thank Prof. Christiane Fuchs for her willingness to act as examiner. I also wish to thank all participants of the ZeSt young researchers workshops at Bielefeld University for the inspiring exchange of ideas in a friendly atmosphere.

In both departments I am affiliated with, I am indebted to my colleagues for making work very enjoyable. The atmosphere in both groups is always supportive and the very opposite of dog-eat-dog. At the time of writing, we were all working from home due to Covid-19, and especially at those days I realised once more how motivating the work together with my colleagues in the office is. Aside from these words of thanks, I would like to give a special thanks to the "Bielefeld kids" Jenny, Sina, and Timo, for the amazing conference trips we had together, especially to the IWSM.

I also owe a more personal thanks to my parents for their unstinting help over the years. Without their support, I would have never gone to university. Anna, thank you for your everlasting encouragement and understanding, even though hours were often long, especially in the final phase.

Statement of contribution

This thesis consists of an introductory part and eight scientific papers. Chapters 2, 3, 4, 5, and 8 are published in peer reviewed journals. Chapters 6, 7, and 9 are currently under revision in peer reviewed journals and additionally available on *arXiv*. Chapter 1 provides an overview of the thesis and a brief summary of Chapters 2 – 9. Seven out of eight papers (Chapters 2 – 8) were written in cooperation with co-authors. The contributions of the authors to the respective papers are listed below. Chapter by chapter, these are:

Chapter 2

Christian Deutscher, Marius Ötting, Sandra Schneemann, Hendrik Scholten (2019): The demand for English Premier League soccer betting. *Journal of Sports Economics*, 20(4), 556–579.

Christian Deutscher initiated the project and conceived the structure of the paper. Hendrik Scholten investigated the literature. Marius Ötting was responsible for organising and cleaning the data, for the exploratory data analysis, and for fitting the GAMLSS models. Sandra Schneemann was responsible for fitting the linear models. The manuscript was written and revised in close cooperation with all co-authors.

Chapter 3

Christian Deutscher, Bernd Frick, Marius Ötting (2019): Betting market inefficiencies are short-lived in German professional football. *Applied Economics*, 50(30), 3240–3246.

Christian Deutscher and Bernd Frick initiated the project and conceived the structure of the paper. Christian Deutscher and Marius Ötting investigated the literature. Marius Ötting was responsible for organising and cleaning the data, and for the data analysis. The manuscript was written by Christian Deutscher and Marius Ötting. All co-authors have contributed to revising the manuscript.

Chapter 4

Marius Ötting, Roland Langrock, Christian Deutscher (2018): Integrating multiple data sources in match-fixing warning systems. *Statistical Modelling*, 18(5-6), 483–504.

Christian Deutscher and Marius Ötting initiated the project and conceived the structure of the paper. Marius Ötting implemented the bivariate Poisson distribution within the R package `gamboostLSS` and performed the data analysis. Marius Ötting wrote the manuscript. Roland Langrock and Christian Deutscher supported the development by discussions on the modelling framework and on betting markets, respectively. Both of them contributed to revising the manuscript. Preliminary work on Chapter 4 was published in the proceedings of the 32nd International Workshop on Statistical Modelling (see *Ötting et al.*, 2017).

Chapter 5

Marius Ötting, Roland Langrock, Christian Deutscher, Vianey Leos-Barajas (2020): The hot hand in professional darts. *Journal of the Royal Statistical Society (Series A)*, 183(2), 565–580.

Christian Deutscher and Roland Langrock initiated the project and conceived the structure of the paper. Marius Ötting implemented the likelihood functions for the state-space models and performed the data analysis. The manuscript was written and revised by Marius Ötting, Roland Langrock, and Christian Deutscher. Vianey Leos-Barajas supported the development by discussions on the modelling framework and by revising the manuscript. Preliminary work on Chapter 5 was published in the proceedings of the 33rd International Workshop on Statistical Modelling (see *Ötting et al.*, 2018a).

Chapter 6

Marius Ötting, Andreas Groll: A regularised hidden Markov model for analysing the 'hot shoe' in football. *arXiv preprint*, arXiv:1911.08138, invitation to revise and resubmit at *Statistical Modelling*.

Andreas Groll initiated the project and conceived the structure of the paper. Marius

Ötting implemented the likelihood functions and performed the data analysis. Andreas Groll supported the development by discussions on the modelling framework, especially on the LASSO, and contributed to the corresponding section in the manuscript. The manuscript was written by Marius Ötting, and Andreas Groll contributed to revising the manuscript.

Chapter 7

Marius Ötting, Roland Langrock, Antonello Maruotti: A copula-based multivariate hidden Markov model for modelling momentum in football. *arXiv preprint*, arXiv:2002.01193, invitation to revise and resubmit at *AStA Advances in Statistical Analysis*.

The project was initiated by Marius Ötting. Marius Ötting implemented the likelihood functions, performed the data analysis, and wrote the manuscript. Roland Langrock and Antonello Maruotti both supported the development by discussions on the modelling framework and contributed to revising the manuscript. Preliminary work on Chapter 7 was published in the proceedings of the 34$^{\text{th}}$ International Workshop on Statistical Modelling (see *Ötting et al.*, 2019).

Chapter 8

Marius Ötting, Christian Deutscher, Sandra Schneemann, Roland Langrock, Sebastian Gehrmann, Hendrik Scholten (2020): Performance under pressure in skill tasks: An analysis of professional darts. *PLOS ONE*, 15(2), e0228870.

Christian Deutscher initiated and conceived the project. Sandra Schneemann, Christian Deutscher, and Hendrik Scholten investigated the literature and the corresponding economic theories. Sebastian Gehrmann was responsible for cleaning the data. Marius Ötting performed the data analysis. The manuscript was written in cooperation with all co-authors, and Roland Langrock contributed to revising the manuscript.

Chapter 9

Marius Ötting: Predicting play calls in the National Football League using hidden Markov models. *arXiv preprint*, arXiv:2003.10791, invitation to revise and resubmit at *Journal of Management Mathematics*.

1 Introduction

In the past decades, the ever-increasing amount of data has undoubtedly revolutionised empirical research. For example, in medicine, gene expression data allow scientists to acquire knowledge about the behaviour of cancer cells (*Sørlie et al.*, 2001); in ecology, movement data help to study animals in the wild (*Gurarie et al.*, 2016); and in marketing, social media data provide insights into consumer behaviour (*Erevelles et al.*, 2016). These are some examples to illustrate how new types of data, and the increase in their magnitude, enable scientists and practitioners in several disciplines to extend their knowledge. However, vast amounts of raw data alone do not provide cutting-edge insights. For that purpose, statistical methods enable to draw conclusions from large data, for example when trying to classify breast carcinomas based on thousands of gene expression patterns (*Sørlie et al.*, 2001). Statistical tools thus help to acquire knowledge from ever-growing data sets. This applies to different disciplines, and sports is no exception.

In sports, the combination of abundant data and statistical tools enables new insights in a variety of fields. Large data sets in sports often cover summary statistics on the performance of teams sampled every minute or even more frequently, such as the number of shots on goal and the running distance in football. These types of data allow managers, for example, to analyse drivers of injuries (*Rossi et al.*, 2018), to improve their scouting (*Barron et al.*, 2018), and to investigate opponents' strategies (*Diquigiovanni and Scarpa*, 2018). Aside from football, the analysis of strategies and tactics has been investigated in different sports, such as in basketball (*Franks et al.*, 2015), but also in individual sports such as marathon running (*Bartolucci and Murphy*, 2015). In addition to the analysis of teams' strategies, several studies focussed on modelling and predicting outcomes of single matches (see, e.g., *Karlis and Ntzoufras*, 2003; *Koopman and Lit*, 2015) and outcomes of international tournaments such as world cups (*Groll et al.*, 2015). Moreover, abundant sports data do not only allow managers to deepen their insights into (e.g.) opponents' strategies and strength, but

also enable to create novel measures thereof, for example in basketball (*Cervone et al.*, 2016). Finally, large data of different sports combined provide cross-sport comparisons of teams and leagues, which is potentially of great interest for teams, managers, and fans (*Lopez et al.*, 2018).

Although teams and managers benefit from exponentially growing data in sports, the use of such data is not restricted to them. The insights provided in studies as summarised above are often more general, as they reveal insights into human behaviour, especially on decision making. The study of decision making remains an ongoing area of research in different disciplines, such as economics, psychology, and sociology. Specifically, in these disciplines, scientists are interested in how humans form decisions, and whether decisions are rational and unbiased. To tackle these questions, sports data offer great opportunities, as athletes' behaviour and decisions are relatively easy to quantify, frequently sampled, and incentives of teams and athletes are well understood (*Kahn*, 2000). The subsequent paragraph briefly summarises previous studies investigating human decision making by analysing data from sports.

In economics, most models assume that humans maximise their utility. The corresponding assumption for firms is to maximise their expected profit. Since data on decisions and strategies of companies are not available, *Romer* (2006) considered data on plays in the National Football League to test the assumption of profit maximisation. Moreover, as there are often interactions between players in sports, those interactions provide further settings for investigating decision making. For example, *Brown* (2011) analyses athletes' performance when competing with a superstar. However, not only players are involved in interactions in sports, but also referees, who should make unbiased decisions. Challenging situations for referees, e.g. supportive crowds favouring the home team, allow for the analysis of potential biases in referees' judgements and corresponding causes (see *Dohmen and Sauermann*, 2016, for a review). Further interactions in sports occur between managers. According to economic theory, managers are expected to make unbiased predictions about the future, as markets are assumed to be efficient. This theory is tested empirically by *Massey and Thaler* (2013), who investigate managers' decisions in the National Football League draft. Furthermore, performance statistics of athletes can be paired with data on salaries and prize money, which are freely available for several sports. This renders sports suitable for labour market research. Corresponding research questions include, for example, the investiga-

tion of race discrimination in salaries (see *Kahn*, 2000, for a review). The business side of sports provides further settings for analysing behaviour and decision making. For match scheduling, teams are interested in the behaviour of spectators when analysing the determinants of attendance (see *Villar and Guerrero*, 2009, for a review). Such data-driven insights can aid teams' decision making, e.g. when deciding on the kickoff time of matches. To maximise profits generated by tickets sold, dynamic pricing system allow for varying ticket prices across (e.g.) weekdays and weather conditions. The application of statistical methods can help here to understand how demand is affected if ticket prices vary (see, e.g., *Paul and Weinbach*, 2013a).

All topics briefly summarised above have in common that the combination of large sports data sets and statistical tools allows for insights into human behaviour. However, in recent years, sports data have become not only larger but also more complex. New technologies allow to measure a great number of details within matches, thus leading to high-dimensional data sets which often comprise hundreds of variables. In addition, time series data covering athletes' performance often exhibit state-switching dynamics. The analysis of decision making based on sports data thus requires the careful consideration of these complex structures when formulating statistical models. This thesis contributes to answering several research questions related to the analysis of decision making based on sports data. The magnitude and complexity of such data allows much more detailed inference than was previously possible. To explicitly account for complex structures, the thesis develops several versatile statistical modelling frameworks, which are flexible enough such that they can be applied to various settings and are not limited to specific sports. Through the development of new statistical methods tailored to the specific complex structures to be modelled, the thesis further paves the way for future empirical work in sports.

The problems studied can be grouped into two categories. First, this thesis considers betting market data to better understand bettors' behaviour, which is then used to detect fraud in betting markets. Second, the thesis investigates in-game data to evaluate athletes' performance during matches. These analyses reveal insights into human behaviour, such as the performance in high-pressure situations. The main topics covered in this thesis — betting fraud detection and evaluation of in-game performance — are further explained and motivated in Sections 1.1 and 1.2, respectively.

1.1 Betting fraud detection

Betting markets have grown considerably in the past decade, with the total gross revenue by bookmakers in 2016 being estimated as 30 billion euro (*IRIS*, 2017). With highly liquid betting markets, substantial amounts of money can be won if outcomes of matches are manipulated. In recent years, several match fixing incidents occurred in different sports, which makes match fixing a growing threat to the integrity of sports. Sports with fixed matches include football (*Federbet*, 2015), tennis (*Gunn and Rees*, 2008), and cricket (*Jewell and Reade*, 2014), to name but a few. Since protecting the integrity of sports is of societal relevance, betting fraud detection systems exist for different sports, e.g. for football (offered by Sportradar), and for tennis (offered by the Tennis Integrity Unit). Such fraud detection systems aim at detecting fixed matches in sports in a data-driven way by monitoring odds movements from bookmakers all over the world. However, in the past, existing betting fraud detection systems failed to detect fixed matches ex-ante, rendering the development of reliable fraud detection systems an active area of research. The growing threat of match fixing to the integrity of sports thus increases the demand for reliable fraud detection systems.

Existing fraud detection systems usually focus on odds movement only, thus neglecting the potential additional information offered by betting volumes. In highly liquid betting markets, very high betting volumes are required to observe odds movements. In such cases, fixed matches in large markets are potentially missed by fraud detection systems due to negligible odds movements. **Chapter 2** presents an approach to model betting volumes, considering data obtained from the online betting exchange *Betfair* for the English Premier League. Since such data were not accessible previously, little is known about a suitable modelling framework and thus about drivers of betting volumes. From a statistical point of view, many challenges arise when modelling betting volumes, as there is substantial heteroscedasticity, and some covariates may have non-linear effects. To explicitly account for these complex patterns, we consider generalised additive models for location, scale and shape (GAMLSS) for modelling betting volumes.

Betting fraud detection based on betting volumes and betting odds requires the consideration of market efficiency. As proposed by *Fama* (1970), market efficiency implies that financial markets comprise all information available, leading to the absence of potential strategies to "beat the market" in the long run. Moreover, aside from

the objective of fraud detection, a test of market efficiency in betting markets is much simpler to perform than in stock markets, as bets have a precise deadline after which their value becomes observable. For the case of fraud detection, if betting markets are inefficient, extreme betting volumes may then be driven by bettors who exploit the market inefficiency to make substantial profits. Thus, in the presence of market inefficiencies it may be impossible to disentangle whether high betting volumes arise from market inefficiencies or fraud. Whereas previous studies have investigated entire seasons to detect betting market inefficiencies, it may be that inefficiencies occur only temporary, for example at the beginning of a season as there is only little information available about the teams. To fill the gap in the literature on temporary inefficiencies, **Chapter 3** investigates short-term betting market inefficiencies. Specifically, we analyse the betting market of the German Bundesliga and consider the beginning of a season where the teams' actual strength is difficult to evaluate for bookmakers.

To avoid match fixing, existing literature and fraud detection systems (such as the one by Sportradar described above) primarily focus on the analysis of betting odds provided by bookmakers. In **Chapter 4**, we suggest to make use of both betting volumes *and* odds to identify potential fixed matches, as odds movement is unlikely to observe in highly liquid betting markets. For that purpose, we make use of the approach presented in Chapter 2 — modelling betting volumes using GAMLSS — to identify outliers, and hence potential fixed matches. In addition to the volumes, we derive betting odds by employing a GAMLSS with bivariate Poisson response to model the number of goals scored by both teams. We then flag suspicious matches using both the derived odds and outliers in betting volumes. As a case study for the approach, we analyse the Italian Serie B as in that case there are several matches where it has been proven that they were fixed.

1.2 Evaluation of in-game performance

Humans are evaluated in quite a few situations in everyday life. Example situations include schools where pupils are graded, selection panels where members decide who to hire, and also sports, where managers and fans evaluate the performance of players frequently. As discussed above, a distinction between sports and several other settings (such as selection panels) is that the performance of players in sports can often be measured fairly accurately. In basketball, for example, coaches can consult the propor-

tion of successful free throws, whereas in most industries no simple summary statistics exist on the performance of job applicants. Data on the performance of athletes can thus aid decision making of teams and managers. Such data are in fact beneficial not only to teams and managers. Data on performance allow for insights into human behaviour, as they can reveal pitfalls in performance evaluation in general, such as the tendency of humans to over-interpret streaks of success and failure. As an illustration, consider the following realisations from independent Bernoulli trials, where (for illustration) the ones can be interpreted as 'success' and the zeros as 'failure':

$$1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 1\ 1\ 1\ 1\ 1\ 0\ 1\ 0\ 1\ 0.$$

As there are two streaks in this sequence with five successes in a row, humans might (falsely) interpret these streaks as evidence for success being more likely followed by another success than by failure. Previous studies have indeed provided evidence that humans tend to misinterpret randomness by over-interpreting streaks of success and failure in such sequences (see, e.g., *Bar-Hillel and Wagenaar*, 1991). The over-interpretation of streaks in such sequences has even been used as a primary example in behavioural economics and psychology for how humans form beliefs and expectations (*Kahneman*, 2011; *Tversky and Kahneman*, 1971, 1974). Since — aside from settings like casino games — the data-generating process is usually not known in practice, proper statistical methods are needed to infer whether those sequences as shown above are realisations of an i.i.d. process or whether there indeed exists serial correlation in performance. Inference on the existence of such streaks can have important consequences on decision making when evaluating performance (*Miller and Sanjurjo*, 2018).

In human performance, and in particular in sports, the concept of the "hot hand" refers to the idea that humans indeed show serial correlation in their performance. Research and debates on the existence of the hot hand started with the seminal paper by *Gilovich et al.* (1985), who analysed free throws in basketball and found no evidence for a hot hand effect. Since then, the hot hand has often been labelled a cognitive illusion. For the analysis of a hot hand effect, several previous studies considered hidden Markov models (HMMs), where the underlying state process serves for a player's form. However, existing studies exhibit some caveats, as they often consider settings with interactions between opponents, thus rendering the analysis of a hot hand effect difficult. For example, when analysing the hot hand of batters in baseball, the

performance of the pitcher is also important but difficult to account for. Aside from caveats of the data, most studies which consider HMMs select two or three states, which may be too coarse for modelling a player's form. In **Chapter 5**, we aim to overcome these caveats by considering a setting without any interaction between opponents, namely professional darts. For the modelling framework, we consider HMMs with a continuous-valued state process to allow for gradual changes in a player's form.

Whereas sports with no interactions between opponents — and hence with at most a few covariates — seem most suitable for the analysis of a hot hand effect, there are also settings where many covariates affect the outcome. When facing such situations, suitable variable selection procedures guide the choice of covariates. For standard regression models, several approaches exist for variable selection, such as the inclusion of covariates based on p-values. Modern regularisation methods such as the LASSO (*Tibshirani*, 1996) and boosting (*Friedman*, 2001) even allow for automated variable selection. Whereas such regularisation methods exist for regression models, the standard approach for variable selection in HMMs is to consult information criteria such as AIC and BIC (*Zucchini et al.*, 2016). **Chapter 6** presents a regularisation approach in HMMs by considering the LASSO, thus allowing for implicit variable selection. To illustrate the usefulness of this method, as a case study we investigate a potential hot hand effect of penalty-takers in football with about 650 covariates in total.

Concepts similar to the hot hand are "momentum" and "momentum shifts". These terms are frequently used by sports commentators and fans in situations where an event — such as a shot hitting the woodwork in a football match — seems to change the dynamics of the match, e.g. in a sense that the supposed underdog suddenly seems to dominate the match. As introduced above, research on the hot hand revealed that humans tend to misinterpret randomness, such that it is to be expected that perceived momentum shifts to some extent are cognitive illusions. **Chapter 7** covers an analysis of potential momentum shifts within football matches. Specifically, we consider HMMs to model minute-by-minute in-game statistics that are potentially subjected to switches in underlying states. Within these HMMs, we formulate multivariate state-dependent distributions using copulas to fully address the given data structure.

A further topic related to performance evaluation is the effect of pressure on human performance. Understanding how humans cope with pressure situations is relevant in various areas of society, such as disaster management, workplace management, and sports. The effect of pressure on human performance is thus a further research question

relevant in different disciplines, and easiest to investigate in sports, as performance in other settings (such as disaster management) is difficult to quantify. However, previous studies on the performance under pressure provided mixed results, potentially as a consequence of neglecting interaction effects between players. **Chapter 8** presents an analysis of pressure on human performance in professional darts. As for the analysis of the hot hand in Chapter 5, professional darts provides a near-ideal setting with no direct interaction between players. For the analysis, we consider generalised linear mixed models to account for the longitudinal data structure.

Driven by the increasing amount of data, teams in different sports are interested in analysing data to investigate opponent teams' strategies. There is a long history of analysing opponents to gain an advantage on the field. Baseball was the first sport where data on performance of players and teams were tracked and analysed. However, nowadays sports data cover not only summary statistics based on complete matches, but also detailed data on in-game dynamics. **Chapter 9** covers the analysis of a comprehensive play-by-play data set of the National Football League (NFL), focusing on the prediction of plays. As previous studies do not account for the time series structure of such data, Chapter 9 considers HMMs for the prediction of plays, thus exploiting the data strucuture at hand.

2 The demand for English Premier League football betting

2.1 Introduction

Sports betting is a rapidly growing market with worldwide turnover of 58 billion euro in 2015 as estimated by the European Gaming and Betting Association. This value exceeds the gross domestic product of countries like Panama and Costa Rica, and such official numbers ignore illegal betting. Although the market size is huge and growing, surprisingly little is known about the demand for sports betting, as existing studies mostly focus on inefficiencies of betting markets or match fixing. One reason for this research gap is connected to the unavailability of corresponding data sets.

While access to betting volume data was not given until now, such data helps to acquire knowledge about betting markets, which is of crucial importance for various reasons. First, to deepen the insights into the economic impact of single sports events, it is crucial to include sports betting into corresponding analyses. So far, existing analyses mostly focus on ticket sales, TV contracts, and merchandise, thus neglecting the economic value of sports betting (*Roberts et al.*, 2016). The data considered in this chapter suggest that the economic importance of sports betting is potentially superior with respect to total revenue. The economic importance of sports competition, hence, tends to be underestimated to date. Second, even though revenue of betting exchanges or bookmakers does not benefit the clubs directly, they profit from increased betting indirectly, e.g. with respect to (shirt) sponsorship engagements of betting companies. Several bookmakers act as a shirt sponsor of a Premier League club. In the season 2016/17, ten out of twenty teams had a shirt sponsorship contract with a company from the gambling industry. Thus, the rise of the betting industry generates an important economic effect on the teams in the Premier League, leading to millions of additional pounds earned through sponsorship deals.

Focusing on sports betting puts economic importance of matches into perspective

and allows for comparisons between teams within a certain league as well as across leagues. Furthermore, analyses of the demand for bets can also improve the understanding of demand for sports contests in general. Even though there is an ever-growing literature on drivers for attendance[1], some ambiguities remain, especially with respect to the impact of outcome uncertainty.[2] The ambiguities with respect to demand for sporting contests and the role of uncertainty of outcome may be related to empirical issues, such as censored data due to sellouts of stadiums, unavailable data on TV or online stream audience, and problems with respect to the inclusion of ticket prices. Data on betting demand reduces these issues as demand is not censored, and exact numbers on demand and prices are available. Corresponding results can thus provide valuable information on determinants of demand for sports events in general.

To the best of our knowledge, the studies by *Humphreys et al.* (2013) and *Paul and Weinbach* (2010) are rare examples that analyse demand for bets. While *Paul and Weinbach* (2010) investigate the number of bets placed on NBA and NHL games, the study by *Humphreys et al.* (2013) is the only that parses volumes of dollars bet. Both studies find that determinants of demand for bets are comparable to the determinants that affect fan behaviour, such as teams' quality, TV coverage, time of day, and outcome uncertainty. Given the scarce literature on the determinants of betting volumes in relation to a fast-growing sports betting market, this chapter closes several gaps in the literature. First, this is the first analysis focussing on any European market. While US betting is dominated by point spread betting, European betting refers mainly to betting on the result (home win, draw, away win). In terms of betting odds, European betting can exhibit heavy underdogs and favourites, whereas in US betting bookmakers set the spread such that both teams usually have the same odds. Second, this is the first analysis of football, which is the premier European sport with the English Premier League being the economically most relevant league. Third, to the best of our knowledge, this is the first research on betting volumes considering data from a betting exchange. Forth, it is the first analysis of a high-turnover betting platform. On average, bettors wager 2.7 million pounds per Premier League match.[3] Our results

[1] For an overview on determinants of demand for sporting contests see e.g. *Borland and MacDonald* (2003) or *Villar and Guerrero* (2009).

[2] There is extensive literature on the concept of uncertainty of outcome or competitive balance of sporting contests. Although the underlying idea of competitive balance seems straightforward, its effect on fan interest remains unclear, as existing empirical studies provide mixed results (see e.g. *Benz et al.*, 2009; *Borland and MacDonald*, 2003; *Forrest et al.*, 2005; *Szymanski*, 2003).

[3] By comparison: the data considered by *Humphreys et al.* (2013) show a mean of $20,584 per NCAA basketball match.

confirm that the strength of participating teams is a major determinant for the demand of bets. In addition, the day of the week, economic factors, and the uncertainty of outcome affect demand significantly.

This chapter is organised as follows: Section 2.2 provides a literature review on sports betting in general. Section 2.3 covers the empirical analysis on drivers of betting volumes in the English Premier League.

2.2 Literature review

Sports betting has drawn huge academic interest in the past and still does today. Major research interests refer to the way bookmakers set their prices, whether odds are biased, and — as of late — what affects demand for bets. Part of academic research on determinants of betting behaviour deals with general motives of bettors. While early studies generally describe bettors as investors whose single interest is to maximise profits, *Samuelson* (1952) is the first to challenge this assumption. Later, *Conlisk* (1993) shows that, in theory, gambling provides utility in itself. The importance of consumption motives are empirically shown by *Paul and Weinbach* (2013b).

More detailed research on the determinants of betting behaviour has long time been limited due to little information available. *Gramm et al.* (2007) investigate drivers of betting volumes in horse racing. They find that the day of the week is important and that the strength of the teams raises betting volumes. *Paul and Weinbach* (2010) analyse factors influencing betting on NBA and NHL games during the 2008/2009 regular season, considering the number of bets placed as response variable. They find that betting behaviour is similar to fan behaviour as the strength of the teams is found to be positively related to the number of bets, and that games which are thought of as more high-scoring attract a higher number of bets. In addition, bettors are further seen to dislike heavily lopsided contests as opposed to televised games, which attract more betting action. *Paul and Weinbach* (2013b) find similar results. In their study of the NBA seasons from 2004/05 to 2006/07 they confirm the positive effects of TV coverage and of teams' strength, respectively. They further find an increased number of bets for weekend games and for games where the most popular teams are participating. *Humphreys et al.* (2013) are the first (and to date only) to analyse the amount of dollars. They investigate determinants of betting volumes of NCAA men's basketball games. Similar to previous research, they find that bettors' behaviour

resemble fan behaviour as the strength of the teams, the uncertainty of outcome, and TV coverage affect betting volumes.

2.3 Empirical analysis

Our data provide betting information on all matches of the English Premier League from seasons 2009/10 to 2015/16. The data was taken from the online betting exchange betfair.com. Betfair is the world's leading sports betting exchange with about five million customers worldwide.[4] Their revenue exceeded £1,500 million in 2016.[5]

Customers at Betfair have two alternatives to bet. They can either bet on an event occurring (backing the event) or an event not occurring (laying the event).[6] Traditional betting refers to the former alternative of backing events. Betfair itself both acts as a traditional bookmaker by setting fixed betting odds and as a betting exchange. At the betting exchange, they match back and lay bets at stated odds and withhold a certain percentage of profits to generate earnings. A bet comes off only in case that a back (lay) bet can be matched to at least one lay (back) bet. The following analysis considers bets placed at the exchange market as information on "traditional" bets placed on betfair.com is unavailable. The corresponding business model differs from that of traditional bookmakers by mainly working as an intermediary between bettors. In contrast to traditional bookmakers, the Betfair exchange platform realises profits independent from the game outcome since they do not wager their own money but charge a commission on winning bets. As they do not take any risk, their commission can be lower compared to other betting platforms, resulting in potentially favourable odds to bettors. Since there is no risk in being the intermediary, betting exchanges do not limit individual wagers.

Customers can either offer bets at individual odds (when placing lay bets) or choose between different odds (when placing back bets). Odds offered at a particular time are identical to all bettors in the market and are fixed once a bet is matched. However — as discussed above — a bet comes off only if another bettor lays or backs, respectively,

[4] The (unrestricted) use of Betfair is not allowed in every country due to national gambling laws or taxes. Betfair is currently legal in 47 countries (several countries have different jurisdiction for different states) and illegal in 30 countries. In all remaining countries placing bets at Betfair is a grey area, since it is neither explicitly illegal nor legal.

[5] https://www.paddypowerbetfair.com/~/media/Files/P/Paddy-Power-Betfair/documents/annual-report-2016.pdf

[6] Appendix A.1 provides further explanations on back and lay bets.

the odds offered. Our data cover the sum of wagers and the number of *matched* bets per match for 2,660 Premier League matches (380 matches per season). However, we restrict our analysis to bets that are placed before the start of a match since in-game betting volumes are changing in accordance to game dynamics which are difficult to cover as time stamps for all bets would be required.

Table 2.1: Matched volume broken down to different betting types.

betting type	empirical proportion
match outcome	0.62
over/under 2.5 goals	0.10
correct score	0.06
over/under 1.5 goals	0.02
over/under 3.5 goals	0.02
half time/full time	0.02
Asian handicap	0.02
over/under 4.5 goals	0.02
over/under 0.5 goals	0.01
other	0.10

More than 200 different types of pre-game bets exist for the majority of matches, e.g. bets on the result (home win, draw, away win), the correct score, the number of goals, or a combined bet of the winner of the first half and the winner of the match, to name but a few. Whereas North American sports betting mostly offers point spread betting, our data include a variety of different betting types. Point spread betting in European football betting markets is typically called "(Asian) Handicap" betting, which is also included in our data. Table 2.1 summarises the most prominent types of bets found in the data, with only two percent of the matched volume referring to "Asian Handicap" bets. However, about 78% of pre-game bets in our sample are placed on the following types of bets: match outcome (home win, draw, away win), over/under 2.5 goals, and the correct score.

We consider the number of *matched* (back and lay) bets (*numberbets*) and the amount of British pounds bet (*poundsbet*) per match (added up over all betting types) as potential response variables. Figure 2.1 shows the amount of pounds bet and the number of matched bets, respectively. Both empirical distributions are similarly skewed to the right. Hence, we use the logarithm of both variables in the subsequent regressions. As both the volume of pounds bet and the number of matched bets reflect demand for bets, we assume that these two variables are highly correlated. Figure 2.2 confirms this assumption. The corresponding correlation coefficient is 0.9. Bettors place on average more than 32,500 bets per game, corresponding to an average

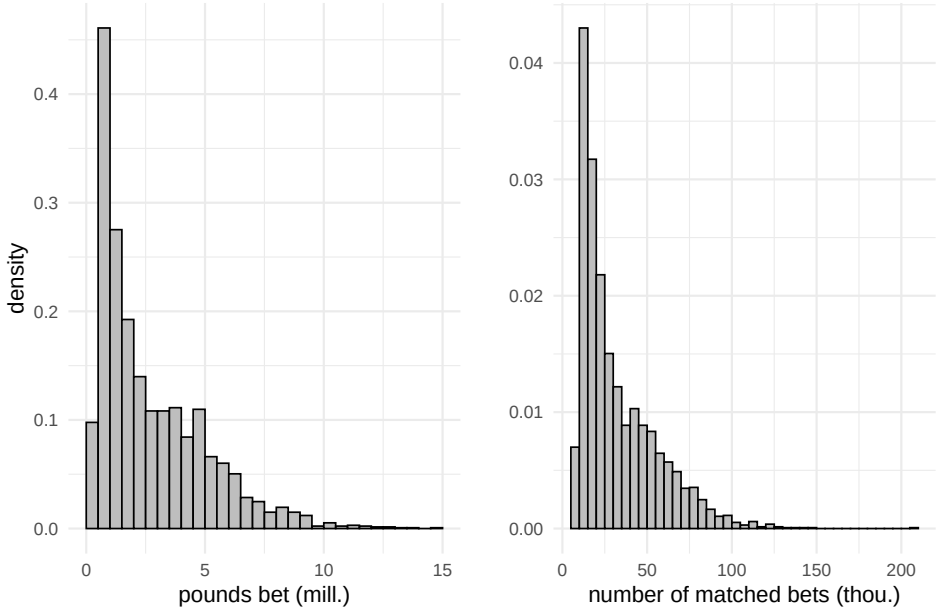


Figure 2.1: Histograms of pounds bet (left panel) and number of matched bets (right panel).

volume of more than 2.7 million pounds and an average wager of about 78 pounds per bet[7], thus indicating a huge economic importance of sports betting.[8]

Similar to *Paul and Weinbach* (2010) and *Humphreys et al.* (2013), we fit two models with the number of matched bets (*numberbets*) and the amount of money bet (*poundsbet*) as response variables, respectively. To analyse and comprehend bettors' behaviour, we focus on covariates which have been suggested to affect demand for bets. In both models, we include covariates reflecting the strength of the home and away team, respectively. In addition, we account for temporal factors, such as the season, the day of the week, and the matchday.

Figure 2.3 shows the empirical proportions of matches taken place on the different weekdays. Most matches (56.5%) take place on Saturdays. Whereas both *Paul and Weinbach* (2010) and *Humphreys et al.* (2013) consider information on the month, we use the matchday instead. This enables us to model the seasonal dynamics in more detail. We assume that demand for bets is higher for matches played at the beginning and at the end of a season. Therefore, we also include the square of the matchday in our model.

Over the years, gambling laws and restrictions have changed in several countries.

[7] Compared to *Humphreys et al.* (2013), this value seems fairly high. A potential explanation may arise from the fact that Betfair sets no upper limit with respective to wagers per bet while traditional bookmakers cap bets.

[8] As discussed in the introduction, the gambling industry with its worldwide turnover of 58 billion euro (2015) is of huge economic importance. To get an idea of how much money is placed at Betfair, Table A1 in Appendix A.2 displays the average betting volumes per team. Bettors place on average over £5 million on matches at which Manchester United participates. Even the club with the lowest average betting volume in our sample, namely Cardiff City, attracts on average more than £1.6 million every match.

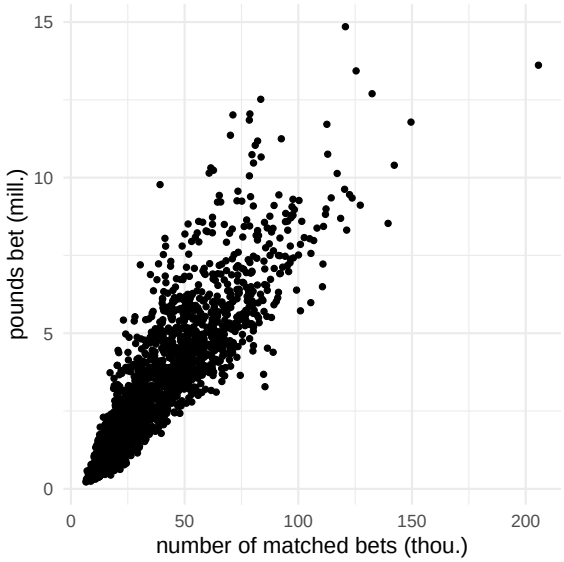


Figure 2.2: Scatter plot of pounds bet and number of matched bets.

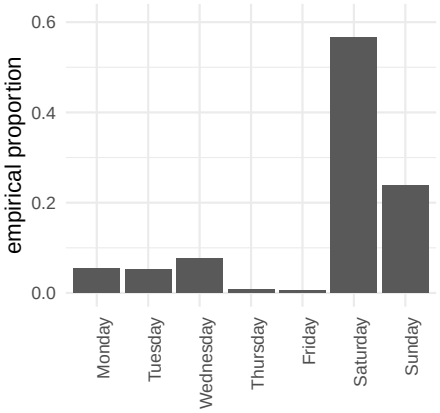


Figure 2.3: Empirical proportions of matches taken place through the week.

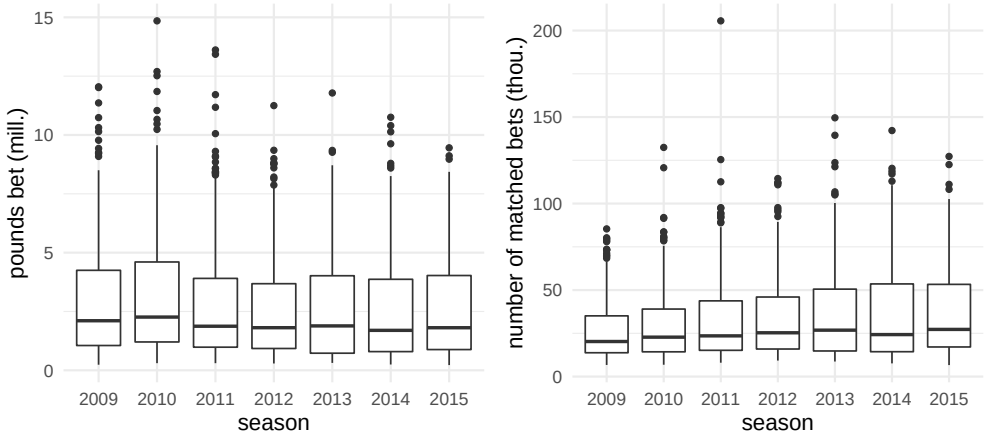


Figure 2.4: Boxplots for the betting volume (left panel) and the average number of bets (right panel) per season.

The number of potential bettors are thus subject to fluctuations over time. Furthermore, demand for bets may be affected by a (global) time trend. Figure 2.4 shows that the median number of matched bets increased from 2009 to 2015, whereas the median of the total amount of money bet decreased. The average wager decreased from 102 pounds in 2009 to 64 pounds in 2015. We therefore include dummy variables for the different seasons in our analysis.

We further consider a match's uncertainty of outcome, which may have two opposing effects on betting action. On the one hand, (un)certainty reflects the financial attractiveness of a bet. The higher the perceived certainty of a match outcome, the more likely the bet on this outcome will succeed. Hence, bettors may choose to place a bet on this outcome. On the other hand, uncertainty may positively affect the decision to place a bet, because these matches are perceived as more "exciting". We include the uncertainty of the match outcome (*certainty*) by using betting odds. Since bettors who place bets at Betfair can choose between different odds, it seems inappropriate to use Betfair odds. Hence, we make use of another popular bookmaker, namely bwin. We measure the match uncertainty as the absolute difference between the winning probabilities of the two teams, which we derive from their respective betting odds.

Another potential determinant of demand for sporting contests refers to the quality. Better teams attract on average more fans than teams of lower quality (see e.g. *Borland and MacDonald*, 2003). As the results provided by *Paul and Weinbach* (2010) and *Humphreys et al.* (2013) suggest that the same is true for demand for sports betting, we include the sum of teams' ranks before the match[9] and the market values of both teams

[9] As a team's standing prior to the first match is not available, in those cases we refer to the teams' final standing in the previous season. The rank of promoted teams are calculated by the teams' final ranks in the Championship plus 20 (since 20 teams participate in the Premier League).

as covariates in the regression. A further covariate affecting the demand for sports betting refers to TV transmissions. The studies mentioned above find a significant and positive effect of TV coverage on betting volumes. However, we omit this variable since Betfair customers — on the contrary to customers of betting platforms analysed in the two studies discussed above — are residents from different countries. Since Premier League matches are partly broadcast in these countries and we are not able to connect betting volumes to countries, it is not possible to analyse the effect of TV coverage on betting volumes for the data at hand. In addition, the availability of online streams could blur the effect of live TV coverage. The internet offers (often illegal) streams for all games covered in our data set. Hence, the effect of TV coverage and online streaming is not investigated.

The availability of substitutes is a further relevant determinant of demand. A potential substitute for placing bets on a certain match is placing bets on other matches. Even though it is possible to bet on future matches, we focus on the number of matches taking place in the Premier League at the same time (*substitutes*) and expect a negative effect of that covariate on betting action.

The decision to place a bet may also be affected by a team's status as being promoted to the Premier League for various reasons. First, promoted teams are often of lower strength than teams that have played in the first division for several seasons. This may reduce the interest in matches with promoted teams. Second, it is more challenging to predict the winning probabilities of promoted teams since no or only few observations are available for matches of these teams against other Premier League teams. Bettors, just like bookmakers, may therefore under- or overestimate winning probabilities of promoted teams which may result in financially more attractive bets (*Deutscher et al.*, 2018). Third, fans of recently promoted teams might be euphoric about the promotion of "their" team, which potentially results in higher betting volumes. Since the data at hand does not allow to distinguish between these potential explanations empirically we simply include dummy variables indicating whether one or both teams were promoted to the Premier League prior to the season (*onepromoted* and *bothpromoted*).

Table 2.2 shows descriptive statistics for the variables considered in our analysis. The information on the variables was taken from `www.www.worldfootball.net`, `www.transfermarkt.com`, and `www.football-data.co.uk`. On average, about 2.7 million pounds are bet per Premier League match. Minimum values of *poundsbet* and

Table 2.2: Descriptive statistics.

	mean	st. dev.	min.	max.
poundsbet (mill.)	2.723	2.245	0.222	14.85
numberbets (thou.)	32.55	22.82	6.646	205.6
pounds per bet	78.24	29.01	25.24	249.4
matchday	19.50	10.97	1	38
$matchday^2$	500.6	441.1	1	1444
Monday	0.055	–	0	1
Tuesday	0.053	–	0	1
Wednesday	0.076	–	0	1
Thursday	0.008	–	0	1
Friday	0.005	–	0	1
Saturday	0.565	–	0	1
Sunday	0.238	–	0	1
mvrelhome	1	0.709	0.124	2.785
mvrelaway	1	0.709	0.124	2.785
certainty	0.307	0.206	0	0.802
$certainty^2$	0.137	0.151	0	0.643
sumranks	21.39	8.147	3	42
substitutes	2.522	2.424	0	9
onepromoted	0.244		0	1
bothpromoted	0.014		0	1

of *numberbets* refer to the match of Aston Villa against Swansea (season 2015/16). The highest betting volume refers to the match Manchester United against Arsenal FC in 2010/11, whereas most bets were placed on the match Manchester City versus Manchester United in 2011/12. The data include both rather balanced and unbalanced matches. Manchester City against Crystal Palace (2013/14) represents the most unbalanced game in the period observed. The mean market value of a Premier League team is 192.9 million euro (Min: Norwich in 2011/12, Max: Chelsea in 2013/14). As the mean market value increased over the years, we use the relative market value per team and season in the subsequent analysis (*mvrelhome* and *mvrelaway*). Table 2.3 displays the correlation coefficients for the main covariates.

Table 2.3: Correlation matrix of main covariates.

	match.	*mvh*	*mva*	*cert.*	*sumra.*	*subst.*	*onepr.*	*bothpr.*
matchday	1							
mvrelhome	-0.0002	1						
mvrelaway	0.0002	-0.053	1					
certainty	0.017	0.549	-0.075	1				
sumranks	-0.021	-0.459	-0.459	-0.223	1			
substitutes	0.085	-0.145	-0.294	-0.013	0.210	1		
onepromoted	-0.007	-0.199	-0.199	-0.030	0.239	0.090	1	
bothpromoted	0.012	-0.103	-0.103	-0.069	0.106	0.071	-0.064	1

2.3.1 Methodology

The empirical analysis focuses on determinants of betting volumes in sports betting. We consider the number of matched bets (*numberbets*) per match and the sum of wagers (*poundsbet*) placed on a match, respectively, as response variables in equation (2.1). Due to the positive skewness of both response variables (cf. Figure 2.1), they enter the model (2.1) in logarithmic form. Several covariates are included in a linear and a quadratic form to account for potential diminishing/increasing marginal effects. In the first part of the analysis, we present the results of a classical linear model. In the second part (see Section 2.3.3), we use GAMLSS to consider potential non-linear effects of non-categorical covariates. Furthermore, the GAMLSS framework allows to model several parameters of the assumed distribution of the response variable simultaneously instead of considering only the mean (*Rigby and Stasinopoulos*, 2005), which in our case is beneficial due to the presence of heteroscedasticity.

2.3.2 Linear model

Equation (2.1) shows the model formulation including all covariates mentioned above. In the following, y represents either the number of matched bets or pounds bet on a match. For notational simplicity, we omit the indexes corresponding to teams, matchdays, and seasons, leading to the following form of our model:

$$\begin{aligned}\log(y) = \beta_0 &+ \beta_1 matchday + \beta_2 matchday^2 + \beta_3 Monday + \beta_4 Tuesday \\ &+ \beta_5 Wednesday + \beta_6 Thursday + \beta_7 Friday + \beta_8 Saturday \\ &+ \beta_9 mvrelhome + \beta_{10} mvrelhome^2 + \beta_{11} mvrelaway \\ &+ \beta_{12} mvrelaway^2 + \beta_{13} certainty + \beta_{14} certainty^2 \\ &+ \beta_{15} sumranks + \beta_{16} substitutes + \beta_{17} onepromoted \\ &+ \beta_{18} bothpromoted + u\end{aligned} \tag{2.1}$$

Table 2.4 displays the results for the response variables *poundsbet* and *numberbets*, respectively. The R^2 is fairly large for both models, indicating that the chosen variables explain a considerable variation of demand for bets. Since a Breusch-Pagan test rejects the null hypothesis of homoscedasticity, we use heteroscedasticity-consistent standard errors in our analysis. Almost all variables are statistically significant on a 5% level.

Due to the high correlation between *poundsbet* and *numberbets*, the results of the fitted models are fairly similar. However, a single difference remains: whereas there is an increasing marginal effect estimated of *certainty* on *poundsbet*, the effect on

Table 2.4: Regression results.

	response variable:	
	$\log(poundsbet)$	$\log(numberbets)$
matchday	0.012	0.009
	[0.006; 0.018]	[0.004; 0.013]
matchday^2	-0.0004	-0.0002
	[-0.001; -0.0003]	[-0.0003; -0.0001]
Monday	0.483	0.414
	[0.407; 0.559]	[0.362; 0.465]
Tuesday	0.022	0.022
	[-0.056; 0.100]	[-0.031; 0.075]
Wednesday	-0.197	-0.158
	[-0.265; -0.129]	[-0.205; -0.112]
Thursday	0.168	0.087
	[-0.014; 0.350]	[-0.036; 0.211]
Friday	0.252	0.165
	[0.019; 0.485]	[0.007; 0.323]
Saturday	-0.194	-0.202
	[-0.237; -0.151]	[-0.231; -0.173]
mvrelhome	0.479	0.455
	[0.362; 0.596]	[0.376; 0.535]
mvrelhome^2	-0.068	-0.078
	[-0.108; -0.028]	[-0.105; -0.051]
mvrelaway	0.746	0.567
	[0.628; 0.865]	[0.487; 0.647]
mvrelaway^2	-0.133	-0.105
	[-0.173; -0.093]	[-0.132; -0.077]
certainty	0.377	0.227
	[0.085; 0.669]	[0.029; 0.425]
certainty^2	0.728	0.099
	[0.305; 1.151]	[-0.188; 0.386]
sumranks	-0.009	-0.008
	[-0.011; -0.006]	[-0.010; -0.006]
substitutes	-0.166	-0.130
	[-0.173; -0.158]	[-0.135; -0.125]
onepromoted	0.118	0.076
	[0.075; 0.162]	[0.046; 0.105]
bothpromoted	0.205	0.152
	[0.062; 0.347]	[0.055; 0.249]
constant	0.344	9.782
	[0.185; 0.503]	[9.674; 9.890]
season dummy variables	yes	yes
observations	2,660	2,660
R^2	0.770	0.814

Note: 95% CIs are shown in brackets.

numberbets is rather linear. However, both variables are positively affected by an increased *certainty* of outcome, i.e. lopsided matches tend to increase the betting volume and the number of bets placed. This contradicts the results of *Paul and Weinbach* (2010) and *Humphreys et al.* (2013) who find that bettors tend to prefer uncertain over lopsided matches. The contradictory result of NCAA bets (as provided by *Humphreys et al.*, 2013) may be explained by the number of "points" scored per game. Football represents a "low-scoring" sports, i.e. a single goal can decide a match. In a basketball match, the favourite has far more possibilities to score. Therefore, small differences in ex-ante winning probabilities may be more decisive in basketball than in football matches.

Table 2.4 also shows that *matchday* has a non-linear (inverted U-shaped) effect on the response variables. The turning point for *poundsbet* is 15, while it is estimated as 21.5 for *numberbets*. Potential explanations for these results are discussed in Section 2.3.3. With respect to the weekdays, our results suggest that highest demand exists on average for matches played on Fridays and lowest demand for matches played on Wednesdays and Saturdays.[10] The negative coefficient for matches played on Saturdays is potentially caused by the fairly high number of substitutes at that time, i.e. matches played in other popular football leagues. According to our results, the higher the number of substitutes, the lower is, on average, the demand. Accordingly, the demand for matches played on a Monday is fairly high, as there are only few matches played in other European football leagues on Mondays. In addition, the quality of the match significantly affects the demand for bets. The better both teams are ranked in the standings prior to a match (*sumranks*) and the higher the relative market value of the home (*mvrelhome*) and away team (*mvrelaway*), the higher is on average the demand for bets. In addition, the betting volume is (significantly) higher on average if promoted teams participate (see Table 2.4).

2.3.3 Flexible approach

Our models include several covariates in both linear and quadratic form (see Eq. 2.1). To investigate whether effects of a higher order might be appropriate, we consider GAMLSS, which allows for smooth functional effects of non-categorical covariates (*Rigby and Stasinopoulos*, 2005). Estimating smooth functional effects enables us to

[10] The reference category is *Sunday*.

examine the relationship between response and covariates in more detail. Furthermore, the GAMLSS framework allows to simultaneously model several parameters of the distribution of the response variable (e.g. mean and variance) instead of modelling only the mean as in a classical linear regression model. Due to our heteroscedastic data, we can potentially improve the model fit by additionally modelling the standard deviation of the betting volumes. Specifically, we apply the semi-parametric additive formulation of GAMLSS (see *Rigby and Stasinopoulos*, 2005), which is given by

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k\boldsymbol{\beta}_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{2.2}$$

where $\boldsymbol{\theta}_k$ is a parameter of the distribution assumed for the response variable $\mathbf{Y}$, $g_k(\cdot)$ is a known link function, $\mathbf{X}_k$ is an $n \times J_k'$ design matrix, $\boldsymbol{\beta}_k$ is a vector of regression coefficients of length J_k', and the h_{jk} are unknown smooth functions.

Another advantage of the GAMLSS framework is that several distributions for the response variable are allowed, which do not have to be part of the exponential family. Due to our skewed betting volumes (cf. Figure 2.1) we initially considered several different right-skewed distributions for our response variable $\mathbf{Y}$, with the result that a log-normal distribution fits best to our data. Thus, the covariates are linked to the mean $\boldsymbol{\mu}\,(=\boldsymbol{\theta}_1)$ and the standard deviation $\boldsymbol{\sigma}\,(=\boldsymbol{\theta}_2)$, of the distribution of $\mathbf{Y}$, leading to the following special case of (2.2):

$$\boldsymbol{\mu} = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} h_{j1}(\mathbf{x}_{j1}) \tag{2.3}$$

$$\log(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} h_{j2}(\mathbf{x}_{j2}). \tag{2.4}$$

The covariates described in Section 2.3 enter the model in (2.3) and (2.4) as follows: the dummy variables considered in the linear model (2.1), i.e. dummy variables for the season, the weekday, and the promoted teams are included in the linear parts both in (2.3) and (2.4). We further consider smooth functional effects for the non-categorical covariates, i.e. for the uncertainty of outcome measured by the difference in winning probabilities (*certainty*), for the market values of the home and away team (*mvrelhome*, *mvrelaway*), respectively, for the matchday (*matchday*), and for the sum of the teams' ranks in the league standings (*sumranks*). Thus, both predictors for $\boldsymbol{\mu}$

and $\log(\boldsymbol{\sigma})$ are of the following form[11]:

$$\begin{aligned}\eta = \mathbf{x}'\boldsymbol{\beta} &+ h_1(matchday) + h_2(mvrelhome) + h_3(mvrelaway) \\ &+ h_4(certainty) + h_5(sumranks)\end{aligned} \tag{2.5}$$

The column vector $\mathbf{x}$ contains a one for the intercept as well as the above mentioned and in (2.1) included dummy variables. The smooth functional effects h_m are estimated using P-Splines (*Eilers and Marx*, 1996) with 20 initial knots of equal space. The smoothing parameter is selected by using a local BIC criterion as the number of observations is fairly large (see *Stasinopoulos et al.*, 2017). The model is fitted using the package `gamlss` (*Rigby and Stasinopoulos*, 2005) which is available for the statistical software `R` (*R Core Team*, 2019).

As the estimated smooth functional effects are similar for both response variables, we present only those for the response variable *poundsbet*. The estimated effects of the dummy variables are fairly close to the ones of the linear model (cf. Table 2.4). Figures 2.5 and 2.6 show the smooth functional effects of the fitted model. For the effects on the mean (Figure 2.5), the effect of *sumranks*, *certainty*, and *mvrelhome* are linear functions with negative slopes. Surprisingly, the marginal effect of the (relative) market value for the away team diminishes, whereas the marginal effect for the market value of the home team is linear. For the *matchday*, a more flexible smooth function is estimated. The estimated effect suggests that the effect of *matchday* varies considerably. There is a local maximum at matchday 20, which refers to matches played between December 28 and January 4. Hence, the high demand for bets in that period may be explained by public holidays worldwide in most countries. Demand for sporting contests is positively affected by holidays which tends to be true for sports betting, too. In some countries employees also receive a Christmas bonus at this time of the year which may increase demand for bets as additional money is spendable. The mean of the betting volume decreases after its maximum and increases again at the end of a season. These effects are hard to accommodate within a linear model. The substantial increase in the betting volumes at the end of the season may refer to the fact that at this time there are often matches which are of crucial importance for some of the teams, e.g. with respect to championship, qualification for the European competitions, and relegation. This great importance of some of the matches may positively

[11] We drop the observation-specific indexes here for simplicity.

affect demand for bets.

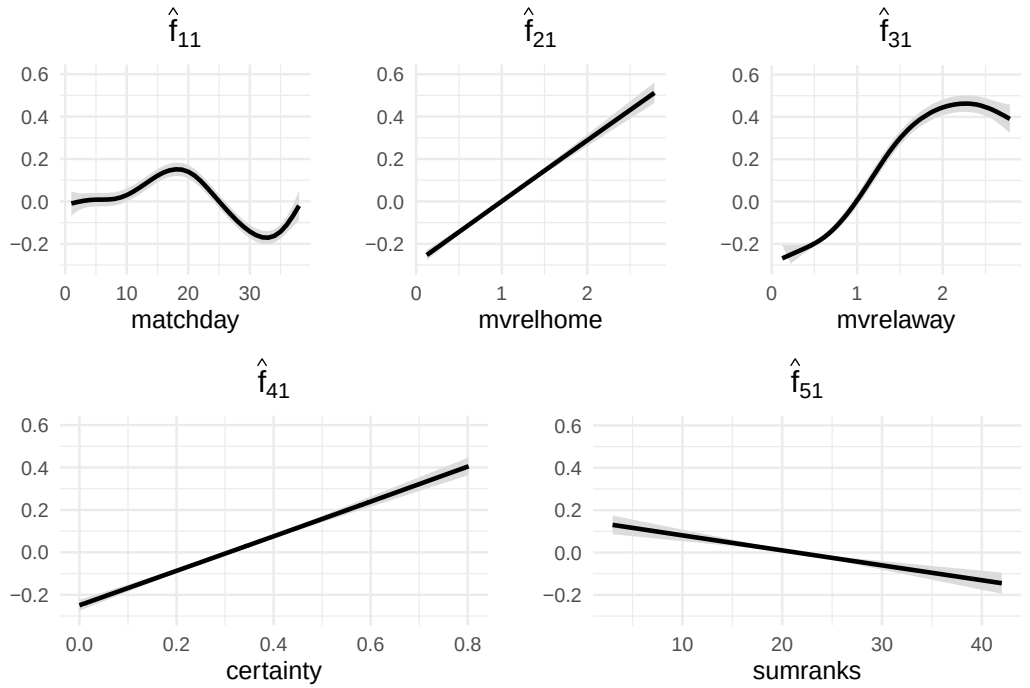


Figure 2.5: Estimated smooth functional effects from model (2.5) on the mean including point-wise confidence intervals.

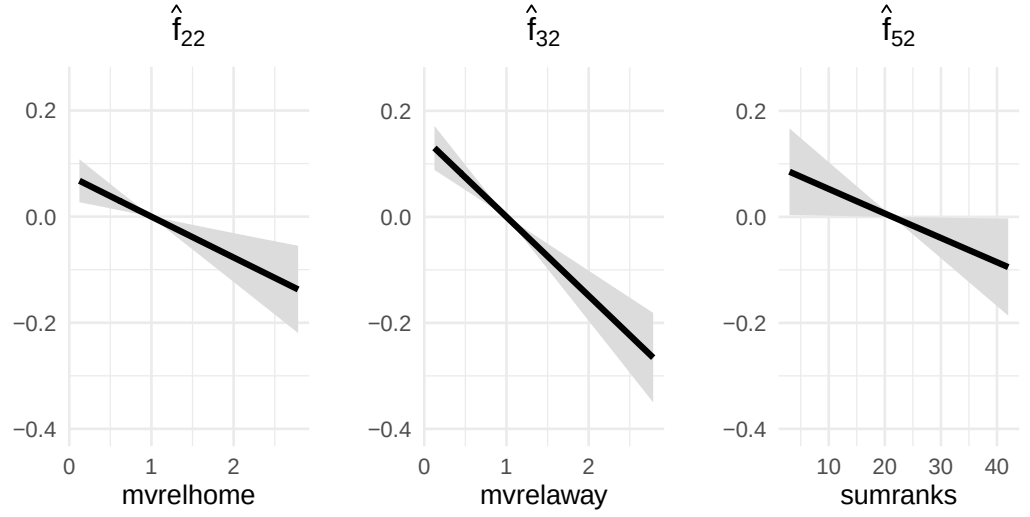


Figure 2.6: Estimated smooth functional effects from model (2.5) on the standard deviation including point-wise confidence intervals.

Modelling the standard deviation as a function of *matchday* and *certainty* did not improve the model fit according to the AIC. Thus, there is no systematic trend for the standard deviation with respect to *matchday* and *certainty*. The estimated smooth functional effects for *mvrelhome*, *mvrelaway*, and *sumranks* on the standard deviation are the only ones which improved the model fit. All effects seem to be rather linear, indicating that the standard deviation in betting volumes tends to be lower for teams with a higher (relative) market value (cf. Figure 2.6). Considering the flexible class of

GAMLSS for modelling betting volumes as proposed in this subsection improves the AIC substantially compared to the simple linear model presented above ($\Delta\text{AIC} = 251$).

2.4 Final remarks

Our results indicate that determinants of demand for sports betting are similar to those for sporting events. Match quality, uncertainty of outcome, and time and economic factors significantly affect bettors' behaviour, confirming previous results on betting demand (*Humphreys et al.*, 2013; *Paul and Weinbach*, 2010). However, Betfair customers wagering on Premier League matches prefer (ex-ante) lopsided matches whereas bettors of NBA, NHL (*Paul and Weinbach*, 2010), and NCAA basketball (*Humphreys et al.*, 2013) matches rather enjoy wagering on close games.

Limitations of our study are given by, first, missing controls for live TV coverage of games, which has shown to influence betting (*Humphreys et al.*, 2013). In addition, by not including TV coverage, our approach might suffer from an omitted variable bias, since the information on TV coverage may correlate with other covariates in our model. TV stations might prefer to cover matches of high-quality teams or matches with a high ex-ante uncertainty of outcome. Thereby, the coefficients corresponding to the covariates *mvrelhome*/*mvrelaway* and *certainty* would be biased. However, betting at Betfair is possible from (almost) all over the world and information on bettors' origin is unavailable. Given that live TV coverage of games differs between countries, it is not feasible for us to link TV coverage to betting behaviour. Second, this chapter considers a single data source, namely Betfair, for the estimation of betting market turnover. While this restriction is caused by the lack of additional data, Betfair is among the largest betting exchanges worldwide. Since betting odds do not vary substantially between betting platforms, we assume similar betting behaviour at other bookmakers. Third, this chapter considers pre-game betting only and does not account for in-game betting, which draws large interest from bettors as well. However, in-game betting critically depends on the dynamics of the game such as early goals. To account for the wide range of potential in-game dynamics, statistical models which account for the time series structure of in-game betting data are required. Fourth, we restricted our analysis to betting volumes aggregated over all betting types. Further research could focus on the betting volumes corresponding to particular bets, such as a home win or the exact number of goals scored. Followed by that point, the effect of some variables,

e.g. the uncertainty of outcome of the match, might differ across betting types.

By deepening the comprehension of determinants of betting volumes it is possible to identify unusual deviations between expected and actual betting volumes. Whereas the usual procedure of fraud detection systems such as Sportradar is to compare fair and actual betting odds to detect fixed matches (see *Forrest and McHale*, 2015), further insights can be achieved by additionally analysing betting volumes. Since in liquid betting markets very high betting volumes are required to observe odds movements, some fixed matches might be missed in large markets due to only little odds movements. Thus, analysing betting volumes in combination with betting odds can serve as a starting point to detect match fixing (*Ötting et al.*, 2018b).

Considering betting volumes from a single betting platform as was done in this chapter gives an idea of the economic importance of a match. The growing number of online bookmakers paired with the large black market for illegal betting suggests that the economic importance of a match is no longer to the participating teams only, but additionally includes the business covering sports betting. It surely does not wonder that team sponsoring by bookmakers has become the rule.

3 Betting market inefficiencies are short-lived in German professional football

3.1 Introduction

Since the publication of Fama's now seminal paper, financial markets are assumed to "digest" all available information and are thus considered efficient (*Fama*, 1970). This assumption implies the complete absence of strategies to "beat the market", i.e. to generate returns exceeding the returns of popular stock indices. Transferred to sports betting, market efficiency implies that betting odds reflect all relevant information on a particular game's outcome. According to *Thaler and Ziemba* (1988), betting markets are weak efficient if no strategies exist to generate positive expected returns.

Whereas previous studies typically restrict themselves to identify market inefficiencies in entire seasons, we argue in this chapter that inefficiencies do exist, but are of temporary nature only. Thus, we focus on periods of a season where the teams' actual strength is difficult to evaluate. We use the specifics of European team sports leagues, where due to promotion and relegation the composition of the teams in a particular division changes between consecutive seasons to document the temporary existence of market inefficiencies. Since teams that have recently been promoted to the top division are hard to assess in terms of their strength, winning probabilities are difficult to determine for two reasons. First, recent performance is only observed against teams from lower divisions. Second, recently promoted teams typically modify the composition of their rosters considerably due to access to more financial resources. These roster changes, in turn, are difficult to evaluate as the newly signed players have to settle and to adapt to their teammates and the team style of play. We argue that a promoted team's strength is revealed after the start of the season, following a reasonable number of appearances. This argument is even more important as in European football — unlike the situation in the North American Major Leagues — no official preseason exists. Instead, teams play friendlies against other teams, typically from

different leagues and countries.

When predicting betting odds, bookmakers have to rely on rather vague information on the actual strength of recently promoted teams. Since the commission rates retained by bookmakers typically do not vary by much between games, we hypothesise that strategies exploiting this lack of information on the strength of recently promoted teams are likely to exist. We assume this betting market inefficiency to be short-lived, since bookmakers collect more reliable information on the teams' strength after the start of the season and process this information efficiently. Analysing betting odds from 14 consecutive seasons of the first division in German football (the Bundesliga), we find positive returns when betting on recently promoted teams at the beginning of the season. This effect disappears around mid-season, suggesting that market inefficiency is a temporary phenomenon.

This chapter is organised as follows: in Section 3.2, we briefly review the literature on biases in betting odds. Section 3.3 starts with the exploratory analysis of the returns on investment when betting on recently promoted teams. In addition, a logistic regression model is fitted to the data.

3.2 Biases in betting odds

Taking the assumption of efficient markets as starting point, a number of empirical studies have provided evidence for the existence of different biases in betting odds, in particular the favourite longshot bias, the home bias, and the sentiment bias.

The favourite longshot bias, which suggests higher returns when betting on favourites than on longshots, has been analysed using betting data from different sports. *Ottaviani and Sørensen* (2008) summarise several potential explanations for this bias, arguing that bettors who are willing to take risks accept a lower expected payout when betting on longshots. Using football betting odds, *Forrest and Simmons* (2008) provide evidence to support the notion of a favourite longshot bias. Other studies, e.g. *Cain et al.* (2003), however, find a reverse favourite longshot bias, implying higher returns when betting on longshots than on favourites. This result is also documented for Major League Baseball betting (see e.g. *Woodland and Woodland*, 1994, 2003).

A further extensively studied bias is the home bias, implying higher returns when betting on the home as opposed to the away team. *Buraimo et al.* (2010) for example

find that referees in the German Bundesliga and in the English Premier League tend to favour home teams by showing, on average, fewer yellow and red cards to players of the home team. Using the same methodology, *Buraimo et al.* (2012) find similar results for the first division in Spain and for the Champions League. *Pollard* (2008) summarises further potential reasons for the existence of a home advantage, such as travel and crowd effects. In an efficient betting market, the home advantage should be priced into betting odds, i.e. bettors should not expect higher returns when placing bets on the home team. However, the evidence regarding the existence of a home bias in betting markets is rather mixed. Using data from the Spanish first division, *Forrest and Simmons* (2008) confirm that betting on the home team significantly increases the probability of winning a bet, whereas (e.g.) *Franck et al.* (2011) fail to find such an effect for the English Premier League.

Further studies have focused on the sentiment bias, the existence of which requires the presence of bettors' sentiment. From a theoretical point of view, popular teams draw increased attention from bettors, which should result in lower expected returns compared to when betting on less popular teams. Analysing betting data from the first divisions in Italy, England, Germany, Spain, and France, *Feddersen et al.* (2017) provide evidence for the sentiment bias. However, as in the case of the favourite longshot bias, the evidence is not clear-cut, as other studies indicate an opposite effect (see, e.g., *Forrest and Simmons*, 2008; *Franck et al.*, 2011).

The studies summarised above have in common that they analyse data covering one or more full seasons to document inefficiencies for a specific season or sport. However, it may well be the case that some biases occur only temporarily. Accordingly, *Borghesi* (2007) as well as *Davis and Krieger* (2017) have analysed data distinguishing between different phases of a season. They argue that at the start of a season typically little information is available about the teams' strength. In addition, *Davis and Krieger* (2017) argue that in pre-season matches in the NBA and NFL, most teams refrain from using their best players. Thus, the winning probabilities of underdogs may be underestimated by bookmakers. By developing a specific betting strategy for these pre-season matches, *Davis and Krieger* (2017) find positive returns when betting on underdogs in particular matches. We argue in a similar direction by assuming that potential inefficiencies occur only temporarily, e.g. at the beginning of a season, as there is less information available on the teams' strength. This is especially the case for recently promoted teams in an open league system, such as the German Bundesliga.

Due to the lack of reliable information on the actual strength of promoted teams, it may be the case that these teams are underrated and have, on average, inappropriate high betting odds. If this holds true, it should be possible to realise positive returns when betting on promoted teams. During the season, as more information becomes available to improve the efficiency of betting odds, such positive returns should disappear. To test this assumption, we use betting odds for matches played in the German Bundesliga during the seasons 2002/03 until 2015/16.

3.3 Evidence from the German Bundesliga

The data considered in our analysis cover all matches of the German Bundesliga during the seasons 2002/03 until 2015/16, totalling in 4,218 matches. Information on the matches was taken from the website www.football-data.co.uk and includes betting odds on each possible game outcome according to the bookmaker www.bet365.com.[1] However, for four matches the information on betting odds is missing. The betting odds are stated in the European standard decimal format. As an illustration, suppose Bayern Munich plays Borussia Dortmund and is quoted at 1.70 to win, betting successfully one euro would result in a payout of 1.70 euro (70 cents profit plus 1 euro return to stake). Since bookmakers cover the uncertainty of match outcomes by adding margins to all odds, the reported odds do not reflect the actual probabilities of match outcomes. Accordingly, one has to correct for the bookmaker commission to calculate the implicit probabilities for all possible game outcomes. Using the odds on a home win, draw, and away win (denoted by o_h, o_d, and o_a, respectively), the probability of a home win p_h is then estimated as:

$$\hat{p}_h = \frac{1/o_h}{1/o_h + 1/o_d + 1/o_a}.$$

The implicit probabilities $\hat{p}_d$ and $\hat{p}_a$ are calculated accordingly. Correcting for the observed average overround of 105%, the implicit winning probability for Bayern Munich in the example above is 56.02%. In the following, we refer to these implicit probabilities as *bookprob*.

Our analysis starts with an exploratory analysis of the returns on investment (ROIs)

[1] www.football-data.co.uk also reports betting odds from other bookmakers that are highly correlated in our sample. We refer to www.bet365.com data in the following as it has the highest coverage for the period considered. For the available periods, pairwise correlations between www.bet365.com and all other bookmakers are at least 0.96.

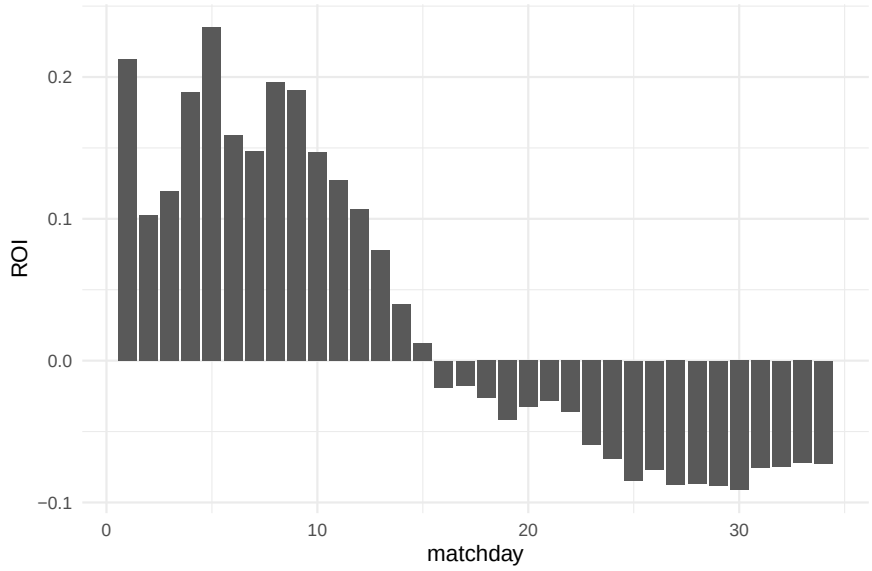


Figure 3.1: ROIs when betting on recently promoted teams in the Bundesliga (2002/03 – 2015/16).

when betting on recently promoted teams. The ROI describes the profits out of a given investment, i.e.

$$\text{ROI} = \frac{\text{payout} - \text{wager}}{\text{wager}}.$$

A positive ROI is equivalent to profits while a negative ROI describes losses. For all ROIs reported in this chapter, we have chosen a wager of 1 euro per bet. Figure 3.1 displays the aggregated ROIs for the seasons considered. It suggests that betting on recently promoted teams is highly profitable at the beginning of a season, but after the first half of a season, the ROIs become negative and remain so until the season ends. The early positive returns are consistent with our theoretical argument developed above, because predicting winning probabilities for recently promoted teams is apparently a particularly challenging task at the beginning of a season when there is little information available on the strength of promoted teams. As during the season more information becomes available, predicting the winning probabilities becomes easier and more accurate, implying that early-season market inefficiencies should disappear.

Due to a potential home bias mentioned above, we further distinguish between recently promoted teams playing at home and playing away. It appears that positive ROIs are observed for recently promoted teams when playing away until matchday 30 (see Figure 3.2). At the same time, we find negative returns only for recently promoted teams playing at home, no matter until which matchday we aggregate our payout.[2]

[2]Since positive returns from betting on promoted teams disappear after the winter break, we checked whether betting *against* promoted teams after the winter break yields positive ROIs. Betting against promoted teams leads to small positive returns for home games only.

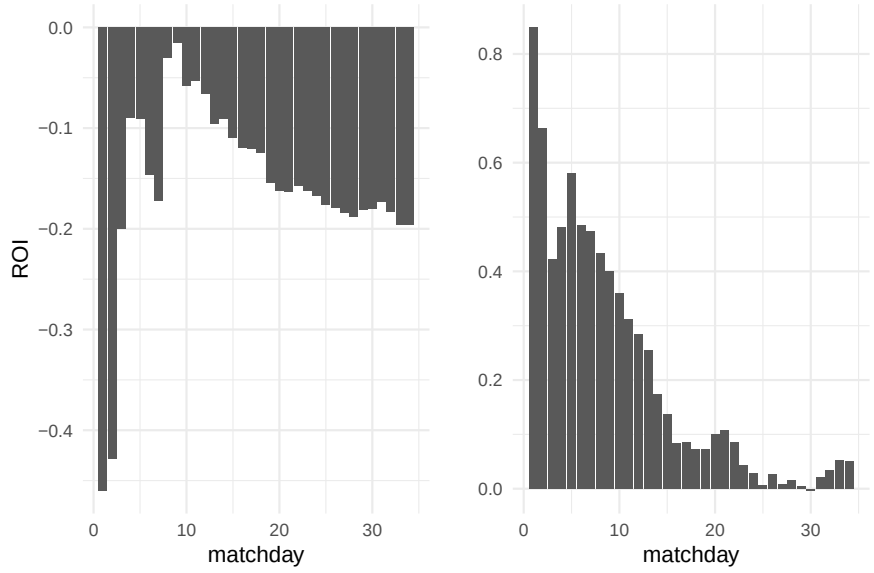


Figure 3.2: ROIs when betting on recently promoted teams playing at home (left panel) and away (right panel) (2002/03 – 2015/16).

Due to the home bias found in other studies, one might have expected the returns for home and away teams in Figure 3.2 to be the other way round. The most plausible explanation for this phenomenon is a psychological one: due to the euphoria around teams that have recently been promoted to the Bundesliga, the players are likely to experience particularly high levels of pressure. This pressure is likely to be higher when playing at home than when playing away, suggesting a better performance on the road than on the home pitch (at least as expected by bookmakers).

To document the robustness of these effects, we fit a logistic regression model. For all matches included in our data set, we focus on bets on the match outcome. Thus, for every match we have two rows in the data: one row for the bet on the home team and one for the bet on the away team. Due to our focus on promoted teams, we delete all 64 matches between two promoted teams, yielding 8,436 observations. The response variable in our analysis is the binary variable *outcome* indicating whether the bet was won ($\mathit{outcome} = 1$) or not ($\mathit{outcome} = 0$). This approach has been considered in several previous studies on biases in betting odds (see, e.g., *Forrest and Simmons*, 2008; *Franck et al.*, 2011).

Since we are particularly interested in the performance of recently promoted teams, we include a dummy variable in our model indicating whether a team has just been promoted (*promoted*). Depending on the outcome of the promotion playoff between the team finished 16^{th} in the first division and the team finished 3^{rd} in the second division, we observe either two or three promoted teams per season. Since we expect

potential profits especially at the beginning of a season — when less reliable information is available about promoted teams — we also include the matchday in our model (*matchday*). In line with existing studies, we control for a potential home bias by including a dummy variable indicating whether a team plays at home (*home*). To account for a potential sentiment bias, we consider the covariate *diffattendance* which covers the difference in the mean home attendance of the two teams in the previous season (this information was taken from www.worldfootball.net). Table 3.1 displays the summary statistics, both for the response and the covariates.

Table 3.1: Summary statistics on the response and the covariates.

	mean	st. dev.	min.	max.
outcome	0.376	–	0	1
promoted	0.134	–	0	1
home	0.5	–	0	1
bookprob	0.371	0.167	0.019	0.916
matchday	17.49	9.811	1	34
diffattendance	0	24.30	−70.53	70.53

As our response variable *outcome* is binary, we fit a logistic regression model. To adequately address the potential biases in betting odds mentioned above, the following covariates are included in the model formulation: *bookprob*, *home*, *promoted*, and *diffattendance*. Since we expect a positive effect for recently promoted teams only for the first few matchdays, we also include an interaction term $promoted \cdot matchday$. In addition, to account for a home bias of promoted teams, we also include an interaction term between *home* and *promoted*. The probability $\Pr(outcome_i = 1)$ is combined with the linear predictor η_i through the logit link function:

$$\text{logit}\big(\Pr(outcome_i = 1)\big) = \eta_i.$$

The linear predictor includes all covariates and interaction terms considered:

$$\begin{aligned}\eta_i =& \beta_0 + \beta_1 bookprob_i + \beta_2 home_i + \beta_3 promoted_i + \beta_4 home_i \cdot promoted_i + \\ & \beta_5 matchday_i + \beta_6 promoted_i \cdot matchday_i + \beta_7 diffattendance_i.\end{aligned}$$

Finally, our observations are likely to be correlated. As discussed above, every match is represented in the data by two rows: if the home team wins, the away team cannot win and vice versa. Hence, we extend our logistic regression by adding a random intercept for each match to account for correlated observations.

Table 3.2: Regression results for the German Bundesliga.

	response variable:
	outcome
bookprob	4.452
	[4.043; 4.866]
promoted	0.424
	[0.088; 0.755]
home	0.095
	[−0.023; 0.213]
matchday	0.004
	[−0.001; 0.009]
promoted · matchday	−0.017
	[−0.032; −0.002]
home · promoted	−0.327
	[−0.617; −0.036]
diffattendance	0.002
	[−0.001; 0.004]
constant	−2.330
	[−2.505; −2.157]
observations	8,436
Pseudo R^2	0.159
Note:	95% CIs are shown in brackets.

Table 3.2 displays the results. To interpret the estimated effects of the covariates *home*, *promoted*, and *matchday*, one has to carefully consider the interaction terms. It appears that betting on recently promoted teams ($promoted = 1$) that did not play at home ($home = 0$) increases the odds of winning a bet significantly. According to our fitted model, the odds for winning a bet can be increased until matchday 24 following this betting strategy. This estimated effect is much smaller for recently promoted teams playing at home: betting on these teams also increases the odds for winning a bet significantly, but according to the coefficients of the interaction terms $promoted \cdot matchday$ and $home \cdot promoted$, the odds are increased until matchday 2 only. Figure 3.3 shows the probabilities for winning a bet — as predicted under the fitted model — when betting on recently promoted teams playing away (*diffattendance* is set to its mean, i.e. 0, in this figure). We see that for matchdays 5, 10, and 15 the estimated probabilities lie above the diagonal (solid black line in Figure 3.3). Hence, according to our model, betting on recently promoted teams in away matches at the beginning of a season leads to higher winning probabilities than is to be expected from the implied odds provided by bookmakers.[3] At the end of a season, the estimated probabilities for these teams tend to move more and more towards the diagonal, i.e. the initial large difference between the estimated probabilities as implied by our fitted

[3]Indeed, according to Figure 3.3, small returns (8% ROI) are possible when betting on recently promoted teams playing on away ground with implicit winning probabilities of less than 15%. This is in-line with the existing literature on the reverse favourite-longshot bias (see e.g. *Woodland and Woodland*, 1994, 2003 and *Gandar et al.*, 2002).

model and the probabilities stated by bookmakers disappears. For teams which were not recently promoted ($promoted = 0$) we find no such effect. Finally, we find no sentiment bias in the German Bundesliga, as the estimated effect for *diffattendance* is not statistically significant.

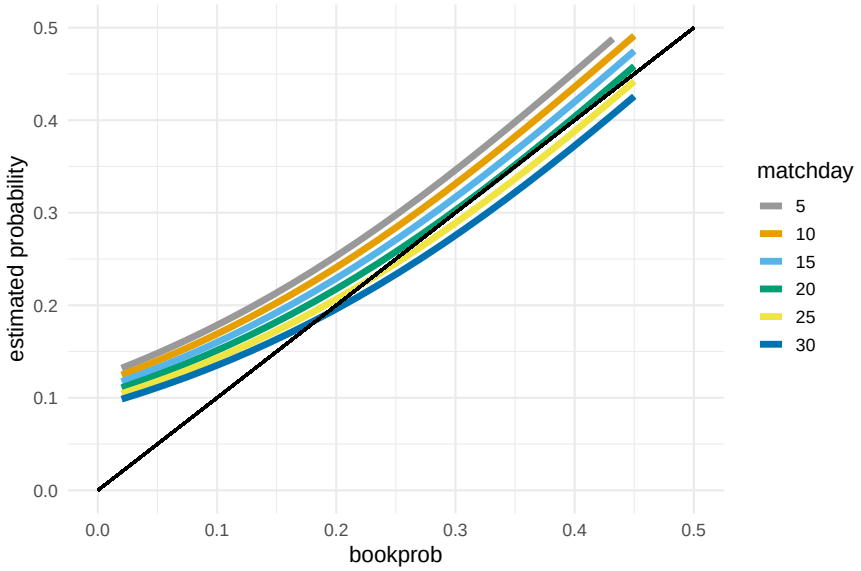


Figure 3.3: Estimated probabilities for betting on recently promoted teams playing on away ground.

3.4 Concluding remarks

Considering data on betting odds from 14 consecutive seasons of the first division in German professional football, we find evidence that bookmakers have difficulties to predict the strength of recently promoted teams. Our results concerning the promoted teams are, first, consistent with the exploratory analysis and, second, consistent with our theoretical considerations above: bookmakers systematically underrate the strength of recently promoted teams, potentially overestimating the teams' need for acclimatisation to tougher competition. While promoted teams play their home games in the same stadium as in the season before, their away games usually take place in larger stadiums and accordingly in front of larger crowds. According to our results, bookmakers overestimate the positive effect of home fans on hosts' results when facing a recently promoted team. From a bettor's point of view, this market inefficiency allows to generate temporary positive ROIs by betting on away wins of promoted teams at the beginning of a season. With more and more information on the promoted teams' strength becoming available during a season the market inefficiency disappears.

It is perhaps surprising to observe market inefficiencies in the German Bundesliga, as it is permanently covered by the media, providing detailed information on player transfers. Moreover, the media covers teams' performance in friendly matches and in training camps prior to the season. For further research, data from lower divisions is of interest as they include both recently promoted and relegated teams. In addition, it may be the case that market inefficiencies are even stronger in leagues that are covered by the media to a lesser extent.

4 Integrating multiple data sources in match-fixing warning systems

4.1 Introduction

The substantial growth of betting markets has led to an increased interest in automated corruption detection systems, in particular such that are data driven, hence exploiting the availability of comprehensive data such as bookmakers' odds. Corresponding research that involves data analysis includes basketball (*Wolfers*, 2006), sumo wrestling (*Duggan and Levitt*, 2002), and football (*Reade and Akie*, 2013), to name but a few. For football, *Forrest* (2012) reports recent match fixing scandals in Europe and overseas. Unfortunately, most ex-post detected fixed matches were not indicated ex-ante by fraud detection systems. Thus, the development of reliable fraud detection systems remains an active area of research. Beneficiaries of such systems include bookmakers, as they would lose customers if match fixing occurs on a regular basis, but also spectators interested in fair competitions, which can also be of economic interest as sponsors do not want to finance sports linked to illegal activity.

Match fixing is used to manipulate the outcome of a certain match to make profits on the betting market. To prevent match fixing, many of the major European football leagues rely on the fraud detection system developed by Sportradar. Sportradar's system compares bookmakers' betting odds to the probabilities as estimated by a statistical model (*Forrest and McHale*, 2015). Because most of the relevant information for estimating match-related probabilities is publicly available, for a match that has not been fixed the probability of any particular outcome stated by bookmakers should be very close to that derived from an adequate statistical model based on all available information (*Reade and Akie*, 2013). Here we argue that focusing on odds only and hence ignoring the betting volumes neglects the importance of market liquidity. High liquidity can potentially result in inflexible odds — i.e. the odds do not adapt as quickly when high volume bets are placed as for markets with a low liquidity — even

for high-volume single bets related to match fixing. Thus, in liquid betting markets it can be very difficult to detect match fixes via odds deviation only. Vice versa, in low liquidity markets singular bets can lead to a large deviation between odds and winning probabilities suggested by a statistical model. Since it is clear that there is relevant information contained in the bets placed on an outcome, we suggest to take into account both betting odds and betting volumes when trying to flag suspicious matches. More specifically, we model both odds and volumes using flexible semi-parametric regression models, and present an approach how these separate models can together be used to flag suspicious matches.

As a case study, we consider the Italian Serie B from season 2009/10 through 2015/16, modelling betting volumes and odds observed on the betting exchange platform Betfair. Such betting exchanges increase incentives to fix matches since they enhance financial opportunities for match fixing due to (e.g.) the absence of limits for wagers and anonymous betting (*Forrest et al.*, 2008). However, since we analyse data from only one betting platform, this chapter is to be regarded as a simple case study of how betting volumes and odds together can be used to detect fixed matches, rather than expecting that all fixed matches are indeed correctly identified.

The main problem in assessing the suitability of any given fraud detection system is that it is generally not known which matches have indeed been fixed, such that studies cannot present accuracy measures such as true and false positive rates. Here we attempt to approximate these rates, exploiting the fact that in case of the Italian Serie B there are several matches where it has been proven that they were fixed. In doing so, we challenge the accuracy of detection systems solely relying on betting odds.

This chapter is organised as follows. Section 4.2 describes the data and models for analysing betting volumes and betting odds. In addition, some special characteristics of betting exchanges are explained. Section 4.3 presents the results, including true and false positive rates for both procedures, i.e. for matches flagged via betting volumes and odds, respectively; as well as for a approach combining both models.

4.2 Building models for betting volumes and odds

We use two separate regression models, for betting volumes and odds, respectively, to detect outliers and hence uncover potential match fixing. For betting volumes, we

analyse the residuals of predicted values to find outliers, while in case of the odds we compare the odds as predicted by the model to the actual betting odds. Data on betting volumes and odds were downloaded from the online betting exchange platform Betfair (www.betfair.com). For football, bets can be placed on several events, e.g. the match outcome, the number of goals, the player scoring the first goal, or even the team being awarded the first throw-in. Liquidity is predominately allocated to the former two betting types. For match fixers, this means that these main types of bets are most attractive, since placing high volume bets on less popular events would be more suspicious. In the more liquid markets, potential manipulations are more likely to remain undetected, since the odds do not adapt as quickly when high volume bets are placed. Thus, in this work, we focus on the most popular types of bets, namely the match outcome (home win, draw, away win) and the total number of goals (specifically the bets "over/under 1.5 goals", "over/under 2.5 goals", and "over/under 3.5 goals", respectively).

4.2.1 Modelling betting volumes

Customers at betting exchanges such as Betfair bet against each other rather than bookmakers. Moreover, customers can either bet on a certain event (backing the event) or against it (laying the event); a more detailed description of the betting process on such betting exchange platforms is provided in Appendix A.1. For a match that has been fixed, say with a predetermined event of at least three goals, this implies opportunities for profits for two strategies, either backing the event "over 2.5 goals" or laying the event "under 2.5 goals". To account for this, we add up the betting volumes in the "over 2.5 goals" and "under 2.5 goals" market. We focus on four types of events, in each case considering the total volume from the back and lay market, respectively:

- match outcome (home win, draw, away win);
- over and under 1.5 goals (*OU1.5*);
- over and under 2.5 goals (*OU2.5*);
- over and under 3.5 goals (*OU3.5*).

We analyse pre-game data, i.e. bets placed before the start of a match. In the Serie B, 462 matches are played per season. Betting volumes are occasionally unavailable for

some betting types, such that we have four or less observations per match, resulting in a total of 11,915 observations from 3,219 matches.

When modelling betting volumes as a function of covariates, several features need to be taken into account: 1) the betting volumes are highly right-skewed, 2) for some covariates it is anything but clear if their effects are linear (e.g. matchday), and 3) there is substantial heteroscedasticity. To address these patterns, we make use of the flexible class of GAMLSS (*Rigby and Stasinopoulos*, 2005). GAMLSS extend generalised additive models (GAM; see *Hastie and Tibshirani*, 1990), allowing not only the conditional mean, but also other possible parameters of the distribution of a response variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$, e.g. the scale or the shape, to be modelled. These distributional parameters can be modelled via smooth functional effects of covariates. More specifically, we apply the semi-parametric additive formulation of GAMLSS,

$$g_k(\boldsymbol{\theta}_k) = \boldsymbol{\eta}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} f_{jk}(\mathbf{x}_{jk}), \tag{4.1}$$

where $\boldsymbol{\theta}_k = (\theta_{k,1}, \ldots, \theta_{k,n})$ is a parameter of the distribution assumed for the response variable $\mathbf{Y}$ (one value for each observation), $g_k(\cdot)$ is a known link function, $\mathbf{X}_k$ is the $n \times J_k'$ design matrix, $\boldsymbol{\beta}_k$ is a vector of regression coefficients of length J_k', and the f_{jk} are smooth functions to be estimated. In (4.1), both g_k and the f_{jk} are evaluated component-wise, e.g. $f_{jk}(\mathbf{x}_{jk}) = \big(f_{jk}(x_{jki}), \ldots, f_{jk}(x_{jkn})\big)$. Depending on the distributional family assumed for $\mathbf{Y}$, several parameters $\boldsymbol{\theta}_k$ and hence several regression equations of the form (4.1) may be considered. Being able to model several parameters of the distribution assumed for the response $\mathbf{Y}$ relaxes the exponential family assumption of classical GAMs, hence for GAMLSS a larger class of distributions can be assumed.

Due to the highly right-skewed distribution of the betting volumes as well as the presence of some outliers, we use a log-normal distribution for our response. Betting volumes are likely to depend on the strength of teams, the uncertainty of the match outcome, and also temporal factors such as matchday ($matchday$) and day of the week (day) (see, e.g., *Humphreys et al.*, 2013; *Paul and Weinbach*, 2010). As a (crude) proxy for the strength of teams we use their market values ($mvhome$ and $mvaway$, respectively) according to www.transfermarkt.com. The market value varies per team and season. The uncertainty of the outcome of a match is taken into account via the implied probabilities of betting odds. Specifically, average betting odds

from up to 52 bookmakers on home and away wins are taken for each match from www.football-data.co.uk and adjusted for the bookmakers' margin. As a measure for the uncertainty of a match, we then use the absolute value of the difference between the average probability of a home win and away win (*certainty*). Finally, to account for the different betting type selections, dummy variables are included with the match outcome selection as reference category (*selections*). Table 4.1 summarises all non-categorical covariates considered for the betting volume model.

Table 4.1: Descriptive statistics for the covariates for the betting volume model.

	mean	st. dev.	min.	max.
volume (in £10^3)	22.63	62.26	0.005	3,383
matchday	–	–	1	42
mvhome (in 10^6 euro)	17.49	10.57	4.4	70.93
mvaway (in 10^6 euro)	17.50	10.64	4.4	70.93
certainty	0.196	0.138	0	0.833

Due to the log-normal distribution assumed for the betting volumes $\mathbf{Y}$, the covariates are linked to the mean $\boldsymbol{\mu}(=\boldsymbol{\theta}_1)$ as well as the standard deviation $\boldsymbol{\sigma}(=\boldsymbol{\theta}_2)$, leading to the following special case of (4.1):

$$\boldsymbol{\mu} = \boldsymbol{\eta}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \sum_{j=1}^{J_1} f_{j1}(\mathbf{x}_{j1}) \tag{4.2}$$

$$\log(\boldsymbol{\sigma}) = \boldsymbol{\eta}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \sum_{j=1}^{J_2} f_{j2}(\mathbf{x}_{j2}). \tag{4.3}$$

Thus, in this model, g_1 is the identity link function, while g_2 is the log link function. Some of the variables discussed above enter the model in (4.2) and (4.3) in the linear part, while for the non-categorical covariates smooth functional effects are estimated. To allow for a different effect of *certainty* on the betting volume for the different betting types, interactions between these variables are also taken into the model. Since we have up to four observations per match contained in our data, the observations coming from the same match may be correlated. Thus, we add a random intercept γ_m for each match m, $m = 1,\ldots,n$. Furthermore, accounting for the up to four observations per match corresponding to the different betting types, $b = 1,\ldots,4$, the predictors of the

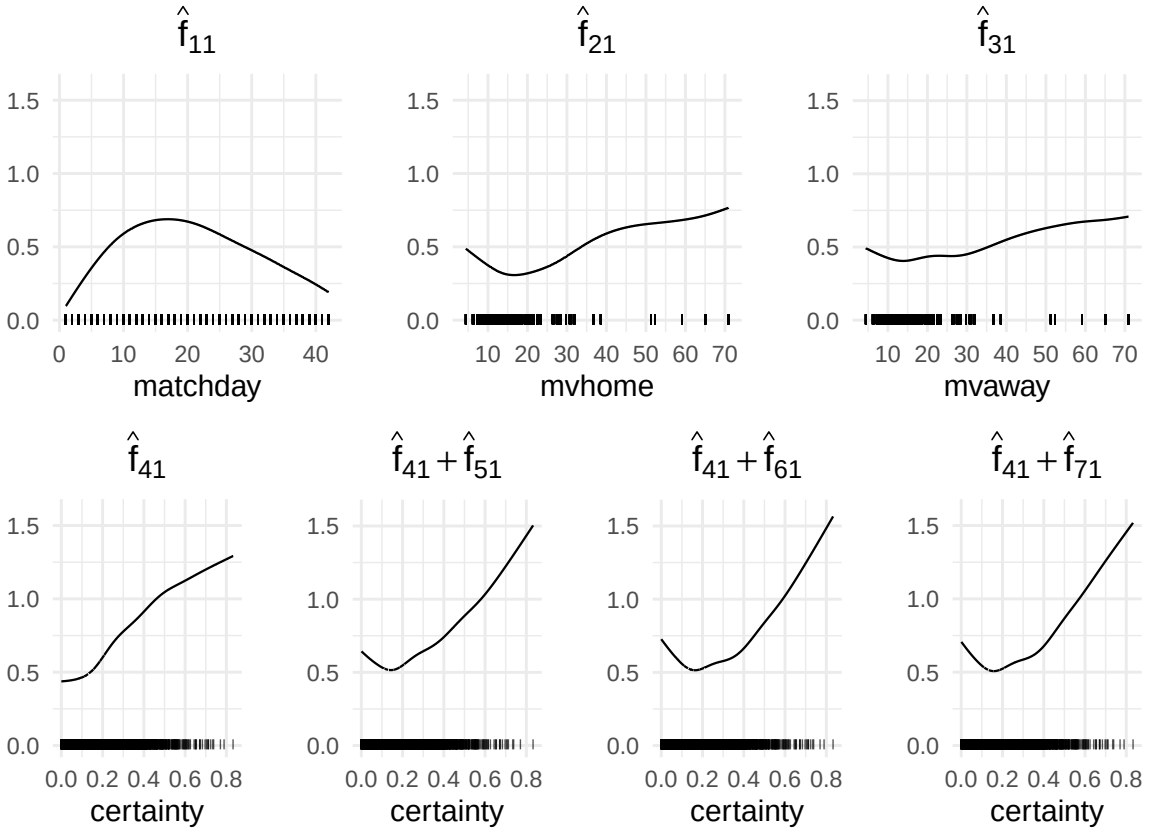
parameters μ_{mb} and σ_{mb} then are of the following form:

$$\begin{aligned}\eta_{mb} = \mathbf{x}'_{mb}\boldsymbol{\beta} + \gamma_m + f_1(matchday_m) + f_2(mvhome_m) \\ + f_3(mvaway_m) + f_4(certainty_m) + f_5(certainty_m) \cdot OU1.5_{mb} \\ + f_6(certainty_m) \cdot OU2.5_{mb} + f_7(certainty_m) \cdot OU3.5_{mb},\end{aligned} \tag{4.4}$$

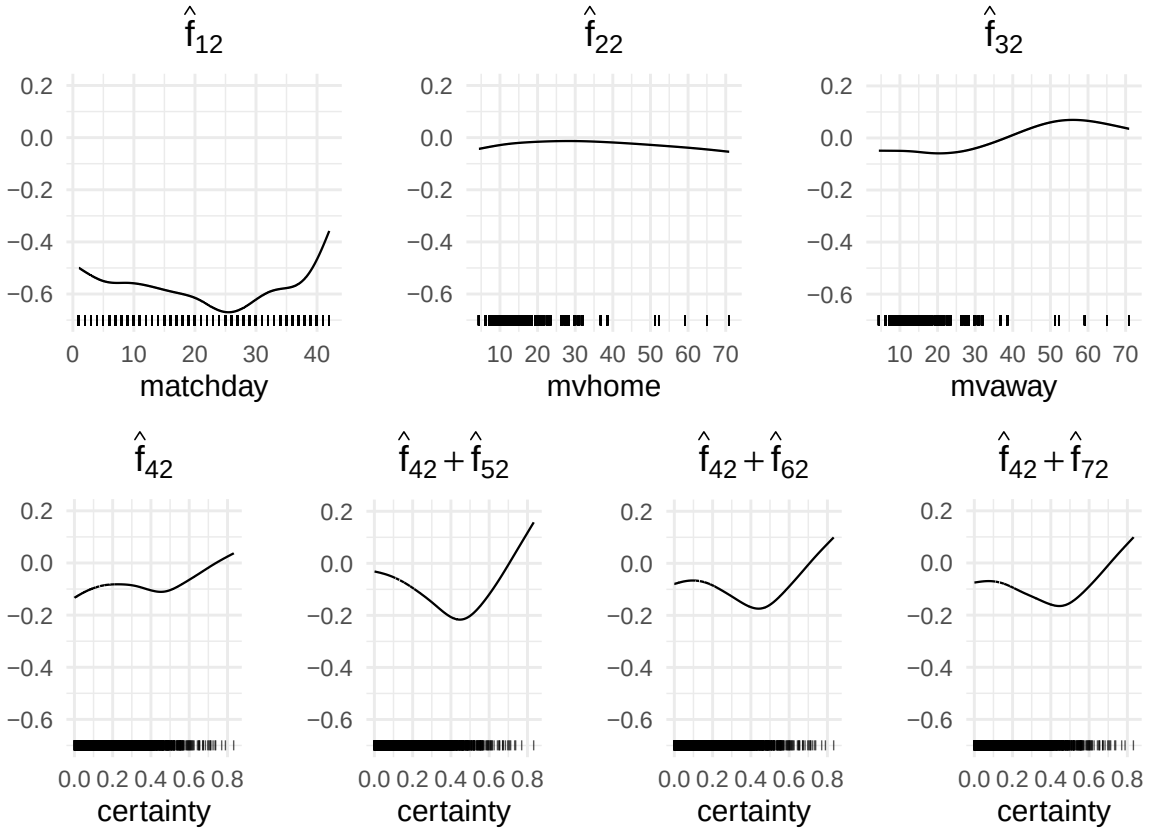
where γ_m is the match-specific random effect, the $\mathbf{x}_{mb}$ is a column vector containing the dummies related to the variables $selection$ and day as well as a one for the intercept, and the f_l are smooth functions. The purpose of our regression analysis is to minimise the prediction error. It will thus be fitted using past data only. Furthermore, variable selection is rather challenging as multiple distributional parameters are modelled via additive predictors. These two points motivate the use of a boosting approach, which performs variable selection during the estimation process and yields more stable predictions due to a reduced estimator variance (see *Mayr et al.*, 2012a). We thus fit the model using gradient boosting with early stopping. The early-stopping algorithm uses cross-validation to calculate the predictive risk (here: the negative log-likelihood) for values of $\mathbf{m}_{\text{stop}} = (m_{\text{stop},1}, m_{\text{stop},2})$ from a grid of reasonable values. We then use the $\mathbf{m}_{\text{stop}}$ that minimises the predictive risk. For each season, all base learners are selected when applying early stopping. In Appendix B.1, we provide a more detailed description of the GAMLSS boosting algorithm.

Figure 4.1 shows the estimated partial smooth functional effects of the selected P-spline base learners for a single season (2015/16), for μ and $\log(\sigma)$, respectively. According to the estimated effect of the covariate *matchday*, betting volumes increase in the first half of a season, reaching a maximum around matchday 20. During the second half of a season, betting volumes tend to decrease. The other estimated effects shown in Figure 4.1a indicate that higher quality teams tend to attract more money, that matches with seemingly more certain outcome attract higher volumes, and that bettors' preferences regarding the uncertainty of the match outcome do not vary much across the different betting types. To facilitate interpretation of the estimated smooth functional interaction effects, for each interaction term we display its sum with the effect of *certainty* for the reference category (see Equation 4.4).

Figure 4.1b shows the estimated smooth functional effects on the standard deviation. The standard deviation of the betting volumes increases for end-of-season matches. This may be due to the fact that there is more variation in the importance



(a) Estimated partial non-parametric effects on $\hat{\mu}$.



(b) Estimated partial non-parametric effects on $\log(\hat{\sigma})$.

Figure 4.1: Estimated partial non-parametric effects for the model specified in (4.4), fitted to data from season 2009/10 to 2014/15 — (a) shows the effects on $\hat{\mu}$*, and (b) on* $\log(\hat{\sigma})$.

of corresponding matches, e.g. with respect to promotion or relegation. In addition, the standard deviation in the betting volumes increases with increasing market values. For the variable *certainty*, we find that lopsided matches tend to increase the standard deviation for all betting types.

	effects on $\hat{\mu}$	effects on $\log(\hat{\sigma})$
	estimate	estimate
over/under 1.5	-3.224	0.325
over/under 2.5	-1.919	0.143
over/under 3.5	-3.649	0.379
Monday	1.573	-0.410
Wednesday	1.083	-0.109
Thursday	0.232	-0.108
Friday	0.930	-0.263
Saturday	-0.192	-0.084
Sunday	0.681	-0.116

Table 4.2: Estimated effects of the dummy variables on the mean (left) and the standard deviation (right) in the betting volume model, fitted to data from season 2009/10 to 2014/15.

Table 4.2 gives the estimated effects of the dummy variables, i.e. the dummies regarding the type of bet and the day of the week. The reference category for the type of bet is the match outcome (home win, draw or away win) while Tuesday is the reference day. On average, bets on the match outcome attract higher betting volumes than over/under bets on the number of goals. In addition, we find that betting volumes vary across the week. The estimated negative coefficient for Saturday matches is most likely caused by substitutes, i.e. matches played concurrently in other popular football leagues, such as the German Bundesliga or the English Premier League. This is in agreement with the relatively large coefficient estimated for Monday games, as there are only few concurrent matches played in other European football leagues on that day.

For outlier detection in any given season, we estimate the model using the data from all previous seasons, then predicting $\hat{\mu}$ and $\hat{\sigma}$ for each match of the season considered. For the season 2009/10, where we do not have any data from previous seasons, we calculate the in-sample predictions. With these predicted parameters we calculate the quantile residuals, which are standard normally distributed if the model is correct (*Dunn and Smyth*, 1996). The QQ-Plot in Figure 4.2 is based on the quantile residuals for the model for season 2015/16, i.e. the model fitted to all data previous to that season, indicating a satisfactory goodness of fit. Especially the large positive quantile

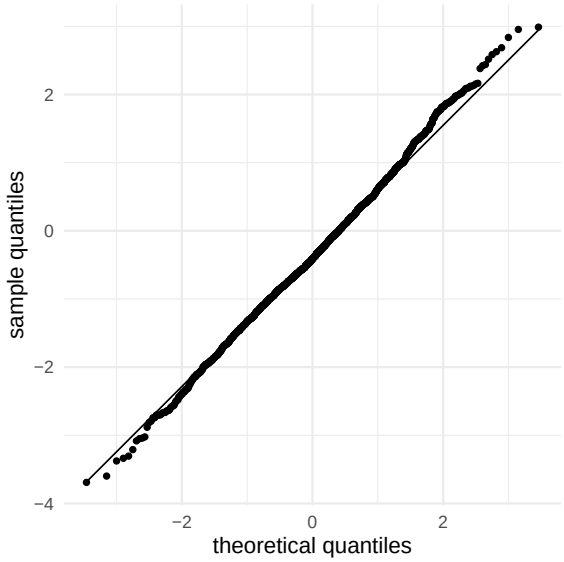


Figure 4.2: QQ-Plot of the quantile residuals for the model fitted to data from season 2009/10 to 2014/15.

residuals — e.g. a quantile residual of 3.03 for the match Albinoleffe vs. Piacenza in 2010, for which it is known that it was manipulated — are of particular interest to us, since they correspond to betting volumes that are much higher than expected under the fitted model. In some cases, such large residuals may be related to match fixing, and in general we simply flag such matches as suspicious. The corresponding observations are further investigated in Section 4.3.

4.2.2 Modelling betting odds

The precise estimation of match outcome probabilities is effectively the bookmakers' business model. Research in this area mostly focuses on testing the efficiency of betting markets (see e.g. *Deutscher et al.*, 2018; *Forrest and Simmons*, 2008) and on modelling specific game events, especially goals (see e.g. *Dixon and Coles*, 1997; *Groll et al.*, 2015; *Maher*, 1982). For the latter, a Poisson distribution is usually assumed for the number of goals scored during a match. However, modelling home and away goals separately as independent Poisson variables would neglect any potential correlation between these random variables. Bivariate Poisson distributions overcome this limitation by explicitly modelling the correlation of two Poisson random variables, in our case home and away goals. We consider the model proposed by *Karlis and Ntzoufras* (2003), where, if Z_k, $k = 1,2,3$, are independent Poisson random variables with parameters $\lambda_k > 0$, then $Y_1 = Z_1 + Z_3$ and $Y_2 = Z_2 + Z_3$ follow the bivariate Poisson distribution:

$$\Pr(Y_1 = y_1, Y_2 = y_2) = e^{-(\lambda_1+\lambda_2+\lambda_3)} \frac{\lambda_1^{y_1}}{y_1!} \frac{\lambda_2^{y_2}}{y_2!} \sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i} \binom{y_2}{i} i! \left(\frac{\lambda_3}{\lambda_1 \lambda_2}\right)^i, \quad (4.5)$$

where the marginal distributions of Y_1 and Y_2 are univariate Poisson with $E(Y_1) = \lambda_1 + \lambda_3$ and $E(Y_2) = \lambda_2 + \lambda_3$, respectively. The parameter λ_3 measures the covariance of Y_1 and Y_2. For football, λ_1 and λ_2 represent the scoring rates of the home and the away team, respectively. *Karlis and Ntzoufras* (2003) allow all parameters in their bivariate Poisson model to depend on covariates. In recent years, the approach has been extended; for example, *McHale and Scarf* (2011) model the dependence between home and away goals via copulas, thus allowing also for negative correlation. They do not use their model for forecasting, but their findings imply that the dependence of home and away goals varies with the ability of the teams.

The GAMLSS seems well-suited to model goal-scoring rates, since the effects of non-categorical covariates, such as the teams' market value or the matchday, are here also likely to be non-linear. While the seminal paper by *Rigby and Stasinopoulos* (2005) on GAMLSS covers univariate distributions only, *Klein et al.* (2015) extended the GAMLSS framework to allow also for multivariate outcomes.

Table 4.3: Descriptive statistics on the covariates for the scored goals model.

	mean	st. dev.	min.	max.
home goals (y_1)	1.389	1.133	0	6
away goals (y_2)	1.052	1.028	0	6
mvhome (in 10^6 euro)	16.24	10.17	1.130	70.93
mvaway (in 10^6 euro)	16.29	10.19	1.130	70.93
avgp4hteam	1.312	0.676	0	3
avgp4ateam	1.373	0.686	0	3
matchday	–	–	1	42

We consider the following covariates for modelling goal-scoring rates: market values of the home and away team (*mvhome* and *mvaway*), their respective average points in the previous four matches, i.e. a moving average over the last four matches, (*avgp4hteam* and *avgp4ateam*), the matchday (*matchday*), and dummy variables indicating whether a team was recently promoted (*promhome* and *promaway*) or recently relegated (*relhome* and *relaway*) (cf. *Groll et al.*, 2015; *Reade and Akie*, 2013). As for the betting volume model, the market values vary per team and season. Table 4.3 summarises the response variables as well as the non-categorical covariates. Our

model for the goal-scoring rates hence is

$$\begin{aligned}\log(\boldsymbol{\lambda}_k) = \boldsymbol{\eta}_{\lambda_k} = \mathbf{x}_k'\boldsymbol{\beta}_k + f_{1k}(mvhome) + f_{2k}(mvaway) + f_{3k}(matchday) \\ + f_{4k}(avgp4hteam) + f_{5k}(avgp4ateam),\end{aligned} \tag{4.6}$$

where the column vector $\mathbf{x}_k$ comprises the four dummy variables (*promhome*, *promaway*, *relhome*, and *relaway*). The purpose of such models, which we fit using past data only, is again to minimise prediction error for future matches. Furthermore, variable selection is again involved in this setting due to now three parameters that are modelled via additive predictors. Thus, we again use gradient boosting to fit the model and to conduct variable selection.

Gradient boosting for GAMLSS with multivariate distribution of the response variable is similar to the univariate case described in Appendix B.1. In our setting, the aim is to find estimates for the additive predictors $\hat{\eta}_{\lambda_k}$, $k = 1,2,3$, that optimise the loss function ρ, which in this case is the negative (log-)likelihood of the bivariate Poisson distribution. As for the univariate case, in each iteration the algorithm is updating the best-fitting base learner for each additive predictor until the respective stopping iteration is reached. For our purpose, i.e. modelling the number of goals in football, the negative log-likelihood of the bivariate Poisson distribution and the partial derivatives with respect to λ_1, λ_2, and λ_3 are implemented as a new `families` object in the `gamboostLSS` (version 2.0-0) package (*Hofner et al.*, 2017). The derivatives are presented in Appendix B.1. For a more detailed description of the boosting algorithm with focus on the bivariate case, see *Groll et al.* (2018).

As for the betting volume model, we apply early stopping in the boosting-based estimation procedure, i.e. the early-stopping algorithm uses cross-validation to calculate the predictive risk for values of $\mathbf{m}_{\text{stop}} = (m_{\text{stop},1}, m_{\text{stop},2}, m_{\text{stop},3})$ from a grid of reasonable values. We apply the early-stopping algorithm for each of the seven season-specific models separately, resulting in different optimal stopping values. Finally, probabilities for the match outcome and for the over/under bets are obtained from the bivariate Poisson distribution with estimated rate parameters $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$.

The number of base learners selected by the boosting algorithm varies by season. Table 4.4 summarises the selected base learners for each model. Market values are almost always selected for all three distribution parameters, whereas the covariates indicating whether a team was promoted or relegated are rarely selected. Figure 4.3

	09/10	10/11	11/12	12/13	13/14	14/15	15/16
mvhome	λ_1, λ_2	λ_1, λ_2	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$
mvaway	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$
matchday	λ_2, λ_3	λ_2, λ_3	$\lambda_1, \lambda_2, \lambda_3$	λ_2, λ_3	$\lambda_1, \lambda_2, \lambda_3$	λ_2, λ_3	λ_2, λ_3
avgp4hteam	λ_2, λ_3	λ_2	$\lambda_1, \lambda_2, \lambda_3$	λ_2, λ_3	$\lambda_1, \lambda_2, \lambda_3$	λ_2	λ_2
avgp4ateam	λ_2, λ_3	$\lambda_1, \lambda_2, \lambda_3$	λ_1	λ_2, λ_3	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$	$\lambda_1, \lambda_2, \lambda_3$
promhome	–	–	λ_1	λ_2	–	–	λ_2
promaway	–	–	λ_1	–	$\lambda_1, \lambda_2, \lambda_3$	λ_2, λ_3	λ_3
relhome	–	–	–	–	–	–	–
relaway	–	–	–	λ_2	λ_2	λ_2	λ_2

Table 4.4: Overview of selected base learners for all seasons.

displays the estimated effects of the selected P-spline base learners for a single season (2015/16), for λ_1, λ_2, and λ_3, respectively. Regarding λ_1, the market value of the home team has a diminishing marginal positive effect, which is intuitively plausible. Unsurprisingly, the effect of the market value of the away team is negative, i.e. higher market values of the away team lead, on average, to less goals of the home team. The effect of the average points in the previous four matches of the away team is also as expected. Interestingly, the average performance during the last four matches of the home team is not selected, as it does not affect, according to this model, the number of goals scored by the home team. Regarding λ_2, the estimated effects for the market values follow the same logic as for λ_1, i.e. higher market values lead, on average, to more goals and higher market values of the opposition lead, on average, to fewer goals. However, while the points of the home team over the last four matches were not selected for λ_1, the points *avgp4ateam* of the away team are chosen for λ_2. We consider this to be an intuitive finding: for away teams, the current form and hence the level of confidence is potentially more relevant than for home teams. The estimated effects of the variable *matchday* indicate more goals on average in end-of-season matches for the away team, although the effect size is rather small. For the covariance between the home and away goals, i.e. λ_3, four P-spline base learners are selected. According to the estimated effects, higher market values are linked to a lower covariance. The fact that the market values, i.e. the overall strengths of teams, affect the *dependence* between home and away goals is in agreement with the findings of *McHale and Scarf* (2011). In addition, the covariance changes during a season with highest covariance for end of season matches, where games are more often either hard-fought competitions (likely with less goals on both sides) or irrelevant (and hence with potentially less effort spent on defending and hence more goals on both sides). The higher the points in the last four matches of the away team, the lower, on average, the

covariance between the home and away goals. Again, this could be due to the level of confidence exhibited by the away team.

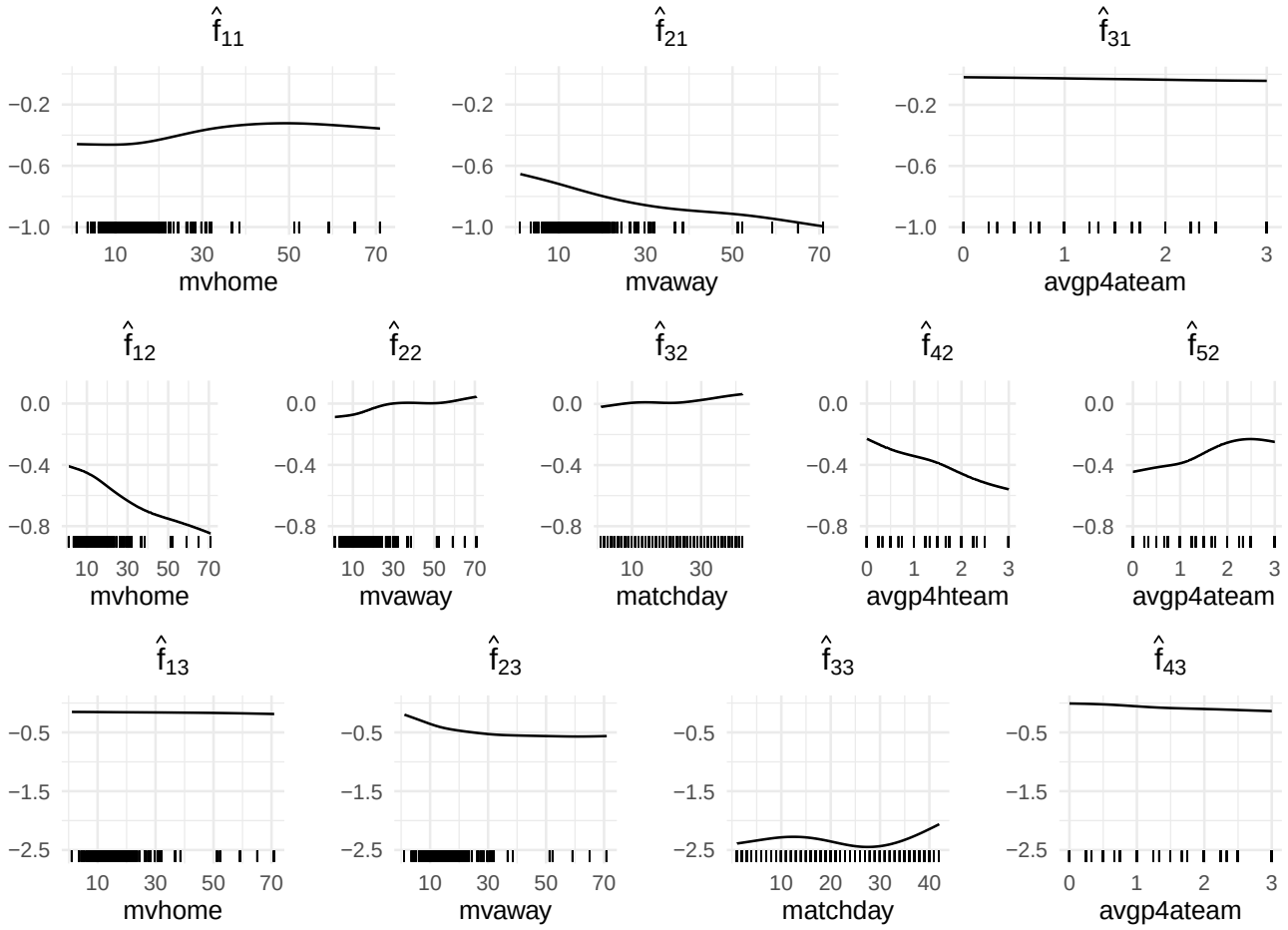


Figure 4.3: Estimated partial non-parametric effects on $\hat{\lambda}_1$ (top row), $\hat{\lambda}_2$ (middle row) and $\hat{\lambda}_3$ (bottom row) for the model fitted to data from season 2009/10 to 2014/15.

When predicting betting odds, it is crucial to compare the out-of-sample prediction accuracy of the model-based odds to the odds stated at the bookmaker/betting exchange, in our case Betfair. In comparing the odds we are able to detect outliers, which we flag as suspicious matches. If the odds derived from the bivariate Poisson GAMLSS model would have poor predictive power, this comparison and flagging of outliers would be pointless. We analyse the odds separately for all betting types. For comparing the prediction performance of the odds, we follow *Rue and Salvesen* (2000) by using pseudo-likelihood statistics, which are equivalent to the geometric means of the probabilities for the actual results. Hence, larger pseudo-likelihood statistics imply higher prediction power. Since no data is available for the "under 1.5 goals" selection at Betfair, we omit this betting type. Table 4.5 displays the pseudo-likelihoods for the Betfair and the model odds, respectively. For each season, the model odds

were derived from the models fitted to all seasons from season 2007/08 until the previous one. For each betting type except the over/under 2.5 goals, the Betfair odds have a slightly better prediction accuracy, but overall we consider the performance of our model to be satisfactory. Clearly, the predictive power could be further improved if more covariate information was available, or if the odds were updated dynamically throughout a season, rather than using the model fitted up to the previous season in the out-of-sample prediction.

	home	draw	away	U1.5	O1.5	U2.5	O2.5	U3.5	O3.5
Betfair	0.454	0.311	0.275	–	0.702	0.568	0.443	0.772	0.236
Model	0.442	0.295	0.268	–	0.668	0.576	0.424	0.771	0.229

Table 4.5: Pseudo-likelihood statistics: Betfair and Poisson-Model odds separated by the different betting types.

4.3 Detection of match fixing

For the seasons from 2009/10 to 2015/16, 24 Serie B matches have been proven to be fixed. This section presents how many of these fixed matches are identified by outlier detection using the betting volume and the odds model, respectively. Furthermore, a procedure to detect fixed matches using both models is presented. Note that the relevant betting type for the proven fixed matches is unknown. For example, a fixed match ending in a 2-2 draw could be either fixed as a draw, for some number of total scored goals, e.g. over 3.5 goals, or both. Hence, if our models flag a match for at least one betting type as suspicious and this match has effectively been proven to be fixed, we group this match into the "correctly predicted" class.

4.3.1 Classification results based on betting volumes

For the betting volumes, quantile residuals are used to detect matches that may have been fixed. To find an optimal cut-off value for the classification based on betting volumes, and to further investigate the accuracy of this classifier, a receiver operating characteristic (ROC) curve is considered (solid line in Figure 4.4). In addition, we make use of Youden's index, which is defined as $\textit{sensitivity} + \textit{specificity} - 1$ (see *Youden*, 1950). The maximum of this index corresponds to an optimal cut-off value. Furthermore, as we only have very few proven fixed matches in our data, we also calculate the positive and negative predicted values (PPV and NPV, respectively) which are defined as *number of true positives/number of positives predicted* and *number of*

false positives/number of negatives predicted, respectively. The corresponding results are displayed in Table A3 in Appendix B.2.

For the quantile residuals of the betting volumes, the maximum Youden index is 0.41, corresponding to an optimal cut-off value for the quantile residuals of 0.754. This leads to the identification of 1,103 suspicious matches, out of which 18 are known to have been fixed. The confusion matrix in Table 4.6 (first values) summarises the classification results obtained from the betting volume model. These results imply a true positive rate of $\frac{18}{24} = 75\%$ and a false positive rate of $\frac{1085}{3195} \approx 33.96\%$. Arguably, a false positive rate of about a third may be too high for the corresponding warning system to be useful in practice.

4.3.2 Classification results based on betting odds

When comparing Betfair's closing odds with our estimated odds, lower odds at Betfair imply that an event is more likely to occur than stated by our estimated odds. To measure this deviance, for each betting type we divide the model-based odds by the corresponding Betfair odds. Matches which exceed the empirical β quantile of this ratio are flagged as suspicious. As for the betting volumes, we make use of a ROC curve to find the optimal cut-off, i.e. the optimal β quantile for flagging matches as suspicious. Figure 4.4 (dotted line) shows the corresponding ROC curve. The maximum of the Youden index for the classification based on betting odds is 0.33, equivalent to an optimal cut-off quantile of $\beta = 0.86$ for the odds fraction. Given this optimal cut-off, our procedure results in 1,621 matches being flagged, from which 20 are actually known to have been fixed. This corresponds to a true positive rate of $\frac{20}{24} \approx 83.3\%$, which is slightly higher compared to the betting volume model, but is also accompanies by a much higher false positive rate of $\frac{1601}{3195} \approx 50.1\%$. Table 4.6 displays the corresponding values for the confusion matrix (values in the middle). The classification results for the betting odds approach via the PPV can be found in Table A3 in Appendix B.2. Comparing the results of the two procedures for their particular optimal cut-off, we find that the betting volume-based detection leads to a lower false positive rate, whereas the true positive rate of the odds model is slightly higher. However, it should be noted here that both confusion matrices depend on the corresponding cut-off values used for outlier detection. To investigate the influence of the cut-off values applied, the ROC curves (see Figure 4.4) show the true and false

positive rates for all possible thresholds. For the quantile residuals, the cut-off values range between -7.3 and 5.6, and from 0 to 1 for the quantile of the odds ratio. The lower the cut-off values, the more matches are flagged, leading to a higher true positive rate for both models, at the cost of a much increased false positive rate. The area under curve (AUC) for the betting volume model is about 0.72 and for the odds model 0.60. Given that the optimal AUC value is 1 and a value of 0.5 represents a random guess, the classification via betting volumes is more accurate according to the AUC.

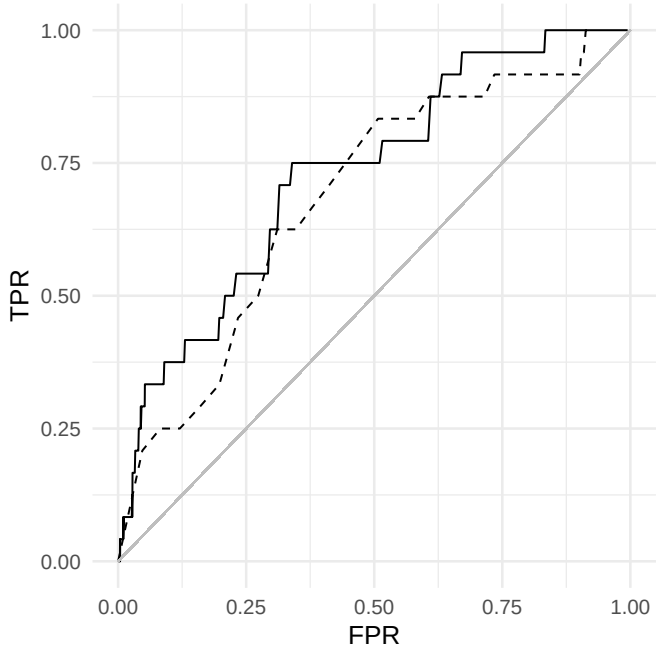


Figure 4.4: ROC curves for all possible cut-off values for the betting volume model (solid) and the odds model (dotted). The diagonal (grey) represents a random guess.

4.3.3 Combining the classification based on volumes and odds

We combine the two approaches described above as follows. First, we define a two dimensional grid containing all combinations of possible cut-off values for the betting volume and the odds model, respectively. Second, for each pair of cut-off values, we flag all matches as suspicious for which the corresponding observations are greater than *at least one* of the cut-off values, i.e. matches which are flagged by at least one model. For the third step, we again make use of both Youden's index and the PPV to find an optimal pair of cut-off values for both models. The results for the classification based on the PPV can again be found in Appendix B.2. The maximum of Youden's index for the combined approach is 0.44, and thus higher than for either of the two approaches when considered separately. The corresponding optimal cut-off values are 0.75 for the

quantile residuals from the betting volume model and 0.99 for the quantiles of the odds fraction. Using these cut-off values, a true positive rate of 79.2% is achieved at the expense of 35.5% false positives. Table 4.6 summarises the classification results. Thus, compared to detecting fixed matches via betting volumes only, combining both approaches slightly improves the true positive rate, whereas the false positive rate remains roughly the same. Compared to the detection relying on odds solely, the true positive rate is slightly lower, but the false positive rate is considerably lower. However, it is worth noting here that this comparison is based on the particular optimal cut-off values for the respective approach.

	actual: fixed	actual: normal	sum
predicted: fixed	18 / 20 / 19	1,085 / 1,601 / 1,133	1,103 / 1,621 / 1,152
predicted: normal	6 / 4 / 5	2,110 / 1,594 / 2,062	2,116 / 1,598 / 2,067
sum	24	3,195	3,219

Table 4.6: Confusion matrix for flagged matches due to betting volumes / betting odds / combined approach based on cut-off values chosen via the Youden index.

4.3.4 Discussion of the results

It is worth noting that — for the respective optimal cut-off values — not all fixed matches found by comparing the odds are also found by the betting volume model, and vice versa. In other words, some fixed matches are correctly classified by one model only. Specifically, there are 14 matches which were identified by both models, whereas four matches are correctly identified only by the volume-based outlier detection, and six other matches only via the odds. Thus, the overall detection rate is improved by analysing betting volumes as well — notably those from only a *single* betting platform — suggesting that using betting volumes in addition to betting odds helps to detect fixed matches. However, both approaches and especially the detection of fixed matches via betting odds leads to a high false positive rate. When combining both approaches using the corresponding optimal pair of cut-off values, the accuracy is comparable to that of the betting volume model only, with a slightly higher true positive rate. Table A2 in Appendix A.3 details the 24 fixed matches contained in our data, together with the largest quantile residual and odds ratio quantile, respectively, across the different betting types.

There are some possible explanations why a fixed match is flagged by only one of

the models. For example, in highly liquid markets, betting activity by match fixers has little effect on betting odds, such that odds do not adapt as quickly when high volume bets are placed, and is hence unlikely to end up being classified as suspicious. In less liquid markets the impact of singular bets in terms of a shift in betting odds is typically more substantial. Since we have data only for one betting platform, it is also possible that match fixers place their bets with other bookmakers. In such a case, we do not observe unusual betting volumes at Betfair, but may still find a deviation from fair betting odds at Betfair: if match fixers place heavy bets with other bookmakers, the odds at these bookmakers start to drop, and Betfair follows with lower odds, because otherwise there would be possibilities for arbitrage.

Although we make use of optimal cut-off values, when using the suggested combination of betting volume model and odds model, then practical considerations will guide the choice of the optimal cut-off values. The fraud detection system developed by Sportradar takes into account the expertise of journalists and other experts who assess whether there may be unmodelled but genuine factors that could explain suspicious odds, e.g. key players being injured, or very one-sided matches where it is likely that key players will be rested. If a fraud detection system does involve such expert elicitation, several false positive matches could be eliminated by taking the opinions of these experts into account. In such a setting, lower cut-off values for all approaches considered may be adequate.

4.4 Conclusions

Using betting odds and betting volumes from seven consecutive seasons in the second division in professional Italian soccer, we demonstrated how statistical modelling can support match-fixing warning systems. We extended the usual procedure of fraud detection systems — which consists of comparing estimated betting odds to the odds stated by bookmakers — by an outlier analysis of betting volumes. The data basis considered does have some caveats, most notably the uncertainty on which other matches additionally to the 24 known ones have been fixed. Nevertheless, our results clearly indicate that modelling betting volumes in addition to betting odds improves the accuracy of detecting fixed soccer matches.

Although for the respective optimal cut-off values a substantial amount of matches are flagged by the different approaches, there are still fixed matches which were not

captured by our models, which to some extent is likely due to restrictions of our analysis. First, we focus on pre-game betting only and do not account for in-game betting. Second, we use data from only one betting platform, Betfair, for the analysis of the betting volumes. This restriction is simply a consequence of the lack of additional data. For the betting odds, the limitation to only one betting platform is of no major importance, since betting odds do not vary substantially between betting platforms. In any case, it is remarkable that modelling betting volumes still yields a true positive rate of approximately 75% (yet at the same time 33.96% false positives) when considering only pre-game betting data from Betfair. Furthermore, it is worth noting that some of the false positives as returned by our analysis may in fact be true positives, i.e. matches that were fixed but (as of yet) are not known to have been fixed.

The main objective of our work was to demonstrate the potential usefulness of extending early-warning systems to also incorporate information on betting volumes. We thus focused primarily on a *comparison* of the two warning systems, corresponding to the outlier detections conducted under the two separate models. However, as we believe that such systems would in most cases lead to a detailed check of a given match already if only one of the two approaches flags a match, the presented approach of combining the two models is adequate for practitioners. Furthermore, the high false positive rate of the detection based solely on odds is reduced by the combined approach, whereas at the same time the true positive rate remains at about the same level. However, as discussed above, these results are based on the respective optimal cut-off values, whereas the choice of cut-off values in practice might rather depend on other factors, such as if other expert knowledge can be taken into account.

Clearly, the accuracy and hence the usefulness of the model-based outlier detection crucially depends on the predictive performance of the model. Thus, future research could focus on further refining the model formulations developed in this work, e.g. by incorporating additional covariates, or by allowing models to be dynamically updated throughout a season. For the betting volume model, one could also consider a multilevel model and allow effects for covariates such as the weekday to vary across the several betting types. Furthermore, it would also be of interest to develop statistical models for betting taking place *during* games. For such in-game betting, odds strongly depend on the dynamics of the game, e.g. early goals, ball possession or running distance (cf. *Schauberger et al.*, 2018). Dynamic and effectively latent factors such as the momentum of the match could be accounted for using time series regression mod-

els such as Markov-switching GAMLSS (cf. *Adam et al.*, 2017). The wide range of potentially relevant game dynamics, together with the natural time series structure of live-betting data, renders outlier detection based on dynamically adjusted odds a challenging task.

5 The hot hand in professional darts

5.1 Introduction

In sports, the concept of the "hot hand" refers to the idea that athletes may enter a state in which they experience exceptional success. For example, in basketball, players are commonly referred to as being "in the zone" or "on fire" when they hit several shots in a row. Although empirical analyses of the hot hand phenomenon tend to focus on sports due to the corresponding data being relatively easily accessible, the notion of the hot hand does in fact apply to much more general settings in which streaks may occur, including human performance in general (*Gilden and Wilson*, 1995a), artistic, cultural, and scientific careers (*Liu et al.*, 2018), the performance of hedge funds (*Edwards and Caglayan*, 2001; *Hendricks et al.*, 1993; *Jagannathan et al.*, 2010), enduring rivalries in international relations (*Colaresi and Thompson*, 2002; *Gartzke and Simon*, 1999), and even gambling activities, against all odds (*Xu and Harvey*, 2014). However, when perceiving such dynamics, people tend to over-interpret streaks of success and failure, respectively (*Bar-Hillel and Wagenaar*, 1991). This phenomenon has been studied intensively by behavioural economists and psychologists (see, e.g., *Tversky and Kahneman*, 1971, 1974), and is regarded as a cognitive illusion that has been considered as a primary example for how humans form beliefs and expectations (*Kahneman*, 2011; *Thaler and Sunstein*, 2009). Especially in gambling settings it has been demonstrated that people strongly believe in the "streakiness" of their performances, while at the same time also acting according to the gambler's fallacy, such that after a streak of identical outcomes an increase in betting volume against the streak is observed despite an i.i.d. random process generating the outcome (*Croson and Sundali*, 2005). Such apparent irrationality underlines the importance of being able to precisely quantify a potential hot hand effect in settings where its existence is highly disputed, e.g. in professional sports. In general, a profound knowledge regarding the existence and magnitude of streakiness in performances can aid general decision-making (*Miller and*

Sanjurjo, 2018).

In their seminal paper, *Gilovich et al.* (1985) analysed basketball free-throw data to find no support for a hot hand, hence coining the notion of the "hot hand fallacy". The alleged fallacy has been attributed in particular to a potential memory bias, with notable streaks in performances being more memorable than outcomes that are perceived as random, but also to general misconceptions regarding chance, with laypeople expecting randomness to lead to performances that are more balanced in terms of successes and failures than is actually the case. Since the landmark paper by *Gilovich et al.* (1985), there has been mixed evidence regarding the hot hand in sports, with some papers claiming to have found indications of a hot hand phenomenon and others disputing its existence. *Bar-Eli et al.* (2006) review the literature on the hot hand in sports, including analyses of data from basketball, baseball, golf, tennis, volleyball and bowling. They summarise 24 studies, from which only 11 studies provide evidence for a hot hand effect. Perhaps due to the availability of increasingly large data sets, most of the more recent studies have found evidence for a hot hand effect (see *Green and Zwiebel*, 2017; *Miller and Sanjurjo*, 2018; *Raab et al.*, 2012; *Shea*, 2014), whereas only some studies dispute its existence by providing mixed results (see *Elmore and Urbaczewski*, 2018; *Wetzels et al.*, 2016).

Two types of approaches have been used to investigate such potential hot hand patterns, namely 1) analyses of the serial correlation of shot *outcomes* (see, e.g., *Dorsey-Palmateer and Smith*, 2004; *Gilovich et al.*, 1985; *Miller and Sanjurjo*, 2014), and 2) such that use a latent variable to describe the form of a player (see, e.g., *Albert*, 1993; *Green and Zwiebel*, 2017; *Sun*, 2004; *Wetzels et al.*, 2016), where the hot hand is understood as serial correlation in shot *probabilities*. While there is no consensus in the literature regarding the definition of the hot hand, *Stone* (2012) and *Miller and Sanjurjo* (2018) show that direct analyses of the correlation in outcomes, as per 1) above, may vastly underestimate the correlation in shot probabilities. For example, a correlation of $\rho_p = 0.4$ in shot probabilities can co-occur with a very much lower correlation of $\rho_r = 0.057$ in shot realisations (*Miller and Sanjurjo*, 2014; *Stone*, 2012). In other words, if a genuine hot hand process is driven by autocorrelation in success probabilities (i.e. in players' forms) — which may very well be the case — then this can easily go undetected if the focus lies on the (much weaker) serial correlation of outcomes. *Stone* (2012) and *Arkes* (2013) thus conclude that it is preferable to

analyse correlation in shot probabilities, as per 2) above. Hence, in this chapter, we focus on approach 2), which we believe is also more aligned with the way terminology related to the hot hand concept (e.g. "on fire", "in the zone") is commonly applied — as argued by *Stone* (2012), it seems most natural to measure players' form by their time-varying success probabilities, rather than (noisy) shot outcomes.

In addition to such conceptual issues regarding the representation of the hot hand in the data-generating process, *Miller and Sanjurjo* (2018) highlight a subtle selection bias that may sneak into analyses of sequential data, which provides a further challenge to the findings of *Gilovich et al.* (1985). Aside from mathematical fallacies, which would already seem to explain many failed attempts to prove the existence of the hot hand, we note that many of the existing studies considered data, e.g. from baseball or basketball, which we believe are hardly suitable for analysing streakiness in performances. For example, when analysing hitting streaks of a batter in baseball, other factors such as the performance of the pitcher are also important but hard to account for. The same applies to basketball, as there are also several factors affecting the probability of a player to make a shot, e.g. the position (of a field goal attempt) or the effort of the defence. In particular, an adjustment of the defensive strategy to stronger focus on a player during a hot hand streak can conceal a possible hot hand phenomenon (*Bocskocsky et al.*, 2014).

To overcome these caveats, here we investigate whether there is a hot hand effect in professional darts, a setting with a high level of standardisation of individual throws. In professional darts, well-trained players repeatedly throw at the dartboard from the exact same position and effectively without any interaction between competitors, making the course of play highly standardised. In the existing literature, there are very few contributions that consider darts data, and almost all of these are restricted to laboratory settings. For example, *Van Raalte et al.* (1995) analyse the effect of positive and negative self-talk on throwing performances, considering the throwing sequences of 60 individuals, each of length 15. The hot hand effect has previously been investigated using darts data by *Gilden and Wilson* (1995b); analysing only 24 throwing sequences of eight volunteers, they find no evidence for a hot hand effect. Here we consider a much larger data set, with $n = 167,492$ throws in total, which allows for comprehensive inference regarding the existence and the magnitude of the hot hand effect. Using state-space models, we evaluate serial dependence in a latent state process, which can

be interpreted as a player's varying form, in line with approaches of type 2) above.

5.2 Data

A dartboard is divided into 20 numbered slices (1 to 20 points) and the centre of the board, the latter with an outer circle (the single bull; 25 points) and an inner circle (the bullseye; 50 points). Each of the 20 slices is further divided into three segments, the singles, doubles and triples, respectively, with the latter two resulting in twice or triple the slice number being awarded as points. All matches in our analysis are played by two players in the *501 up* format. In this format, both players start with 501 points and make their turns one after another. Within each turn, a player throws three darts in quick succession, with the value of the segment hit by each dart being reduced from the current score. The first player to reach exactly 0 points wins a "leg". The last dart used to reduce the score to 0 must hit a double or the bullseye ("double out"). To win the match, a player must be the first to win a pre-specified number of legs (typically between 7 and 15).[1] If a player wins the match, he advances to the next round of the tournament.

Data was extracted from `http://live.dartsdata.com/`, covering all professional darts tournaments organised by the Professional Darts Corporation (PDC) between April 2017 and January 2018. In our analysis, we include all players who played at least 50 legs during the time period considered. This leads to a total of 8,310 legs and 167,492 dart throws (from 833 matches played across 25 tournaments).

At the beginning of a leg, players consistently aim at high numbers to quickly reduce their points. The maximum score in a single throw is 60 as in a triple 20 (T20), which players usually aim at, but the data indicate the triples of the numbers 11–19 (T11 to T19), and bullseye (Bull), to be targeted in the initial phase of a leg as well. In fact, *Tibshirani et al.* (2011) showed that T20 is not necessarily the best segment to aim at, depending on the precision of a player's throws. In addition, when the score before the last throw of a turn is slightly above or around 180, then with that throw players commonly try to avoid "bogey numbers", i.e. scores below 170 points which cannot be reduced to 0 within a player's turn. For example, if the score is 182 points before the third dart of a turn, then aiming at T20 but hitting the single 20 would reduce the

[1] In some matches, a player must win a pre-specified number of "sets", where each set is played as best of five legs. Whether the match is played in sets or in legs is not relevant for our analysis.

score to 162 points, which is a bogey number. Hence, with 182 points left, instead of aiming at the T20, players may aim at T12, since if they fail and hit the single 12, the score reduces to 170 points which can still be checked out with three darts during the next turn. Thus, in the initial phase of a leg we regard any throw to land in the set $H = \{\text{T11}, \text{T12}, \text{T13}, \text{T14}, \text{T15}, \text{T16}, \text{T17}, \text{T18}, \text{T19}, \text{T20}, \text{Bull}\}$ as success. The corresponding empirical proportions of throws hitting the elements of H are displayed in Table 5.1. Since a leg is won once a player reaches exactly 0 points, players do not always target H towards the end of legs, but rather numbers that make it easier for them to reduce to 0. To retain a high level of standardisation and comparability across throws, we thus truncate our time series data, excluding throws where the remaining score was less than $c = 180$ points.

We consider binary time series $\{y_t^{p,l}\}_{t=1,\ldots,T_{p,l}}$, indicating the throwing success of player p within his l–th leg in the data set, with

$$y_t^{p,l} = \begin{cases} 1 & \text{if the } t\text{–th throw lands in } H; \\ 0 & \text{otherwise,} \end{cases}$$

where the $T_{p,l}$–th throw is the last throw of player p in his l–th leg with the player's remaining score still greater than or equal to $c = 180$. The final data set then comprises $n = 167,492$ throws by $P = 73$ players. To illustrate the structure as well as typical patterns of the data, we display Gary Anderson's throwing success histories throughout his first 10 legs in the data:

001 011 011
111 110 0
000 111 101
010 000 101 01
000 110 101
111 000 010 0
110 100 101
100 010 010 00
101 010 000 1
110 100 101

Each row corresponds to one leg — truncated when the score fell below 180 — and gaps between blocks of three successive dart throws indicate a break in Anderson's

play due to the opponent taking his turn. Next we formulate a model that enables us to potentially reveal any unusual streakiness in the data, i.e. a possible hot hand effect.

Table 5.1: Absolute frequencies and proportions for the different outcomes of H in our data set.

outcome	frequency	proportion
T11	1	0.00001
T12	3	0.00002
T13	1	0.00001
T14	0	0
T15	6	0.00004
T16	4	0.00002
T17	546	0.003
T18	1,709	0.010
T19	9,897	0.059
T20	53,509	0.319
Bull	108	0.0006

5.3 Modelling the hot hand in darts

5.3.1 State-space model of the hot hand

We aim at explicitly incorporating any potential hot hand phenomenon into a statistical model for throwing success. As discussed in Section 1, it seems desirable to focus on potential serial correlation in success probabilities. Conceptually, a corresponding hot hand phenomenon naturally translates into a latent, serially correlated state process, which for any player considered measures his varying form. For average values of the state process, we would observe normal throwing success, whereas for high (low) values of the state process, we would observe unusually high (low) percentages of successful attempts. Figuratively speaking, the state process serves as a proxy for the player's "hotness" — alternatively, it can simply be regarded as the player's varying form. The magnitude of the serial correlation in the state process then indicates the strength of any potential hot hand effect. A similar approach was indeed used by *Wetzels et al.* (2016) and by *Green and Zwiebel* (2017), who use discrete-state hidden Markov models to measure the form. While there is some appeal in a discrete-state model formulation, most notably mathematical convenience and ease of interpretation (with cold vs. normal vs. hot states), we doubt that players traverse through only finitely many states, and advocate a continuously varying underlying state variable

instead, thus allowing for gradual changes in a player's form. Specifically, dropping the superscripts p and l for notational simplicity, we consider models of the following form:

$$y_t \sim \text{Bern}(\pi_t), \quad \text{logit}(\pi_t) = \eta_t(s_t), \quad s_t = h_t(s_{t-1}) + \varepsilon_t, \tag{5.1}$$

where $\{y_t\}_{t=1,\ldots,T}$ is the observed binary sequence indicating throwing success, and $\{s_t\}_{t=1,\ldots,T}$ is the unobserved continuous-valued state process indicating a player's varying form. We thus model throwing success using a logistic regression model in which the predictor $\eta_t(s_t)$ for the success probability π_t depends, among other things, on the current form as measured by s_t. The unobserved form process $\{s_t\}$ is modelled using an autoregressive process.

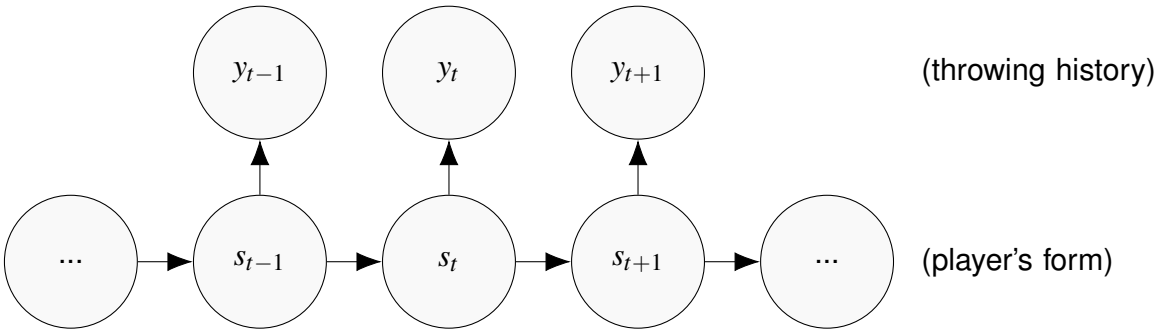


Figure 5.1: Dependence structure of the modelling framework used to investigate a potential hot hand effect. The throwing history $\{y_t\}$ is an observed binary sequence indicating whether or not individual throws were successful, while the player's form $\{s_t\}$ is an unobserved continuous-valued variable which drives the player's time-varying success probabilities.

Model (5.1) is a special case of a state-space model (SSM), with the dependence structure illustrated in Figure 5.1. Under this model, the outcomes of the throws are *conditionally* independent, given the player's form process. However, if there is serial correlation in the form process s_t, then this will induce serial correlation also in the throwing performance. This model formulation is thus in line with the suggestion by *Stone* (2012) to focus on autocorrelation of success probabilities. Crucially, the model includes the possibility to be reduced to the nested special case of uncorrelated success probabilities, and hence independent observations, corresponding to absence of any hot hand phenomenon. Before we dicuss how to conduct maximum likelihood estimation within these type of models, we first present the models considered in the empirical analysis, i.e. the exact forms of $\eta_t(s_t)$ and of h_t that we use to analyse the darts data.

5.3.2 Model specifications

For our hot hand analysis, we formulate three different models, which are all special cases of the general formulation stated in equation (5.1). We start with formulating a benchmark model (*Model 1*) which corresponds to the absence of any hot hand effect. To account for differences between the players, this model includes player-specific intercepts $\beta_{0,p}$ in the predictor, which indicate the individual players' proportions of throwing success (on the logit scale). In principle, a player's overall performance level may change over his career, but since our data covers less than one year of tournaments, it is reasonable to assume this quantity to be constant for each player over the observation period considered. Furthermore, we include a categorical covariate D_t, $D_t \in \{1,2,3\}$, indicating the position of the dart thrown at time t within the player's current turn (first, second or third). This categorical covariate is included since the relative frequency of hitting H, i.e. of throwing success in the early stages of a leg, does in fact differ notably across the three throws within a player's turn, with the empirical proportions of hitting H in our data found to be $0.355, 0.409$ and 0.420 for the first, second and third throw, respectively. The substantial improvement after the first throw within a player's turn is partly due to the necessary re-calibration at the start of a turn (but see further discussion below). Hence, the predictor in *Model 1*, a basic logistic regression model, is given by

$$\text{logit}(\pi_t) = \beta_{0,p} + \beta_1 I_{\{D_t=2\}} + \beta_2 I_{\{D_t=3\}},$$

with $I_{\{\cdot\}}$ denoting the indicator function, and $\beta_{0,p}$ player p's baseline level for the first dart within any given turn. In *Model 2*, a state variable $\{s_t\}$ is included to account for potential serial correlation in success probabilities (i.e. a hot hand effect). We assume $\{s_t\}$ to follow an autoregressive process of order 1,

$$\begin{aligned} \text{logit}(\pi_t) &= \beta_{0,p} + \beta_1 I_{\{D_t=2\}} + \beta_2 I_{\{D_t=3\}} + s_t; \\ s_t &= \phi s_{t-1} + \sigma \varepsilon_t, \end{aligned}$$

with $\varepsilon_t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. Effectively this is a Bernoulli model for throwing success in which the success probability fluctuates around the players' baseline levels — $\beta_{0,p}$, $\beta_{0,p}+\beta_1$ and $\beta_{0,p}+\beta_2$ for within-turn throws one, two and three, respectively — according to the autoregressive process $\{s_t\}$. The process $\{s_t\}$ can be interpreted as a player's time-

varying form. Thus, at the time of an individual throw, the success probability of hitting H depends on the general ability of the player, on the position of the throw within a turn, and on the current (underlying) form as modelled by the state process. Within this model formulation, a hot hand is defined as autocorrelation in the state process, i.e. in the current form. For $\phi = 0$, the model reduces to a model without autocorrelation in the underlying form (i.e. absence of a hot hand), which is similar to *Model 1*, but includes an additional stochastic component in players' abilities. Otherwise, if there is autocorrelation in the underlying form, i.e. if $\phi > 0$, then this would provide evidence in favour of a hot hand effect in the form of positively correlated success probabilities. For the beginning of a new leg, we explicitly account for the result of the last leg, by assuming $s_1 \sim \mathcal{N}(\mu_{\delta_{\text{won}}}, \sigma_\delta)$ if the last leg was won, and $s_1 \sim \mathcal{N}(\mu_{\delta_{\text{lost}}}, \sigma_\delta)$ if the last leg was lost, i.e. we assume that a player's form in a new leg may depend on the result of the last leg, as expressed by the mean of the initial distribution. It should be noted here that by pooling the data of different players, the parameters $\beta_1, \beta_2, \phi, \sigma, \mu_{\delta_{\text{won}}}, \mu_{\delta_{\text{lost}}}$ and σ_δ are assumed to be equal across players, whereas the ability (as measured by $\beta_{0,p}$) varies between players. In Chapter 5.5 we discuss how these assumptions could be modified.

To consider the structure of a darts match in more detail, we further distinguish between transitions *within* a player's turn to throw three darts (e.g. between first and second, or between second and third throw) and those *across* the player's turns (e.g. between third and fourth throw). This extension is considered in *Model 3* and accounts for the fact that there is a short break in a player's action between his turns, with the next turn starting with an empty board, whereas within a single turn the three darts are thrown in very quick succession — it thus seems plausible that any possible hot hand effect may show different time series dynamics within turns than across turns. Specifically, within *Model 3* we assume a periodic autoregressive process of order 1 (PAR(1); *Franses and Paap*, 2004) to describe a player's time-varying form:

$$\begin{aligned} \text{logit}(\pi_t) &= \beta_{0,p} + \beta_1 I_{\{D_t=2\}} + \beta_2 I_{\{D_t=3\}} + s_t, \\ s_t &= \begin{cases} \phi_a s_{t-1} + \sigma_a \varepsilon_t & \text{if } t \text{ mod } 3 = 1; \\ \phi_w s_{t-1} + \sigma_w \varepsilon_t & \text{otherwise.} \end{cases} \end{aligned}$$

Despite distinguishing between transitions within and across turns, this model still allows for longer term hot hand effects that last through the whole leg. Before we

present the results of the different models in Chapter 5.4, in the next section we first discuss how to conduct maximum likelihood estimation within the general formulation given in Eq. (5.1).

5.3.3 Maximum likelihood estimation

The likelihood of a model as in (5.1) involves analytically intractable integration over the possible realisations of s_t, $t = 1, \dots, T$. We use a combination of numerical integration and recursive computing, as first suggested by *Kitagawa* (1987), to obtain an arbitrarily fine approximation of this multiple integral. Specifically, we finely discretise the state space, defining a range of possible values $[b_0, b_m]$ and splitting this range into m intervals $B_i = (b_{i-1}, b_i)$, $i = 1, \dots, m$, of length $(b_m - b_0)/m$. The likelihood of a single throwing history can then be approximated as follows:

$$\begin{aligned} L_T &= \int \cdots \int p(y_1, \dots, y_T, s_1, \dots, s_T) d_{s_T} \dots d_{s_1} \\ &\approx \sum_{i_1=1}^{m} \dots \sum_{i_T=1}^{m} \Pr(s_1 \in B_{i_1}) \Pr(y_1 | s_1 = b_{i_1}^\star) \prod_{t=2}^{T} \Pr(s_t \in B_{i_t} | s_{t-1} = b_{i_{t-1}}^\star) \Pr(y_t | s_t = b_{i_t}^\star), \end{aligned} \tag{5.2}$$

with $b_i^\star$ denoting the midpoint of B_i. This is just one of several possible ways in which the multiple integral can be approximated (see, e.g., *Zucchini et al.*, 2016, Chapter 11). In practice, we simply require that m be sufficiently large. With the specification as logistic regression model as in (5.1), we have that

$$\Pr(y_t | s_t = b_{i_t}^\star) = \left\{ \text{logit}^{-1}\left(\eta_t(b_{i_t}^\star)\right) \right\}^{y_t} \cdot \left\{ 1 - \text{logit}^{-1}\left(\eta_t(b_{i_t}^\star)\right) \right\}^{1-y_t}.$$

The approximate probability of the state process transitioning from interval $B_{i_{t-1}}$ to interval B_{i_t}, $\Pr(s_t \in B_{i_t} | s_{t-1} = b_{i_{t-1}}^\star)$, follows immediately from the specification of h_t and the distribution of the noise, ε_t.

The computational cost of evaluating the right hand side of Equation (5.2) is of order $\mathcal{O}(Tm^T)$. However, the discretisation of the state space effectively transforms the SSM into a HMM, with a large but finite number of states, such that we can apply the corresponding efficient machinery. In particular, for this approximating HMM, the forward algorithm can be applied to calculate its likelihood at a cost of order $\mathcal{O}(Tm^2)$ only (*Zucchini et al.*, 2016, Chapter 11). More specifically, defining $\boldsymbol{\delta} = (\delta_1, \dots, \delta_m)$

with $\delta_i = \Pr(s_1 \in B_i)$, $i = 1, \ldots, m$, the transition probability matrix (t.p.m) $\mathbf{\Gamma} = (\gamma_{ij})$ with $\gamma_{ij} = \Pr(s_t \in B_j | s_{t-1} = b_i^\star)$, $i, j = 1, \ldots, m$, and $m \times m$ diagonal matrix $\mathbf{P}(y_t)$ with i–th diagonal entry equal to $\Pr(y_t | s_t = b_i^\star)$, the right hand side of Equation (5.2) can be calculated as

$$L_T \approx \boldsymbol{\delta}\mathbf{P}(y_1)\mathbf{\Gamma}\mathbf{P}(y_2)\ldots\mathbf{\Gamma}\mathbf{P}(y_{T-1})\mathbf{\Gamma}\mathbf{P}(y_T)\mathbf{1}, \tag{5.3}$$

with column vector $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^m$. Equation (5.3) applies to a single leg played by one player. Assuming independence of the individual leg's throwing histories, the likelihood of the full data set is obtained as

$$L = \prod_{p=1}^{73} \prod_{l_p=1}^{L_p} \boldsymbol{\delta}\mathbf{P}(y_1^{p,l_p})\mathbf{\Gamma}\mathbf{P}(y_2^{p,l_p})\ldots\mathbf{\Gamma}\mathbf{P}(y_{T_{p,l}}^{p,l_p})\mathbf{1}. \tag{5.4}$$

We estimate the model parameters by numerically maximising the approximate likelihood, subject to the usual technical issues as detailed in *Zucchini et al.* (2016).

5.4 Results

For *Model 1*, the benchmark model corresponding to the absence of a hot hand effect, the estimated player-specific baseline levels for the first within-turn throws, $\beta_{0,1}, \ldots, \beta_{0,73}$, range from -1.021 to -0.295, corresponding to throwing success probabilities ranging from 0.265 to 0.427. The coefficients β_1 and β_2, which correspond to the increase of throwing success probabilities after the first throw within a player's turn (on the logistic scale), are estimated as 0.228 and 0.276, respectively. From the first to the second within-turn throw, there is thus a strong increase in the chance of success, followed by a further, smaller increase from the second to the third throw.

Table 5.2: Parameter estimates with 95% confidence intervals for Model 2.

parameter	estimate	95% CI
ϕ	0.494	[0.438; 0.550]
σ	0.659	[0.565; 0.768]
β_1	0.248	[0.221; 0.274]
β_2	0.297	[0.269; 0.325]
$\mu_{\delta_{\text{won}}}$	-0.051	[-0.102; 0.001]
$\mu_{\delta_{\text{lost}}}$	-0.068	[-0.117; -0.019]
σ_δ	0.701	[0.659; 0.745]

Model 2, which unlike *Model 1* can capture a potential hot hand effect, was fitted using $m = 150$ and $-b_0 = b_m = 2.5$ in the likelihood approximation, monitoring

the likely ranges of the process $\{s_t\}$ to ensure the range considered is sufficiently wide given the parameter estimates. Table 5.2 displays the parameter estimates (except the player-specific intercepts) including 95% confidence intervals based on the observed Fisher information. Crucially, the estimate $\hat{\phi} = 0.493$ seems to support the hot hand hypothesis, indicating considerable serial correlation in players' forms, with the associated confidence interval not containing 0. The AIC clearly favours the state-space formulation, *Model 2*, over the benchmark model assuming independent throws, *Model 1* ($\Delta\text{AIC} = 550$). However, the state-space model as it stands makes the implicit assumption that observations are regularly sampled, here such that the sampling unit is one throw. This fails to acknowledge the actual structure of a player's sequence of throws, with blocks of three darts being thrown in quick succession, with breaks of a couple of seconds between blocks (due to the opponent taking his turn).

Table 5.3: Parameter estimates with 95% confidence intervals for Model 3.

parameter	estimate	95% CI
ϕ_w	0.727	[0.642; 0.811]
ϕ_a	0.058	[-0.010; 0.125]
σ_w	0.461	[0.350; 0.607]
σ_a	0.789	[0.699; 0.892]
β_1	0.270	[0.242; 0.297]
β_2	0.331	[0.302; 0.360]
$\mu_{\delta_{\text{won}}}$	-0.025	[-0.069; 0.019]
$\mu_{\delta_{\text{lost}}}$	-0.043	[-0.084; -0.001]
σ_δ	0.691	[0.648; 0.736]

To better reflect the grouping of darts, *Model 3* distinguishes between transitions *within* a player's turn to throw three darts and those *across* the player's turns. For the (approximate) likelihood of *Model 3*, the t.p.m. $\mathbf{\Gamma}$ is then not constant across time anymore, but equal to either a within-turn t.p.m. $\mathbf{\Gamma}^{(w)}$ or an across-turn t.p.m. $\mathbf{\Gamma}^{(a)}$. For *Model 3*, which is clearly favoured over *Model 2* by the AIC ($\Delta\text{AIC} = 242$), the parameter estimates as well as the associated confidence intervals are displayed in Table 5.3. The estimate of the persistence parameter of the AR(1) process active within a player's turn, $\hat{\phi}_w = 0.727$ (95% CI: $[0.642; 0.811]$), corresponds to fairly strong serial correlation. However, the estimate $\hat{\phi}_a = 0.058$ indicates only minimal persistence in the players' forms across turns. In fact, when at time t a player begins a new set of three darts within a leg, then the underlying form variable is drawn from an $\mathcal{N}(0.058 s_{t-1}, 0.789^2)$ distribution, which is notably close to the two possible initial distributions of the AR(1) process ($\mathcal{N}(-0.025, 0.691^2)$ and $\mathcal{N}(-0.043, 0.691^2)$,

respectively) which determine the form at the start of a leg. We cannot rule out that there may be a weak carry-over effect across turns — our results show no conclusive evidence in this regard, with the 95% confidence interval for ϕ_a only just containing the 0. We thus find strong evidence of serial correlation within turns, whereas the evidence regarding potential carry-over effects across turns is inconclusive — the relevance of these findings with regard to the hot hand effect will be discussed in Section 5.

Table 5.4: Overview of Models 1–3.

	no. param.	AIC	state process	description
Model 1	75	223,200	–	player-specific intercepts and dummy variables for the throw within a player's turn
Model 2	79	222,651	AR(1)	*Model 1* + AR(1) state process for the form
Model 3	81	222,400	PAR(1)	*Model 1* + PAR(1) state process, distinguishing transitions within and across a player's turn

Table 5.5: Relative frequencies of the eight possible throwing success histories within a player's turn. The second column gives the proportions found in the data, while columns 3–5 give the proportions as predicted under the various models fitted, for data structured exactly as the real data.

sequence	emp. prop.	*Model 1*	*Model 2*	*Model 3*
0 0 0	0.252	0.222	0.239	0.250
0 0 1	0.151	0.159	0.153	0.150
0 1 0	0.130	0.152	0.139	0.136
0 1 1	0.114	0.111	0.113	0.110
1 0 0	0.103	0.121	0.114	0.108
1 0 1	0.080	0.088	0.082	0.082
1 1 0	0.086	0.084	0.083	0.082
1 1 1	0.084	0.063	0.075	0.083

Table 5.4 provides an overview of the three models fitted, detailing the number of parameters, the AIC values, the type of state process (if any) and a short description. To check the goodness of fit and the adequacy of our models, and also to obtain a more detailed picture of the (short-term) serial correlation implied by *Models 2* and *3*, in Table 5.5 we compare the empirical relative frequencies of the eight possible throwing success sequences within players' turns — 000, 001, 010, 011, 100, 101, 110, and 111 — to the corresponding frequencies as expected under the three different models that were fitted. We restricted this comparison to the first two turns of players within each leg, and used Monte Carlo simulation to obtain the model-based frequencies of

the eight sequences. The benchmark model (*Model 1*), which corresponds to complete absence of any hot hand pattern, clearly underestimates the proportion of 000 and 111 sequences, with deviations of up to 0.03. This indicates correlation in throwing performances within a turn. *Model 2* better reflects the cumulation of 000 and 111 sequences, with a maximum deviation of 0.013. Finally, *Model 3*, which is favoured by the AIC, almost perfectly captures the proportion of 000 and 111 sequences, with the main mismatch in proportions (0.006) found for 010 sequences. While Table 5.5 deals with outcomes *within* turns, Table 5.6 provides similar summary statistics corresponding to outcomes *across* turns. Considering the first two turns by a player within a leg, the table displays the empirical and model-derived proportions of the 16 possible pairs, $(a,b) \in \{0,1,2,3\}^2$, corresponding to the number of successes within two consecutive turns; for example, the pair $(1,3)$ indicates that the number of successful throws in the first and second turn are 1 and 3, respectively. We find that *Model 3* is clearly superior to *Model 2* in terms of capturing the observed proportions of pairs of turns with very different success (e.g. $(0,3)$ and $(2,0)$). This is due to *Model 2* assuming the parameter ϕ to be constant across time, such that the corresponding estimate represents a compromise between within-turn and across-turn correlation found in the data. As a consequence, the former is underestimated, while the latter is overestimated, such that the model predicts pairs of turns with very different success rates to occur less often than is actually the case. In other words, *Model 2* overestimates the magnitude of a potential hot hand effect across turns. On the other hand, comparing the empirical proportions of the pairs $(0,0)$ and $(3,3)$ to the model-derived proportions, we find that *Model 1*, representing absence of a hot hand effect, underestimates persistence in success rates not only within turns, as shown in Table 5.5, but also across turns. Overall, *Model 3* thus comes closest to capturing both within-turn and across-turn correlation in performances. For the observed and expected frequencies shown in Table 5.6, a corresponding χ^2 goodness-of-fit test rejects the null hypothesis at the 1% level for *Model 1* ($\chi^2 = 507.1, \text{df} = 15$) and for *Model 2* ($\chi^2 = 130.2, \text{df} = 15$), while for *Model 3* the test fails to reject the null hypothesis ($\chi^2 = 29.80, \text{df} = 15$).

To further investigate typical patterns of the hidden process $\{s_t\}$, we calculate, under *Model 3*, the most likely trajectory of the latent state (i.e. form) for each player

Table 5.6: Relative frequencies of the 16 possible combinations of numbers of success during the first two turns. The second column gives the proportions found in the data, while columns 3–5 give the proportions as predicted under the various models fitted, for data structured exactly as the real data.

pair	emp. prop.	*Model 1*	*Model 2*	*Model 3*
$(0,0)$	0.067	0.051	0.063	0.064
$(0,1)$	0.097	0.096	0.100	0.098
$(0,2)$	0.072	0.062	0.063	0.067
$(0,3)$	0.022	0.013	0.016	0.021
$(1,0)$	0.092	0.096	0.098	0.100
$(1,1)$	0.150	0.186	0.166	0.155
$(1,2)$	0.110	0.122	0.114	0.109
$(1,3)$	0.032	0.027	0.031	0.034
$(2,0)$	0.067	0.062	0.060	0.066
$(2,1)$	0.106	0.122	0.110	0.105
$(2,2)$	0.079	0.082	0.081	0.076
$(2,3)$	0.026	0.018	0.024	0.025
$(3,0)$	0.020	0.013	0.015	0.019
$(3,1)$	0.031	0.027	0.029	0.030
$(3,2)$	0.023	0.018	0.023	0.023
$(3,3)$	0.007	0.004	0.007	0.007

and leg. Specifically, again dropping the superscripts p and l, we seek

$$(s_1^*,\ldots,s_T^*) = \underset{s_1,\ldots,s_T}{\operatorname{argmax}} \Pr(s_1,\ldots,s_T|y_1,\ldots,y_T),$$

i.e. the most likely state sequence, given the observations. After discretising the state space into m intervals, maximising this probability is equivalent to finding the optimal of m^T possible state sequences. This can be achieved at computational cost $\mathcal{O}(Tm^2)$ using the Viterbi algorithm. We then calculate the corresponding trajectories $\pi_1^*,\ldots,\pi_T^*$ of the most likely success probabilities to have given rise to the observed throwing success histories, taking into account also the player-specific abilities and the dummy variables. Figure 5.2 displays the decoded sequences for six players from the data set. Since there are only $2^3 = 8$ different possible sequences of observations within a player's turn, and since players start each turn almost unaffected by previous performances (cf. $\hat{\phi}_a = 0.058$), there is only limited variation in the *most likely* sequences. The actual sequences may of course differ from these most likely sequences. The player-specific intercepts for within-turn throw one measure the difference in the players' abilities; in Figure 5.2, the corresponding success probabilities range from 0.283 (Zoran Lerchbacher) to 0.420 (Michael van Gerwen). The probability of hitting H increases after the first throw within a turn due to the two dummy variables. We also see

confirmed that the form is not retained across turns.

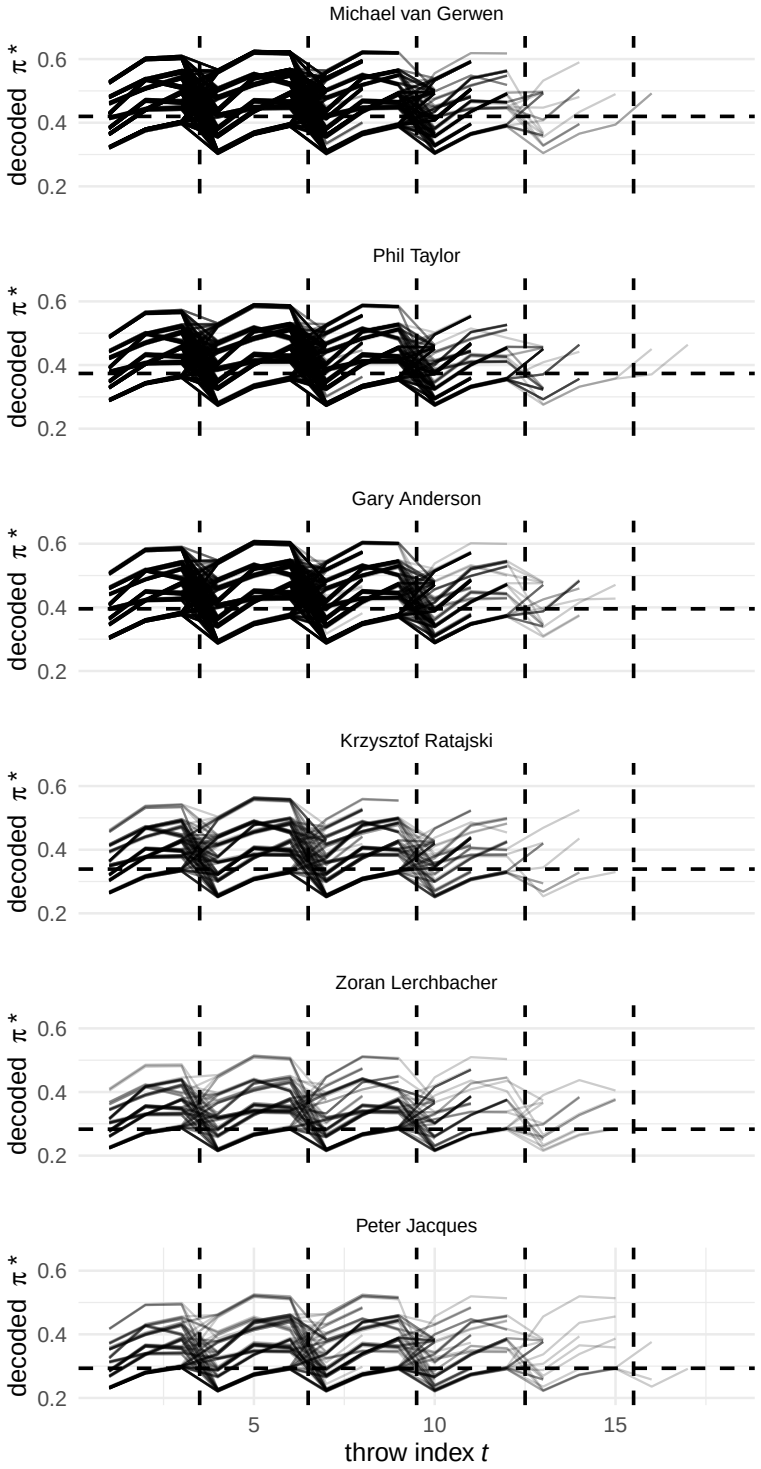


Figure 5.2: Decoded most likely sequences of throwing success probabilities according to Model 3, for > 100 *legs played by each of six players from the data set. The horizontal dashed lines indicate the player-specific intercepts for the respective player's within-turn throw one, and the vertical dashed lines denote the transition between a players' turn of three darts each.*

5.5 Discussion

Our results indicate that within a player's turn, involving three darts thrown in quick succession, there is strong correlation in the latent state process. However, short breaks, which in the given setting result from the opponent taking his turn, effectively result in a fresh start of the process describing the player's form. From a purely statistical point of view, if the hot hand phenomenon is understood as the presence of serial correlation in individuals' forms, then our findings would seem to provide strong evidence in favour of the hot hand. However, some strategic aspects in darts need to be considered when interpreting our results with regard to a potential hot hand effect. In particular, depending on the exact position of a turn's first dart within or close to a triple segment, this dart can potentially be used as a "marker". Darts two and three within the same turn can then be aimed at the marker and may be deflected into the target — the marker thus effectively increases the target area. The coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$, which indicate systematic differences in the success probabilities of the second and third throw within a turn, relative to the first throw, do in fact indicate that the success probability for hitting H increases during a player's turn (see Tables 5.2 and 5.3). However, the corresponding dummy variables do not incorporate any information on whether or not the first dart is a marker dart — if it is, then the success probability may *increase* even further than indicated by $\hat{\beta}_1$ and $\hat{\beta}_2$. Similarly, the first dart may also end up blocking a target area, thus *decreasing* the success probability of subsequent throws. Within our modelling approach, these effects seem to have been captured by the within-turn serial correlation induced by the latent autoregressive process. In other words, while we conceptualised the latent state process as a proxy variable measuring a player's time-varying form, it seems likely that within turns this process actually accommodates other not directly observed effects, namely such related to marker or blocker darts. The strong serial correlation within a player's turn, hence, may be partially tied to the effect of marker darts rather than a hot hand effect. Since our data does not contain corresponding information that would allow us to disentangle these two possible causes, there is no definitive conclusion whether the strong serial correlation within turns provides evidence of a hot hand effect, or rather is a consequence of marker darts. In any case, it is at least questionable whether serial correlation within a sequence of only three darts, thrown in quick succession, is what sports commentators, fans and athletes have in mind when referring to the hot

hand. In conclusion, while we find strong serial correlation for within-hand throws, the persistence parameter measuring potential correlation across turns was estimated to be very small, such that overall we do not find conclusive evidence for a hot hand.

Further research specifically on the hot hand in darts could focus on explicitly addressing player heterogeneity. In addition to the baseline level of π_t, the parameters ϕ_w, ϕ_a, σ_w and σ_a, and hence the magnitude of the hot hand effect, may vary across players. This could reveal that for some players the hot hand effect lasts longer than for others, and potentially also across turns. Modelling this individual variability could be achieved using covariates or, if no suitable covariates are available to explain the heterogeneity, via random effects. However, fitting state-space models such as those presented here already involves a high computational cost, with the numerical maximisation taking several days on a usual desktop computer. This would be further increased when incorporating random effects due to the required integration over possible values of the random effects. Possible parameter estimation approaches moving forward with random effects include: (i) using the Laplace approximation to evaluate the marginal likelihood via the template model builder package in R, (ii) an approximate Bayesian inference approach using an integrated nested Laplace approximation or (iii) using MCMC in a Bayesian framework, where the realisations of the random effects are sampled alongside the other model parameters. While the focus of our analyses was on the hot hand hypothesis, darts does in fact provide an excellent setting for studying several other performance-related hypotheses, due to the highly standardised actions and the absence of interactions between opponents. For example, darts data have recently been used to investigate how individuals perform in high-pressure situations (see *Klein Teeselink et al.*, 2020; *Ötting et al.*, 2020a).

The modelling framework developed in the present chapter, with a continuous-valued latent process representing a player's time-varying form, can easily be tailored to other sports — or in fact any sequential performance measures — for further investigations into the existence and magnitude of the hot hand. A caveat of the existing study is the binary nature of the observations, which corresponds to a rather noisy measure of the actual form. In sports such as archery and shooting, performance can be measured more precisely by considering the continuous-valued distance between a shot and the middle of the target. While these sports easily lend themselves to a time series analysis due to the structured way in which actions take place, there are many other sports of interest where time intervals between actions are irregular, including

immensely popular sports such as football, American football, rugby or basketball. With the attention given to these, it would be of interest to transfer our modelling approach also to corresponding settings with irregular time intervals between actions. Conceptually, this can relatively easily be achieved by replacing the AR(1) process used in this work to represent the latent form of a player by its continuous-time analogue, the Ornstein-Uhlenbeck process. Via discretising the state space, the tools available for *continuous-time* HMMs can then be applied for making inference (*Jackson et al.*, 2003). However, given that evidence from the most recent literature points to a small hot hand effect (see, e.g., *Green and Zwiebel*, 2017; *Miller and Sanjurjo*, 2018), we believe sports settings with no direct interaction between opponents — e.g. archery, shooting, or free throws in basketball in an experimental setting as in *Gilovich et al.* (1985) — to be best suited for analyses related to the hot hand, as otherwise it can be difficult to disentangle hot hand patterns from potential confounding factors.

6 A regularised hidden Markov model for analysing the 'hot shoe' in football

6.1 Introduction

In sports, the performance of players is frequently discussed by fans and journalists. An often discussed phenomenon in several sports is the "hot hand", meaning that players may enter a state where they experience extraordinary success. For example, the former German football player Gerd Müller potentially was in a "hot" state when scoring 11 penalties in a row between 1975 and 1976. However, with 3 penalties missed in a row earlier in 1971, he potentially was in a "cold" state when taking these penalty kicks.

Academic research on the hot hand started by *Gilovich et al.* (1985). In their seminal paper, they analysed basketball free-throw data and provided no evidence for the hot hand, arguing that people tend to belief in the hot hand due to memory bias. In the past decade, however, some studies provided evidence for the hot hand while others failed to find such an effect (see *Bar-Eli et al.*, 2006, for a review). The existence of a hot hand effect thus remains an open question.

In our analysis, we investigate a potential "hot shoe" effect of penalty takers in the German Bundesliga. Our data set comprises all penalties taken in the Bundesliga from the first season (1963/64) until season 2016/17, totaling in $n = 3{,}482$ observations. Specifically, to explicitly account for the underlying (latent) form of a player, we consider HMMs to investigate a potential hot shoe effect. Using HMMs to investigate the hot hand was first done by *Albert* (1993) for an analysis in baseball, but also more recently by *Green and Zwiebel* (2017) who also analyse data from baseball and by *Ötting et al.* (2020b) who analyse data from darts.

There are several potential confounding factors when analysing the outcome of penalty kicks, such as the score of the match and the abilities of the two involved players, i.e. the penalty taker and the opposing team's goal keeper. Accounting for

these factors leads to a large number of covariates, some of them also exhibiting a noteworthy amount of correlation/multicollinearity, which makes model fitting and interpretation of parameters difficult. Hence, sparser models are desirable. To tackle these problems, variable selection is performed here by applying a LASSO penalisation approach (see *Tibshirani*, 1996). Our results suggest two different states, which can be tied to a cold and a hot state. In addition, the results shed some light on exceptionally well-performing goalkeepers.

The remainder of the chapter is structured as follows. The data on penalty kicks from the German Bundesliga is described in Section 6.2. In Section 6.3 the methodology considered is presented, namely a LASSO penalisation technique for HMMs. The proposed approach is further investigated in a short simulation study in Section 6.4 and the results of our hot shoe analysis are presented in Section 6.5.

6.2 Data

The data set considered comprises all penalty kicks taken in the German Bundesliga from its first season 1963/1964 until the end of the season 2016/2017. Parts of the data have already been used in *Bornkamp et al.* (2009). In the analysis, we include all players who took at least 5 penalty kicks during the time period considered, resulting in $n = 3{,}482$ penalty kicks taken by 310 different players. For these penalty kicks considered, 327 different goalkeepers were involved. The resulting variable of interest is a binary variable indicating whether the player scored the penalty or not. Hence, we consider binary time series $\{y_{p,t}\}_{t=1,\dots,T_p}$, with T_p denoting the total number of penalties taken by player p, indicating whether player p scored the penalty at attempt t, i.e.:

$$y_{p,t} = \begin{cases} 1, & \text{if the } t\text{–th penalty kick is scored;} \\ 0, & \text{otherwise.} \end{cases}$$

Since several other factors potentially affect the outcome of a penalty kick (such as the score of the match), we consider further covariates. For the choice of covariates, we follow *Dohmen* (2008), who analysed the effect of pressure when taking penalty kicks and, hence, accounts for several potential confounders. These additional covariates include a dummy indicating whether the match was played at home, the matchday, the minute where the penalty was taken, the experience of both the penalty taker and

the goalkeeper (quantified by the number of years played for a professional team) and the categorised score difference, with categories more than 2 goals behind, 2 goals behind, 1 goal behind, 1 goal ahead, 2 goals ahead, or more than 2 goals ahead. Since the effect of the score might depend on the minute of the match, we further include interaction terms between the categories of the score difference and the minute. To consider rule changes for penalty kicks (see *Dohmen*, 2008, for more details), we include dummy variables for different time intervals (season 1985/86 and before, between season 1986/87 and season 1995/96, season 1996/1997, and from season 1997/1998 up to season 2016/2017). Table 6.1 summarises descriptive statistics for all metric covariates considered as well as for our response variable.

Table 6.1: Descriptive statistics.

	mean	st. dev.	min.	max.
successful penalty	0.780	0.414	0	1
matchday	–	–	1	38
home	0.316	0.465	0	1
experience (penalty taker)	6.323	3.793	0	19
experience (goalkeeper)	5.343	4.187	0	19
minute	51.92	24.91	1	90

Finally, to explicitly account for player-specific characteristics, we include intercepts for all penalty takers as well as for all goalkeepers considered in our sample. These parameters can be interpreted as the players' penalty abilities (i.e., the penalty shooting skill for the penalty taker and the (negative) penalty saving skill for the goalkeeper). To illustrate the typical structure of our data, an example time series from our sample of the famous German attacker Gerd Müller, who played in the Bundesliga for Bayern Munich from 1964 until 1979, is shown in Figure 6.1. The corresponding part in the data set is shown in Tables 6.2 and 6.3.

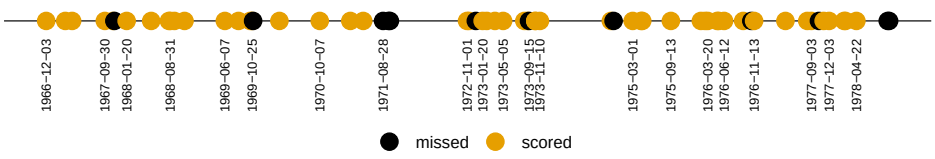


Figure 6.1: Penalty history over time of the player Gerd Müller for the time period from 1964 until 1979; a successful penalty is shown in yellow, a failure in black.

Table 6.2: Part of the data set corresponding to the metric covariates.

player	successful penalty	matchday	home	experience (penalty taker)	experience (goalkeeper)	minute	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯
Gerd Müller	1	15	0	1	3	90	⋯
Gerd Müller	1	25	0	1	3	81	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯
Gerd Müller	0	7	0	13	2	37	⋯
Gerd Müller	0	8	1	13	8	68	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

Table 6.3: Part of the data set corresponding to the player- and goalkeeper specific effects.

Hans Müller (player)	Gerd Müller (player)	Ludwig Müller (player)	⋯	Günter Bernard (goalkeeper)	Wolfgang Schnarr (goalkeeper)	⋯	Dieter Burdenski (goalkeeper)	Wolfgang Kneib (goalkeeper)
⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋮
0	1	0	⋯	0	1	⋯	0	0
0	1	0	⋯	1	0	⋯	0	0
⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋮
0	1	0	⋯	0	0	⋯	0	1
0	1	0	⋯	0	0	⋯	1	0
⋮	⋮	⋮	⋯	⋮	⋮	⋯	⋮	⋱

6.3 Methods

Figure 6.1 indicates that there are phases in the career of Gerd Müller where he scored several penalty kicks in a row, e.g. between 1975 and 1976 as discussed in the introduction. At some parts of his career, however, successful penalty kicks were occasionally followed by one or more missed penalty kicks. To explicitly account for such phases we consider HMMs, where the latent state process can be interpreted as the underlying varying form of a player. Moreover, *Stone* (2012) argues that HMMs are more suitable for analysing the hot hand than analysing serial correlation of outcomes, since the latter mentioned outcomes are only noisy measures of the underlying (latent) form of a player.

6.3.1 Hidden Markov models

In HMMs, the observations $y_{p,t}$ are assumed to be driven by an underlying state process $s_{p,t}$, in a sense that the $y_{p,t}$ are generated by one of N distributions according to the Markov chain. In our application, the state process $s_{p,t}$ serves for the underlying varying form of a player. For notational simplicity, we drop the player-specific subscript p in the following. Switching between the states is taken into account by the transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$, with $\gamma_{ij} = \Pr(s_t = j | s_{t-1} = i), i, j = 1, \ldots, N$.

We further allow for additional covariates at time t, $\mathbf{x}_t = (x_{1t}, \ldots, x_{Kt})'$, each of which assumed to have the same effect in each state, whereas the intercept is assumed to vary across the states, leading to the following linear state-dependent predictor:

$$\eta_t^{(s_t)} = \beta_0^{(s_t)} + \beta_1 x_{1t} + \ldots + \beta_k x_{Kt}.$$

In fact, this is a simple Markov-switching regression model, where only the intercept varies across the states (see, e.g., *Goldfeld and Quandt*, 1973). The dependence structure of the HMM considered is shown in Figure 6.2.

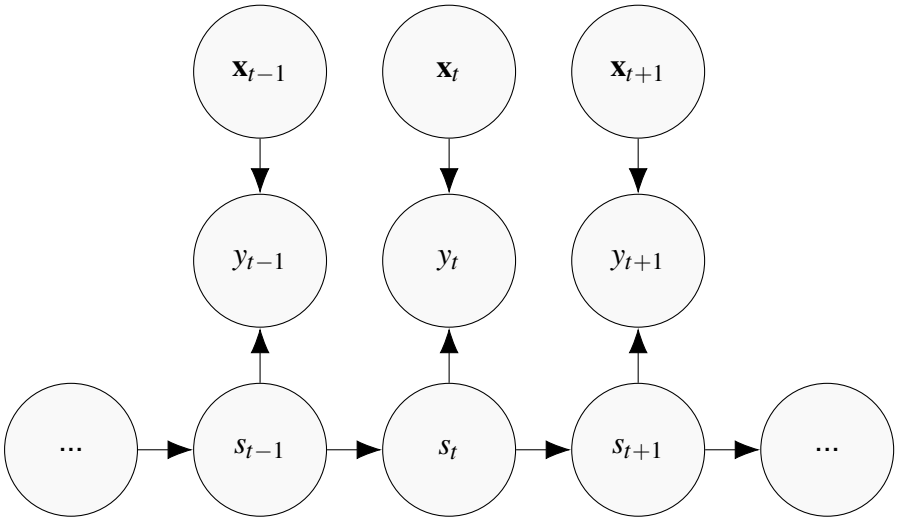


Figure 6.2: Dependence structure of the HMM considered. Each observation y_t is assumed to be generated by one of N distributions according to the state process s_t, which serves for the underlying form of a player. In addition, covariates $\mathbf{x}_t$ are assumed to affect y_t.

For our response variable y_t, indicating whether the penalty attempt t was successful or not, we assume $y_t \sim \text{Bern}(\pi_t^{(s_t)})$ and link $\pi_t^{(s_t)}$ to our state-dependent linear predictor $\eta_t^{(s_t)}$ using the logit link function, i.e. $\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)}$. Defining an $N \times N$ diagonal matrix $\mathbf{P}(y_t)$ with i–th diagonal element equal to $\Pr(y_t \,|\, s_t = i)$ and assuming that the initial distribution $\boldsymbol{\delta}$ of a player is equal to the stationary distribution, i.e. the solution to $\boldsymbol{\Gamma\delta} = \boldsymbol{\delta}$ subject to $\sum_{i=1}^{N} \delta_i = 1$, the likelihood for a single player p is given by

$$L_p(\boldsymbol{\alpha}) = \boldsymbol{\delta}\mathbf{P}(y_{p,1})\boldsymbol{\Gamma}\mathbf{P}(y_{p,2})\ldots\boldsymbol{\Gamma}\mathbf{P}(y_{p,T_p})\mathbf{1},$$

with column vector $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^N$ (see *Zucchini et al.*, 2016) and parameter vector $\boldsymbol{\alpha} = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1N}, \ldots, \gamma_{NN}, \beta_0^{(1)}, \ldots, \beta_0^{(N)}, \beta_1, \ldots, \beta_k)'$ collecting all unknown parameters. Specifically, formulating the likelihood as above amounts to running the forward algorithm, which allows to calculate the likelihood recursively at computational cost $\mathcal{O}(TN^2)$ only, thus rendering numerical maximisation of the likelihood feasible

(*Zucchini et al.*, 2016). To obtain the full likelihood for all 310 players considered in the sample, we assume independence between the individual players such that the likelihood is calculated by the product of the individual likelihoods of the players:

$$L(\boldsymbol{\alpha}) = \prod_{p=1}^{310} L_p(\boldsymbol{\alpha}) = \prod_{p=1}^{310} \boldsymbol{\delta}\mathbf{P}(y_{p,1})\boldsymbol{\Gamma}\mathbf{P}(y_{p,2})\ldots\boldsymbol{\Gamma}\mathbf{P}(y_{p,T_p})\mathbf{1}.$$

For our analysis of a potential hot shoe effect, we initially select $N = 2$ states, which potentially are aligned to a "hot" and a "cold" state, i.e. states with superior and poor performance, respectively. The parameter vector $\boldsymbol{\alpha}$, hence, reduces to $\boldsymbol{\alpha} = (\gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \beta_0^{(1)}, \beta_0^{(2)}, \beta_1, \ldots, \beta_k)'$. The choice of N will be further discussed in Section 6.5.

Parameter estimation is done by maximising the likelihood numerically using `nlm()` in R (*R Core Team*, 2019). However, if we consider all covariates introduced in our model formulation from Section 6.2, the model gets rather complex, is hard to interpret, and multicollinearity issues might occur. Hence, we propose to employ a penalised likelihood approach based on a LASSO penalty, which is described in the next subsection.

6.3.2 Variable selection by the LASSO

To obtain a sparse and interpretable model, the estimation of the covariate effects will be performed by a regularised estimation approach. The idea is to first set up a model with a rather large number of possibly influential variables (in particular, with regard to the player-specific ability parameters) and then to regularise the effect of the single covariates. This way, the variance of the parameter estimator is diminished and, hence, usually lower prediction error is achieved than with the unregularised maximum likelihood (ML) estimator. The basic concept of regularisation is to maximise a penalised version of the likelihood $\ell(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha}))$. More precisely, one maximises the penalised log-likelihood

$$\ell_{\text{pen}}(\boldsymbol{\alpha}) = \log(L(\boldsymbol{\alpha})) - \lambda \sum_{k=1}^{K} |\beta_k|, \tag{6.1}$$

where λ represents a tuning parameter, which controls the strength of the penalisation. The optimal value for this tuning parameter has to be chosen either by cross-validation

or suitable model selection criteria. The latter usually constitute a compromise between the model fit (e.g., in terms of the likelihood) and the complexity of the model. Frequently used are the AIC or BIC. In the context of LASSO, the effective degrees of freedom for the AIC and BIC are estimated as the number of non-zero coefficients (see *Zou et al.*, 2007). Since our longitudinal data structure with multiple short time series from 310 individuals renders cross-validation rather difficult, we select the tuning parameter λ in the following by information criteria.

Note that in contrast to the ridge penalty, which penalises the squared coefficients and shrinks them towards zero (see *Hoerl and Kennard*, 1970), the LASSO penalty on the absolute values of the coefficients, first proposed by *Tibshirani* (1996), can set coefficients to exactly zero and, hence, enforces variable selection. Another advantage of the employed penalisation is the way correlated predictors are treated. For example, if two (or more) predictors are highly correlated, parameter estimates are stabilised by the penalisation. In such scenarios, the LASSO penalty tends to include only one of the predictors and only includes a second predictor if it entails additional information for the response variable. Therefore, if in our case study variables possibly contain information on the outcome of the penalty, they can be used simultaneously.

To fully incorporate the LASSO penalty in our setting, the non differentiable L_1 norm $|\beta_k|$ in Eq. (6.1) is approximated as suggested by *Oelker and Tutz* (2017). Specifically, the L_1 norm is approximated by $\sqrt{(\beta_k + c)^2}$, where c is a small positive number (say c $= 10^{-5}$). With the approximation of the penalty, the corresponding likelihood is still maximised numerically using `nlm()` in R as denoted above.

In the simulation study from the subsequent section, we also investigate a relaxed LASSO-type version of our fitting scheme. The relaxed LASSO (*Meinshausen*, 2007) is known to often produce sparser models with equal or lower prediction loss than the regular LASSO. To be more precise, for each value of the tuning parameter λ, in a final step we fit an (unregularised) model that includes only the variables corresponding to the non-zero parameters of the preceding LASSO estimates.

6.4 A short simulation study

We consider a simulation scenario similar to our real-data application, with a Bernoulli-distributed response variable, an underlying two-state Markov chain and 50 covariates,

47 of which being noise covariates:

$$y_t \sim \text{Bern}(\pi_t^{(s_t)}),$$

with

$$\text{logit}(\pi_t^{(s_t)}) = \eta_t^{(s_t)} = \beta_0^{(s_t)} + 0.5 \cdot x_{1t} + 0.7 \cdot x_{2t} - 0.8 \cdot x_{3t} + \sum_{j=4}^{47} 0 \cdot x_{jt}\,.$$

We further set $\beta_0^{(1)} = \text{logit}(0.75), \beta_0^{(2)} = \text{logit}(0.35)$ and

$$\mathbf{\Gamma} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}.$$

The covariate values were drawn independently from a uniform distribution within the interval $[-2,2]$. The interval boundaries as well as the corresponding effects β_1, β_2, and β_3 were chosen such that reasonable values for the response are obtained (i.e., moderate proportions of ones and zeros). We conduct 100 simulation runs, in each run generating $T = 5100$ observations $y_t, t = 1, \ldots, 5100$, from the model specified above, with the sample size being about the same size as for the real data application. Out of these 5100 simulated observations, the first 5000 are used for model fitting, whereas for the last 100 observations (which are denoted by y_t^{test}), the predictive performance of several different models is compared (see below).

For the choice of the tuning parameter λ, we consider a (logarithmic) grid of length 50, $\Lambda = \{5000, \ldots, 0.0001\}$. To compare the performance of the above described LASSO-type estimation, we consider the following five fitting schemes:

- HMM without penalisation (i.e., with $\lambda = 0$)
- LASSO-HMM with λ selected by AIC
- LASSO-HMM with λ selected by BIC
- relaxed-LASSO-HMM with λ selected by AIC
- relaxed-LASSO-HMM with λ selected by BIC

For all five methods considered, we calculate the mean squared error (MSE) of the coefficients $\beta_1, \ldots, \beta_{50}$:

$$\text{MSE}_{\boldsymbol{\beta}} = \frac{1}{50}\sum_{k=1}^{50}(\hat{\beta}_k - \beta_k)^2.$$

We also calculate the MSE for the state-dependent intercepts $\beta_0^{(1)}$ and $\beta_0^{(2)}$, and for the entries of the t.p.m., γ_{11} and γ_{22}, which is done analogously to the MSE for $\beta_1, \ldots, \beta_{50}$ shown above. To further compare the predictive performance of the fitting schemes considered, we predict the distribution for each of the 100 out-of-sample observations, i.e. the success probabilities $\hat{\pi}_t^{\text{pred}}$, and compare these to the 100 remaining simulated observations y_t^{test}. For that purpose, we calculate the Brier score and the average predicted probability, which are given as follows:

$$B = \frac{1}{100}\sum_{t=1}^{100}(\hat{\pi}_t^{\text{pred}} - y_t^{\text{test}})^2$$

$$A = \frac{1}{100}\sum_{t=1}^{100}\left(\hat{\pi}_t^{\text{pred}}\mathbb{1}_{\{y_t^{\text{test}}=1\}} + (1-\hat{\pi}_t^{\text{pred}})\mathbb{1}_{\{y_t^{\text{test}}=0\}}\right),$$

with $\mathbb{1}_{\{.\}}$ denoting the indicator function. For the Brier score B, more accurate predictions correspond to lower values, with the lowest possible value being 0. For the average predicted probability A, higher values correspond to more precise predictions. In addition, the average predicted probability can be directly interpreted as the probability for a correct prediction.

The boxplots showing the MSEs over the 100 simulation runs, and the boxplots showing the Brier score and the average predicted probability for the predictive performance, are shown in Figures 6.3 and 6.4, respectively. In both figures, the HMM without penalisation is denoted by "MLE", the LASSO-HMM with λ selected by AIC and BIC are denoted by "AIC" and "BIC", respectively, and the relaxed-LASSO-HMM with λ selected by AIC and BIC are denoted by "AIC relaxed" and "BIC relaxed", respectively. For the state-dependent intercepts $\beta_0^{(1)}$ and $\beta_0^{(2)}$, we see that the median MSE for the models with λ chosen by the BIC is fairly high compared to all other models considered. A similar behaviour is observed for the entries of the t.p.m., γ_{11} and γ_{22}.

The middle row in Figure 6.3 shows the MSE for $\beta_1, \ldots \beta_{50}$ as well as the corresponding true and false positive rates (denoted by TPR and FPR, respectively). The simulation results indicate that the non-noise covariates are detected by all models, whereas especially the LASSO-HMM with λ selected by the AIC detects several noise

covariates. A fairly low number of noise coefficients is selected by the relaxed-LASSO-HMM fitting scheme with λ selected by the AIC and by the LASSO-HMM with λ selected by the BIC. The corresponding medians for the FPR are 0.149 and 0.170, respectively. The most promising results are given by the relaxed-LASSO-HMM with λ chosen by the BIC. In 84 out of 100 simulations, no noise covariates were selected by this model.

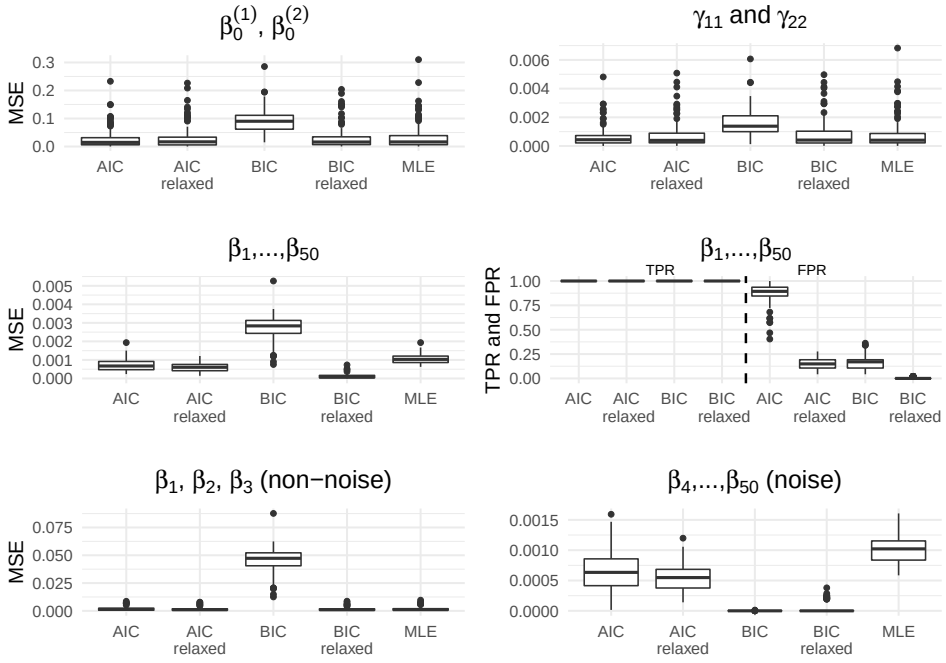


Figure 6.3: Boxplots of the MSE, TPR, and FPR obtained in 100 simulation runs. "AIC" and "BIC" denote the LASSO-HMM fitting scheme with λ chosen by AIC and BIC, respectively. "AIC relaxed" and "BIC relaxed" denote the relaxed-LASSO-HMM fitting scheme with λ chosen by AIC and BIC, respectively. "MLE" denots the HMM without penalisation.

The left plot in the middle row of Figure 6.3 shows that the median MSE of the coefficients β_j for the LASSO-HMM with λ selected by the BIC is higher than the median MSE of all other models considered. This arises since the BIC tends to select a rather high λ, which can be seen from the median MSE separated for the noise and non-noise coefficients (last row of Figure 6.3). With the fairly high λ chosen by the BIC, i.e. with more shrinkage involved, the MSE for the non-noise coefficients is rather large. At the same time, since only a few covariates are selected with a rather high λ, the MSE for the noise coefficients is very low.

For the predictive performance of the methods considered, we see that visually there is no clear difference in the Brier score between the models, which is shown in the top panel of Figure 6.4. The result for the average predicted probability — shown in the bottom panel of Figure 6.4 — confirm that the LASSO-HMM with λ chosen

by the BIC and the HMM without penalisation perform worse than the other models considered.

The results of the simulation study are very encouraging, with the LASSO penalty allowing for variable selection. The performance in terms of MSE, TPR, and FPR suggest that the LASSO-HMM with λ selected by the BIC performs worst, with the MSE being higher than for the HMM without penalisation. However, the LASSO-HMM with λ selected by the AIC as well as the relaxed-LASSO-HMM with λ selected by AIC and BIC, respectively, (partly substantially) outperform the HMM without penalisation in terms of MSE. The relaxed-LASSO-HMM with λ selected by the BIC performs best in terms of MSE, TPR, FPR and the predictive performance. Finally, in all simulation runs the overall pattern was captured with regard to the true underlying state-dependent intercepts $\beta_0^{(1)}, \beta_0^{(2)}$ and diagonal entries of the t.p.m., i.e. γ_{11} and γ_{22}. Fitting the LASSO-HMM and the relaxed-LASSO-HMM on the grid containing 50 different tuning parameters λ took on average 47 minutes using a 3.4 GHz Intel© Core™ i7 CPU. This is remarkably fast, considering that both the LASSO-HMM and the relaxed-LASSO-HMM are fitted to the data for each value of λ, leading to 100 fitted models in total.

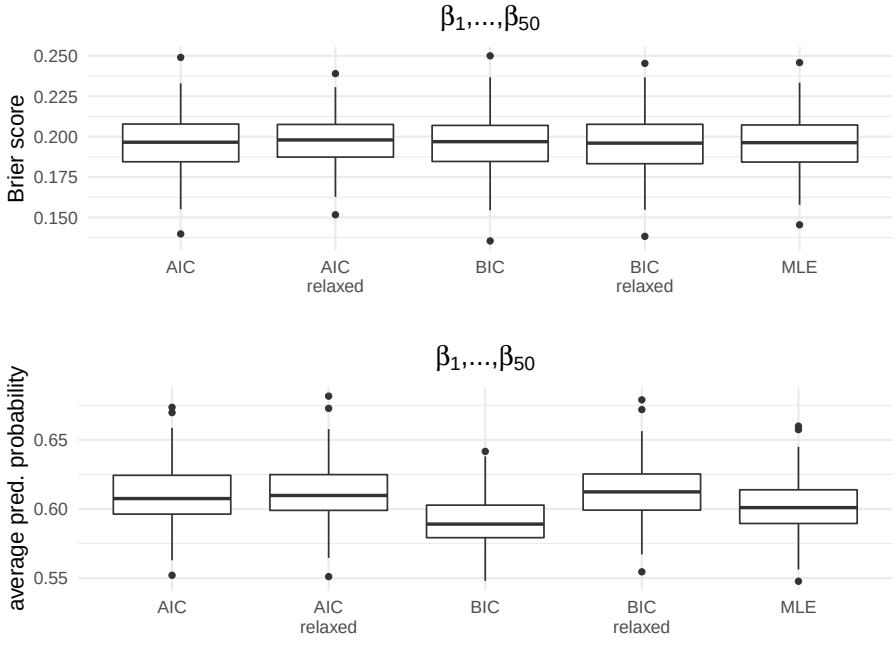


Figure 6.4: Boxplots of the Brier score (top panel) and the average predicted probability (bottom panel) obtained in 100 simulation runs."AIC" and "BIC" denote the LASSO-HMM fitting scheme with λ chosen by AIC and BIC, respectively. "AIC relaxed" and "BIC relaxed" denote the relaxed-LASSO-HMM fitting scheme with λ chosen by AIC and BIC, respectively. "MLE" denots the HMM without penalisation.

6.5 Results

We now apply our LASSO-HMM approach to the German Bundesliga penalty data. For the analysis of a potential hot shoe effect, we include all covariates from Section 6.2 into the predictor and choose $N = 2$ for the number of states, representing potential hot and cold states, respectively.[1] This yields the following linear state-dependent predictor[2]:

$$\begin{aligned}\text{logit}(\pi_t^{(s_t)}) = \beta_0^{(s_t)} + \beta_1 \text{home}_t + \beta_2 \text{minute}_t + \ldots + \\ \beta_{100} \text{GerdMueller}_t + \ldots + \beta_{656} \text{WolfgangKneib}_t \,.\end{aligned}$$

Since the simulation study above indicates that the unpenalised maximum likelihood estimator is not appropriate for such a large number of covariates, we only consider the LASSO-type fitting schemes for the real data application. Specifically, since the relaxed-LASSO-HMM with λ selected by the BIC showed the most promising results in the simulation, we focus on the results obtained by this fitting scheme, but we also present the results obtained by the two LASSO-HMM fitting schemes.[3] The parameter estimates obtained (on the logit scale) indicate that the baseline level for scoring a penalty is higher in the model's state 1 than in state 2 ($\hat{\beta}_0^{(1)} = 1.422 > \hat{\beta}_0^{(2)} = -14.50$). However, the relevance of these results regarding a potential hot shoe effect is discussed in Section 5.5. With the t.p.m. estimated as

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.978 & 0.022 \\ 0.680 & 0.320 \end{pmatrix},$$

there is high persistence in state 1, whereas state 2 is a transient state, where switching to state 1 is most likely. In addition, the model is slightly favoured by the AIC over a single-state model, i.e. a standard logit model without a potential hot shoe effect ($\text{AIC}_{\text{hotshoe}} = 3664$, $\text{AIC}_{\text{logit}} = 3670$). The stationary distribution as implied by the estimated t.p.m. is $\hat{\boldsymbol{\delta}} = (0.969, 0.031)$, i.e., according to the fitted model, players are in about 96.9% of the time in state 1 and in about 3.1% in state 2. The diag-

[1] A psychological reason for being hot or cold may be a higher/lower level of self confidence.

[2] Throughout this chapter, all metric covariates are considered as linear. Since the main interest of this chapter is to investigate the LASSO penalty in HMMs, future research on the hot shoe could focus also on non-linear effects, for example for the matchday and the minute.

[3] The relaxed-LASSO-HMM with optimal λ selected by the AIC yielded a rather unreasonable model, where almost all of the more than 600 covariates were selected and with partly unrealistically large corresponding estimated covariate effects, indicating some tendency of overfitting. For this reason, we excluded this model from the analysis.

onal elements of the t.p.m. for the other fitting schemes are obtained as $\hat{\gamma}_{11,\text{AIC}} = 0.989$, $\hat{\gamma}_{22,\text{AIC}} = 0.386$ and $\hat{\gamma}_{11,\text{BIC}} = 0.987$, $\hat{\gamma}_{22,\text{BIC}} = 0.368$, respectively. The corresponding state-dependent intercepts are obtained as $\hat{\beta}^{(1)}_{0,\text{AIC}} = -14.71$, $\hat{\beta}^{(2)}_{0,\text{AIC}} = 1.347$ and $\hat{\beta}^{(1)}_{0,\text{BIC}} = -18.83$, $\hat{\beta}^{(2)}_{0,\text{BIC}} = 1.360$, respectively. The results of the other fitting schemes considered, hence, are fairly similar to those obtained by the relaxed-LASSO-HMM with λ selected by the BIC.

For the grid of potential tuning parameters λ, Figure 6.5 shows the progress of the AIC and BIC for the LASSO-HMM, indicating that the AIC selects a lower tuning parameter than the BIC. The corresponding coefficient paths of the LASSO-HMM approach together with the associated optimal tuning parameters selected by the AIC and BIC, respectively, are shown in Figure 6.6.[4] No covariates are selected by the LASSO-HMM with λ selected by the BIC, whereas Jean-Marie Pfaff, a former goalkeeper of Bayern Munich, is selected by the LASSO-HMM with λ selected by the AIC and by the relaxed-LASSO-HMM based on BIC (see Table 6.4). The negative coefficient implies that the odds for scoring a penalty decrease if Jean-Marie Pfaff is the goalkeeper of the opponent's team. The coefficient is substantially larger in magnitude for the relaxed-LASSO model, since for this fitting schemes the model is re-fitted with $\lambda = 0$ on the set of selected coefficients from the first model fit (see above).

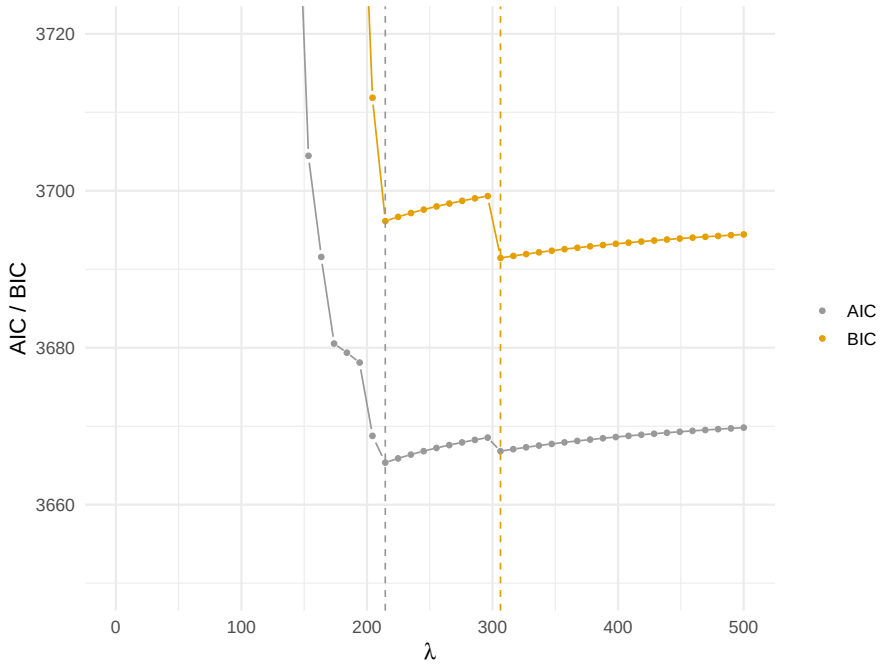


Figure 6.5: Paths of AIC and BIC in the LASSO-HMM models. The vertical lines indicate the optimal penalty parameters λ selected by AIC and BIC, respectively.

[4] We abstain from showing the coefficient paths plot for the relaxed LASSO-HMM model, because due to the unpenalised re-fit the paths look rather irregular.

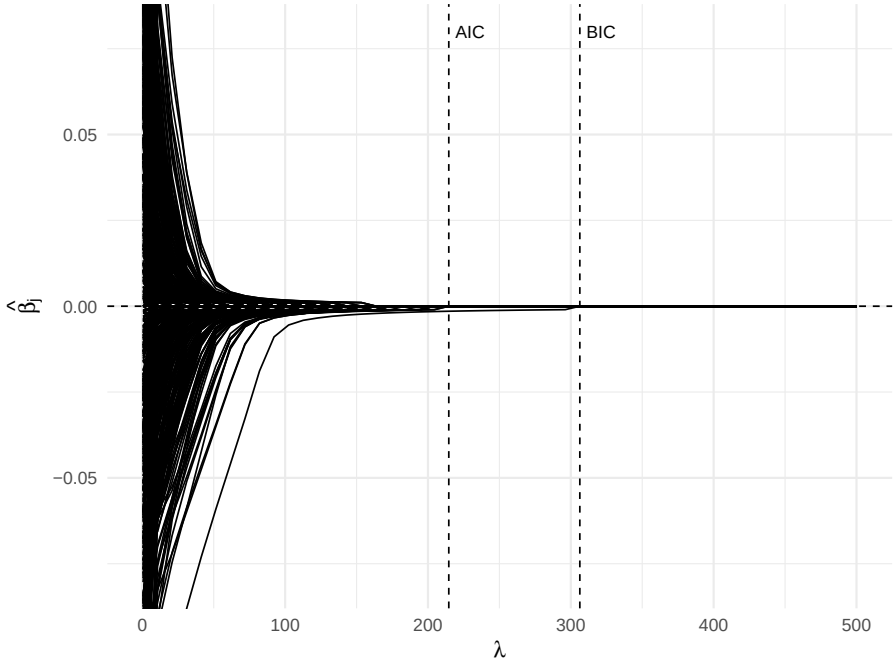


Figure 6.6: Coefficient paths of all covariates considered in the LASSO-HMM models. The dashed lines indicate the optimal penalty parameters λ selected by AIC and BIC, respectively. For the λ selected by the BIC, no covariates are selected, whereas for the AIC one covariate is selected (see also Table 6.4).

Table 6.4: Overview of selected players and goalkeepers by all models considered.

	BIC	**AIC**	**BIC relaxed**
Jean-Marie Pfaff (goalkeeper)	0.000	-0.001	-0.125
Rudolf Kargus (goalkeeper)	0.000	0.000	0.000
⋮	⋮	⋮	⋮
Manuel Neuer (goalkeeper)	0.000	0.000	0.000
Lothar Matthäus (player)	0.000	0.000	0.000
⋮	⋮	⋮	⋮
Nuri Sahin (player)	0.000	0.000	0.000

6.6 Discussion

The modelling framework developed in this chapter, a (relaxed-)LASSO-HMM, allows for implicit variable selection in the state-dependent process of HMMs. The performance of the variable selection is first investigated in a simulation study, indicating that the relaxed-LASSO-HMM with the corresponding tuning parameter selected by the BIC is the best-performing fitting scheme considered.

For the analysis of a hot shoe effect, we fit both LASSO-HMMs and relaxed-LASSO-HMMs to data on penalty kicks in the German Bundesliga. Factors potentially

affecting the performance of penalty-takers, such as the current score of the match, are included in the predictor. In addition, dummy variables for the penalty-takers as well as for the goalkeepers are included. Our results suggest two states with different levels of performance, and shed some light on exceptionally performing players such as Jean-Marie Pfaff, a former goalkeeper of Bayern Munich, who has been selected by our fitting schemes.

A clear limitation of the real data application considered is the problem of self selection. Since the manager (or the team) can decide which player has to take the penalty, players who have been rather unsuccessful in the past may not take penalty kicks anymore. However, several teams have demonstrated in the past that they rely on and trust in a certain player for taking penalty kicks, regardless of the outcome of the kick. Whereas penalty kicks in football yield to a time series due to the way in which penalties take place, the corresponding time intervals between actions are irregular. Although our data cover all attempts in the German Bundesliga, there are sometimes several month between two attempts. Moreover, some players might be involved in penalty kicks in matches from other competitions such as the UEFA Champions League, the UEFA European Cup, or in matches with their national teams. From this perspective, the time series of Bundesliga penalties could be considered as partly incomplete for some players.

From a methodological point, the number of states selected (i.e., $N = 2$) may be too coarse for modelling the underlying form of a player. Considering a continuously varying underlying state variable instead may be more realistic, since gradual changes in a player's form could then be captured. This could be achieved by considering models with an underlying continuous state process, where regularised estimation approaches are a first point for further research. The motivation in this chapter for $N = 2$ states, however, was to approximate the potential psychological states in a simple manner for ease of interpretation, e.g., in the sense of hot ("player is confident") or cold ("player is nervous") states. Moreover, our main focus was to show the usefulness of our method developed in a rather simple setting with two states. Our results should, hence, be treated with caution regarding the existence of a potential hot shoe effect. Further points for future research include regularisation approaches in HMMs where not only the intercept (as considered here), but also the parameters β_j are allowed to depend on the current state. Regularisation in this model formulation could be taken into account by applying so-called fused LASSO techniques (see, e.g., *Gertheiss and Tutz*,

2010), where the parameters could either be shrunk to zero or to the same size for all states considered.

The modelling framework developed here can easily be tailored to other applications, where implicit variable selection in HMMs is desired. For the application considered in this chapter, i.e. an analysis of a potential hot hand/hot shoe effect, other sports such as basketball or hockey could be analysed. Potential covariates — whose corresponding effects are penalised — in these sports cover the shot type, shot origin, and game score, to name but a few.

7 A copula-based multivariate hidden Markov model for modelling momentum in football

7.1 Introduction

Sports commentators and fans frequently use vocabulary such as "momentum", "momentum shift", or related terms to refer to change points in the dynamics of a match. Usage of such terms is typically associated with situations during a match where an event — such as a shot hitting the woodwork in a football match — seems to change the dynamics of the match, e.g. in a sense that a team which prior to the event had been pinned back in its own half suddenly seems to dominate the match. A prominent example is the 2005 Champions League final between Milan and Liverpool, within which Liverpool was trailing by three goals after the first half, but fought back after half time and eventually won by penalty shootout.

Despite the widespread belief in momentum shifts in sports, it is not always clear to what extent *perceived* shifts in the momentum are genuine. From the literature on the "hot hand" — i.e. research on serial correlation in human performances — it is well known that most people do not have a good intuition of randomness, and in particular tend to overinterpret streaks of success and failure, respectively (see, e.g., *Kahneman*, 2011; *Thaler and Sunstein*, 2009). It is thus to be expected that many perceived momentum shifts are in fact cognitive illusions in the sense that the observed shift in a competition's dynamics is driven by chance only.

Momentum shifts have been investigated in qualitative psychological studies, e.g. by interviewing athletes, who reported momentum shifts during matches (see, e.g., *Jones and Harwood*, 2008; *Richardson et al.*, 1988). Fuelled by the rapidly growing amount of freely available sports data, quantitative studies have investigated the drivers of ball possession in football (*Lago-Peñas and Dellal*, 2010), the detection of main playing

styles and tactics (*Diquigiovanni and Scarpa*, 2018; *Gonçalves et al.*, 2017) and the effects of momentum on risk-taking (*Lehman and Hahn*, 2013). In some of the existing studies, e.g. in *Lehman and Hahn* (2013), momentum is not investigated in a purely data-driven way, but rather pre-defined as winning several matches in a row.

In this chapter, we analyse potential momentum shifts within football matches. Specifically, we investigate the potential occurrence of momentum shifts by analysing minute-by-minute bivariate summary statistics from the German Bundesliga using HMMs. The corresponding data is described in Section 7.2. Within the HMMs, we consider copulas to allow for within-state dependence of the variables considered. The corresponding methodology is presented in Section 7.3. Our results, which are presented in Section 7.4, suggest states which can be tied to different levels of control in a match. In addition, we investigate the causes of momentum shifts, e.g. the current score of the match. This type of insight could be of great interest to managers, bookmakers and sports fans.

7.2 Data

We analyse minute-by-minute in-game statistics of Bundesliga matches, taken from www.whoscored.com, to investigate to what extent momentum shifts in a football match are genuine, and what kind of events lead to a shift. Since the strength and tactics differ between the teams, we do not pool data from multiple teams, but consider data from a single team. Throughout this chapter, we consider data from Borussia Dortmund. In Appendix C, we present the same analysis for Hannover 96.

As proxy measures for the current momentum within a football match, we consider the number of shots on goal and the number of ball touches, with both variables sampled on a minute-by-minute basis. For match m, $m = 1, \ldots, 34$, this results in a bivariate time series $\{\mathbf{y}_{mt}\}_{t=1,2,\ldots,T_m}$, with $\mathbf{y}_{mt} = (y_{mt1}, y_{mt2})$ the pair of variables observed at time t (out of T_m minutes played) during the match.

Due to injury times being added to the regular match length of 90 minutes, the lengths of the time series considered range from 91 to 100 minutes. The final data set then comprises $n = 3{,}214$ bivariate observations from $m = 34$ matches of the season 2017/18. In addition, since the underlying dynamics of a match, from Borussia Dortmund's perspective, potentially depend on characteristics of the opponent (such as the strength of the squad) as well as events in the match (such as goals), the

following four covariates are considered:

- the market value of the opponent team (taken from www.transfermarkt.com);
- the goal difference in the current score;
- a dummy variable indicating whether the match is played at home or away;
- the current minute of the match.

The first covariate considered is a (crude) proxy for the strength of teams and does not vary for a team in the given period of time. The difference in the current score is calculated from Borussia Dortmunds point of view, i.e. positive values refer to a lead of Dortmund whereas negative values represent that Dortmund is trailing. The dummy indicating whether the match is played at home is included since several studies provided evidence for a home field advantage, because of (e.g.) crowd effects and psychological advantage when playing at home (see, e.g., *Pollard*, 2008). Finally, to account for the potential state of exhaustion of players, the minute of the match is also included. The variables considered are summarised in Table 7.1.

Table 7.1: Descriptive statistics of the variables analysed, 'shots' and 'ball touches', as well as the covariates 'market value' and 'score difference'.

	mean	st. dev.	min.	max.
shots	0.150	0.412	0	3
ball touches	6.101	5.036	0	28
market value (in 10^6 euro)	142.6	127.1	48.80	610.3
score difference	0.253	1.500	-6	5

One example bivariate time series from the data set, corresponding to the in-game statistics observed for Borussia Dortmund in the match against FC Schalke 04 played in November 2017 is shown in Figure 7.1. In the media, this match was said to have a momentum shift, since Borussia Dortmund was in a 4:0 lead at half time, but Schalke 04 scored four goals in the second half such that the match resulted in a draw.

7.3 Modelling momentum

Figure 7.1 underlines that there are periods in the match where Borussia Dortmund's number of ball touches and the number of shots on goal are fairly low (e.g. around minute 75–90), as well as periods with relatively many ball touches and shots on goal

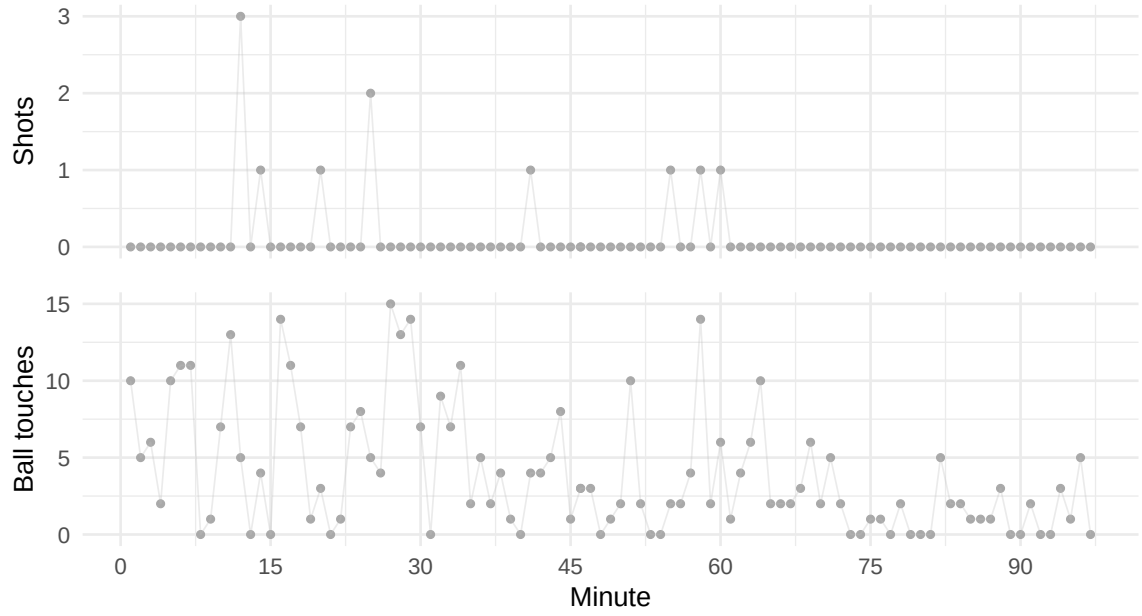


Figure 7.1: Bivariate time series of the number of shots on goal (top) and the ball touches (bottom) of Borussia Dortmund for one example match from the data set (Borussia Dortmund vs. FC Schalke 04).

(e.g. around minute 15–30). HMMs hence constitute a natural modelling approach for the minute-by-minute bivariate time series data, as they accommodate the idea of a match progressing through different phases, with potentially changing momentum. The states can be interpreted as the underlying momentum, i.e. as potentially different levels of control of the team considered. In the most simple model formulation with two states, the states could, for example, be interpreted as either the team considered or the opponent having a high level of control (i.e. dominating the match). In this section, the basic HMM model formulation will be introduced (Section 7.3.1) and extended to allow for within-state dependence using copulas (Section 7.3.2). The latter is desirable since the potential within-state dependence may lead to a more comprehensive interpretation of the states regarding the underlying momentum. Finally, for the model formulation presented in Section 7.3.2, covariates will be included (Section 7.3.3).

7.3.1 A baseline model

HMMs involve two components: an unobserved Markov chain with N possible states, and an observed state-dependent process, whose observations are assumed to be generated by one of N distributions as selected by the Markov chain. For the data considered in this chapter, the observations and the state process are denoted by $\mathbf{y}_{mt}$ and $\{s_{mt}\}_{t=1,2,\ldots,T_m}$, respectively. Switches between the state are modelled by the transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$, where $\gamma_{ij} = \Pr(s_{mt} = j | s_{m,t-1} = i)$, $i, j = 1, \ldots, N$.

Figure 7.2 shows the model structure as directed graph. For the model formulation of an HMM to be completed, the number of states N and the class(es) of state-dependent distribution(s) have to be selected. While choosing state-dependent distribution(s) is straightforward for univariate time series, it is generally not straightforward to define a multivariate distribution to allow for within-state dependence of the variables considered, unless a multivariate normal distribution can be assumed. Hence, for the vector of observations $\mathbf{y}_{mt}$, in the baseline model formulation we assume that the joint probability is obtained by the product of the marginal distributions,

$$f(\mathbf{y}_{mt}|s_{mt}) = \prod_{k=1}^{K} f(y_{mtk}\,|\,s_{mt}), \tag{7.1}$$

with $K = 2$ here. This assumption (also known as *contemporaneous conditional independence*) is often used in practice (see, e.g., *DeRuiter et al.*, 2017; *Punzo et al.*, 2018; *van Beest et al.*, 2019; *Wall and Li*, 2009). In Eq. (7.1) f denotes a p.m.f. since we deal with discrete data, but in principle f could also denote a density without any further changes in the baseline model formulation. The contemporaneous conditional independence assumption will be modified in the next subsection.

Since both the number of shots on goal and the number of ball touches are count data, the Poisson distribution would be a standard choice for either of the two variables. Here, to account for possible over- and underdispersion in the data, a Conway-Maxwell-Poisson (CMP) distribution is assumed both for the number of shots on goal and the number of ball touches, with p.m.f.

$$\Pr(X = x) = \frac{1}{Z(\lambda,\nu)}\frac{\lambda^x}{(x!)^\nu},$$

with $Z(\lambda,\nu) = \sum_{k=0}^{\infty}\lambda^k/(k!)^\nu$, $\lambda > 0$ and $\nu \geq 0$ (*Conway and Maxwell*, 1961). The CMP distribution contains some well-known discrete distributions:

- for $\nu = 1$, $Z(\lambda,\nu) = e^\lambda$, and the CMP distribution simply reduces to the ordinary Poisson(λ);
- for $\nu \to \infty$, $Z(\lambda,\nu) \to 1+\lambda$, and the CMP distribution approaches the Bernoulli with parameter $\lambda(1+\lambda)^{-1}$;

- for $\nu = 0$ and $0 < \lambda < 1$, $Z(\lambda, \nu)$ is a geometric sum

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \lambda^j = \frac{1}{1-\lambda}$$

  and, accordingly, the CMP distribution reduces to the geometric distribution $p_x = \lambda^x(1-\lambda)$;

- for $\nu = 0$ and $\lambda \geq 1$, $Z(\lambda, \nu)$ does not converge, leading to an undefined distribution.

In general, the normalising constant $Z(\lambda, \nu)$ does not reduce to such a simple closed-form expression. Asymptotic results are however available (*Gillispie and Green*, 2015).

To formulate the likelihood for the baseline model, the $i-$th diagonal element of the $N \times N$ diagonal matrix $\mathbf{P}(\mathbf{y}_{mt})$ consists of the joint probability of the observations y_{mt1} and y_{mt2} given state i, i.e. $f(y_{mt1} \,|\, s_{mt} = i) \cdot f(y_{mt2} \,|\, s_{mt} = i)$. Since the Conway-Maxwell-Poisson distribution contains an infinite sum in the normalising constant, the evaluation of the p.m.f. is not straightforward. Here, the R package `COMPoissonReg` was used for this purpose (*Sellers et al.*, 2018). Since stationarity cannot reasonably be assumed in our setting, we estimate the initial distribution $\boldsymbol{\delta} = \big(\Pr(s_{m1} = 1), \ldots, \Pr(s_{m1} = N)\big)$, regarding the parameters of $\boldsymbol{\delta}$ as $N-1$ additional parameters to be estimated. With these quantities defined, the likelihood for a single match m is given by:

$$L = \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_{m1})\boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_{m2}) \ldots \boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_{mT_m})\mathbf{1}$$

with column vector $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^N$ (see *Zucchini et al.*, 2016). Calculation of this matrix product expression amounts to the application of the forward algorithm, which is a powerful recursive technique for efficiently calculating the likelihood of an HMM at computational cost $\mathcal{O}(TN^2)$ only (*Zucchini et al.*, 2016). To obtain the likelihood for the full data set, we assume independence between the individual matches such that the likelihood is given by the product of likelihoods for the individual matches:

$$L = \prod_{m=1}^{34} \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_{m1})\boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_{m2}) \ldots \boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_{mT_m})\mathbf{1} \tag{7.2}$$

The model formulation presented here could be extended to account for momentum carry-over effects across matches, but this is not investigated in the present work since

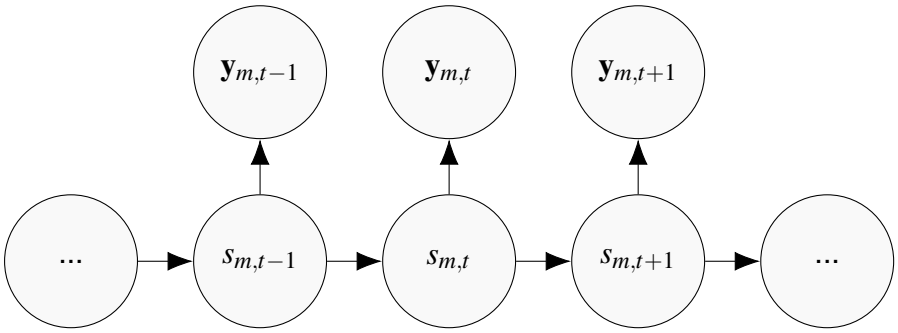


Figure 7.2: Dependence structure of the HMM considered: each pair of observations $\mathbf{y}_{mt}$ is assumed to be generated by one of N (bivariate) distributions according to the state process s_{mt}.

there is usually a time difference of 5-7 days between matches. The model parameters are estimated by numerical maximum likelihood estimation using the function `nlm()` in R (*R Core Team*, 2019). To avoid local maxima, we carefully selected starting values for the numerical maximisation by drawing random numbers from uniform distributions several times and choosing the model with the best likelihood. In addition, to speed up computation time, we implemented the forward algorithm in C++ using the R-package `Rcpp` (*Eddelbuettel*, 2013). For a model with $N = 2$ states, it takes less than a minute to numerically maximise the likelihood on a usual desktop computer.

7.3.2 Modelling within-state dependence using copulas

In the baseline model formulation, we assume contemporaneous conditional independence, i.e. that there is no within-state correlation between the two variables considered. However, when modelling momentum in football, it is of interest to explicitly model any within-state dependence to draw a comprehensive picture of the dynamics of a match. For example, high ball possession can be linked to both an attacking phase with lots of shots on goal, but also much less goal-oriented tactics, where the main aim is simply to control the match by keeping the ball, without much pressure on goal. The between-variable correlation would likely be very different in those two scenarios. By estimating the within-state correlation between the two variables, we are better able to distinguish between such fairly subtle differences in a team's style of play.

To modify the contemporaneous conditional independence assumption, a multivariate distribution needs to be assumed to specify the dependence structure between the variables considered within states. Here, we allow for within-state correlation of our variables $\mathbf{y}_{mt}$ by formulating a bivariate distribution as state-dependent distribution

using a copula. A copula is a multivariate probability distribution with uniform margins. As introduced by *Sklar* (1959), the idea of a copula is to split a multivariate distribution into its univariate margins and the dependence structure, where the latter depends on the copula considered. Within the class of HMMs, copulas have previously been used by *Härdle et al.* (2015) to model within-state dependence in financial data, and by *Brunel and Pieczynski* (2005) and *Lanchantin et al.* (2011) for image analysis. For our modelling approach, we again consider the Conway-Maxwell-Poisson both for the number of shots on goal and the number of ball touches as marginal distribution. With $F_1(y_{mt1}|s_{mt})$ and $F_2(y_{mt2}|s_{mt})$ denoting the (state-dependent) c.d.f. of the marginals, the bivariate state-dependent distribution is given by

$$F(\mathbf{y}_{mt}\,|\,s_{mt}) = C\big(F_1(y_{mt1}\,|\,s_{mt}), F_2(y_{mt2}\,|\,s_{mt})\big),$$

where $C(.,.)$ is a bivariate copula. When deriving the corresponding p.m.f., differences are needed rather than derivatives, since the marginals are discrete (see, e.g., *Nikoloulopoulos*, 2013). Thus, the bivariate p.m.f. of $\mathbf{y}_{mt}$ given state s_{mt} is given by

$$\begin{aligned} f(\mathbf{y}_{mt}\,|\,s_{mt}) = & C\big(F_1(y_{mt1}\,|\,s_{mt}), F_2(y_{mt2}\,|\,s_{mt})\big) \\ & - C\big(F_1(y_{mt1}-1\,|\,s_{mt}), F_2(y_{mt2}\,|\,s_{mt})\big) \\ & - C\big(F_1(y_{mt1}\,|\,s_{mt}), F_2(y_{mt2}-1\,|\,s_{mt})\big) \\ & + C\big(F_1(y_{mt1}-1\,|\,s_{mt}), F_2(y_{mt2}-1\,|\,s_{mt})\big). \end{aligned} \tag{7.3}$$

The copula $C(.,.)$ needs to be selected from the large number of possible copula functions available in the literature. Here, we focus on copulas that can model positive and negative dependence. Archimedean copulas (see, e.g., *Nelsen*, 2006, p. 116 for an overview) are convenient for this modelling purpose. We consider three different families of copulas, comparing their fit to the data in Section 7.4: first, the Frank-copula, which for two marginals u_1 and u_2 defined as

$$C(u_1,u_2) = -\frac{1}{\theta}\log\left(1+\frac{(\exp(-\theta u_1)-1)(\exp(-\theta u_2)-1)}{\exp(-\theta)-1}\right),$$

second, the Clayton-copula,

$$C(u_1,u_2) = \Big(\max\{u_1^{-\theta}+u_2^{-\theta}-1;0\}\Big)^{-1/\theta},$$

and third, the Ali-Mikhail-Haq (AMH) copula,

$$C(u_1,u_2)=\frac{u_1u_2}{1-\theta(1-u_1)(1-u_2)},$$

where for each copula considered the dependence parameter is denoted by θ. With these quantities defined, the diagonal matrix $\mathbf{P}(\mathbf{y}_{mt})$ in the HMM likelihood (see Eq. 7.2) changes slightly. The i–th diagonal entry is now equal to $f(\mathbf{y}_{mt}|s_{mt}=i)$ as defined in Eq. (7.3) instead of the product of the marginals. The corresponding likelihood is then again numerically maximised using the function `nlm()` in R.

7.3.3 A model including covariates

In the previous subsections, the transition probabilities γ_{ij} were assumed to be constant over time. To account for possible events which may lead to state-switching, and hence to possible momentum shifts, we modify this assumption by explicitly allowing the transition probabilities γ_{ij} to depend on covariates at time t. This is done by linking $\gamma_{ij}^{(t)}$ to covariates $x_1^{(t)},\dots,x_p^{(t)}$ using the multinomial logit link:

$$\gamma_{ij}^{(t)}=\frac{\exp(\eta_{ij}^{(t)})}{\sum_{k=1}^{N}\exp(\eta_{ik}^{(t)})}$$

with

$$\eta_{ij}^{(t)}=\begin{cases}\eta_{ij}^{(t)}=\beta_0^{(ij)}+\sum_{l=1}^{p}\beta_l^{(ij)}x_l^{(t)} & \text{if } i\neq j;\\ 0 & \text{otherwise.}\end{cases}$$

Since the transition probabilities depend on covariates, the t.p.m. $\mathbf{\Gamma}_t$ is not constant across time anymore, i.e. the Markov chain is non-homogeneous. However, the structure of the HMM likelihood as stated in Eq. (7.2) is unaffected, such that the likelihood can still be maximised numerically.

7.4 Results

In this section, the different models presented in Section 7.3 are fitted to data on the matches of Borussia Dortmund in the 2017/18 Bundesliga season. To further illustrate

the methodology, in particular for lower-ranked teams, in Appendix C we provide the results also for Hannover 96.

Baseline model

For the baseline model, we make use of the contemporaneous conditional independence assumption, cf. Eq. (7.1), initially focusing on the case of $N = 2$ states. The corresponding parameter estimates associated with the number of shots on goal are $\hat{\boldsymbol{\lambda}}_{\text{shots}} = (0.125, 0.149)$, $\hat{\boldsymbol{\nu}}_{\text{shots}} = (0.206, 0.001)$, while for the number of ball touches, they are $\hat{\boldsymbol{\lambda}}_{\text{touches}} = (0.971, 2.381)$, $\hat{\boldsymbol{\nu}}_{\text{touches}} = (0.102, 0.390)$. It is not straightforward here to compute the means of the fitted distributions due to the infinite sum in the normalising constant. *MacDonald and Bhamani* (2018) discuss several approaches and suggest to calculate the mean by $\frac{1}{Z(\lambda,\nu)} \sum_{k=0}^{d} k\lambda^k/(k!)^\nu$ using a very large d (say $d = 100$). Following this approach, the means of the number of shots on goal are 0.138 and 0.175 for states 1 and 2, respectively. For the ball touches, the means are 4.080 (state 1) and 10.104 (state 2), respectively. Thus, state 2 can be interpreted as the team considered, Borussia Dortmund, being more dominant, i.e. having a higher level of control over the match, than when being in state 1. The t.p.m. is estimated as

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.867 & 0.133 \\ 0.280 & 0.720 \end{pmatrix},$$

and the initial distribution as $\hat{\boldsymbol{\delta}} = (0.258, 0.742)$. According to the t.p.m. of the fitted model, there is some persistence in both states. Although this is the most simple model formulation considered here, the fitted model comprises interpretable states which refer to different levels of control over the match. The model can thus be regarded as a simple baseline model for capturing momentum shifts. We will now gradually increase its complexity to more fully capture the in-game dynamics.

Copula-based HMM with N = 2

To capture possible within-state correlation of the variables, a multivariate distribution needs to be considered. For Poisson marginals, the bivariate Poisson as proposed by *Karlis and Ntzoufras* (2003) would be a possible candidate. However, as discussed in Section 7.3.1, this approach would have two limitations, namely the inability to capture

overdispersion (and underdispersion), and the restriction to positive between-variable correlation. Instead we use more flexible CMP distributions for the marginals, stitching them together using a copula as described in Section 7.3.2.

First, we investigate the consequences of relaxing the contemporaneous conditional independence assumption. To this end, Figure 7.3 displays the estimated state-dependent distributions of two-state copula-based HMM formulations, using the Frank, Clayton and AMH copula, respectively. While visually there is no clear difference between the different copula functions considered, the application of the Clayton copula led to the highest likelihood of the fitted model. Compared to the baseline model, the copula-based model shows a clear improvement in the fit ($\Delta\text{AIC} = 48; \Delta\text{BIC} = 35$). The fitted state-dependent distributions can again be interpreted as Borussia Dortmund exhibiting different levels of control, with state 1 corresponding to situations where the game is balanced, whereas state 2 refers to a high level of control. As for the baseline model, there is a fairly high persistence in the states, with the diagonal elements of the t.p.m. estimated as $\hat{\gamma}_{11} = 0.852$ and $\hat{\gamma}_{22} = 0.706$.

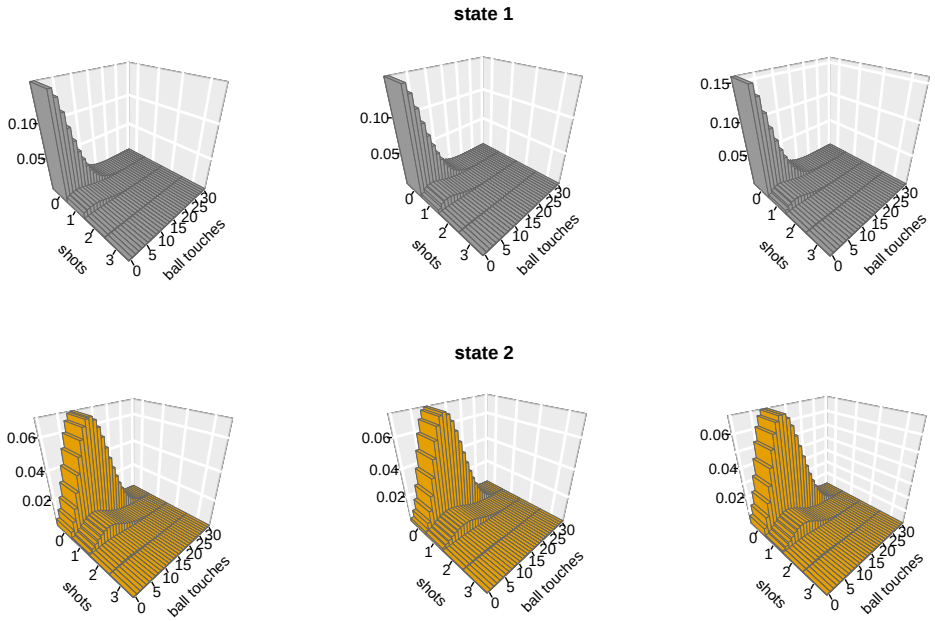


Figure 7.3: Fitted state-dependent distributions for the baseline two-state HMM for Borussia Dortmund. From left to right: Frank-, Clayton- and AMH-copula, respectively.

Choosing the number of states

For the choice of the number of states, it is anything but clear how many states a given team may exhibit in a football match. To choose an appropriate number of states, and also a copula, we first consult the AIC and the BIC for the copula-based HMMs using different numbers of states and the three copulas considered above. The corresponding results are displayed in Table 7.2. Starting with the choice of the copula, the Clayton copula is preferred by both AIC and BIC. Hence, from now on, we use the Clayton copula. Choosing the number of states is not as conclusive: according to the AIC, the five-state model is preferred, whereas the BIC selects three states. As it is well-known that the AIC tends to select too many states in a HMM (see *Pohle et al.*, 2017), a choice of $N = 3$ seems more appropriate based on these formal criteria. To make an informed choice based also on interpretability of the resulting model states, in Figure 7.4 we further inspect the fitted models with three and four states, respectively, by means of their estimated state-dependent distributions. Figure 7.4 illustrates that the general patterns of the state-dependent distributions from the three-state model are also included in the four-state model, whereas the state-dependent distribution of state 2 in the four-state model seems to refer to an underlying level of control which is not included in the three-state model. However, at closer inspection of the distributional shapes in the four-state model, there is a substantial overlap between the state-dependent distributions of state 2 and state 3, respectively. Hence, given that the BIC points to the three-state model, and since we do not see meaningful additional information in a potential fourth state, from now on we focus exclusively on three-state models.

Copula-based HMM with N = 3

For the Clayton-copula HMM with three states, Table 7.3 displays the estimated parameters of the marginal distributions as well as the dependence parameter of the copula. Deriving the corresponding means for the marginal distributions as described above yields means for the number of shots of 0.226, 0.132 and 0.147 for state 1, 2 and 3, respectively. For the number of ball touches, the corresponding means are 2.032 (state 1), 4.583 (state 2) and 9.732 (state 3). Based on the means and the corresponding distributional shapes (see top row in Figure 7.4), the different states

Table 7.2: AIC and BIC for copula-based HMMs with different numbers of states.

	Frank		Clayton		AMH	
	AIC	BIC	AIC	BIC	AIC	BIC
2 states	20,954	21,033	20,941	21,020	20,943	21,022
3 states	20,865	**21,005**	20,839	**20,979**	20,861	**21,001**
4 states	20,836	21,049	20,817	21,030	**20,831**	21,043
5 states	**20,814**	21,112	**20,801**	21,098	20,834	21,132

Table 7.3: Parameter estimates for the state-dependent distributions of the Clayton-copula HMM with three states.

	state 1	state 2	state 3
shots on goal	$\hat{\lambda} = 0.212, \hat{\nu} = 0.631$	$\hat{\lambda} = 0.117, \hat{\nu} \approx 0$	$\hat{\lambda} = 0.128, \hat{\nu} = 0.002$
ball touches	$\hat{\lambda} = 0.670, \hat{\nu} \approx 0$	$\hat{\lambda} = 1.093, \hat{\nu} = 0.149$	$\hat{\lambda} = 2.145, \hat{\nu} = 0.352$
dependence	$\hat{\theta} = 1.721$	$\hat{\theta} = 0.510$	$\hat{\theta} = -0.048$

can be interpreted as Borussia Dortmund showing different levels of control over the match: low control with counter attacks in state 1, a fairly balanced match in state 2, and high control with lots of ball possession in state 3. In state 3, the estimated negative dependence between the number of shots and ball touches may result from two different styles of high-control play: either Borussia Dortmund is controlling and passing the ball without much pressure on goal, or they go effectively straight for goal, without much passing. In addition, the t.p.m. is estimated as

$$\hat{\mathbf{\Gamma}} = \begin{pmatrix} 0.471 & 0.054 & 0.475 \\ 0.006 & 0.988 & 0.006 \\ 0.195 & \approx 0 & 0.805 \end{pmatrix}.$$

Here, with $\hat{\gamma}_{22} = 0.988$ and $\hat{\gamma}_{33} = 0.805$, there is very high persistence in state 2 (balanced state) and moderately high persistence in state 3 (high-control state). State 1 (low control and counter attacks) is a transient state with $\hat{\gamma}_{11} = 0.471$, where switching to the high-control state is most likely. Up next we will present the results for the model including covariates in the state process.

A model including covariates

The models presented so far already provide interesting insights into the dynamics of football matches, since the state-dependent distributions can be tied to different levels of control of the team considered. To gain further insights, we incorporate covariates to investigate potential drivers of momentum shifts. According to the AIC, the

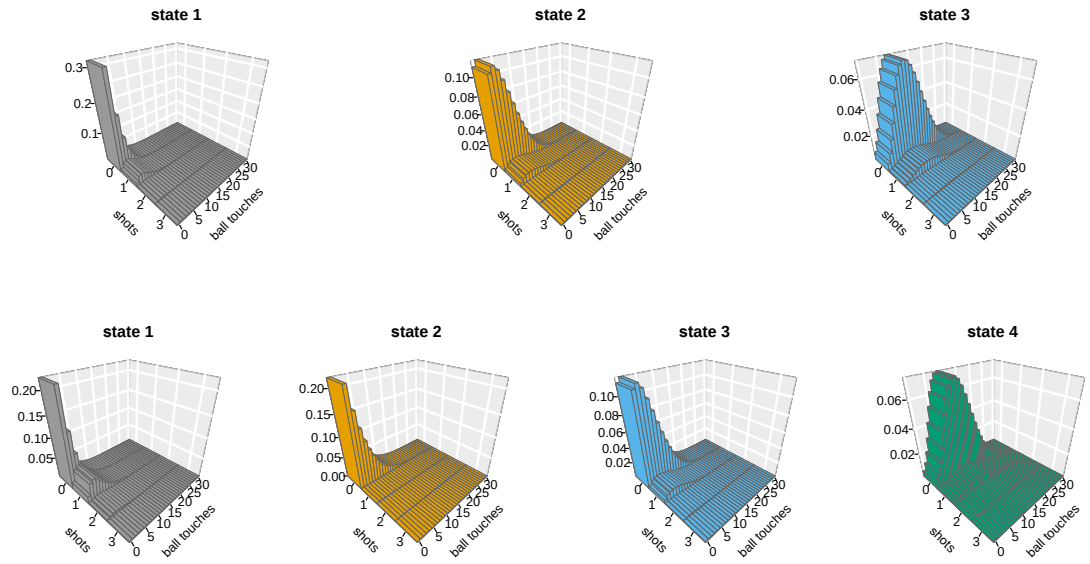


Figure 7.4: State-dependent distributions for the three-state (top row) and four-state (bottom row) Clayton-copula HMM, respectively.

model including all covariates considered is preferred over the model without covariates ($\Delta\text{AIC} = 51$); we do not conduct variable selection as we regard this analysis step as explanatory (rather than an attempt to find the best model).

For ease of interpretation, we suggest to visualise the estimated transition probabilities as functions of covariates, and present the theoretical stationary distributions of the Markov state process when fixing the covariate values at certain levels. The theoretical stationary distributions indicate how state occupancy, i.e. how much time is spent in a state, varies across different values of the covariate considered (*Patterson et al.*, 2009). To illustrate these two approaches, we present (i) the transition probabilities as functions of the covariate minute, and (ii) the stationary distributions with respect to the score difference. Table A4 in Appendix C displays the estimated $\beta_0^{(ij)}, \ldots, \beta_p^{(ij)}$ and their 95% CIs.

For (i), as displayed in Figure 7.5, the values of the score difference and the market value of the opponent are set to 0 and 200, respectively, corresponding to situations where the score is even and the opponent's strength is about average. In addition, we focus on home matches only, since the corresponding dummy variable in the linear predictor does not affect the overall pattern regarding the direction of the effect. The confidence intervals (indicated by the dashed lines) are obtained based on Monte Carlo simulation from the approximate multivariate normal distribution of the estimator. According to the estimated effects, switching from state 1 (low control and counter

attacks) and state 2 (balanced state) to state 3 (high-control state), respectively, becomes more likely at the end of matches. In addition, staying in state 3 also becomes more likely at the end of matches.

The stationary distributions for the score difference are shown in Table 7.4. The values of the minute and the market value of the opponent are fixed at 80 and 200, respectively, corresponding to situations in the final stage of a match with the opponent's strength being about average. The stationary distributions indicate that there is a high probability for Borussia Dortmund to be in state 3 (high-control state) either if they have a clear lead or if they are trailing. In contrast, if they hold only a slender lead, then the probability of being in state 1 (low control and counter attacks) is highest.

To further investigate typical patterns of momentum shifts according to the state process $\{s_{mt}\}$, we calculate the most likely trajectory of the states for each match m. Specifically, for a given match m, we seek

$$(s^*_{m1}, \ldots, s^*_{mT_m}) = \underset{s_{m1},\ldots,s_{mT_m}}{\operatorname{argmax}} \Pr(s_{m1}, \ldots, s_{mT_m} | \mathbf{y}_{m1}, \ldots, \mathbf{y}_{mT_m}),$$

i.e. the most likely state sequence, given the observations. Maximising this probability is equivalent to finding the optimal of N^{T_m} possible state sequences. This can be achieved at computational cost $\mathcal{O}(T_m N^2)$ using the Viterbi algorithm (*Zucchini et al.*, 2016). Figure 7.6 displays the decoded sequences for the match Borussia Dortmund against Schalke 04 which was already shown in Figure 7.1. We see confirmed that Borussia Dortmund started the match in the high control state with occasional switches to the low control state with counter attacks. According to the decoded state sequence, Borussia Dortmund is predominantly staying in the low control state with counter attacks after the half time, with occasional changing level of control around minute 70, where they switched to the balanced state. However, at the end of the match, they mostly stayed in the low control state with counter attacks and conceded two more goals.

7.5 Discussion

There is wide interest in the dynamics of football matches, and specifically in potential momentum shifts, in particular by fans and the media. From a managerial perspec-

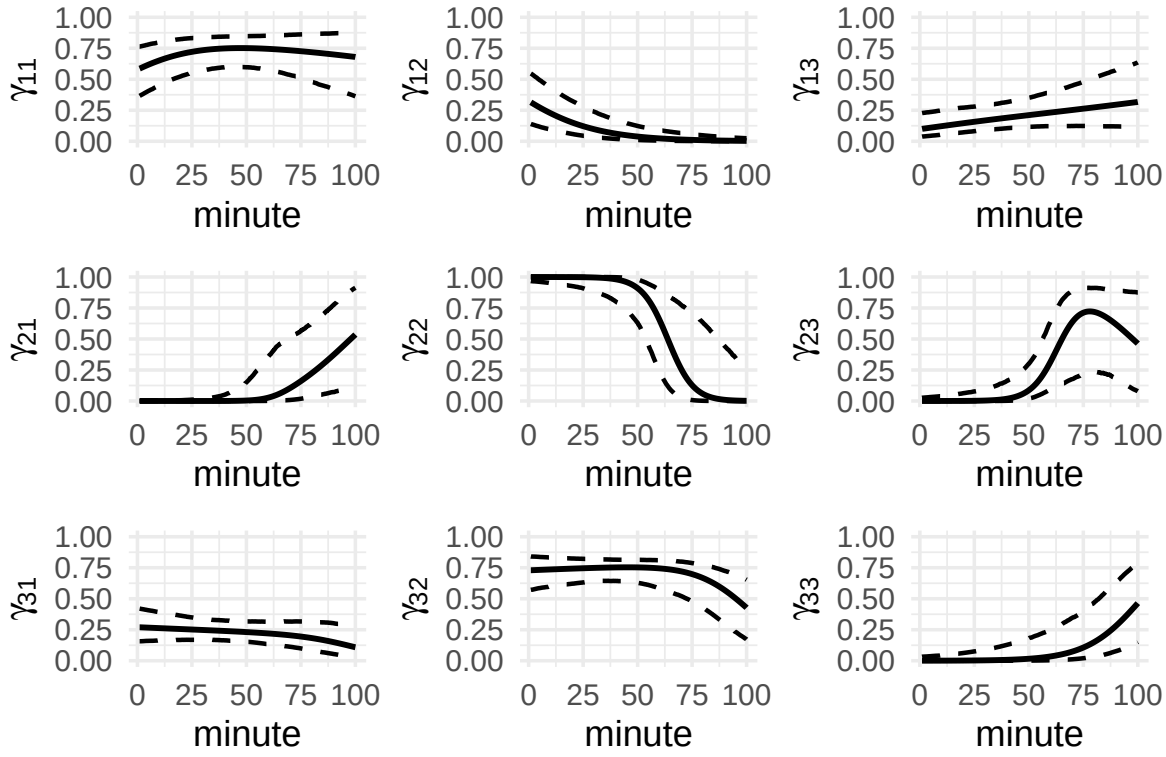


Figure 7.5: Transition probabilities as functions of the covariate minute. The dashed lines indicate confidence intervals (obtained based on Monte Carlo simulation). The values of the score difference and the market value of the opponent are set to 0 and 200, respectively. Table A4 in Appendix C displays the coefficients of the multinomial logistic regression underlying this figure.

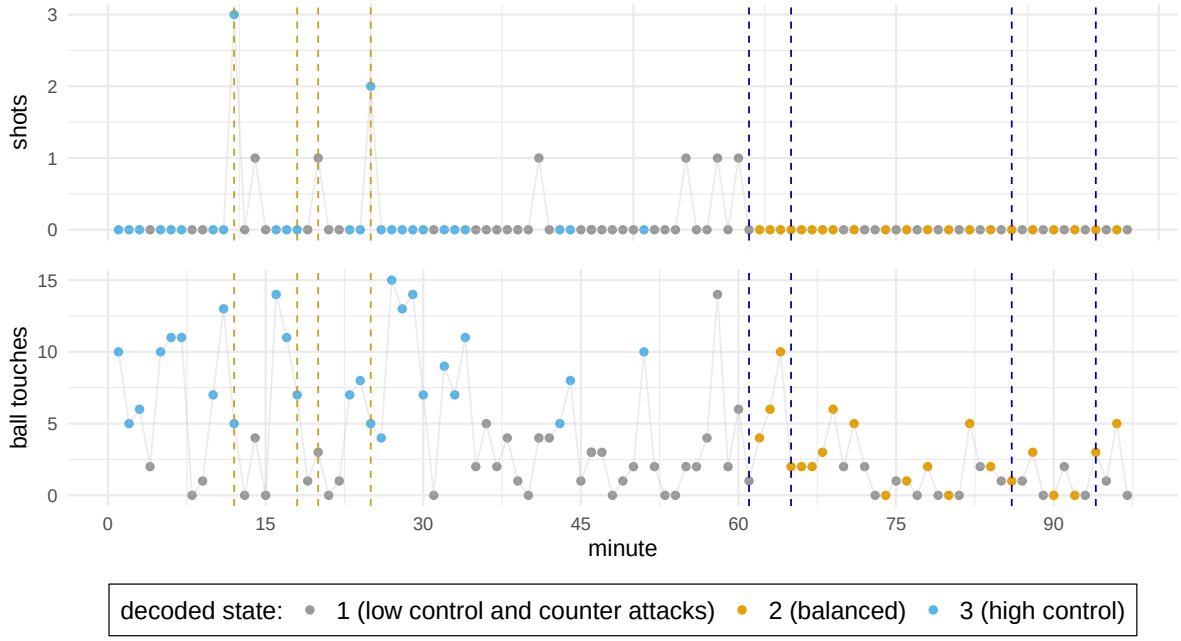


Figure 7.6: Decoded most likely state sequence of the match Borussia Dortmund against Schalke 04 according to the three-state Clayton-copula HMM including covariates. The vertical dashed lines denote goals scored by Borussia Dortmund (yellow lines) and Schalke 04 (blue lines).

tive, it is important to understand the causes of such shifts, and hence also how to potentially exert an influence on the match outcome. With data sets on in-game sum-

Table 7.4: Stationary distributions when fixing the score difference at certain levels. Probabilities were calculated for each value of the score difference, with the market value of the opponent and the minute of the match fixed at 200 and 80, respectively, corresponding to situations in the final stage of a match against an opponent team of average strength.

	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5
state 1	0.073	0.100	0.134	0.175	0.222	0.280	0.523	0.732	0.705	0.642	0.560	0.475
state 2	0.391	0.364	0.334	0.301	0.267	0.234	0.206	0.175	0.147	0.122	0.098	0.076
state 3	0.535	0.535	0.532	0.524	0.511	0.486	0.271	0.094	0.148	0.236	0.342	0.450

mary statistics becoming freely available, we now have the opportunity to statistically investigate the corresponding processes. To that end, here we provide a modelling framework — copula-based multivariate HMMs — which naturally accommodates potential changes in the dynamics of a match by relating the observed in-game match statistics to latent states. A key strength of the proposed approach is that we not only partition a given match into different phases but also allow for the investigation into drivers of how a match unfolds dynamically over time.

In our proof-of-concept case study, we tested the feasibility of our approach by analysing minute-by-minute data on matches of one particular team, namely Borussia Dortmund. The underlying states of the fitted model correspond to match phases where Borussia Dortmund exhibits a low level of control with counter attacks, to phases where the match is balanced, and to those with high level of control, respectively. In addition, the estimated effects of the covariates shed some light on what kind of events may lead to switches between those states. Specifically, we found that Borussia Dortmund has the highest probability of being in the high-control state when having a clear lead or when trailing.

Although the states of the fitted models are tied to different levels of control, it remains unclear whether these are clearly attributed to shifts in the underlying momentum. Specifically, some of the reported effects may arise due to tactical considerations rather than momentum shifts. For example, for one-goal leads, being in the low control and counter attacks state may be a tactical consideration rather than a shift in the underlying momentum. The data considered here does not allow us to disentangle these two possible causes, rendering a definitive conclusion whether the switches between the states are momentum shifts or tactical considerations impossible. However, with the states and effects of the covariates considered (cf. Figure 7.5 and Table 7.4) being easy to interpret, they still provide interesting insights to dynamics of football matches. In addition, using copula-based HMMs as presented in this chapter

may be helpful for bookmakers to obtain more precise estimations of betting odds. For instance, when modelling the time until the next goal during a football match, bookmakers could take into account the latent dynamics of a match as modelled here.

A clear limitation of the approach as presented here is that we focus on the in-game dynamics of only one of the two teams involved in a match, when in fact it is clear that the dynamics of a match result from the combination of both teams' actions. It seems conceptually desirable to extend the approach to allow for the joint modelling of both teams' in-game statistics. This could be achieved using a bivariate Markov chain to represent both teams' underlying states, resulting in N^2 combinations of states (see, e.g., *Sherlock et al.*, 2013). To further improve the realism of these models, it would be beneficial to also include tracking data, e.g. by considering the distances run per minute as covariate information.

The modelling framework used in this chapter, i.e. copula-based HMMs for modelling football minute-by-minute data, can easily be transferred to other sports for further investigations and possible characteristics of momentum shifts. These sports include, e.g., basketball, where the variables to be modelled comprise, for example, the number of points/shots, the number of rebounds, and the number of blocks/steals. More general, sports with two individuals or teams competing against each other and multiple variables measured on a fine-grained scale are best suitable for analysing momentum shifts using the modelling framework provided here.

8 Performance under pressure in skill tasks: An analysis of professional darts

8.1 Introduction

The effect of pressure on human performance is relevant in various areas of the society, including sports competitions (*Hill et al.*, 2010), political crises (*Boin et al.*, 2016), and performance-based payment in workplaces (*Ariely et al.*, 2009), to name but a few. A broad distinction differentiates between effort and skill tasks. Success in effort tasks is dependent on motivation to perform while skill task outcomes underlie precision of (often automatic) execution. For effort tasks, such as counting digits (*Konow*, 2000) or filling envelopes (*Abeler et al.*, 2011), individuals will typically respond to increased pressure (e.g. resulting from performance-related payment schemes) by investing more effort, which given the nature of such tasks will improve their performance (*Lazear*, 2000; *Paarsch and Shearer*, 1999, 2000; *Prendergast*, 1999). However, the literature on the impact of pressure on performance in skill tasks, e.g. juggling a football (*Ali*, 2011), is inconsistent and effectively divided into two different strands of research.

On the one hand, the existing literature related to potential "choking under pressure" indicates broad agreement that performance in skill tasks declines in high-pressure or decisive situations. An individual is said to be choking under pressure when their performance is worse than expected given their capabilities and past performances (*Beilock and Gray*, 2007). While there may also be random fluctuations in skill levels, choking under pressure refers to systematic suboptimal performance in high-pressure situations. The associated empirical findings — both such that are based on experimental data but also those using field data — consistently confirm a negative impact of pressure on skill tasks. On the other hand, and to some extent in contrast to the literature related to choking under pressure, the literature related to the concept of "social facilitation" refers to potential negative but also potential positive effects of (social) pressure on performance — depending on circumstances associated with the

performance. The social facilitation literature explicitly incorporates characteristics of the task and individuals' level of expertise into their analyses, and generally states that the circumstances surrounding performance play an important role regarding the impact of pressure on performance. Existing contributions focusing on potential choking have largely neglected the corresponding more comprehensive picture drawn by the social facilitation literature, by simply relating performance decrements to changes in the execution of actions, or simply distraction, generated either by rewards in case of success (*Ariely et al.*, 2009; *Baumeister*, 1984) or potential penalties in case of failure (*Kleine et al.*, 1988).

Our empirical investigation of individual's performance in pressure situations is based on a large data set from a skill task, namely professional darts, comprising 32,274 individual dart throws, for a comprehensive empirical test of performance under pressure. For the professional darts players analysed in this study, playing darts is a full time job. The top players regularly earn prize money exceeding one million euro per year. In professional darts, highly skilled players repeatedly throw at the dartboard from the exact same position effectively without any interaction between competitors, making the task highly standardised. The amount of data available on throwing performances not only allows for comprehensive inference on the existence and the magnitude of any potential effect of pressure on performance, but also enables to track the variability of the effect across players. The literature on choking would suggest that performance of professional darts players declines in high-pressure situations. However, when considering the highly standardised task to be performed and players' high level of expertise, we do not expect dart players to choke under pressure.

The chapter is structured as follows. Section 8.2 reviews the literature on performance under pressure, and in particular details what we consider to be advantages of the darts setting with respect to investigating performance under pressure. In Section 8.3, we explain the rules of darts and define what constitutes pressure situations in darts. Section 8.4 presents the empirical approach and results.

8.2 Performance under pressure

8.2.1 Terminology

Pressure results from individuals' ambitions to perform in an optimal way in situations where high-level performance is in demand (*Baumeister*, 1984). Performance under pressure could in principle go either way, i.e. high expectations towards (the own) performance could impact performance in a negative (choking) or a positive (clutch) way — or not at all. To measure the impact of pressure, performance in pressure situations is compared to performance in non-pressure situations. Choking under pressure refers specifically to a *negative* impact of high performance expectations (*Baumeister and Showers*, 1986; *Hill et al.*, 2009) while clutch performance is described as "any performance increment or superior performance that occurs under pressure circumstances" (*Otten*, 2009, p. 584).

8.2.2 Potential effects of pressure

The impact of pressure on performance crucially depends on the type of task to be performed. Tasks can be such that performance is determined mostly by effort, or alternatively tasks can be such that the skill level is the key factor for success. For effort tasks, pressure situations result in increased effort and hence improved performance (*Rosen*, 1986). For skill tasks, performance has been demonstrated to be both impaired (choking) and increased (clutching) by pressure — or not affected at all. While the effect of pressure on effort tasks is obvious and well documented, in skill tasks the potential psychological factors at play are likely more complex, such that we focus on these tasks in the following.

Choking

Choking under pressure in skill tasks may be related to various drivers. In particular, different skills may make use of different memory functions, namely explicit and procedural memory, respectively (*Beilock*, 2010). Explicit memory enables the intentional recollection of factual information, while procedural memory works without conscious awareness and helps at performing tasks. Two classes of attentional theories capture choking under pressure, distraction theories and explicit monitoring theories (*DeCaro*

et al., 2011; *Hill et al.*, 2010).[1] Distraction theories claim high-pressure situations to harm performance by putting individuals attention to task irrelevant thoughts (*Beilock and Carr*, 2001; *Lewis and Linder*, 1997). Put in a nutshell, individuals concern about two tasks at once, since the situation-related thoughts add to the task to be performed. Given the restricted working memory individuals performance declines as focus is drawn away from the main task (*Engle*, 2002).

On the other hand, self-focus or explicit monitoring theories explicitly predict that pressure increases self-consciousness to a point where it harms performance (overattention). It can cause the skilled performer to deviate from routine actions (*Markman et al.*, 2006). Instead, closer attention is paid to the single processes of performance and their step-by-step control. This ties in with the concept of skill acquisition: when initially learning a skill, performance is controlled consciously by explicit knowledge as actions are executed step-by-step (*Anderson*, 1982). Over time and through practice, skills become internalised and usage of conscious control decreases. Pressure can interfere with this now automated control processes of skilled performers (*Wulf and Su*, 2007). Under pressure, actions are no longer executed automatically as attention is redirected to task execution (*DeCaro et al.*, 2011). The overall sequence of actions is broken down into step-by-step control as in early stages of learning, resulting in impaired performance (*Masters*, 1992). Consequently, individuals consciously monitor and control a skill they would perform automatically in non-pressure situations (*DeCaro et al.*, 2011; *Jackson et al.*, 2006).

Other potential effects

An alternative strand of literature suggests that 'pressure' situations do not inevitably affect performance in a negative way but may also have a positive impact on task performance — or no effect at all. The corresponding notion of social facilitation is one of the oldest paradigms within experimental social psychology (see, e.g., *Geen and Gange*, 1977; *Zajonc*, 1965): "Generally, social facilitation refers to performance enhancement and impairment effects engendered by the presence of others either as coactors or, more typically, as observers or an audience" (*Blascovich et al.*, 1999, p. 75). A potential theoretical explanation for the opposing effects of audience is that

[1] Some authors argue that distraction and explicit monitoring theories are not necessarily mutually exclusive, but rather complementary (see e.g. *Beilock and Carr*, 2001; *Sanders and Walia*, 2012).

social presence facilitates *dominant* behaviour (*Zajonc*, 1965).[2] Hence, whether audience facilitates $[+]$ or impairs $[-]$ performance depends on the type of task (simple $[+]$ vs. complex $[-]$) and/or individuals' level of expertise (expert $[+]$ vs. beginner $[-]$) (*Harkins*, 1987). The presence of others increases the individuals' (physiological) arousal or drive level which in turn impairs or enhances task performance, respectively (*Zajonc*, 1965). A review of 12 years of research following the drive theory suggests that their propositions are still valid (*Geen and Gange*, 1977). Nonetheless, alternatives to drive theory have evolved in the following decades. While some non-drive theories relate audience effects to self-awareness (*Bond and Titus*, 1983), others refer to (cognitive) attention focus (*Huguet et al.*, 1999). Though experimental research uniformly confirms that social presence affects individuals' performance, it remains unclear which mechanism mainly drives behaviour. As the presence of others represents a particular case of pressure, it hence seems perfectly possible that pressure *enhances* performance — depending on the type of task and the individuals' level of expertise.

8.2.3 Empirical findings for performance under pressure in skill tasks

As this chapter analyses performance under pressure in a sport-related skill task, this section is devoted to previous findings from sports.[3] Golf putting performance is investigated in an experimental setting, suggesting performance to be worse when subjects are put under pressure (*Lewis and Linder*, 1997). However, in high-pressure situations participants who are distracted by a secondary task (counting down from 100) outperform subjects who solely concentrate on the putting task. The latter result is explained by too much focus on the task execution induced by the additional motivation to perform well in high-pressure conditions. The additional focus disturbs task execution which normally is performed automatically. There is also further evidence for diminishing golf putting performance under pressure provided by asking 108 undergraduate students with little or no golf experience to putt a golf ball as close to a target as possible (*Beilock and Carr*, 2001). Considering different kinds of intervention methods, pressure-like situations using monetary incentives are created. Results

[2] Dominant behaviour refers to the kind of response which is more likely: correct or incorrect. In case of, e.g., simple tasks it is more likely to perform the task correctly while individuals tend to make more mistakes when executing more complex tasks (*Bond and Titus*, 1983).

[3] There are also early non-sport studies (*Baumeister*, 1984; *Heaton and Sigall*, 1991).

generally confirm decreasing performance for high-pressure situations. However, the authors show putting accuracy to slightly increase under pressure when subjects had made their practice putts under self-consciousness-raising conditions.

Based on the assumption that pressure increases left-hemispheric activation which in turn is related to the controlled execution of a task and thereby to performance decrements, participants of a previous study performed a sport-related motor skill task in three blocks (in football, tea kwon do, or badminton) (*Beckmann et al.*, 2013). While the first two trials serve as for the introduction of pressure, the third trial is performed after participants have squeezed a softball for 30 seconds. Thereby, half of the participants activated their right hemisphere by squeezing the ball in their left hand, before again performing the task under pressure. Overall, the findings indicate performance deterioration when pressure is introduced but that the activation of the right hemisphere can eliminate this effect, thus preventing choking under pressure. However, they find no evidence for increased performance under pressure.

In a further study, a throwing task had to be performed by the participants to analyse novices' performances (*McKay et al.*, 2012). During the experiment, the performance expectancy within the experimental group regarding the ability to perform under pressure is manipulated. The results show a significant performance increase of the experimental group when pressure is applied, while the performance of the participants in the control group does not alter before and during pressure situations.

For a hockey dribbling task with 34 experienced participants, performance is found to be worse in high-pressure situations (*Jackson et al.*, 2006). Results further show that within high and low-pressure conditions subjects perform better when not concentrating explicitly on the task execution. By analysing a hockey dribbling setting with experienced hockey players, additional evidence for declining performance in pressure situations is found. However, it is demonstrated that in a high-pressure priming condition, performances are equal to those in a low-pressure situation and better (thus faster) than in a high-pressure non-priming condition (*Ashford and Jackson*, 2010).

For basketball novices, decreasing free throw success in pressure situations is shown (*Jackson et al.*, 2006). This result only applies to those subjects who are asked to pay close attention to the execution process during the practising phase. Analysing free throw performances of competitive basketball players instead of novices supports the results (*Wang et al.*, 2004). Thus, participants suffer a significant decrease in free throw success when performing in a high-pressure situation induced by the introduction

of an audience, videotaping, and offering financial rewards for improved performance.

A further study analyses the impact of fear of negative evaluation on performance, investigating success rates of throwing a basketball from a short distance (*Mesagno et al.*, 2012).[4] The authors find decreasing performance (thus choking) only for participants who were anxious about being evaluated negatively. For other subjects no significant differences in success rates are found.

Outside of experiments, field studies take advantage of the wealth of data on actual market participants who repeatedly perform almost identical tasks but under varying degrees of pressure. Pressure in these instances is determined by factors such as the importance of the competition considered, the current score in the competition, and the time left to play in a match.

Penalty kicks in football are considered to be a prototype pressure situation, as they critically affect the match outcome and the expectation to score a goal is very high. In line with the hypothesis of individuals tending to choke under pressure at skill tasks, success rates of penalty kicks in professional football are found to decline with increasing importance of success, i.e. as pressure increases (*Dohmen*, 2008). However, contradictory to these results, success rates in penalty shootouts are found to increase with pressure in the German cup competition confirming clutch performance (*Kocher et al.*, 2008). In addition, several studies focus on the "last-mover disadvantage", i.e. whether teams that go first in a shootout have an advantage over the other team resulting from higher pressure from trailing (*Apesteguia and Palacios-Huerta*, 2010; *Arrondel et al.*, 2019; *Kocher et al.*, 2012). One of these studies finds that last-mover teams indeed suffer from this kind of pressure (*Apesteguia and Palacios-Huerta*, 2010), the other studies refute this finding and speculate the contradictory results to be a consequence of data issues (*Arrondel et al.*, 2019; *Kocher et al.*, 2012). Potential reasons for varying success in penalty shootouts between players are that players from high-status countries a) generally perform worse and b) engage more in escapist self-regulation strategies than players from low status-countries (*Jordet*, 2009).

In golf, performance under pressure is analysed for putting (*Clark III*, 2002a,b). Analysing the impact of the current leaderboard situation on performance, the author finds that interim results are irrelevant for performance. In particular players who are in the lead or close to the lead in the final round do not perform worse than those who are further behind. Furthermore, players' performances are constant across

[4] Shots are taken from five different spots which all are placed at the distance of the free throw line.

rounds. Between-athlete comparisons may explain this finding, which is not in line with the widely accepted hypothesis of individuals choking under pressure (*Wells and Skowronski*, 2012). Considering also within-golfer comparisons, such findings cannot be replicated, and corresponding studies instead do find athletes to choke under pressure (*Wells and Skowronski*, 2012). Relating choking under pressure to golfers' age, an inverted U-shaped relationship on the professionals' tour with performance under pressure peaking at age 36 is shown (*Fried and Tauer*, 2011). The success rate at the final putt of a golf tournament is found to decrease as the value associated with that shot increases (*Hickman and Metz*, 2015). Finally, golfer currently with the lead are found to underperform at the end of close contests (*Hickman et al.*, 2019).

Basketball free throws constitute another scenario that is often investigated to analyse performance under pressure. Considering data from the National Basketball Association (NBA), and modelling free throw success rates as a function of the current score, players are shown to perform much worse when their team is either trailing by 1 or 2 points, or in the lead with 1 point. Attempts are more successful when the score is tied (which equals less pressure since a miss would end in an overtime and not a loss) (*Worthy et al.*, 2009). Further evidence for choking under pressure in professional basketball is reported with performance declining with additional pressure (*Cao et al.*, 2011). However, the authors show performance to be unaffected by the crowd size, the tournament round, and whether or not it is a home game for the player considered. Examining the determinants of choking under pressure, overall lower free-throw success rates are found for different groups (containing females and males, and amateurs and professionals) in case of high-pressure situations (*Toma*, 2017). Analysing the performance of professional basketball players who had been categorised as "clutch players" by basketball experts is also part of previous research (*Solomonov et al.*, 2015). Results show that clutch players are indeed able to increase their performance[5] in high-pressure situations such as the final minutes of close games, while performance of other players is not affected by pressure. Therefore, results provide evidence that clutch performers actually do exist. However, the analysis further shows no differences for clutch players' field goal percentage between low-pressure and high-pressure situations. It is also reported that professional basketball players who maintain their performance under pressure earn higher salaries (*Deutscher et al.*, 2013).

While some contradictory results have been reported, overall there still seems to

[5]Performance is measured by points scored and fouls drawn.

be fairly strong evidence that professional athletes do choke under pressure, at least in some scenarios.

8.2.4 Task features of the darts setting

Empirical advantages

Despite the effort that has already gone into studying the impact of pressure on performance, we believe that the setting of professional darts is an important addition to the existing body of literature. While we do not claim the following features to be unique to darts — as they effectively also apply to bowling, archery etc. — they are important to mention as they improve the reliability of any results obtained, compared to other more complex settings which have regularly been analysed in past research.

First, in darts, players cannot interfere the performance of the opponent directly. To precisely measure the impact of pressure, analyses need to focus on such performance that is not affected by others (*Baumeister and Steinhilber*, 1984). In many other settings, such as penalty kicks in football, opponents can impact each other's success. As a matter of fact missing a penalty shot can be caused by the kicker's or the goalkeeper's performance, respectively, or both. The individualistic nature of darts reduces variance caused by interference of opponents present in other settings.

Second, subjects in our data are highly trained in the task they perform. Such experience is obtained from training and previous competition, the latter may or may not be covered in our sample. Observing experienced professionals vastly reduces the noise to be expected for inexperienced players with large fluctuations in performance. The separation of the impact of pressure on performance is hence much clearer in professional sports settings (compared to lab experience with amateurs).

Third and closely related to the previous point, the task to be performed in a pressure situation is more or less identical to the only task the players perform throughout the contest. The only difference is given by the specific field the player attempts to hit. In comparison, penalty shots only account for a very small fraction of actions a football player need to perform (*Feri et al.*, 2013). In line with our previous argument, estimating skill levels in pressure situations requires such separation of signal and (potentially very large) noise. If pressure is closely related to the task at hand (e.g. a penalty shot) it is hard to separate between pressure generated by the task and pressure generated by the situation.

Fourth, all players in darts are repeatedly confronted with high pressure situations. For penalty kicks or free throws, team managers may rely on the same set of players when confronted with pressure situations, namely those who they have faith in to deal with the pressure or are very skilled in the specific task. Such sample selection can be detrimental to the quality of the results and occurs especially for very specific tasks.

Overall, we believe that professional darts offers a nearly optimal empirical setting to investigate the impact of pressure on performance. Players repeatedly perform highly standardised actions, with no interference by an opponent or any teammates involved, and hardly any relevant external factors.

Characteristics of task / darts players

As already discussed above, the social facilitation literature suggests that the circumstances surrounding performance affects the consequences of pressure. These circumstances mainly refer to the individuals' level of expertise and complexity / difficulty of the task. As our data set includes professional dart players who are highly trained in throwing darts, we observe individuals of high expertise.

Throwing darts is a skill task which requires high motor skills in order to perform well (*McEwan et al.*, 2013). There is a high level of standardisation of individual throws as well as many repetitions of almost identical actions, performed by professionals. Even though hitting a specific slice of the dartboard requires a high precision of movements, we assume that throwing a dart at a dartboard is less complex than, e.g., taking a penalty kick (football), throwing at a basket (basketball), or putting a ball (golf). The more the task relies on simple, well-rehearsed responses, the smaller the chances of performance decrements. Hence, we expect performance of dart players to be unaffected by pressure. In contrast to the literature related to social facilitation, the choking literature would predict that performance in darts declines as pressure increases.

8.3 Pressure situations in darts

For readers who may be unfamiliar with the rules of darts, we here provide a short description. The dartboard consists of 20 different slices, which differ with respect to their value (ranging from 1 to 20), and the center of the board, which is composed of two fields, namely the single bull and the bullseye. Each slice is further divided into

three different parts: two single, one double and one triple field. The bullseye is the double field of the single bull. Figure 8.1 shows the layout of a standard dartboard, highlighting the single five segment, the double and triple eight, respectively, and the single bull together with the bullseye. The inside width of the triple and double fields is 8mm, whereas the diameter of the bullseye is 12.7mm. A darts match is typically played by two players. (There are cases of team competitions in darts but these are not considered in our analysis.) Players are standing 2.37m away from the dartboard (at the "oche"), the height of which is 1.73m (from the ground to the center of the bullseye).

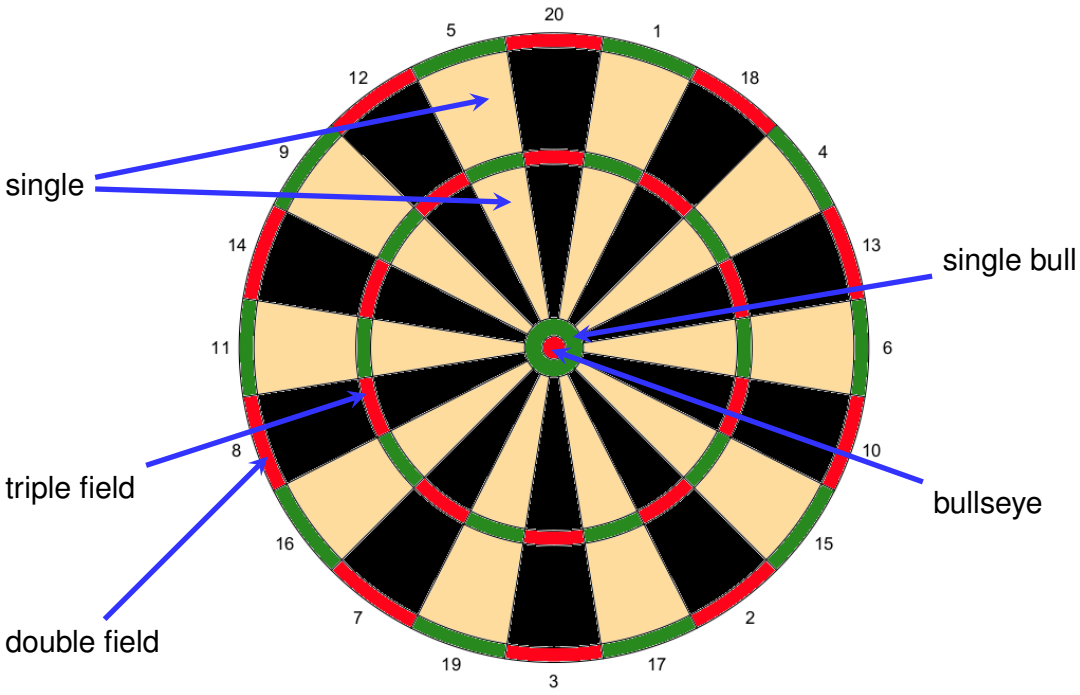


Figure 8.1: Dartboard layout.

While there are many possible games in darts, professional darts commonly follow the *501 up* format. To win a corresponding match, a player must be the first to win a pre-specified number of legs (typically between 7 and 15). Both players start each leg with 501 points and the opening throws in a new leg alternate between the two players. The first player to reach exactly zero points wins the leg, with the restriction that the dart that ultimately reduces the points to zero must hit a double field. For instance, in case a player throws a dart at the single/double/triple field of segment 20, 20/40/60 points are deducted from the player's current score.[6] The players take turns

[6] If a player hits a field that reduces his score below zero it is called a bust. The player starts with the number of points he had before he busted at his following turn.

to throw three darts in quick succession. At the beginning of a leg, players consistently aim at high numbers — usually triple 20 or triple 19 — to quickly reduce their points. The maximum score per dart is 60 (triple 20) and hence 180 for a set of three darts.

Once a player has the possibility to finish a leg (i.e. reach exactly zero points) with three darts (or less) during his turn, he is in the *finish region*. If he takes the opportunity and finishes the leg, this is called a *checkout*. As the last single dart has to hit a double field, the highest possible checkout is 170: two darts at triple 20 ($2 \times 60 = 120$) followed by a dart into the bullseye (50 points). The highest checkout not requiring a bullseye is 160 (two triple 20 followed by a double 20). For some scores below 170 there are multiple combinations for a checkout while there are none for others (e.g. 159 points as there is no three darts combination that leads to exactly zero points with the last dart hitting a double field[7]).

We determine the likelihood of a player checking out for any given number of points left. To do so, we use information on all attempts for the given score to determine the success rate (see below). The checkout proportions for the individual scores are shown in Figure 8.2, which in addition indicates whether (at least) 1, 2, or 3 darts are needed for a checkout. It is important to note that there is a strategic element to the game, where players sometimes deliberately attempt to set their score to a certain number for their next turn instead of checking out immediately. If, for example, Player A has a fairly high number of points to check out, say 160 points, but Player B has no finish with his next turn, then Player A could set up an easier checkout for his next turn rather than going straight for a checkout. The occurrence of such strategic behaviour is corroborated by Figure 8.3, which shows the checkout proportions in the data for those situations. For scores above 120 the checkout proportion for Player A is usually higher if Player B has a finish (compared to situations where Player B has no finish). When having a high score left to finish, players tend to set up an easier checkout if their opponents have no chance to finish in the next turn. Such strategic behaviour becomes less relevant for lower scores. For scores below 50, many of which can be checked out with one dart only, such that setting up a score is less relevant, the checkout proportions do not differ substantially between situations where the opponent had a finish and those where he had no finish. We explicitly account for such strategic

[7] 159 points could be reduced to exactly zero points with three darts if the last dart does not need to hit a double field, e.g. by triple 20 — triple 20 — triple 13. However, since all tournaments in our data are played as "double out", 159 points can not be reduced to zero within a players' turn.

considerations by restricting our sample to those observations where the opponent also has a finish.

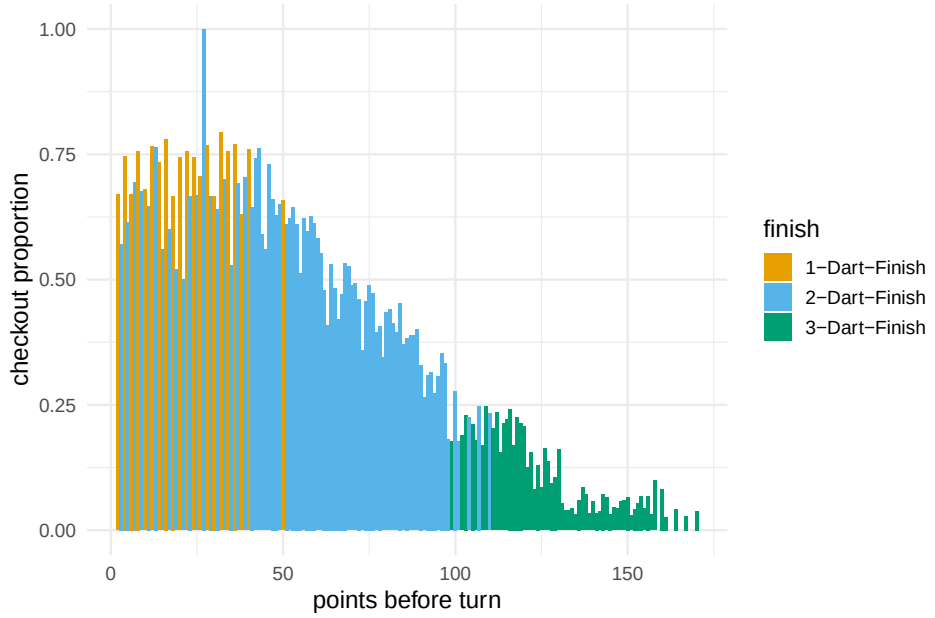


Figure 8.2: Checkout proportions for the individual scores before a player's turn. Colours indicate whether for the given score 1, 2, or 3 darts are needed for a checkout.

For any given turn of a player, the level of pressure is a result from the player's own likelihood of finishing within the current turn as well as that of the opponent finishing within his next turn. Respective probabilities are estimated by the corresponding empirical proportions as described above. Following the literature, the intermediate score of a match can also generate pressure. We hence also analyse a sub-sample of throws, which are performed in situations that are very crucial to the outcome of the match. More specifically, we investigate *decider legs*, referred to as legs where both players only need one more leg to win the match. Winning such a leg hence results in winning the match, whereas losing such a leg would result in losing the match. For example, in a best-of-19 leg match, a decider match occurs when the score is tied at nine legs apiece and leg number 19 decides the winner of the match. Pressure in decider legs is thus higher.

8.4 Empirical analysis

The data — extracted from http://live.dartsdata.com — cover all professional darts tournaments organised by the Professional Darts Corporation (PDC) between April 2017 and September 2018. Based on the raw data we reconstruct which player

makes a throw, the score before each dart, how many legs have been played in the match, which player had the first throw in any leg considered and, of course, if the player making a throw checks out. In the data we analyse, each row, i.e. observation, corresponds to a player's turn to throw (at most) three darts. From those rows, i.e. from all sets of three darts played by a player, we consider only those instances where both the player and the opponent have the chance to check out within the given and the next turn, respectively. To ensure reliable inference on player-specific effects, we further reduced the data set to consider only those players who had at least 50 attempts to check out. The final data set comprises information on the checkout performances of $m = 122$ different players, totalling to $n = 32,274$ observations (checkout yes/no).

8.4.1 Descriptive statistics

Our response variable *checkout* indicates whether a player managed to check out (coded as "*checkout* = 1") or not ("*checkout* = 0"). As detailed above, we measure the degree of pressure on a player by differentiating between his and the opponents' chances to finish a leg prior to his turn. The chance of a player checking out is quantified by the checkout proportions of all finishes from the player's current score (*checkoutproportion*). For the opponent, the corresponding covariate *checkoutproportionopp* indicates the checkout proportion of the opponent's current score.[8] To account for the ex-ante heterogeneity of players' chances to win the match, the competitive balance (*cb*) indicates the absolute difference in the winning probabilities. Based on betting odds taken from www.oddsportal.com, and after correcting for the bookmakers' margin, *cb* can take values between 0 and 1. High values of *cb* imply that the match is lopsided, whereas the value 0 means that both players have equal winning probabilities. Finally, as our data contains trained athletes, we are able to further control for the experience of the athlete (*exper*), proxied by the number of years the player belongs to a professional darts organisation (British Darts Organisation or PDC).

Table 8.1 summarises all covariates considered. Overall, about 42% of all checkout attempts are successful. However, the probability to successfully complete a checkout is highly dependent on the number of points required: the more points are needed,

[8] In an alternative model specification, we replaced the *checkoutproportionopp* variable by a dummy indicating whether or not the opponent had a chance to check out with his next attempt, restricting the sample to 1-dart finishes for comparable checkout proportions. The corresponding results (not shown) were consistent with the ones presented here.

Table 8.1: Descriptive statistics for the covariates.

	mean	st. dev.	min.	max.
checkout	0.420	–	0	1
checkoutproportion	0.419	0.279	0.027	1
checkoutproportionopp	0.486	0.266	0.027	1
exper	13.15	7.050	0	36
cb	0.363	0.228	0	0.899

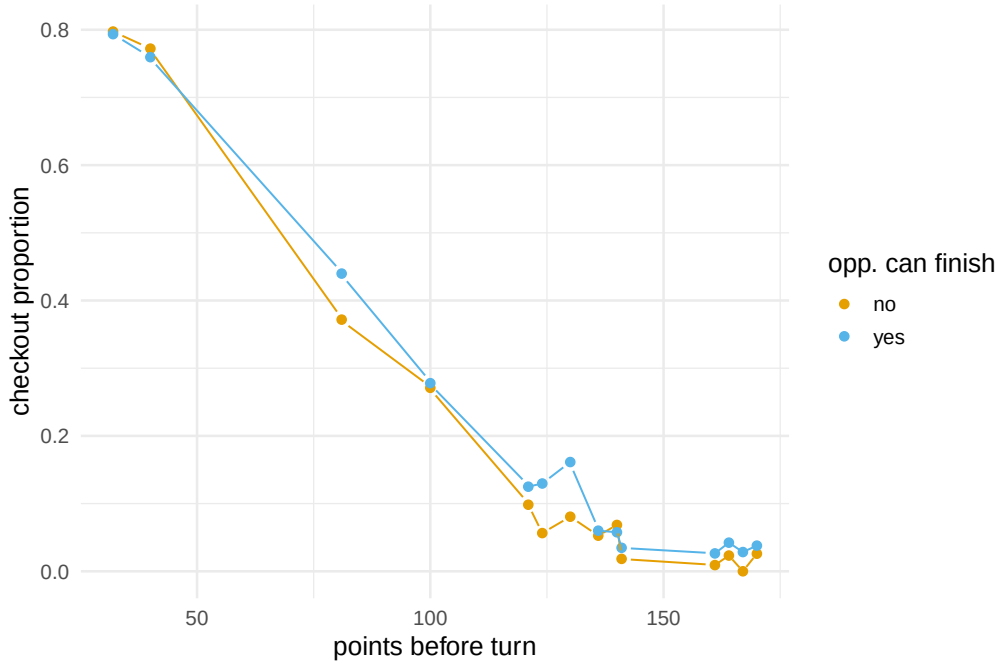


Figure 8.3: Checkout proportions for situations where the opponent had a finish (blue) and those where the opponent had no finish (yellow). Finishes with at least 100 observations in each category are shown.

the less likely is a checkout (see Figure 8.2).

To investigate the impact of pressure on performance, Figure 8.4 shows the checkout proportions for different levels of pressure, which are indicated by the colours. Due to the potential strategic adjustments discussed above, only those observations where the opponent can also finish are included. For scores above 100, the checkout proportions seem to increase with increasing likelihood of the opponent checking out, i.e. the more a player is under pressure. For lower scores there is no such clear trend.

In addition, the pressure as indicated by decider legs is investigated in Figure 8.5 by comparing the empirical checkout proportions in decider vs. non-decider legs. Since in only about half of the finishes the checkout proportion is higher in decider legs, there is no clear pattern indicated by these summary statistics.

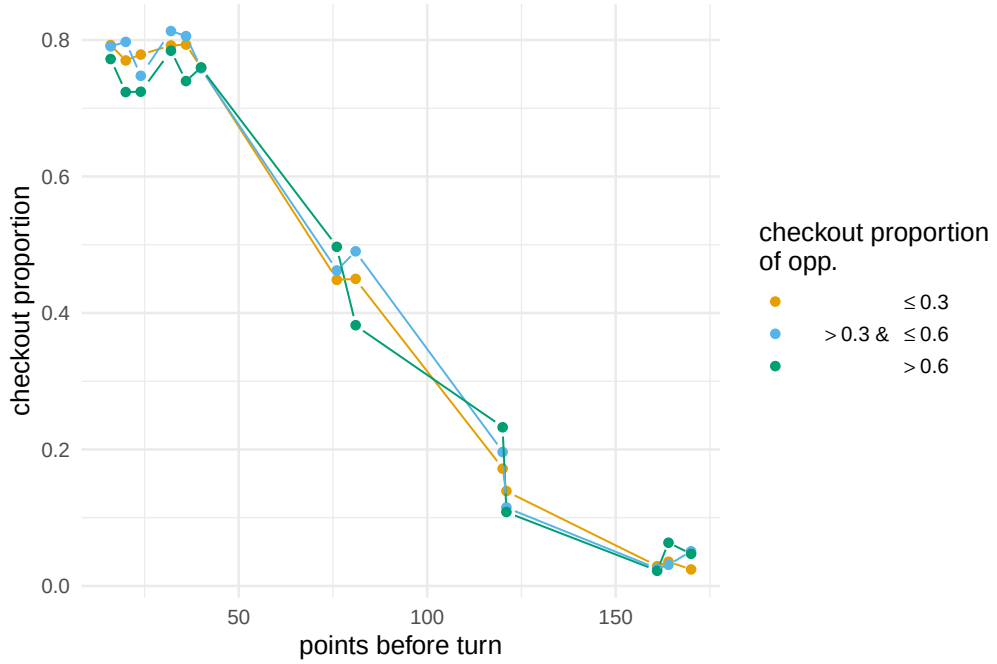


Figure 8.4: Checkout proportions in pressure vs. non-pressure situation. Specifically, checkout proportions are separated for different categories of checkout proportions of the opponent. Only scores with at least 100 observations per category are shown.

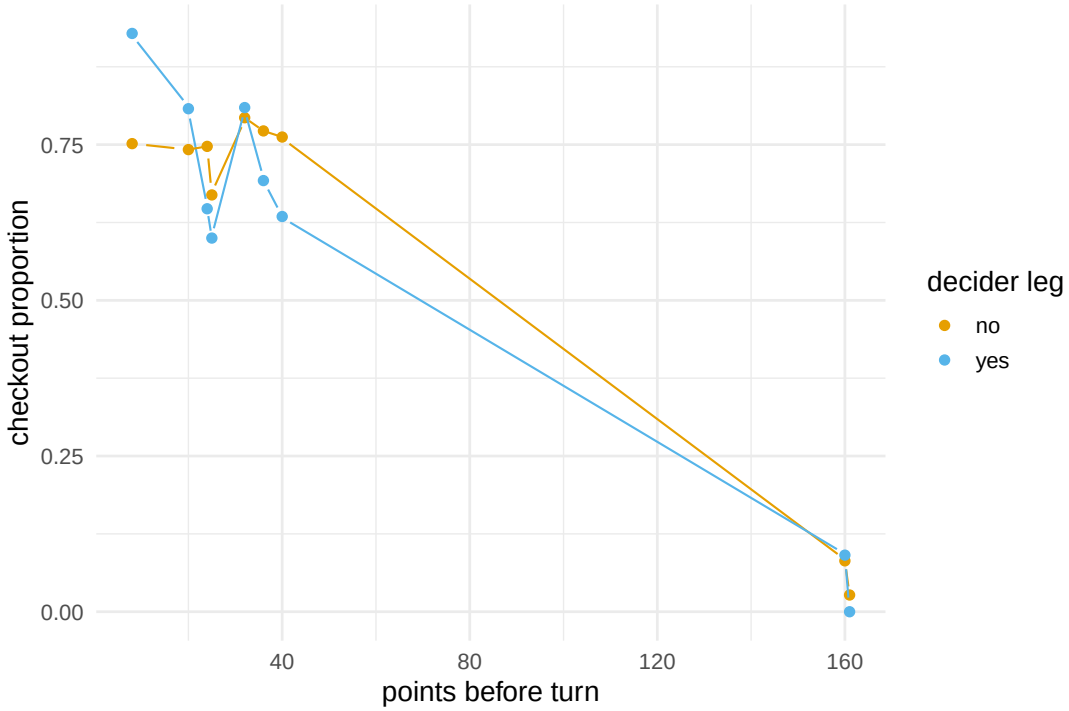


Figure 8.5: Checkout proportions in pressure vs. non-pressure situation as indicated by decider legs. Scores with at least 10 observations per category are shown.

8.4.2 Modelling checkout performance

The structure of the data considered is longitudinal, as we model the binary response variable $checkout_{ij}$, indicating whether or not the i–th player $(i = 1, \ldots, m)$ checked out

($checkout_{ij} = 1$) on the j–th attempt ($j = 1, \ldots, n_i$). To cover player-specific effects, and also to account for the fact that each individual player's observations are likely to be correlated, we apply generalised linear mixed models where the linear predictor η_{ij} contains a vector of fixed effects $\boldsymbol{\beta}$ as well as a vector of zero-mean random effects $\boldsymbol{\gamma}_i$:

$$\eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{u}_{ij}'\boldsymbol{\gamma}_i, \qquad i = 1, \ldots, m, \quad j = 1, \ldots, n_i,$$

with $\mathbf{x}_{ij} = (1, checkoutproportion_{ij}, \ldots)'$, and $\mathbf{u}_{ij}'$ the subvector of $\mathbf{x}_{ij}'$ with those covariates for which we assume individual-specific effects. The logit function links the binary response variable, $checkout_{ij}$, to the linear predictor:

$$\text{logit}\big(\Pr(checkout_{ij} = 1|\boldsymbol{\gamma}_i)\big) = \eta_{ij} = \mathbf{x}_{ij}'\boldsymbol{\beta} + \mathbf{u}_{ij}'\boldsymbol{\gamma}_i.$$

The linear predictor includes all covariates considered as well as a random intercept for each player to account for player-specific effects:

$$\eta_{ij} = \beta_0 + \beta_1 checkoutproportion_{ij} + \beta_2 checkoutproportionopp_{ij} + \beta_3 exper_i + \beta_4 cb_{ij} + \gamma_{0i}.$$

The random intercept γ_{0i} displays the player-specific deviation from the average intercept β_0 — further individual-specific effects will be considered below. These models are fitted by maximum likelihood estimation using the package `lme4` in R (*Bates et al.*, 2015; *R Core Team*, 2019). Table 8.2 displays the results for the corresponding fixed effects.

Table 8.2: Estimation results for the fixed effects of the turn-level model.

	response variable:		
	checkout		
	all attempts	no deciders	deciders
checkoutproportion	5.132	5.128	5.715
	[5.019; 5.245]	[5.014; 5.242]	[4.665; 6.764]
checkoutproportionopp	0.014	0.016	−0.108
	[−0.091; 0.118]	[−0.089; 0.122]	[−0.938; 0.722]
exper	0.005	0.005	0.007
	[−0.001; 0.012]	[−0.001; 0.012]	[−0.028; 0.043]
cb	−0.018	−0.014	0.107
	[−0.151; 0.114]	[−0.147; 0.120]	[−0.961; 1.175]
constant	−2.799	−2.797	−3.084
	[−2.920; −2.678]	[−2.918; −2.677]	[−3.977; −2.191]
observations	32,274	31,715	559
Note:			95% CIs are shown in brackets.

The estimated coefficients associated with *checkoutproportionopp* are of main in-

terest here as they display the impact of the opponent's chance of checking out during his next attempt on the player's chance to check out during his current attempt. To identify different levels of pressure connected to the intermediate score of the game, we fitted the model to different samples, distinguishing non-decider legs and decider legs. Perhaps somewhat surprisingly, evaluating the effect of *checkoutproportionopp* across the first two models, the more pressure a player is exposed to, i.e. the more likely the checkout of the opponent, the higher is the increase in the corresponding odds for a checkout. However, the corresponding effects are not statistically significant. For the third model, the effect is also statistically insignificant. Hence, we find no evidence that pressure impacts performance. This is also supported by a different model formulation where we pooled all attempts and introduced a dummy variable indicating if the throw occurred in a decider leg. The corresponding coefficient is insignificant, again providing no evidence for an effect of pressure on performance (results not shown). The player-specific random intercepts $\hat{\gamma}_{0i}$, i.e. the player-specific deviations from the intercept $\hat{\beta}_0$, range (on the logistic scale) from -0.217 to 0.398.

To conduct a more fine-grained analysis of the throwing performance, we ran a second analysis in which we changed the sampling unit to single throws instead of a complete turn of three throws. When analysing single throws instead of turns in darts, additional strategic adjustments have to be considered. If players can reduce their score to 0 with a single dart (e.g. if their score is 32), players often throw a "marker dart" with their first dart of a turn just outside of the board, such that the second dart is aimed at the marker and may be deflected into the target. To again account for such strategic adjustments, we only consider the third dart of a turn, since no marker darts are thrown with the third throw. The covariate *checkoutproportion* is then built from the score-specific checkout proportion of the third dart of a turn. The results of fitting the model to data of single throws are shown in Table 8.3. As was done also for the previous analysis based on turns (see Table 8.2), we fitted the model to data of all attempts, to non-decider legs, and to decider legs separately. The results again indicate that pressure does not impact performance in professional darts.

Since in the current model formulation we only allow for player heterogeneity in the baseline throwing performance, we further consider an extension where potential additional variation in the performance-in-pressure situations across players is investigated. The corresponding (and again insignificant) results are presented in Appendix D.

Table 8.3: Estimation results for the fixed effects of the model fitted to data of single throws.

	response variable:		
		checkout	
	all attempts	no deciders	deciders
checkoutproportion	4.534	4.562	2.749
	[3.930; 5.137]	[3.953; 5.170]	[−2.319; 7.817]
checkoutproportionopp	0.076	0.084	−0.327
	[−0.054; 0.205]	[−0.047; 0.215]	[−1.322; 0.669]
exper	0.005	0.006	−0.001
	[0.0004; 0.011]	[0.0005; 0.011]	[−0.043; 0.041]
cb	0.150	0.148	0.710
	[−0.011; 0.310]	[−0.014; 0.309]	[−0.579; 2.000]
constant	−2.394	−2.408	−1.570
	[−2.646; −2.141]	[−2.662; −2.153]	[−3.611; 0.471]
observations	14,849	14,590	259
Note:			95% CIs are shown in brackets.

8.5 Discussion

We find no evidence that professional darts players are impacted by (high) pressure situations. While player-specific effects for performance under pressure indicate that some professional players in our sample may improve, and some may worsen their performance in pressure situations, the average effect over all players is not statistically significant. Hence, our results do not corroborate studies supporting the choking hypothesis which states that overall performance in skill tasks decreases with increasing pressure.

The difference between our findings and previous studies on performance under pressure may partly be due to the fact that in our study we consider very highly skilled individuals who have to deal with the considered type of pressure situations on a regular basis. Professional darts players are at the very top of their profession and cannot fluke out of pressure situations, which is possible in team settings where tasks can be assigned to different team members. In fact, darts players face pressure situations on a regular basis and hence gain experience in dealing with these. While throwing darts is the one skill required in the setting considered, in other professions the set of tasks is much more diverse, often combing the requirement of both, skill and effort.

The literature on social facilitation offers a possible explanation for the absence of any choking effect. Social facilitation suggests that the type of task and level of expertise greatly affect the consequences of audiences or general pressure. As all players in our data set are professionals, pressure situations should affect performance

positively. However, we find positive effects only for some but not all players.[9] On the one hand, "ceiling effects by performing a well-learned task" (*Blascovich et al.*, 1999, p. 75) may lead to such insignificant performance effects. Hence, future research on darts players should also observe less experienced subjects to circumvent such ceiling effects. On the other hand, players may differ with respect to personal variables, such as self-confidence. Thus, pressure may affect performance differently depending on personal attributes. Further research on performance under pressure would benefit from including more information on personal characteristics.

Investigating semi-professional players (such as youth players) may further be beneficial with respect to a potential selection bias. Our sample may to some extent be the result of selection effects of subjects who can withstand pressure and become professionals, such that only those individuals who do not choke in pressure situations succeeded in the profession at hand and made it to the top (and hence into our sample).

The importance of coping with pressure situations has been investigated by in a qualitative study by interviewing ten international top athletes (*Jones*, 2002). In this study, several attributes are stated as important factors for being "mental tough", such as to be in control under pressure. In a further study, again several former Olympic or world championship winning athletes are interviewed as well as sport psychologists and coaches, finding that mentally tough athletes can not only cope with pressure situations, but even use it to raise their performance (*Jones et al.*, 2007). An explanation for this is that individuals are either entering a "competition state" or a "threat state" when forced to pressure situations, where the former helps their performance and the latter does not (*Jones et al.*, 2009). Thus, to not choke under pressure is not a conscious decision but rather a state of mind which is reached subconsciously.

Throwing darts arguably is a very specific task, much less complex than other actions required to perform in under pressure situations. Our finding of individuals not choking under pressure may be due to this specific task feature. Thus, future research on performance under pressure should include characteristics of the task and individuals into their considerations as these drive pressure effects. While the setting itself would be ideal to test gender differences in performance under pressure in a specific task, women's darts does not offer the data necessary to draw comparisons.

[9] Accordingly, results cannot be attributed to the type of task. Since all of the players are of high expertise and execute the same task, the type of task should have the same effect on all players.

Empirical comparisons in line with the research by *Gneezy et al.* (2003) are thus not possible at this time. Given the high number of observations for each player, further research could tackle the question if there is a memory for choking under pressure. More precisely, one could determine if choking under pressure impacts future choking under pressure, similar to a hot hand phenomenon particularly concerning pressure situations (*Miller and Sanjurjo*, 2018; *Ötting et al.*, 2020b).

Even though the social facilitation literature helps to understand the inconsistent impact of pressure on individuals' behaviour, it may be the case that pressure resulting from, e.g., competing for large monetary rewards or championship titles differs from pressure due to the presence of others. Whether individuals react to pressure with enhanced or impaired performance may hence also depend on the kind of pressure they experience while performing a certain task. It would be interesting to test whether dart players react differently to pressure situations (due to interim results) when playing before an audience or no spectators, respectively. However, this scenario would only be testable in laboratory settings as there are no contests taking place without spectators.

9 Predicting play calls in the National Football League using hidden Markov models

9.1 Introduction

Unpredictability of play calls is widely accepted to be a key ingredient to success in the NFL. For example, according to several players of the 2017 Dallas Cowboys, being too predictable regarding their play calling may have been one reason for their elimination from the playoff contention of the 2017 NFL season. Being unpredictable hence is desirable, and, vice versa, it is clearly also of interest to be able to accurately predict the opponent's next play call. In earlier studies, play call predictions were carried out by simple arithmetics, such as calculating the relative frequencies of runs and passes of previous matches (*Heiny and Blevins*, 2011). Driven by the availability of play-by-play NFL data, several studies considered statistical models for play call predictions. These studies can be divided in those where play-by-play data only is considered (see, *Heiny and Blevins*, 2011; *Teich et al.*, 2016) and those who consider additional data on the players on the field, such as the number of offensive players for a certain position and player ratings (see *Joash Fernandes et al.*, 2020; *Lee et al.*, 2017). The former report prediction accuracies of about 0.67, whereas the latter provide accuracies of about 0.75.

Most of these studies use basic statistical models, e.g. linear discriminant analysis, logistic regression, or decision trees, which do not account for the time series structure of the data at hand. This chapter considers HMMs for modelling and forecasting NFL play calls. In the recent past, HMMs have been applied in different areas of research for forecasting, including stock markets (see, e.g., *De Angelis and Paas*, 2013; *Dias et al.*, 2015), environmental science (see, e.g., *Chambers et al.*, 2012; *Tseng et al.*, 2020), and political conflicts (*Schrodt*, 2006). Within HMMs, the observations are assumed to be driven by an underlying state variable. In the context of play calling, the underlying states serve as a proxy for the team's current propensity to make a pass (as

opposed to a run). The state sequence is modelled as a Markov chain, thereby inducing correlation in the observations and hence accounting for the time series structure of the data. HMMs are fitted to data from seasons 2009 to 2017 to predict the play calls for season 2018. In practice, these predictions are helpful for defense coordinators to make adjustments in real time on the field. Offense coordinators may also benefit from these models, since they allow them to check the predictability of their own play calls.

This chapter is organised as follows: Section 9.2 describes the play-by-play data and provides exploratory data analysis. Section 9.3 explains HMMs in further detail, and Section 9.4 presents the results.

9.2 Data

The data for predicting play calls in the NFL were taken from www.kaggle.com, covering (almost) all plays of regular season matches between 2009 to 2018. In total, $m = 2,526$ matches are considered[1], each of which is split up into two time series (one for each team's offense), totalling in 5,052 time series containing 318,691 plays. The observed time series $\{y_{m,p}\}_{p=1,\dots,P_m}$ indicates whether a run or a pass play has been called in the p-th play in match m, with

$$y_{m,p} = \begin{cases} 1, & \text{if } p\text{–th play is a pass;} \\ 0, & \text{otherwise} \end{cases}$$

and P_m denoting the total number of plays in match m. For all matches considered, other plays such as field goals and kickoffs, which occur typically at the beginning or the end of drives, are ignored here. Since the main goal is to predict play calls, we divide the data into a training and a test data set. The data set for training the models cover all matches from seasons 2009 – 2017, comprising 2,302 matches and 289,191 plays. The test data cover 224 matches, totalling in 29,500 plays. For the full data set, about 58.4% of play calls were passes.

Since the play of the offense is likely affected by intermediate information on the match (such as the current score), several covariates are considered, which have also been considered by previous studies on predicting play calls summarised above: a dummy indicating whether the match is played at home (*home*), the yards to go for a

[1] The data comprises 2,526 regular-season matches out of 2,560 matches which have taken place in the time period considered.

first down (*ydstogo*), the current down number (*down1*, *down2*, *down3*, and *down4*), a dummy indicating whether the formation is shotgun (*shotgun*), a dummy indicating whether the play is a no-huddle play (*no-huddle*), the difference in the intermediate score (own score minus the opponent's score) (*scorediff*), a dummy indicating whether the current play is a goal-to-go play (*goaltogo*), and a dummy indicating whether the team is starting within 10 yards of their own end zone (*yardline90*). Table 9.1 summarises the covariates and displays corresponding descriptive statistics (for the full data set).

Table 9.1: Descriptive statistics of the covariates.

	mean	st. dev.	min.	max.
pass (response)	0.584	0.493	0	1
home	0.503	0.500	0	1
ydstogo	8.634	3.931	1	50
down1	0.443	0.497	0	1
down2	0.333	0.471	0	1
down3	0.209	0.407	0	1
down4	0.015	0.121	0	1
shotgun	0.525	0.499	0	1
no-huddle	0.087	0.282	0	1
scorediff	−1.458	10.84	−59	59
goaltogo	0.057	0.232	0	1
yardline90	0.033	0.178	0	1

To investigate how play calling varies with different downs and the shotgun formation, Figure 9.1 shows the empirical proportions for a pass found in the data, separated for the different downs and the shotgun formation. As indicated by the figure, a pass becomes more likely with increasing number of downs, and there is a substantial increase in passes observed if the team is in shotgun formation. However, whether a run or a pass is called is also likely to depend on the yards to go for a first down, which is shown in Figure 9.2, indicating that a pass becomes more likely the more yards are needed for a first down. The colours in Figure 9.2 indicate the (categorised) score difference, suggesting that a pass becomes more likely if teams are trailing.

In addition to the covariates potentially affecting the decision to call a pass or a run, one example time series from the data set, corresponding to the play calls observed for the New Orleans Saints in the match against the New York Giants played in November 2015 is shown in Figure 9.3. With 101 points scored in total, this match is one of the highest scoring NFL games. The plays shown in the figure underline that there are periods with a fairly high number of passing plays (e.g. around play 20), and those where more runs are called (e.g. around play 30).

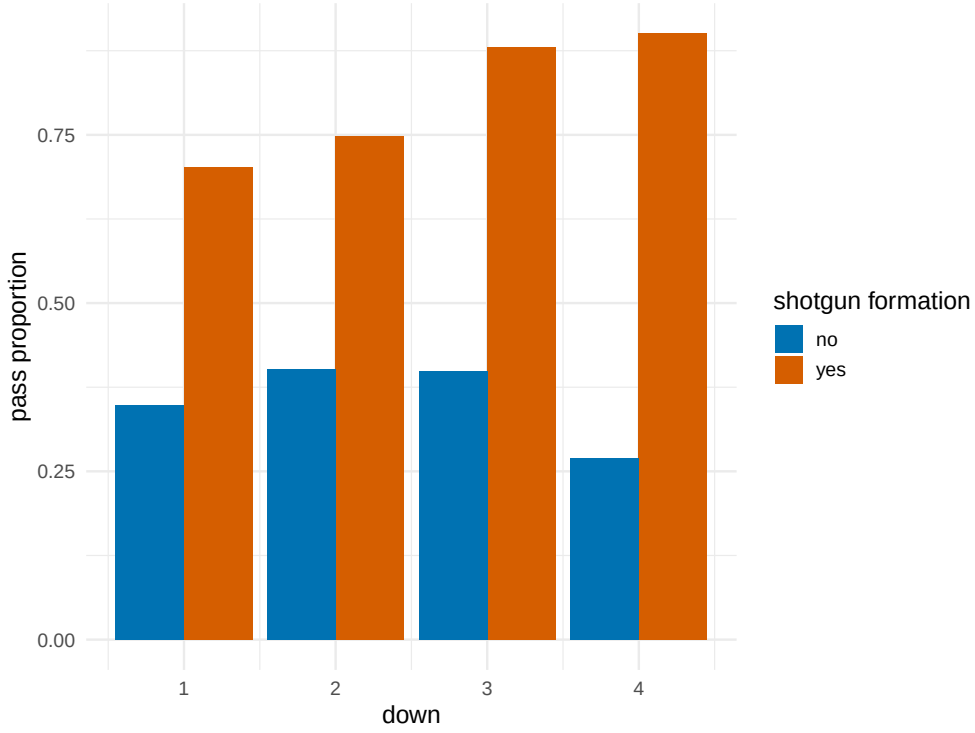


Figure 9.1: Empirical proportions for a pass found in the data for different downs and the shotgun formation.

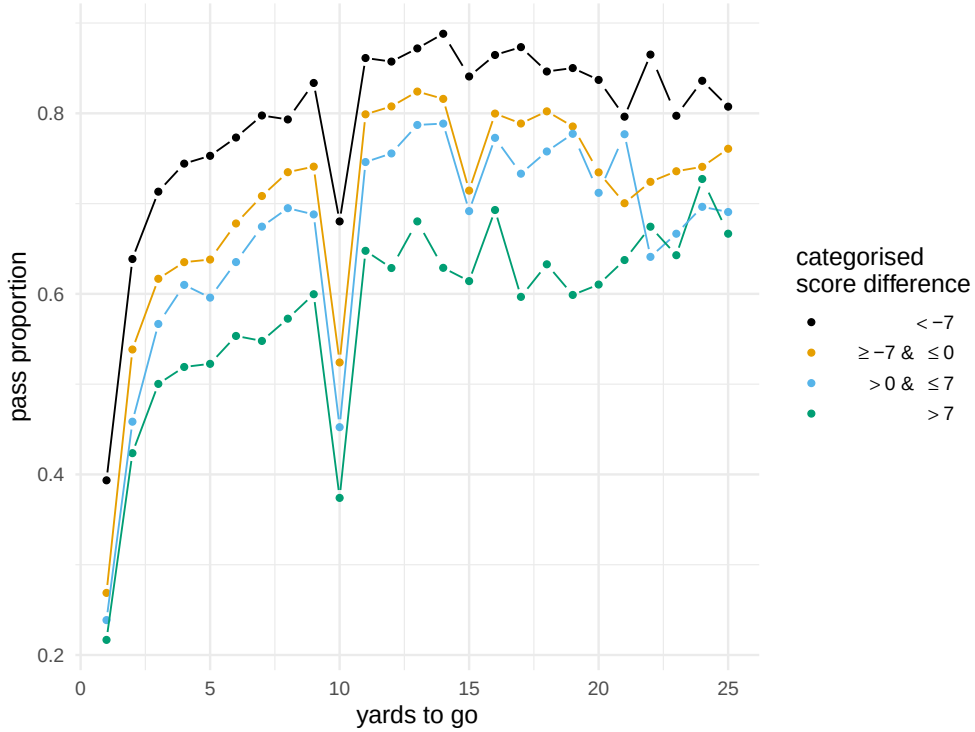


Figure 9.2: Empirical proportions for a pass found in the data for the different yards to go for a first down. Colours indicate the (categorised) score difference. The proportion for a pass for 10 yards to go is relatively low, since most of these observations correspond to a first down, where a run is more likely. Observations with more than 25 yards to go are excluded (the number of observations for each of these categories is less than 100).

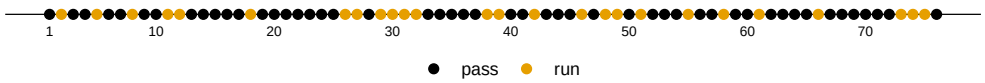


Figure 9.3: Example time series found in the data: the play calls of the New Orleans Saints observed for the match against the New York Giants played on November 1, 2015.

9.3 Modelling and forecasting play calls

To account for the periods of passes and runs as indicated by Figure 9.3, HMMs are considered for modelling and forecasting play calls. The underlying states can be interpreted as the propensity to make a pass (as opposed to a run) of the team considered. A HMM involves two components, namely an observed state-dependent process and an unobserved Markov chain with N states, assuming that the observations are generated by one of N pre-specified state-dependent distributions. The dependence structure of the HMM considered is shown in Figure 9.4. Here, the observed time series are the play calls $\{y_{m,p}\}_{p=1,\ldots,P_m}$, which are denoted from now on by y_p for notational simplicity. The unobserved state process, modelled by a N-state Markov chain, is denoted by $\{s_p\}_{p=1,\ldots,P_m}$. For the state transitions, a transition probability matrix (t.p.m.) $\mathbf{\Gamma} = (\gamma_{ij})$ is defined, with $\gamma_{ij} = \Pr(s_p = j | s_{p-1} = i)$, i.e. the probability of switching from state i at play $p-1$ to state j in play p. For the model formulation of a HMM to be completed, the number of states N and the class of the state-dependent distribution have to be selected. Since the play calls are binary, the Bernoulli distribution is chosen here. The corresponding probabilities of the observation given state i, i.e. $f(y_p \,|\, s_p = i)$ are comprised in the $i-$th diagonal element of the $N \times N$ diagonal matrix $\mathbf{P}(y_p)$. Since assuming a team to start in its stationary distribution at the beginning of an American football match is fairly unrealistic, we estimate the initial distribution $\boldsymbol{\delta} = \big(\Pr(s_p = 1), \ldots, \Pr(s_p = N)\big)$.

To include the covariates introduced above which may lead to state-switching, we allow the transition probabilities γ_{ij} to depend on covariates at play p. This is done by linking $\gamma_{ij}^{(p)}$ to covariates (denoted by $x_1^{(p)}, \ldots, x_k^{(p)}$) using the multinomial logit link:

$$\gamma_{ij}^{(p)} = \frac{\exp(\eta_{ij}^{(p)})}{\sum_{k=1}^{N} \exp(\eta_{ik}^{(p)})}$$

with

$$\eta_{ij}^{(p)} = \begin{cases} \beta_0^{(ij)} + \sum_{l=1}^{K} \beta_l^{(ij)} x_l^{(p)} & \text{if } i \neq j; \\ 0 & \text{otherwise.} \end{cases}$$

Since the transition probabilities depend on covariates, the t.p.m. as introduced above is not constant across time, and hence denoted by $\mathbf{\Gamma}^{(p)}$. To formulate the likelihood, we apply the forward algorithm, which allows to calculate the likelihood recursively at low computational cost (*Zucchini et al.*, 2016). The likelihood for a single match m is then given by:

$$L = \boldsymbol{\delta}\mathbf{P}(y_{m,1})\mathbf{\Gamma}^{(m,2)}\mathbf{P}(y_{m,2})\dots\mathbf{\Gamma}^{(m,P_m)}\mathbf{P}(y_{m,P_m})\mathbf{1}$$

with column vector $\mathbf{1} = (1,\dots,1)' \in \mathbb{R}^N$ (*Zucchini et al.*, 2016). To obtain the likelihood for the full data set, we assume independence between the individual matches such that the likelihood is given by the product of likelihoods for the individual matches:

$$L = \prod_{m=1}^{M} \boldsymbol{\delta}\mathbf{P}(y_{m,1})\mathbf{\Gamma}^{(m,2)}\mathbf{P}(y_{m,2})\dots\mathbf{\Gamma}^{(m,P_m)}\mathbf{P}(y_{m,P_m})\mathbf{1},$$

where M denotes the total number of matches. The model parameters are estimated by numerically maximising the likelihood using `nlm()` in R (*R Core Team*, 2019). Subsequently, we predict play calls for the test data using the fitted models. Specifically, to forecast play calls, the forecast distribution is considered, which is for a single match given as a ratio of likelihoods (dropping the subscript m for notational simplicity):

$$\Pr(y_{P+1} = y \,|\, \mathbf{y}^{(P)}) = \frac{\boldsymbol{\delta}\mathbf{P}(y_1)\mathbf{\Gamma}^{(2)}\mathbf{P}(y_2)\cdots\mathbf{\Gamma}^{(P)}\mathbf{P}(y_P)\mathbf{\Gamma}^{(y)}\mathbf{P}(y)\mathbf{1}}{\boldsymbol{\delta}\mathbf{P}(y_1)\mathbf{\Gamma}^{(2)}\mathbf{P}(y_2)\cdots\mathbf{\Gamma}^{(P)}\mathbf{P}(y_P)\mathbf{1}},$$

where $\mathbf{\Gamma}^{(y)}$ and $\mathbf{y}^{(P)}$ denote the t.p.m. as implied by the new covariates and the vector of all preceding observations of the match considered, respectively (*Zucchini et al.*, 2016). The play which is most likely under the forecast distribution is then taken as the one-step-ahead forecast. To address heterogeneity between teams, the models are fitted to data of each team individually instead of pooling the data of all teams. The corresponding results are presented in the next section.

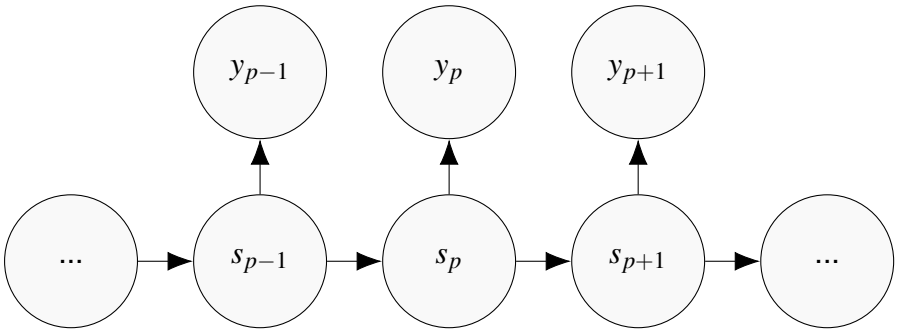


Figure 9.4: Dependence structure of the HMM considered. Each observation y_p is assumed to be generated by one of N distributions according to the state process s_p, which serves for the team's current propensity to make a pass (as opposed to a run).

9.4 Results

Before presenting the results on the prediction of play calls, the number of states N and the covariates have to be selected. As the number of parameters (due to the inclusion of covariates) increases considerably fast compared to the number of observations per team, we select $N = 2$ states here to avoid numerical instability. We apply a forward selection of the covariates described in Section 9.2 based on the AIC. In addition, we also include several interactions between the covariates, such as an interaction between *ydstogo* and *scorediff*, which was already indicated by in Figure 9.2. Based on further explanatory data analysis, the following additional interaction terms are considered: interactions between the different downs and *ydstogo*, between *shotgun* and *ydstogo*, between *nohudlle* and *scorediff*, and between *nohuddle* and *shotgun*. The AIC-based forward covariate selection is then applied for each team individually, with the covariates selected being slightly different between the teams.

The play call forecasts are evaluated by the prediction accuracy (i.e. the proportion of correct predictions), the precision (i.e. the proportion of predicted runs/passes that were actually correct) and the recall (i.e. the proportion of actual runs/passes that were identified correctly). The weighted average of the prediction accuracy over all teams is obtained as 0.715. This is a substantial improvement compared to existing studies that were also based on play-by-play data only (i.e. without including information on the players on the field). Moreover, the prediction accuracy obtained here is only slightly lower than the ones reported by *Lee et al.* (2017) and *Joash Fernandes et al.* (2020) (which are about 75%), notably *without* taking into account information about the players on the field.

The prediction accuracy for the individual teams is shown in Figure 9.5, indicating

that the lowest and highest prediction accuracy are obtained for the Seattle Seahawks (0.602) and the New England Patriots (0.779), respectively. In addition, the precision rates for a run range from 0.532 (Green Bay Packers) to 0.763 (Houston Texans), which can be interpreted as follows: when our model predicts a run for the Houston Texans (Green Bay Packers), it is correct in about 76.3% (53.2%) of all predicted runs. The recall rates for a run range from 0.324 (Baltimore Ravens) to 0.886 (Los Angeles Rams) — in other words, our model correctly predicts 88.6% of all runs for the Los Angeles Rams. For passing plays, precision and recall range from 0.559 (Seattle Seahawks) to 0.9 (Los Angeles Rams), and from 0.664 (Los Angeles Rams) to 0.922 (Pittsburgh Steelers), respectively. These summary statistics on the predicted play calls reveal that there are substantial differences in the predictive power with regard to the individual teams. Section 9.5 discusses practical implications following from these summary statistics. It took us on avarage 7 hours to conduct the AIC-based forward selection for the covariates on a standard desktop computer. However, using the fitted models to predict play calls takes less than a second for a single match, thus rendering the approach considered suitable for application in practice.

9.5 Discussion

The use of HMMs to predict play calls in the NFL indicates that the accuracy of the predictions is increased — compared to similar previous studies — by accounting for the time series structure of the data. We split the data into a training set (seasons 2009–2017) and a test set (season 2018), and fitted HMMs to the (training) data of all teams individually, which yields 71.5% correctly predicted out-of-sample play calls. The prediction accuracy for the individual teams range from 60.2% to 77.9%, with the highest prediction accuracy obtained for the New England Patriots (see Figure 9.5).

Practitioners have to take into account the variation in the prediction accuracy across teams and plays. For example, if a pass is predicted for the Los Angeles Rams, it is fairly likely that the actual play will indeed be a pass (according to our model), since the corresponding precision is obtained as 90%. On the other hand, if a pass is predicted for the Seattle Seahawks, this forecast has to be treated with caution, as the precision is obtained as 55.9%. Additional aspects for practitioners are the costs of an incorrect decision. For example, if teams want to avoid that a pass is anticipated although the actual play of the opponent's offense is a run, then coaches

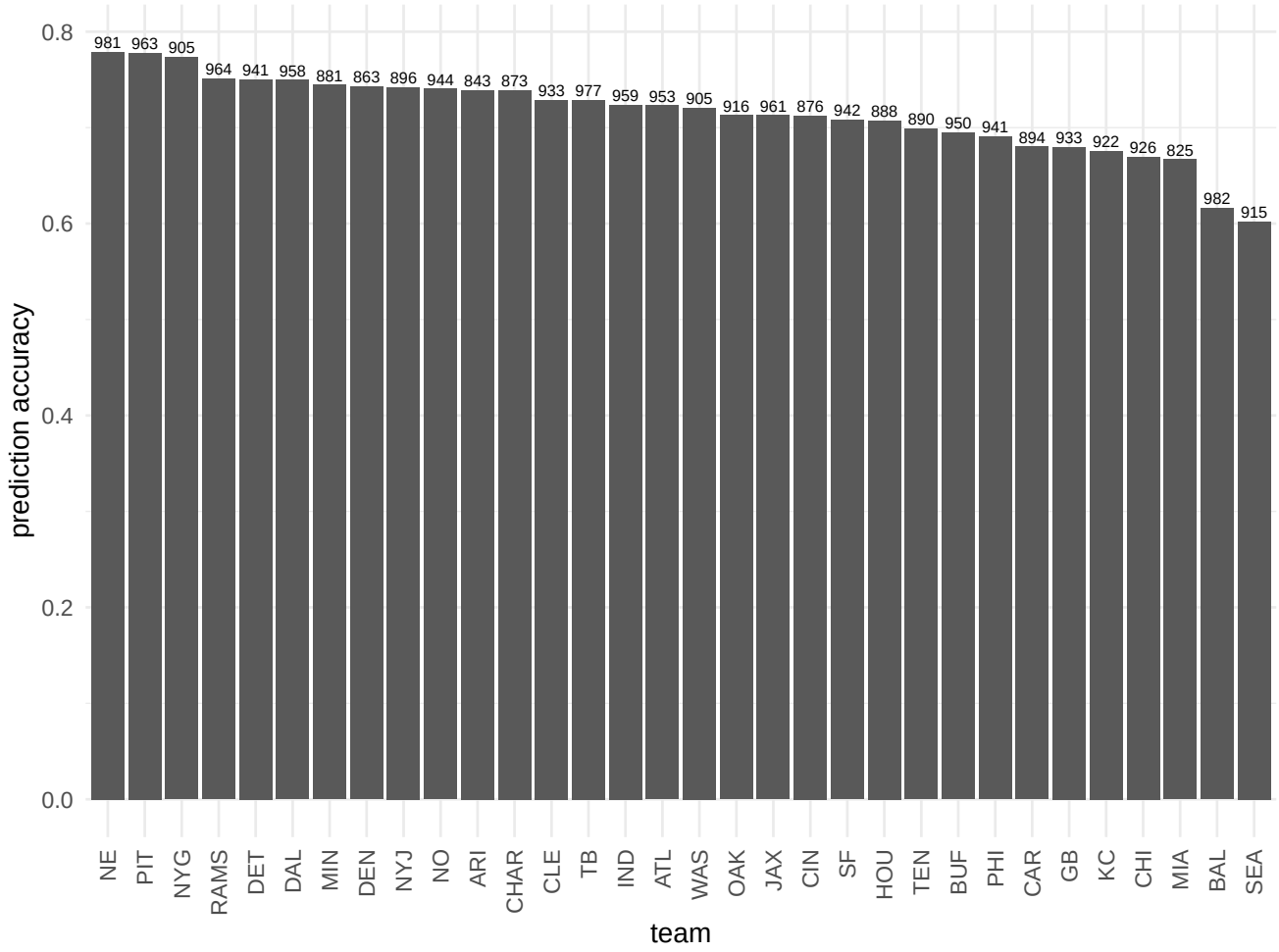


Figure 9.5: Prediction accuracy for the individual teams. The number of out-of-sample observations (i.e. of predicted plays) is shown at the top of the bars.

should carefully consider the corresponding precision rates. Since the models presented here provide probabilistic forecasts and not only binary classifications, coaches could consult the forecasts only if the predicted probability exceeds a chosen threshold. In any case, practitioners should not regard these models as a tool which delivers defense adjustments for each play automatically, but rather as an additional help to make better defense and offense plays, respectively.

Further research could focus on including additional covariates to improve the predictive power, such as the personnel of the team, i.e. the information on how many running backs/fullbacks, tight ends and wide receiver are on the field. In addition, the current strength of the team is not captured yet. This could be quantified by, for instance, the player ratings provided by the video game Madden, which was also done by *Lee et al.* (2017) and *Joash Fernandes et al.* (2020). However, it is at least questionable whether information on players can indeed be used on the field in practice, since players are substituted fairly frequently during a match. Finally, updating the

model throughout the 2018 season dynamically, rather than using the model fitted up to season 2018 in the out-of-sample prediction would further improve the predictive power.

10 Summary and outlook

This chapter summarises the main results of Chapters 2 – 9 and provides discussions on further research. Chapters 2 – 4 cover studies on betting markets, including determinants of betting volumes, market inefficiencies, and fraud detection. Section 10.1 summarises the main findings of these chapters and provides an outlook for further research. Chapters 5 – 9 cover several analyses on the evaluation of in-game performance. Section 10.2 summarises the main findings of these chapters and provides additional points for further research in this field.

10.1 Betting fraud detection

Several match fixing incidents occurred in the past decade, leading to an increased demand for fraud detection systems. Whereas existing fraud detection systems focus primarily on odds movements, this thesis argues that both betting volumes and betting odds should be monitored. For the former, statistical models have to account for the complex patterns present in the data, which include heteroscedasticity and non-linear effects. In Chapter 2 we use the flexible class of GAMLSS to explicitly account for these patterns. As a case study, we analyse betting volumes of the English Premier League. The results suggest that the matchday, the weekday, the strength of the teams (quantified by both teams' market values), and the uncertainty of outcome affect both the mean and the standard deviation of betting volumes. Compared to a classical linear model, the GAMLSS improves the model fit substantially in terms of the AIC. When using these models for examining betting fraud, the concept of market inefficiencies becomes important, as extreme betting volumes may arise due to inefficiencies and fraud, respectively. This renders fraud detection by analysing betting volumes difficult in the presence of market inefficiencies. The results presented in Chapter 3 suggest that inefficiencies exist in German betting markets. These inefficiencies occur at the beginning of a season, as bookmakers have only little information on the strength of recently promoted teams at that time. However, with more information on teams'

strength becoming available during the season, the inefficiencies disappear. For the analysis of match fixing in Chapter 4, the model formulation developed in Chapter 2 is used to detect fixed matches by identifying outliers in betting volumes. In addition to the betting volumes, betting odds are predicted using a GAMLSS with bivariate Poisson response. Considering data from the Italian Serie B, we achieve a true positive rate of about 75% for the detection of fixed matches, with at the same time only 34% false positives. These results suggest that monitoring betting volumes *and* betting odds can lead to more reliable detection of fixed matches.

The field of betting markets provides several points for further research. The investigation of in-game betting markets constitutes a natural extension of the pre-game analysis as presented in Chapters 2 – 4. When analysing the determinants of in-game betting volumes, the modelling framework has to account for the time series structure of the data, as in-game volumes are sampled at high frequencies. For the analysis of inefficiencies in in-game betting markets, betting odds could be investigated after certain events such as red cards and scored goals. However, as for the analysis of in-game betting volumes, statistical models accounting for the time series structure are also required when investigating in-game betting odds. Using both in-game betting odds and volumes, extending the fraud detection approach presented in Chapter 4 to in-game betting constitutes a further point for future research. Such fraud detection systems for in-game betting are highly relevant, as nowadays about 70% of the total betting volume is generated by in-game bets (*Forrest and McHale*, 2019).

10.2 Evaluation of in-game performance

Due to increasing amounts of data in sports, managers aim at evaluating the performance of players in an objective manner. However, although it is to be expected that an increased amount of data is beneficial for this purpose, humans often succumb to cognitive biases when evaluating performance. For example, when analysing potential hot hand patterns, humans tend to over-interpret streaks of success and failure. In academic research on the hot hand, it still remains an open question whether such an effect exists. In Chapter 5, we developed a state-space modelling framework for analysing a hot hand effect, and fitted our model to data from professional darts. Although the corresponding results are inconclusive regarding a hot hand effect, the modelling framework can be applied to several sports. Whereas this is a first point

for further research, an additional point is to explicitly address heterogeneity between players. As sports commentators and journalists often presume that certain players are more likely to experience streaks, investigating such differences and potential causes — such as players' experience — is of great interest. A further point for future research relates to the computational cost of model fitting. Speeding up computation time could be achieved by implementing the for-loops in C++ using the package `Rcpp` (*Eddelbuettel*, 2013), as the code used for model fitting was purely written in R. In addition, since the likelihood for the complete data set is given by the product of individual likelihoods, parallelising the computation of the individual likelihoods would further speed up the computation time.

When considering HMMs for the analysis of a potential hot hand effect in settings with a large number of covariates, regularisation approaches as presented in Chapter 6 facilitate variable selection. We incorporated the (relaxed-)LASSO into HMMs, which has shown to be a computationally fast and accurate approach for variable selection. However, as Chapter 6 only covers the simple LASSO penalty, further research should focus on other penalties aligned with the LASSO. These include, for example, penalties for a group of coefficients (see, e.g., *Hastie et al.*, 2015). Such penalties are useful in the context of HMMs with covariates for selecting the number of states. Specifically, if a group of coefficients belonging to a certain state is not selected, it can be removed from the model, thus leading to a simpler model with fewer states. Moreover, the use of the LASSO in HMMs is not restricted to the analysis of sports data. Researchers in many fields can benefit from regularisation methods in HMMs in settings where many covariates are included in the model.

Chapter 7 covers the development of a modelling framework for analysing in-game data with a focus on momentum shifts. The framework developed is very flexible, as it allows to model multivariate in-game data using HMMs in combination with copulas. The case study presented in Chapter 7 provides insights into the dynamics of football matches. However, the modelling framework is flexible enough such that it can be applied to various sports and is not restricted to football data. The application of the modelling framework to other sports is thus a first point for further research. Moreover, future research includes a more fine-grained analysis of matches by considering tracking data. In several sports, modern technologies enable tracking (x,y) coordinates of all players on the pitch, often sampled at frequencies of at least 1 Hz. For the case of momentum shifts, data on player movements enable a more detailed analysis, as (e.g.)

in football it makes a huge difference whether the team considered has much ball possession in their own half or in the opponent's half. In addition, tracking data allow for new response variables, as players' paths could be modelled to deepen the insights into momentum shifts. As HMMs do currently not allow for responses that are curves (such as players' path), future method development is required to fully address such high-frequency tracking data. Similar to the LASSO-HMMs discussed above, HMMs with curves as response may not only be beneficial for the analysis of sports data, but also for researchers in different fields working with such functional data.

A further phenomenon related to performance evaluation is choking under pressure. Chapter 8 presents a corresponding analysis of professional darts. The results provide no evidence for either choking or excelling under pressure. The points for further research in this field are similar to those for Chapter 5 discussed above. Gender differences are of interest when analysing performance under pressure, and including personal characteristics of players might deliver new insights on the determinants of performance in pressure situations.

A potential benefit for managers in sports when analysing in-game data is to gain an advantage over the opponent by, e.g., analysing the opponent teams' tactics. Chapter 9 presents a corresponding study using play-by-play data from the NFL to predict play calls. As it was done in the previous chapters, the time series structure of the data is explicitly taken into account here by using HMMs. The model developed achieves a out-of-sample prediction accuracy of 71.5%. The prediction accuracy is thus increased compared to previous studies where simple statistical models are used. Similar to the outlook for Chapter 7, tracking data may improve the prediction accuracy of play calls. Specifically, by considering such data, we could use the teams' personnel as an additional important covariate. In addition, as discussed above in the outlook for Chapter 7, by using tracking data we could model and predict routes of players instead of the play (run/pass) only.

Appendices

A Further betting-related information

A.1 Details on Betfair

Betfair offers traditional bookmaking as well as the world's largest online betting exchange. In Chapters 2 and 4, we make use of data from the Betfair betting exchange. Put simply, Betfair betting exchange is a broker for bets. It merges supply and demand for bets and takes a commission for matching both market sides. Betting volume hence is generated only when bets are matched. When a match "opens" for betting on Betfair, there is no volume matched, since at that point nobody has had the opportunity to place a bet yet. To place a stake on a certain outcome of a match, bettors have to choose a wager and odds for the bet. Betfair will display this "offer" on its website for other market participants to place a certain stake *against* this bet at the odds the first bettor stated. Accordingly, the betting odds are provided by the bettors themselves. If a bettor accepts the odds offered and places a stake against it, the bets match and the matched betting volume increases correspondingly. If an offer remains unmatched then it does not increase the betting volume.

A.2 Betting volumes per team (Chapter 2)

Table A1: Mean betting volumes and number of bets per team (2009/10 until 2015/16).

team	mean of *poundsbet*	mean of *numberbets*
Manchester United	5,044,876	55,670
Manchester City	4,351,260	47,261
FC Liverpool	4,328,456	49,583
FC Arsenal	4,271,753	48,420
FC Chelsea	4,234,363	46,646
Tottenham Hotspur	3,298,788	40,945
FC Portsmouth	2,648,104	21,414
Newcastle United	2,549,732	33,242
FC Blackpool	2,477,045	25,640
FC Everton	2,394,599	31,098
Aston Villa	2,351,609	28,994
Leicester City	2,287,525	32,032
West Ham United	2,274,285	28,199
AFC Bournemouth	2,253,632	30,038
Birmingham City	2,209,694	22,162
Blackburn Rovers	2,190,458	22,513
Queens Park Rangers	2,104,151	28,130
Wolverhampton Wanderers	2,017,532	21,013
FC Fulham	1,995,686	23,550
Bolton Wanderers	1,987,743	21,029
FC Southampton	1,980,386	28,332
AFC Sunderland	1,968,522	25,296
Crystal Palace	1,902,075	27,582
Swansea City	1,894,234	26,320
FC Burnley	1,887,839	21,750
Hull City	1,844,026	22,476
West Bromwich Albion	1,843,745	23,735
FC Reading	1,817,930	23,852
Wigan Athletic	1,803,402	20,975
Stoke City	1,801,054	22,783
FC Watford	1,749,590	25,434
Norwich City	1,687,900	23,037
Cardiff City	1,663,929	23,467

A.3 List of fixed matches (Chapter 4)

season	home	away	volume model – max. quant. residual	odds model – max. quantile
2009	1	29	1.32	1.00
2009	8	25	0.43	0.86
2009	13	25	1.04	0.59
2009	16	25	0.76	0.90
2009	25	26	0.86	0.93
2010	1	33	3.03	1.00
2010	3	4	1.78	2.25
2010	4	24	1.20	0.86
2010	4	33	3.73	0.99
2010	28	40	0.26	0.99
2010	40	33	2.12	0.95
2010	40	46	2.02	0.92
2013	5	36	0.81	0.99
2013	13	16	1.95	0.99
2013	15	44	1.08	0.89
2013	26	5	1.53	0.77
2013	29	10	0.21	0.88
2014	11	13	-0.32	0.94
2014	11	24	1.00	0.89
2014	11	42	0.11	0.94
2014	11	44	0.26	0.93
2014	22	11	0.81	0.96
2014	44	11	1.86	0.96
2014	46	11	2.23	0.61

Table A2: Proven fixed matches (anonymised) together with the largest quantile residual (betting volume model) and the largest quantile of the odds ratio (odds model) across the betting types considered.

B Additional details on Chapter 4

B.1 Gradient boosting GAMLSS

The following brief overview of gradient boosting is based on *Mayr et al.* (2012a). For a two-parameter distribution of the response variable, the aim is to find estimates for the

additive predictors $\hat{\eta}_k$, $k = 1,2$, that optimise a loss function ρ, which is the negative (log-)likelihood of the assumed distribution for the response variable. In addition, for each additive predictor k and covariate $j, j = 1,\ldots,J_k$, we specify so-called base learners $g_{kj}(\cdot)$. A base learner is a regression function which determines the type of effect that is assumed for a certain covariate, e.g. a linear base learner (which would lead to $g_{kj}(\mathbf{x}_{kj}) = \beta_{kj}\mathbf{x}_{kj}$) or a P-Spline (*Eilers and Marx*, 1996) base learner. In our implementation, one separate base learner is associated with each covariate and for both additive predictors the same set of base learners is used. The steps detailed in the following are conducted iteratively for the additive predictors $\hat{\eta}_1$ and $\hat{\eta}_2$, updating the predictor considered before moving on to the next predictor, and skipping $\hat{\eta}_k$ only if the corresponding maximum number of boosting iterations, $m_{\text{stop},k}$, has been reached. When applying gradient boosting, the first step is to compute the negative gradient vector, which for the m–th boosting iteration is defined as

$$\mathbf{u}_k^{[m]} = \left(-\frac{\partial}{\partial \eta_k} \rho\left(y_i, \hat{\eta}_1^{[m-1]}, \hat{\eta}_2^{[m-1]}\right) \right)_{i=1,\ldots,n}.$$

In the next step, the base learners $g_{kj}(\cdot)$ are fitted separately to the negative gradient vector, i.e.

$$\mathbf{u}_k^{[m]} = g_{kj}(\mathbf{x}_{kj}) + \boldsymbol{\varepsilon}_{kj},$$

resulting in $\hat{g}_{k1}^{[m]}, \ldots, \hat{g}_{kJ}^{[m]}$. Subsequently, the best base learner in terms of improvement of the residual sum of squares is selected. We denote the selected base learner by $\hat{g}_{kj^\star}^{[m]}$. Finally, the resulting additive predictor $\hat{\eta}_{\lambda_k}^{[m]}$ is updated, such that

$$\hat{\eta}_k^{[m]} = \hat{\eta}_k^{[m-1]} + \nu \cdot \hat{g}_{kj^\star}^{[m]},$$

where $0 < \nu \leq 1$ is a step-length tuning parameter, typically chosen as $\nu = 0.1$. In each iteration only the base learner with the best fit is selected, such that some base learners may never be updated, leading to automated variable selection. Boosting in the GAMLSS framework is implemented in the R-Package `gamboostLSS` (version 2.0-0). For a more detailed description of the boosting algorithm in `gamboostLSS` as well as a corresponding tutorial, see *Mayr et al.* (2012a) and *Hofner et al.* (2016), respectively.

We apply early stopping in the boosting-based estimation procedure, which means that the boosting algorithm does not run until convergence, resulting in more stable

predictions. In addition, early stopping reduces the model complexity for the distribution parameters (*Mayr et al.*, 2012b). In practice, the early-stopping algorithm uses cross-validation to calculate the predictive risk (here: the negative log-likelihood) for values of $\mathbf{m}_{\text{stop}} = (m_{\text{stop},1}, m_{\text{stop},2})$ from a grid of reasonable values. We then use the $\mathbf{m}_{\text{stop}}$ that minimises the predictive risk.

Derivatives for the bivariate Poisson distribution

This paragraph provides the derivatives of the negative log-likelihood of the bivariate Poisson distribution. For simplicity, all equations are derived for a single bivariate observation. The implementation of the bivariate Poisson distribution within the R package `gamboostLSS` is uploaded to GitHub (https://github.com/marius-oetting/match-fixing-warning-systems).

The loss function ρ required for gradient boosting is the negative log-likelihood of the bivariate Poisson distribution (see Eq. 4.5):

$$\begin{aligned}\rho(y_1, y_2, \lambda_1, \lambda_2, \lambda_3) =& \lambda_1 + \lambda_2 + \lambda_3 - y_1 \cdot \log(\lambda_1) + \log(y_1!) - y_2 \cdot \log(\lambda_2) \\ &+ \log(y_2!) - \log\left(\sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i! \left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^i\right)\end{aligned}$$

As detailed in Chapter 4, the parameters λ_1, λ_2, and λ_3 are modelled via additive predictors $\eta_{\lambda_1}, \eta_{\lambda_2}$, and η_{λ_3}, respectively, using the log link function. Replacing thus λ_i in the loss function $\rho(\cdot)$ by $\exp(\eta_{\lambda_i}), i = 1,2,3$, gives

$$\begin{aligned}\rho(y_1, y_2, \lambda_1, \lambda_2, \lambda_3) =& \exp(\eta_{\lambda_1}) + \exp(\eta_{\lambda_2}) + \exp(\eta_{\lambda_3}) - y_1 \cdot \eta_{\lambda_1} + \log(y_1!) - y_2 \cdot \eta_{\lambda_2} \\ &+ \log(y_2!) - \log\left(\sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i! \left(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\right)^i\right).\end{aligned}$$

In the following, the negative derivatives with respect to η_{λ_1}, η_{λ_2}, and η_{λ_3} are presented.

$$-\frac{\partial \rho(\cdot)}{\partial \eta_{\lambda_1}} = -\exp(\eta_{\lambda_1}) + y_1 - \sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i \cdot i! \left(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\right)^i \cdot \frac{1}{\sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i! \left(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\right)^i}$$

$$-\frac{\partial \rho(\cdot)}{\partial \eta_{\lambda_2}} = -\exp(\eta_{\lambda_2}) + y_2 - \sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i \cdot i! \Big(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\Big)^i \cdot \frac{1}{\sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i! \Big(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\Big)^i}$$

$$-\frac{\partial \rho(\cdot)}{\partial \eta_{\lambda_3}} = -\exp(\eta_{\lambda_3}) + \sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i \cdot i! \Big(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\Big)^i \cdot \frac{1}{\sum_{i=0}^{\min(y_1,y_2)} \binom{y_1}{i}\binom{y_2}{i} i! \Big(\frac{\exp(\eta_{\lambda_3})}{\exp(\eta_{\lambda_1})\exp(\eta_{\lambda_2})}\Big)^i}$$

B.2 Classification results for cut-off values chosen via the PPV

Since the classification results based on the Youden index as presented in Chapter 4 leads to a high FPR due to the unbalanced data of only a few matches for which a fix has been proven, we also present the classification based on the positive predicted values (PPV). For the classification based on betting volumes, applying the PPV, we choose the cut-off value for the corresponding quantile which maximises the PPV. This leads to a cut-off value for the quantile of 3.61, implying a PPV of 0.077 and an NPV of 0.99. Hence, according to this result, if a match is predicted as fixed it is only in about 7.7% of all cases a proven fixed match. However, the high NPV implies that if a match is classified as non-fixed, there is a high chance that the corresponding match is indeed not fixed.

Regarding the classification based on betting odds, when applying the maximum of the PPV as a measure to find the optimal cut-off, 151 matches are flagged, from which five are actually fixed proven matches. This yields a PPV of 0.033, i.e. if a match is predicted as fixed, it is in about 3% a proven fixed match. At the same time, the corresponding NPV shows — as for the classification based on betting volumes — a value of 0.99. When comparing the PPV and NPV values of both approaches, the NPV is the same for both models, whereas the PPV for the classification based on betting volumes is slightly higher (about 4 percentage points).

For the approach combining odds and volumes, the choice of the cut-off value based on the PPV leads to a maximum PPV of 0.039, corresponding to an optimal cut-off

value for the quantile residuals from the betting volume model of 2.00 and 0.99 for the quantiles of the odds fraction. At the same time, the NPV is 0.99. However, compared to the classification based only relying on betting volumes, the PPV is lower whereas the NPV is at about the same level. Compared to the classification relying on odds solely, the PPV is increased whereas the NPV is also for this classification procedure on the same level.

Thus, applying the PPV measure to find an optimal cut-off value leads — compared to the classification based on the Youden index — to less flagged matches and hence to a slightly higher predictive power for fixed matches. Regarding the latter, the PPV lies between 0.033 and 0.077, whereas for the classification based on the Youden index the PPV varies between 0.012 and 0.016. However, at the same time, the NPV is always about 0.99. Hence, when comparing these measures, there is only a slight difference in the PPV regarding both approaches. The slightly higher PPV for the PPV-based approach arises mainly due to flagging only very few matches as fixed compared to the approach based on Youden's index (cf. Tables 4.6 and A3, respectively). Thus, as the PPV-based approach flags less matches, the TPR is considerably lower than for the classification by the Youden index. However, this also leads to a much decreased FPR. Hence, as discussed above in Section 4.3.4, practical considerations will guide the choice of the cut-off values by deciding whether a high number of matches flagged as fixed is applicable.

	Actual: fixed	actual: normal	sum
predicted: fixed	1 / 5 / 10	12 / 146 / 245	13 / 151 / 255
predicted: normal	23 / 19 / 14	3,183 / 3,049 / 2,950	3,206 / 3,068 / 2,964
sum	24	3,195	3,219

Table A3: Confusion matrix for flagged matches due to betting volumes / betting odds / combined approach based on cut-off values chosen via the PPV.

C Additional results for Chapter 7

C.1 Coefficients in the model for Borussia Dortmund

Table A4: Estimates of the coefficients determining the state transition probabilities as functions of covariates, in the final three-state Clayton copula HMM for the Borussia Dortmund data; 95% confidence intervals in brackets.

	1→2	1→3	2→1	2→3	3→1	3→2
intercept	-1.447	-7.749	-1.918	-4.922	-1.474	-4.111
	[-1.844; -1.049]	[-12.14; -3.362]	[-2.754; -1.082]	[-7.339; -2.505]	[-2.147; -0.801]	[-6.430; -1.791]
score difference	0.074	1.310	0.812	-4.504	-0.240	-0.410
	[-0.207; 0.355]	[-0.140; 2.760]	[0.197; 1.426]	[-7.993; -1.015]	[-0.803; 0.324]	[-0.952; 0.133]
home	0.099	0.412	1.101	-0.553	-0.228	0.763
	[-0.412; 0.610]	[-2.051; 2.875]	[0.234; 1.968]	[-2.233; 1.128]	[-1.315; 0.858]	[-1.064; 2.590]
market value	0.634	4.403	-1.823	3.211	0.312	0.047
	[0.279; 0.989]	[1.721; 7.086]	[-2.955; -0.690]	[1.438; 4.983]	[-0.110; 0.733]	[-0.830; 0.925]
minute	-0.104	6.239	-1.318	4.451	0.278	2.148
	[-0.443; 0.235]	[2.483; 9.995]	[-1.876; -0.760]	[1.231; 7.670]	[-0.225; 0.780]	[0.905; 3.391]

C.2 Additional analysis of Hannover 96 data

For the analysis of Hannover 96, we use the same copula-based HMM model formulation as in Chapter 7 for Borussia Dortmund. The state-dependent distributions for the fitted baseline model are shown in Figure A1. As for Borussia Dortmund, the choice of the copula function considered does not seem to change the shape of the distribution remarkably. Compared to the state-dependent distributions of Borussia Dortmund (see Figure 7.3), Hannover 96 has less number of ball touches and shots on goal, which is intuitively plausible. For all copulas considered, state 1 refers to a high level of control, whereas state 2 can be interpreted as a low level of control.

To select a model for Hannover 96, we again compare the AIC and BIC values for different number of states and copulas, which are shown in Table A5. For the model selected by the BIC, i.e. the AMH-copula-based HMM with two states, the transition probabilities as functions of the covariate minute are shown in Figure A2. As chosen above for Borussia Dortmund, the values for the score difference and the market value of the opponent are fixed at 0 and 200, respectively. According to the estimated effects, staying in state 1 (high level of control) becomes less likely at the end of such matches, whereas staying in state 2 (low level of control) becomes more likely. The stationary distributions for given values of the score difference are shown in Table A6. The values of the minute and the market value of the opponent are again fixed at 80 and 200, respectively. We see that the probability for being in state 1 (high-control state) increases if Hannover is trailing. If the score is even or if they are leading, it is

Table A5: AIC and BIC for copula-based HMMs with different numbers of states (Hannover 96).

	Frank		Clayton		AMH	
	AIC	BIC	AIC	BIC	AIC	BIC
2 states	18,951	**19,030**	19,024	19,103	18,949	**19,027**
3 states	18,949	19,089	18,950	**19,090**	18,948	19,088
4 states	**18,888**	19,101	18,911	19,123	18,920	19,132
5 states	18,891	19,789	**18,899**	19,197	**18,886**	19,184

more likely that they are in state 2 (low control state) than in state 1, which again is intuitively plausible.

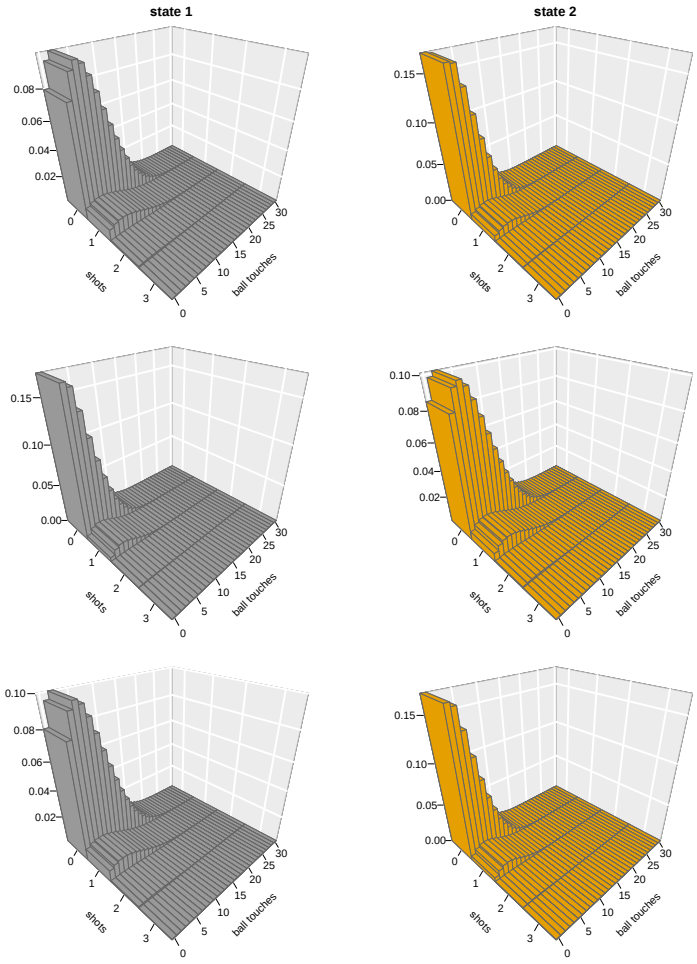


Figure A1: Fitted state-dependent distributions for the baseline two-state HMM for Hannover 96. From top to bottom: Frank-, Clayton- and AMH-copula, respectively.

D Additional results for Chapter 8

In Appendix D, we present the analysis of potential additional variation in the performance in pressure situations across players. This is investigated by analysing throwing performance based on individual throws. While the model presented here provides some insights regarding player-specific performances under pressure, it should be noted that

Table A6: Stationary distributions when fixing the score difference at certain levels. Probabilities were calculated for each value of the score difference, with the market value of the opponent and the minute of the match fixed at 200 and 80, respectively, corresponding to situations in the final stage of a match against an opponent team of average strength.

	-4	-3	-2	-1	0	1	2	3
state 1	0.638	0.642	0.626	0.539	0.320	0.111	0.028	0.006
state 2	0.362	0.358	0.374	0.461	0.680	0.889	0.972	0.994

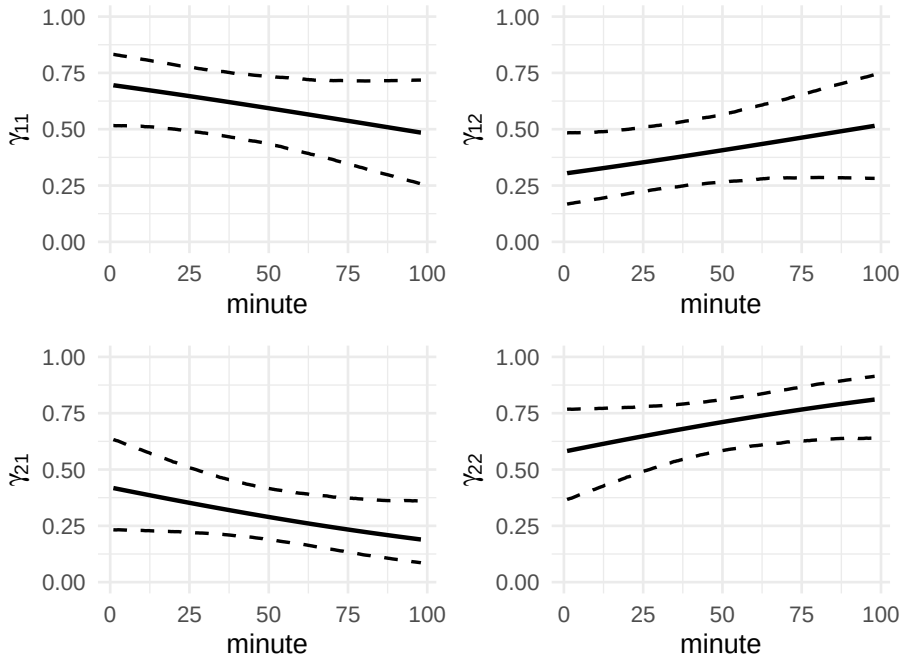


Figure A2: Transition probabilities as functions of the covariate minute.

it does not yield an improvement in the AIC compared to the individual-throw model considered above. To analyse scores which are of about the same difficulty, we consider the scores 2, 8, 16, 22, 32 and 36. The corresponding checkout proportion of these scores with the third dart of a turn vary between 0.408 and 0.476.[1] Considering these finishes for third throws where the opponent also had a finish accounts for $n = 4,773$ single dart throws. A first comparison of the performance under pressure situation between players is investigated in Figure A3. The colours indicate whether the opponent also has a remaining score of 2, 8, 16, 22, 32 or 36, thus indicating pressure situations for the player (denoted by *oppcanfinish* below). Remarkably, there are substantial differences between the players. To extend the model formulation considered above, we include additional zero-mean random effects, γ_{1i}, which represent the player-specific deviations from the fixed effect of *oppcanfinish*, leading to the following

[1]The checkout proportion for all scores which can be finished with a single dart vary between 0.231 (34 points) and 0.476 (2 points). To make the throws comparable, we restrict our analysis to the above mentioned scores with checkout proportion of at least 0.4.

linear predictor:

$$\eta_{ij} = \beta_0 + \beta_1 oppcanfinish_{ij} + \beta_2 exper_i + \beta_3 cb_{ij} + \gamma_{0i} + \gamma_{1i} oppcanfinish_{ij}.$$

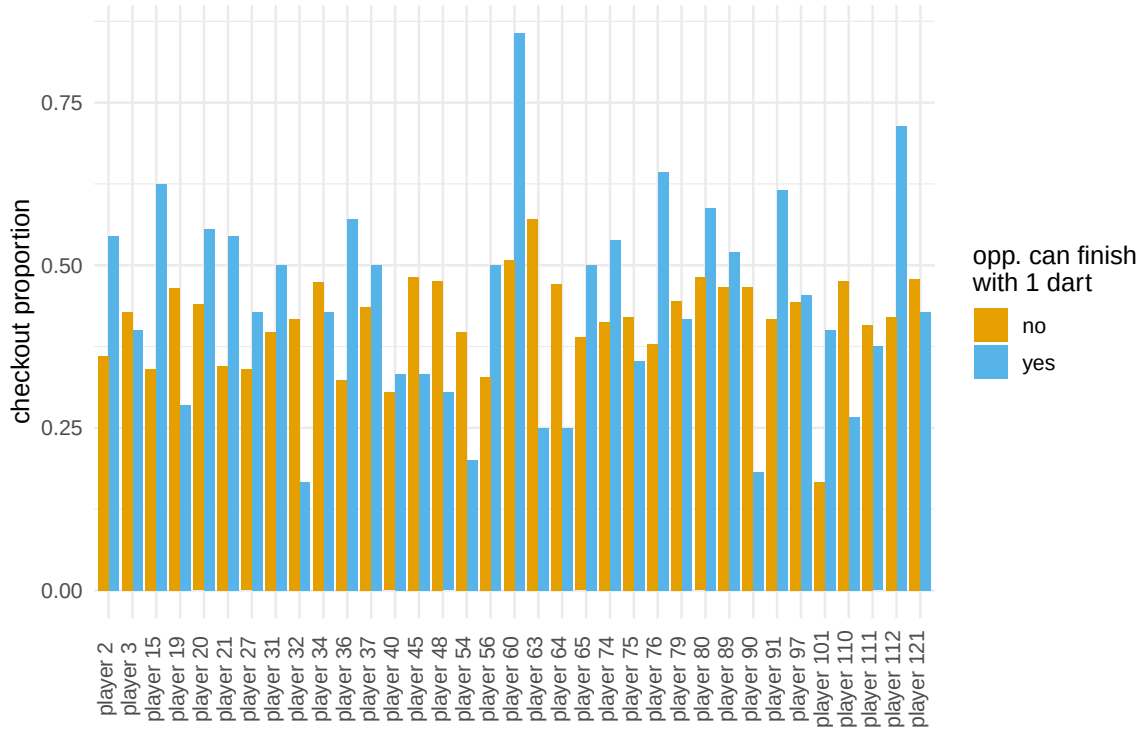


Figure A3: Checkout proportions for situations with 2, 8, 16, 22, 32 or 36 points to checkout before the third throw of a turn. Colours indicate whether the opponent also had 2, 8, 16, 22, 32 or 36 points left. Checkout proportions are shown for players with at least 10 observations in the corresponding subsample, i.e. third throws of non-decider legs with 2, 8, 16, 22, 32 or 36 points left.

As was done also for the previous analyses (see Tables 8.2 and 8.3), we fitted the model to data of all attempts, to non-decider legs, and to decider legs separately. The estimated fixed effects are displayed in Table A7. The particular pressure situation defined above, as indicated by *oppcanfinish*, i.e. the situations where the opponent also has 2, 8, 16, 22, 32 or 36 points left, does not have a statistically significant effect on the checkout performance. The estimated random effects $\hat{\gamma}_{1i}$ are further investigated in Table A8, displaying the sum of the estimated fixed effect of *oppcanfinish*, $\hat{\beta}_1$, and the corresponding player-specific random effect $\hat{\gamma}_{1i}$. As already indicated by Figure A3, the checkout performance in pressure situations varies substantially between players, but the model fit is not improved compared to the models presented above without additional random effects for the performance under pressure.

Table A7: Results of the individual-throw model with random slopes.

	response variable:		
		checkout	
	all attempts	no deciders	deciders
oppcanfinish	0.025	0.037	−1.244
	[−0.155 ; 0.204]	[−0.143 ; 0.218]	[−3.518 ; 1.031]
exper	0.002	0.002	−0.034
	[−0.006 ; 0.009]	[−0.005 ; 0.010]	[−0.097 ; 0.028]
cb	0.426	0.432	0.440
	[0.178 ; 0.673]	[0.183 ; 0.681]	[−1.826 ; 2.706]
constant	−0.473	−0.488	0.202
	[−0.619 ; −0.328]	[−0.635 ; −0.340]	[−0.882 ; 1.286]
observations	4,773	4,698	75
Note:			95% CIs are shown in brackets.

Table A8: Estimated fixed effects of oppcanfinish with the added corresponding random slope.

	$\hat{\beta}_1 + \hat{\gamma}_{1i}$
player with largest performance improvement	0.161
player with 2nd largest performance improvement	0.147
player with 3rd largest performance improvement	0.136
⋮	⋮
player with 3rd largest performance decline	-0.128
player with 2nd largest performance decline	-0.135
player with largest performance decline	-0.139

Bibliography

Abeler, J., A. Falk, L. Goette, and D. Huffman (2011), Reference points and effort provision, *American Economic Review*, *101*(2), 470–492.

Adam, T., A. Mayr, and T. Kneib (2017), Gradient boosting in Markov-switching generalized additive models for location, scale and shape, *arXiv preprint arXiv:1710.02385*.

Albert, J. (1993), A statistical analysis of hitting streaks in baseball: comment, *Journal of the American Statistical Association*, *88*(424), 1184–1188.

Ali, A. (2011), Measuring soccer skill performance: a review, *Scandinavian Journal of Medicine & Science in Sports*, *21*(2), 170–183.

Anderson, J. R. (1982), Acquisition of cognitive skill, *Psychological Review*, *89*(4), 369–406.

Apesteguia, J., and I. Palacios-Huerta (2010), Psychological pressure in competitive environments: evidence from a randomized natural experiment, *American Economic Review*, *100*(5), 2548–2564.

Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar (2009), Large stakes and big mistakes, *The Review of Economic Studies*, *76*(2), 451–469.

Arkes, J. (2013), Misses in "hot hand" research, *Journal of Sports Economics*, *14*(4), 401–410.

Arrondel, L., R. Duhautois, and J.-F. Laslier (2019), Decision under psychological pressure: the shooter's anxiety at the penalty kick, *Journal of Economic Psychology*, *70*, 22–35.

Ashford, K. J., and R. C. Jackson (2010), Priming as a means of preventing skill failure under pressure, *Journal of Sport and Exercise Psychology*, *32*(4), 518–536.

Bar-Eli, M., S. Avugos, and M. Raab (2006), Twenty years of "hot hand" research: review and critique, *Psychology of Sport and Exercise*, *7*(6), 525–553.

Bar-Hillel, M., and W. A. Wagenaar (1991), The perception of randomness, *Advances in Applied Mathematics*, *12*(4), 428–454.

Barron, D., G. Ball, M. Robins, and C. Sunderland (2018), Artificial neural networks and player recruitment in professional soccer, *PloS one*, *13*(10), e0205818.

Bartolucci, F., and T. B. Murphy (2015), A finite mixture latent trajectory model for modeling ultrarunners' behavior in a 24-hour race, *Journal of Quantitative Analysis in Sports*, *11*(4), 193–203.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015), Fitting linear mixed-effects models using lme4, *Journal of Statistical Software*, *67*(1), 1–48.

Baumeister, R. F. (1984), Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance, *Journal of Personality and Social Psychology*, *46*(3), 610–620.

Baumeister, R. F., and C. J. Showers (1986), A review of paradoxical performance effects: choking under pressure in sports and mental tests, *European Journal of Social Psychology*, *16*(4), 361–383.

Baumeister, R. F., and A. Steinhilber (1984), Paradoxical effects of supportive audiences on performance under pressure: the home field disadvantage in sports championships, *Journal of Personality and Social Psychology*, *47*(1), 85–93.

Beckmann, J., P. Gröpel, and F. Ehrlenspiel (2013), Preventing motor skill failure through hemisphere-specific priming: cases from choking under pressure, *Journal of Experimental Psychology: General*, *142*(3), 679–691.

Beilock, S. L. (2010), *Choke: What the Secrets of the Brain Reveal About Getting It Right When You Have To*, New York: Simon and Schuster.

Beilock, S. L., and T. H. Carr (2001), On the fragility of skilled performance: what governs choking under pressure?, *Journal of Experimental Psychology: General*, *130*(4), 701–725.

Beilock, S. L., and R. Gray (2007), Why do athletes choke under pressure?, in *Handbook of Sport Psychology*, edited by G. Tenenbaum and R. C. Eklund, pp. 425–444, Hoboken: John Wiley & Sons Inc.

Benz, M.-A., L. Brandes, and E. Franck (2009), Do soccer associations really spend on a good thing? Empirical evidence on heterogeneity in the consumer response to match uncertainty of outcome, *Contemporary Economic Policy*, *27*(2), 216–235.

Blascovich, J., W. B. Mendes, S. B. Hunter, and K. Salomon (1999), Social "facilitation" as challenge and threat, *Journal of Personality and Social Psychology*, *77*(1), 68–77.

Bocskocsky, A., J. Ezekowitz, and C. Stein (2014), The hot hand: a new approach to an old "fallacy", *8th Annual MIT Sloan Sports Analytics Conference*.

Boin, A., E. Stern, and B. Sundelius (2016), *The Politics of Crisis Management: Public Leadership Under Pressure*, Cambridge: Cambridge University Press.

Bond, C. F., and L. J. Titus (1983), Social facilitation: a meta-analysis of 241 studies, *Psychological Bulletin*, *94*(2), 265–292.

Borghesi, R. (2007), The late-season bias: explaining the NFL's home-underdog effect, *Applied Economics*, *39*(15), 1889–1903.

Borland, J., and R. MacDonald (2003), Demand for sport, *Oxford Review of Economic Policy*, *19*(4), 478–502.

Bornkamp, B., A. Fritsch, O. Kuss, and K. Ickstadt (2009), Penalty specialists among goalkeepers: a nonparametric Bayesian analysis of 44 years of German Bundesliga, in *Statistical Inference, Econometric Analysis and Matrix Algebra*, edited by B. Schipp and W. Kräer, pp. 63–76, Heidelberg: Physica-Verlag.

Brown, J. (2011), Quitters never win: the (adverse) incentive effects of competing with superstars, *Journal of Political Economy*, *119*(5), 982–1013.

Brunel, N., and W. Pieczynski (2005), Unsupervised signal restoration using hidden Markov chains with copulas, *Signal Processing*, *85*(12), 2304–2315.

Buraimo, B., D. Forrest, and R. Simmons (2010), The 12th man? Refereeing bias in English and German soccer, *Journal of the Royal Statistical Society (Series A)*, *173*(2), 431–449.

Buraimo, B., R. Simmons, and M. Maciaszczyk (2012), Favoritism and referee bias in European soccer: evidence from the Spanish league and the UEFA Champions League, *Contemporary Economic Policy*, *30*(3), 329–343.

Cain, M., D. Law, and D. Peel (2003), The favourite-longshot bias, bookmaker margins and insider trading in a variety of betting markets, *Bulletin of Economic Research*, *55*(3), 263–273.

Cao, Z., J. Price, and D. F. Stone (2011), Performance under pressure in the NBA, *Journal of Sports Economics*, *12*(3), 231–252.

Cervone, D., A. D'Amour, L. Bornn, and K. Goldsberry (2016), A multiresolution stochastic process model for predicting basketball possession outcomes, *Journal of the American Statistical Association*, *111*(514), 585–599.

Chambers, D. W., J. A. Baglivo, J. E. Ebel, and A. L. Kafka (2012), Earthquake forecasting using hidden Markov models, *Pure and Applied Geophysics*, *169*(4), 625–639.

Clark III, R. D. (2002a), Do professional golfers "choke"?, *Perceptual and Motor Skills*, *94*(3), 1124–1130.

Clark III, R. D. (2002b), Evaluating the phenomenon of choking in professional golfers, *Perceptual and Motor Skills*, *95*(3), 1287–1294.

Colaresi, M. P., and W. R. Thompson (2002), Hot spots or hot hands? Serial crisis behavior, escalating risks, and rivalry, *Journal of Politics*, *64*(4), 1175–1198.

Conlisk, J. (1993), The utility of gambling, *Journal of Risk and Uncertainty*, *6*(3), 255–275.

Conway, R. W., and W. L. Maxwell (1961), A queuing model with state dependent service rates, *The Journal of Industrial Engineering*, *12*(2), 132–136.

Croson, R., and J. Sundali (2005), The gambler's fallacy and the hot hand: empirical data from casinos, *Journal of Risk and Uncertainty*, *30*(3), 195–209.

Davis, J. L., and K. Krieger (2017), Preseason bias in the NFL and NBA betting markets, *Applied Economics*, *49*(12), 1204–1212.

De Angelis, L., and L. J. Paas (2013), A dynamic analysis of stock markets using a hidden Markov model, *Journal of Applied Statistics*, *40*(8), 1682–1700.

DeCaro, M. S., R. D. Thomas, N. B. Albert, and S. L. Beilock (2011), Choking under pressure: multiple routes to skill failure, *Journal of Experimental Psychology: General*, *140*(3), 390–406.

DeRuiter, S. L., R. Langrock, T. Skirbutas, J. A. Goldbogen, J. Calambokidis, A. S. Friedlaender, and B. L. Southall (2017), A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure, *The Annals of Applied Statistics*, *11*(1), 362–392.

Deutscher, C., B. Frick, and J. Prinz (2013), Performance under pressure: estimating the returns to mental strength in professional basketball, *European Sport Management Quarterly*, *13*(2), 216–231.

Deutscher, C., B. Frick, and M. Ötting (2018), Betting market inefficiencies are short-lived in German professional football, *Applied Economics*, *50*(30), 3240–3246.

Dias, J. G., J. K. Vermunt, and S. Ramos (2015), Clustering financial time series: new insights from an extended hidden Markov model, *European Journal of Operational Research*, *243*(3), 852–864.

Diquigiovanni, J., and B. Scarpa (2018), Analysis of association football playing styles: an innovative method to cluster networks, *Statistical Modelling*, *19*(1), 1–27.

Dixon, M. J., and S. G. Coles (1997), Modelling association football scores and inefficiencies in the football betting market, *Journal of the Royal Statistical Society (Series C)*, *46*(2), 265–280.

Dohmen, T. (2008), Do professionals choke under pressure?, *Journal of Economic Behavior & Organization*, *65*(3-4), 636–653.

Dohmen, T., and J. Sauermann (2016), Referee bias, *Journal of Economic Surveys*, *30*(4), 679–695.

Dorsey-Palmateer, R., and G. Smith (2004), Bowlers' hot hands, *The American Statistician*, *58*(1), 38–45.

Duggan, M., and S. D. Levitt (2002), Winning isn't everything: corruption in sumo wrestling, *American Economic Review*, *92*(5), 1594–1605.

Dunn, P. K., and G. K. Smyth (1996), Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Eddelbuettel, D. (2013), *Seamless R and C++ Integration with Rcpp*, New York: Springer.

Edwards, F. R., and M. O. Caglayan (2001), Hedge fund performance and manager skill, *Journal of Futures Markets: Futures, Options, and Other Derivative Products*, *21*(11), 1003–1028.

Eilers, P. H., and B. D. Marx (1996), Flexible smoothing with B-splines and penalties, *Statistical Science*, *11*(2), 89–102.

Elmore, R., and A. Urbaczewski (2018), Hot and cold hands on the PGA tour: do they exist?, *Journal of Sports Analytics*, *4*(4), 275–286.

Engle, R. W. (2002), Working memory capacity as executive attention, *Current Directions in Psychological Science*, *11*(1), 19–23.

Erevelles, S., N. Fukawa, and L. Swayne (2016), Big data consumer analytics and the transformation of marketing, *Journal of Business Research*, *69*(2), 897–904.

Fama, E. F. (1970), Efficient capital markets: a review of theory and empirical work, *The Journal of Finance*, *25*(2), 383–417.

Feddersen, A., B. R. Humphreys, and B. P. Soebbing (2017), Sentiment bias and asset prices: evidence from sports betting markets and social media, *Economic Inquiry*, *55*(2), 1119–1129.

Federbet (2015), *Annual Fixed Matches. Report 2015*, https://bit.ly/333ttPI.

Feri, F., A. Innocenti, and P. Pin (2013), Is there psychological pressure in competitive environments?, *Journal of Economic Psychology*, *39*, 249–256.

Forrest, D. (2012), The threat to football from betting-related corruption, *International Journal of Sport Finance*, *7*(2), 99–116.

Forrest, D., and I. McHale (2015), University of Liverpool study into the FDS, https://bit.ly/38Avd4g.

Forrest, D., and I. McHale (2019), Using statistics to detect match fixing in sport, *Journal of Management Mathematics*, *30*(4), 431–449.

Forrest, D., and R. Simmons (2008), Sentiment in the betting market on Spanish football, *Applied Economics*, *40*(1), 119–126.

Forrest, D., R. Simmons, and B. Buraimo (2005), Outcome uncertainty and the couch potato audience, *Scottish Journal of Political Economy*, *52*(4), 641–661.

Forrest, D., I. McHale, and K. McAuley (2008), "Say it ain't so": betting-related malpractice in sport, *International Journal of Sport Finance*, *3*(3), 156–166.

Franck, E., E. Verbeek, and S. Nüesch (2011), Sentimental preferences and the organizational regime of betting markets, *Southern Economic Journal*, *78*(2), 502–518.

Franks, A., A. Miller, L. Bornn, and K. Goldsberry (2015), Characterizing the spatial structure of defensive skill in professional basketball, *The Annals of Applied Statistics*, *9*(1), 94–121.

Franses, P. H., and R. Paap (2004), *Periodic Time Series Models*, Oxford: Oxford University Press.

Fried, H. O., and L. W. Tauer (2011), The impact of age on the ability to perform under pressure: golfers on the PGA Tour, *Journal of Productivity Analysis*, *35*(1), 75–84.

Friedman, J. H. (2001), Greedy function approximation: a gradient boosting machine, *Annals of Statistics*, *29*(5), 1189–1232.

Gandar, J. M., R. A. Zuber, R. S. Johnson, and W. Dare (2002), Re-examining the betting market on Major League Baseball games: is there a reverse favourite-longshot bias?, *Applied Economics*, *34*(10), 1309–1317.

Gartzke, E., and M. W. Simon (1999), "Hot hand": a critical analysis of enduring rivalries, *The Journal of Politics*, *61*(3), 777–798.

Geen, R. G., and J. J. Gange (1977), Drive theory of social facilitation: twelve years of theory and research, *Psychological Bulletin*, *84*(6), 1267–1288.

Gertheiss, J., and G. Tutz (2010), Sparse modeling of categorial explanatory variables, *The Annals of Applied Statistics*, *4*(4), 2150–2180.

Gilden, D. L., and S. G. Wilson (1995a), On the nature of streaks in signal detection, *Cognitive Psychology*, *28*(1), 17–64.

Gilden, D. L., and S. G. Wilson (1995b), Streaks in skilled performance, *Psychonomic Bulletin & Review*, *2*(2), 260–265.

Gillispie, S. B., and C. G. Green (2015), Approximating the Conway–Maxwell–Poisson distribution normalization constant, *Statistics*, *49*(5), 1062–1073.

Gilovich, T., R. Vallone, and A. Tversky (1985), The hot hand in basketball: on the misperception of random sequences, *Cognitive Psychology*, *17*(3), 295–314.

Gneezy, U., M. Niederle, and A. Rustichini (2003), Performance in competitive environments: gender differences, *The Quarterly Journal of Economics*, *118*(3), 1049–1074.

Goldfeld, S. M., and R. E. Quandt (1973), A Markov model for switching regressions, *Journal of Econometrics*, *1*(1), 3–16.

Gonçalves, B., D. Coutinho, S. Santos, C. Lago-Penas, S. Jiménez, and J. Sampaio (2017), Exploring team passing networks and player movement dynamics in youth association football, *PloS one*, *12*(1), e0171156.

Gramm, M., C. N. McKinney, D. H. Owens, and M. E. Ryan (2007), What do bettors want? Determinants of pari-mutuel betting preference, *American Journal of Economics and Sociology*, *66*(3), 465–491.

Green, B., and J. Zwiebel (2017), The hot-hand fallacy: cognitive mistakes or equilibrium adjustments? Evidence from Major League Baseball, *Management Science*, *64*(11), 4967–5460.

Groll, A., G. Schauberger, and G. Tutz (2015), Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA World Cup 2014, *Journal of Quantitative Analysis in Sports*, *11*(2), 97–115.

Groll, A., T. Kneib, A. Mayr, and G. Schauberger (2018), On the dependency of soccer scores – a sparse bivariate Poisson model for the UEFA European Football Championship 2016, *Journal of Quantitative Analysis in Sports*, *14*(2), 65–79.

Gunn, G., and J. Rees (2008), *Environmental Review of Integrity in Professional Tennis*, London: International Tennis Federation.

Gurarie, E., C. Bracis, M. Delgado, T. D. Meckley, I. Kojola, and C. M. Wagner (2016), What is the animal doing? Tools for exploring behavioural structure in animal movements, *Journal of Animal Ecology*, *85*(1), 69–84.

Härdle, W. K., O. Okhrin, and W. Wang (2015), Hidden Markov structures for dynamic copulae, *Econometric Theory*, *31*(5), 981–1015.

Harkins, S. G. (1987), Social loafing and social facilitation, *Journal of Experimental Social Psychology*, *23*(1), 1–18.

Hastie, T., and R. Tibshirani (1990), *Generalized Additive Models*, London: Chapman & Hall.

Hastie, T., R. Tibshirani, and M. Wainwright (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Boca Raton: Chapman & Hall/CRC.

Heaton, A. W., and H. Sigall (1991), Self-consciousness, self-presentation, and performance under pressure: who chokes, and when?, *Journal of Applied Social Psychology*, *21*(3), 175–188.

Heiny, E. L., and D. Blevins (2011), Predicting the Atlanta Falcons play-calling using discriminant analysis, *Journal of Quantitative Analysis in Sports*, *7*(3), 1–12.

Hendricks, D., J. Patel, and R. Zeckhauser (1993), Hot hands in mutual funds: short-run persistence of relative performance, 1974–1988, *The Journal of Finance*, *48*(1), 93–130.

Hickman, D. C., and N. Metz (2015), The impact of pressure on performance: evidence from the PGA Tour, *Journal of Economic Behavior & Organization*, *116*, 319–330.

Hickman, D. C., C. Kerr, and N. Metz (2019), Rank and performance in dynamic tournaments: evidence from the PGA tour, *Journal of Sports Economics*, *20*(4), 509–534.

Hill, D. M., S. Hanton, S. Fleming, and N. Matthews (2009), A re-examination of choking in sport, *European Journal of Sport Science*, *9*(4), 203–212.

Hill, D. M., S. Hanton, N. Matthews, and S. Fleming (2010), Choking in sport: a review, *International Review of Sport and Exercise Psychology*, *3*(1), 24–39.

Hoerl, A. E., and R. W. Kennard (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, *12*, 55–67.

Hofner, B., A. Mayr, and M. Schmid (2016), gamboostLSS: an R package for model building and variable selection in the GAMLSS framework, *Journal of Statistical Software*, *74*(1), 1–31.

Hofner, B., A. Mayr, N. Fenske, and M. Schmid (2017), *gamboostLSS: boosting methods for GAMLSS models*, R package version 2.0-0.

Huguet, P., M. P. Galvaing, J. M. Monteil, and F. Dumas (1999), Social presence effects in the stroop task: further evidence for an attentional view of social facilitation, *Journal of Personality and Social Psychology*, *77*(5), 1011–1025.

Humphreys, B. R., R. J. Paul, and A. P. Weinbach (2013), Consumption benefits and gambling: evidence from the NCAA basketball betting market, *Journal of Economic Psychology*, *39*, 376–386.

IRIS (2017), *Preventing Criminal Risks Linked to the Sports Betting Market*, Paris: The French Institute for International and Strategic Affairs (IRIS).

Jackson, C. H., L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto (2003), Multistate Markov models for disease progression with classification error, *Journal of the Royal Statistical Society (Series D)*, *52*(2), 193–209.

Jackson, R. C., K. J. Ashford, and G. Norsworthy (2006), Attentional focus, dispositional reinvestment, and skilled motor performance under pressure, *Journal of Sport and Exercise Psychology*, *28*(1), 49–68.

Jagannathan, R., A. Malakhov, and D. Novikov (2010), Do hot hands exist among hedge fund managers? An empirical evaluation, *The Journal of Finance*, *65*(1), 217–255.

Jewell, S., and J. J. Reade (2014), On fixing international cricket matches, *University of Reading, Department of Economics, Discussion Paper No. 113*.

Joash Fernandes, C., R. Yakubov, Y. Li, A. K. Prasad, and T. C. Chan (2020), Predicting plays in the National Football League, *Journal of Sports Analytics*, *6*(1), 35–43.

Jones, G. (2002), What is this thing called mental toughness? An investigation of elite sport performers, *Journal of Applied Sport Psychology*, *14*(3), 205–218.

Jones, G., S. Hanton, and D. Connaughton (2007), A framework of mental toughness in the worlds best performers, *The Sport Psychologist*, *21*(2), 243–264.

Jones, M., C. Meijen, P. J. McCarthy, and D. Sheffield (2009), A theory of challenge and threat states in athletes, *International Review of Sport and Exercise Psychology*, *2*(2), 161–180.

Jones, M. I., and C. Harwood (2008), Psychological momentum within competitive soccer: players' perspectives, *Journal of Applied Sport Psychology*, *20*(1), 57–72.

Jordet, G. (2009), Why do English players fail in soccer penalty shootouts? A study of team status, self-regulation, and choking under pressure, *Journal of Sports Sciences*, *27*(2), 97–106.

Kahn, L. M. (2000), The sports business as a labor market laboratory, *Journal of Economic Perspectives*, *14*(3), 75–94.

Kahneman, D. (2011), *Thinking, Fast and Slow*, New York: Farrar, Straus and Giroux.

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

Kitagawa, G. (1987), Non-Gaussian state-space modeling of nonstationary time series, *Journal of the American Statistical Association*, *82*(400), 1032–1063.

Klein, N., T. Kneib, S. Klasen, and S. Lang (2015), Bayesian structured additive distributional regression for multivariate responses, *Journal of the Royal Statistical Society (Series C)*, *64*(4), 569–591.

Klein Teeselink, B., P. van Loon, R. J. Dave, M. J. van den Assem, and D. van Dolder (2020), Incentives, performance and choking in darts, *Journal of Economic Behavior and Organization*, *169*, 38–52.

Kleine, D., R. M. Sampedro, and S. L. Melo (1988), Anxiety and performance in runners: effects of stress and anxiety on physical performance, *Anxiety Research*, *1*(3), 235–246.

Kocher, M. G., M. V. Lenz, and M. Sutter (2008), Performance under pressure: the case of penalty shootouts in football, in *The Economics and Psychology of the World's Greatest Sport*, edited by C. Schmidt, P. Ayton, and P. Andersson, pp. 61–72, Cambridge: Cambridge Scholars Publishing.

Kocher, M. G., M. V. Lenz, and M. Sutter (2012), Psychological pressure in competitive environments: new evidence from randomized natural experiments, *Management Science*, *58*(8), 1585–1591.

Konow, J. (2000), Fair shares: accountability and cognitive dissonance in allocation decisions, *American Economic Review*, *90*(4), 1072–1091.

Koopman, S. J., and R. Lit (2015), A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League, *Journal of the Royal Statistical Society (Series A)*, *178*(1), 167–186.

Lago-Peñas, C., and A. Dellal (2010), Ball possession strategies in elite soccer according to the evolution of the match-score: the influence of situational variables, *Journal of Human Kinetics*, *25*, 93–100.

Lanchantin, P., J. Lapuyade-Lahorgue, and W. Pieczynski (2011), Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise, *Signal Processing*, *91*(2), 163–175.

Lazear, E. P. (2000), Performance pay and productivity, *American Economic Review*, *90*(5), 1346–1361.

Lee, P., R. Chen, and V. Lakshman (2017), Predicting offensive play types in the National Football League, https://tinyurl.com/ya3q4w6q.

Lehman, D. W., and J. Hahn (2013), Momentum and organizational risk taking: evidence from the National Football League, *Management Science*, *59*(4), 852–868.

Lewis, B. P., and D. E. Linder (1997), Thinking about choking? Attentional processes and paradoxical performance, *Personality and Social Psychology Bulletin*, *23*(9), 937–944.

Liu, L., Y. Wang, R. Sinatra, C. L. Giles, C. Song, and D. Wang (2018), Hot streaks in artistic, cultural, and scientific careers, *Nature*, *559*(7714), 396–399.

Lopez, M. J., G. J. Matthews, and B. S. Baumer (2018), How often does the best team win? A unified approach to understanding randomness in North American sport, *The Annals of Applied Statistics*, *12*(4), 2483–2516.

MacDonald, I. L., and F. Bhamani (2018), A time-series model for underdispersed or overdispersed counts, *The American Statistician*, DOI: 10.1080/00031305.2018.1505656.

Maher, M. J. (1982), Modelling association football scores, *Statistica Neerlandica*, *36*(3), 109–118.

Markman, A. B., W. T. Maddox, and D. A. Worthy (2006), Choking and excelling under pressure, *Psychological Science*, *17*(11), 944–948.

Massey, C., and R. H. Thaler (2013), The loser's curse: decision making and market efficiency in the National Football League draft, *Management Science*, *59*(7), 1479–1495.

Masters, R. S. W. (1992), Knowledge, knerves and know-how: the role of explicit versus implicit knowledge in the breakdown of a complex motor skill under pressure, *British Journal of Psychology*, *83*(3), 343–358.

Mayr, A., N. Fenske, B. Hofner, T. Kneib, and M. Schmid (2012a), Generalized additive models for location, scale and shape for high dimensional data: a flexible approach based on boosting, *Journal of the Royal Statistical Society (Series C)*, *61*(3), 403–427.

Mayr, A., B. Hofner, and M. Schmid (2012b), The importance of knowing when to stop, *Methods of Information in Medicine*, *51*(2), 178–186.

McEwan, D., K. A. M. Ginis, and S. R. Bray (2013), The effects of depleted self-control strength on skill-based task performance, *Journal of Sport and Exercise Psychology*, *35*(3), 239–249.

McHale, I., and P. Scarf (2011), Modelling the dependence of goals scored by opposing teams in international soccer matches, *Statistical Modelling*, *11*(3), 219–236.

McKay, B., R. Lewthwaite, and G. Wulf (2012), Enhanced expectancies improve performance under pressure, *Frontiers in Psychology*, *3*, 1–5.

Meinshausen, N. (2007), Relaxed lasso, *Computational Statistics & Data Analysis*, *52*(1), 374–393.

Mesagno, C., J. T. Harvey, and C. M. Janelle (2012), Choking under pressure: the role of fear of negative evaluation, *Psychology of Sport and Exercise*, *13*(1), 60–68.

Miller, J. B., and A. Sanjurjo (2014), A cold shower for the hot hand fallacy, *IGIER Working Paper Series*, *Paper No. 518*.

Miller, J. B., and A. Sanjurjo (2018), Surprised by the hot hand fallacy? A truth in the law of small numbers, *Econometrica*, *86*(6), 2019–2047.

Nelsen, R. B. (2006), *An Introduction to Copulas*, New York: Springer.

Nikoloulopoulos, A. K. (2013), Copula-based models for multivariate discrete response data, in *Copulae in Mathematical and Quantitative Finance*, edited by P. Jaworski, F. Durante, and W. K. Härdle, pp. 231–249, New York: Springer.

Oelker, M. R., and G. Tutz (2017), A uniform framework for the combination of penalties in generalized structured models, *Advances in Data Analysis and Classification*, *11*(1), 97–120.

Ottaviani, M., and P. N. Sørensen (2008), The favorite-longshot bias: an overview of the main explanations, in *Handbook of Sports and Lottery Markets*, edited by W. T. Ziemba and D. B. Hausch, pp. 83–101, Amsterdam: North Holland.

Otten, M. (2009), Choking vs. clutch performance: a study of sport performance under pressure, *Journal of Sport and Exercise Psychology*, *31*(5), 583–601.

Ötting, M., C. Deutscher, and R. Langrock (2017), Detecting match-fixing in the Italian Serie B using flexible regression, *Proceedings of the 32nd IWSM*, *Vol. 2*, 243–246.

Ötting, M., C. Deutscher, R. Langrock, and V. Leos-Barajas (2018a), The hot hand in professional darts, *Proceedings of the 33rd IWSM*, *Vol. 1*, 231–236.

Ötting, M., R. Langrock, and C. Deutscher (2018b), Integrating multiple data sources in match-fixing warning systems, *Statistical Modelling*, *18*(5-6), 483–504.

Ötting, M., , R. Langrock, and A. Maruotti (2019), A copula-based multivariate hidden Markov model for modelling momentum in football, *Proceedings of the 34th IWSM*, *Vol. 1*, 125–129.

Ötting, M., C. Deutscher, S. Schneemann, R. Langrock, S. Gehrmann, and H. Scholten (2020a), Performance under pressure in skill tasks: an analysis of professional darts, *PloS one*, *15*(2), e0228870.

Ötting, M., R. Langrock, C. Deutscher, and V. Leos-Barajas (2020b), The hot hand in professional darts, *Journal of the Royal Statistical Society (Series A)*, *183*(2), 565–580.

Paarsch, H. J., and B. S. Shearer (1999), The response of worker effort to piece rates: evidence from the British Columbia tree-planting industry, *Journal of Human Resources*, *34*(4), 643–667.

Paarsch, H. J., and B. S. Shearer (2000), Piece rates, fixed wages, and incentive effects: statistical evidence from payroll records, *International Economic Review*, *41*(1), 59–92.

Patterson, T. A., M. Basson, M. V. Bravington, and J. S. Gunn (2009), Classifying movement behaviour in relation to environmental conditions using hidden Markov models, *Journal of Animal Ecology*, *78*(6), 1113–1123.

Paul, R. J., and A. P. Weinbach (2010), The determinants of betting volume for sports in North America: evidence of sports betting as consumption in the NBA and NHL, *International Journal of Sport Finance*, *5*(2), 128–140.

Paul, R. J., and A. P. Weinbach (2013a), Determinants of dynamic pricing premiums in Major League Baseball, *Sport Marketing Quarterly*, *22*(3), 152–165.

Paul, R. J., and A. P. Weinbach (2013b), Baseball: a poor substitute for football – more evidence of sports gambling as consumption, *Journal of Sports Economics*, *14*(2), 115–132.

Pohle, J., R. Langrock, F. M. van Beest, and N. M. Schmidt (2017), Selecting the number of states in hidden Markov models: pragmatic solutions illustrated using animal movement, *Journal of Agricultural, Biological and Environmental Statistics*, *22*(3), 270–293.

Pollard, R. (2008), Home advantage in football: a current review of an unsolved puzzle, *The Open Sports Sciences Journal*, *1*(1), 12–14.

Prendergast, C. (1999), The provision of incentives in firms, *Journal of Economic Literature*, *37*(1), 7–63.

Punzo, A., S. Ingrassia, and A. Maruotti (2018), Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population, *Statistics in Medicine*, *37*(19), 2797–2808.

R Core Team (2019), *R: A Language and Environment for Statistical Computing*, Vienna: R Foundation for Statistical Computing.

Raab, M., B. Gula, and G. Gigerenzer (2012), The hot hand exists in volleyball and is used for allocation decisions, *Journal of Experimental Psychology: Applied*, *18*(1), 81–94.

Reade, J. J., and S. Akie (2013), Using forecasting to detect corruption in international football, *Proceedings of the 4th International Conference on Mathematics in Sport*.

Richardson, P. A., W. Adler, and D. Hankes (1988), Game, set, match: psychological momentum in tennis, *The Sport Psychologist*, *2*(1), 69–76.

Rigby, R. A., and D. M. Stasinopoulos (2005), Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society (Series C)*, *54*(3), 507–554.

Roberts, A., N. Roche, C. Jones, and M. Munday (2016), What is the value of a Premier League football club to a regional economy?, *European Sport Management Quarterly*, *16*(5), 575–591.

Romer, D. (2006), Do firms maximize? Evidence from professional football, *Journal of Political Economy*, *114*(2), 340–365.

Rosen, S. (1986), Prizes and incentives in elimination tournaments, *American Economic Review*, *76*(4), 701–715.

Rossi, A., L. Pappalardo, P. Cintia, F. M. Iaia, J. Fernández, and D. Medina (2018), Effective injury forecasting in soccer with GPS training data and machine learning, *PloS one*, *13*(7), e0201264.

Rue, H., and O. Salvesen (2000), Prediction and retrospective analysis of soccer matches in a league, *Journal of the Royal Statistical Society (Series D)*, *49*(3), 399–418.

Samuelson, P. A. (1952), Probability, utility, and the independence axiom, *Econometrica*, *20*(4), 670–678.

Sanders, S., and B. Walia (2012), Shirking and "choking" under incentive-based pressure: a behavioral economic theory of performance production, *Economics Letters*, *116*(3), 363–366.

Schauberger, G., A. Groll, and G. Tutz (2018), Analysis of the importance of on-field covariates in the German Bundesliga, *Journal of Applied Statistics*, *45*(9), 1561–1578.

Schrodt, P. A. (2006), Forecasting conflict in the Balkans using hidden Markov models, in *Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention*, edited by R. Trappl, pp. 161–184, Dordrecht: Kluwer Academic.

Sellers, K., T. Lotze, and A. Raim (2018), *COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression*, R package version 0.6.1.

Shea, S. (2014), In support of a hot hand in professional basketball and baseball, *PsyCh Journal*, *3*(2), 159–164.

Sherlock, C., T. Xifara, S. Telfer, and M. Begon (2013), A coupled hidden Markov model for disease interactions, *Journal of the Royal Statistical Society (Series C)*, *62*(4), 609–627.

Sklar, M. (1959), Fonctions de repartition an dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris*, *8*, 229–231.

Solomonov, Y., S. Avugos, and M. Bar-Eli (2015), Do clutch players win the game? Testing the validity of the clutch player's reputation in basketball, *Psychology of Sport and Exercise*, *16*, 130–138.

Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. Van De Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, and A.-L. Børresen-Dale (2001), Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications, *Proceedings of the National Academy of Sciences*, *98*(19), 10,869–10,874.

Stasinopoulos, M. D., R. A. Rigby, G. Z. Heller, V. Voudouris, and F. De Bastiani (2017), *Flexible Regression and Smoothing: Using GAMLSS in R*, Boca Raton: Chapman & Hall/CRC.

Stone, D. F. (2012), Measurement error and the hot hand, *The American Statistician*, *66*(1), 61–66.

Sun, Y. (2004), Detecting the hot hand: an alternative model, *Proceedings of the Annual Meeting of the Cognitive Science Society*, *Vol. 26*.

Szymanski, S. (2003), The economic design of sporting contests, *Journal of Economic Literature*, *41*(4), 1137–1187.

Teich, B., R. Lutz, and V. Kassarnig (2016), NFL play prediction, *arXiv preprint arXiv:1601.00574*.

Thaler, R. H., and C. R. Sunstein (2009), *Nudge: Improving Decisions about Health, Wealth, and Happiness*, London: Penguin.

Thaler, R. H., and W. T. Ziemba (1988), Anomalies: parimutuel betting markets: racetracks and lotteries, *The Journal of Economic Perspectives*, *2*(2), 161–174.

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society (Series B)*, *58*(1), 267–288.

Tibshirani, R. J., A. Price, and J. Taylor (2011), A statistician plays darts, *Journal of the Royal Statistical Society (Series A)*, *174*(1), 213–226.

Toma, M. (2017), Missed shots at the free-throw line: analyzing the determinants of choking under pressure, *Journal of Sports Economics*, *18*(6), 539–559.

Tseng, Y.-T., S. Kawashima, S. Kobayashi, S. Takeuchi, and K. Nakamura (2020), Forecasting the seasonal pollen index by using a hidden Markov model combining meteorological and biological factors, *Science of The Total Environment*, *698*, 134246.

Tversky, A., and D. Kahneman (1971), Belief in the law of small numbers, *Psychological Bulletin*, *76*(2), 105–110.

Tversky, A., and D. Kahneman (1974), Judgment under uncertainty: heuristics and biases, *Science*, *185*(4157), 1124–1131.

van Beest, F. M., S. Mews, S. Elkenkamp, P. Schuhmann, D. Tsolak, T. Wobbe, V. Bartolino, F. Bastardie, R. Dietz, C. von Dorrien, A. Galatius, O. Karlsson, B. McConnell, J. Nabe-Nielsen, M. Tange Olsen, J. Teilmann, and R. Langrock (2019), Classifying grey seal behaviour in relation to environmental variability and commercial fishing activity – a multivariate hidden Markov model, *Scientific Reports*, *9*(1), 5642.

Van Raalte, J. L., B. W. Brewer, B. P. Lewis, D. E. Linder, G. Wildman, and J. Kozimor (1995), Cork! The effects of positive and negative self-talk on dart throwing performance, *Journal of Sport Behavior*, *18*(1), 50–57.

Villar, J. G., and P. R. Guerrero (2009), Sports attendance: a survey of the literature 1973-2007, *Rivista di Diritto e di Economia dello Sport*, *5*(2), 112–151.

Wall, M. M., and R. Li (2009), Multiple indicator hidden Markov model with an application to medical utilization data, *Statistics in Medicine*, *28*(2), 293–310.

Wang, J., D. Marchant, T. Morris, and P. Gibbs (2004), Self-consciousness and trait anxiety as predictors of choking in sport, *Journal of Science and Medicine in Sport*, *7*(2), 174–185.

Wells, B. M., and J. J. Skowronski (2012), Evidence of choking under pressure on the PGA Tour, *Basic and Applied Social Psychology*, *34*(2), 175–182.

Wetzels, R., D. Tutschkow, C. Dolan, S. van der Sluis, G. Dutilh, and E.-J. Wagenmakers (2016), A Bayesian test for the hot hand phenomenon, *Journal of Mathematical Psychology*, *72*, 200–209.

Wolfers, J. (2006), Point shaving: corruption in NCAA basketball, *American Economic Review*, *96*(2), 279–283.

Woodland, L. M., and B. M. Woodland (1994), Market efficiency and the favorite-longshot bias: the baseball betting market, *The Journal of Finance*, *49*(1), 269–279.

Woodland, L. M., and B. M. Woodland (2003), The reverse favourite–longshot bias and market efficiency in Major League Baseball: an update, *Bulletin of Economic Research*, *55*(2), 113–123.

Worthy, D. A., A. B. Markman, and W. T. Maddox (2009), Choking and excelling at the free throw line, *The International Journal of Creativity & Problem Solving*, *19*(1), 53–58.

Wulf, G., and J. Su (2007), An external focus of attention enhances golf shot accuracy in beginners and experts, *Research Quarterly for Exercise and Sport*, *78*(4), 384–389.

Xu, J., and N. Harvey (2014), Carry on winning: the gamblers' fallacy creates hot hand effects in online gambling, *Cognition*, *131*(2), 173–180.

Youden, W. J. (1950), Index for rating diagnostic tests, *Cancer*, *3*(1), 32–35.

Zajonc, R. B. (1965), Social facilitation, *Science*, *149*(3681), 269–274.

Zou, H., T. Hastie, and R. Tibshirani (2007), On the "degrees of freedom" of the lasso, *The Annals of Statistics*, *35*(5), 2173–2192.

Zucchini, W., I. L. MacDonald, and R. Langrock (2016), *Hidden Markov Models for Time Series: An Introduction Using R*, Boca Raton: Chapman & Hall/CRC.

Short CV

Education

- 10/2014 – 02/2017 Statistical Science (M.Sc.), Bielefeld University
- 10/2011 – 09/2014 Economics (B.Sc.), Bielefeld University

Publications

1. **Ötting, M.**, Langrock, R., Deutscher, C., Leos-Barajas, V. (2020) The hot hand in professional darts. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(2), 565–580.
2. **Ötting, M.**, Deutscher, C., Schneemann, S., Langrock, R., Gehrmann, S., Scholten, H. (2020) Performance under pressure in skill tasks: An analysis of professional darts. *PLOS ONE*, 15(3), e0230528.
3. Deutscher, C., **Ötting, M.**, Schneemann, S., Scholten, H. (2019) The demand for English Premier League soccer betting. *Journal of Sports Economics*, 20(4), 556–579.
4. **Ötting, M.**, Langrock, R., Deutscher, C. (2018) Integrating multiple data sources in match-fixing warning systems. *Statistical Modelling*, 18(5-6), 483–504.
5. Deutscher, C., Frick, B., **Ötting, M.** (2018) Betting market inefficiencies are short-lived in German professional football. *Applied Economics*, 50(30), 3240–3246.

Awards

- Publication award for junior scientists from the Department of Business Administration and Economics, Bielefeld University
- DAAD scholarship for the *CMStatistics* 2019 conference in London