Christoph Leitner

## Modeling Consensus and (Dis)agreement in Rating Processes

Thesis

DISSERTATION

# Modeling Consensus and (Dis)agreement in Rating Processes

Christoph Leitner

Mat.Nr.: 0251689

Christoph.Leitner@wu.ac.at

Supervisor: Kurt Hornik (Kurt.Hornik@wu.ac.at)

Institute for Statistics and Mathematics, WU Wien

October 2010

# Danksagung

Nach drei Jahren Arbeit an der WU Wien im Rahmen eines von der Österreichischen Nationalbank geförderten Projekts unter dem Titel "Modellieren von Ratingprozessen und Ratingvalidierung" und im Zuge eines wissenschaftlichen Mitarbeiterpostens am Institut für Statistik und Mathematik konnte ich nun die vorliegende Dissertation abschließen. Die erfolgreiche Beendigung wäre ohne die Hilfe und Unterstützung vieler Kollegen und Freunde nicht möglich gewesen. Zuerst möchte ich all jenen danken, die durch gemeinsame Forschung die wissenschaftliche Basis für diese Arbeit gelegt haben: An erster Stelle danke ich Kurt Hornik, der mir im Oktober 2005 die Möglichkeit gab am oben erwähnten OeNB Projekt mitzuwirken. Dieses Projekt, welches mir vorerst ermöglichte meine Diplomarbeit für mein Magisterstudium zu erarbeiten, erweckte erst mein Interesse an Ratingsystemen. Dieses Themengebiet bildet die Grundlage dieser Dissertation, die von Kurt Hornik als 1. Betreuer betreut wurde. Kurt Hornik stand mir als wissenschaftlicher Mentor immer zur Seite. Er gab mir aber auch die Freiheit meine Forschung nach eigenen Interessen auszurichten. Dieser offene Zugang meines Betreuers zum Thema Forschung gab mir die Möglichkeit meine sportlichen Interessen in diese Arbeit einfliessen zu lassen.

Weiters danke ich meinem 2. Dissertationsbetreuer Stefan Pichler. Stefan Pichler habe ich auch durch das gemeinsame OeNB Projekt kennen und schätzen gelernt. Von beiden Betreuern habe ich gelernt wie man gute Ideen aufgreift und zuwissenschaftlichen Publikationen umsetzt. Trotz aller inhaltlicher Vorgaben gaben mir beide Betreuer immer die Möglichkeit selbständig zu arbeiten.

Besonders danken möchte ich auch Achim Zeileis der mir durch seine in der Regel kritischen aber immer konstruktiven Beiträge sehr viel beigebracht hat. Weiters danke ich Bettina Grün einer weiteren Instituskollegin und Co-Authorin. Außerdem waren sie und Achim Zeileis zu jeder Tages- und Nachtzeit bereit mir bei inhaltlichen sowie auch technischen Problemen weiter zu helfen. Paul Hofmarcher hat mit seiner lusitgen und lockeren Art mein letztes Froschungsjahr an der WU Wien auf besondere Art und Weise belebt. Ich möche ihm sowie allen anderen Institutskollegen danken, dass durch die gute Gemeinschaft am Institut der Unialltag in einer sehr angenehmen und abwechslungsreich Atmosphäre ablief.

Durch das oben erwähnte Froschungsprojekt durfte ich auch noch Rainer Jankowitsch, Manuel Lingo und Gerhard Winkler kennenlernen. Auch dieses Trio hat mich jederzeit unterstützt und meinen Forschungshorizont erweitert.

Alle wissenschaftlichen Leistungen wären aber nicht viel wert, ohne die Menschen, die mir im Leben außerhalb der Universität Kraft und Rückhalt geben. Deshalb möchte ich besonders meiner Freundin Karina Breuer danken, die mich jederzeit unterstützt hat und für mich da war. Auch für extreme Arbeitszeiten zeigte sie Verständnis. Weiters haben auch meine Freunde immer wieder für die notwendige Abwechslung gesorgt. Zu guter Letzt danke ich meiner Familie, ohne die ich hier nicht stehen könnte. Diese hat mich in jeder Hinsicht unterstützt und gab mir so die Basis für all das.

# Abstract

This dissertation introduces a general framework modeling common rating processes in order to aggregate rating information stemming from a variety of raters or rating sources. Ratings play an increasingly important role in our life. They are used to evaluate a variety of objects and activities all over the world. Here we apply our model framework to two different "ratings", the credit ratings and the bookmakers odds. Whereas credit ratings represent the evaluation of credit customers or firms by banks or external rating agencies, bookmakers odds are prospective ratings of the performance of the participating players or teams in a sports competition. Despite the fact that these ratings are used in different kind of areas, both rating systems have a very similar underlying rating process. In both rating processes each rater estimates an underlying numerical variable which represent a probability or is directly related to a probability. In the case of credit ratings this probability is the probability of default (PD) of a credit customer or a firm and in the case of bookmakers odds this probability is the probability of winning a specific sports competition. The proposed model framework is then used to solve the aggregation problem of the two rating processes for different applications yielding different model specifications. Finally, the model results are used to validate the different underlying rating systems as well as for forecasting.

**Keywords:** Credit ratings, bookmakers odds, consensus, agreement.

# Zusammenfassung

Diese Arbeit stellt eine generelle Methode zur Modellierung von bekannten Ratingprozessen dar. Das Ziel ist dabei die Ratinginformation von verschiedenen Ratern oder Ratingquellen zu aggregieren. Ratings spiel heutzutage eine immer wichtigere Rolle. Diese werden rund um den ganzen Globus genutzt um Dinge und Aktivitäten zu evaluieren und zu bewerten. Im Speziellen, wenden wir hier die vorgestellte Methode auf zwei verschiedene Ratings, den Kreditratings und den Buchmacherquoten an. Während bei Kreditratings Banken oder exterene Ratingagenturen Kreditnehmer oder Firmen bewerten, stellen Buchmacherquoten zukünftige Erwartungen der Buchmacher über die Leistungen von Teilnehmern eines Sporttuniers dar. Im Gegensatz dazu, dass diese Ratings in verschiedenen Anwendungsgebieten verwendet werden, haben beide Ratingsysteme einen sehr ähnlichen zugrundeliegenden Ratingprozess. Bei beiden Ratingprozessen schätzt jeder Rater eine zugrundeliegende numerische Variable welche eine Wahrscheinlichkeit repräsentiert. Im Falle der Kreditratings ist diese Wahrscheinlichkeit die Ausfallswahrscheinlichkeit eines Kreditnehmers oder einer Firma und im Falle der Wettquoten die Gewinnwahrscheinlichkeit ein spezielles Sporttunier zu gewinnen. Das hier vorgeschlagene Model wird verwendet um das Aggregationsproblem für diese beiden Ratingprozesse und für verschiedene Anwendungen, welche zu verschiedenen Modelspezifikationen führen, zu lösen. Die Modelergebnisse werden danach genutzt um die zugrundeliegenden Ratingsysteme zu validieren, sowie um Vorhersagen zu machen.

# Contents

# Chapter 1

# Introduction

This dissertation focuses on two common rating instruments, the credit ratings and the bookmakers odds. Analyzing credit ratings as well as their underlying process have been among the most active areas of recent financial research. The analysis of the betting market has also been of increasing interest in the recent economic and sports research. Nevertheless, for these kinds of ratings the literature do not provide a viable methodology to aggregate rating information stemming from different raters/rating sources.

We first motivate the application of credit ratings in Section 1.1. Section 1.2 explains the role of bookmakers odds as sports ratings. Section 1.3 summarizes the literature about consensus and agreement, the measures which are needed to solve the aggregation problem. Section 1.4 gives an overview of this dissertation.

## 1.1 Credit ratings

Credit ratings are one of the most important ratings all over the world. They are used to evaluate the creditworthiness of firms or credit customers. Due to the new regulatory Basel II framework for financial institutions (Bank for International Settlements, 2004) and the subprime crisis along with the

subsequent financial crisis (Rousseau, 2009) the role of credit rating processes, also called credit rating systems has increased.

There are two major parties providing credit ratings, financial institutions (e.g., banks) and external rating agencies. According to the new regulatory Basel II framework banks have several incentives to make use of internal rating systems to estimate risk parameters which are the essential input to calculate their regulatory capital requirements (Bank for International Settlements, 2004). Since modern credit risk pricing applies individual risk parameters, like rating implied default probabilities, the credit ratings of the big three external rating agencies Standard&Poor's, Moody's and Fitch play nowadays an even more prominent role in financial regulations. All big three rating agencies provide ordinal ratings, but they use different rating systems with different granularity as well as different labels (typically, a combination of letters, numbers and/or modifiers). What makes credit ratings comparable is the fact, that in the view of the agencies, likelihood of default is the centerpiece of creditworthiness and therefore, consistent with the goal of an ordinal rating scale, issuers with a lower rating should have a higher probability of default than issuers with a high rating (Coughlin et al., 2009; Erlenmaier, 2006; Cantor and Packer, 1997). There is a growing literature on analysis of the external rating agencies. Krahnen and Weber (2001) present a comprehensive framework for evaluating the quality of standard rating systems and present several principles that ought to be met by an appropriate rating process. Moon and Stotsky (1993) examine the determinants of municipal bond ratings for Moody's and Standard and Poor's. Stolper (2009) studies a principal–agent problem in which a regulator approves credit rating agencies and derives an approval scheme which induces rating agencies to assign correct ratings. Altman and Rijken (2004) give a model for dealing the problem of rating stability of the agencies. Several other academic studies have examined differences in the rating outcomes of rating agencies for a set of issuers (Jewell and Linvingston, 2002; Cantor and Packer, 1995).

Along with the increasing role of rating processes, the validation of credit rating systems has become an important field of research (Krahnen and Weber,

2001; Crouhy et al., 2001). Traditional methods of validating credit rating systems focus primarily on the discriminatory power, i.e., the ability to ex ante distinguish between defaulting and non-defaulting obligors. Among the best-known methods of this type are analyses based on the Accuracy Ratio and the Receiver Operating Characteristic (e.g., Bank for International Settlements, 2005). Unfortunately, these methods do not provide any conclusive information about the accuracy of the PD estimates. Consider a hypothetical rating system which systematically underestimates the true PDs by one half. Obviously, such a rating system, though remarkably inaccurate, has maximal theoretical discriminatory power. On the other hand, in a population of obligors with identical PDs even a perfectly accurate PD estimation will show zero discriminatory power. Recent studies also deal with the shortcomings of the use of these concepts for rating validation (e.g., Lingo and Winkler, 2008). As pointed out by the regulatory bodies (Bank for International Settlements, 2005), validation methods have to aim at directly assessing the *calibration quality*, i.e., the accuracy and reliability of PD estimates (e.g., Stein, 2002).

In contrast to *backtesting* methods where ex ante estimates are compared to ex post realizations our general model framework is based on the existence of contemporaneous ratings for the same obligor provided by different rating sources. Hornik et al. (2007) call such a dataset *multi-rater panel*. Such a *benchmarking* approach is particularly helpful when sufficient default observations are not available or competing rating systems are to be compared (for a discussion of benchmarking approaches see Bank for International Settlements, 2005; Hornik et al., 2007).

## 1.2 Bookmakers odds

In addition to credit ratings, we apply our general model framework to a specific kind of sport ratings, bookmakers odds. Sport ratings or rankings are typically derived by suitably aggregating the competitors' previous performances and are often found to provide predictive power in forecasting tasks. Boulier and Stekler (1999) show that rankings provide forcasting informa-

tion for basketball tournaments and tennis matches. Lebovic and Sigelman (2001) analyze the predictive accuracy of college football rankings. Suzuki and Ohmori (2008) use the FIFA/Coca Cola World rating (Fédération Internationale de Football Association, 2008), one of the most popular rating system in soccer, as a forecasting tool for the last four FIFA World Cups (1994, 1998, 2002, 2006). In addition, Dyte and Clarke (2000) use the FIFA ratings to predict the distribution of scores in international soccer matches. Another popular rating system is the Elo rating system, originally developed to calculate the relative skills of chess players (e.g., Elo, 2008), which has subsequently also been applied to various other sports including soccer. Song et al. (2009) apply it as one method to forecast the winner of single American Football games. Edmans et al. (2007) select important soccer games based on the World Football Elo Ratings.

Bookmakers odds represent a rather different type of rating compared to the methods above. In the course of growing popularity of online sports betting, the analysis of betting markets has been receiving increased interest, often focusing on two types of analyses: (1) testing the forecasting power of the bookmakers, and (2) testing the efficiency of the betting market. The bookmakers publish odds for a variety of players and teams for winning sports competitions and tournaments. Based on the bookmakers' expert judgments (which typically include, but are not limited to, knowledge about past performances) the odds reflect expected outcomes in a particular competition where the bookmakers have strong economic incentives to rate the competitors correctly. A bias (in either direction, too good or too bad) will cost them money, or, in other words, will reduce their profits. Hence, bookmakers can be seen as experts in the matter of sports rating (see Pope and Peel, 1989) and are likely to provide good predictions (Forrest and Simmons, 2000). This is confirmed by various empirical studies in which fixed odds are found to be an efficient forecasting instrument for the outcome of single matches (e.g., Vlastakis et al., 2009; Spann and Skiera, 2009; Song et al., 2007; Forrest et al., 2005b; Dixon and Pope, 2004; Boulier and Stekler, 2003).

One advantage of employing bookmakers odds is that winning probabilities

for the corresponding competition can be derived easily while this is not straightforward for many of the ability ratings. However, if abilities are measured on a ratio scale (or can be transformed to such), winning probabilities for pairwise matches can be derived using the approach of the Bradley and Terry (1952) model. Notable in this respect is the Elo rating from which pairwise winning probabilities for single matches can be obtained (e.g., Stefani and Pollard, 2007; Edmans et al., 2007). Thus, when the competition of interest is a single match, forecasts based on ability ratings and bookmakers odds can be compared easily. The same is not true if the competition is a more complex tournament for which the bookmakers odds, by their prospective nature, can include additional effects such as group draws or seedings.

## 1.3 Consensus information and agreement

By analyzing the two common rating processes (described above) we are interested in the aggregated rating information as well as the agreement across different forecasters, here raters. Therefore, a measure of "consensus" and a measure of "agreement" are needed. Zarnowitz and Lambros (1987) define "consensus" as the degree of agreement among point predictions aimed at the same target by different individuals and "uncertainty" as the diffuseness of the corresponding probability distributions. The "consensus" measure can be computed as the median (Su and Su, 1975) or the mean of all the forecasts in the sample (Zarnowitz and Lambros, 1987). In order to measure "uncertainty" or "disagreement" the standard deviations of the predictive probability distributions could be used (e.g, Clements, 2008; Zarnowitz and Lambros, 1987; Lahiri and Teigland, 1987). These strategies are applied to sports competitions by Song et al. (2009, 2007). There is no similar application to credit ratings. Alternative strategies for the aggregation of forecasts are discussed by Kolb and Stekler (1996); Schnader and Stekler (1991).

## 1.4 Overview of the dissertation

The general aim of this dissertation is to solve the aggregation problem of rating information stemming from a variety of raters or rating sources. In particular, we are interested in the consensus and the agreement information across raters/rating sources. This problem represents a major research issue for different rating areas. In this dissertation we solve this problem for two common rating processes. These two processes are the credit rating process where financial institutions or external rating agencies evaluate credit customers or firms and the bookmakers' evaluations of teams' or players' performances in a sports competition. In order to obtain consensus as well as (dis)agreement information across different raters/rating sources, we investigate a general model framework modeling these two common rating processes.

Chapter 2 presents a general *mixed-effects model* framework (e.g., Pinheiro and Bates, 2000) to obtain consensus as well as (dis)agreement information across raters/rating sources. The notation of this class of model is introduced and the estimation within a frequentist maximum likelihood setting or using Bayesian estimation techniques is described. Furthermore, we specify the different applications as well as the variety of research questions which are answered in the following two chapters.

In Chapter 3 we extend our general model framework into a static latent variable model for the probability of default (PD) of firms or credit customers using PD rating information stemming from a variety of raters of only one time point (see Hornik et al., 2008, 2010) and into a dynamic latent variable model for ordinal ratings using rating information stemming from a variety of raters over a specific time period (see Grün et al., 2010). For both applications we derive consensus ratings as well as information about the rater agreement to validate the underlying rating processes.

Chapter 4 applies the general model framework to bookmakers odds and investigate a general model specification for bookmakers odds in order to forecast the outcome and analyze the bookmakers agreement of three differ-

ent sport tournaments, the UEFA EURO 2008 (see Leitner et al., 2008a,b, 2010a), Wimbledon 2009 (see Leitner et al., 2009c,d), and the UEFA Champions League 2008/09 (see Leitner et al., 2009a,b, 2010b). For all applications the forecasting power of the bookmaker consensus is compared to other rating systems in ex post analysis.

Chapter 5 concludes this dissertation.

# Chapter 2

# General Model framework

The basic assumption of our general model framework is that raters estimate a numerical variable—representing information about the underlying rating subject—in an internal rating process. Due to general informational asymmetry between the rater and the rating subject the rater cannot estimate the "true" numerical variable. This asymmetry can be due to limited access to the existing information, such as incomplete information, or delayed observations of the driving factors. Generally, the modeler is entirely free in the functional specification of the relationship between the estimation error and the "true" numerical variable.

In our model the "true" numerical variable is taken as a latent variable. Rating outcomes from different sources are treated as noisy estimations/observations of this latent variable. A parametric specification of such a model has to include the following components:

- The distribution of the latent variable,
- the formal relation by which the noise or error terms are linked to the latent variable, and
- the distribution of the error terms.

The means and (co)variances of these distributions are the key outcome of

the model. The mean parameters indicate the rating *bias*, i.e., the expected shift of the estimated numerical variable of a certain rating system compared to the average numerical variable across all rating sources. The variance parameters reflect the general size of undirected estimation errors, i.e., they reflect the *precision* of the rating system. Finally, the covariances convey information about potential error dependencies across rating systems.

Denoting the underlying numerical variable as the *rating score* $\ell(r_{ij}(t))$ of the rating $r_{ij}(t)$ of rating subject $i$ by rater $j$ at time $t$, the relationship between the estimated rating score $\ell(r_{ij}(t))$ (estimated in an internal rating process by rater $j$) and the latent rating score $\ell(r_i(t))$ can be written as

$$\ell(r_{ij}(t)) = \ell(r_i(t)) + \epsilon_{ij}(t), \tag{2.1}$$

where $\ell$ is a suitable *link function* transforming the rating to the real axis so that the error terms are related additively to the latent rating score.

This relationship builds our general model framework which can be used to validate the different raters/rating sources by analyzing the rating errors and to compute *consensus* information across all rating systems, which particularly in the case of only partially available rating information, i.e., not for all rating subjects rating observations by each rater or rating source is available, is non trivial (see e.g., Cook et al., 1986, 2007). To the best of our knowledge there is no viable methodology available to obtain consensus ratings for these kinds of data sets. The estimation of consensus ratings is thus one of the major contributions of this dissertation. The obtained consensus information can be then used for forecasting issues. Deviations between observed ratings and consensus ratings can be analyzed on an atomistic case by case basis. Further, these differences can be aggregated on rating source or rating subject levels, which can aid in detecting potential systematic patterns or anomalies in rating behavior.

**Estimation.** In order to estimate the model parameters of the specific models we employ standard maximum likelihood estimation (Lehmann and

Casella, 1998) as well as Bayesian estimation techniques (Spiegelhalter et al., 2002; Carlin and Louis, 2009).

**Application.** The general model framework (see above) offers a variety of model specifications and can therefore be used for many different application. We apply it to the following issues:

- A static latent variable model for PD estimation in order to validate 13 Austrian banks,
- a dynamic latent variable model for ordinal credit ratings in order to validate the big three external credit rating agencies (Standard&Poor's, Moody's and Fitch),
- a general model specification for bookmakers odds in order to forecast the outcome and analyze the bookmakers agreement of the following three sport tournaments
  - UEFA EURO 2008,
  - Wimbledon 2009, and
  - UEFA Champions League 2008/09.

By obtaining consensus as well as (dis)agreement from the specific models we can validate the different raters/rating sources as well as use the aggregated information for forecasting issues. Whereas the applications on credit ratings focus mainly on the validation of the raters, we use the consensus of the bookmakers to predict the outcome of a tournament.

In particular, we try to find answers to the following research questions:

- Are there marked rating differences in the Austrian credit market across banks?
- Are there marked rating differences in the Austrian credit market across industries (or industry/bank combinations)?

- Can some rating differences in the Austrian credit market be explained by the obligors' legal form or exposure size?
- Are there marked rating differences across the big three external credit rating agencies (Standard&Poor's, Moody's and Fitch)?
- Is there a relationship between the consensus ratings of the iTraxx Europe companies and the real markets (e.g., the Dow Jones EURO STOXX 50 index)?
- Is the consensus information across bookmakers a good forecasting tool for the outcome of a specific sports tournament, like the UEFA EURO 2008, Wimbledon 2009, and the UEFA Champions League 2008/09?
- How is the (dis)agreement across bookmakers for a specific tournament, like the UEFA Champions League 2008/09?
- Is there a relationship between the (dis)agreement across bookmakers and some team-specific characteristics (e.g., the teams' association)?

# Chapter 3

# Modeling credit ratings

Credit ratings, evaluating the creditworthiness of firms or credit customers are usually recorded on an ordinal scale and thus are not directly usable measures of the firm's or credit customer's default probability (Carey and Hrycay, 2001). Due to Basel II, it is more and more common for raters, e.g., banks to estimate the probability of default (PD) directly and provide PD ratings (Bank for International Settlements, 2004).

Here, we extend our general model framework (Equation 2.1) into a static latent variable model for PD estimation using PD rating information stemming from a variety of raters of only one time point (Section 3.1) and into a dynamic latent variable model for ordinal ratings using rating information stemming from a variety of raters over a specific time period (Section 3.2). We derive consensus ratings as well as information about the rater agreement to validate the underlying rating processes.

## 3.1 A static latent variable model for PD estimation

### 3.1.1 Model specification

One of the most important credit risk parameters is the probability of default (PD) which is defined to measure the likelihood of the occurrence of a default event for a certain obligor over a one year horizon. Modern credit risk management is crucially based on the risk-adjusted pricing of loans and other credit-risk contingent claims which again heavily relies on a valid and accurate PD estimation methodology. The accuracy of PD estimates is of particular importance for virtually all pricing models for structured credit derivatives. While some pricing models need accurate measures of the average PD of a bond or loan portfolio as an input, some more advanced models require the distribution of individual PDs of such a portfolio which imposes even higher challenges on the validity of the estimation models.

Here, we use our general model framework (Equation 2.1) to assess the accuracy of PD estimates. Therefore, the "true" PD is taken as the latent variable. Rating outcomes from different sources (e.g., banks, rating tools, or rating agencies) are treated as noisy observations of this latent variable, because we assume that the raters cannot observe the "true" PD of the obligor due to informational asymmetry between firm owners and debt holders which constitutes the cornerstone of modern corporate finance (e.g., Leland and Pyle, 1977; Berk and DeMarzo, 2007). Possible reasons for this asymmetry are limited access to the existing information, such as incomplete accounting information (Duffie and Lando, 2001), and delayed observations of the driving risk factors (Guo et al., 2009).

The motivation for the specification used in this context builds on the main concept of *structural* or *firm value* models (e.g., Merton, 1974; Lando, 2004). The standard models assume the firm value to be the only driving factor of credit risk and to follow a Geometric Brownian Motion. As a consequence, an important stylized property of these models is that the probit of the PD

is linear in the natural log of the firm value. Let $V_i$ be the log asset value of firm $i$ and $PD_i$ its probability of default. The basic model of Merton (1974) can be written as

$$PD_i = \Phi(-DD_i), \qquad DD_i = a_i + b_i V_i,$$

where $\Phi$ is the distribution function of the standard normal distribution, $a_i$ and $b_i$ are constants independent of $V_i$, and $DD_i$ is the so-called *distance to default* of firm $i$ (Crosbie and Bohn, 2003; Bharath and Shumway, 2008). Along the lines of Duffie and Lando (2001), we assume that the error in the observation of the firm value is normal and additive to the log of the firm value. This assumption can also be justified by the wide-spread use of structural models for PD estimation in the banking industry which have gained additional importance by the introduction of the Basel II supervisory framework. The most prominent industry model was developed by Moody's KMV (Crosbie and Bohn, 2003) and is used in several extensions and modifications by many financial institutions. In this class of models the distance to default is derived from stock market and accounting data and used as the key input to the PD estimation. It thus seems natural to assume that erroneous observations and incomplete information lead to normally distributed errors which are additive to the distance to default.

The "estimate" $PD_{ij}$ as derived by bank $j$ for the true PD of firm $i$ is thus of the form

$$PD_{ij} = \Phi(-DD_i + \epsilon_{ij})$$

where $\epsilon_{ij}$ is the corresponding error (which depends on $j$ in particular as raters might have access to different information sets for firm $i$). Equivalently,

$$\Phi^{-1}(PD_{ij}) = -DD_i + \epsilon_{ij} = \Phi^{-1}(PD_i) + \epsilon_{ij}$$

A second important class of industry models estimate PDs based on a probit (or logit) regression of observed default indicators on a set of risk characteristics of the firm (e.g., Blume et al., 1998; Nickell et al., 2000). A linear

combination of the risk characteristics of the firm where the resulting regression coefficients are used as weights constitutes the rating score of the firm. The PD estimate is then obtained by transforming the score to the unit interval by the corresponding transformation. In the case of the probit transformation this approach is fully consistent with our framework, because the error term in the regression is normal and additive to the score.

Consolidating the Merton type and the regression approaches, we are led to apply our general model framework (Equation 2.1) and relate the raters' estimated (observed) PDs to the (unobservable) true PDs in the form

$$\ell(PD_{ij}) = \ell(PD_i) + \epsilon_{ij}.$$

for a suitable (strictly monotonically increasing) link function $\ell$ mapping the $(0, 1)$ PD scale to $(-\infty, +\infty)$. Via $\ell$, the PDs are mapped to corresponding scores. On the score scale, the rating errors are modeled additively. Let $S_{ij} = \ell(PD_{ij})$ and $S_i = \ell(PD_i)$ denote the observed and latent scores, respectively. For Merton type models, $\ell = \Phi^{-1}$ and $S_i = -DD_i$. Writing $\mu_{ij}$ and $\sigma_{ij}$ denote the mean and standard deviation of $\epsilon_{ij}$, respectively, the above latent trait model can be written as

$$S_{ij} = S_i + \mu_{ij} + \sigma_{ij} Z_{ij}$$

where the standardized rating errors $Z_{ij} = (\epsilon_{ij} - \mu_{ij})/\sigma_{ij}$ have mean zero and unit variance. We prefer to think of $PD_i$ and hence the corresponding scores $S_i$ as drawn randomly from an underlying obligor population, and assume that rating errors are independent of the true PDs and that true PDs and rating errors are i.i.d. across obligors. (Of course, these assumptions could be relaxed if more involved probabilistic specifications are desired.) We thus obtain a mixed-effects model (e.g., Pinheiro and Bates, 2000) for the observed $S_{ij}$ where the latent true PD scores enter as random effects. We refer to this model as the *latent trait model* for the multi-rater panel of PD estimates.

Given parametric models for the $\mu_{ij}$ and $\sigma_{ij}$ and the distributions of true PD scores and standardized rating errors, the latent trait model can be esti-

mated using e.g. marginal maximum likelihood (provided that the marginal distributions of the $S_{ij}$, i.e., the convolutions of the true PD score and rating error distributions, can be computed well enough), or Bayesian techniques.

A very simple parametric specification of the bias/variance structure of the rating errors is $\mu_{ij} = \mu_j$ and $\sigma_{ij} = \sigma_j$, in which case the rating errors would be independent of the obligors and their characteristics (in particular, their creditworthiness itself). We suggest the employment of flexible models of the form

$$\mu_{ij} = \mu_{g(i),j}, \qquad \sigma_{ij} = \sigma_{g(i),j},$$

where $g(i) \in \{1, \ldots, G\}$ is the group of obligor $i$, for a suitable grouping of obligors relative to which raters exhibit homogeneous rating error characteristics. Note that we use $\mu$ for the means of the rating errors and the respective model parameters, with $\mu_{ij} = \mathbb{E}(\epsilon_{ij})$ and $\mu_{g,j}$ the parameter for group $g$ and rater $j$. The $\sigma$ notation is analogous. Such groups can e.g. be defined by industry, obligor "type", size, and legal form, or combinations thereof. The importance of accounting for industry group effects in the analysis of creditworthiness patterns has e.g. been emphasized in Crouhy et al. (2001).

As

$$\mathbb{E}(S_{ij}) = \mathbb{E}(S_i) + \mu_{g(i),j},$$

conditions relating the rating biases to the mean observed PD scores are required to ensure identifiability. More generally, note that common random effects in the rating errors cannot be separated from the true PD scores. A natural condition consistent with the interpretation as rating bias relative to an underlying unbiased truth is $\sum_j \mu_{g,j} = 0$, i.e., that the raters' average rating bias within each obligor group is zero. With this identifiability constraint, the marginal group effects in the means are absorbed into the means of the latent scores, for which we shall employ the basic model $\mathbb{E}(S_i) = \nu_{g(i)}$ consistent with the identifiability constraint. Possible models for the variances of the PD scores include $\mathrm{var}(S_i) = \tau^2$ (constant for all obligors) or $\mathrm{var}(S_i) = \tau^2_{g(i)}$ (constant for all obligors in the same group).

With these specifications, the consensus score for obligor $i$ is given by $\hat{S}_i$, the estimated random effect in the fitted mixed-effects model. Consensus PD estimates are readily obtained by transforming the consensus scores back to the PD scale using the inverse $\ell^{-1}$ of the link function, i.e., as $\widehat{PD}_i = \ell^{-1}(\hat{S}_i)$. Finally, residuals are the part of the observations unexplained by the model and given by $S_{ij} - \hat{S}_i - \hat{\mu}_{ij}$, the observed scores minus the estimated random and fixed effects.

In what follows, we assume that true PD scores and rating errors have a (multivariate) normal distribution. Possible extensions are discussed in Section 3.1.3. With these assumptions, the latent trait model can be estimated using standard software for mixed-effects models, provided that these allow for sufficiently flexible specifications of the error covariance structure.

Our framework is applicable in the context of banking supervision and development of rating models. Banking supervisors are interested in the parameters of the error terms in order to assess the calibration quality of the internal rating system of a supervised bank. Supervisors might also be interested in the consensus PD estimates for analyzing financial stability of a banking system (see Elsinger et al., 2006). Finally, developers of rating systems such as banks and rating agencies have a natural interest in comparing their outcomes to peers at different stages of the development. Note, however, that as for any benchmarking method our model does only provide information about the relative rather than the absolute rating errors since actual default information is not incorporated.

The model presented here is related to other studies on benchmarking credit rating systems (e.g., Hornik et al., 2007; Stein, 2002; Carey, 2001). In contrast to these contributions, our model explicitly proposes a probabilistic framework which reflects the stochastic nature of the true PDs and the rating errors incurred in the process of PD estimation. This allows to directly estimate parameters reflecting the calibration quality of rating systems and derive consensus PD estimates consistent with the suggested framework.

### 3.1.2 Application

#### Data Description

For the empirical example we employ a data set on rating information provided by Oesterreichische Nationalbank, the Austrian central bank. The data contain rating information (one-year PD estimates) from 13 major Austrian banks on 2090 obligors in September 2007 and cover a significant share of the Austrian credit market. For each obligor, at least two PD estimates are available. The number of *co-ratings* (occurrences of ratings of a single obligor by two different banks) is 5460. In addition to the PD estimates we have cross-sectional information about the obligors, like legal form, industry affiliation and outstanding exposure. Table 3.1 reports descriptive statistics of the data set.

| | **Min.** | **Median** | **Max.** | **Mean** |
|---|---|---|---|---|
| Number of obligors per bank | 70 | 182 | 1700 | 420 |
| Number of ratings per obligor | 2 | 2 | 11 | 2.5 |
| Size of banks measured by their total assets (in Euro billions) | 1.0 | 8.7 | 128.5 | 11.8 |

Table 3.1: Descriptive statistics of the characteristics of the rating information and the 13 Austrian banks in the data set.

Note that even the smallest bank has at least 70 obligors in common with one or more of the other institutions. Apart from looking at the number of co-ratings on a bank level, we also compute the number of co-ratings on an obligor level. The median number of these co-ratings is 2, suggesting that most obligors have business relations to only a small number of banks.

For a deeper analysis we group all obligors by their industry affiliation and their legal form. Based on the NACE codes (European Commission, 2008) we classify obligors to nine main industries. Table 3.2 shows the distribution of the obligors across the industries.

Table 3.2 shows that the total numbers of co-ratings (5460) is not uniformly

| Label | Industry | No. of co-ratings | No. of co-ratings (%) |
|---|---|---|---|
| Manufac | Manufacturing | 938 | 17.2 |
| Energy | Energy & Environment | 180 | 3.3 |
| Constr | Construction | 184 | 3.4 |
| Trading | Trading | 641 | 11.7 |
| Finance | Financial Intermediation | 1737 | 31.8 |
| RealEst | Real Estate & Renting | 754 | 13.8 |
| Public | Public Sector | 344 | 6.3 |
| Service | Service | 435 | 8.0 |
| Private | Private Individuals | 247 | 4.5 |
| Total | | 5460 | 100.0 |

Table 3.2: Distribution of the co-ratings of the 13 Austrian banks across industries.

distributed across the nine industries, ranging from 180 co-ratings in Energy & Environment ("Energy") to 1737 co-ratings in Financial Intermediation ("Finance"). With 13 banks and 9 industries there are 117 possible sub-portfolios to be analyzed. However in 17 of these there are no observations.

In addition the obligors can be grouped with aspect to their legal form yielding that 79.6% of the obligors are limited companies, 12.2% unlimited companies, and 8.2% are private individuals.

Finally, we use information on the banks' relative exposures against each obligor. Relative exposure is measured as the outstanding amount against the obligor expressed as a fraction of the total volume of outstanding loans of a specific bank and serves as a rough indicator for the size of the obligor.

### Results

This section describes the model selection process and the empirical results. The general model class allows for many competing model specifications based on the data structure (see Section 3.1.1). Our selection process yields a model using the industry as grouping variable. We present bank and indus-

try specific analyses based on rating biases and error variances derived from this specification. Furthermore, we analyze the errors from the consensus rating of this model on the obligor level based on exposure and legal form.

**Model Selection and Parameter Estimates.** Equation 2.1 in Section 3.1.1 describes a very general model class allowing for a variety of specifications for the means and variances of the normal distributions of the latent PD scores and rating errors. We use parametric models which group obligors based on the available co-variates (for the data set at hand, industry affiliation, legal form and exposure). For the error distributions, bank effects as well as group/bank interaction terms are considered. We also investigate models allowing for general correlation patterns between rating errors. For each fitted model, we compute the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC, also known as Schwarz's Bayesian criterion). The best model is then selected based on these criteria.

The parameters of the mixed-effects models are estimated via maximum likelihood (e.g., Pinheiro and Bates, 2000) The best model found by the model selection procedure uses industry affiliation as the sole grouping variable and a single variance parameter PD score, and is given by

$$S_{ij} = S_i + \mu_{g(i),j} + \sigma_{g(i),j} Z_{ij}, \qquad S_i \sim N(\nu_{g(i)}, \tau^2), \tag{3.1}$$

where $\mu_{g,j}$ is the rating bias to the mean PD score of bank $j$ for obligors in industry $g$, $\sigma_{g,j}$ is the standard deviation of the rating error of bank $j$ for obligors in industry $g$, and $\nu_g$ is the mean PD score in industry $g$. This model forms the basis for further analysis. Note that the $\mu$ and $\sigma$ parameters are unestimable for industry/bank combinations with no observations.

We begin our analysis of the estimation results by showing the parameters describing the distribution of the true latent scores. These are the industry specific means $\nu_g$ and the standard deviation $\tau$. For ease of interpretation we additionally show the images under the inverse link function of the mean PD scores for each industry and the respective one standard deviation intervals

(see Table 3.3).

| **Industry** | $\nu_g$ | $\Phi(\nu_g)$ | $\Phi(\nu_g - \tau)$ | $\Phi(\nu_g + \tau)$ |
|---|---|---|---|---|
| Manufac | −2.542 | 55.1 | 17.7 | 151.1 |
| Energy | −2.993 | 13.8 | 3.8 | 44.2 |
| Constr | −2.448 | 71.8 | 23.8 | 190.9 |
| Trading | −2.375 | 87.7 | 29.8 | 227.5 |
| Finance | −3.256 | 5.6 | 1.5 | 19.2 |
| RealEst | −2.474 | 66.8 | 22.6 | 174.7 |
| Public | −3.330 | 4.3 | 1.1 | 15.1 |
| Service | −2.517 | 59.2 | 19.8 | 157.0 |
| Private | −2.296 | 108.4 | 40.0 | 267.4 |

Table 3.3: Industry specific means $\nu_g$ and PD intervals measured in basis points ($10^{-4}$). Intervals are obtained by applying the standard normal distribution to $\nu_g \pm \tau$.

We infer from Table 3.3 that on an aggregate level the portfolio of our sample banks might exhibit important differences in average credit quality across industries ranging from 4.3 basis points (bp; one bp corresponds to $10^{-4}$) measured in terms of PDs for public obligors to 108.4bp for private obligors. Furthermore, a standard deviation $\tau$ of 0.357 on the score level yields intervals of different width on the PD scale depending on the industry specific $\nu_g$. E.g., for public obligors we obtain an interval ranging from 1.1bp to 15.1bp whereas it is much wider among private obligors spanning from 40.0bp to 267.4bp.

Analyzing the rating biases, Table 3.4 shows the bank specific bias estimates ($\mu_{g,j}$). A number of conclusions can be drawn from this analysis. First one might want to interpret the results from an aggregate bank specific perspective. Evidently, in columns 1 and 3 most parameter estimates show a negative sign, whereas the opposite holds for column 13 with all estimates being positive. Banks 1 and 3 thus seem to be too optimistic in their credit assessment whereas bank 13 might exhibit an extremely conservative rating behavior.

We also employ so-called relationship plots to visualize the estimated rating

| | **Bank** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Manufac | −0.214 | NA | −0.313 | −0.002 | NA | 0.120 | 0.097 | −0.053 | 0.042 | −0.138 | −0.011 | NA | 0.472 |
| Energy | −0.301 | −0.148 | −0.191 | 0.166 | NA | NA | 0.168 | 0.282 | −0.063 | −0.113 | 0.200 | NA | NA |
| Constr | −0.211 | −0.026 | −0.123 | 0.053 | NA | NA | NA | −0.113 | −0.058 | −0.012 | −0.117 | NA | 0.607 |
| Trading | −0.253 | NA | −0.192 | 0.071 | 0.156 | 0.129 | −0.122 | 0.176 | −0.173 | −0.082 | 0.048 | −0.163 | 0.403 |
| Finance | 0.029 | 0.068 | 0.434 | −0.259 | −0.267 | 0.154 | −0.259 | −0.259 | 0.171 | 0.304 | −0.259 | −0.259 | 0.402 |
| RealEst | −0.249 | −0.145 | −0.337 | −0.004 | 0.234 | 0.008 | 0.080 | 0.094 | −0.008 | −0.251 | 0.013 | 0.090 | 0.474 |
| Public | −0.084 | 0.148 | 0.297 | −0.219 | NA | 0.130 | −0.224 | −0.216 | 0.162 | 0.048 | NA | −0.236 | 0.194 |
| Service | −0.214 | −0.129 | −0.316 | −0.019 | 0.390 | −0.156 | 0.068 | 0.221 | −0.099 | −0.132 | 0.023 | NA | 0.361 |
| Private | −0.197 | −0.085 | −0.617 | 0.240 | 0.032 | −0.029 | 0.285 | NA | −0.195 | 0.169 | 0.193 | 0.202 | NA |

Table 3.4: Rating bias $\mu_{g,j}$ for bank/industry combinations of the 13 Austrian banks. In case of no observations the bias cannot be estimated, hence the NAs.

bias and error variance parameters. These plots visualize the relationship between measurements of a quantitative variable (here: estimated model parameters) and the interaction of two qualitative factors (here: industry/bank combinations). Each combination of factor levels is represented by a rectangular cell shaded by gray values representing the corresponding measurement values.

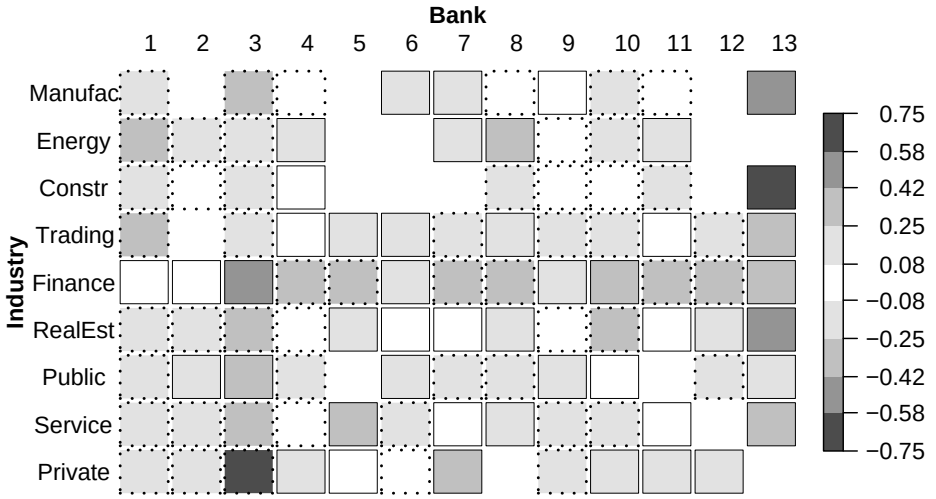



Figure 3.1: Rating bias for bank/industry combinations $\mu_{g,j}$ of the 13 Austrian banks. A dark cell represents a high absolute value, whereas a light cell represents a very low absolute value. If the underlying value is negative, the border of the cell is dotted instead of lined. In case of no observations the bias cannot be estimated, hence the missing cells.

We can also use the industries to better understand the general tendency of identified potential "outlier banks". Bank 1 rather generally overestimates credit quality relative to the other banks (particularly in the Energy, Trading, and Real Estate industries). Bank 3 possibly overestimates the credit quality for private individuals. Conversely, for bank 13 we note that it potentially acts over-cautiously in the Construction industry.

One of the key strengths of our framework is its ability to estimate the standard deviations of the bank specific errors and thus measure the precision

| | Bank | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Manufac | 0.447 | NA | 0.394 | 0.270 | NA | 0.407 | 0.098 | 0.033 | 0.140 | 0.249 | 0.136 | NA | 0.383 |
| Energy | 0.152 | 0.095 | 0.401 | 0.096 | NA | NA | 0.248 | 0.355 | 0.038 | 0.355 | 0.029 | NA | NA |
| Constr | 0.039 | 0.149 | 0.306 | 0.199 | NA | NA | NA | 0.252 | 0.284 | 0.256 | 0.155 | NA | 0.343 |
| Trading | 0.206 | NA | 0.521 | 0.175 | 0.411 | 0.054 | 0.212 | 0.109 | 0.160 | 0.296 | 0.221 | 0.028 | 0.215 |
| Finance | 0.328 | 0.166 | 0.412 | 0.012 | 0.041 | 0.446 | 0.018 | 0.020 | 0.337 | 0.380 | 0.020 | 0.026 | 0.360 |
| RealEst | 0.503 | 0.530 | 0.462 | 0.181 | 0.319 | 0.245 | 0.183 | 0.282 | 0.344 | 0.329 | 0.163 | 0.282 | 0.414 |
| Public | 0.221 | 0.204 | 0.331 | 0.036 | NA | 0.271 | 0.035 | 0.006 | 0.014 | 0.297 | NA | 0.048 | 0.242 |
| Service | 0.349 | 0.346 | 0.397 | 0.238 | 0.051 | 0.247 | 0.169 | 0.522 | 0.210 | 0.336 | 0.233 | NA | 0.149 |
| Private | 0.041 | 0.287 | 0.160 | 0.386 | 0.386 | 0.259 | 0.332 | NA | 0.133 | 0.428 | 0.283 | 0.612 | NA |

Table 3.5: Standard deviations $\sigma_{g,j}$ of the rating errors for bank/industry combinations of the 13 Austrian banks. In case of no observations the standard deviation cannot be estimated, hence the NAs.

of the respective rating systems. Table 3.5 contains the results and Figure 3.2 shows the corresponding relationship plot. The values range from 0.006 for bank 8 for public obligors to a maximum of 0.612 observed for bank 12 for private individuals indicating that the rating tool employed by this bank might be inappropriate for this industry. From a bank-wide perspective Figure 3.2 furthermore suggests that the level of precision is particularly low for bank 3. We observe the highest standard deviation levels for the Real Estate industry, indicating that the banks have particular difficulties in accurately assessing the credit quality and suggesting that informational asymmetries are rather pronounced in this industry.

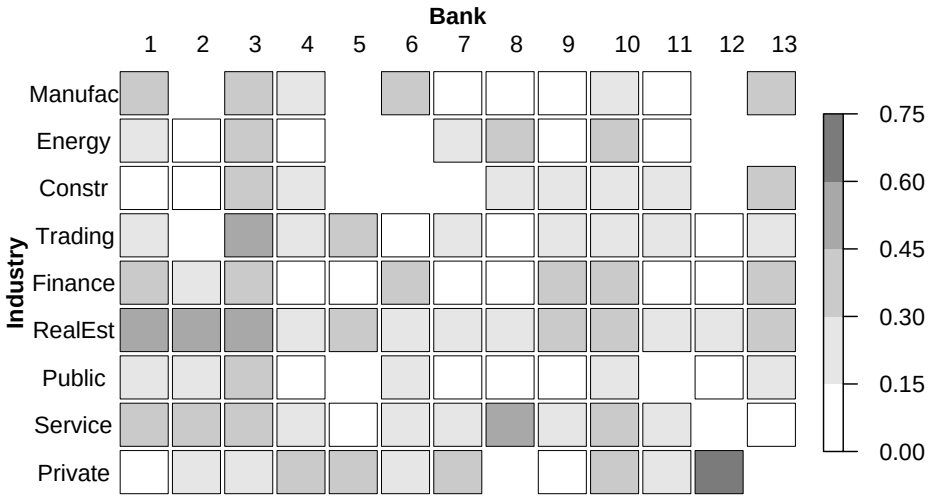



Figure 3.2: Standard deviations $\sigma_{g,j}$ of the rating errors for bank/industry combinations of the 13 Austrian banks. A dark cell represents a high absolute value, whereas a light cell represents a very low absolute value. In case of no observations the deviation can't be estimated, hence the missing cells.

**Consensus Rating and Residual Analysis.** The calibration results of the model presented in the previous section allows us to estimate a consensus rating for each obligor (see Section 3.1.1). The consensus rating itself can be used for many applications where deviations of individual raters or ratings

from an aggregated rating is of interest, e.g., in banking supervision. In this section, the consensus ratings are used to calculate residuals for each rating which allows for a deeper economic analysis. The two variables not employed for grouping in the model specification, i.e., legal form and exposure, are used in this section to illustrate possible further analysis.

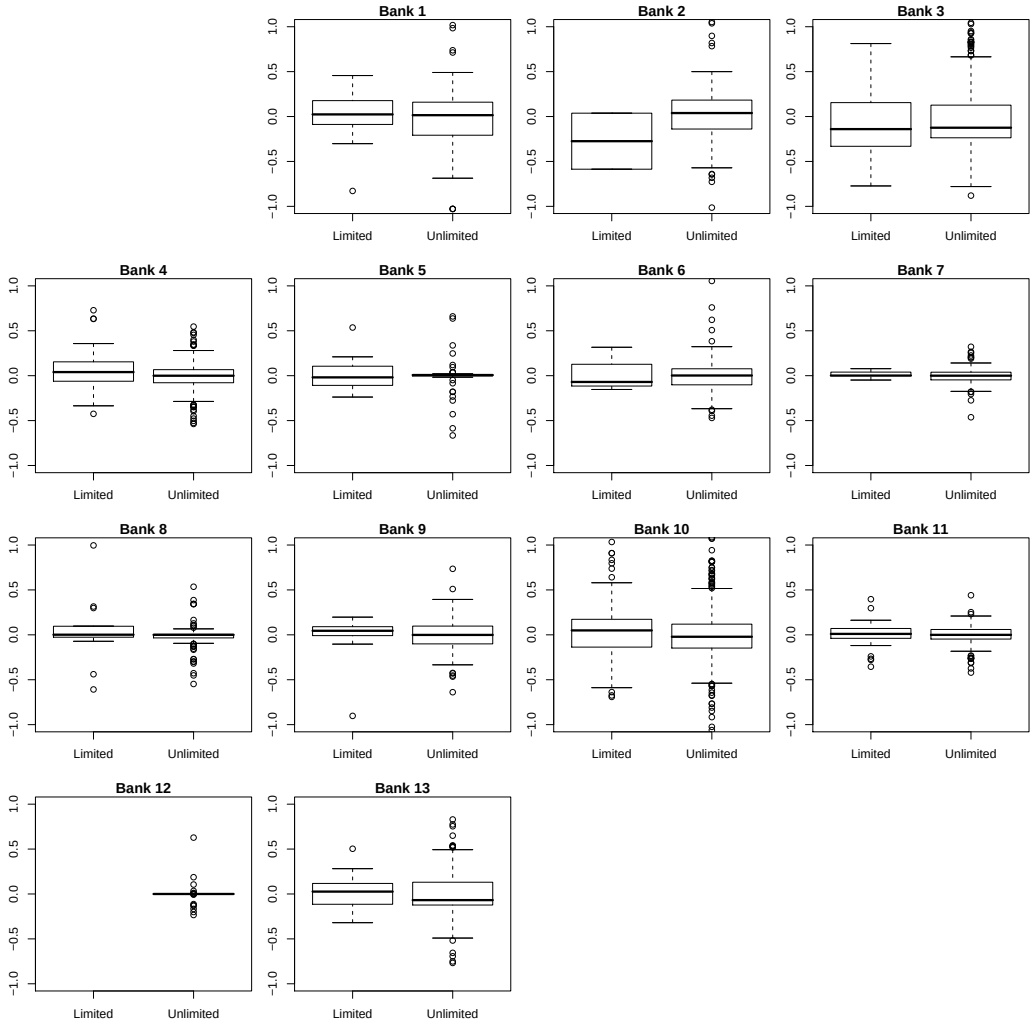



Figure 3.3: Residual analysis for all 13 banks across the legal forms: limited and unlimited companies.

Figure 3.3 presents the residuals for each bank for corporate obligors with limited and unlimited liability. Based on this representation we analyze the

locations and dispersions of the residuals on the bank level. In general, the medians of the residuals are small in absolute terms and we find no structural effect over all banks, i.e., there is no general difference in terms of the median residuals between limited and unlimited liability obligors. On a bank specific level we find residual medians that markedly differ between the two legal forms, e.g., for banks 2 and 13. Bank 2 assigns favorable ratings for limited liability obligors and vice versa for bank 13. We also see no systematic difference in residual dispersion between limited and unlimited corporates, but note that individual banks exhibit rather marked levels of residual dispersion for a specific legal form. Such results could be particularly interesting for supervisors, as it might allow to identify problem areas of individual banks.

As a second illustrative example, we analyze the residuals with respect to the relative exposure size of the obligors. The relative exposure shows the importance of an obligor for the bank. Thus we are particularly interested to detect obligors with large relative exposures and too favorable ratings. Figure 3.4 shows residuals against relative exposures for two selected banks (bank 13 and bank 8). Bank 13 rates two obligors (marked in Figure 3.4) with high relative exposures (more than 2.5% of the total exposure) rather too favorable relative to the market consensus. For bank 8, however, we cannot find comparable outliers. Such an outlier analysis is very important, as it might help supervisors to identify problem loans which have a significant size within a bank's credit portfolio.

### 3.1.3 Discussion

In this section we propose a new probabilistic framework for credit rating model validation in a multi-rater setup, i.e., in situations where PD estimates from different sources for the same obligors are available. In our model the unobservable true PD of all obligors is treated as a latent variable and raters obtain only noisy observations of the latent true PD. In the general framework three ingredients are to be specified: The distribution of the latent PD, the

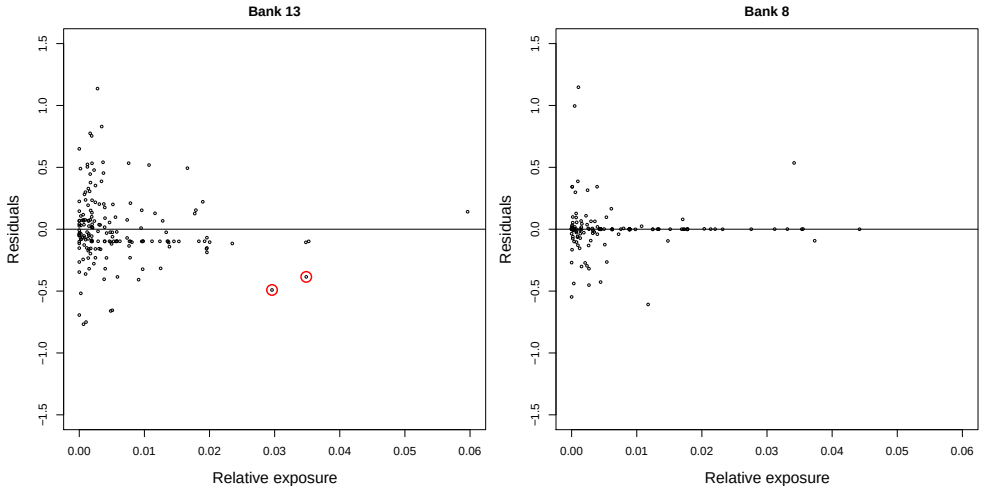



Figure 3.4: Residual analysis for two banks (bank 13 and bank 8) across the relative exposure.

distribution of the error terms, and a suitable link function which transforms the PD to the real axis such that the error terms are additive. Building on the theory of structural credit risk models we propose a specification with normal error terms which uses the probit as a link function. In an empirical example we estimate suitable parametrizations of this model and present results on parameter estimates and possible economic interpretation. Our framework has a variety of potential applications: Banking supervisors might be interested in parameters of the error terms to assess the calibration quality of bank internal rating systems. Developers of rating systems such as banks and rating agencies have a natural interest in benchmarking their rating outcomes to competing models.

Results of benchmarking analyses always need to be assessed relative to the representativity of the available data panel. In our case, it obviously needs to be ensured that raters in the panel follow mutually distinct rating procedures. Data in the panel must be measurements of the *same* underlying entity. In the application framework of this section, banks must supply ratings for the same underlying notion of obligor creditworthiness, which in our case are one-year issuer-specific point-in-time probabilities of default. If different notions

are used, e.g. when trying to incorporate ordinal ratings of creditworthiness such as those provided by the major rating agencies, one could attempt to map these to their one-year PD equivalents. In this case, rating errors will include the corresponding mapping errors.

The suggested framework for modeling multi-rater panels of PD estimates is very general and allows for a variety of possible enhancements. We already indicated the possibility of including additional terms in the parametric specifications of the means and variances of the PD scores and rating errors, or allowing for correlations of rating errors across raters (again, note that a common error "factor" is indistinguishable from the latent PD score). In addition, one could aim at employing more flexible models for the distributions of the PD scores or rating errors, e.g., via suitable mixtures of normals. One could also try to model potential censoring effects mandated by regulatory frameworks. E.g., Paragraph 331 of Bank for International Settlements (2004) states that "the PD for retail exposures is the greater of the one-year PD associated with the internal borrower grade to which the pool of retail exposures is assigned or 0.03%", suggesting to enhance Equation 2.1 along the lines of $S_{ij} = \max(S_i + \epsilon_{ij}, c_i)$ with *known* obligor-specific cutoffs $c_i$. One should note, however, that in many applications co-rating patterns are rather sparse, limiting the flexibility of statistical models which can be inferred from available data. Finally, one could think of extending the cross-sectional setup to a dynamic framework where the PD estimates are also observed at different points in time and hence the latent PDs and the error terms have to be modeled by suitable stochastic processes. Such a framework would allow forecasting future PDs as well as a lead-lag analysis across different raters.

## 3.2 A dynamic latent variable model for ordinal credit ratings

### 3.2.1 Model specification

In this section we develop a model framework to derive a consensus rating for raters providing ordinal rating information, e.g., external agency ratings. Our model is designed for a dynamic framework capturing a time dependent rating process. Despite the fact that the raters publish ordinal ratings, we assume that they estimate a numerical variable—representing the creditworthiness of the firm—in an internal rating process. Each firm is then assigned to a particular rating class if this variable lies within a certain interval (e.g., McNeil and Wendin, 2007; Stefanescu et al., 2009). In general, the specific rating process including both the estimation as well as the scale of the variable (representing the creditworthiness) is unknown. In the literature, modeling the creditworthiness, was first discussed by Altman (1968) who introduces the Z-score. Z-scores are used to predict corporate defaults and are an easy-to-calculate control measure for the financial distress status of companies. The Z-score uses multiple corporate income and balance sheet values to measure the financial health of a company. Furthermore, Merton (1974) assumes that the creditworthiness can be reflected by the distance-to-default (DD) capturing the distance of the firm's asset value to its default threshold on the real line. Alternatively, the creditworthiness variable can also be the result of a ordered probit or logit regression model (e.g., Altman and Rijken, 2004). To obtain ordinal ratings, the estimated DD, the Z-score, or any other numerical variable representing the creditworthiness—which is in the following referred to as "rating score"—is mapped onto an ordinal rating scale by the raters.

Let $\{1, \dots, K_j\}$ be the set of possible non-default rating classes of rater $j$ in descending creditworthiness. That is, 1 denotes the best credit quality and $K_j$ the worst non-default rating class of rater $j$. Further, $S_{ij}(t)$ denotes the estimated rating score (e.g., negative DD, Z-score) and $r_{ij}(t)$ the associated

observed ordinal rating of firm $i$ by rater $j$ at time $t$. The relationship between $r_{ij}(t)$ and $S_{ij}(t)$ is given by

$$r_{ij}(t) = k \Leftrightarrow S_{ij}(t) \in [\lambda_{k-1,j}, \lambda_{k,j}), \tag{3.2}$$

for a monotonically increasing sequence $\lambda_{k,j}$ with $k = 1, \ldots, K_j$. The class boundaries are assumed to be constant over time. The data consists of observations for $J$ raters and $T$ time points. Observing rating $k$ for a firm by rater $j$ means that its rating score lies somewhere in the interval $[\lambda_{k-1,j}, \lambda_{k,j})$.

In general, the thresholds $\lambda_{k,j}$ are not provided by the raters. One possibility to obtain $\lambda_{k,j}$ is to relate the ratings to the observable empirical default rates. In particular, the thresholds can be computed by using the empirical default rates on an appropriate scale[1]. Assuming that the scores of empirical default rates, $S_{ij}(t)$, are defined on the real line we have to fix the lower as well as the upper threshold ($\lambda_{0,j} = -\infty$ and $\lambda_{K_j,j} = +\infty$, respectively). The length of the intervals need not be equal and may differ from rater to rater. Nevertheless, it is expected that firms within the same interval will exhibit roughly the same creditworthiness (Stefanescu et al., 2009).

Due to general informational asymmetry between firm owners and raters[2] which can be due to limited access to the existing information, such as incomplete accounting information (Duffie and Lando, 2001), or delayed observations of the driving risk factors (Guo et al., 2009) the raters cannot estimate the "true" score (reflecting the creditworthiness) of a firm. Assuming that the yielding rating errors can be modeled additively[3] and following Equation 3.2 the relationship between the estimated rating score $S_{ij}(t)$ and

---

[1]Beside this, we assume that raters do not change their rating technology during the desired time period, i.e, they are always measuring creditworthiness on the same scale. This assumption justifies time independent $\lambda_{k,j}$.

[2]The general informational asymmetry between firm owners and raters constitutes the cornerstone of modern corporate finance (e.g., Leland and Pyle, 1977; Berk and DeMarzo, 2007).

[3]This is in line with Duffie and Lando (2001) who build their model on a Merton-type log normal firm value process and assume that the error in the observation of the firm value is normal and additive to the log of the firm value.

the latent score $S_i(t)$ on the score scale is given by

$$S_{ij}(t) = S_i(t) + \epsilon_{ij}(t), \tag{3.3}$$

where $\epsilon_{ij}(t)$ denotes the rating error for firm $i$ by rater $j$ at time $t$. In the following, the latent score $S_i(t)$ is also referred to as the *consensus* score.

On the right hand side of Equation 3.3 we find two terms, which have to be specified: (1) The latent score $S_i(t)$ which describes the consensus creditworthiness and (2) the error term $\epsilon_{ij}(t)$ which captures the accuracy of the rating system of a specific rater. In the following those terms are specified for both the dynamic latent trait model and the benchmark approach.

Despite the fact that the scores $S_{ij}(t)$ are unknown, the latent scores $S_i(t)$ and the bias/variance structure of the rating errors can be estimated in our framework by specifying the distribution of the rating errors and using the interval thresholds $\lambda_{\cdot,j}$ along with the relationship of Equation 3.2. The estimated consensus scores $S_i(t)$ can then be mapped on the rater-specific ordinal scale to derive the consensus ratings $r^*_{ij}(t)$ which obviously depend on the used rating system (of rater $j$). Since $r_{ij}(t)$ and $r^*_{ij}(t)$ for all $i$ and $j$ are on the same rating scale one can easily compare these ratings and derive inference about the quality of the ratings $r_{ij}(t)$.

### Dynamic latent trait model

**Latent consensus score.** In order to specify the latent scores $S_i(t)$, we follow the lines of McNeil and Wendin (2007); Stefanescu et al. (2009) and assume that the scores are driven by market- (systematic risk) as well as firm-specific effects (idiosyncratic risk). We define a time-dependent process $m_i(t)$ capturing the idiosyncratic changes and a latent market factor $f(t)$ capturing the systematic development of the latent scores $S_i(t)$. The idiosyncratic changes $m_i(t)$ capture the firm-specific risk and can be modeled as an adequate time series process to cope with repeated observations. The latent market $f(t)$, capturing the development of the market, implies a cor-

relation structure between the different firms and can also be modeled by an adequate time-dependent process, e.g., a stationary auto-regressive process or a random walk. Let $\nu_i$ be the firm specific long-term mean of firm $i$ which can be interpreted as the historical average creditworthiness of the firm. The development of the latent scores $S_i(t)$ on the score scale is given by

$$S_i(t) = \nu_i + m_i(t) + \alpha f(t), \tag{3.4}$$

where the factor loading $\alpha$ captures the dependence of $S_i(t)$ on $f(t)$.

In order to estimate the consensus scores $S_i(t)$ we have to specify the underlying processes and distributions of this framework. We specify the time-dependent processes, describing the development of $S_i(t)$ (Equation 3.4), the firm-specific changes $m_i(t)$ and the latent market factor $f(t)$ as AR(1) processes

$$\begin{aligned} m_i(t) &= \beta_i m_i(t-1) + \omega_i(t), & (3.5)\\ f(t) &= \gamma f(t-1) + \xi(t). & (3.6) \end{aligned}$$

$m_i(t)$ and $f(t)$ are assumed to start with zero at $t = 0$. $\omega_i(t)$ is a normal distributed error term with mean zero and a constant variance across time and firms, and $\xi(t)$ is a standard normal distributed error term. $\beta_i$ ($|\beta_i| < 1$) and $\gamma$ ($|\gamma| < 1$) reflect the dependence on period $t-1$ (inter-temporal correlation).

**Rating error.** In order to specify the rating errors $\epsilon_{ij}(t)$, we assume that they are independent of the firms and their characteristics (in particular, their creditworthiness itself) and the general rating process does not change over time $t$ (see Section 3.1). Assuming that $\mu_j$ and $\sigma_j$ denote the mean and standard deviation of the rating errors $\epsilon_{ij}(t)$, respectively, the rating errors $\epsilon_{ij}(t)$ are given by

$$\epsilon_{ij}(t) = \mu_j + \sigma_j Z_{ij}(t) \tag{3.7}$$

where $Z_{ij}(t)$ is assumed to be independent standard normal distributed over $i$, $j$ and $t$.

### Benchmark Model

In addition to the dynamic latent trait model, we define an intuitive benchmark approach and compare it with our dynamic latent trait model. Being conservative, one could consider to take the companies' worst rating as the benchmark. This is inappropriate for two reasons. Firstly, such an approach disregards the information contained in the other available rating sources. Secondly, from an economic point of view a rated company must be convinced that its creditquality lies somewhere *between* its ratings and is not represented by the worst rating. Otherwise there would be little reason to obtain several ratings (Hsueh and Kidwell, 1988). Hence, without any rater specific characteristics, the "mean" of the observed ratings could serve as a consensus benchmark.

**Latent consensus score.** Our benchmark model follows the idea that for any time $t$, the consensus score $S_i(t)$ of a company is simply the *mean* over rating scores $S_{ij}(t)$. In doing so, we do not assume any time-dependent process driving the development of $S_i(t)$, i.e., for any time $t$, $S_i(t)$ is independent of $S_i(t-1)$.

**Rating errors.** For the rating errors, we assume that there are no rater specific error terms $\mu_j$ and $\sigma_j$, but a constant standard deviation $\sigma$ of the rating errors between the raters. This implies that all raters are weighted equally in the estimation process. Within our model framework the relationship between consensus score $S_i(t)$ and the estimated scores $S_{ij}(t)$ for the benchmark model is given by

$$S_{ij}(t) = S_i(t) + \sigma Z_{ij}(t), \tag{3.8}$$

with $Z_{ij}(t)$ distributed as in the dynamic case.

One drawback of this model specification is that these assumptions for the rating errors and the latent scores may lead to distorted results for rating data including missings, i.e., some companies are not rated by all agencies (see Figure 3.6).

### 3.2.2 Application

We apply our dynamic latent variable model for ordinal rating information to the iTraxx Europe companies. For this portfolio we obtain rating information from the big three rating agencies Standard&Poor's, Moody's and Fitch. The big three external rating agencies use different rating systems with different granularity as well as different labels (typically, a combination of letters, numbers and/or modifiers). Nowadays, almost all large corporates are nowadays rated by at least one of the big three rating agencies, Standard&Poor's, Moody's and Fitch (e.g. Kliger and Sarig, 2000). Obviously, they do not always agree on the creditworthiness of the firms, and therefore differ in their opinion about the default probabilities of these corporates (Jewell and Linvingston, 2002; Cantor and Packer, 1995).

#### Data Description

**Ordinal ratings of the iTraxx Europe companies.** We use historical long-term issuer ratings of the constituents of the iTraxx Europe index (Series 10) from February 2007 to January 2009 provided by the big three external rating agencies Standard&Poor's, Fitch and Moody's. The iTraxx Europe index series consists of the 125 most-liquid CDS referencing European investment-grade entities and a new series is determined by dealer liquidity poll every six months. Most of the 125 names in the indexes are large multinationals and have traded equity. We choose the iTraxx Europe index, because it forms a representative contingent of the overall European credit derivative market and its constituents have a high number of co-ratings (occurrences of ratings of a single firm by two different raters) from the big three rating agencies. The time series was constructed using historical ordinal rating an-

nouncements taken from Reuters Credit Views. We exclude all companies for which we do not have rating information of at least two agencies for the complete time period, i.e., those with withdrawn ratings and entities which acquire a rating for the first time within the selected time frame. This process yields a sample of 5616 monthly ratings for 95 companies over 24 months (February 2007 to January 2009). Table 3.6 shows the co-ratings structure of the three raters. The average number of ratings for each month is 2.46.

| | Fitch | Moody's | S&P |
|---|---|---|---|
| Fitch | 88 | 44 | 88 |
| Moody's | 44 | 51 | 51 |
| S&P | 88 | 51 | 95 |

Table 3.6: Co-ratings structure for 95 out of the 125 iTraxx Europe (Series 10) companies of the big three external rating agencies Fitch, Moody's and Standard&Poor's (S&P).

As described above, the three rating agencies use different rating systems. Moody's rating system for global corporates contains 20 non-default rating categories, ranging from $Aaa$ to $C$ and is so in the near default ratings more granular than the rating systems of Fitch and Standard&Poor's (Emery and Ou, 2009). These two agencies assign 17 non-default rating categories ($AAA$ to $CCC/C$) to global corporates (Needham and Verde, 2009; Vazza et al., 2009). Table 3.7 shows the number of ratings (per rating category and rater) of the monthly ratings from February 2007 to January 2009 for the rating agencies Fitch, Moody's and Standard&Poor's.

According to the three rating distributions of this rating data, only one firm is rated as a non-investment firm (ContinentalAG) and this only by Standard&Poor's (see Crouhy et al., 2001, for a description of investment grades and speculative grades). The distributions show also that the granularity of the three rating systems is equal in the relevant segment of this rating data.

The rating history of 57 firms (60%) changed over the considered time period. Fitch changed the ratings of 35 firms, where 29 firms were downgraded and 4 firms were upgraded. The remaining two companies experienced a down-

| | Fitch | | Moody's | | S&P | |
|---|---|---|---|---|---|---|
| | label | no. | label | no. | label | no. |
| 1 | AAA | 6 | Aaa | 18 | AAA | 0 |
| 2 | AA+ | 85 | Aa1 | 176 | AA+ | 45 |
| 3 | AA | 148 | Aa2 | 41 | AA | 167 |
| 4 | AA− | 193 | Aa3 | 54 | AA− | 233 |
| 5 | A+ | 226 | A1 | 79 | A+ | 170 |
| 6 | A | 243 | A2 | 153 | A | 251 |
| 7 | A− | 410 | A3 | 225 | A− | 473 |
| 8 | BBB+ | 454 | Baa1 | 231 | BBB+ | 576 |
| 9 | BBB | 315 | Baa2 | 183 | BBB | 292 |
| 10 | BBB− | 30 | Baa3 | 64 | BBB− | 72 |
| 11 | BB+ | 2 | Ba1 | 0 | BB+ | 0 |
| 12 | BB | 0 | Ba2 | 0 | BB | 1 |
| 13 | worse | 0 | worse | 0 | worse | 0 |

Table 3.7: Number of ratings (per rating category and rater) of the 95 out of the 125 iTraxx Europe companies.

grade as well as an upgrade. Moody's changed the ratings of 17 firms, where 8 firms were downgraded and 8 firms were upgraded (the remaining company experienced two upgrades as well as two downgrades). Standard&Poor's changed the ratings of 45 firms, where 29 firms were downgraded and 12 firms were upgraded (the remaining four company experienced upgrade(s) as well as downgrade(s)). Hence, a clear tendency of downgrading is observable in this period.

In order to model the consensus ratings (Equation 3.3), each ordinal rating is identified with a numerical interval reflecting the upper and lower bound of the creditworthiness on the real line (see Equation 3.2). Here, we estimate the thresholds for the ordinal ratings using the empirical default rates (1990–2006) provided by the external raters (Needham and Verde, 2009; Emery and Ou, 2009; Vazza et al., 2009) by following the approach proposed by Neagu et al. (2009). They relate empirical PDs to ratings on an appropriate score scale. The score variable represents a rank ordering of risk of default over some future time horizon (we use a one year future time period). The task is

to find a transformation of the score variable into an empirical PD. In other words, this method aims at finding a function $F$ such that:

$$\text{PD} = F(\text{score}),$$

which can be written by using a default indicator as:

$$\text{Prob}(\text{default indicator} = 1) = F(\text{score})$$

and gives the base formulation for the binary response class of models. Different types of models, utilizing different forms for the function $F$, can be fit. Neagu et al. (2009) suggest to try the three most commonly used binary response models: logit, probit, and complementary log-log (CLL) models. These models can be applied directly to the score data, but in real-world applications the score data tends to exhibit a high degree of skewness. In this case it is recommended that a transformation of the score variable is made: a Box-Cox power transformation (Fox, 1997) or a Box-Tidwell transformation (Granger and Newbold, 1977).

In particular, we use the published historical empirical global corporate default rates of the three external rating agencies from 1990 to 2006 (Needham and Verde, 2009; Emery and Ou, 2009; Vazza et al., 2009). In order to yield one-year empirical default rates we compute the averages over the time period. We then fit all combinations of binary response class models (probit, logit, and CLL) and transformations (Box-Cox power and Box-Tidwell) to the average default rates. A probit score model with Box-Tidwell transformation is selected as the best method according to the Hosmer-Lemeshow statistic (Hosmer and Lemeshow, 2000). Figure 3.5 shows the estimated "mapping" lines using a probit score model with Box-Tidwell transformation for the three different rating systems of Fitch, Moody's, and Standard & Poor's using the empirical default rates from 1990 to 2006. Note, that the rating system of Moody's is finer on the upper side, i.e., assigning four more rating grades to the high PD segment than the other two raters.

Whereas the empirical default rates and the PD mapping of Fitch and Stan-

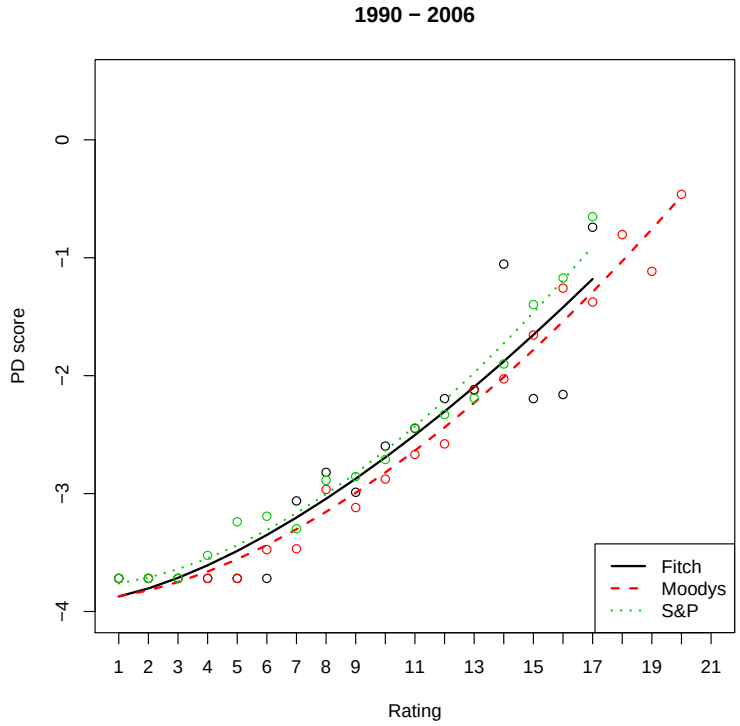



Figure 3.5: Mapping of the empirical default rates stemming from the three raters on the score scale based on a probit score model with Box-Tidwell transformation using the empirical default rates from 1990 to 2006.

dard&Poor's seem to be rather similar, Moody's empirical default rates and mapping line is clearly below the other two. E.g., in average the difference on the probit scale between the investment grades of Standard&Poor's and Moody's is 0.139.

In order to cleave to the ordinal structure of ratings, thresholds for the mapping PDs derived from the empirical default rates have to be computed. We compute the thresholds by the means of two adjacent mapping PDs on the logit scale for each rater $j$. I.e., the upper threshold $\lambda_k$ of rating class $k = 1, \ldots, K_j - 1$ of rater $j$ is given by $\lambda_k = 1/2(\text{logit}(\text{PD}_{k+1}) + \text{logit}(\text{PD}_k))$ and the "lower" threshold of the best rating class is $-\infty$ and the "upper" threshold of the worst rating class is $+\infty$ (Altman and Rijken, 2004).

**Dow Jones EURO STOXX 50.** For comparison reason we use the Dow Jones EURO STOXX 50 as a representative market development of the iTraxx Europe portfolio from February 2007 to January 2009 (see Figure 3.7). The Dow Jones EURO STOXX 50 is the leading stock (price) index for the Eurozone and covers 50 stocks from 12 Eurozone countries: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, the Netherlands, Portugal, and Spain. At January 2009, stocks of 30 out of the 95 companies are contained in the EURO STOXX 50.

### Analysis of the big three rating agencies using their ratings for the iTraxx Europe companies

**Model estimation.** Using the available ordinal ratings $r_{ij}(t)$ for each company $i = 1, \ldots, 95$ (out of the 125 iTraxx Europe companies) and external rating agency $j = \{F, M, SP\}$ from $t = 1, \ldots, 24$ (February 2007 to January 2009) and the associated thresholds $\lambda_{j,k}$ for $k = 1, \ldots, K_j$ with $K_F = 17$, $K_M = 20$, and $K_{SP} = 17$ we estimate the model parameters of our dynamic latent trait model as well as the parameters of our benchmark model. For the estimation frequentist as well as Bayesian techniques can be used. E.g., the static model in Section 3.1 is estimated by standard maximum likelihood estimation. Here, we follow McNeil and Wendin (2007) and Stefanescu et al. (2009) and choose a Bayesian estimation approach using Markov chain Monte Carlo methods (MCMC) and Gibbs sampling (Carlin and Louis, 2009). Such an approach requires prior distributions to be chosen for the parameter set. In order to minimize the influence of the prior distributions on the posterior distribution we have specified non-informative priors for all our parameters.

In particular, we run four parallel Markov chains, each initialized with a different seed and a different random number generator. The Gibbs sampler ran for $50,000$ iterations, using a thinning of 10 whereby the first $5,000$ were discarded as burn-in period. This yields $4,500$ draws from the posterior for each parameter for each chain. Trace plots as well as the Geweke diagnostic and the Gelman Rubin's convergence diagnostic indicated satisfactory

convergence of all chains (e.g., Gelman and Rubin, 1992; Plummer et al., 2008).

**Model selection.** In order to compare our dynamic latent trait model with the benchmark model we use the *deviance information criterion* (DIC; according to Spiegelhalter et al., 2002). The DIC is a generalization of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) for hierarchical models. In contrast to the AIC and BIC, DIC allows to compare Bayesian hierarchical models where the effective number of parameters is not clearly defined. Similar to the other information criteria a trade-off between model fit and model complexity is evaluated. The DIC contains one penalty term for the effective number of parameters used measuring model complexity and one term equal to the deviance of the likelihood measuring model fit. A lower DIC value indicates a better model fit. According to Spiegelhalter et al. (2002), if the difference in DIC is greater than 10, then the model with the larger DIC value has considerably less support than the model with the lower DIC value.

For our models, the lower DIC value of our dynamic latent trait model (DIC = 9485.77) indicates that this model dominates in the terms of model fit as well as model complexity the obvious benchmark model (DIC = 12319.82).

### Results for the dynamic latent trait model

**Rating errors.** We begin our analysis of the estimation results with the rating errors. Our dynamic latent trait model captures estimates for the rating bias $\mu_j$ and the standard deviation $\sigma_j$ of the rating error of the big three external rating agencies on the score scale. Table 3.8 shows the results for the estimated posterior distribution of the parameters for the three raters $\mu_j$ and $\sigma_j$, respectively. The posterior distributions of the parameters are characterized by the mean values (mean) and the standard deviations (SD) of the $18,000$ $(4 \times 4,500)$ posterior draws.

We infer from Table 3.8 that Fitch has the smallest absolute rating bias from

| | $\mu_j$ | | $\sigma_j$ | |
|---|---|---|---|---|
| | mean | SD | mean | SD |
| Fitch | 0.0155 | 0.0018 | 0.0752 | 0.0021 |
| Moody's | −0.0887 | 0.0024 | 0.1013 | 0.0029 |
| S&P | 0.0732 | 0.0017 | 0.0641 | 0.0017 |

Table 3.8: Estimated rating bias $\mu_j$ and standard deviations $\sigma_j$ for the rating errors (on the score scale) of the big three external rating agencies Fitch, Moody's and Standard&Poor's. The posterior distributions of the parameters are characterized by the mean values (mean) and the standard deviations (SD) of the $18,000$ $(4 \times 4,500)$ posterior draws.

the consensus on the score scale with respect to the posterior mean (0.0155). Moody's clearly seems to be too optimistic in its credit assessment yielding a posterior mean for the rating bias $\mu$ of $-0.089$ on the score scale. Note, that our model is based on the thresholds $\lambda_{j,k}$ (and therefore PD equivalents) which are clearly lower for Moody's than the other two raters. Despite the high difference (on the score scale: 0.139) in the PD equivalents of Moody's and Standard&Poor's indicated in the Appendix (see Figure 3.5), Moody's is still more optimistic by rating investment-grade firms than Standard&Poor's. In this study, Standar&Poor's is with a posterior mean of the rating bias of 0.073 the most conservative rater out of the three considered rating agencies.

In addition to the rating biases, our model captures the standard deviation (precision) of the rating errors of the three raters (Table 3.8). Whereas the posterior mean of the standard deviation $\sigma$ of the rating errors is rather similar for Fitch and Standard&Poor's (0.075, 0.064), Moody's has a higher posterior mean of the standard deviation (0.101), indicating that its ratings deviate more strongly from the consensus ratings.

**Consensus score.** In addition to the analysis of the bias/variance structure of the rating errors, we analyze the estimated consensus scores of our dynamic latent trait model. Instead of showing the consensus scores of all iTraxx Europe companies, Figure 3.6 shows the estimated consensus rating scores of four sample companies (ENELSPA, NESTLE, GLENCORE INT.

AG, ROYAL BANK OF SCOTLAND) and compares them with the original ratings (mapped onto the score scale) of the three raters Fitch, Moody's and Standard&Poor's as well as with the mean rating score of the three raters.

Figure 3.6: Estimated consensus score, the mean score, and the original ratings mapped onto the score scale of the big three external rating agencies Fitch (F), Moody's (M) and Standard&Poor's (S).

Due to the fact that the companies ENELSPA and NESTLE are rated by all three raters, the consensus score (solid line) is very similar to the mean score (dashed line). In the case of the two other companies GLENCORE INT. AG and ROYAL BANK OF SCOTLAND where for each company ratings of only two raters are available, Figure 3.6 shows remarkable differences between the consensus and the mean score. Due to rater specific error terms, our latent consensus score is able to incorporate such a missingness structure.

Furthermore, we can confirm the need of a latent market factor in our dynamic latent trait model by showing the strong relationship between our latent market $f(t)$ and a reference market, the Dow Jones EURO STOXX 50 index (correlation: $-0.946$)[4].

**Consensus rating.** In addition to the analysis of the consensus scores, we can use the consensus ratings derived by mapping the scores onto the rater's rating scales to analyze the rating agreement of the raters.

An intuitive way for this is the Hit-Miss-Match (HMM) Matrix which counts how many consensus ratings exactly match the ratings provided by a rater. Table 3.9, 3.10 and 3.11 show the HMM matrix for each rater.

In Table 3.9 most ratings are on the main diagonal or one rating notch below or above indicating a high agreement between Fitch's ratings and the consensus ratings. Table 3.10 shows that Moody's ratings are rather one or more rating notches below the consensus ratings, confirming the negative rating

[4]Note, that the negative correlation is due to the fact that an increase in $f(t)$ on the score scale indicates a decrease in the creditworthiness.

| Consensus rating | Fitch rating | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AAA | AA+ | AA | AA− | A+ | A | A− | BBB+ | BBB | BBB− | BB+ |
| AAA | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA+ | 0 | 33 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA | 6 | 52 | 124 | 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA− | 0 | 0 | 21 | 157 | 44 | 14 | 0 | 0 | 0 | 0 | 0 |
| A+ | 0 | 0 | 3 | 19 | 149 | 50 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 0 | 0 | 33 | 166 | 33 | 0 | 0 | 0 | 0 |
| A− | 0 | 0 | 0 | 0 | 0 | 13 | 309 | 82 | 3 | 0 | 0 |
| BBB+ | 0 | 0 | 0 | 0 | 0 | 0 | 68 | 350 | 93 | 0 | 0 |
| BBB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | 218 | 4 | 0 |
| BBB− | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 26 | 2 |
| BB+ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3.9: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Fitch, measured on the Fitch rating scale.

| Consensus rating | Moody's rating | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Aaa | Aa1 | Aa2 | Aa3 | A1 | A2 | A3 | Baa1 | Baa2 | Baa3 | Ba1 |
| Aaa | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aa1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aa2 | 10 | 80 | 7 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Aa3 | 7 | 96 | 31 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| A1 | 0 | 0 | 3 | 16 | 33 | 24 | 0 | 0 | 0 | 0 | 0 |
| A2 | 0 | 0 | 0 | 33 | 19 | 3 | 0 | 0 | 0 | 0 | 0 |
| A3 | 0 | 0 | 0 | 0 | 3 | 126 | 73 | 0 | 0 | 0 | 0 |
| Baa1 | 0 | 0 | 0 | 0 | 24 | 0 | 150 | 74 | 4 | 21 | 0 |
| Baa2 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 157 | 101 | 3 | 0 |
| Baa3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 76 | 35 | 0 |
| Ba1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 5 | 0 |

Table 3.10: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Moody's, measured on the Moody's rating scale.

| Consensus rating | Standard&Poor's rating | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AAA | AA+ | AA | AA− | A+ | A | A− | BBB+ | BBB | BBB− | BB+ | BB |
| AAA | 0 | 43 | 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA+ | 0 | 2 | 93 | 7 | 0 | 6 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA | 0 | 0 | 28 | 136 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| AA− | 0 | 0 | 0 | 90 | 109 | 9 | 0 | 0 | 0 | 0 | 0 | 0 |
| A+ | 0 | 0 | 0 | 0 | 58 | 122 | 2 | 0 | 0 | 0 | 0 | 0 |
| A | 0 | 0 | 0 | 0 | 0 | 98 | 159 | 0 | 0 | 0 | 0 | 0 |
| A− | 0 | 0 | 0 | 0 | 0 | 16 | 311 | 185 | 0 | 0 | 0 | 0 |
| BBB+ | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 391 | 91 | 0 | 0 | 0 |
| BBB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 201 | 39 | 0 | 0 |
| BBB− | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 33 | 0 | 1 |
| BB+ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| BB | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Table 3.11: Hit-Miss-Match Matrix between the estimated consensus ratings and the ratings provided by Standard&Poor's, measured on the Standard&Poor's rating scale.

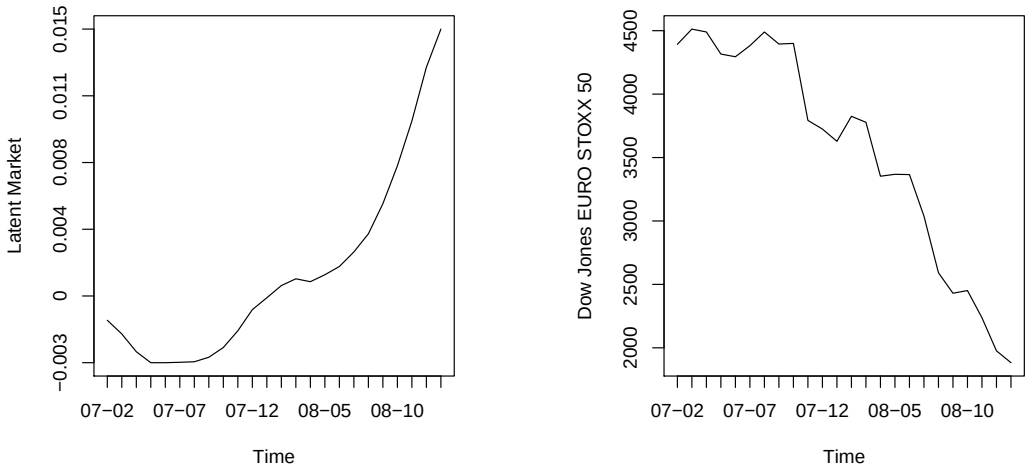



Figure 3.7: Estimated latent market factor $f(t)$ and the Dow Jones EURO STOXX 50 index over the full time period (2007-02 to 2009-01).

bias shown in Table 3.8. In contrast to Moody's ratings, Standard&Poor's ratings are rather one or more rating notches above the consensus ratings, confirming the positive rating bias shown in Table 3.8.

Furthermore, we can compute the proportion of ratings for each rating deviation (measured in rating notches) between the consensus ratings and the ratings provided by the raters (shown in Table 3.12).

| | $-4$ | $-3$ | $-2$ | $-1$ | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|---|
| Fitch | 0.000 | 0.000 | 0.008 | 0.154 | 0.725 | 0.108 | 0.004 | 0.000 |
| Moody's | 0.000 | 0.000 | 0.017 | 0.027 | 0.273 | 0.547 | 0.115 | 0.020 |
| S&P | 0.003 | 0.030 | 0.429 | 0.532 | 0.007 | 0.000 | 0.000 | 0.003 |

Table 3.12: Proportion of ratings per rating class deviation between the consensus ratings and the origin ratings provided by the big three rating agencies Fitch, Moody's and Standard&Poor's.

Table 3.12 shows that Fitch's ratings have a very high accordance (72.5%) with the estimated consensus ratings. According to the estimated rating

biases (see Table 3.8) Moody's is rather too "optimistic" than the other raters. These effect is also seen in Table 3.12. Only 27.3% of Moody's ratings exactly hit the consensus rating. 84.7% are within one rating notch and 68.2% are more optimistic, i.e., are at least one rating category better than our estimated consensus rating. For Standard&Poor's we obtain that 53.9% are within one rating category in comparison to the consensus rating. In contrast to Fitch, Standard&Poor's have few ratings which are even 4 rating classes below the estimated consensus rating.

### 3.2.3 Discussion

In this chapter we investigate a new dynamic framework for aggregating credit rating information in a multi-rater set-up, i.e., in situations where ordinal ratings from different sources for the same firm are available. In our model we assume that the raters do not directly estimate the ordinal ratings, but they estimate a numerical variable—representing the creditworthiness of the firm—in an internal rating process. We treat the true unobservable numerical variable of a firm as a latent variable and model its dynamic by using systematic as well as idiosyncratic changes. In contrast to other methods, our model class allows missingness in the data and captures the panel structure of the data.

In addition to the solution for the aggregation problem, our model is useful in the validation of the different sources. The analysis of the mean/variance structure of the rating errors yields to rater-specific rating biases as well as the precision of the different rating systems.

The suggested framework for modeling consensus of a multi-rater panel is very general and allows for a variety of possible enhancements. We could aim at employing more flexible models for the distributions of the rating scores and rating errors, e.g., via suitable mixtures of normals. We could also allow more flexibility in the specification of the factor loading $\alpha$ capturing the dependence between the latent scores and the latent market (see Equation 3.4) by using a firm- or industry-specific factor loading. In ad-

dition, it would be interesting to allow for industry-specific parameters for the rating bias, the standard deviation of the rating error and the long-term mean (see Section 3.1). We could also try to use an external market factor (e.g., the Dow Jones EURO STOXX 50) instead of a latent market factor to describe the systematic changes of the latent scores. The use of Bayesian estimation techniques allows very flexible specification of models, so that we intend to explore these possible enhancements in our future research.

By using the ratings for the iTraxx Europe companies (Series 10) provided by the big three rating agencies Fitch, Moody's and Standard&Poor's we compute a more informative rating, the consensus rating for each company and show that there are remarkable differences in the rating behavior and rating systems of the three raters. In particular, we infer from our results, that Moody's is the most favorable and Standard&Poor's the most pessimistic rater.

# Chapter 4

# Modeling bookmakers odds

In the course of growing popularity of online sports betting, the analysis and forecasting of competitive sports has been receiving increasing interest. Forecasts of sports events are often based on one of two types of information: ratings or rankings of the competitors' ability/strength, and bookmakers odds for winning a competition of two or more contestants. Here, we show how both types of forecasts—winning probabilities and underlying abilities—can be derived from both sources of information—ability ratings and bookmakers odds.

Sports ratings or rankings are typically derived by suitably aggregating the competitors' previous performances and are often found to provide predictive power in forecasting tasks. Boulier and Stekler (1999) show that rankings provide forcasting information for basketball tournaments and tennis matches. Lebovic and Sigelman (2001) analyze the predictive accuracy of college football rankings. Suzuki and Ohmori (2008) use the FIFA/Coca Cola World rating (Fédération Internationale de Football Association, 2008), one of the most popular rating system in soccer, as a forecasting tool for the last four FIFA World Cups (1994, 1998, 2002, 2006). In addition, Dyte and Clarke (2000) use the FIFA ratings to predict the distribution of scores in international soccer matches. Another popular rating system is the Elo rating system, originally developed to calculate the relative skills of chess players

(e.g., Elo, 2008), which has subsequently also been applied to various other sports including soccer. Song et al. (2009) apply it as one method to forecast the winner of single American Football games. Edmans et al. (2007) select important soccer games based on the World Football Elo Ratings.

Bookmakers odds represent a rather different type of rating compared to the methods above. Based on the bookmakers' expert judgments (which typically include, but are not limited to, knowledge about past performances) the odds reflect expected outcomes in a particular competition where the bookmakers have strong economic incentives to rate the competitors correctly. A bias (in either direction, too good or too bad) will cost them money, or, in other words, will reduce their profits. Hence, bookmakers can be seen as experts in the matter of sports rating (see Pope and Peel, 1989) and are likely to provide good predictions (Forrest and Simmons, 2000). This is confirmed by various empirical studies in which fixed odds are found to be an efficient forecasting instrument for the outcome of single matches (e.g., Vlastakis et al., 2009; Spann and Skiera, 2009; Song et al., 2007; Forrest et al., 2005b; Dixon and Pope, 2004; Boulier and Stekler, 2003).

One advantage of employing bookmakers odds is that winning probabilities for the corresponding competition can be derived easily while this is not straightforward for many of the ability ratings. However, if abilities are measured on a ratio scale (or can be transformed to such), winning probabilities for pairwise matches can be derived using the approach of the Bradley and Terry (1952) model. Notable in this respect is the Elo rating from which pairwise winning probabilities for single matches can be obtained (e.g., Stefani and Pollard, 2007; Edmans et al., 2007). Thus, when the competition of interest is a single match, forecasts based on ability ratings and bookmakers odds can be compared easily. The same is not true if the competition is a more complex tournament for which the bookmakers odds, by their prospective nature, can include additional effects such as group draws or seedings. To link forecasts of abilities (associated with pairwise winning probabilities) and winning probabilities for sports tournaments, we suggest a simulation approach that allows to (approximately) map abilities to winning probabilities

and vice versa.

## 4.1 General model specification for bookmakers odds

As an alternative application, we use the above introduced general model framework to model the rating process of bookmakers. By publishing odds, bookmakers rate the players' or teams' chances of winning a competition or tournament. Hence, bookmakers odds can be seen as prospective ratings of the performance of the participating players or teams in a sports competition.

The raw quoted bookmakers odds are no "honest" odds, but are the payout amounts for successful bets which has two important implications: (1) They still contain the stake, i.e., the payment for placing the bet (the "1" in Equation 4.1 below). (2) More importantly, the bookmakers odds contain a profit margin, the so-called "overround", which means that the "true" underlying odds are actually larger (see e.g., Henery, 1999; Forrest et al., 2005b). Assuming that the overround $\delta$ is constant across all possible outcomes (e.g., the same for all competitors winning a tournament), it can be computed by restricting the corresponding probabilities to sum to unity. More precisely, the raw quoted odds $rawodds_{i,b}$ for event $i$ by bookmaker $b$ can be adjusted to $odds_{i,b}$ and then transformed to probabilities $p_{i,b}$ via:

$$odds_{i,b} = (rawodds_{i,b} - 1)\,\delta_b, \tag{4.1}$$

$$p_{i,b} = 1 - \frac{odds_{i,b}}{1 + odds_{i,b}}. \tag{4.2}$$

Then, $\delta_b$ for bookmaker $b$ can be chosen such that $\sum_i p_{i,b} = 1$. (Note, that the complementary probabilities have to be used as the bookmakers odds represent expectations for an outcome not to occur.) In the case of winning odds for a tournament, this means that the implied winning probabilities can be easily derived from the quoted odds for all competitors. When appropriately adjusted and transformed, the bookmakers odds yield expected

winning probabilities $p_{i,b}$ for each team/player $i = 1, \ldots, I$ and bookmaker $b = 1, \ldots, B$.

We can then adopt our general model framework (Equation 2.1) and relate the expected winning probabilities $p_{i,b}$ to the latent winning probability for team or player $i$ on a logit scale

$$\text{logit}(p_{i,b}) \quad = \quad \text{logit}(p_i) + \epsilon_{i,b}, \tag{4.3}$$

where $\epsilon_{i,b}$ is the deviation of bookmaker $b$ for team or player $i$. The (unobservable) "true" winning logits $\text{logit}(p_i)$ for team or player $i$ reflect the bookmakers consensus and the additional (unobservable) "error" term $\epsilon_{i,b}$ of bookmaker $b$ for team or player $i$ reflects the disagreement across the bookmakers. According to the specific competition or tournament this model can be refined by team/player- or bookmaker-specific effects.

**Applications.** In the following we use the general model specification for bookmakers odds in order to forecast the outcome and analyze the bookmakers agreement of three different and very popular sport tournaments, the UEFA EURO 2008 (Section 4.2), Wimbledon 2009 (Section 4.3), and the UEFA Champions League 2008/09 (Section 4.4) .

## 4.2 UEFA EURO 2008

We first apply our general model framework for bookmakers odds stemming from a variety of bookmakers (Equation 4.3) to the UEFA EURO 2008, one of the world's biggest sports events that took place in June 2008 in Austria and Switzerland.

After deriving the bookmaker consensus from the general model, the consensus information is compared to the forecasts from the World Football Elo rating (also considered in a note from UBS Wealth Management Research Switzerland, 2008, for prediction of the EURO 2008) and the ranking im-

plied by the FIFA/Coca Cola World rating (also employed in a note from Raiffeisen Zentralbank, 2008), both also obtained on 2008-04-21.

This section is organized as follows: We first discuss some basic features of sports ratings, bookmakers odds, and sports tournaments in Section 4.2.1. Section 4.2.3 provides a data and tournament description for the EURO 2008 for which the various forecasts (using the bookmakers' expectations and using the World Football Elo ratings) are obtained and assessed in Section 4.2.4. Section 4.2.5 concludes the analysis of the UEFA EURO 2008.

### 4.2.1 Ratings of (prob)abilities in sports tournaments

#### Sports ratings

**Ratings of "abilities" or "strengths".** In competitive sports, players or teams as well as their supporters are interested in ratings of the competitors as a measure of their abilities or strengths. A common strategy for deriving suitable ratings employs adaptive schemes which update assessments based on historic performances upon availability of data about current performances. Typical examples for this include the FIFA/Coca Cola World rating in soccer or the ATP (Association of Tennis Professionals) rating in tennis (see Stefani, 1997, for an overview). Some ratings are based on a simple point system while others employ statistical models, e.g., the Elo rating (Elo, 2008) implies pairwise winning expectancies (see Joe, 1991). A natural application of ability ratings is to employ them for forecasting performances in future matches (e.g., Song et al., 2009). In some sports, ratings are also used for deriving seedings which in turn can be used for forecasting as in Boulier and Stekler (1999).

**Bookmakers odds as ratings of winning probabilities.** A rather different source of "ratings" of competitors in sports are bookmakers odds: Unlike the ratings discussed above these are not derived directly from past performances but emerge from "expert" knowledge. Of course, this typically

encompasses knowledge about past results but may also take into account expectations about future events. Due to the increasing popularity of online sports betting, bookmakers odds are a type of data that is abundant and easily available and that has been successfully employed in forecasts of single matches (e.g., Vlastakis et al., 2009; Spann and Skiera, 2009; Song et al., 2007; Forrest et al., 2005b; Dixon and Pope, 2004; Boulier and Stekler, 2003). Another important difference between bookmakers odds and the ability ratings discussed above is that they are an assessment of outcome probabilities (e.g., winning probabilities in the case of sports tournaments) rather than of the underlying abilities.

### 4.2.2 Sports tournaments

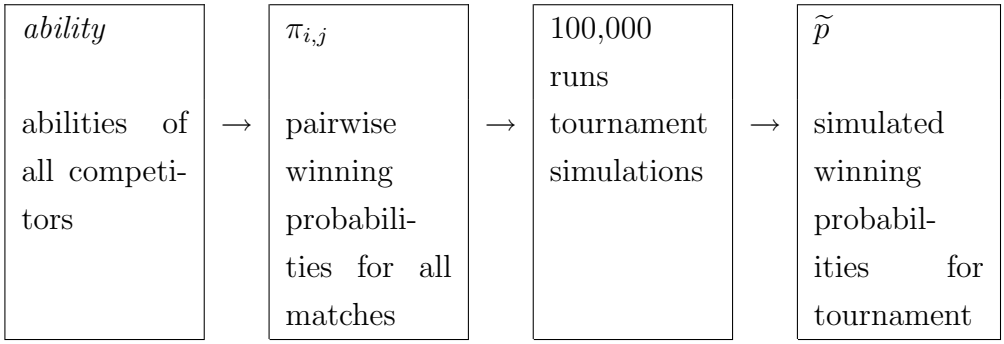
**Pairwise comparisons.** In many sports disciplines, winners and losers are determined by pairwise comparisons, called matches or games. Clearly, the outcome from a match depends on the current abilities of the two competitors. Given abilities measured on a ratio scale, the classical method for computing winning probabilities from abilities is the Bradley and Terry (1952) approach which derives the probability for competitor $i$ beating competitor $j$ as:

$$\pi_{i,j} = \frac{ability_i}{ability_i + ability_j} \qquad (i \neq j), \tag{4.4}$$

where $ability_i$ is the ability for team $i$ on a ratio scale. However, for many sports rating systems it is not clear what the underlying measurement scale is. A notable exception is the Elo rating (Elo, 2008) which uses a similar approach for obtaining winning expectancies. Hence, as discussed in detail below, Elo ratings can easily be transformed to abilities in the sense of Equation 4.4.

**Tournament schedule.** If a winner shall be determined from a group (rather than just a pair) of competitors, this is typically accomplished by using a sequence of pairwise comparisons, called tournament. Various designs are available for constructing suitable schedules for such a tournament

(see Scarf and Bilbao, 2006, for a discussion). In a round-robin tournament, where each competitor (or player or team) plays each other, it is obvious that the strongest competitor has the highest winning probability in each pairwise comparison and therefore the highest chance to win the tournament, followed by the second strongest competitor and so on. However, for other tournament schedules the strongest competitor does not necessarily have the highest probability of winning. For example, if the tournament schedule is based on a draw of a group phase and/or a knockout phase, some competitors might be favored/discriminated by being drawn together with relatively weak/strong competitors. However, when the tournament schedule and the abilities of its participants are known, it is (in principle) straightforward to compute the associated winning probabilities based on the pairwise probabilities from Equation 4.4 by applying conditional probabilities to all possible tournament "paths". As explicit enumeration of all paths can be burdensome, the winning probabilities can also be approximated easily by simulating a large number of tournament runs (100,000 say) and then assessing the empirical winning proportions $\widetilde{p}$ for each competitor:




The resulting (approximated) winning probabilities $\widetilde{p}(\textit{ability})$ then also capture all "tournament effects" induced by the schedule. Note that this approach models the contestants' abilities as constant over the course of the competition and might be further enhanced to accommodate hypothesized patterns of change in abilities. Also, this generic simulation setup might require adaptation to some details of a specific tournament, e.g., for EURO 2008 potential ties after the group phase need to be resolved (as described in detail in Section 4.2.4).

### 4.2.3 EURO 2008: Data and tournament description

#### Data

**Elo ratings.** The World Football Elo Ratings (Advanced Satellite Consulting Ltd, 2008), Elo ratings for short, for all 16 teams participating in the EURO 2008 have been collected from `http://www.eloratings.net/` (accessed 2008-04-21). In contrast to many other sports rating systems (such as the FIFA ratings below), the Elo ratings imply winning expectancies for pairwise comparisons (see Elo, 2008, Equation 46). The probability that team $i$ beats team $j$ can be related to

$$\pi_{i,j} = \frac{1}{10^{-(Elo_i - Elo_j)/400} + 1} \qquad (i \neq j), \tag{4.5}$$

where $Elo_i$ and $Elo_j$ are the Elo ratings for teams $i$ and $j$, respectively. For home teams (i.e., Austria and Switzerland in the EURO 2008), 100 rating points are added to the Elo rating (Advanced Satellite Consulting Ltd, 2008). Thus, the Elo ratings are essentially on a $\log_{10}$ scale which is somewhat different from the standard Bradley and Terry (1952) model. However, using Equations 4.4 and 4.5, it is easy to provide a transformation to log-abilities in the Bradley-Terry sense which imply the same pairwise winning probabilities $\pi_{i,j}$. As the log-abilities are just defined up to a constant $\gamma$, we choose $\gamma$ such that they are on a logit scale:

$$\log\left(ability_i^{(ELO)}\right) = \frac{\log(10)}{400} Elo_i + \gamma, \tag{4.6}$$

$$\sum_i \text{logit}^{-1}\left(\log\left(ability_i^{(ELO)}\right)\right) = 1, \tag{4.7}$$

where Equation 4.7 implies $\gamma = -13.496$ for the EURO 2008 data, log is the natural logarithm, and $\text{logit}^{-1}$ denotes the inverse of the logit function. The resulting Elo log-abilities are provided in Table 4.1 where the logit scale facilitates comparison with logits of tournament winning probabilities derived in the following.

**Bookmakers odds.** Longterm odds for winning the EURO 2008 were obtained from the websites of 45 international bookmakers for all 16 participating teams on 2008-04-21. These are all of 50 European top-selling online sports bookmakers who offered odds for this event. Prior to all further analysis, the odds are adjusted by removing the stake and a bookmaker-specific overround (see Equation 4.1) and then transformed to winning probabilities by means of Equation 4.2. This yields tournament winning probabilities $p_{i,b}$ for $i = 1, \ldots, 16$ teams and $b = 1, \ldots, 45$ bookmakers which reflect the bookmakers' beliefs about the outcome of the EURO 2008.

**FIFA ratings.** The FIFA/Coca Cola World ratings (Fédération Internationale de Football Association, 2008), FIFA ratings for short, for all 16 participating teams were retrieved from `http://www.fifa.com/` on 2008-04-21. These ratings capture abilities of the teams but on an unknown scale so that it is not straightforward to compute pairwise winning probabilities $\pi_{i,j}$ or tournament winning probabilities $p_i$ (see McHale and Davies, 2007, for an approach for building more complex statistical models based on the FIFA rating). Therefore, in the following, the FIFA ratings are employed only for comparison as a ranking (rather than rating).

### The tournament

The UEFA EURO 2008 is a tournament where 52 European national football teams (UEFA's members) compete in a multi-stage modus (qualification, group and knockout stage) to determine the European champion. First, 16 teams are determined via a qualification stage for the group stage, i.e., the main EURO 2008 tournament carried out in June 2008 in Austria and Switzerland. Table 4.1 lists the 16 teams as drawn into four groups, labeled A through D. Each group of four plays a round-robin—every team plays every other team, for a total of six matches within the group—and the top two teams in each group advance to the next stage, the quarter-final. The winner of group A plays against the second best team of group B (first quarter-final) and the winner of group B plays against the second best team of group A

(second quarter-final). Analogously, the winner of group C plays against the second best team of group D (third quarter-final) and the winner of group D plays against the second best team of group C (forth quarter-final). The four winners of the quarter-finals reach the semi-finals, where the winner of the first quarter-final plays against the winner of the second one and the winner of the third quarter-final plays against the winner of the forth. The winners of the semi-finals then play the final and the winner of the final is the European football champion (Union of European Football Associations, 2009a).

### 4.2.4 Forecasting of the EURO 2008

In this section, forecasts of team (log-)abilities and winning probabilities for the EURO 2008 tournament are obtained based on the Elo ratings and the bookmakers odds, respectively. The resulting four quantities are compared with the actual result of the tournament and the best-performing method is analyzed in some more detail.

#### Forecasting based on the Elo ratings

As argued in Sections 4.2.1 and 4.2.3, the Elo ratings $Elo_i$ $(i = 1, \ldots, 16)$ represent an assessment of the current ability/strength of the teams participating the EURO 2008. By construction, pairwise probabilities $\pi_{i,j}$ for all combinations of participants can be obtained. Furthermore, to approximate winning probabilities that include tournament effects such as the group draw, the empirical winning proportions from 100,000 simulated tournaments are used:

$$ability_i^{(ELO)} = \exp\left(\frac{\log(10)}{400} Elo_i - 13.496\right), \tag{4.8}$$

$$p_i^{(ELO)} = \widetilde{p}\left(ability^{(ELO)}\right)_i. \tag{4.9}$$

Thus, $ability^{(ELO)}$ is the vector of abilities (in the Bradley-Terry sense) based on which the tournament simulations are carried out. The results for all teams are reported in Table 4.1.

By adopting the classical Bradley-Terry model, the simulation of each match yields only a winner and a loser without the possibility of a tie and without further information about the number of goals or the goal difference. This is sufficient for the knock-out stage of the tournament as it reflects that the actual matches always have a winner (if necessary in overtime and penalties). However, for the group phase within the simulation this approach might result in tied teams. If necessary, we resolve such ties by additional "fictitious" matches between the tied teams to obtain unique winners and the runner-ups of the groups.

Our simulation method could be extended by using more elaborate simulation techniques including ties and number of goals, e.g., a model where the team scores follow independent Poisson distributions (e.g., Maher, 1982; Dixon and Coles, 1997; Dyte and Clarke, 2000), or an ordered probit regression model (Goddard and Asimakopoulos, 2004).[1]

According to the Elo rating, Italy is the strongest team ($\log(ability^{(ELO)}) = -1.97$) and also has the highest probability for winning the tournament ($p^{(ELO)} = 18.28\%$). However, the second strongest team France has only the third highest winning probability ($\log(ability^{(ELO)}) = -2.09$, $p^{(ELO)} = 14.08\%$) while Germany is only the fifth strongest but has the second highest winning probability ($\log(ability^{(ELO)}) = -2.34$, $p^{(ELO)} = 15.99\%$). Thus, team Germany clearly profits from being drawn in a group (B) with weaker competitors while France has a certain disadvantage from being placed in a group (C) with strong competitors such as Italy. This tournament effect can be conveniently assessed by comparing differences between the teams' log-abilities and their winning logits, respectively (as both measurements have been constructed such that they are on a logit scale). For example, Italy's margin over Germany of 0.37 ($= -1.97 - (-2.34)$) is reduced to 0.16

[1]However, all approaches should give reasonable approximations of the probabilities for being promoted to the next round.

| | $\log(ability_i)$ | | $p_i(\%)$ | | $\text{logit}(p_i)$ | | |
|---|---|---|---|---|---|---|---|
| | ELO | BCM | ELO | BCM | ELO | BCM | Group |
| Germany | −2.34 | −2.33 | 15.99 | 17.45 | −1.66 | −1.55 | B |
| Spain | −2.25 | −2.41 | 13.14 | 12.21 | −1.89 | −1.97 | D |
| Italy | −1.97 | −2.40 | 18.28 | 11.34 | −1.50 | −2.06 | C |
| Portugal | −2.95 | −2.54 | 3.36 | 9.97 | −3.36 | −2.20 | A |
| France | −2.09 | −2.50 | 14.08 | 9.14 | −1.81 | −2.30 | C |
| Netherlands | −2.33 | −2.62 | 8.29 | 6.77 | −2.40 | −2.62 | C |
| Croatia | −2.86 | −2.77 | 5.03 | 6.72 | −2.94 | −2.63 | B |
| Czech Republic | −2.67 | −2.74 | 7.17 | 5.88 | −2.56 | −2.77 | A |
| Switzerland | −2.79 | −2.88 | 5.18 | 3.92 | −2.91 | −3.20 | A |
| Greece | −2.93 | −2.91 | 2.76 | 3.31 | −3.56 | −3.37 | D |
| Sweden | −3.32 | −2.98 | 0.77 | 2.87 | −4.86 | −3.52 | D |
| Russia | −3.42 | −3.00 | 0.55 | 2.72 | −5.20 | −3.58 | D |
| Turkey | −3.27 | −3.06 | 1.30 | 2.26 | −4.33 | −3.77 | A |
| Romania | −2.72 | −3.04 | 2.77 | 2.12 | −3.56 | −3.83 | C |
| Poland | −3.35 | −3.19 | 1.19 | 2.05 | −4.42 | −3.87 | B |
| Austria | −3.93 | −3.85 | 0.14 | 0.93 | −6.55 | −4.67 | B |

Table 4.1: Log-abilities, winning proabilities, and corresponding logits of all teams for the EURO 2008 based on the Elo rating (ELO) and on the bookmaker consensus model (BCM). The ELO log-abilities are directly computed from the Elo ratings and winning probabilities are derived via simulation. The BCM logits are estimated by team-wise means of bookmaker log-odds, the corresponding log-abilities are found by "inverse" simulation. The rows are sorted by the BCM winning probabilities.

$(= -1.5 - (-1.66))$ by including tournament effects while France's margin over Germany of 0.25 is reversed to −0.15. Furthermore, it is worth noting that team Spain, the favorite in group D, has the fourth highest winning probability ($p^{(ELO)} = 13.14\%$) while Austria has the lowest chances of winning the EURO 2008 ($p^{(ELO)} = 0.14\%$), nonwithstanding its potential home advantage (see e.g., Forrest et al., 2005a; Clarke and Norman, 1995).

**Forecasting based on bookmakers odds**

When appropriately adjusted and transformed, as described above, the bookmakers odds yield expected winning probabilities $p_{i,b}$ for each team $i = 1, \ldots, 16$ and bookmaker $b = 1, \ldots, 45$. In the following, a single forecast for the winning probability of each team is obtained by aggregation of the $p_{i,b}$ across bookmakers. Subsequently, a vector of underlying team abilities is found by "inverse" application of the simulation approach adopted above.

The bookmakers odds are prospective ratings of the performance of the 16 participating teams in the EURO 2008 which vary between 45 bookmakers. In order to obtain an aggregated measure for each team some sort of consensus between we use the general model framework for bookmakers odds stemming from a variety of bookmakers introduced above (Equation 4.3).

As the bookmakers' expectations about the EURO 2008 are rather homogeneous a straightforward fixed-effects model with zero-mean deviations $\epsilon_{i,b}$ should be appropriate. Thus, the consensus winning logits are simply means across bookmakers:

$$\widehat{\text{logit}(p_i)} = \frac{1}{45} \sum_{b=1}^{45} \text{logit}\left(p_{i,b}\right).$$

Transforming these winning logits back to the probability scale yields the bookmakers' consensus winning probabilities $p_i^{(BCM)}$. Both probabilities and corresponding logits for this bookmakers consensus model (BCM), are shown in Table 4.1. The model captures 98.21% of the variance of the $p_{i,b}$, the associated estimated standard error of $\epsilon_{i,b}$ is 0.11396.

Although forecasting the winning probabilities for the EURO 2008 is the main concern in our investigation, there is also interest in the team abilities underlying the bookmakers' expectations. The tournament schedule was already known at the time the bookmakers odds were retrieved, and hence should be included in the expectations about the outcome of the tournament. In order to strip the "tournament effects" (see Section 4.2.1) from this

measure, we employ an "inverse" application of the simulation approached described in the previous sections. More precisely, we want to find a set of team abilities $ability_i$ $(i = 1, \ldots, 16)$ that result in simulated winning probabilities $\widetilde{p}(ability)_i$ that are as similar as possible to the consensus winning probabilities $p_i^{(BCM)}$:

$$p_i^{(BCM)} = \text{logit}^{-1}\left(\widehat{\text{logit}(p_i)}\right), \tag{4.10}$$

$$ability^{(BCM)} = \underset{ability}{\text{argmin}} \sum_{i=1}^{16} \left| p_i^{(BCM)} - \widetilde{p}\,(ability)_i \right|. \tag{4.11}$$

The minimum in the second line is determined using a local search strategy for the full vector $ability^{(BCM)}$ where 100,000 tournament runs are employed in each evaluation of $\widetilde{p}(\cdot)$. The results are reported in Table 4.1.

According to the BCM, Germany has the highest chances of winning the EURO 2008 ($p^{(BCM)} = 17.45\%$) with some margin over Spain (12.21%) and Italy (11.34%). Thus, although there is considerable overlap among the top five teams obtained from BCM and Elo results, the ranking and associated winning probabilities of these teams are rather different. Also, France (which was the second strongest team according to the Elo rating) has only the fifth largest winning probability (9.14%). Finally, host country Austria is again expected to have the lowest winning probability (0.93%) but it is somewhat larger in absolute terms compared to the Elo forecast.

In order to investigate the tournament effect, differences in the teams' winning logits can again be compared with differences in their log-abilities. Again, this shows that Germany greatly profits from the group draw because its margin in terms of winning logits over Spain or Italy (0.42 and 0.51, respectively) is greatly reduced in terms of log-abilities (0.08 and 0.07). Note also that this reduction is larger for Italy than for Spain, conveying that Italy suffers particularly from being drawn in the strong group C (often referred to as the "group of death").

### Ex post comparison of all forecasts

The previous subsections present two different types of forecasts (abilities and winning probabilities) derived from two different types of ratings (Elo rating and bookmakers odds). As usual in forecasting, it is of central interest which strategy performs best in practice. Although this is difficult to answer because there are no "real" replications of the tournament, we can compare the forecasts with the single real outcome of the EURO 2008.

Table 4.2 assesses the predictive performance of all four forecasts by comparing them with the actual tournament outcomes using Spearman's rank correlation. For the actual results, a total ranking including ties is employed, as commonly used in rankings of such incomplete tournaments.[2] First, this shows that the winning probabilities (including the tournament effects) have higher correlation with the actual outcome (0.525 for BCM and 0.304 for ELO, respectively) compared to the corresponding (log-)abilities (0.441 and 0.203). Second, the forecasts based on the bookmakers odds clearly outperform those based on the Elo ratings. This conveys that the prospective ratings of experts (i.e., the bookmakers) have been more useful than the retrospective performance-based Elo ratings.

In addition to the four forecasts derived in this section, Table 4.2 also provides correlations with the ranking implied by the FIFA/Coca Cola World rating. Interestingly, this has a higher Spearman correlation (0.373) with the tournament outcome than the Elo forecasts. Furthermore, it is more closely associated with both (log-)ability measurements (0.841 and 0.815) than with the corresponding winning probabilities (0.809 and 0.809). This confirms that the (retrospective) FIFA rating is an assessment of the teams' current ability and conveys that its predictive power could be enhanced if the corresponding winning probabilities could be computed or simulated. However, as no rigorous method for computing pairwise winning probabilities $\pi_{i,j}$ based on the FIFA rating is known to us, we cannot pursue this approach here.

---

[2]Various strategies for dissolving the ties have been explored but did not lead to qualitatively different results.

| | $p^{(BCM)}$ | $ability^{(BCM)}$ | $p^{(ELO)}$ | $ability^{(ELO)}$ | FIFA |
|---|---|---|---|---|---|
| Tournament ranking | 0.525 | 0.441 | 0.304 | 0.203 | 0.373 |
| $p^{(BCM)}$ | | 0.988 | 0.871 | 0.771 | 0.809 |
| $ability^{(BCM)}$ | | | 0.909 | 0.826 | 0.841 |
| $p^{(ELO)}$ | | | | 0.956 | 0.809 |
| $ability^{(ELO)}$ | | | | | 0.815 |

Table 4.2: Spearman's rank correlation between the actual tournament ranking and rankings according to the estimated BCM winning probabilities and (log-)abilities, simulated Elo winning probabilities and (log-)abilities (equivalent to the original Elo rating), and the FIFA/Coca Cola World rating.

In order to investigate the sources of the good performance of the BCM for the winning probabilities, it is useful to extract the two best-ranked teams from each group in Table 4.1. This shows that the consensus winning probabilities correctly predict five teams (Germany, Spain, Italy, Portugal, Croatia) which played the quarter-finals, as well as the actual final (played by the teams Germany and Spain). The big surprises of the tournament were teams Russia and Turkey which both reached the semi-finals rather unexpectedly. Whereas the BCM ranked team Russia better than the Elo and the FIFA rating, the converse is true for team Turkey. Furthermore, France surprisingly did not reach the quarter-finals which was neither expected by the bookmakers nor using the Elo or FIFA ratings. However, it was somewhat more likely using the BCM.

**Tournament analysis based on the BCM forecast**

In addition to the team abilities and winning probabilities (Table 4.1), some further insights can be gained from the best-performing BCM forecast due to adoption of the simulation approach. So far, we have only considered the empirical winning proportions of each team in the 100,000 tournament runs. But, of course, the empirical proportions of reaching the quarter-final, semi-final, and final can be extracted as well. Figure 4.1 shows the performance of

each team in the simulations based on $ability^{(BCM)}$ as a performance curve (or "survival" curve over the course of the tournament). The endpoints of the curves are the simulated winning probabilities, which are by construction (Equation 4.14) (roughly) identical to the probabilities derived from the BCM (Table 4.1).

The performance curves in Figure 4.1 show that groups B and D are rather heterogeneous with weaker teams and clear favorites (Germany and Spain, respectively) while groups A and C are rather homogeneous. This group effect can also be quantified on an aggregated level by considering deviations of the mean group winning logits (computed from Table 4.1) from the overall mean winning logits across all teams. Despite the fact that group B includes the bookmakers' favorite of winning the European championship (Germany), group B clearly is the weakest group and has the smallest chance to include the winner (with a deviation of $-0.187$ on the logit scale). This is followed by group D with a deviation of $-0.116$. Group C, on the other hand, is clearly the toughest group and has the greatest probability of including the champion (0.293). Group A can be interpreted as the average group with a deviation of 0.010 from the overall mean.

The simulation also provides information about the most likely coupling for the final: A match of Germany and Spain, the actual final, occurs with the highest probability of 20.45%. Given this coupling in the final, the winning probabilities of both teams are given by the Bradley-Terry model (Equation 4.4) based on the teams' estimated abilities $ability_i^{(BCM)}$. Although team Germany has a slight advantage with a winning probability of 52.08%, this essentially conveys that no clear favorite exists in this final. This is confirmed by the actual EURO 2008 final which ended with a very close result: Germany 0, Spain 1.

### 4.2.5 Discussion

We embedded various methods for rating players/teams in competitive sports into a common framework that allows for forecasting winning probabilities in

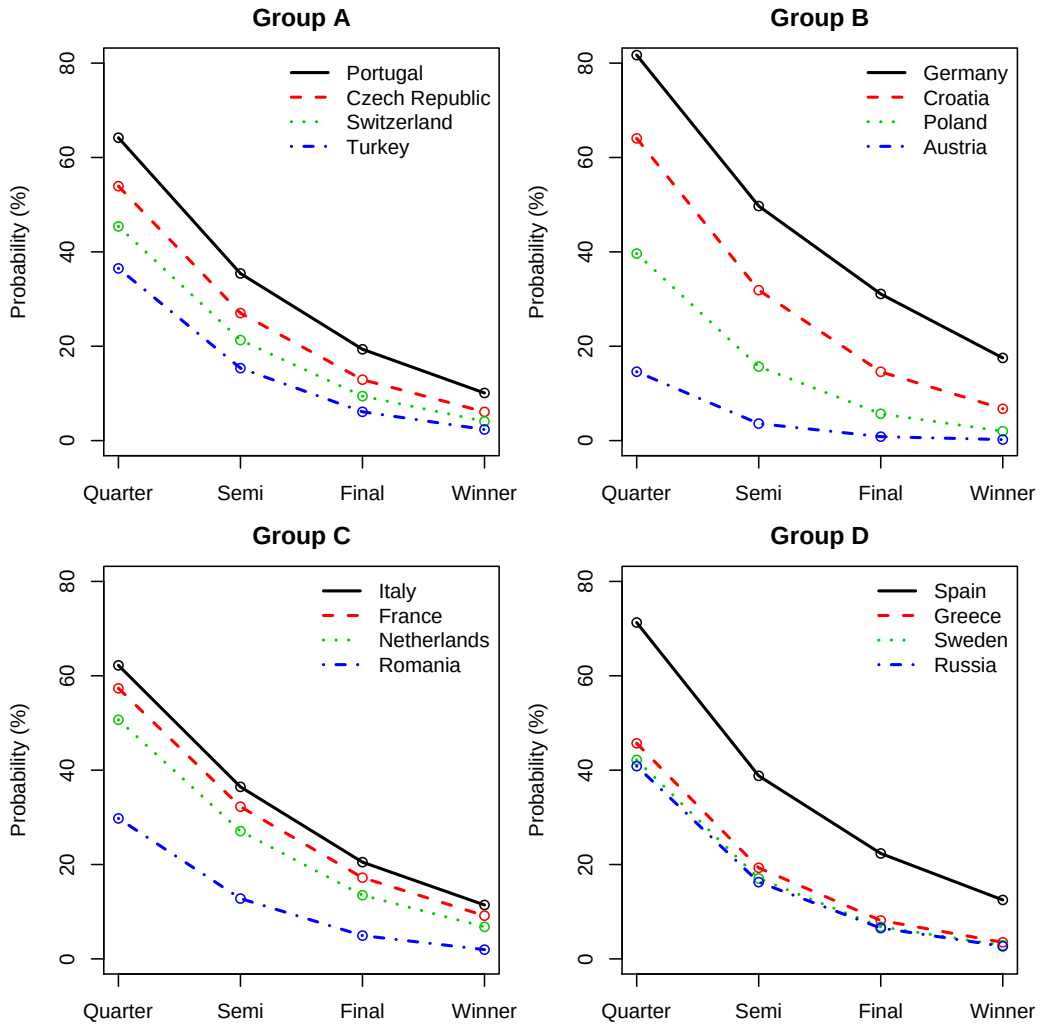



Figure 4.1: Simulated probabilities (from 100,000 tournament runs based on the BCM consensus abilities) for reaching the quarter-final, the semi-final, the final, and for winning the EURO 2008.

sports tournaments (rather than single matches) and obtaining the competitors' underlying strengths/abilities. The link between abilities and winning probabilities is established by means of a simulation approach that takes into account potential tournament effects such as group draws or seedings. Specifically, these methods are applied to the World Football Elo rating and the odds from a set of international bookmakers and assessed using forecasts of the European football championship 2008. A consensus model for the

bookmakers odds performs best in this comparison, correctly predicting the actual final of the tournament and revealing clear tournament effects due to the group draw.

Although the model forecasts provide promising results for the EURO 2008, various improvements are conceivable and deserve further study: The tournament simulation could be enhanced to provide not only winners and losers but more realistic results (such as goals or goal differences in a soccer tournament). The bookmaker consensus model adopted here only includes a fixed team effect but could be extended to encompass further fixed or random effects capturing, for example, group strengths, bookmaker bias, or differences in variance.

Furthermore, our results convey that the prospective rating based on aggregated expert judgment in the bookmaker consensus model provides more accurate forecasts of sports tournament outcomes compared to retrospective ratings that derive current team/player abilities from past performances. Application of both approaches to future tournaments will continue to explore the potential of these methods and help to establish a more complete picture.

## 4.3 Wimbledon 2009

Furthermore, we apply our general model framework for bookmakers odds stemming from a variety of bookmakers (Equation 4.3) to the Men's singles of Wimbledon 2009. Wimbledon is the oldest tennis tournament, being held at the All England Club in the London suburb of Wimbledon since 1877. It is the most popular tournament played on grass in the world and belongs to the four annual major tennis tournaments, the Grand Slams, along with the Australian Open, the French Open and the US Open (Wimbledon, 2009).

In the Men's singles of Wimbledon 2009 the top seeded and defending champion Rafael Nadal withdrew from the tournament due to injury days prior to the tournament. Here, we analyze the effects of this withdrawal, especially on the expected ability of the bookmakers' favorite Federer.

This section is organized as follows: Section 4.3.1 provides a data and tournament description of Wimbledon 2009 for which the players' abilities are modeled and analyzed in Section 4.3.2. Section 4.3.3 concludes the analysis of Wimbeldon 2009.

### 4.3.1 Wimbledon 2009: Tournament and Data Description

#### The tournament

In the Men's Singles of Wimbledon 2009, a total of 128 international tennis players compete in a single elimination tournament modus (knockout system) to determine the "best" tennis player on grass. Players wishing to enter Wimbledon are required to submit their entry on a special form. The organizing committee evaluates all applications for entry, and use ATP rankings to determine which players will be admitted directly into the tournament, those who have to qualify and those who are rejected. A player without a high enough ATP ranking can be admitted as a "wild card" by the committee. Wild cards are usually offered on the basis of past performance at Wimbledon or to increase British interest. A player who neither has a high enough ranking nor receives a wild card can participate in a qualifying tournament (a three-round event) held one week before Wimbledon. The players who win all three rounds will progress. "Lucky losers" are losers from the final round of qualifying competitions — chosen in order of ATP rankings — to fill any vacancy which occurs in the draw before the first round has been completed. The committee seeds the top 32 players based on their ATP rankings in order to make sure that the top 32 players do not meet each other in the tournament before the third round. The seedings can also be changed due to players' previous grass court performance by the committee (see Wimbledon, 2009).

## Data

**Bookmakers Odds.** Long-term odds of winning Wimbledon 2009 (Men's Singles) were obtained from the website `http://odds.bestbetting.com` which compares odds of a variety of international bookmakers. We obtained all available odds on two different dates, 2009-06-16 (before the tournament draw and before Nadal's withdrawal; henceforth called W1) and on 2009-06-22 (before the tournament started, but after the draw; henceforth called W2). The first dataset contains odds of 17 international bookmakers for 96 players who are expected to participate in Wimbledon 2009. The latter dataset contains odds of 15 international bookmakers for 105 participating players.

In order to recover the underlying beliefs of the bookmakers, we adjust the quoted odds as described above (4.1). This adjustment is done separately for all bookmakers yielding bookmaker-specific overrounds and expected winning probabilities $p_{i,b}$ for each player $i$ and bookmaker $b$ derived from the adjusted odds.

**ATP Rankings (Singles).** The South African Airways ATP rankings (singles) is based on the players' results (measured in points) at the four Grand Slams, the eight mandatory ATP World Tour Masters 1000 tournaments and the Barclays ATP World Tour Finals of the ranking period, and the best four results from all ATP World Tour 500 tournaments played in the calendar year. We obtained the points assigned to the rankings (henceforth called ATP ratings) from 2009-06-22 from ATP's website for all 128 participating players and for the injured player Rafael Nadal (Association of Tennis Professionals, 2009).

**Seeding and Draw for Wimbledon 2009.** As described above, the Wimbledon organizing committee seeds the top 32 players of the tournament based on their ATP rankings and their previous grass court performance. We obtained the seeding for Wimbledon 2009 from 2009-06-17 and from 2009-

06-19 (after Nadal's withdrawal) from the Wimbledon webpage (Wimbledon, 2009). Additionally, we obtained the draw from 2009-06-19. According to the Wimbledon seeding from 2009-06-17 Nadal was the top seeded player, followed by Federer, Murray, Djokovic, and Del Potro. Due to Nadal's withdrawal after the draw, the committee left the top position blank, and seeded the previously unseeded player Kiefer as 33 and included Thiago Alves as a lucky loser to the draw. The draw changed in that way, that Del Potro (seeded on 5) took the place from Nadal, Blake seeded as 17 took Del Potro's place, and Kiefer took Blake's place.

### 4.3.2 Modeling Players' Abilities

The focus here is to analyze the effect of Nadal's withdrawal from Wimbledon 2009, especially on the expected abilities of the main competitor Federer. It is obvious that Nadal's withdrawal increases, on average, the chance of winning the tournament of all other players. However, the ability/strength of each player should not change. Thus, the winning probability for a specific match, e.g., Federer beating Murray in a potential Wimbledon 2009 final, should not be affected by Nadal's withdrawal. The "true" abilities of the players are unknown, but an approximation can be derived from performance measures or winning expectancies, like the ATP rating, the seedings, or the bookmakers odds. Here, we compare all three rating strategies in a forecasting study for Wimbledon 2009. Aa above, we find that a consensus derived from the (prospective) bookmakers odds has higher predictive power than retrospective ratings based on historical results (in this study, the Wimbledon seeding and the ATP rankings, see Table 4.4). Subsequently, we estimate players' abilities based on bookmakers odds using two different odds sets: one including winning expectancies for Nadal and one obtained after his withdrawal. The resulting expected abilities are compared to assess the effect on Nadal's withdrawal. Furthermore, we use the players' abilities in order to compare different tournament designs in a simulation study.

### Consensus Information

Since the bookmakers' expectations about Wimbledon 2009 are rather homogeneous, we use again the very straightforward aggregation strategy computing the means of the winning logits (i.e., winning log-odds) to find appropriate consensus measures of all bookmakers:

$$\widehat{\text{logit}(p_i)} = \frac{1}{B}\sum_{b=1}^{B} \text{logit}\left(p_{i,b}\right), \tag{4.12}$$

where $B$ is the number of bookmakers.

Transforming these consensus winning logits back to the probability scale yields the bookmakers' consensus winning probabilities $\widehat{p_i}$ for each player $i$ for whom odds are available.

Table 4.3 shows the estimated winning probabilities $\widehat{p_i}$ and their associated winning logits $\widehat{\text{logit}(p_i)}$ of the top ten participating players of Wimbledon 2009 using the winning odds W1 and W2.

According to the BCM for W1 and W2, Federer has the highest chance of winning Wimbledon 2009 (W1: 38.52%, and W2: 45.95%). Federer, is followed by Murray with a clear distance (W1: 18.31%, and W2: 23.00%). The expected winning probability of the top seeded player Nadal is clearly below the top two (14.19%). His withdrawal increases the winning probabilities of both players strongly, whereas the winning probabilities of all other players do not change as clearly.

In order to test the predictive power of the bookmaker consensus we compare the consensus winning logits including the last available information (W2) with the actual tournament outcome, the Wimbledon seeding, and the ATP ranking of the top ten players using Spearman's rank correlation (Table 4.4).

Although the correlation between the bookmaker consensus winning probabilities and the actual tournament outcome is rather low (0.109) the BCM still performs better than the Wimbledon seeding ($-0.156$) and the ATP ranking ($-0.185$). Both, the seeding and the ATP ranking have a negative

| | $\widehat{p}_i(\%)$ | | $\widehat{\text{logit}(p_i)}$ | | $\log(ability_i)$ | | $\widetilde{p}_i(\%)$ | |
|---|---|---|---|---|---|---|---|---|
| | W1 | W2 | W1 | W2 | W1 | W2 | W1 | W2 |
| Federer | 38.52 | 45.95 | −0.47 | −0.16 | −3.63 | −3.32 | 38.68 | 46.17 |
| Murray | 18.31 | 23.00 | −1.50 | −1.21 | −4.41 | −4.03 | 18.50 | 23.04 |
| Nadal | 14.19 | | −1.80 | | −4.49 | | 14.40 | |
| Djokovic | 5.94 | 5.68 | −2.76 | −2.81 | −4.67 | −4.75 | 6.09 | 5.84 |
| Roddick | 2.53 | 3.30 | −3.65 | −3.38 | −5.05 | −4.88 | 2.62 | 3.50 |
| Del Potro | 2.74 | 3.03 | −3.57 | −3.47 | −5.07 | −4.91 | 2.96 | 3.29 |
| Tsonga | 3.33 | 3.01 | −3.37 | −3.47 | −4.91 | −4.84 | 3.49 | 3.16 |
| Söderling | 1.84 | 1.29 | −3.98 | −4.34 | −5.07 | −5.07 | 2.04 | 1.42 |
| Verdasco | 1.43 | 1.22 | −4.23 | −4.40 | −5.34 | −5.23 | 1.61 | 1.38 |
| Haas | 0.81 | 1.12 | −4.81 | −4.49 | −5.65 | −5.33 | 1.01 | 1.27 |
| Hewitt | 0.43 | 0.78 | −5.44 | −4.84 | −5.38 | −5.30 | 0.62 | 0.92 |

Table 4.3: Estimated winning probabilities $\widehat{p}_i$, their associated winning logits $\widehat{\text{logit}(p_i)}$, estimated log-abilities $\log(ability_i)$ and associated simulated winning probabilities $\widetilde{p}_i$ of the top ten participating players of Wimbledon 2009 and Nadal using their winning odds from 2009-06-16 (W1) and from 2009-06-22 (W2).

Spearman's rank correlation with the actual tournament outcome, assigning rather high ranks to two players who reach the quarter-finals (Hewitt) or the semi-finals (Haas).

In addition to the correlation, we analyze the correctly predicted participants of each round (third round to winner). Table 4.5 shows that the BCM correctly predicts nine players of the last 16, whereas the Wimbledon seeding predicts only seven and the ATP ranking only eight players correctly. Furthermore, the BCM correctly predicts five of the last eight and three of the last four, everytime one more than the Wimbledon seeding and the ATP ranking. All three approaches forecast the actual winner Federer correctly, but expected Murray who was beaten by Roddick in the semi-finals, as the runner-up.

Nevertheless, the ex post analysis shows that the correlation between the bookmakers expectancies for Wimbledon 2009 and the actual tournament

| | BCM | Seeding | ATP |
|---|---|---|---|
| Tournament ranking | 0.109 | −0.156 | −0.185 |
| BCM | | 0.688 | 0.792 |
| Seeding | | | 0.956 |

Table 4.4: Spearman's rank correlation between the actual tournament ranking and rankings according to the estimated BCM winning probabilities, the seeding, and the ATP rating of the top ten participating players of Wimbledon 2009.

| | Round of last ... | | | | |
|---|---|---|---|---|---|
| | 16 | 8 | 4 | 2 | 1 |
| BCM | 9 | 5 | 3 | 1 | 1 |
| Seeding | 7 | 4 | 2 | 1 | 1 |
| ATP | 8 | 4 | 2 | 1 | 1 |

Table 4.5: Correctly prediction of the last $16, 8, 4, 2$, and the winner using the (log-)abilities, the seeding, and the ATP raking of the top 128 participating players of Wimbledon 2009.

outcome is not high, but the bookmakers perform better than the Wimbledon seeding and the ATP ranking. The reasons for the difficulties in forecasting tennis are twofold. First, tennis is an individual sport competition and the outcome of a match/tournament depend only on one individual who can easily have a day off or an injury rather than a whole team. Second, in the tennis tournament design (single elimination tournament) every single match is important, if a player loses one match he is eliminated from the tournament.

### Estimation of Abilities

With the winning logits and associated winning probabilities we have computed measures for the specific tournament, Wimbledon 2009, including information about the tournament design (in W1 and W2) and including the

original draw (in W2). In order to obtain measures of the unknown "true" abilities of the players we have to adjust the winning logits by the tournament effects (tournament schedule and draw). I.e., we try to estimate the abilities which correspond with the winning logits. For this we employ again the well known Bradley and Terry (1952) model which measures abilities on a ratio scale and for which the probability $\pi_{i,j}$ for competitor $i$ beating competitor $j$ is given by:

$$\pi_{i,j} = \frac{ability_i}{ability_i + ability_j} \qquad (i \neq j), \tag{4.13}$$

where $ability_i$ is the ability for competitor $i$.

Given the abilities of all players and the tournament schedule, we can compute the associated winning probabilities based on the pairwise probabilities from Equation 4.13. Alternatively, we can simulate a large number of tournament runs (100,000 say) and then assessing the empirical winning proportions $\widetilde{p}$ for each competitor (see 4.2.1). I.e., for given abilities $ability_i$ $(i = 1, \ldots, 128)$ for all competitors we obtain the simulated winning probability $\widetilde{p}(ability)_i$ for competitor $i$. We can try to estimate the unknown "true" abilities by choosing them in a way that the $\widetilde{p}(ability)_i$ match the Bookmaker Consensus Model winning probabilities $p_i$ as closely as possible. In our case, we minimize the total absolute deviation between $p$ and $\widetilde{p}$, i.e., we solve the optimization for

$$ability \quad = \quad \underset{ability}{\operatorname{argmin}} \sum_{i=1}^{n} \left| p_i - \widetilde{p}\,(ability)_i \right|, \tag{4.14}$$

using a local search strategy.

In order to estimate the ability for each player, we need winning logits for all players. Due to the fact that not all players are assigned to odds, we do not obtain winning logits for all players derived from the BCM. Therefore, we use a simple linear model modeling the relationship between the ATP ratings on the log-scale and the consensus winning logits:

$$\text{logit}(p_i) \quad = \quad \beta_0 + \beta_1 \cdot \log(ATP), \tag{4.15}$$

and predicted the consensus winning logits of the "unrated" players. The relationships have a high correlation for both W1 and W2 (W1: 0.828, W2: 0.836). and the estimated model parameters for the slope $\beta_1$ and the intercept $\beta_0$ are 1.73 and $-18.79$ for W1, and 1.71 and $-18.66$ for W2. For ease of comparison, we show the estimated abilities on the log-scale and their associated simulated winning probabilities $\widetilde{p}_i$ (which match the winning probabilities $\widehat{p}_i$ derived from the BCM) of the top players of Wimbledon 2009 for W1 and W2 in Table 4.3. According to the estimated log-abilities Federer is still the best player of Wimbledon 2009 (W1: $-3.627$, W2: $-3.315$), followed again by Murray (W1:$-4.409$, W2: $-4.030$). If Nadal had played Wimbledon 2009, he was expected to be the third strongest player of the tournament (W1: $-4.492$, with an associated simulated winning probability of 14.40%).

In order to assess whether the ability of a player was altered due to Nadal's withdrawal, we compare the players estimated log-abilities by subtracting the log-abilities of a reference player. We choose Söderling, because he has rather similar log-abilities for W1 and W2. Thus, Figure 4.2 shows for each top ten player if the chance of beating Söderling increases or decreases after Nadal's withdrawal. The comparison of the log-abilities shows that the abilities of almost all top ten players (except Djokovic) increases, but primarily the abilities of Federer, Murray, and Haas. E.g., the probability that Federer beat Söderling increases from 80.84% to 85.25%.

The changes in the (log)abilities of the top two, Federer and Murray, show that the bookmakers do not react on Nadal's withdrawal and its consequential changes of the draw as expected. Apparently, they have not considered the whole tournament again and instead just increased Federer's and Murray's winning probabilities—presumably because they expected much more punters betting on a tournament by one of the two players. In any case, this explanation for the increase in Federer's and Murray's expected abilities seems to be far more plausible than interpreting the results literally as an increase in their abilities. In the latter case, one would have to argue that Federer and Murray are so relieved by the drop-out of Nadal that they even play stronger in matches against other players (such as Söderling). Further-

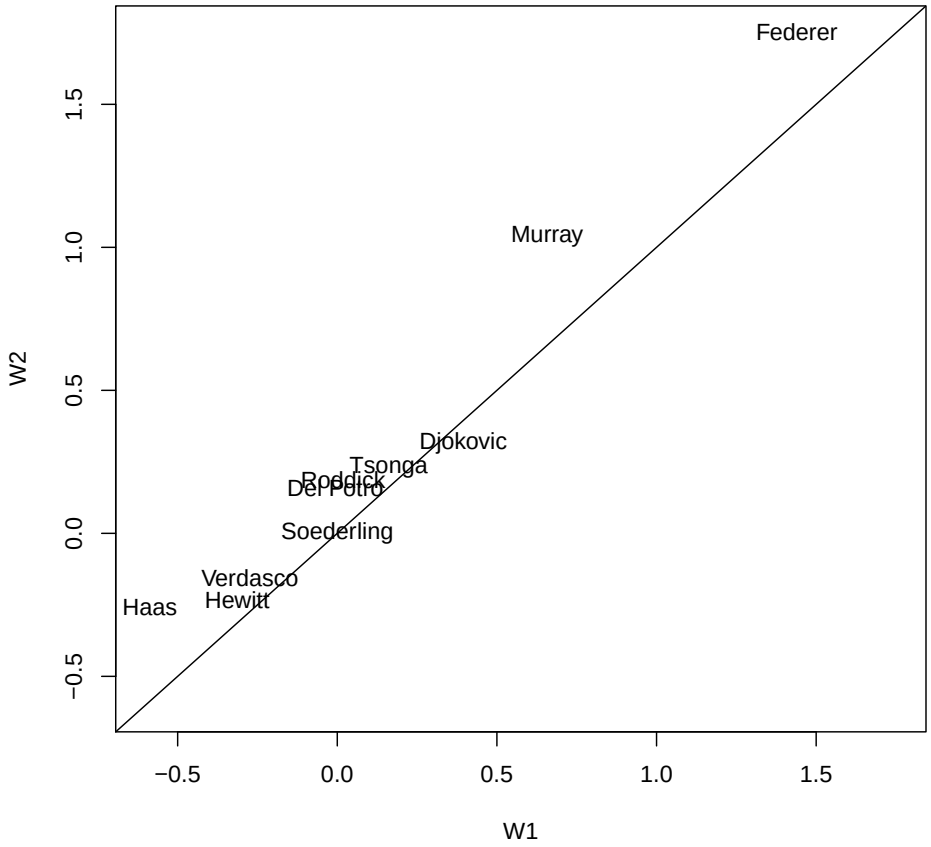



Figure 4.2: Comparison of the estimated log-abilities $\log(ability_i)$ of the top ten participating players of Wimbledon 2009 using the winning odds from 2009-06-16 (W1) and from 2009-06-22 (W2).

more, the changes in the abilities of Haas and Djokovic can be explained by a delayed reaction to the outcome of the Wimbledon warm up tournament in Halle, where Haas beat Djokovic rather clearly (6-3 6-7(4) 6-1) in the final. Although this information had already been available at time W1, it appears to have only been used in the odds at time W2—potentially due to a change in the punters' betting behavior in the week between the tournaments of Halle and Wimbledon.

### Effects of the Tournament Design

With the estimated abilities of the players a measure adjusted for the tournament effects is now available and we are able to determine the effects of different tournament designs by simulating winning probabilities of all participants. A tennis tournament is typically a single elimination tournament and so each match plays an important role. A player with the ambition of winning the tournament is not able to have a day off. Furthermore, in a tennis tournament like Wimbledon a specific number of players is seeded.

In order to determine the effects of the tennis tournament with its seeding, we compare three different designs: (1) a single elimination tournament with the original seeding and draw of Wimbledon 2009, (2) a single elimination tournament without seeding and random draw, and (3) a round-robin tournament, where each player plays each other ones. We use the estimated abilities from all 128 players of Wimbledon 2009 derived from the BCM (W2) and simulate their chances of winning the tournament according to the above described simulation approach (100,000 runs). For comparison reason we transform the empirical propabilities into winning logits and compare them for the top ten players in Figure 4.3. The winning logits of the single elimination tournament with seeding and without seeding differ not really much. Only some winning logits slightly increase (for a few of the weaker players) and some slightly decrease (e.g., Murrary and Djokovic) if the single elimination tournament is played without seeding. However, overall these differences are minor signalling that in the long run, the seeding does not have a large effect on the tournament outcome. In contrast, if we consider a round-robin where instead of 127 matches 8128 matches have to be played, the winning probability of the player with the highest ability (here: Federer) increases strongly compared to the single elimination tournaments. The winning logits of all other players (except the second strongest player Murray) decrease sharply. In general, we can conclude that a single elimination tournament is clearly more exciting than a round-robin tournament. Whereas in a round-robin with 128 players each player has to play 127 matches, in a single elimination tournament the final participants have to play seven matches. Nevertheless,

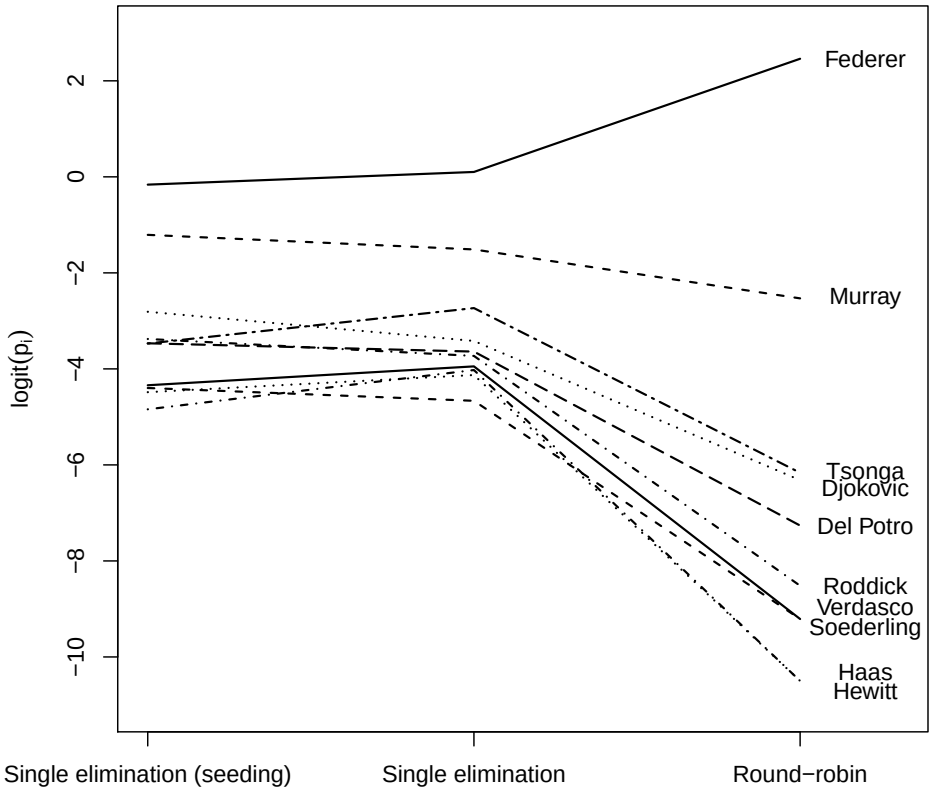



Figure 4.3: Winning probabilities of the top ten players simulated by three different tournament designs (single elimination tournament with seeding, single elimination tournament without seeding and a round-robin tournament) using the estimated abilities of all 128 participating players of Wimbledon 2009.

a round-robin tournament would be more favorable to top players.

### 4.3.3 Discussion

In this application we investigate a strategy for estimating the expected players' abilities of a tennis tournament (Wimbledon 2009) using bookmakers ex-

pectancies for winning the tournament. A comparison of the estimated abilities for two datasets incorporating different information about the (expected) participants of the tournament shows that the bookmakers do not react appropriately on a rapid change of the tournament (here: Nadal's withdrawal). The abilities of the main competitors (Federer and Murray) increase. We also investigate the effect of the tournament schedule on top players' chances of winning the tournament by a simulation study, comparing three different tournament designs.

## 4.4 UEFA Champions League 2008/09

In the next application we extend the general model framework for bookmakers odds stemming from a variety of bookmakers and model the bookmakers consensus as well as the (dis)agreement across the bookmakers for the UEFA Champions League 2008/09.

This section is organized as follows: Section 4.4.1 provides a tournament and data description for the UEFA Champions League 2008/09 for which the bookmakers consensus and agreement are modeled in Section 4.4.2 and analyzed in Section 4.4.3. Section 4.4.4 concludes the paper.

### 4.4.1 UEFA Champions League 2008/09: Tournament and data description

#### The tournament

The UEFA Champions League is the most prestigious club competition of the Union of European Football Associations (UEFA) and so one of the most popular annual sports tournaments all over the world. Every year, a selection of European football clubs compete in a multi-stage modus (qualification, group, and knockout stage) to determine the "best" European team. First, 32 teams are determined via three qualification rounds for the group stage

and drawn into eight groups (A–H). The number of eligible teams is determined by UEFA's Coefficient Ranking System for its member associations (see below, for more details). In the 2008/09 season, teams from 17 associations out of UEFA's 53 members qualified for the group stage. The four teams of each group play a round-robin—every team plays every other team twice (one home and one away match), for a total of twelve games within the group—and the group winners and runners-up advance to the knockout stages. In the knock-out stage, each round pairings are determined by means of a draw and played under the cup (knock-out) system, on a home-and-away basis, where the winners advance to the next round until two teams remain. The two teams play the final as one single match at a neutral venue yielding the winner of the UEFA Champions League (Union of European Football Associations, 2009b).

### Data

**Bookmakers odds.** Long-term odds of winning the UEFA Champions League 2008/09 were obtained from the websites of 31 international bookmakers for all 32 participating teams on 2008-09-01 (before the tournament started, but after the group draw). The 31 bookmakers are all bookmakers who offer odds for this event out of 50 European top-selling online sports bookmakers. The odds are again adjusted by removing the stake and a bookmaker-specific overround (see Equation 4.1) and then transformed to winning probabilities by means of Equation 4.2. This yields tournament winning probabilities $p_{i,b}$ for $i = 1, \ldots, 16$ teams and $b = 1, \ldots, 32$ bookmakers which reflect the bookmakers' beliefs about the outcome of the UEFA Champions League 2008/09. Figure 4.4 shows the quoted odds (on a log-axis) for all 32 participating teams of the UEFA Champions League 2008/09 by the 31 bookmakers. It can be seen that the heterogeneity increases along with the level of the quotes odds.

For our dataset we obtain a mean overround of 23.58% across all bookmakers with an interquartile range from 19.71% to 26.89%.

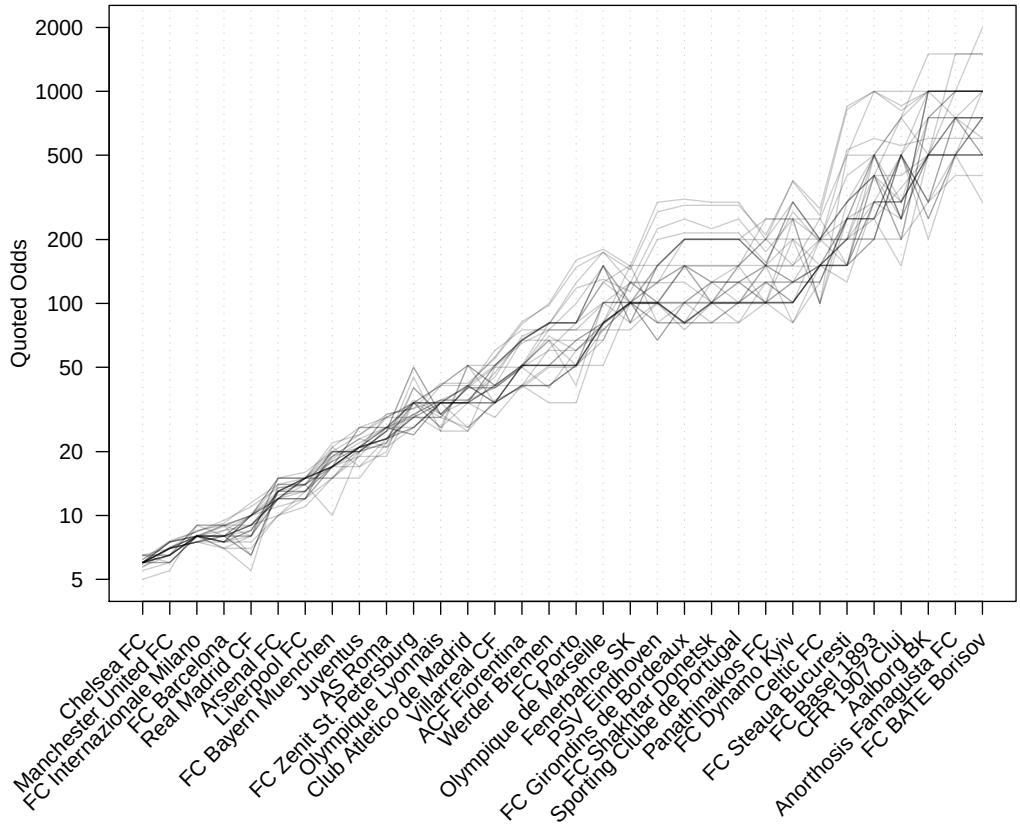



Figure 4.4: Quoted odds (on log-axis) for all 32 participating teams of the UEFA Champions League 2008/09 by the 31 bookmakers.

**UEFA's club coefficient and seeding.** The UEFA also announces their expectancies for the tournament outcome prior to the tournament by publishing a group draw seeding which is a ranking that is very similar to the ranking of UEFA's club coefficient of the teams. The UEFA's club coefficient is determined by the results of a club in European club competitions in the last five seasons, and the league coefficient. The latter is also used to determine the number of eligible teams for the UEFA Champions League where the best three associations have four teams in the tournament (for more details see Union of European Football Associations, 2009b). We obtained the UEFA's club coefficient and seeding for the group draw on 2008-08-28 from

UEFA's website for all 32 participating teams and, in Section 4.4.3, compare both to the ranking derived from the bookmakers' consensus forecast.

### 4.4.2 Modeling consensus and agreement

#### Model class

In order to model the expected winning probabilities $p_{i,b}$ for each team $i = 1, \ldots, 32$ and bookmaker $b = 1, \ldots, 31$, as derived from the raw quoted odds, straightforward linear models are not appropriate as the $p_{i,b}$ necessarily lie within the unit interval. Therefore, we follow the standard technique of employing a suitable link function to transform probabilities to the real line and then using linear models for the transformed data. Various link functions are conceivable; standard choices include the logit or probit link function. In the following, we employ the logit link throughout; using the probit link instead would lead to qualitatively similar results.

On the transformed logit scale, an intuitive and straightforward strategy would be to compute team-wise means for the consensus and team-wise standard deviations for the disagreement across bookmakers (as suggested by, e.g., Zarnowitz and Lambros, 1987). In our application, this simple strategy might be appropriate because we could expect the teams to be sufficiently different and the bookmakers to have rather similar information about the teams. However, from a statistical point of view one should investigate whether this simple strategy is sufficient or can be improved by including additional effects (e.g., pertaining to the bookmakers), or by using a more parsimonious parametrization still giving a good approximation of the underlying data-generating process. Therefore, we propose a stochastic model class that captures the underlying probability distribution on a logit scale and contains the simple strategy as a special case. We assume additive and normally distributed "errors" on the logit scale, providing a natural way for assessment of means and variances in the models.

The model relates the expected winning logits $\text{logit}(p_{i,b})$ to the (unobservable)

"true" winning logits $\text{logit}(p_i)$ for team $i$, reflecting the bookmakers consensus, plus an additional (unobservable) normally-distributed error term $\epsilon_{i,b}$ of bookmaker $b$ for team $i$, reflecting the disagreement across the bookmakers. In order to capture these latent quantities by a linear mixed-effects model, we allow the true winning logits to depend on a team effect $\alpha_i$, an association effect $\beta_{a(i)}$ for association $a$ of team $i$, as well as an overall intercept $\nu$. The error can additionally depend on $\mu_b$, the mean effect of bookmaker $b$. We allow also different specifications of the deviation $\epsilon_{i,b}$ of bookmaker $b$ for team $i$. In summary, this can be written as

$$\begin{aligned} \text{logit}(p_{i,b}) &= \text{logit}(p_i) + \epsilon_{i,b} && (4.16) \\ &= \nu + \alpha_i + \beta_{a(i)} + \mu_b + \sigma_{i,b} Z_{i,b}, && (4.17) \end{aligned}$$

where $Z_{i,b}$ is a standardized error and $\sigma_{i,b}$ is the standard deviation which can be either constant ($\sigma_{i,b} = \sigma$), or constant within a specific group ($\sigma_{i,b} = \sigma_i$: team-specific standard deviation; $\sigma_{i,b} = \sigma_b$: bookmaker-specific; or $\sigma_{i,b} = \sigma_{a(i)}$: association-specific). Even if contrasts are employed, this model is overspecified when all three effects $\alpha_i$, $\beta_{a(i)}$, and $\mu_b$ are included as fixed effects due to the dependence of association $a(i)$ on the team $i$.

In order to overcome this methodological issue, there are various conceivable solutions which can also be motivated by subject-matter considerations: (a) The association effects could be omitted signalling that all teams are sufficiently different. Note that the full team effect then still captures association differences. (b) Alternatively, the team effect could be specified as a random effect (with zero mean) conveying that the winning logits for each team deviate randomly from the mean as captured by the remaining effects (e.g., by fixed association differences). (c) A random effect for the bookmakers would be conceivable implying that the bookmakers' odds deviate randomly from the mean as captured by the remaining effects. (d) Finally, the four different specifications of the deviation $\epsilon_{i,b}$ of bookmaker $b$ for team $i$ represent different views on the sources of variation and thus disagreement. For example, even if there is a fixed team effect $\alpha_i$ in the consensus, it would be conceivable that the amount of disagreement is only driven by the team's association be-

cause bookmakers might have a similar degree of information about teams in the same association. Combinations of the ideas (a)–(d) lead to 20 different mixed-effects models. Table 4.6 specifies the different effects and standard deviations of $\epsilon_{i,b}$ for each model. In order to find a parsimonious model which still gives a good approximation of the underlying data-generating process, standard model selection methods can be employed. We use the Bayesian information criterion (BIC; Pinheiro and Bates, 2000).

### Model selection

Fitting the 20 conceivable mixed-effects models discussed in the previous sections yields the results in Table 4.6 which provides the log-likelihood, number of parameters, and associated BIC. In general, the model selection approach shows that all models including fixed team effects perform clearly better than models with a random team effect, even if an additional association effect is included. Furthermore, the models with constant standard deviation are worse than all models using other standard deviation specifications. With respect to the BIC, the best model emerging from Models 1–20 is Model 3 (BIC = 82.13), containing only a fixed team effect (and hence no additional association) and a team-specific standard deviation. The second best model (Model 7) includes an additional random effect for the bookmakers, capturing bookmaker differences. The best four models (Models 3, 4, 7, and 8) have a fixed team effect and a team- or association-specific standard deviation. In summary, this conveys that, as expected, the main differences are across individual teams which require a full fixed effect (and can not be sufficiently captured by more parsimonious parametrizations such as a fixed association effect plus a random team effect). Furthermore, the fact that the bookmaker effect can be omitted or captured as a random effect suggests that there are no large systematic deviations between the bookmakers. Similarly, a team-specific standard deviation is necessary to obtain the best model fit. However, models including association-specific standard deviations are only slightly worse, implying that agreement across bookmakers is driven to a large extent by the association differences.

| | Team | Bookmaker | Association | Deviation | logLik | df | BIC |
|---|---|---|---|---|---|---|---|
| 1 | fixed | fixed | none | const | −3.20 | 63 | 441.09 |
| 2 | fixed | none | none | const | −121.71 | 33 | 471.11 |
| 3 | fixed | none | none | team | 179.73 | 64 | 82.13 |
| 4 | fixed | none | none | association | 121.48 | 49 | 95.12 |
| 5 | fixed | random | none | const | −51.88 | 34 | 338.34 |
| 6 | fixed | random | none | bookmaker | 12.61 | 64 | 416.37 |
| 7 | fixed | random | none | team | 179.73 | 65 | 89.03 |
| 8 | fixed | random | none | association | 121.63 | 50 | 101.72 |
| 9 | random | fixed | none | const | −130.99 | 33 | 489.68 |
| 10 | random | fixed | fixed | const | −96.30 | 49 | 530.69 |
| 11 | random | fixed | none | bookmaker | −69.91 | 63 | 574.51 |
| 12 | random | fixed | fixed | bookmaker | −35.35 | 79 | 615.78 |
| 13 | random | fixed | none | team | 59.08 | 64 | 323.41 |
| 14 | random | fixed | fixed | team | 93.68 | 80 | 364.62 |
| 15 | random | fixed | none | association | 12.88 | 49 | 312.32 |
| 16 | random | fixed | fixed | association | 47.49 | 65 | 353.50 |
| 17 | random | none | none | const | −245.68 | 3 | 512.05 |
| 18 | random | none | none | bookmaker | −163.39 | 33 | 554.47 |
| 19 | random | none | none | team | 46.04 | 34 | 142.51 |
| 20 | random | none | none | association | −10.33 | 19 | 151.75 |
| 21 | fixed | none | none | linear | 83.35 | 34 | 67.88 |
| 22 | fixed | none | none | power | 113.47 | 35 | 14.56 |

Table 4.6: Effect and standard deviation specifications of the mixed-effects models for $\text{logit}(p_{i,b})$ of team $i$ by bookmaker $b$. Each model is evaluated by the log-likelihood value (logLik), the number of estimated parameters (df), and the BIC.

Model 3 confirms the simple strategy of employing team-specific means for the consensus and team-specific standard deviations for agreement across bookmakers. It is reassuring that this intuitive model has been selected from a more general class of models, where some of the alternatives would have also had appealing interpretations. In Section 4.4.3 it is shown how the parametrization of the standard deviation can be made more parsimonious while retaining the same consensus (Models 21 and 22 of Table 4.6).

### 4.4.3 Analysis of the UEFA Champions League 2008/09

#### Consensus

The bookmaker consensus for the UEFA Champions League 2008/09 can be derived from the best model (Model 3) by using the estimated winning logits $\text{logit}(\widehat{p}_i) = \widehat{\nu} + \widehat{\alpha}_i$ which equal the team-specific means of the winning logits across the bookmakers for each team ($= 1/31 \sum_{b=1}^{31} \text{logit}(p_{i,b})$). This consensus information on the logit scale can easily be transformed to the associated winning probabilities $\widehat{p}_i$ of winning the tournament for all 32 participating teams which are shown in Table 4.7. Additionally, the eight origin groups of the preliminaries, and the football association of the teams are shown.

Chelsea FC is seen as the best team of the 32 teams and has the highest probability (13.52%) of winning the tournament. The expected runner-up (if the tournament schedule allows such a final) comes also from England, Manchester United FC (winning probability: 12.00%). The top two are followed by the champion of the "Serie A" FC Internazionale Milano (10.10%) and the champion of the "Primera Division" FC Barcelona (10.05%). The last four teams are participating for the first time in the tournament and have just a winning probability of 0.20% or less. Four teams out of the first seven ranked teams are from England which implies that England is the strongest European association. Three teams out of the first eleven are members of group H, but only two of them can advance to the next round. Using the group information in combination with the winning probabilities of the participating teams (Table 4.7) the following 16 teams (eight group-winners and eight runners-up) are expected to play the first knock-out round: Chelsea FC, AS Roma (group A), FC Internazionale Milano, Werder Bremen (B), FC Barcelona, FC Shakhtar Donetsk (C), Liverpool FC, Club Atlético de Madrid (D), Manchester United FC, Villarreal GF (E), FC Bayern München, Olympique Lyonnais (F), Arsenal FC, FC Porto (G), Real Madrid CF, and Juventus (H). These 16 teams are not the 16 participants with the highest winning probabilities implying that the group drawn has an effect to the

tournament outcome. In summary, the bookmaker consensus gives winning probabilities of the teams which can be used to predict the winner of the tournament. See Section 4.2.1 on how this forecast can be complemented for dynamics of such tournaments by a simulation approach.

In order to show how well the bookmaker consensus performs in practice, we compare the forecast with the real outcome of the UEFA Champions League 2008/09. Table 4.8 assesses the predictive performance of the bookmaker consensus by comparing them with the actual tournament outcome using Spearman's rank correlation. For the actual results, a total ranking including ties is employed, as commonly used in rankings of such incomplete tournaments. Various strategies for resolving the ties have been explored but did not lead to qualitatively different results. In addition, Table 4.8 also provides correlations with the ranking implied by the UEFA's seeding and UEFA's club coefficient of the teams (prior to the group drawn).

This shows that the bookmakers consensus has a very high correlation with the actual outcome (0.798) and performs somewhat better than the rankings based on the UEFA's seeding (0.756) and UEFA's club coefficient (0.754) of the teams. In particular, the bookmaker consensus correctly predicts three of four semifinalists (Chelsea FC, Manchester United, FC Barcelona) and 14 of 16 teams which played the first knockout round.

**Agreement**

In addition to the consensus of the bookmaker we can also derive the team-specific standard deviations of Model 3. As discussed above, the estimated standard deviations $\widehat{\sigma}_i$ captures the disagreement across the bookmakers. A low standard deviation for a team reflects a low disagreement across the bookmakers, whereas a high standard deviation implies a high disagreement across the bookmakers. The standard deviations $\sigma_i$ for team $i$ for all 32 participating teams are shown in Table 4.7.

In general, the team-specific standard deviations are low implying a low disagreement across the 31 bookmakers. The team with the lowest disagreement

across the bookmakers is one of the top teams, FC Barcelona, with a standard deviation of 0.065 on the logit scale. Conversely, the team with the highest disagreement (standard deviation 0.494) is Aalborg BK which has a low consensus winning probability. Taking a closer look (see Figure 4.5), we can see that the agreement increases with increasing winning logits of the teams. By exploiting this information, our current best model (Model 3) can be improved further by fitting a relationship between the team-specific standard deviations and the fitted values on the logit scale:

$$\sigma_{i,b} = \sigma_i = \gamma_1 + \gamma_2 \text{logit}\left(p_i\right)^{\gamma_3}, \tag{4.18}$$

where $\gamma_1$, $\gamma_2$, and $\gamma_3$ are the function parameters which are estimated by the model (jointly along with the parameters specifying the consensus $\text{logit}(p_i)$).

In addition to the power specification above we also investigate a linear specification ($\gamma_3 = 1$). By using a linear relationship a much more parsimonious model, reducing the number of estimated parameters from 64 $(32 + 32)$ to 34 $(32 + 2)$ and improving the model selection criteria (BIC = 67.88) can be fitted (see Model 21 of Table 4.6). The estimated function parameters of the linear relationship are: $\gamma_1 = 0.000$ and $\gamma_2 = 0.055$. By estimating one more model parameter for the power $\gamma_3$ of a non-linear relationship the model can be improved again yielding a BIC of 14.56 (see Model 22 of Table 4.6). The estimated function parameters of the non-linear relationship are: $\gamma_1 = 0.065$, $\gamma_2 = 0.005$, and $\gamma_3 = 2.375$. Figure 4.5 shows the team-specific relationship of Model 3 (points), as well as the linear relationship of Model 21 (dashed line) and the non-linear relationship of Model 22 (solid line). Note that in all three models (Models 3, 21, and 22) all parameters are estimated simultaneously yielding the same estimated bookmaker consensus, but different specifications of disagreement across the bookmakers.

### Team's association

According to the bookmaker consensus (Table 4.7) four teams out of the first seven ranked teams are from England which implies that England is

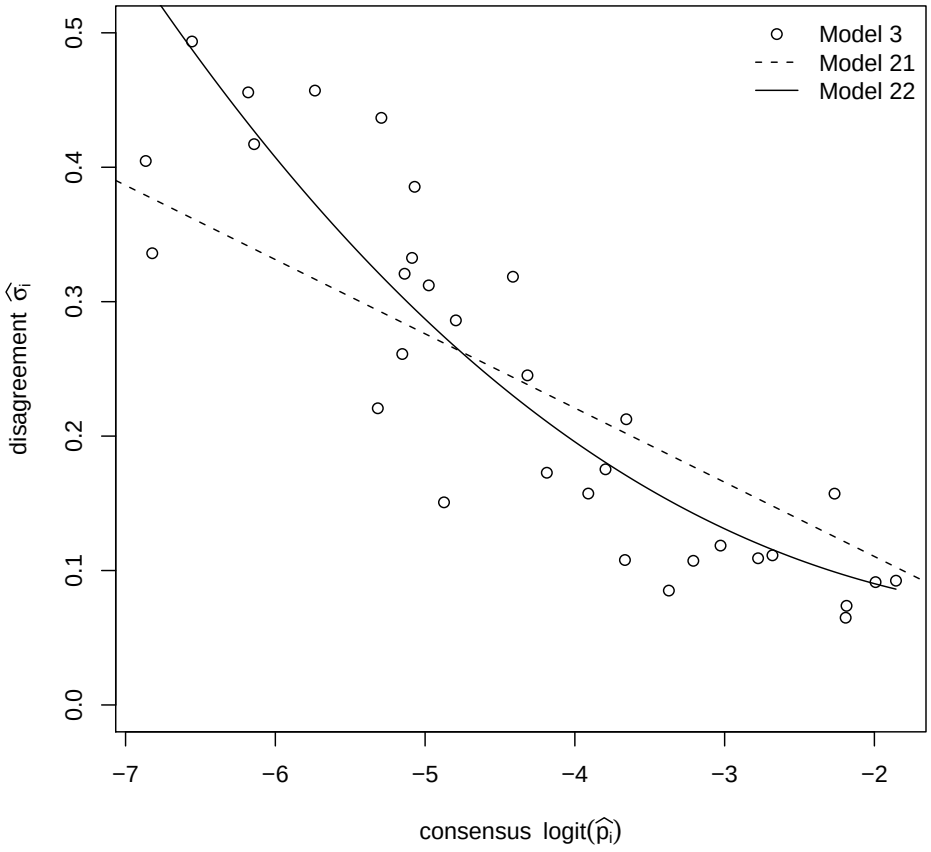



Figure 4.5: Relationship between the estimated bookmaker consensus $\text{logit}(\widehat{p}_i)$ and different specifications of disagreement $\widehat{\sigma}_i$ for all 32 participating teams of the UEFA Champions League 2008/09. The points show the team-specific, the dashed line the linear and the solid line the non-linear relationship captured by the Models 3, 21 and 22 of Table 4.6.

the strongest European association. But what about the other associations? The estimated consensus can also be used to rank the 17 associations of the participating teams. Therefore, we compute the means of the winning logits $\text{logit}(\widehat{p}_i)$ of all teams coming from an association $a$ (see Table 4.9). The difference of these means and the overall mean $\nu$ of all 32 participating teams can be seen as an implied "association effect" on the logit scale. In addition

to the average consensus of an association, Table 4.9 shows the average disagreement (average standard deviations) and the number of qualified teams of the 17 associations.

There is a strong correlation between the average consensus on the logit scale and the number of qualified teams (0.75) implying that strong associations according to the bookmakers consensus have a higher number of qualified teams (cf., UEFA's determination strategy for the number of eligible teams in Union of European Football Associations, 2009b). England, Spain and Italy have the maximum number of qualified teams (four), but England with the highest average consensus on the logit scale ($-2.33$) is the strongest European association. Russia with only one team (FC Zenit St. Petersburg) is rated better than Germany (two teams), France (three teams) and Portugal (two teams). The association with the weakest (average) consensus is clearly Belarus where the team with the lowest probability of winning the Champions League (FC BATE Borisov) comes from.

In addition to the relationship between the association effects and the number of qualified teams, we can also show the relationship between the agreement of the teams and their associations. Table 4.9 shows that the disagreement across the 31 bookmakers is very low for the teams coming from the top three associations (England, Spain and Italy) and increases with the increasing average consensus.

### 4.4.4 Discussion

Based on quoted bookmakers odds for the occurrences of a certain set of events (such as players/teams winning a particular sports match/tournament), this section extends the general model class for the unknown "true" logits of the occurrence of the events. It is applied to the assessment of consensus and (dis)agreement among 31 international bookmakers for the UEFA Champions League 2008/09. A linear mixed-effects model framework capturing different effects for the teams, the bookmakers as well as for the team's associations and allowing different specifications for

the standard deviation leads to a variety of models. According to a model selection approach using the BIC, the natural strategy of using the means of the winning logits as consensus and the team-specific standard deviation as measure for disagreement is appropriate. The estimated winning probabilities derived from the bookmaker consensus predicts the actual outcome very well (correlation of 0.798), somewhat better than UEFA's expectations (UEFA's seeding and UEFA's club coefficient). In particular, the bookmaker consensus model correctly predict three of four semifinalists (Chelsea FC, Manchester United FC, FC Barcelona) and 14 of 16 teams which played the first knockout round. Furthermore, the analysis of the bookmakers agreement implies a negative relationship between the estimated winning probabilities of a team and the disagreement across the bookmakers which can be modeled by a linear relationship or a non-linear relationship. Both extended models capturing these relationships reduce the number of estimated parameters of the model substantially and improve the model selection criteria. By analyzing the team's associations, we show a strong positive relationship between the number of teams coming from an association and the average consensus of the respective association. This reflects UEFA's strategy of allocating more fixed and qualifying slots to "stronger" associations. Finally, we find a strong negative relationship between the disagreement across the bookmakers and the average consensus of an association.

| | $\widehat{p}_i(\%)$ | $\text{logit}(\widehat{p}_i)$ | $\widehat{\sigma}_i$ | Group | Association |
|---|---|---|---|---|---|
| Chelsea FC | 13.52 | −1.86 | 0.092 | A | England |
| Manchester United FC | 12.00 | −1.99 | 0.091 | E | England |
| FC Internazionale Milano | 10.10 | −2.19 | 0.074 | B | Italy |
| FC Barcelona | 10.05 | −2.19 | 0.065 | C | Spain |
| Real Madrid CF | 9.40 | −2.27 | 0.157 | H | Spain |
| Arsenal FC | 6.41 | −2.68 | 0.111 | G | England |
| Liverpool FC | 5.86 | −2.78 | 0.109 | D | England |
| FC Bayern München | 4.62 | −3.03 | 0.119 | F | Germany |
| Juventus | 3.88 | −3.21 | 0.107 | H | Italy |
| AS Roma | 3.32 | −3.37 | 0.085 | A | Italy |
| FC Zenit St. Petersburg | 2.52 | −3.66 | 0.213 | H | Russia |
| Olympique Lyonnais | 2.49 | −3.67 | 0.108 | F | France |
| Club Atlético de Madrid | 2.20 | −3.80 | 0.175 | D | Spain |
| Villarreal CF | 1.96 | −3.91 | 0.157 | E | Spain |
| ACF Fiorentina | 1.50 | −4.19 | 0.173 | F | Italy |
| Werder Bremen | 1.32 | −4.32 | 0.245 | B | Germany |
| FC Porto | 1.20 | −4.41 | 0.319 | G | Portugal |
| Olympique de Marseille | 0.82 | −4.79 | 0.286 | D | France |
| Fenerbahçe SK | 0.76 | −4.87 | 0.151 | G | Turkey |
| PSV Eindhoven | 0.69 | −4.98 | 0.312 | D | Netherlands |
| FC Girondins de Bordeaux | 0.62 | −5.07 | 0.385 | A | France |
| FC Shakhtar Donetsk | 0.61 | −5.09 | 0.333 | C | Ukraine |
| Sporting Clube de Portugal | 0.58 | −5.14 | 0.321 | C | Portugal |
| Panathinaikos FC | 0.58 | −5.15 | 0.261 | B | Greece |
| FC Dynamo Kyiv | 0.50 | −5.29 | 0.437 | G | Ukraine |
| Celtic FC | 0.49 | −5.32 | 0.221 | E | Scotland |
| FC Steaua Bucureşti | 0.32 | −5.74 | 0.457 | F | Romania |
| FC Basel 1893 | 0.21 | −6.14 | 0.417 | C | Switzerland |
| CFR 1907 Cluj | 0.21 | −6.18 | 0.456 | A | Romania |
| Aalborg BK | 0.14 | −6.56 | 0.494 | E | Denmark |
| Anorthosis Famagusta FC | 0.11 | −6.82 | 0.336 | B | Cyprus |
| FC BATE Borisov | 0.10 | −6.87 | 0.405 | H | Belarus |

Table 4.7: Estimated winning probabilities $\widehat{p}_i$, associated winning logits $\text{logit}(\widehat{p}_i)$ (reflecting the bookmakers consensus), and standard deviations $\widehat{\sigma}_i$ (reflecting the agreement across the bookmakers) for all 32 participating teams of the UEFA Champions League 2008/09. Additionally, the eight origin groups of the preliminaries, and the football association of the teams are shown.

| | Bookmaker | Seeding | Coefficient |
|---|---|---|---|
| Tournament ranking | 0.798 | 0.756 | 0.754 |
| Bookmaker | | 0.843 | 0.841 |
| Seeding | | | 0.996 |

Table 4.8: Spearman's rank correlation between the actual tournament ranking, the ranking of the bookmaker consensus, the UEFA's seeding and the UEFA's club coefficient of the 32 participating teams.

| | No. of teams | Av. consensus | Av. disagreement |
|---|---|---|---|
| England | 4 | −2.33 | 0.101 |
| Spain | 4 | −3.04 | 0.139 |
| Italy | 4 | −3.24 | 0.110 |
| Russia | 1 | −3.66 | 0.213 |
| Germany | 2 | −3.67 | 0.182 |
| France | 3 | −4.51 | 0.260 |
| Portugal | 2 | −4.78 | 0.320 |
| Turkey | 1 | −4.87 | 0.151 |
| Netherlands | 1 | −4.98 | 0.312 |
| Greece | 1 | −5.15 | 0.261 |
| Ukraine | 2 | −5.19 | 0.385 |
| Scotland | 1 | −5.32 | 0.221 |
| Romania | 2 | −5.96 | 0.456 |
| Switzerland | 1 | −6.14 | 0.417 |
| Denmark | 1 | −6.56 | 0.494 |
| Cyprus | 1 | −6.82 | 0.336 |
| Belarus | 1 | −6.87 | 0.405 |

Table 4.9: Number of qualified teams, average consensus (in winning logits) and average disagreement (average standard deviation) for the 17 associations of all 32 participating teams of the UEFA Champions League 2008/09.

# Chapter 5

# Conclusion

This dissertation introduces a new general framework modeling common rating processes in order to aggregate rating information stemming from a variety of raters or rating sources. The current literature do not provide a viable strategy to solve this aggregation problem.

In order to model the rating processes, our general model framework is based on the assumption that raters estimate a numerical variable—representing information about the underlying rating subject—in an internal rating process. Due to general informational asymmetry between the rater and the rating subject the rater cannot estimate the "true" numerical variable. So we model the numerical variable as a latent variable. Rating outcomes from different sources are treated as noisy estimations/observations of this latent variable. In order to estimate the latent variable the distribution of the latent variable, the formal relation by which the noise or error terms are linked to the latent variable, and the distribution of the error terms have to be specified. In addition to the latent variable, the means and (co)variances of these distributions are the key outcome of the model. The mean and variance of the rating errors can then be used to validate the underlying rating processes. Furthermore, the estimates of the latent variable denoted as consensus information can be used for forecasting issues.

In order to show the performance of this general model framework, we model

two common rating processes: the credit ratings and the bookmakers odds. In particular, we apply the proposed model to five different applications yielding different model specifications.

The first application investigate a static latent variable model for a multi-rater panel provided by the Austrian central bank where PD estimates are observed for a variety of obligors by 13 banks. In order to estimate the parameters of the distributions of the errors and latent PDs standard maximum likelihood techniques is used. The results of this empirical example show that the framework is suitable to identify bank specific regularities with respect to grouping variables, like industry affiliation, legal form and exposure size, and to conduct an outlier analysis of the estimated rating errors of individual banks.

In the second application we extend the static latent variable model form the first application to a dynamic latent variable model for ordinal rating information. Here, the true unobservable numerical variable is treated as a latent variable and its dynamic is modeled by using systematic (latent market factor) as well as idiosyncratic changes. This model is then used to aggregate the rating information of the bigthree external credit rating agencies (Standard&Poor's, Moody's and Fitch) and to validate their rating behavior. Due to the complexity of this model, Bayesian techniques is employed to estimate the model's parameter. We infer from the different rating biases and standard deviations of the rating errors that there are important differences in the rating systems/rating behavior of the three rating agencies. For the ratings of the iTraxx Europe firms, Standard&Poor's has the smallest absolute rating bias from the consensus. Whereas Moody's clearly seems to be too favorable in its credit assessment, Fitch might exhibit a more conservative rating behavior of these firms. Furthermore, we show that the estimated latent market and the time-dependent mean score of all consensus scores are highly correlated with the Dow Jones EURO STOXX 50 index (used as a reference market).

For the next three applications, the general model framework is applied to bookmakers odds in order to forecast the outcome and analyze the bookmak-

ers agreement of three very popular sport tournaments, the UEFA EURO 2008, Wimbledon 2009, and the UEFA Champions League 2008/09.

For the UEFA EURO 2008, one of the world's biggest sports events that took place in June 2008 in Austria and Switzerland, the quoted bookmakers odds for all 16 participating teams by 45 international bookmakers prior to the tournament are used. The main outcome of the model is the bookmaker consensus which is used to predict the winner of the EURO 2008. The bookmaker consensus is compared to the forecasts from the World Football Elo rating and the ranking implied by the FIFA/Coca Cola World rating. In this ex post comparison, the bookmaker consensus performs best and predicts the correct final (Germany vs. Spain). Furthermore, the results provide many further insights into the effects of the group draw in the tournament, clearly showing that the two finalists come from groups with relatively weak competitors.

In the next application the general model framework for bookmakers odds stemming from a variety of bookmakers is applied to a very popular tennis tournament, the Men's singles of Wimbledon 2009. After showing that the bookmakers odds which are prospective ratings of the participating players' performance perform better, in terms of forecasting the tournament outcome, than the Wimbledon seeding and the ATP ranking, we estimate the abilities of each participating player for two different odds sets. The comparison of the estimated abilities shows that Federer's and Murray's chance of winning Wimbledon 2009 was overestimated by the bookmakers after Nadal's withdrawal. Furthermore, we use all estimated abilities to simulate the outcome of three different tournament designs, showing that in the long run the seeding has not that much influence and a round-robin tournament would be more favorable to top players than the origin single elimination tournament.

In the last application we extend the framework for modeling bookmakers odds to a more general model class. Based on bookmakers odds for the occurrence of a set of events (e.g., players/teams winning a particular match/tournament), a natural strategy for the computation of consensus and (dis)agreement are event-specific means and variances across the differ-

ent bookmakers. The statistical modeling framework outlined above contains this strategy as a special case –namely fixed event effects for both means and variances – but also allows exploration of a wider range of model specifications. For example, potential advantages of random vs. fixed effects can be investigated, or effects pertaining to the bookmaker, grouping effects for the different events, or associations between means and variances can be exploited to specify more parsimonious models. In the application to the UEFA Champions League 2008/09, it can be shown that the straightforward strategy of event-specific means (also used above for the UEFA EUR 2008 and Wimbledon 2009) and variances performs well in a wide range of models. However it can be improved even further when the association between means and variances is incorporated, i.e., when considering that events with higher probability of occurrence also have a higher level of agreement. The resulting bookmaker consensus forecast for the UEFA Champions League 2008/09 performs well in practice, exhibiting a high correlation with the actual tournament outcome.

# Computational details

All computations were carried out in the R system (version 2.8.1) for statistical computing (R Development Core Team, 2009). In particular, the R package nlme version 3.1-90 (Pinheiro et al., 2008) was used for the maximum likelihood estimation of the mixed-effects models (see Pinheiro and Bates, 2000) and the R-package rjags version 1.0.3-5 (Plummer, 2009) was used for the Bayesian estimation of the dynamic latent trait model.

We also employ so-called relationship plots to visualize the estimated rating bias and error variance parameters. These plots use the strucplot framework according to Meyer et al. (2006) and the corresponding strucplot function of the R package vcd (Meyer et al., 2008) to visualize the relationship between measurements of a quantitative variable (here: estimated model parameters) and the interaction of two qualitative factors (here: industry/bank combinations). Each combination of factor levels is represented by a rectangular cell shaded by gray values representing the corresponding measurement values.

# List of Figures

# List of Tables

# Bibliography


Advanced Satellite Consulting Ltd. The World Football Elo Rating System, 2008. URL `http://www.eloratings.net/`. [Online; accessed 2008-04-21].

E. I. Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *Journal of Finance*, 23:189–209, 1968.

E. I. Altman and H. A. Rijken. How rating agencies achive rating stability. *Journal of Banking and Finance*, 28:2679–2714, 2004.

Association of Tennis Professionals. Tennis ATP World Tour, 2009. URL `http://www.atpworldtour.com/`. [Online; accessed 2009-06-22].

Bank for International Settlements. International convergence of capital measurement and capital standards: A revised framework, 2004. URL `http://www.bis.org/publ/bcbs107.htm`.

Bank for International Settlements. Studies on the validation of internal rating systems (revised), 2005. URL `http://www.bis.org/publ/bcbs_wp14.pdf`.

J. Berk and P. DeMarzo. *Corporate Finance*. Statistics and Computing. Pearson International Edition, Boston, USA, 2007. ISBN 0-321-41680-5.

S. T. Bharath and T. Shumway. Forecasting default with the Merton distance to default model. *Review of Financial Studies*, 21(3):1339–1369, 2008.

M. Blume, F. Lim, and A. MacKinlay. The declining credit quality of US corporate debt: Myth or reality. *Journal of Finance*, 53:1389–1413, 1998.

B. L. Boulier and H. O. Stekler. Are sports seedings good predictors?: An evaluation. *International Journal of Forecasting*, 15:83–91, 1999.

B. L. Boulier and H. O. Stekler. Predicting the outcomes of National Football League games. *International Journal of Forecasting*, 19:257–270, 2003.

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39:324–345, 1952.

R. Cantor and F. Packer. Souvereign credit ratings. *Current Issues in Economics and Finance*, 1(3), 1995.

R. Cantor and F. Packer. Differences of opinion and selection bias in the credit rating industry. *Journal of Banking and Finance*, 21:1395–1417, 1997.

M. Carey. Some evidence on the consistency of banks' internal credit ratings. Technical report, Federal Reserve Board, 2001.

M. Carey and M. Hrycay. Parameterizing credit risk model with rating data. *Journal of Banking and Finance*, 25:197–270, 2001.

B. P. Carlin and T. A. Louis. *Bayesian Methods for Data Analysis*. Chapman & Hall/CRC, 2009.

S. R. Clarke and J. M. Norman. Focus on sport – home ground advantage of individual clubs in English soccer. *Journal of the Royal Statistical Society D*, 44(4):509–521, 1995.

M. P. Clements. Consensus and uncertainty: Using forecast probabilities of output declines. *International Journal of Forecasting*, 24:76–86, 2008.

W. D. Cook, M. Kress, and L. M. Seiford. Information and preference in partial orders: A bimatrix representation. *Psychometrika*, 51:197–207, 1986.

W. D. Cook, B. Golany, M. Penn, and T. Raviv. Creating a consensus ranking of proposals from reviewers' partial ordinal rankings. *Computers & Operations Research*, 34:954–965, 2007.

P. Coughlin, N. Bukspan, and D. Wyss. Understanding Standard&Poor's rating definitions. Ratings direct, Standard&Poor's, 2009. URL `http://www.standardandpoors.com/ratingsdirect`.

P. Crosbie and J. Bohn. Modeling default risk. Technical report, Moody's KMV, December 2003.

M. Crouhy, D. Galai, and R. Mark. Prototype risk-rating system. *Journal of Banking and Finance*, 25:47–95, 2001.

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

M. Dixon and P. Pope. The value of statistical forecasts in the UK association football betting market. *International Journal of Forecasting*, 20:697–711, 2004.

D. Duffie and D. Lando. Term structures of credit spreads with incomplete information. *Econometrica*, 69:633–664, 2001.

D. Dyte and S. Clarke. A rating based poisson model for World Cup soccer simulation. *The Journal of the Operational Research Society*, 51(8):993–998, 2000.

A. Edmans, D. García, and O. Norli. Sports sentiment and stock returns. *Journal of Finance*, 62(4):1967–1998, 2007.

A. E. Elo. *The Rating of Chess Players, Past and Present*. Ishi Press, San Rafael, United States, 2008.

H. Elsinger, A. Lehar, and M. Summer. Risk assessment for banking systems. *Management Science*, 52(9):1301–1314, 2006.

K. Emery and S. Ou. Corporate default and recovery rates, 1920–2008. Moody's global credit policy, Moody's, New York, USA, 2009. URL `http://www.moodys.com`.

U. Erlenmaier. *The Basel II Risk Parameters*, chapter IV. The Shadow Rating Approach—Experience from Banking Practice, pages 39–77. Springer Berlin Heidelberg, 2006.

European Commission. Statistical classification of economic activities in the european community, 2008. URL `http://ec.europa.eu/environment/emas/documents/nace_en.htm`.

Fédération Internationale de Football Association. FIFA/Coca-Cola World Ranking, 2008. URL `http://www.fifa.com/`. [Online; accessed 2008-04-21].

D. Forrest and R. Simmons. Forecasting sport: The behaviour and performance of football tipsters. *International Journal of Forecasting*, 16: 317–331, 2000.

D. Forrest, J. Beaumont, J. Goddard, and R. Simmons. Home advantage and the deabate about competitive balance in professional sports leagues. *Journal of Sports Sciences*, 23(4):439–445, 2005a.

D. Forrest, J. Goddard, and R. Simmons. Odds-setters as forecasters: The case of English football. *International Journal of Forecasting*, 21:551–564, 2005b.

J. Fox. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage, London, 1997.

A. Gelman and D. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

J. Goddard and I. Asimakopoulos. Modelling football match results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23:51–66, 2004.

C. Granger and P. Newbold. *Forecasting Economic Time Series*. Academic, New York, 1977.

B. Grün, P. Hofmarcher, K. Hornik, C. Leitner, and S. Pichler. Deriving consensus ratings of the big three rating agencies. Report 99, Institute of

Statistics and Mathematics, WU Wirtschaftsuniversität Wien, Research Report Series, March 2010. URL http://epub.wu.ac.at/.

X. Guo, R. A. Jarrow, and Y. Zeng. Credit risk models with incomplete information. *Mathematics of Operations Research*, 34(2):320–332, 2009.

R. J. Henery. Measures of over-round in performance index betting. *Journal of the Royal Statistical Society D*, 48(3):435–439, 1999.

K. Hornik, R. Jankowitsch, M. Lingo, S. Pichler, and G. Winkler. Validation of credit rating systems using multi-rater information. *Journal of Credit Risk*, 3:3–29, 2007.

K. Hornik, R. Jankowitsch, C. Leitner, M. Lingo, S. Pichler, and G. Winkler. A latent variable approach to validate credit rating systems. *Working Paper*, 2008. URL http://ssrn.com/abstract=1269306.

K. Hornik, R. Jankowitsch, C. Leitner, M. Lingo, S. Pichler, and G. Winkler. A latent variable approach to validate credit rating systems. In D. Rösch and H. Scheule, editors, *Model Risk in Financial Crises*, pages 277–296. Risk Books, London, 2010.

D. Hosmer and S. Lemeshow. *Applied Logistic Regression*. Wiley, New York, 2nd edition, 2000.

P. Hsueh and D. Kidwell. Are two better than one? *Financial Management*, 17:46–53, 1988.

J. Jewell and M. Linvingston. A comparsion of bond ratings from Moody's, S&P and Fitch IBCA. *Financial Markets, Institutions & Instruments*, 8: 1–45, 2002.

H. Joe. Rating systems based on paird comparison models. *Statistics & Probability Letters*, 11:343–347, 1991.

D. Kliger and O. Sarig. The information value of bond ratings. *The Journal of Finance*, 6:2879–2902, 2000.

R. Kolb and H. O. Stekler. Is there a consensus among financial forecasters. *International Journal of Forecasting*, 12:455–464, 1996.

J. P. Krahnen and M. Weber. Generally accepted rating principles: A primer. *Journal of Banking and Finance*, 25:3–23, 2001.

K. Lahiri and C. Teigland. On the normality of probability distributions of inflation and GNP forecasts. *International Journal of Forecasting*, 3: 269–279, 1987.

D. Lando. *Credit Risk Modeling*. Princeton University Press, Princeton, NJ, 1st edition, 2004.

J. H. Lebovic and L. Sigelman. The forecasting accuarcy and determinants of football rankings. *International Journal of Forecasting*, 17:105–120, 2001.

E. L. Lehmann and G. Casella. *BTheory of Point Estimation*. Springer, 2nd edition, 1998. ISBN: 0-387-98502-6.

C. Leitner, A. Zeileis, and K. Hornik. Who is going to win the EURO 2008? (A statistical investigation of bookmakers odds). Report 65, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, May 2008a. URL `http://epub.wu.ac.at/`.

C. Leitner, A. Zeileis, and K. Hornik. Predicting the winner of the EURO 2008 (A statistical investigation of bookmakers odds). Report 76, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, October 2008b. URL `http://epub.wu-wien.ac.at/`.

C. Leitner, A. Zeileis, and K. Hornik. Bookmaker consensus and agreement for the UEFA Champions League 2008/09. Report 88, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, August 2009a. URL `http://epub.wu.ac.at/`.

C. Leitner, A. Zeileis, and K. Hornik. Forecasting the winner of the UEFA Champions League 2008/09. In R. Koning and P. Scarf, editors, *Proceed-*

*ings of the 2nd International Conference on Mathematics in Sport – IMA Sport 2009*, pages 94–99, 2009b. ISBN: 979-0-905091-21-1.

C. Leitner, A. Zeileis, and K. Hornik. Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. Report 94, Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, Research Report Series, November 2009c. URL `http://epub.wu.ac.at/`.

C. Leitner, A. Zeileis, and K. Hornik. Is Federer stronger in a tournament without Nadal? An evaluation of odds and seedings for Wimbledon 2009. *Austrian Journal of Statistics*, 38(4):277–286, 2009d.

C. Leitner, A. Zeileis, and K. Hornik. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the EURO 2008. *International Journal of Forecasting*, 26(3):471–481, 2010a. doi: 10.1016/j.ijforecast.2009.10.001.

C. Leitner, A. Zeileis, and K. Hornik. Bookmaker consensus and agreement for the UEFA Champions League 2008/09. *IMA Journal of Management Mathematics*, 2010b. doi: 10.1093/imaman/DPQ016. Forthcoming.

H. E. Leland and D. H. Pyle. Informational asymmetries, financial structure, and financial intermediation. *Journal of Finance*, 32:371–387, 1977.

M. Lingo and G. Winkler. Discriminatory power: An obsolete validation criterion? *Journal of Risk Model Validation*, 2(1):1–27, 2008.

M. Maher. Modelling association football scores. *Statistica Neerlandica*, 36: 109–118, 1982.

I. McHale and S. Davies. Statistical analysis of the effectiveness of the FIFA World Rankings. In J. Albert and R. H. Koning, editors, *Statistical Thinking in Sports*, pages 77–90. Chapman & Hall/CRC, Boca Raton, Florida, 2007.

A. J. McNeil and J. P. Wendin. Bayesian inference for generalized linear mixed model of portfolio credit risk. *Journal of Empirical Finance*, 14: 131–149, 2007.

R. C. Merton. On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, 29:449–470, 1974.

D. Meyer, A. Zeileis, and K. Hornik. The strucplot framework: Visualizing multi-way contingency tables with vcd. *Journal of Statistical Software*, 17 (3):1–48, 2006. URL `http://www.jstatsoft.org/v17/i03/`.

D. Meyer, A. Zeileis, and K. Hornik. *vcd: Visualizing Categorical Data*, 2008. URL `http://CRAN.R-project.org/package=vcd`. R package version 1.1-1.

C. G. Moon and J. G. Stotsky. Testing the differences between the determinants of Moody's and Standard & Poors' ratings. *Journal of Applied Econometrics*, 8:51–69, 1993.

R. Neagu, S. Keenan, and K. Chalermkraivuth. Internal credit rating systems: Methodology and economic value. *The Journal of Risk Model Validation*, 3(2):11–34, 2009.

C. L. Needham and M. Verde. Fitch ratings global corporate finance 2008 transition and default study. Credit market research, Fitch Ratings, 2009. URL `http://www.fitchratings.com`.

P. Nickell, W. Perraudin, and S. Varotto. Stability of rating transitions. *Journal of Banking and Finance*, 24:203–227, 2000.

J. Pinheiro and D. Bates. *Mixed-Effects Models in S and S-PLUS*. Statistics and Computing. Springer-Verlag, New York, USA, 2000. ISBN 0-387-98957-9.

J. Pinheiro, D. Bates, S. DebRoy, and D. Sarkar. *nlme: Linear and Nonlinear Mixed Effects Models*, 2008. URL `http://CRAN.R-project.org/package=nlme`. R package version 3.1-92.

M. Plummer. *rjags: Bayesian Graphical Models Using MCMC*, 2009. URL `http://mcmc-jags.sourceforge.net`. R package version 1.0.3-5.

M. Plummer, N. Best, K. Cowles, and K. Vines. *coda: Output Analysis and Diagnostics for MCMC*, 2008. URL `http://CRAN.R-project.org/package=coda`. R package version 0.13-3.

P. F. Pope and D. A. Peel. Information, prices and efficiency in fixed-odds betting market. *Economica*, 56:323–341, 1989.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Raiffeisen Zentralbank. RZB-Analyse: Wer wird Fußball-Europameister?, 2008. URL `http://www.rzb.at/`. [Online; accessed 2008-05-09].

S. Rousseau. Regulating credit rating agencies after the financial crisis: The long and winding road toward accountability. Research paper, Capital Markets Institute, Rotman School of Management, University of Toronto, 2009.

P. Scarf and M. Bilbao. The optimal design of sporting contests. Report 320, Salford Business School, Working Paper Series, 2006. URL `http://www.mams.salford.ac.uk/`.

M. Schnader and H. O. Stekler. Do consensus forecasts exist? *International Journal of Forecasting*, 7:165–170, 1991.

C. Song, B. L. Boulier, and H. O. Stekler. The comparative accuracy of judgmental and model forecasts of american football games. *International Journal of Forecasting*, 23:405–413, 2007.

C. Song, B. L. Boulier, and H. O. Stekler. Measuring consensus in binary forecasts: NFL game predictions. *International Journal of Forecasting*, 25: 182–191, 2009.

M. Spann and B. Skiera. Sports forecasting: A comparison of the forecast accuracy of prediciton markets, betting odds and tipsters. *Journal of Forecasting*, 28:55–72, 2009.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model comlexity and fit. *J.R.Statist. Soc.B*, 4:583–639, 2002.

C. Stefanescu, R. Tunaru, and S. Turnbull. The credit rating process and estimation of transition probabilities: A Bayesian approach. *Journal of Empirical Finance*, 16:216–234, 2009.

R. T. Stefani. Survey of the major world sports rating systems. *Journal of Applied Statistics*, 24(6):635–646, 1997.

R. T. Stefani and R. Pollard. Football rating systems for top-level competition: A critical survey. *Journal of Quantitative Analysis in Sports*, 3(3): 1–20, 2007.

R. Stein. Benchmarking default prediction models: Pitfalls and remedies in model validation. Technical Report #020305, Moody's KMV, 2002. URL `http://riskcalc.moodysrms.com/us/research/crm/Validation_Tech_Report_020305.pdf`.

A. Stolper. Regulation of credit rating agencies. *Journal of Banking & Finance*, 33:1266–1273, 2009.

V. Su and J. Su. An evaluation of ASA/NBER business outlook survey forecasts. *Explorations in Economic Research*, 2:588–618, 1975.

K. Suzuki and K. Ohmori. Effectiveness of FIFA/Coca-Cola World Ranking in predicting the results of FIFA World Cup finals. *Football Science*, 5: 18–25, 2008.

UBS Wealth Management Research Switzerland. European champions for 2008 will be..., 2008. URL `http://www.ubs.com/`. [Online; accessed 2008-04-21].

Union of European Football Associations. UEFA Champions League, 2009a. URL http://en.euro2008.uefa.com/tournament/index.html. [Online; accessed 2009-04-23].

Union of European Football Associations. UEFA Champions League, 2009b. URL http://www.uefa.com/competitions/ucl/. [Online; accessed 2009-03-23].

D. Vazza, D. Aurora, and N. Kraemer. 2008 annual global corporate default study and rating transition. Ratings direct, Standard&Poor's, 2009. URL http://www.standardandpoors.com/ratingsdirect.

N. Vlastakis, G. Dotsis, and R. N. Markellos. How efficent is the European football betting market? Evidence from arbitrage and trading strategies. *Journal of Forecasting*, 28:426–444, 2009.

Wimbledon. The Championships, Wimbledon 2009 – Official site, 2009. URL http://www.wimbledon.org/. [Online; accessed 2009-09-25].

V. Zarnowitz and L. A. Lambros. Consensus and uncertainty in economic prediction. *Journal of Political Economy*, 95:561–621, 1987.